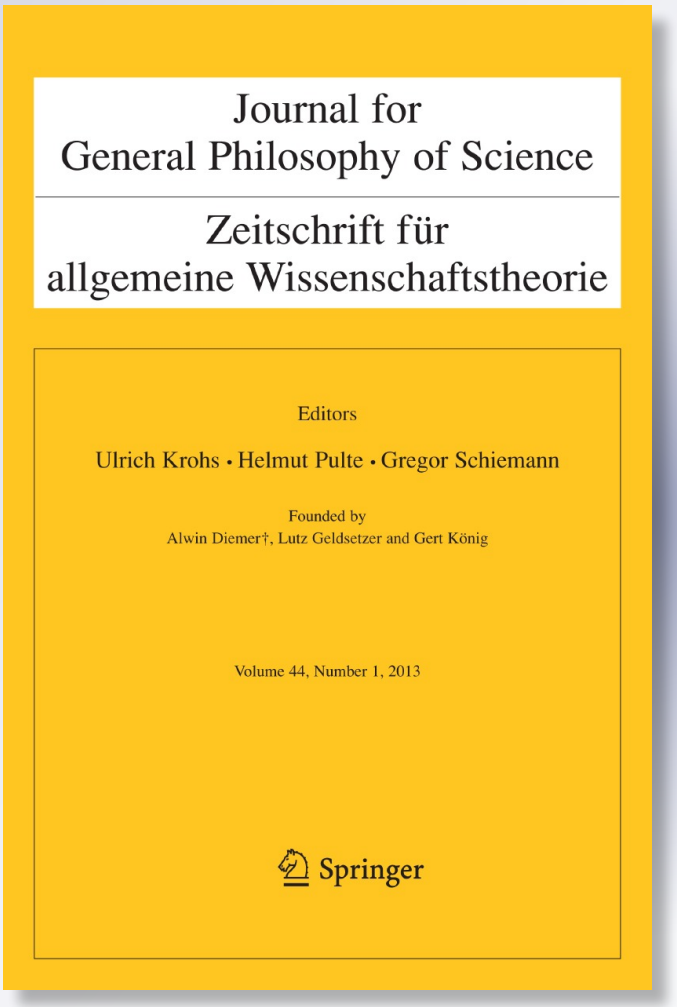
On the Impossibility of Amalgamating Evidence

Aki Lehtinen

Journal for General Philosophy of Science

ISSN 0925-4560
Volume 44
Number 1

J Gen Philos Sci (2013) 44:101-110
DOI 10.1007/s10838-013-9209-5



Springer

J Gen Philos Sci (2013) 44:101–110
DOI 10.1007/s10838-013-9209-5

ARTICLE

On the Impossibility of Amalgamating Evidence

Aki Lehtinen

Published online: 19 March 2013

Abstract It is argued in this paper that amalgamating confirmation from various sources is relevantly different from social-choice contexts, and that proving an impossibility theorem for aggregating confirmation measures directs attention to irrelevant issues.

Keywords Amalgamating evidence · Arrow's impossibility theorem · Interpersonal comparisons · Intensity · Reliability · Relevance · Stegenga

1 Introduction

In a recent paper Jacob Stegenga (forthcoming) posits that amalgamating evidence from different modes[1] into a single index of overall confirmation is analogous to amalgamating individual preferences as in social choice theory, and presents an impossibility theorem in this context. If such an analogy were to hold to a sufficient degree, then an impossibility theorem would mean that amalgamating confirmation measures is fraught with deep epistemic problems. I argue, however, that the analogy fails in some crucial respects, and thus that the apparently gloomy conclusion with respect to confirmation cannot be drawn. Furthermore, since the disanalogies imply that not all of the conditions for an Arrow-like impossibility theorem are normatively acceptable, I argue that the relevance of Stegenga's theorem is practically non-existent.

2 Interpersonal Comparability and Inter-Mode Comparability

Stegenga compares single modes of evidence to individual voters in social choice theory. One of Arrow's (1963) constraints for social welfare functions and thus the whole

[1] A mode is a "particular way of finding out about the world; a type of evidence; a technique or a study design" (Stegenga 2012, 208).

A. Lehtinen (✉)
Department of Political and Economic Studies/Practical Philosophy Unit,
University of Helsinki, P.O. Box 24, 00014 Helsinki, Finland
e-mail: aki.lehtinen@helsinki.fi

social-choice framework was that one could never compare one individual's utility to that of another. Although none of the conditions in Arrow's theorem explicitly rule out interpersonal comparisons, social choice theory eschews them in the sense that none of the conditions incorporates such comparisons. The reasons for eschewing such comparisons were primarily epistemological: it is difficult to measure preference intensities, and even more difficult to obtain reliable information on the individual differences and similarities. Robbins (1952) argues that it is impossible to make a *scientific* judgment about interpersonal comparisons. This argument, in turn, is based on the idea that there are no individual choices that provide information about such comparisons. The meaningful application of (standard) social choice methods presupposes that there are good reasons for disregarding cardinal information and assuming non-comparability. The rationale for using the ordinal social choice framework thus crucially hinges on there being such epistemic problems. If the problem of inter-mode comparability were similar to that involving interpersonal comparability in social choice theory, it would mean that there are similar epistemological difficulties in comparing evidence from different sources or modes.

What would non-comparability mean in the context of aggregating confirmation measures? In the context of social choice theory interpersonal comparisons of utilities can be conceptualized in terms of giving weights to individual utilities in the aggregation function. Similarly, inter-mode comparability means that one can assign weights to the different modes. There seem to be two considerations that should determine the weights: the *reliability* of the different modes in generating the evidence, and the *relevance* of the various pieces of evidence. Relevance, in turn, depends on how closely the evidence is related to the hypothesis (as measured in terms of likelihood in Bayesian confirmation theory, for example). If there is a problem with inter-mode comparability, it must consist in the impossibility of evaluating the reliability and relevance of the various pieces of evidence so that it is impossible to assign weights to the various modes. Stegenga (2009) argues in another paper that scientists lack universal criteria determining the relevance of various pieces of evidence, and that many scientific disagreements can be traced back to different views about the relevance of different kinds of evidence. I agree, but if this is taken as an argument for the *non-comparability* of modes on which the impossibility theorem hinges, it is too weak. We do have some comparability information, and we should use it.

The very idea of constructing an amalgamation function for some hypothesis or theory presupposes that some elementary inter-mode comparisons have already been made. Let me explain. Suppose I should try to evaluate the confirmatory status of two hypotheses H_1 and H_2 by way of amalgamating evidence from various modes. How would I decide which modes of evidence to include in the amalgamation? It seems clear that such a decision requires a judgment concerning the relevance of the various kinds of evidence. However, if I have been able to determine that e_1 and e_2 are relevant, I must already have made the judgment that they ought to have at least *some* non-zero weight in the aggregation function. Then again, this just means that I have already made the inter-mode judgment that e_1 and e_2 have some weight but some others, such as e_3 and e_4, have none. It is true that these inter-mode judgments do not necessarily indicate exactly how to compare the different pieces of evidence deriving from the various sources, but their necessity nevertheless shows that whenever one considers amalgamating evidence for some hypothesis or hypotheses, one always already knows at least something about the relevance of the various pieces, and thus about their inter-mode comparability.

Table 1 Confirmation scores

Mode 1	Mode 2	Mode 3
x (1)	y (1)	z (1)
y (0.9)	z (0.3)	x (0.1)
z (0)	x (0)	y (0)

3 Ordinality

Stegenga provides two arguments for the ordinality of confirmation measures and for the Independence of Irrelevant Alternatives (IIA). Both are borrowed from social choice theory, but he only attempts to apply the second to amalgamating evidence. I will now discuss each argument in turn. The first is that it is not obvious that one might want to include preference intensity information in the aggregation function because this would imply benefiting the fanatics at the expense of the moderates. Whatever the plausibility of the argument in social choice theory,[2] the analogous argument for amalgamating evidence fails. Consider what corresponds to the social-choice notion of intensity of preference in amalgamating evidence. In social choice theory, intensity is conceptualized in terms of a cardinal scale for utilities. Under this measurability assumption it is meaningful to compare utility differences. Let us consider an example (Table 1).

The numbers in parentheses represent confirmation scores obtained from the various modes. It is impossible to use the information contained in those numbers without assuming inter-mode comparability. Let us thus assume that each mode is equally reliable and gets a weight of exactly 1, the aggregate scores are then 1.1, 1.9, and 1.3 for x, y, and z, respectively. Stegenga's argument that one should not benefit the fanatics at the expense of the moderates implies that one should not take the differences in scores into account, and rather declare the confirmation scores equal because the orderings yield a cycle. Under such assumptions I would rather settle for y for the time being because it has the highest score. Assume now that one obtains new information about the reliability of the various modes, and it turns out that mode 3 is extremely reliable whereas modes 1 and 2 are both rather unreliable. Suppose that one gives weights 0.1, 0.1 and 1 for modes 1, 2, and 3, respectively. The aggregate scores are now $0.1 \times 1 + 0.1 \times 0 + 1 \times 0.1 = 0.2$ for x, 0.1 for y, and 1.03 for z. The benefiting-fanatics argument is based on the idea that there is something intrinsically wrong with taking preference intensity into account. However, in the context of amalgamating evidence, 'fanaticism' refers to the reliability of the modes and the intensity of their results, and it seems rather illegitimate not to take them into account.[3] It is true that the available information is not usually as exact as in the example, but if there are various sources of evidence, there is typically at least some knowledge

[2] Arrow does not endorse this argument (see Arrow 1977).

[3] In fact, it puzzles me why Stegenga presents this argument at all, given that he does not seem to think that it could be employed in evidence amalgamation. He writes as follows in another paper: "Another *important desideratum* of evidential assessment is salience, which refers to the strength or intensity of results from a mode, or the impact of a unit of evidence on our credence. For example, when testing the efficacy of a new drug to treat depression, if the symptoms in the treatment group improve by five percent compared to the placebo group, that would be less salient than if the symptoms in the treatment group improve by fifty percent compared to the placebo group" (Stegenga 2012, 221 modified emphases). I posit that 'fanaticism' is to be translated into reliability and intensity of modes because this vague notion may be taken to refer to both *inter*personal and *intra*personal preference intensity. These translations, inter-mode and intra-mode intensity, respectively, should refer to the reliability and intensity ('salience') of modes.

about the reliability with which the various sources provide it. The analogy with social choice thus fails with respect to preference intensity.[4]

Social choice theorists defending the second argument posit that it is impossible to obtain reliable information about interpersonal differences and similarities in preference intensities. Stegenga admits, however, that one sometimes has 'absolute' (this corresponds roughly to 'cardinal and interpersonally comparable' in the social choice context) measures for confirmation, and qualifies his claim to concern cases in which one does not have such measures. He then argues that because one cannot always have absolute measures of confirmation, 'it is better to assume that for particular cases we only have a confirmation ordering' (Stegenga forthcoming, fn. 16). As a reviewer pointed out, evidence may be described in terms of a whole range of intermediate measurability assumptions between absolute measures and orderings: ratio scale and cardinal scale, for example. Stegenga argues, correctly I think, that data expressed in terms of a more precise measurability assumption can always be expressed in terms of the more austere orderings. However, he then concludes that if absolute measures are not available, one should limit one's methods to what is always possible, and it is always possible to have confirmation orderings. Thus, if one does not have absolute measures for all of the modes, one should only use orderings in the amalgamation. The problem with this argument is that it makes little sense to limit one's methods to what is always possible in cases in which any of the modes provides information that is more precise than orderings. There seems to be no good reason to throw away the extra information from the modes that do yield more than orderings. Thus, Stegenga's argument is plausible only in cases in which *all* the different modes have been found only to provide orderings. Stegenga argues in an earlier paper as follows:

> Evidence from different types of experiments is often written in different 'languages'. Petri dishes suggest x, test tubes suggest y, mice suggest z, monkeys suggest 0.8z, mathematical models suggest 2z, clinical experience suggests that sometimes y occurs, and human case control studies suggest y while randomized trials suggest $\sim$y (Stegenga 2009, 654).

In other words, evidence from various sources may well be expressed in terms of different measurability assumptions. Note that there may be literally millions of voters in an election, but it seems unlikely that evidence from more than a dozen modes is ever considered. In such circumstances it makes sense to study whether or not the more precise data should carry more weight than orderings rather than throw away all that valuable information by transforming such data into orderings. In and of itself, I agree with Stegenga that amalgamating evidence is a difficult task for the reasons taken up in the above quotation. It is just that the difficulty has very little to do with the possibility of having cyclic confirmation orderings from various modes, as his impossibility theorem would suggest (more on this below).

[4] In his written response to a version of this paper, Stegenga objected to this argument on the grounds that it presupposes inter-mode comparability, i.e., precisely what he had argued one does not have. I have two responses to this. First, the fanatics argument is based on interpersonal comparability because one cannot even say who is a fanatic and who is a moderate without making interpersonal comparisons. Thus, dismissing my argument on the grounds that it is based on inter-mode comparisons amounts to allowing an appeal to such comparisons in formulating the fanatics argument but denying it in criticizing it. Secondly, although I use exact numerical values for confirmation measures in my counterargument, the argument itself in no way depends on such exact comparability: it is all about what corresponds to fanaticism in the context of amalgamating evidence, and the numerical values merely provide a convenient means of making the point easier to grasp.

It may be instructive to compare Stegenga's treatment of the impossibility theorem to that provided by Okasha (2011) for theory choice. Okasha compares individuals to the theory-choice criteria, and proves an Arrow-like impossibility theorem in this context. He then tries to solve the problem through the application of Amartya Sen's ideas about broadening the informational basis of the aggregation function. As is known from social choice theory (Sen 1970), the impossibility theorem does not disappear given only cardinal utility information, and interpersonal comparability is also needed. If the thorny problem in amalgamating evidence is that the different modes provide evidence expressed in terms of different measurability assumptions, employing Sen's social welfare functional[5] approach is unlikely to help: in social choice the problem would correspond to the idea that different *individual voters* give information under different measurability assumptions. It is a difficult problem that enriching the informational base will not solve because it is always assumed that all the individual transformations affect the functional in some determinate way. In other words, the approach assumes that there are no interpersonal differences in measurability. The only way in which the problem can be treated mechanically is to throw away the extra information—a move Stegenga is willing to employ, but which I have found unacceptable.

Okasha's case is stronger than Stegenga's also because it is reasonable to think that there are never circumstances under which people agree on the weights for the different theory-choice criteria. In contrast, there are cases in which relatively similar evidence from various sources is evaluated. For example, climate researchers study evidence of global warming by way of investigating different simulation models (see e.g., Parker 2006, 2009). Yet, each model provides, among other data, an estimate of the average temperature in the future. Although the exact numerical values differ from one model to another, and even though the data may be rather discordant, comparing the temperatures does not seem to present any particular epistemological difficulties. There are thus circumstances in which one can compare confirmation measures across the modes and evaluate them on an absolute scale. This does not mean that amalgamating evidence is always easy. Indeed, I agree with Okasha's judgment and its analogue in the context of amalgamating confirmation: there is no algorithm for judging the relative importance of the various theory-choice criteria, and no algorithm for judging the relative importance and reliability of different modes of evidence.

Baumann (2005) criticizes the transitivity of the criteria for theory choice on the basis of an example that involves a cyclic ordering 360 Arrow's. The lack of agreement about the weights for the different criteria gives some plausibility to his account, although it is also susceptible to my critique of the need to make a decision about using this or that criterion in the first place. Baumann uses explanatory power, empirical evidence and simplicity, whereas Okasha uses accuracy, consistency, scope, simplicity and fruitfulness. The problem lies in the fact that these two sets of criteria are different for Okasha and Baumann, and this very difference indicates that selecting a particular set of criteria is contestable. Then again, if there is no way of weighing the importance of the various criteria against each other, how is one supposed to decide that some particular criteria are to be accepted as input in the aggregation in the first place? If it is impossible to weigh the criteria, it must be impossible to select them in the first place. Although the difficulty of selecting the relevant modes could be considered an additional problem that afflicts amalgamating evidence, my main point here is that it is a problem for Stegenga, Okasha and Baumann because it shows that the fundamental problem is not the possibility of cyclic profiles[6]: if one is allowed to fiddle freely with

[5] A social welfare functional takes utility functions as its arguments—hence the term 'functional'.

[6] A profile is a list of orderings, one for each individual, mode or theory-choice criterion.

the number of modes or criteria, and if one knows the orderings they provide, it is easy to construct a profile that will make the cycle disappear. This is crucial to the relevance of impossibility theorems because all their proofs rely on there being a cyclic profile. In other words, if the profile is not cyclic, all the conditions are satisfied. All proofs of Arrow's theorem require a cyclic profile such as the one in the Condorcet paradox. This is why the practical relevance of all impossibility theorems depends on there being an actual cycle, even though in principle the theorems apply generally. Hence there is a huge body of literature on the likelihood of cyclic profiles in social choice theory (see e.g., Gehrlein 2006).

Note that it is realistic to assume that whoever considers amalgamating evidence from various sources knows what the orderings are. He or she can go back to them even after having tried to amalgamate the evidence. Elections provide information on individual preferences only once, and although it is usually morally questionable to re-run an election (unless the original was conducted in a non-democratic way), one can always go back to the various modes to see what evidence they give.

It will not do to retort that IIA is precisely the condition that specifies which theory-selection criteria or modes are relevant because Hansson (1973) showed that IIA does not even distinguish between the candidates who are running in the elections and those who never stood. It would be nice, of course, if it did have that property because it could then solve the problem of selecting the modes. But it does not have it. Note that there is no similar problem in the context of social choice: the relevant voters are those who are eligible and find their way to the voting booth. Whoever organizes an election does not have a similar need to use judgment to find out which voters are allowed to participate before letting them register their votes.

4 Under What Conditions Should We Amalgamate?

Employing an aggregation function in the context of social choice theory is justified on the grounds that even if people disagree, one has to make societal decisions. In principle, the situation is similar in the context of amalgamating evidence. For example, policy makers may face a similar predicament. They may need to evaluate evidence from various sources even when the sources give discordant data. The point of using an aggregation function is that it provides a *mechanical* way of making decisions in various kinds of circumstances (see Lagerspetz 2010). This means that the conditions imposed on the functions must be acceptable in all circumstances, not just when the various modes only provide orderings. If my argument that one should use the extra information whenever it is available is valid, it follows that there should be no such thing as a mechanical application of the aggregation function for confirmation because it is more rational to look at the data the various modes provide before making up one's mind about *whether* it makes sense to amalgamate them.

In contrast, aggregation functions in social choice theory are usually thought to represent voting rules. In practice, then, even though one might assume that there will be differences in how much various individuals care about the outcome, aggregation functions usually give equal weight to each individual because most voting rules provide exactly one vote for each voter. Although the one-man-one-vote principle has been justified on the basis of such interpersonal comparability (e.g., Mackie 2003), it could also be justified without invoking interpersonal comparability of utilities, and rather appealing to the equality of voters (e.g., McGann 2006). Note that evidence—at least evidence that could be used in amalgamation—unlike preference is presumably observable. This makes a crucial difference because justifying the idea that each voter has the same number of votes need

not be based on the interpersonal similarity judgment. In the context of amalgamating evidence, however, only the assumption of inter-mode comparability could be taken to justify the equal weights. Modes of evidence do not have particular rights based on equality, and the only reason why one might use an anonymous function is that the relevance and reliability of the different modes is considered the same.

Arrow's framework does not require the equality of voters (or more precisely, anonymity of the social-welfare function), and Stegenga's framework does not require equal weights for the various modes. Yet, the functioning of all *actual* ordinal amalgamation rules depends on this assumption. If one uses the Borda count, for example, or some majoritarian criterion to compute the winning hypothesis from confirmation orderings in various modes, these amalgamation functions presuppose that the weight of each mode is equally important. Then again, if Stegenga claims that his theorem applies to any such actual amalgamation rules, using them re-introduces inter-mode comparability. This would be having one's cake and eating it. There are two possible responses. The first is to allow that Stegenga's theorem does not apply to any such rules, thereby admitting that it is entirely irrelevant. The second is to argue that even though employing such amalgamation functions forces the use of some comparability assumptions, the assumption of equal weights is without proper justification: in other words, it is made just because some comparability assumption must always be made, but if one does not *know* anything about inter-mode comparability one could make this particular assumption for reasons of analytical tractability or simplicity.

This response brings the inter-mode comparability into proper view. If there is any knowledge of the differences and similarities in relevance and reliability of the modes, those differences and similarities should first be evaluated, and only if the resulting judgment is that all the modes are equal in that respect should one consider whether it makes sense to try to amalgamate the information into an aggregate confirmation ordering with some anonymous function. One might also have reason to think that the various modes should not have equal weights. For example, the evidence-based medicine movement tends to regard randomized controlled trials as more reliable than other forms of evidence. In both cases, if there is well-justified information about such differences or similarities between the different modes, then there is information on inter-mode comparability. Now comes the crucial point: if there is no information on inter-mode comparability, what should be done? In my view, one should refrain from amalgamating altogether and thus refrain from using unjustified comparability assumptions, and rather concentrate one's efforts on finding some information on inter-mode comparability. Given what Stegenga (2011) says about meta-analysis, for example, this is what he could have argued in the first place. His impossibility theorem is misleading, however, because it can be taken to presuppose that we *never* have such comparability information, or to presume that one must use a mechanical (and anonymous) amalgamation function even when there is no inter-mode comparability information.

5 The Independence Condition

I will now take a closer look at Stegenga's treatment of the independence condition. Unfortunately, although he defines IIA correctly in Section 4 and uses it correctly in his impossibility proof in the Appendix, his justifying examples in Sections 5.1 and 6.4.1 do not involve Arrow's IIA at all, but rather a *consistency condition* for choice. Choice consistency conditions specify what ought to happen if the set of available alternatives

changes in a specified way.[7] IIA, on the one hand, specifies what happens to social choice between two alternatives if the set of available alternatives remains the same but some preference orderings for alternatives other than the pair in question change.[8] Furthermore, Stegenga's discussion presents IIA as a condition of individual preferences (or modes of evidence), but Arrow's IIA is an inter-profile condition, in other words a condition that specifies what happens to social (rather than individual) choice if two profiles (i.e. collections of individual orderings) agree on a pair of alternatives but some other alternatives have different positions in the two profiles. Whatever plausibility Stegenga's arguments have for such choice-consistency conditions, they cannot be used as arguments for IIA.

The criterion for determining what IIA means is whether or not one can prove an Arrow-like impossibility theorem with it. Given that Sen (1993) proved such a theorem *without* a choice-consistency condition, and given that there is consensus that IIA is necessary for such theorems, discussing choice consistency will not do as an argument for IIA, and he has yet to present a convincing argument for its independence aspect (in the context of amalgamating evidence).

This is what IIA would mean in the context of amalgamating evidence from various sources: for all alternatives in a fixed set, if the two profiles of confirmation orderings agree on a pair of hypotheses, then the social choice for this pair ought to be the same from the two profiles. What it really means is that the input into the aggregation function must be in the form of pair-wise comparisons. The IIA property of binary comparison is the one that is needed for proving the impossibility theorem.[9]

Although seemingly innocuous, IIA is not without problems (see Mackie 2003; Lehtinen 2011 for extended discussions). Quite understandably, Stegenga relies in part on the authority of social choice theorists with respect to the arguments for the crucial IIA condition. This is why I also refer to some arguments as to why present-day theorists no longer unanimously endorse it. As the real content of IIA has become better known, many scholars investigating social choice theory have rejected it. As Saari (1998, and various later publications) argues, the problem with IIA is that it renders useless the imposition of the transitivity of individual orderings: it severs the connection between pairs of alternatives [e.g., (x, y) and (y, z)] such that even though transitivity is explicitly required of the individual orderings, it is impossible to use that information in the aggregation. The result is that an aggregation function that satisfies IIA cannot distinguish between the intransitive confirmation orderings of individual modes of evidence and the intransitivity of the aggregate confirmation ordering.

6 Conclusion

What, then, does Stegenga's impossibility theorem prove? I would put it like this. If there is only information on confirmation orderings from various modes, if it is impossible to make any judgments concerning the relevance and reliability of the various modes, and if the information derived from these various modes is discordant, then it is difficult to say

[7] In that Stegenga's example involves adding new alternatives, the relevant consistency condition is one that deals with 'expansion'. For example, condition β is violated (see e.g., Sen 1986).

[8] Stegenga is definitely not the only author to make the mistake. Mackay (1980) is one of many to confuse the two. Ray (1973) proved the logical independence of these two kinds of conditions, and a useful discussion is to be found in Bordes and Tideman (1991) and Mackie (2003).

[9] Blau (1972) proves an impossibility theorem for m-ary IIA, i.e. for a set of alternatives, but this proof crucially depends on using a choice-consistency condition (for society rather than for individuals) imposed on the social welfare ordering (see Grether and Plott 1982).

what the aggregate confirmation ordering is. It is difficult to think of a context in which such a constellation of assumptions would hold because being able to determine that some particular source is to be accepted as an argument in the aggregation function already indicates at least something about its relevance. If there is no information about inter-mode comparisons, one should refrain from amalgamating altogether. Furthermore, if there is any information beyond confirmation orderings it should not be thrown away. The theorem steers attention away from the real problems of aggregating confirmation measures because it suggests that there is a problem with cyclic profiles. The real problems include the following: the various modes provide information in incommensurate forms (e.g., ordinal, cardinal, intervals, goodness of fit and econometric results versus calibration), and what exactly is being measured in the first place may differ from one mode to another. Given that Stegenga himself has identified such real problems of evidence amalgamation, the general lesson to be learned is that although one may be well be able to prove an analogue for Arrow's theorem in a variety of circumstances, the plausibility of such exercises depends not only on being able to provide some arguments for the analogous conditions. One also needs convincing arguments that the ordinal non-comparative framework as a whole makes sense, and that the conditions of applying the analogues to preference-aggregation functions are satisfied.

What should one conclude if it is agreed that the conditions for an impossibility theorem are not plausible in the context of amalgamating evidence? The standard response in social choice theory is that the conditions have to be weakened or that one must provide more structure to the problem, as in Sen's informational-base approach. One can then prove further impossibility or characterization theorems. I do not think such efforts would make sense in the case of amalgamating evidence for the reason that if different modes provide data with different measurability assumptions it seems impossible to come up with a characterization theorem. Of course, I would be happy to see someone prove me wrong. Finally, if my argument about not throwing away information is tenable, the conclusion should be that we should usually refrain from trying to amalgamate evidence with a mechanical function in the first place, and rather look more closely at the evidence and the procedures that produced it.

Acknowledgments This paper was presented at the PSA 2012 conference in San Diego. Jacob Stegenga wrote an extensive response to a version of this paper. This response and our conversations have clearly made the paper a lot better. Yet, given that I have not been able to respond to all his concerns and some disagreements remain, he should not be held responsible for anything in the paper. I am also grateful to Alessandra Basso, Jaakko Kuorikoski, Petri Ylikoski and two anonymous reviewers for JGPS for their helpful comments on the manuscript. The usual disclaimer applies to these scholars, too.

References

Arrow, K. J. (1977). Current developments in theory of social choice. *Social Research, 44*(4), 607–622.

Arrow, K. J. (1963). *Social Choice and Individual Values* (2nd ed.). New Haven: Yale University Press.

Baumann, P. (2005). Theory choice and the intransitivity of 'Is a better theory than'. *Philosophy of Science, 72*(1), 231–240.

Blau, J. H. (1972). A direct proof of arrow's theorem. *Econometrica, 40*(1), 61–67.

Bordes, G., & Tideman, N. T. (1991). Independence of irrelevant alternatives in the theory of voting. *Theory and Decision, 30*(2), 163–186.

Gehrlein, W. V. (2006). *Condorcet's Paradox*. Heidelberg: Springer.

Grether, D. M., & Plott, C. R. (1982). Nonbinary social choice: An impossibility theorem. *Review of Economic Studies, 49*(1), 143–149.

Hansson, B. (1973). The independence condition in the theory of social choice. *Theory and Decision, 4*, 25–49.
Lagerspetz, E. (2010). Wisdom and numbers. *Social Science Information, 49*(1), 29–59.
Lehtinen, A. (2011). A welfarist critique of social choice theory. *Journal of Theoretical Politics, 23*(3), 359–381.
MacKay, A. F. (1980). *Arrow's theorem, the Paradox of social choice: A case study in the philosophy of economics*. New Haven: Yale University Press.
Mackie, G. (2003). *Democracy Defended*. Cambridge: Cambridge University Press.
McGann, A. (2006). *The logic of democracy: Reconciling equality, deliberation, and minority protection*. Ann Arbor: University of Michigan Press.
Okasha, S. (2011). Theory choice and social choice: Kuhn versus Arrow. *Mind, 120*(477), 83–115.
Parker, W. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Proceedings of the Aristotelian Society, LXXXIII*, 233–249.
Parker, W. (2006). Understanding pluralism in climate modeling. *Foundations of Science, 11*(4), 349–368.
Ray, P. (1973). Independence of irrelevant alternatives. *Econometrica, 41*(5), 987–991.
Robbins, L. (1952). *An essay on the nature & significance of economic science* (2nd ed.). Macmillan: London.
Saari, D. G. (1998). Connecting and resolving Sen's and Arrow's theorems. *Social Choice and Welfare, 15*(2), 239–261.
Sen, A. K. (1993). Internal consistency of choice. *Econometrica, 61*(3), 495–521.
Sen, A. K. (1986). Social choice theory. In K. J. Arrow & M. Intriligator (Eds.), *Handbook of mathematical economics* (pp. 1073–1181). Heidelberg: Elsevier Science publishers B.V.
Sen, A. K. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.
Stegenga, J. (forthcoming). An impossibility theorem for amalgamating evidence. *Synthese*.
Stegenga, J. (2012). Rerum concordia discors: Robustness and discordant multimodal evidence. In L. Soler, E. Trizio, T. Nickles, & W. C. Wimsatt (Eds.), *Characterizing the robustness of science* (pp. 207–226). Springer: London.
Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences, 42*(4), 497–507.
Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science, 76*(5), 650–661.