Environmental Science and Policy 84 (2018) 80–87

Contents lists available at ScienceDirect

Environmental Science and Policy

journal homepage: www.elsevier.com/locate/envsci

Quality of epidemiological studies: Procedural rules for uncertain science for policy, a case study on bisphenol-A

Laura Maxim[a,*], Jeroen Van der Sluijs[b,c,d]

[a] Institut des Sciences de la Communication, UMS3665, CNRS/Université Paris-Sorbonne/UPMC, 20 rue Berbier du Mets, Paris, 75013, France
[b] Centre for the Studies for the Sciences and the Humanities, University of Bergen, Postboks 7805, Bergen, 5020, Norway
[c] Department of Chemistry, University of Bergen, Postboks 7805, Bergen, 5020, Norway
[d] Copernicus Institute of Sustainable Development, Environmental Sciences, Utrecht University, P.O. Box 80115, TC Utrecht, 3508, The Netherlands

ARTICLE INFO

Keywords:
Quality
Epidemiologic
Chemical
Risk
Uncertainty
Advisory

ABSTRACT

This paper proposes a method for in-depth mapping of heterogeneity in expert judgment, in the evaluation of the quality of epidemiological studies used in regulatory chemical risk assessment. Whereas consensus in scientific advisory groups provides legitimation for subsequent political action, it can also have unintended effects on the quality of regulatory risk assessment.

Based on empirical testing of our method, called Qualichem_epi, with ten experts and two epidemiological case studies about bisphenol A (BPA)'s effects on human health, we have shown that expert judgment plays an essential role in managing uncertainty and deciding what "quality" of a study actually means. We found substantial heterogeneity of scientists' judgments about the quality of epidemiological studies, even if the same criteria were used for the assessment. This heterogeneity is not present anymore in reports produced by expert groups, where results are presented under the collective signature of all the scientists involved. We argue that flattening heterogeneity can be an important problem when it is not the result of true scientific agreement but only a secondary effect of consensus-based working procedures of agencies that experts have to follow.

Qualichem_epi provides an easy to understand color-based picture of both majority and minority opinions in a scientific advisory group. We suggest that it could be used on a regular basis for communicating quality assessments of epidemiological studies in regulatory chemical risk assessment.

1. Introduction

What counts as valid "evidence", along with implicit and explicit criteria used for appraising its quality, is the crux of the scientific advisory process. However, in practice this may differ largely among scientists and expert groups active at the science-policy interface. For example, Beronius et al. (2010) compared ten risk assessments produced by expert groups in the European Union (EU), the United States (USA), Canada and Japan, and seven of them (published between 2002 and 2008) found no risk to the general population. One, namely the Chapel Hill experts who gathered in 2006 at a meeting sponsored by National Institute of Environmental Health Sciences (NIEHS) and US Environmental Protection Agency (US EPA), concluded that there is a risk to the entire population at current exposure levels. The remaining two committees published their results in 2008 and expressed concern about some risks, primarily to fetuses and infants.

These divergent conclusions are characteristic for the controversy about the effects of endocrine disrupters on human health and the environment. The core of the disagreement is about the health effects of small doses of chemicals, potentially following nonmonotonic dose-response patterns, which might act on the endocrine system and hence affect a wide range of body functions, on the long term (exposure of mothers leading to lifelong effects on the child to born) (Vandenberg et al., 2009). In a regulatory framework, two problems fuel the controversy: such effects might be hard to grasp by standardized OECD testing protocols that have nevertheless a status of "recognized evidence" (Maxim and Van der Sluijs, 2014), and the definition of endocrine disrupters is debated (Horel and Bienkowski, 2013).

More recent assessments of BPA risk by ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) and EFSA (European Food Safety Authority) reveal the same pattern: ANSES (2013) concluded that "…*handling thermal paper receipts leads to risk situations for the four types of effects considered: mammary gland, brain and behaviour, the female reproductive system, metabolism and obesity*." (p. 6) Two years later, EFSA (2015) was publishing an opposite conclusion: "*there is no health concern for any age group from dietary exposure and low*

* Corresponding author.
E-mail addresses: laura.maxim@cnrs.fr (L. Maxim), Jeroen.Sluijs@uib.no (J. Van der Sluijs).

https://doi.org/10.1016/j.envsci.2018.03.010
Received 21 December 2017; Received in revised form 12 March 2018; Accepted 12 March 2018
Available online 20 March 2018
1462-9011/

health concern from aggregated exposure" (p. 1).

Different scientists may evaluate a given published study (or even raw data) as having very different quality and relevance (Rudén, 2001). Similarly, Maxim and Van der Sluijs (2014) showed that the discipline could influence the expert judgment about the quality of a scientific article on BPA. What is considered a "good paper" can be very different between specialists in endocrinology and scientists trained in other disciplines. Furthermore, at the science–policy interface, scientific quality can be assessed according to regulatory standards, such as OECD's standardized protocols, which can come in conflict with academic standards (Demortain, 2013; Maxim and Van der Sluijs, 2014). Van der Sluijs and Van Eijndhoven (1998) investigated variability in the climate risk assessment reports by expert groups in the Netherlands during the pre-IPCC period. They found remarkable differences in the conclusions of two expert groups despite a large overlap in composition. They showed that the context in which the experts operated and the commitments they made in each setting were key factors in explaining the variability in framing, judgments and conclusions. In practice, there is a lot of flexibility an expert group may introduce into the argumentative strategy when new scientific data or new practical situations arise. Van Eijndhoven and Groenewegen (1991) showed that despite the availability of scientific data that may call for a change in the assessment, the context can drive expert groups to stick to their former conclusions, whereas from the same data other conclusions can be constructed if the context changes.

For highly complex issues such as chemical risks, experts must reach a conclusion despite uncertainty, and expert judgment is used to fill in the gaps (Van der Sluijs et al., 2008). How uncertainty is dealt with in a particular context and by a particular group of scientists determines what is considered to be "good science" (and "evidence"), which has the political function to decide if the risk is "acceptable" or not (Jasanoff, 1990).

The essential role of uncertainty in science for policy has long been recognized and extensively addressed in the field of post-normal science (see Strand, 2017 for a recent overview). Funtowicz and Ravetz (1993) showed that present day complex issues at the science-policy interface exhibit characteristics that make them hard to tackle with normal scientific procedures. This requires new ways of interfacing science and policy (Funtowicz and Ravetz,1990). Funtowicz and Ravetz (1993) have called this class of problems post-normal, where 'normal' refers to Kuhn's (1962) concept of normal science: the practice of uncritical puzzle solving within an unquestioned framework or 'paradigm'.

Funtowicz and Ravetz (1993) signalised that normal science runs into serious limitations when addressing societal issues (in that time nuclear reactor safety) where scientific evidence is highly contested and plagued by uncertainties while decisions are urgent, stakes high, and values in dispute. The available knowledge is typically characterised by imperfect understanding of the complex systems involved. Models, scenarios, and assumptions dominate assessments of such issues, and many (hidden) value loadings reside in problem frames, indicators chosen, and assumptions made.

Scientific assessments of complex risks are thus unavoidably based on a mixture of knowledge, assumptions, models, scenarios, extrapolations and known and unknown unknowns. Consequently, scientific assessments will unavoidably use expert judgements. It comprises bits and pieces of knowledge that differ in status, covering the entire spectrum from well-established knowledge to judgments, educated guesses, tentative assumptions and even crude speculations (Van der Sluijs et al., 2005, 2008). Knowledge utilisation for risk governance requires a full and public awareness of the various sorts of uncertainty and underlying assumptions. To perform this task, Knowledge Quality Assessment (KQA) tools are essential (Van der Sluijs et al., 2008; Maxim and Van der Sluijs, 2011; Maxim and Van der Sluijs, 2014). KQA seeks to systematically reflect on the limits of knowledge in relation to its fitness for function. It comprises systematic analysis of, and critical reflection on, uncertainty, assumptions and dissent in scientific assessments in its societal and institutional context.

Despite the central role of study-quality in underpinning chemical risk policies, there is currently no structured framework for assessing the quality of epidemiologic studies regularly used in agencies. Whereas for toxicological studies several frameworks have been proposed in the literature[1] (Maxim and van der Sluijs, 2014; Samuel et al., 2016; Roth and Ciffroy, 2016) and the Klimisch score is recommended – while disputed - in regulatory contexts such as REACH,[2] some similar attempts were published for epidemiologic studies (World Health Organization, 2000; Deeks et al., 2003; Vandenbroucke et al., 2007; Briggs et al., 2009; Dor et al., 2009; Westreich, 2012) but no guidelines are available for their use in the advisory practice, in chemical risk assessment, at European level. Each agency may then use its own reviewing guideline, which may include the criteria considered relevant at that moment and in that specific institutional and socio-political context.

The choice of the types of studies to be included or excluded in the risk assessment is also implicitly or explicitly decided during the advisory activity and depends on each topic and their contextual/regulatory framework: whereas exclusively published studies could be used in some groups (Barthes, 2014), both published and unpublished literature, like ad-hoc reports funded by agencies for filling in gaps in the existing exposure knowledge, were used by ANSES (2013) and by EFSA (2015). Furthermore, the endpoints considered relevant for measuring the health or environmental impact of a substance can differ from one expert group to another (e.g., different endpoints were considered by ANSES and EFSA for deciding about the risk of BPA).

1.1. The consensus rule and the phenomenon of "the spiral of silence"

The work of expert groups is organized along formal and informal procedures in force in health and environmental agencies (henceforth "agencies"). A very common procedure is judgment by consensus among the members. While in practice part of the advisory work may include exchanges of contradictory evidence and contradictory interpretations of evidence, true or at least apparent-but-undisputed consensus is encouraged, for producing final conclusions.

Besides providing legitimacy for decision-making, the pursuit of consensus in expert groups has a second political role, which is to give a public sign of adhesion to a conclusion for all the members of the group. Jasanoff (1990) has shown how conclusions of expert groups represent the result of a negotiation between the scientists involved, and how expertise in itself may represent for agencies a way to lower or prevent controversies through involvement of the relevant scientists in advisory procedures. The exchange of arguments in an expert group can also lead to sound argument closure (Beauchamp, 1987) on parts of the scientific disputes. Further, dissident scientists for example might have the opportunity to express their views in a process that avoids deployment of their arguments in the public arena and subsequent criticism.

However, consensus is sometimes only an outward appearance (see for example the case of Love Canal described by Jasanoff, 1990). In any group, strong personalities can greatly influence collective discussions and limit the ability of other individuals to express critical opinions. When consensus is overtly favored, some individuals might be reluctant to express criticism when their opinions disagree with the group's majority view and/or chairman, a phenomenon coined "the spiral of silence" (Noelle-Neumann, 1986). The chairman of an expert group has a major role in balancing the different views but he/she might unconsciously favor those views that agree with his own or his institution.

It might be argued that recent procedures in agencies allow the expression of minority opinions. For example, ANSES recently

[1] Criteria for assessing quality are specific to toxicological studies, hence these frameworks cannot be used as such for assessing the quality of epidemiological studies.

[2] Registration, Evaluation, Authorisation and Restriction of Chemicals.

encouraged the expression of minority opinions, supported by French legislation that ground scientific advisory activities in the "contradiction principle" (ANSES, 2012, 2016). Minority opinions can hence be included as an annex to the main scientific output (e.g., a report), a similar procedure to that of EFSA (EFSA, 2017).

However, in practice, very few experts use this opportunity. Indeed, expressing a minority opinion remains exceptional in advisory procedures, as it produces isolation of the expert from the rest of the group and demands his/her strong commitment for using this procedure. Furthermore, it is adapted to few, salient aspects of the work, but not for managing regular heterogeneity of experts' judgments. Whereas it could theoretically reinforce the robustness of the group results, regular criticism can also be perceived as individual inability to work in a group, or even worse, a questioning of the scientific qualities of group colleagues (Barthes, 2014). In extreme cases, such an overtly and repeated criticism can contribute to attributing a label of troublemaker slowing down the collective work to the "guilty" expert. The resulting attitude from those experts can be to prioritize their criticism and focus only on some aspects, the others on which he/she might be critical being concealed to the harmony of the group functioning. In addition, other personal aspects come into play and discourage overt criticism, such as sympathy and respect that often creates during informal interactions (during lunches, coffee breaks, etc.).

In all these ways, the informal rule of consensus leads to losing the deviating views and judgments of certain individuals in the expert groups' discussions, despite their potentially significant contribution to the quality of the final conclusions. Indeed, minority (individual) views in a group are not necessarily minority views in science, but can simply be an artifact of the criteria used to choose experts to include in that group (Maxim and van der Sluijs, 2014).

In all, downplaying "minority" views can have important, if not dramatic consequences, as shown for the Fukushima case by Fujigaki and Tsukahara (2011). Even if some scientists had predicted earthquakes and tsunamis-related nuclear crises similar to what finally happened, those responsible for atomic policies ignored them.

The push for consensus can negatively influence the final quality of the advisory work through still another mechanism: one expert with undeclared conflicts of interests, but with sufficient discursive skills, can be enough to influence the whole judgment of the group, as the other members are collegially striving to consider his/her opinion for reaching consensus.

1.2. Consensus and selection of experts

The criteria used to select the experts to be included influence the ability to reach final consensus and its content. In practice, these criteria are rather general, referring to scientific competence, conflicts of interests, available time to be dedicated to advisory work and balance among different disciplines in a group.[3]

Similarly, the relative weight given to the competence of potential experts, compared to their personality, previous public positions taken on the issues to be addressed in the advisory activities or their scientific discipline is a patchwork job, specific to each situation of advising. Certain characters might be preferred, e.g., those who are more easily reaching consensus with their colleagues, and who adapt easier to the institutional framework of expertise, to procedural rules and to group functioning - all very different of those in academia. The choice of the experts may also depend on the very particular setting of the environmental topic assessed. For example, for the controversial issue of radio frequencies on human health in France, AFSSET built the group with the objective to find a balance between scientists who previously took public positions against NGOs, and those who hadn't, such an offset aiming to attaining somehow the "*group impartiality*" (Barthes, 2014). At EFSA, the group must reflect "*a balanced representation of skills and qualities and a broad and deep range of expertise and scientific perspectives*"[4].

The overall objective in agencies is balance between the experts in a group, which additionally contributes to flatten heterogeneity among scientists relevant for a particular topic. Aiming at attaining balance among different disciplines and even public views on a topic increases the probability that few - if not single - specialists of particular issues in a discipline or a research field are present in a group. Given that the perceived level of uncertainty in a given body of knowledge depends on the "distance from site of production" - i.e., the degree of specialization and knowledge about that issue (MacKenzie, 1990) - allowing the expression of heterogeneity in a group in of key importance to avoid biases. Indeed, compromise may lead to exclusion of specific knowledge that only one member of the group detains, which in some cases might even contradict the views of other members of the group which are not specialists of that specific subject (e.g., the case of statistical analysis of data in scientific papers, for which a specialist can provide a highly qualified insight that most generalist users of statistics do not have).

1.3. BPA as a case study

The BPA case study is particularly appropriate for our objectives: suspected to be an endocrine disrupter, BPA has made headlines all over the world—particularly in the USA and the EU. During the last years, beyond ANSES' and EFSA's reports addressed in this paper, the risk of BPA has been intensively assessed by many advisory groups and agencies, e.g., the European Scientific Committee on Food in 2002, the European Chemicals Bureau in 2003, the EFSA in 2006, 2008, 2009, 2010, 2011, 2015 and 2016, by Environment Canada and Health Canada in 2008, by WHO and FAO in 2009 and 2010, by FDA in 2010 and 2014, by the Swiss federal health authority in 2016, by the Japanese Research Institute of Science for Safety and Sustainability and the National Institute of Advanced Industrial Science and Technology in 2007 and 2011, by the Danish EPA in 2011, by RIVM in 2016. These numerous assessments responded institutionally to the intense sociopolitical controversy over the potential negative impacts of BPA present in many products (baby bottles and other baby food containers, cash receipts, epoxy resins, coatings of cans for food and beverages, electronic equipment housing units, dental sealants, etc.). Exposure during pregnancy was suspected to produce endocrine-related damages in the babies of the women concerned, including cancer, metabolic diseases such as obesity and neurobehavioral problems.

1.4. Objectives

Explicitly addressing the heterogeneity of expert groups can contribute to reinforcing high quality professional scientific judgment, based on continuous contradiction[5] and critical peer-review. Indeed, scientific work is based on the principle of peer-review as an essential contributor to quality and robustness. Accounting for criticism and diverging opinions is not contradictory with the pursuit of consensus, but aims at avoiding that consensus is reached for the wrong reasons (e.g., in scientific advisory activities, undue influence from one particular expert, or experts striving collectively to agree on a conclusion that nevertheless remains scientifically unsatisfactory for some of them).

In case where divergences remain, reporting them can be an option. The assumption that scientific legitimacy can only be based on consensus is based on the untenable linear model of the relationship

[3] https://www.anses.fr/fr/content/comit%C3%A9s-dexperts-sp%C3%A9cialis%C3%A9s-et-groupes-de-travail.

[4] http://www.efsa.europa.eu/sites/default/files/efsa_rep/blobserver_assets/expertselection.pdf.

[5] See for example ANSES, Avis n° 2016-2 relatif à la prise en compte des positions minoritaires [Saisine 14], https://www.anses.fr/fr/system/files/DEON-Ft-2016002.pdf.

between science and policy, which - in spite of its demonstrated unrealism - is still very present and produces problematic underexposure of policy-relevant scientific dissent (Van der Sluijs et al., 2010). For an agency, the perverse effect of flattening heterogeneity can be to give by itself all the reasons why it could be criticized by those who are not involved in its work – who will inevitably exist, given the large number of scientists working on some controversial issues, to the point that the agency cannot all include them in its groups, - or even by those involved but uncomfortable with the conclusions produced by the group (Barthes, 2014).

Inspired by post-normal science, we propose a tool allowing the expression of the disagreement between experts, in addition to points of convergence, for the review of epidemiological studies considered in chemical hasard and risk assessments. Such a tool could be used for strengthening the quality and transparency of the group's work and/or for communicating remaining uncertainties and dissent.

2. Methods

2.1. An original typology

To enable testing of our hypothesis, we combined the analysis of documents produced by ANSES (2011) and EFSA (2010, 2014) with an empirical setting that involved 10 scientists in academia — which is a sample of suitable size (Knol et al., 2010). We thus compared the evaluation of study quality by academic scientists alone (who were not subject to any procedural rules) with quality assessment of the same studies made by expert groups in two agencies, the EFSA and ANSES (where experts worked according to specific procedural rules).

For the empirical setting, we built on a method implemented in a recent study (Maxim and Van der Sluijs, 2014), called Qualichem (described in detail in Appendix A in Supplementary material). That method was entitled *Qualichem in vivo*, so we adapted the method here under the name *Qualichem_epi*. Using this method, we developed the typology of quality criteria adapted to epidemiologic studies (Appendix B in Supplementary material) iteratively, drawing on several sources of information: the main steps in the process of knowledge production in epidemiological studies; analysis of study quality evaluation expressed by agencies like ANSES and EFSA; and previous literature on the quality of epidemiologic studies (World Health Organization, 2000; Deeks et al., 2003; Vandenbroucke et al., 2007; Briggs et al., 2009; Dor et al., 2009; Westreich, 2012). In these sources, we identified the criteria used to criticize, argue in favor of, report, or evaluate the scientific robustness of epidemiologic studies. We considered the various lines of argumentation identified as expressions of expert judgments about epidemiologic studies, and that were therefore relevant criteria to include in our typology. To check the robustness of our typology, our interview protocol contained a final question about the need to exclude criteria or include new ones.

As the case studies, we used two epidemiologic studies of the effects of BPA, i.e., Sugiura-Ogasawara et al. (2005) and Mok-Lin et al. (2010). The first studied the association between serum BPA and miscarriages in 45 patients with a history of three or more consecutive first-trimester miscarriages, and 32 women with no history of live births and infertility. The authors concluded that exposure to BPA is associated with recurrent miscarriages. The second, Mok-Lin et al. (2010), was a prospective cohort study looking at the association between urinary BPA and ovarian response in women undergoing in vitro fertilisation. The authors concluded that BPA was inversely associated with the number of oocytes retrieved.

The 37 criteria of *Qualichem_epi* were assembled into eleven different classes (Appendix B in Supplementary material) that fell into two general categories: "Protocol" and "Results". The Protocol part of the typology include criteria that are relevant to the technical and methodological aspects of the quality of a study. The Results part include one class for technical and methodological quality (i.e., the results analysis), one class pertaining to communication quality (i.e., study reporting) and three classes pertaining to normative quality (i.e., results interpretation).

2.2. Elicitation protocol

We interviewed each respondent individually in the period 2012 to 2014. To prepare the interviews, we pasted relevant text from each of the two studies below each question, which saved respondents from having to search through the study for elements needed to answer the question.

Each respondent assessed one of the two case studies. To solicit their judgment, we presented each respondent with a question related to each of the criteria included in our typology. The elicitation protocols are available upon request. For example, the first question of our protocol was: "*Is the hypothesis of the study precise enough, in accordance with the best scientific knowledge and practices?"* The text from the study that refers to the hypothesis was copied below the question. The respondent was invited to express their judgment by answering according to an ordinal, Likert scale (Appendix C in Supplementary material) and to explain his/her response (Appendix C in Supplementary material). Interviews were recorded and transcribed. We used the transcripts to analyze the results (Appendix D in Supplementary material).

2.3. Choice of respondents

Respondents were chosen among the authors of the epidemiological studies identified by ANSES (2011) and of studies published after 2011, identified through an extensive literature search. One respondent was chosen based on membership in one of the European expert groups that worked on BPA. Following this process, we contacted sixty five experts by email. Ten agreed to participate, all of them were academics.

2.4. Choice of case studies

The two papers used as case studies were chosen following their mention in ANSES (2011). That report deemed Sugiura-Ogasawara et al. (2005) to be of low quality ("*study not taken into consideration since it has major methodological limitations*"), whereas Mok-Lin et al. (2010) was considered to be a study of '*high quality having no major methodological limitations*'. Among all the studies available that we could have chosen, we chose these two studies because of their different quality estimations (i.e., a "good" and a "bad" study). Among the ten respondents, three assessed the Mok-Lin et al. (2010) and seven the Sugiura-Ogasawara et al. (2005).

3. Results: Qualichem_epi for two published papers and comparison with assessments by ANSES and EFSA

Within the application of Qualichem we distinguish two levels of quality: aggregated quality and level of confidence in the whole study (see details in Appendix A in Supplementary material). They provide a way to represent both majority and minority opinions, and hence those results should be considered together. We represent the overall scores in graphs that are divided into three colored areas: red (including scores and median scores < 3), orange (for scores and median scores between 3 and 4) and green (for scores and median scores > 4). For each criterion, a line covers the full range, from the lowest score to the highest score in the group of responding experts. The median score is represented by an "x" and the interquartile range is represented by a rectangle. The aggregated quality of an individual criterion is assigned as follows:

- High aggregated quality: median in the green area (> 4)
- Average aggregated quality: median in the orange area (ranging from 3 to 4)

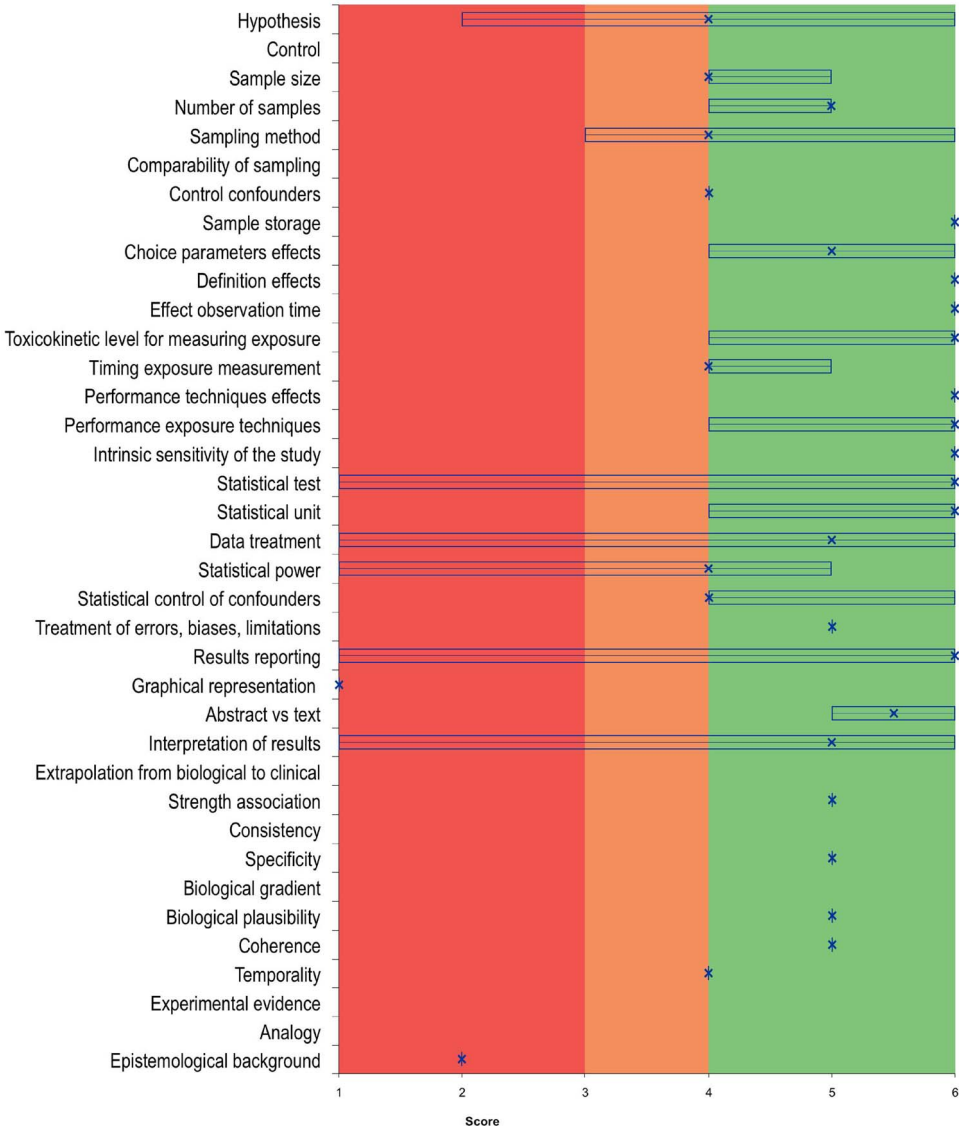

Fig. 1. Quality assessment of Mok-Lin et al. (2010), using Qualichem_epi with three respondents.

- Low aggregated quality: median in the red area (< 3).

Among the expert responses to the case study of Mok-Lin et al. (2010), only two criteria received a median score in the red area (scores 1 and 2), whereas all others fell in the orange or green area, these median scores being equal or higher than 4 (Fig. 1); this indicated a high aggregated quality for most criteria. However, scores on many criteria showed controversy (14 of the 37), and 14 of the 37 criteria were critical according to the definition given by Maxim and Van der Sluijs (2014) (also provided in Appendix A in Supplementary material). These findings mean that the study of Mok-Lin et al. (2010) can be defined as "**average quality**" according to our expert elicitation, showing that minority opinions are numerous and important. In contrast to our assessment, the expert group that produced by ANSES (2011) considered this study to be "**of high quality or having no major methodological limitations**", while pointing out three specific issues related to quality criteria:

- First, the point in time when BPA was measured would reflect exposure at that point, and not at the intended time of follicular maturation.
- Second, the report considered it difficult to extrapolate the results observed in a sample of infertile women who were undergoing in vitro fertilization to the general population.
- Third, the sample size (n = 84) was considered rather limited (p. 101).

By comparison, for these same three criteria, our Qualichem respondents considered that:

- The timing of exposure measurement was rather well dealt with in the paper (average score of 4 and inter-expert differences of only 1 point, showing a consensual judgment)
- The sampling method was equally criticized by one of the respondents, who gave a low score of 3, equally disagreeing with the chosen sample's representativeness of the study population; according to the respondent, the sample should have been selected from the general population instead. However, another respondent gave a maximum score for this criterion without any comment.
- The sample size received an average score of 4, inter-expert maximum difference being of one point.

This comparison shows sensibly different weights given to the different quality criteria by the expert(s) having analysed the quality of

the Mok-Lin et al. (2010) study for ANSES and our interviewees. Two of the criteria having received very low scores for Qualichem have been ignored by ANSES's experts. The difference is even more striking when we compare the results obtained with Qualichem and the quality assessment produced by EFSA for this study. In its BPA risk assessment published in 2014, the EFSA expert group did not consider the study as being relevant enough for assessing the risk of BPA. The report only briefly mentioned Mok-Lin et al.'s study, which they included together with other studies in the weight-of-evidence approach (EFSA, 2010). Contrary to the ANSES's report (2011) and to our Qualichem assessment, the EFSA's report from 2010 deemed Mok-Lin et al. (2010) to have "*no relevance for the assessment of BPA associated health effects in humans*" (p. 58). The EFSA's main argument was that "*there is not supporting evidence from animal studies on the biological plausibility of the relationship between BPA low-exposure and female fertility*". Further criticism of the study's quality referred to the treatment of confounding factors, the number of BPA samples, the mistiming of BPA sampling, and the limited generalizability of the study's findings to women who are not receiving fertility treatment.

Our second case study, Sugiura-Ogasawara et al. (2005), shows a similar pattern as regards quality judgments expressed by different experts in different procedural contexts. Based on Qualichem responses, scores on all 37 criteria showed controversy and 34 of them were also critical, it was therefore a **low quality study** (Fig. 2). The study was not mentioned in the EFSA reports (2010, 2014).[6] Our Qualichem characterization was consistent with the level of quality expressed by ANSES's experts,[7] who noted "*study not taken into consideration since it has major methodological limitations*". Their report detailed flaws according to several quality criteria:

- small population size,
- inappropriate choice of the control group,
- inappropriate analytical method for measuring BPA (ELISA),
- lack of consideration given to confounding factors,
- incorrect choice of statistical test for analyzing the results,
- misinterpretation of the results (median serum levels identical in both groups) (8, p. 101).

All these criteria were also identified to be of low quality (median score in the red area) with Qualichem, which additionally identified 13 others in the same situation (aggregated low scores showing convergence among experts about their low quality). Whereas the global assessment of the study was similar between ANSES's experts and Qualichem interviewees ("bad quality"), the underlying reasons were nevertheless relatively different.

4. Discussion and conclusion

Our comparison of Qualichem responses with the reports of ANSES (2011) and EFSA (2010, 2014) showed that the level of heterogeneity in our respondents' answers is much higher than what was reflected in these documents, where study quality assessment is reported as a result of the whole expert group, without differentiation between the views of particular scientists. That heterogeneity is a normal feature of any expert group, which nevertheless is lost during the processes of consensus-based institutionalized risk assessment.

Our work confirms previous findings published in the literature about the heterogeneity of scientists' responses about the quality of a study (see Introduction). Even when they use the same criteria, and when their professional background is similar, scientists can make largely different judgments about what constitute "best scientific practices" and "quality". Any expert group will inherently show such heterogeneity. However, divergent viewpoints are the very source of intellectual richness and strive for the most robust knowledge in science, so they are not problematic in themselves. What becomes problematic however is when they are disappearing behind forced consensus while the actual scientific disagreement is unresolved.

The treatment of divergent viewpoints in global environmental assessments (GEA) has been recently addressed by Kowarscha et al. (2017), but their focus was on the various actors involved, including scientists but also industry, NGOs and different institutions. Furthermore, the whole GEA process was considered, from the development of the mandate and scope to writing the summary for policy makers. Our paper focuses very particularly on the treatment of divergent opinions between scientists, exclusively in their technical, scientific expression (even while value-laden assumptions may be embodied, while unexpressed, in such views).

Using an original empirical setting called Qualichem_epi, and two epidemiological case studies about BPA's effects on human health, we have shown that expert judgment plays an essential role in managing uncertainty and deciding what "quality" of a study actually means, and in selecting the scientific information that will be labeled as "reliable evidence" for inclusion in the decision-making process.

We compared the evaluation of study quality by academic scientists alone (who were not subject to any procedural rules) with quality assessment of the same studies made by expert groups in two agencies, the EFSA and ANSES (where experts worked according to procedural rules specific to each agency). Thus, we highlighted, by difference, the role of procedural rules in force in agencies. Indeed, conclusions of the expert groups differed between agencies and from our interviewees'. Concerning Mok-Lin et al. (2010), whereas the judgment of our interviewees indicated that it was of "average quality", ANSES (2011) considered this study to be "*of high quality or having no major methodological limitations*", whereas EFSA (2010) judged it to have "*no relevance for the assessment of BPA associated health effects in humans*". Concerning Sugiura-Ogasawara et al. (2005), both our experts and ANSES (2011) considered its quality to be low, whereas EFSA did not mention it.

The main finding of our analysis is the existence of substantial heterogeneity of scientists' judgements about the quality of epidemiological studies. This heterogeneity is not present anymore in reports produced by expert groups, where results are presented under the collective signature of all the scientists involved. Artificial reduction of the experts' judgmental heterogeneity, as current procedures in health agencies do, may lead to loosing essential information for the quality of risk assessments.

While using BPA as a case study for reasons of convenience (easiness to identify the relevant regulatory documents and experts), our objective is to propose a method that is generic enough for being adapted to any other case of regulatory chemical risk analysis.

Both judgments on uncertainty and procedures contribute to communication between experts and with the agency, allowing mutual adaptation between the different competences, disciplines and worldviews in the group and among the agency's employees. The "side effect" is that the combination of group-specific patterns of uncertainty management and particular working procedures inevitably produces context-dependent results, which may thus differ from one group to another and from one agency to another, and opens the way to criticism from the academic community.

[6] EFSA (2010) report deals with the literature between 2007 and 2010, as a previous opinion had been expressed by EFSA in 2006. However, the EFSA (2006) report, Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the Commission related to 2,2-BIS(4-HYDROXYPHENYL)PROPANE (Bisphenol A), did not include any reference to the Sugiura-Ogasawara et al. (2005) study either.

[7] In Appendix D in Supplementary material, we detail the arguments made by our expert respondents regarding the critical criteria identified for each of the two case studies, which enabled comparison with the criticisms (outlined above) made by the expert groups of ANSES and EFSA.

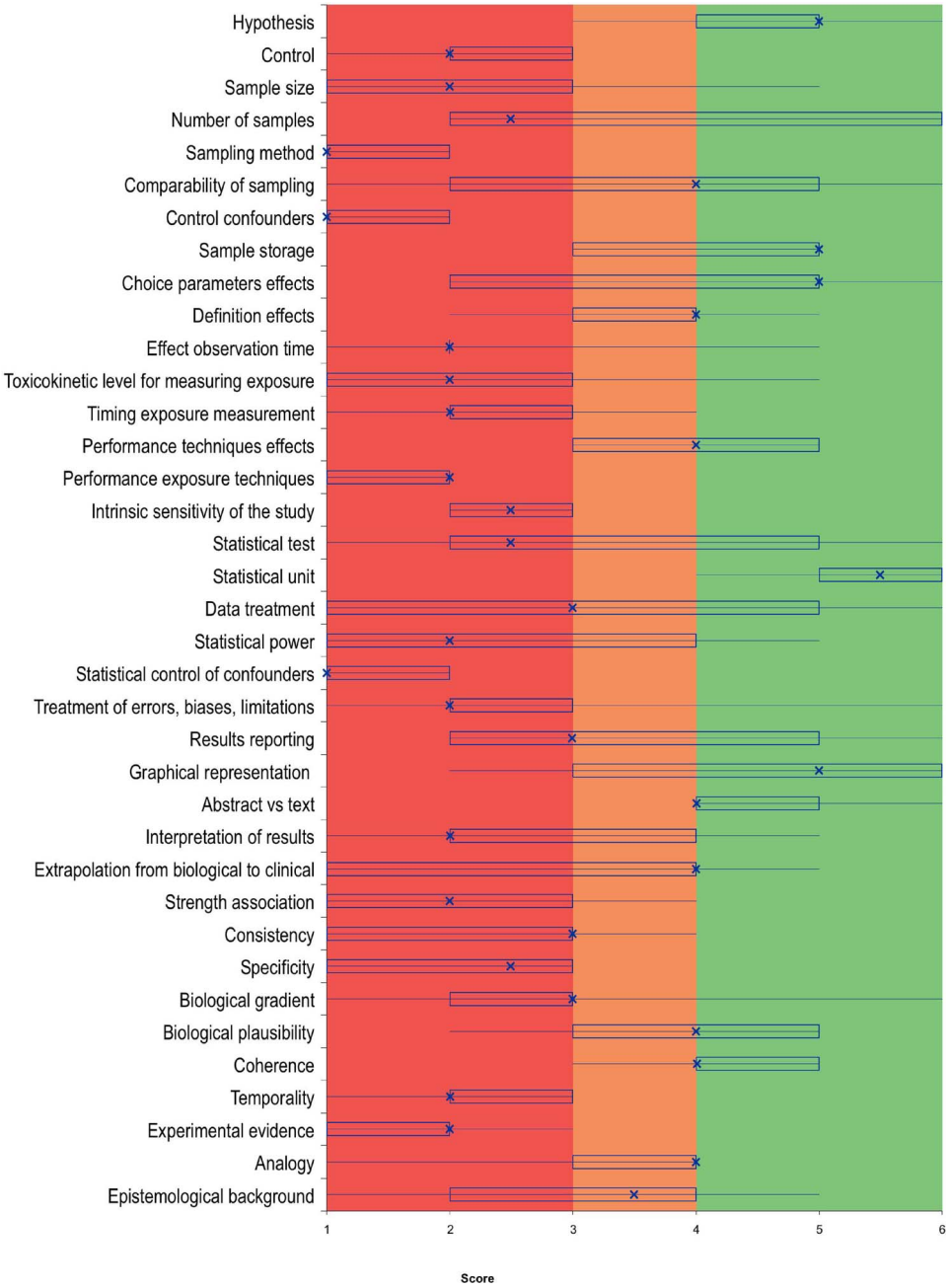

Fig. 2. Quality assessment of Sugiura-Ogasawara et al. (2005), using Qualichem_epi with seven respondents.

Acknowledgments

This work was supported by the French Ministry of Ecology, of Sustainable Development, of Transport and Housing (MEDDTL) in the framework of the PNRPE 2010 programme (URL: http://www.pnrpe.fr/), as part of the project "Toolkit for uncertainty and knowledge quality analysis of endocrine disruptors' risk assessments: the case study of Bisphenol A" (DICO-Risk). We are grateful to Céline Vaslin for help with the figures, to Kara Lefevre for stylistic and linguistic improvements and to two anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.envsci.2018.03.010.

References

ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail), 2011. Health effects of bisphenol A. URL: https://www.anses.fr/fr/system/files/CHIM-Ra-BisphenolAEN.pdf, Accessed 21 December 2017.

ANSES (Agence Nationale de Sécurité Sanitaire, Alimentation, Environnement, Travail), 2012. Code de déontologie de l'expertise. URL : https://www.anses.fr/fr/system/files/ANSES-Ft-CodeDeontologie.pdf. Accessed on 1st March 2018.

ANSES (Agence Nationale de Sécurité Sanitaire, Alimentation, Environnement, Travail), 2013. *Perturbateurs Endocriniens – Évaluation des risques du bisphénol A (BPA) pour la santé humaine. Tome 1*. URL: http://www.anses.fr/fr/content/bisph%C3%A9nol-l%E2%80%99anses-met-en-%C3%A9vidence-des-risques-potentiels-pour-la-sant%C3%A9-et-confirme-la. Accessed 21 December 2017.

ANSES (Agence Nationale de Sécurité Sanitaire, Alimentation, Environnement, Travail), 2016. Avis n° 2016-2 relatif à la prise en compte des positions minoritaires [Saisine 14]. URL : https://www.anses.fr/fr/system/files/DEON-Ft-2016002.pdf. Accessed on 1st March 2018.

Barthes, Y., 2014. L'expertise scientifique vue de l'intérieur : le groupe de travail « Radiofréquences » de l'Afsset (2008–2009). Environment, risque et santé 13 (1), 28–39. http://dx.doi.org/10.1684/ers.2013.0673.

Beauchamp, T.L., 1987. Ethical theory and the problem of closure. In: Engelhardt Jr.H.T.,

Caplan, A.L. (Eds.), Scientific Controversies: Case Studies in the Resolution and Closure of Disputes in Science and Technology. Cambridge University Press, Cambridge, pp. 27–48.
Beronius, A., Ruden, C., Hakansson, H., Hanberg, A., 2010. Risk to all or none? A comparative analysis of controversies in the health risk assessment of bisphenol A. Reprod. Toxicol. 29 (2), 132–146. http://dx.doi.org/10.1016/j.reprotox.2009.11.007.
Briggs, D.J., Sabel, C.E., Lee, K., 2009. Uncertainty in epidemiology and health risk and impact assessment. Environ. Geochem. Health 31, 189–203. http://dx.doi.org/10.1007/s10653-008-9214-5.
Deeks, J.J., Dinnes, J., D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M., Altman, D.G., International Stroke Trial Collaborative Group, European Carotid Surgery Trial Collaborative Group, 2003. Evaluating non-randomised intervention studies. Health Technol. Assess. 7 (27), iii–x. http://dx.doi.org/10.3310/hta7270. 1-173.
Demortain, D., 2013. Regulatory toxicology in controversy. Sci. Technol. Hum. Values. http://dx.doi.org/10.1177/0162243913490201.
Dor, F., Multinger, L., Doornaert, B., Lafon, D., Duboudin, C., Empereur-Bissonnet, P., Lévy, P., Bonvallot, N., 2009. The French approach to deriving toxicity reference values: an example using reprotoxic effects. Regul. Toxicol. Pharmacol. 55, 353–360. http://dx.doi.org/10.1016/j.yrtph.2009.08.006.
EFSA (European Food Safety Authority), 2010. Scientific opinion on bisphenol A: evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the Danish risk assessment of bisphenol A. EFSA J. 8 (9), 1829. URL: http://www.efsa.europa.eu/fr/efsajournal/doc/1829.pdf Accessed 21 December 2017.
EFSA (European Food Safety Authority), 2014. Draft Scientific Opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. Endorsed for public consultation draft scientific opinion URL: http://www.efsa.europa.eu/fr/consultations/call/140117.pdf Accessed 21 December 2017.
EFSA, 2015. Scientific opinion on the risks to public health related to the presence of bisphenol A (BPA) in foodstuffs. EFSA J. 13 (1), 3978. URL: http://onlinelibrary.wiley.com/doi/10.1002/efs2.2015.13.issue-1/issuetoc Accessed 21 December 2017.
EFSA, 2017. Decision of the Management Board of the European Food Safety Authority concerning the establishment and operations of the Scientific Committee, Scientific Panels and of their Working Groups. URL: https://www.efsa.europa.eu/sites/default/files/paneloperation170601.pdf. Accessed 1st March 2018.
Funtowicz, S., Ravetz, J., 1990. Uncertainty and Quality in Science for Policy. Kluwer Academic Publishers, Dordrecht.
Fujigaki, Y., Tsukahara, T., 2011. STS implications of Japan's 3/11 crisis. East Asian Sci. Technol. Soc.: An Int. J. 5, 1–15. http://dx.doi.org/10.1215/18752160-1411264.
Funtowicz, S.O., Ravetz, J.R., 1993. Science for the post-normal age. Futures 25 (7), 739–755. http://dx.doi.org/10.1007/978-94-011-0451-7_10.
Horel, S., Bienkowski, B., 2013. Special report: Scientists critical of EU chemical policy have industry ties. Environmental Health News. September 23.
Jasanoff, S., 1990. The Fifth Branch. Science Advisers as Policymakers. Harvard University Press, Cambridge, Massachusetts, London, England.
Knol, A.B., Slottje, P., Van der Sluijs, J., Lebret, E., 2010. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environ. Health 9 (19). http://dx.doi.org/10.1186/1476-069X-9-19.
Kowarscha, M., Flachslanda, C., Gararda, J., Jabbour, J., Riousset, P., 2017. The treatment of divergent viewpoints in global environmental assessments. Environ. Sci. Policy 77, 225–234. http://dx.doi.org/10.1016/j.envsci.2017.04.001.
MacKenzie, D., 1990. Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance. MIT Press.
Maxim, L., Van der Sluijs, J.P., 2011. Quality in environmental science for policy: assessing uncertainty as a component of policy analysis. Environ. Sci. Policy 14 (4), 482–492. http://dx.doi.org/10.1016/j.envsci.2011.01.003.
Maxim, L., Van der Sluijs, J., 2014. Qualichem in vivo: a tool for assessing the quality of In vivo studies and its application for bisphenol A. PLOS One. http://dx.doi.org/10.1371/journal.pone.0087738.
Mok-Lin, E., Ehrlich, S., Williams, P.L., Petrozza, J., Wright, D.L., Calafat, A.M., Ye, X., Hauser, R., 2010. Urinary bisphenol A concentrations and ovarian response among women undergoing IVF. Int. J. Androl. 33, 385–393. http://dx.doi.org/10.1111/j.1365-2605.2009.01014.x.
Noelle-Neumann, Elisabeth, 1986. The Spiral of Silence. University of Chicago Press, Chicago.
Roth, N., Ciffroy, P., 2016. A critical review of frameworks used for evaluating reliability and relevance of (eco)toxicity data: perspectives for an integrated eco-human decision-making framework. Environ. Int. 95, 16–29. http://dx.doi.org/10.1016/j.envint.2016.07.011.
Rudén, C., 2001. Interpretations of primary carcinogenicity data in 29 trichloroethylene risk assessments. Toxicology 169, 209–225. http://dx.doi.org/10.1016/S0300-483X(01)00525-X.
Samuel, G.O., Hoffmann, S., Wright, R.A., Lalu, M.M., Patlewicz, G., Becker, R.A., DeGeorge, G.L., Fergusson, D., Hartung, T., Lewis, R.J., Stephens, M.L., 2016. Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: a scoping review. Environ. Int. 92, 630–646. http://dx.doi.org/10.1016/j.envint.2016.03.010.
Strand, R., 2017. Post-normal science. In: Spash, C.L. (Ed.), Handbook of Ecological Economics: Nature and Society. Routlegde, London, pp. 288–298.
Sugiura-Ogasawara, M., Ozaki, Y., Sonta, S., Makino, T., Suzumori, K., 2005. Exposure to bisphenol A is associated with recurrent miscarriage. Hum. Reprod. 20 (8), 2325–2329. http://dx.doi.org/10.1093/humrep/deh888.
Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gøtzsche, P.C., Mulrow, C.D., Pocock, S.J., Poole, C., Schlesselman, J.J., Egger, M., for the STROBE initiative, 2007. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Ann. Intern. Med. 147 (8). http://dx.doi.org/10.1371/journal.pmed.0040297. W-163-W-194.
Van Eijndhoven, J.C.M., Groenewegen, P., 1991. The construction of expert advice on health risks. Soc. Stud. Sci. 21, 257–278.
Van der Sluijs, J.P., Van Eijndhoven, J.C.M., 1998. Closure of disputes in the assessments of climate change in the Netherlands arena. Environ. Manage. 22 (4), 597–609. http://dx.doi.org/10.1007/s002679900131.
Van der Sluijs, J.P., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J., Risbey, J., 2005. Combining quantitative and qualitative measures of uncertainty in model based environmental assessment: the NUSAP system. Risk Anal. 25 (2), 481–492. http://dx.doi.org/10.1111/j.1539-6924.2005.00604.x.
Van der Sluijs, J.P., Petersen, A.C., Janssen, P.H.M., Risbey, J.S., Ravetz, J.R., 2008. Exploring the quality of evidence for complex and contested policy decisions. Environ. Res. Lett. 3 (024008), 9. http://dx.doi.org/10.1088/1748-9326/3/2/024008.
Van der Sluijs, J.P., Van Est, R., Riphagen, M., 2010. Beyond consensus: reflections from a democratic perspective on the interaction between climate politics and science. Curr. Opin. Environ. Sustain. 2 (5-6), 409–415. http://dx.doi.org/10.1016/j.cosust.2010.10.003.
Vandenberg, L.N., Maffini, M.V., Sonnenschein, C., Rubin, B.S., Soto, A.M., 2009. Bisphenol-A and the great divide: a review of controversies in the field of endocrine disruption. Endocr. Rev. 30 (1), 75–95. http://dx.doi.org/10.1210/er.2008-0021.
Westreich, D., 2012. Berkson's bias, selection bias, and missing data. Epidemiology 23 (1), 159–164. http://dx.doi.org/10.1097/EDE.0b013e31823b6296.
World Health Organization, 2000. Evaluation and Use of Epidemiological Evidence for Environmental Health Risk Assessment. Guideline Document. URL: 2017URL:. Accessed on 21 December 2017. http://www.euro.who.int/_data/assets/pdf_file/0006/74733/E68940.pdf.