Found Sci
DOI 10.1007/s10699-014-9374-y

Three Criteria for Consensus Conferences

Jacob Stegenga

Abstract Consensus conferences are social techniques which involve bringing together a group of scientific experts, and sometimes also non-experts, in order to increase the public role in science and related policy, to amalgamate diverse and often contradictory evidence for a hypothesis of interest, and to achieve scientific consensus or at least the appearance of consensus among scientists. For consensus conferences that set out to amalgamate evidence, I propose three desiderata: Inclusivity (the consideration of all available evidence), Constraint (the achievement of some agreement of intersubjective assessments of the hypothesis of interest), and Evidential Complexity (the evaluation of available evidence based on a plurality of relevant evidential criteria). Two examples suggest that consensus conferences can readily satisfy Inclusivity and Evidential Complexity, but consensus conferences do not as easily satisfy Constraint. I end by discussing the relation between social inclusivity and the three desiderata.

Keywords Consensus conferences · Evidence · Evidence amalgamation · Multimodal evidence

1 Introduction

Amalgamating diverse evidence is immensely important for contemporary science. But such practice rests on a shaky methodological foundation. Consider the following example. Cartwright (2006) describes work by the epidemiologist Marmot (2004), who argues that low socioeconomic status is bad for one's health in situations in which such status leads to

Forthcoming in *Foundations of Science*.

J. Stegenga (✉)
Department of Philosophy, University of Utah, 215 South Central Campus Drive, Carolyn Tanner Irish Humanities Building, Salt Lake City, UT 84112, USA
e-mail: jacob.stegenga@utah.edu

J. Stegenga
Department of Philosophy, University of Johannesburg, Johannesburg, Gauteng, South Africa

Published online: 17 October 2014

increased stress and a decreased sense of control. To support this general conclusion Marmot cites "a great deal of evidence of different kinds" (Cartwright 2006). For example, Marmot's own work on British civil servants, based on longitudinal studies over 20 years, showed that the highest paid British civil servants have twice the chance of living until the age of sixty than do the lowest paid British civil servants. Marmot also cites evidence from interviews and surveys on job stress and professional status, evidence from laboratory experiments showing associations between stress and physiological reactions, and evidence from other disciplines altogether, such as primatology and anthropology. Cartwright claims that this case illustrates both the importance and the challenge of amalgamating evidence:

> Altogether, informally, it is an impressive package. Where did he publish it? That helps to make my point—in one of those high-caliber 'semipopular' books. For this is not the kind of thing that goes into a serious journal, and in a sense rightly so. Even review articles in journals tend to cite studies that have a great deal of commonality of language and method—that way they can be adequately policed by the experts in the field. That is just the problem. We have no experts on combining disparate kinds of evidence (apart from some neat metastatistical techniques, which do not stretch very far). But doing so is at the heart of scientific epistemology when that epistemology is directed at establishing results we can use. (Cartwright 2006)

Marmot's appeal to a diverse range of evidence makes his case more compelling than if he had simply cited a single kind of study. Cartwright's concern, however, is that Marmot did not employ a compelling method for amalgamating his diverse evidence.

We often have available a variety of kinds of evidence for a hypothesis of interest. Diverse evidence like that mustered by Marmot for his hypothesis—which comes from a variety of different sources all relevant to a hypothesis of interest—I call *multimodal evidence*; a "mode" is a particular way of finding out about the world: a technique, apparatus, study design, experimental or observational set-up. It is often claimed that if multimodal evidence for a hypothesis is concordant, that hypothesis is robust.[1]

To determine the support that multimodal evidence provides to a hypothesis, such evidence must somehow be amalgamated. Several terms have been coined for the process of amalgamation and the study of amalgamation methods: 'research synthesis', 'mixed research synthesis', 'data fusion', and 'data integration' are terms that have been used for the amalgamation of multimodal evidence; 'multi-method research' and 'mixed methods research' are two terms that have been used to describe research that explicitly sets out to gather multimodal evidence. The term 'systematic review' refers to an amalgamation of a set of diverse evidence according to prospectively specified procedures; for instance, the technique referred to as meta-analysis is one type of systematic review. I will refer to any such method as an 'amalgamation method' (AM). A common AM in contemporary science is the consensus conference.

A consensus conference is a social, deliberative process of amalgamating evidence. Denmark was an early proponent of consensus conferences.[2] The Danish model, initiated in

[1] Robustness-style arguments have been frequently appealed to as grounds for objectivity; concordant multimodal evidence has been seen as a way to avoid worries about the fallibility of single modes of evidence and as a way to resist skeptical arguments. See discussions of robustness (or synonyms) in Wimsatt (1981), Cartwright (1983), Salmon (1984), Culp (1994), Chang (2004), Weber (2005), Kosso (2006), Stegenga (2009), Kuorikoski et al. (2010), and Stegenga (2011a).

[2] The process of bringing together experts in an attempt to resolve disagreement and settle on a fact of the matter is probably as old as organized humanity. One of the more infamous examples of a consensus conference is the 1616 meeting of the commission of theologians, or Qualifiers, who came to a formal consensus that the hypothesis of a moving earth is "foolish and absurd in philosophy" (see Westman 2011).

the 1980s, invites randomly selected citizens to engage with experts on scientific questions that have a technological or policy implication, such as genetic engineering for agriculture, air pollution, or genetic testing for health insurance (Joss and Durant 1995). Evidence from multiple disciplines or professions is considered, and these professions often have conflicting standards of evidence and policy interests. Moreover, features of dispute other than evidence are often considered, including social values that relate to policies based on the hypothesis in question (such as cost, safety, and equity), and other pragmatic issues such as operational feasibility of policies based on the hypothesis and the importance of public perception of the dispute or agreement. However, at least some consensus conferences are primarily attempts to resolve disputes regarding scientific hypotheses, and such disputes are often fueled by discordant multimodal evidence. Section 3 provides two examples of such consensus conferences.

There is little common structure to consensus conferences: variable features include the disciplines represented, the methods of deliberation, the variety of evidence considered, the length of deliberations, and the format of communicating the conclusions. Consensus conferences are relatively unsystematic compared to techniques such as meta-analysis, though consensus conferences can involve structured techniques of deliberation. Moreover, consensus conferences sometimes include formal or quantitative methods, such as meta-analysis, as part of the process of assessing multimodal evidence.[3]

Today consensus conferences are important in the biomedical and social sciences, because these disciplines have many hypotheses for which there is available a massive amount of discordant multimodal evidence. The U.S. National Institute of Health (NIH) was among the first institutions that employed consensus conferences; NIH consensus conferences differ from the Danish model in that they are usually convened to settle on factual conclusions rather than policy guidance.[4] It is this epistemic goal of consensus conferences which is my interest here.[5] Specifically, I am concerned with the use of consensus conferences as a means of determining what the 'weight' of evidence indicates—to use a discomfiting metaphor—for a set of competing hypotheses.

The primary contribution of this paper is to propose three desiderata for assessing AMs, and to employ these desiderata in an evaluation of consensus conferences (§2). These desiderata are *Inclusivity* (an AM should include as much of the available evidence as possible), *Constraint* (an AM should constrain intersubjective assessments of the hypothesis of interest), and *Evidential Complexity* (an AM should assess the available evidence on a plurality of relevant evidential criteria). I present two examples of consensus conferences to illustrate the ways in which these desiderata are variably satisfied (§3). I argue that consensus conferences

[3] However, some advocates of deliberative approaches to amalgamating evidence have been critical of formal methods of evidence amalgamation. A long-time critic of formal amalgamation methods such as meta-analysis has suggested that personal judgment is necessary to properly amalgamate evidence:

> A good review is based on intimate personal knowledge of the field, the participants, the problems that arise, the reputation of different laboratories, the likely trustworthiness of individual scientists, and other partly subjective but extremely relevant considerations. Meta-analysis rules out any such subjective factors. (Eysenck 1994)

For a critical account of meta-analysis, see Stegenga (2011b).

[4] As of 2007 the NIH had produced 118 consensus statements (Solomon 2007).

[5] Not much, I think, should be placed on this distinction. Often the policy implications of an epistemic conclusion are clear to the participants of a consensus conference, it is usually policy makers who organize consensus conferences, and policies themselves involve predictions on some epistemic basis or other. Thus, like the Danish model, the U.S. model is often employed for guidance with policy formulation, albeit perhaps less directly.

readily satisfy Inclusivity and Evidential Complexity, but fare worse on Constraint. In §4 I discuss another property with which one can evaluate consensus conferences—the inclusion of participants representing a broad range of disciplines and interests (*Social Inclusivity*)—and I evaluate the relation between social inclusivity and the three desiderata. I end with a brief comparison with more formal methods of amalgamating evidence (§5).

2 Criteria of Evaluation

Philosophical subtleties aside, truth (or one of its less metaphysically-laden cognates, such as empirical adequacy) is typically thought to be the goal of scientific inquiry. As such, the various methods available to scientists can be evaluated, in principle, by the extent to which they are truth-conducive. However, in certain contexts this way of assessing methods cannot be employed for assessing methods of amalgamating evidence, because the sum total of all available evidence for a hypothesis, appropriately amalgamated, is often thought to be the ultimate epistemic arbiter available to *us*. Put another way: to determine the veracity of an AM for any particular hypothesis, we would need an independent indication that the AM got at the truth about the hypothesis. But this independent indication would be, presumably, just some other kind of evidence about the hypothesis in question, which would then be just more evidence to add to the original evidence in judging the hypothesis in question. The independent indication would itself add to the available multimodal evidence (since it would provide evidence from yet another mode); the new evidence (the evidence thought to be the independent indicator of the veracity of the hypothesis) would have to be amalgamated with the previous evidence by the AM. But now, in order to know that this new output of the AM was true, we would need a new independent indication of the veracity of the AM. Thus truth-conduciveness *simpliciter* is not usually a criterion of evaluation available to us for assessing AMs.[6] However, we can assess AMs on metrics more modest than conduciveness to truth.

We should want our AMs to settle on some constrained range of beliefs in competing hypotheses, and to aid in mitigating disagreement among relevant parties. This could result, for instance, in a narrowing of the range of justified credence for a hypothesis, or a narrowing of the warranted range of estimations of the value of some parameter, or an increase in the number of scientists who roughly share a belief regarding some particular subject. I will call this first desideratum for consensus conferences Constraint (C). If (C) is not achieved by

[6] This problem does not arise in contexts in which there is an independent indicator of the truth. An anonymous reviewer suggests that in situations in which evidence is amalgamated in order to make predictions, we have such an independent indicator of the truth, since the AM can be tested against the frequency with which its predictions are borne out. However, often in the contexts in which AMs are used, the track record of an AM can only be evaluated by appeal to further evidence relevant to the hypothesis in question, and when such new evidence is itself inconclusive (which is ubuitous in such contexts) then the above circular argument applies. For instance, suppose our hypothesis (H) is "drug x alleviates symptoms y", and we use an AM to amalgamate the available evidence regarding H, and then come to affirm H as probable. Further suppose that H warrants a prediction that if x were to be used in clinical practice it would alleviate y. We then use $x$in clinical practice with the hope that it alleviates y. But the evidence regarding H that becomes available from the use of x in clinical practice is only one kind of evidence relevant to H, and indeed such evidence is, in some widely recognized respects, inferior to the initial evidence that was amalgamated by the AM in the first place (because, for example, evidence from clinical practice is not controlled, and is liable to confounding by expectation bias and confirmation bias). So even after the prediction is made based on H and evidence is gathered about the prediction, our epistemic state regarding H is not different in kind than it was prior to the prediction, and the inferior evidence gathered after the prediction cannot be an arbiter of the veracity of the AM.

an AM, then one of the central functions of any scientific method—guidance for belief—is not satisfied. This desideratum is very general and should not be interpreted too strongly. For instance, it might be vanity of rigor to think that an AM could constrain intersubjective assessments of hypotheses to a precision suggested by probabilities. (C) should be construed loosely—intersubjective assessments of competing hypotheses, say, or estimations of the magnitude of a parameter, should be constrained to at least some (unspecified) degree by an AM.

There are plenty of unsatisfactory ways to achieve (C). A consensus conference could stipulate that the most-warranted hypothesis is the one supported by the oldest participant, or the one supported by the person living closest to the island of Kiribati, or the one supported by the highest number of participants who were born in the month of November. A less extreme but still unsatisfactory way to achieve (C) would be to only consider evidence from a single mode: when multimodal evidence is available for a hypothesis, an AM might achieve (C) by only taking into account a single mode of evidence.[7] But surely it is irrational to ignore evidence from other relevant modes, when such evidence is available. An AM should include as much relevant evidence as feasible; I will call this desideratum Inclusivity (I). Evidence can come in many forms, and any relevant evidence ought to be considered when forming our beliefs.[8]

To some this may sound like a platitude—it is akin to Carnap's 'principle of total evidence', often expressed as a norm of inductive reasoning—but it is a controversial claim with respect to the actual practice of science, especially when discordant multimodal evidence is available. A common way of dealing with multimodal evidence is to simply consider evidence from what is considered the mode of highest quality, or the mode closest to the practices of one's own disciplinary background; this is not often explicitly stated or justified, but has occasionally been argued for and codified in standards of evidential assessment. For example, in both the biomedical and social sciences, when conducting systematic reviews, some claim that only evidence from randomized controlled trials (RCTs) should be included.[9] The most general defense that I can discern for the practice of ignoring what is considered to be lower quality modes and including only high quality modes in an AM is that lower quality modes have more inductive risk, and so are more likely to support false conclusions. It is also, perhaps, simpler to consider only evidence from a single mode.

In contrast, one justification for (I) is broadly Bayesian: if one ignores relevant evidence for a hypothesis, one might commit the base-rate fallacy. The possibility of 'defeating' evidence provides further reason why one ought to consider all available evidence. For example, if Beth, a specialist in marsupial physiology, tells me that wombat scat has a conical shape, then I have some evidence that indeed wombat scat is conical; but if I later get evidence that Beth is a compulsive liar then I have lost my reason to believe that wombat scat is conical. Attending to some of my evidence (Beth's initial claim) and ignoring other evidence (about Beth's honesty) leads me to believe something false. To reliably assess the veracity of some hypothesis, then, an AM ought to satisfy (I). Moreover, there is no reason why an AM cannot both satisfy (I) *and* appropriately assess the inductive risk of lower quality modes: evidence from lower quality modes would simply not change our credence in the hypothesis to the same degree that evidence from higher quality modes would.

[7] Even this, though, is overly optimistic: elsewhere I argue that even when assessing a single mode of evidence, constraint is not necessarily achieved, because there are numerous features of evidence that must be assessed, which can be variably (but rationally) prioritized.

[8] A caveat: much hinges on what evidence is deemed 'relevant', and this is often a matter of dispute.

[9] Criticisms of this include Worrall (2002), Cartwright (2007) and Worrall (2007). See also my discussion of the relation between social inclusivity and (I) in §4 below.

Consider the following analogy. Suppose that my vision is blurry, and is not as reliable at informing me about my surroundings as my hearing is, but my vision nevertheless accurately informs me about my surroundings more often than not. If I hear the voice of a colleague in the hall *and* see her in the hall, this should give me more reason to believe that she is in the hall than if I only hear her voice in the hall (or than if I hear her voice in the hall but see her in her office, or than if I hear her voice in the hall but see Gandhi in the hall in the place from where her voice comes). It is nearly a dictate of reason to consider both my hearing and my vision when evidence from both sense modalities is available, and this is true regardless of the blurriness of my vision (under the supposition that my vision is more reliable than merely guessing).[10] Similarly, Inclusivity is a basic requirement for an AM.

There are a variety of factors that should be considered when assessing any single mode of evidence. General features of a mode to be assessed, prior to a consideration of the evidence itself, include the quality of the mode (how free the mode is of systematic biases and methodological errors) and the relevance of the mode to the hypothesis. General features of evidence from a mode to be assessed include the degree to which there are reproducible patterns in the evidence, the degree to which the evidence is concordant with evidence from other modes, and the sheer plausibility of the evidence.[11] Finally, particular disciplines and even particular modes have numerous content-specific features that must be assessed. For example, the epidemiologist Sir Austin Bradford Hill developed a list of nine considerations to help assess whether or not epidemiological data support a particular causal hypothesis, one of which is the presence of a dose-response relationship between the purported cause and the purported effect: for example, if the odds of developing lung cancer are higher for those people that smoke more cigarettes, that suggests (but of course does not prove) a causal relation between smoking and lung cancer. In short, there are a variety of features of evidence that must be assessed, and the associated desideratum for AMs—I will call it Evidential Complexity (E)—is that an AM should somehow take into account the variety of features of evidence.[12]

Consider a hypothetical case of the neglect of (E). Suppose we want to know if a particular chemical has adverse health effects on children. We gather the available evidence from all relevant studies, including cell culture studies, epidemiological studies with a range of designs, and animal experiments; in short, (I) is satisfied. Our chosen AM is straightforward: it merely combines the observed 'effect sizes' from the various studies into an average effect size. From this single effect size, we could assess the hypothesis that this chemical has adverse health effects for children. But the paucity of information on the general features of the modes or the particular features of the evidence would render the output of this AM relatively meaningless. The cell culture studies might have been contaminated. The human studies might have included only adults. The animal experiments may have included only species which are resistant to the chemical or especially sensitive to the chemical. The studies may have had striking patterns in their data which the associated effect sizes did not reveal. In short, (E) would have been neglected by the AM in this hypothetical scenario, and the

[10] If a method provides information that is no more reliable than a randomizer, then such information should not be considered 'evidence'. If two methods are both somewhat reliable but their degrees of reliabilities differ, then evidence from such methods should be weighted accordingly by an AM. Elsewhere I investigate such weighting methodologies for evidence in clinical research.

[11] For examples of the plurality of features of evidence that scientists consider, see, for example, Franklin (2002).

[12] (E) is a kind of epistemic inclusiveness at the level of the plurality of features of evidence, rather than at the level of the plurality of kinds of evidence available (which is accounted for by (I)).

concomitant shortcomings of the AM show that such neglect is unreasonable. An AM should satisfy (E).

In sum, desiderata that AMs should meet include:

(I) Inclusivity
An AM should include all available evidence.

(C) Constraint
An AM should constrain intersubjective assessments of the hypothesis of interest.

(E) Evidential Complexity An AM should assess evidence on multiple relevant evidential criteria.

As described, these desiderata are highly schematic. Besides noting some extreme ways in which a consensus conference could fail to satisfy the desiderata, I have not described specific ways in which a consensus conference could satisfy them. This would be a large undertaking for any of the individual desiderata, and is beyond the scope of the present paper.[13]

This list of desiderata is not meant to be exhaustive. Particular AMs may have other uses and some users of AMs may have other aims, which could motivate additional or alternative desiderata. For instance, given that consensus conferences are sometimes used to attain ends other than constraining justified credence about a hypothesis, non-epistemic standards can sometimes be used to assess them. To cite one example, since consensus conferences often involve the participation of non-expert citizens, and since the political legitimacy of a decision is an aim of some consensus conferences, Douglas (2005) poses the following evaluative question: "Has citizen involvement helped to bring citizen values into the heart of technical judgment?" Although this is an important question, it is not my aim to evaluate AMs with Douglas' criterion. Similarly, Solomon (2007) notes that consensus conferences often have been accused of bias in the selection of consensus participants; such criticisms assume that a more inclusive and transparent selection of participants ought to be a desideratum of consensus conferences.[14]

(I), (C), and (E) are more modest than 'ability to get at the truth'. Solomon (2007) claims specifically about NIH consensus conferences that they have "never been assessed for the accuracy of outcomes. No-one has investigated, for example, whether the outcomes are better—more 'true' or whatever—than those achieved by other methods such as non-neutral panels or formal meta-analysis of evidence." Solomon is right to be worried about the paucity of assessments of consensus conferences, given their prominent role in policy deliberation and evaluations of complex hypotheses. The trouble with holding an AM to a standard of 'more true' than competitor AMs is that the central purpose of gathering diverse evidence with a consensus conference is to gather our best indicators of the truth. As per the argument raised above, in the contexts in which AMs are typically employed we cannot assess the accuracy of the outcomes of an AM, nor their ability to get at the truth, unless we have an independent indicator of the truth, in which case we would have no need for the AM in the first place.[15] This consideration applies even when the independent indicator of truth is itself another AM (like meta-analysis): judging the veracity of consensus conferences based on

[13] For instance, one way to substantiate (E) would be to consider the extensive philosophical literature on experimentation, which among many works include Hacking (1983), Franklin (2002), and Weber (2005)

[14] This desideratum, a kind of social inclusiveness, is distinct from my (I) above, meant to be inclusiveness of an epistemic kind only. Nevertheless, as Longino (1990) and others have argued, one way to help achieve the epistemic virtues that I am concerned with might be to guarantee social inclusiveness in the process of consensus formation. I return to the relation between social inclusivity and my three desiderata for AMs in §4.

[15] Some argue that knowledge is what an ideal epistemic community would, in the long run, eventually agree on (for instance, this is one interpretation of Peirce's notion of convergence to the truth). Others argue that knowledge is just what an actual epistemic community settles on (see, for example, Kusch 2002), and

their agreement with meta-analyses assumes that the meta-analyses are themselves veracious. Thus AMs should be assessed not based on whether they are true or accurate, or even more true or more accurate then competitor AMs, but at the very least we can hold AMs like consensus conferences to standards such as (I), (C), and (E).

Miller (2013) argues that a necessary condition for cases of consensus to be considered knowledge-based—as opposed to consensus which arises for non-epistemic reasons (such as group-think), or consensus which arises for epistemic reasons but nevertheless fails to arrive at a sufficiently warranted conclusion—is that there be an apparent consilience of evidence for the hypothesis on which there is consensus.[16] I agree that consilience is important, though I do not explicitly demand it as a desideratum of consensus conferences.[17] Its importance can be understood in relation to the three desiderata presented here: (I) is a necessary condition for a thoroughgoing consilience (since an easy but unsatisfying way of achieving consilience would be to only include evidence which is concordant and exclude all evidence which is discordant); to the extent that consilience is a truly compelling epistemic desideratum, it can be construed as a sufficient condition for achieving (C) (since if all available evidence confirms one hypothesis over its rivals, we would demand of a consensus conference that its conclusions be constrained to support this hypothesis over its rivals). Consilience, however, is not a necessary condition for achieving (C), because even if different modes of evidence support different hypotheses, the consensus conference could have techniques (analytic or otherwise) for achieving constraint despite the discordant evidence.[18]

Thus, Constraint, Inclusivity, and Evidential Complexity are three desiderata that consensus conferences should meet. Although these desiderata may not be exhaustive, they do incorporate the important epistemic aspects of the desiderata that Miller (2013) and others demand of knowledge-based consensus. Indeed, the above considerations suggested that (I), (C), and (E) are desiderata that any method of amalgamating evidence should meet. For instance, I have previously argued that a central goal of meta-analysis (which is often not met) is (C) (Stegenga 2011b).

Whether or not consensus conferences meet (I), (C), and (E) is an empirical matter. However, to my knowledge there are no empirical assessments of consensus conferences with respect to these desiderata (indeed, there are few if any assessments of consensus conferences at all).[19] In the following section I present two examples of consensus conferences in an attempt to begin such an assessment.

Footnote 15 continued
so if intersubjective assessment of hypotheses were tightly constrained, then knowledge would be achieved. Though I will not argue the point here, since many others have done so, the conflation between consensus and knowledge should be rejected. See also Miller (2013).

16 There is a growing body of literature concerned with the epistemic value of consensus, of which Miller (2013) is a recent valuable addition. Since the primary focus of the present paper is on consensus conferences rather than on consensus per se, I avoid an exposition of this literature, but for a sampling, see also Gilbert (1987), Tuomela (1992), Wray (2001), and Tucker (2003).

17 However, for a critique of the assumed epistemic value of consilience, see Stegenga (2009). Consilience is often called 'robustness' (see also footnote 1).

18 The consensus achieved by the Intergovernmental Panel on Climate Change could be described as an example of achieved constraint despite discordant evidence.

19 Though Solomon (2007) notes that consensus conferences have been assessed based on their freedom from bias, by the Rand Corporation in 1983, a group at University of Michigan in 1987, and the NIH in 1999.

3 Two Examples of Consensus Conferences

3.1 Influenza Transmission

A consensus conference was convened in 2006 by the Public Health Agency of Canada to address how the influenza virus is transmitted from person to person. There are two main ways in which any virus, including influenza, is thought to spread from person to person: the Contact Hypothesis is that influenza is transmitted by direct or indirect touching, and the Airborne Hypothesis is that influenza can be transmitted on fine droplets over long distances through the air. The trouble is that available evidence on influenza transmission is discordant: some modes of evidence support the Contact Hypothesis while other modes support the Airborne Hypothesis. Many infectious-disease physicians consider the Airborne Hypothesis to be unlikely, and they appeal to their clinical experience as justification. Many occupational health experts and some virologists, on the other hand, tend to regard the Airborne Hypothesis as more likely than do infectious disease physicians, and they appeal to mathematical models, experiments on animals, and anecdotal accounts (such as the spread of influenza on airplanes) as justification. Carefully controlled human experiments cannot be performed given the obvious potential harm to research subjects.[20]

The consensus conference involved representatives from relevant disciplines, including virology, clinical infectious diseases, occupational health, and epidemiology. Thus the organizers of the consensus conference tried to ensure something like social inclusiveness by including representatives from several distinct professions in the consensus deliberations, with the aim, presumably, of satisfying (I). Participants were invited to present their arguments and evidence in an attempt to develop consensus regarding the mode of influenza transmission. The two-day forum included opportunities to scrutinize the available evidence on multiple criteria, thereby going some way toward satisfying (E). For instance, some participants criticized the available anecdotal evidence as merely observational, and liable to suffer from confounding biases. Similarly, some participants criticized the mathematical models of influenza transmission as relying on many assumptions. Conversely, the proponents that introduced such evidence defended the evidence to the extent that they could (for example, the modellers defended some of the model assumptions on the grounds that they had been empirically tested, and they defended other model assumptions on the grounds that they were theoretically plausible).

However, the disputants did not come to a settled view regarding how influenza is transmitted from person to person, and thus (C) was not satisfied. Moreover, similar consensus conferences on the same topic were held around the world in various jurisdictions, including the United States, the United Kingdom, and the World Health Organization, none of which had a settled outcome. In short, (I) and (E) were satisfied, but (C) was not satisfied.

3.2 Radiation Safety

Beatty (2006) describes a group of geneticists in the 1950s who had the task of estimating the minimum threshold of radiation that humans could safely be exposed to before undergoing genetic mutation. Hermann Muller was one of the geneticists most worried about mutation, and had won the Nobel Prize in 1946 for related work. Other prominent geneticists had argued that genetic variation was valuable for evolution, and so (they argued) it was difficult to know just how detrimental new radiation-induced variation would be. But the disagreement between

[20] However, there were some experiments performed on prisoner 'volunteers' in the 1960s, with mixed results.

the geneticists as a group and the Atomic Energy Commission (AEC) was even greater; the stakes for the AEC were high:

> AEC officials sometimes claimed that the biological (including genetic) effects of radiation exposure from bomb testing and other sources were negligible … they most chose to rebut Muller by emphasizing the lack of consensus among geneticists... (Beatty 2006)

To respond to the AEC, the U.S. National Academy of Sciences (NAS) organized a panel of geneticists, chaired by the Rockefeller official (and non-geneticist) Warren Weaver, to develop guidance on an acceptable radiation level. Although this was prior to the coinage of the term 'consensus conference', the aim and method of this meeting was essentially the same as those meetings that later were called by that term. The meeting began with Weaver encouraging the geneticists to communicate a sense of certainty and agreement to the public. However, during and after the meeting, given the discordant multimodal evidence which was available for this question, the NAS panel disagreed by over three orders of magnitude on estimates of radiation danger. But the final report, published in the *New York Times*, claimed that there was "*no disagreement as to fundamental conclusions*" (emphasis in original). The primary goal of the NAS panel was epistemic, but Beatty argues that a secondary goal was to mask disagreement (which they succeeded at), and to develop guidelines before another group, especially the AEC, did so.

The various geneticists appealed to diverse considerations when estimating the potential danger of radiation, including general considerations based on evolutionary theory, experience with humans irradiated by atomic bombs at Hiroshima and Nagasaki, and mathematic models. In short, these geneticists attempted to consider the relevant evidence that was at their disposal. The various lines of reasoning and modes of evidence mustered for support were exposed to the scrutiny of the group of geneticists. Beatty describes extensive interaction between the geneticists regarding assessing the available evidence. However, as noted above, this consensus conference failed to achieve agreement regarding the minimum threshold of radiation that humans could safely be exposed to. Thus it is reasonable to say of this consensus conference that it satisfied (I) and (E), but not (C).

4 Social Inclusivity

Organizers of consensus conferences are faced with the question of who to invite as participants. Critics of consensus conferences note that a conclusion that a consensus conference might reach is sensitive to who is included in the conference. More generally, many have thought that there is a relation between the inclusion of the perspectives of scientists of diverse backgrounds—call this *social inclusivity*—and scientific accuracy. A prominent example is Longino's argument that social inclusivity is an aid to achieving objectivity (1990). Since consensus conferences are social methods of amalgamating evidence, which involve gathering individuals together to assess multimodal evidence, social inclusivity might be considered another desideratum with which to assess consensus conferences. Since different kinds of evidence are often generated by different scientific sub-disciplines, this desideratum could be achieved by ensuring that representatives of all relevant sub-disciplines are included in a consensus conference. With Longino (and many others) I agree that social inclusivity is an important property that science should generally strive for, and so, specifically, given their social nature, consensus conferences ought to achieve social inclusivity. However, I argue

here that in one important sense the three desiderata argued for in §2 are more fundamental (and I note too that Longino herself suggested this very point in her original discussion of social inclusivity and objectivity).

Social inclusivity has both epistemic and non-epistemic functions. The non-epistemic functions of social inclusivity—in science generally and for consensus conferences in particular—can include the democratic control of research agendas and transparency when forming policies about the technological products of science. The epistemic function of social inclusivity is that it can help to achieve (I) and (E). The consensus conference on influenza transmission described in §3 achieved (I) and (E) at least in part because a wide variety of scientific disciplines and professionals were included as participants. However, in principle the satisfaction of social inclusivity can trade off against the satisfaction of at least one of (I), (C), and (E), and if it does, these desiderata are more fundamental, at least with respect to the epistemic functions of consensus conferences.

Consider the relation between social inclusivity and (I). Satisfying social inclusivity is neither necessary nor sufficient for satisfying (I), since in principle a consensus conference could include representatives from a narrow range of relevant disciplines and yet adequately consider all of the various kinds of relevant evidence, and in principle a consensus conference could succeed in including representatives from all the relevant disciplines and yet not adequately consider all of the various kinds of relevant evidence. Although satisfying social inclusivity is neither necessary nor sufficient for satisfying (I), the satisfaction of the two may be positively correlated; satisfying the former may, on average, tend to aid in, or raise the probability of, satisfying the latter.[21] At least some consensus conferences are organized with this in mind. The organizers of the consensus conference on influenza transmission explicitly aimed at inviting representatives from all of the relevant scientific sub-disciplines.

The intuition that social diversity is positively correlated with epistemic diversity is robust, and so it might be hard to imagine cases in which the two are inversely correlated. In fact the influenza transmission case provides a concrete example. When evaluating the available evidence, some infectious disease physicians argued that "There is no direct evidence such as from randomized control trials. There is some clinical evidence, but it is scarce and unconvincing." This claim expresses the widely held assumption by many of those involved in the so-called evidence-based medicine movement that only evidence from randomized control trials is convincing.[22] Evidence from other kinds of studies (or modes, to use my terminology above) was deemed 'unconvincing' (and many in the evidence-based medicine movement hold that such evidence ought to be ignored). In other words, the inclusion of a particular participant from a particular discipline created a threat to (I). As long as one's aim is to learn whatever one can about a particular hypothesis based on the totality of the available relevant evidence, to the extent that the inclusion of a particular scientific sub-discipline in a consensus conference leads to the neglect of a certain kind of evidence, the problem is with the inclusion of the sub-discipline and not with the evidence.

Similarly, social inclusivity can aid in satisfying (E), since representatives of different sub-disciplines might evaluate different aspects of the available evidence in different ways, leading to a fuller assessment of the plurality of evidential features. Like the relation between social inclusivity and (I), the relation between the satisfaction of social inclusivity and the satisfaction of (E) is not one of necessity or sufficiency, but rather of positive correlation

[21] Mathematical models have been employed to show that groups of diverse problem solvers outperform groups of high-ability problem solvers; see, for example, Hong and Page (2004). For a sociological study of the value of social inclusiveness, see Collins and Evans (2002).

[22] This view has been heavily criticized. See footnote 9 for references.

due to the presumed tendency of social inclusivity to aid in the satisfaction of (E). Longino's canonical account of the role of social inclusivity as an aid to achieving objectivity granted that the social condition for objectivity is not primarily the diversity of the *scientists* involved in assessing a hypothesis, but rather is the diversity of *points of view* relied upon in assessing a hypothesis, where a point of view is not an epistemic state held by one and only one particular person, but rather is an epistemic state that can be held by multiple people, and more importantly multiple such epistemic states can be held by a single person:

> "Many individuals (sharing assumptions and points of view) may be involved in testing a hypothesis (and commonly are in contemporary experiments). And though this is much rarer, one individual may be able to criticize her or his own evidential reasoning and background assumptions from other points of view." (1990)

To use the terminology employed above, Longino's claim is that satisfying (I) and (E) can trade-off against satisfying social inclusivity. My claim above that social inclusivity is only positively correlated with the satisfaction of (I) and (E) is consistent with Longino's claim that though such trade-offs are possible, they are rare (the frequency of this trade-off is an empirical matter, though I am unaware of anything more than intuitive appeal to warrant the claim that it is rare).

The more diverse the set of participants is in a consensus conference, the more difficult it can be to achieve (C). This is seen in contemporary science when a particular community of scientists has managed to achieve (C) for one of their hypotheses or theories, but 'outsiders' to the community claim that the hypothesis is less well-founded than the community claims. Examples abound (climate change skeptics, doubters of the smoking-cancer link, HIV-AIDS deniers, promoters of intelligent design).[23] These are cases in which a defined scientific community achieved (C) or at least satisfied (C) to some degree, but had their community been expanded to include such outsiders, (C) would not have been achieved. Thus, to the extent that social inclusivity is positively correlated with satisfying (I) and (E), consensus conferences should strive for a degree of social inclusivity up to the point that (I) and (E) are maximally satisfied, but no more, since too much social inclusivity can make it more difficult to achieve (C).

5 Conclusion

I have argued that consensus conferences have difficulty achieving (C), and I illustrated this with two examples. Solomon (2007) also suggests that consensus conferences fail to achieve (C). Her argument can be construed as a dilemma: consensus conferences occur either (i) before scientific consensus has been achieved, or (ii) after scientific consensus has been achieved (through means other than consensus conferences). If (ii), then the consensus conference did not achieve (C), because (C) was already achieved; if (i), then the consensus

[23] Dissenting outsiders are often non-scientists, and so are not 'insiders' to any scientific community. But sometimes such outsiders can be respected scientists in one community and be vocal dissenters to a consensus established by another community. The HIV-AIDS deniers (those who deny that HIV causes AIDS) are a salient example. Some are fully outsiders (Thabo Mbeki, the former president of South Africa is one of the most prominent examples). But others include Peter Duesberg, a molecular biologist at University of California, Berkeley, and Kary Mullis, winner of a Nobel Prize in chemistry. This raises the question: among professional scientists, what constitutes membership in this or that community? A cynical Kuhn-inspired answer might be: assent to the hypothesis about which consensus is in question. A less cynical answer might be: performing active research on that hypothesis. Duesberg and Mullis were outsiders to AIDS research on either answer. See van Rijn (2006).

conference cannot achieve (C), because the evidence is still too contentious for (C) to be reached. So either way consensus conferences do not achieve (C): "the window for usefulness [of consensus conferences] is small—after there is enough evidence to reach a conclusion but before the research community itself has reached consensus" (169). Indeed, there are plenty of cases of consensus conferences which seem to achieve (C), in contrast to the two cases presented in §3. But such cases are not necessarily counter-examples to the general claim that consensus conferences fail to achieve (C), since, following Solomon, one could argue that it was not the consensus conference which achieved (C) in these seemingly favorable cases, but rather it was the concordant evidence garnered prior to the consensus conference.[24]

Some might object to (C) as a desideratum, since if (C) were met, then insights of dissenting opinions and evidence could be lost. This concern is expressed, for instance, by Solomon (2006), and of course famously by Mill (1859). There is something attractive about letting a thousand flowers bloom. Loss of dissent is indeed a concern, but only if (C) is achieved cheaply, by ignoring the complexity and diversity of evidence (that is, by ignoring (I) and (E)). If (I) and (E) are satisfied, then, whether or not (C) is met, by stipulation the diversity and complexity of evidence is not ignored.[25] Regardless, the satisfaction of (C) need not involve the silencing of dissent.

Deliberative approaches may use formal or quantitative methods as part of the process of assessing multimodal evidence. The NAS geneticists described by Beatty (2006), for example, calculated numerical averages of their estimates for minimum acceptable radiation levels and for estimating a number of genetic defects given certain levels of radiation. Of course, quantitative or formal methods can be more sophisticated than this. However, some advocates of deliberative approaches have been critical of formal approaches; one criticism is that technical algorithms for assessing and amalgamating multimodal evidence "bury under a series of assumptions many value judgments" (Lomas et al. 2003). In situations of discordant multimodal evidence, critics of formal approaches to amalgamation have claimed that discordance "cannot be resolved by an appeal to science," and when faced with discordance, "the search for some formula or set of principles designed to provide decision-making rules will always prove elusive" (Klein and Williams 2000). In a *New York Times* article discussing recent studies which purport to show that antidepressants are no more effective than placebo, the author complained about reviews based on formal methods: "in the end, the much heralded overview analyses look to be editorials with numbers attached" (Kramer 2011).

Conversely, others argue that formal AMs are more objective than social AMs such as consensus conferences, since these approaches counter the subjective biases and uncertainty present in the latter. Solomon (2007), for instance, suggests that rather than relying on consensus conference to amalgamate evidence, it would be "quicker, more timely, and at least as good to do a meta-analysis of the available evidence" (169). However, elsewhere (2011) I argue that meta-analysis is unable to counter subjective biases, and that multiple meta-analyses of the same hypothesis can reach contradictory conclusions (and hence meta-analyses can also

[24] Solomon (2007) gives several examples of such consensus conferences, including a 1994 conference titled "Helicobacter Pylori in Peptic Ulcer Disease" and a 2002 conference titled "Management of Hepatitis C". Both of these conferences appeared to achieve (C), but in fact the conferences took place some time after the relevant scientific communities had already achieved consensus. For a criticism of Solomon's argument, see Kosolosky (2012).

[25] I do not mean to suggest that it is simple to avoid the subtle biases that arise in group deliberative processes. Janis (1982) argued that groups are liable to come to incorrect conclusions in certain circumstances; peer pressure and authoritative pressure stifles dissent and quiets the discussion of discordant evidence. In contrast, Tollefsen (2006) argues that scientists can engage in collaborative deliberation without engaging in groupthink or stifling dissent.

fail to meet (C)). Thus, although consensus conferences have difficulty achieving (C), more formal techniques are not necessarily better at achieving (C).

In short, I have argued that consensus conferences are good at satisfying what I have called Inclusivity (I) and Evidential Complexity (E), but not as good at satisfying Constraint (C).

Acknowledgments I am grateful to Nancy Cartwright, Boaz Miller, Alex Broadbent, Laszlo Kosolosky, Miriam Solomon, Anton Froeyman, Jeroen Van Bouwel, Heather Douglas, and two anonymous reviewers for detailed feedback on versions of this paper. Financial support was provided by the Banting Postdoctoral Fellowships Program administered by the Social Sciences and Humanities Research Council of Canada.

References

Beatty, J. (2006). Masking disagreement among scientific experts. *Episteme*, *3*, 52–67.
Cartwright, N. (2006). Well-ordered science: Evidence for use. *Philosophy of Science*, *73*, 981–990.
Cartwright, N. (2007). Are RCTs the gold standard? *Biosocieties*, *2*, 11–20.
Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Clarendon Press.
Chang, H. (2004). *Inventing temperature*. Oxford: Oxford University Press.
Collins, H. M., & Evans, Robert. (2002). The third wave of science studies: Studies of expertise and experience. *Social Studies of Science*, *32*(2), 235–296.
Culp, Sylvia. (1994). Defending robustness: The bacterial mesosome as a test case. *PSA*, *1*, 46–57.
Douglas, H. (2005). Inserting the public into science. In S. Maasen, & P. Weingart (Eds.), *Democratization of expertise? Exploring novel forms of scientific advice in political decision-making*. Netherlands: Springer.
Eysenck, H. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal*, *309*, 789–792.
Franklin, A. (2002). *Selectivity and discord: Two problems of experiment*. Pittsburgh: Pittsburgh University Press.
Gilbert, M. (1987). Modeling collective belief. *Synthese*, *73*(1), 185–204.
Hong, L., & Page, S. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389.
Janis, I. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes*. Houghton Mifflin.
Joss, S., & Durant, J. (Eds.). (1995). *Public participation in science: The role of consensus conferences in Europe*. UK: Science Museum.
Klein R., & Williams, A. (2000). Setting priorities: what is holding us back—inadequate information or inadequate institutions? In A. Coulter, & C. Ham (Eds.), *The global challenge of health care rationing*. Buckingham: Open University Press.
Kosolosky, L. (2012). The Intended window of epistemic opportunity: A comment on Miriam Solomon. In B. Van Kerkhove, T. Libert, G. Vanpaemel, & P. Marage, (Eds.), *Logic, philosophy and history of science in Belgium II*. Koninklijke Vlaamse Academie van België.
Kosso, P. (2006). Detecting extrasolar planets. *Studies in History and Philosophy of Science*, *37*, 224–236.
Kramer, P. (2011). In defense of antidepressants. New York Times July 9.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic modeling as robustness analysis. *The British Journal for the Philosophy of Science*, *61*, 541–567.
Kusch, M. (2002). Knowledge by agreement: the programme of communitarian epistemology. Clarendon.
Lomas, J., Fulop, N., Gagnon, D., & Allen, P. (2003). On being a good listener: Setting priorities for applied health services research. *Milbank Quarterly*, *81*(3), 363–388.
Marmot, M. (2004). *The status syndrome*. New York: Times Books.
Mill, J. S. (1859). On liberty.
Miller, B. (2013) When is consensus knowledge-based? Distinguishing shared knowledge from mere agreement . *Synthese, 190*, 1293–1316.
Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
Solomon, M. (2007). The social epistemology of NIH consensus conferences. In H. Kincaid, & J. McKitrick (Eds.), *Establishing medical reality: Essays in the metaphysics and epistemology of biomedical science*. Springer.
Solomon, M. (2006). Groupthink versus the Wisdom of Crowds: The social epistemology of deliberation and dissent. *The Southern Journal of Philosophy*, *44*, 28–42.
Stegenga, J. (2009). Robustness discordance, and relevance. *Philosophy of Science*, *76*, 650–661.

Stegenga, J. (2011a). An impossibility theorem for amalgamating evidence. *Synthese*, *190*, 2391–2411.
Stegenga, J. (2011b). Is meta-analysis the platinum standard? *Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507.
Tollefsen, D. P. (2006). Group deliberation, social cohesion, and scientific teamwork: Is there room for dissent? *Episteme*, *3*, 37–51.
Tucker, A. (2003). The epistemic significance of consensus. *Inquiry*, *46*(4), 501–521.
Tuomela, R. (1992). Group beliefs. *Synthese*, *91*(3), 285–318.
van Rijn, K. (2006). The politics of uncertainty: The AIDS debate, Thabo Mbeki and the South African government response. *Social History of Medicine*, *19*(3), 521–538.
Weber, M. (2005). *Philosophy of experimental biology*. Cambridge: Cambridge University Press.
Westman, R. (2011). *The copernican question: Prognostication, skepticism, and celestial order*. Berkeley: University of California Press.
Wimsatt, W. (1981). Robustness, reliability, and overdetermination. In M. B. Brewer, & B. E. Collins (Eds.), *Scientific inquiry and the social sciences*. Jossey-Bass.
Worrall, J. (2007). Why there's no cause to randomize. *British Journal for the Philosophy of Science*, *58*, 451–588.
Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science*, *69*, S316–S330.
Wray, K. B. (2001). Collective belief and acceptance. *Synthese*, *129*(3), 319–333.

Jacob Stegenga is an Assistant Professor in the Department of Philosophy at the University of Utah. His area of research is philosophy of science, including methodological problems of medical research, conceptual topics in evolutionary biology, and fundamental issues in reasoning and rationality. He is currently completing a book tentatively titled Strange Pill: Evidence, Values, and Medical Nihilism, which defends a skeptical thesis about the products of medical research. A native of Victoria (Canada), Stegenga completed his Ph.D. in philosophy at the University of California, San Diego, and a postdoctoral fellowship at the University of Toronto. Details about his research can be found on his website: http://individual.utoronto.ca/jstegenga.