Routledge
Taylor & Francis Group

Journal of Environmental Planning and Management

ISSN: 0964-0568 (Print) 1360-0559 (Online) Journal homepage: http://www.tandfonline.com/loi/cjep20

Procedural influences on scientific advisory work: the case of chemical hazard characterization

Laura Maxim

To cite this article: Laura Maxim (2018): Procedural influences on scientific advisory work: the case of chemical hazard characterization, Journal of Environmental Planning and Management, DOI: 10.1080/09640568.2017.1407299

To link to this article: https://doi.org/10.1080/09640568.2017.1407299

Published online: 08 Jan 2018.

Journal of Environmental Planning and Management, 2018
https://doi.org/10.1080/09640568.2017.1407299

Procedural influences on scientific advisory work: the case of chemical hazard characterization

Laura Maxim *

Institut des Sciences de la Communication, CNRS/Université Paris Sorbonne/UPMC, Paris, France

(*Received 21 February 2017; final version received 10 November 2017*)

The quality of science for policy depends as much on the robustness of available scientific knowledge as it does on the procedural settings and working procedures in safety agencies. Using a report on Bisphenol A as a case study, and a set of original criteria, we provide an understanding of procedural influences on the results of scientific advisory committees and about literature reviews for chemical hazard characterization. Expert elicitation revealed that three aspects are critically important for the results of the advisory activity and for the selected case study: the method used to combine different studies, the interpretation of the review results in terms of level of evidence and conclusiveness, and the choice of uncertainty factors. Our results also show how procedural settings and working procedures can promote the invisible influence of values and policy on scientific advisory activities.

Keywords: chemical risk; uncertainty; quality; science for policy; risk assessment

1. Introduction

For policy-makers, the scientific advisory activity is a tricky tool, which can be useful for displaying a legitimate basis for decisions related to scientific and technical questions, but which could also have a boomerang effect if experts produced results that are contrary to an intended political line. According to a rhetoric image of science for policy, expert committees identify and synthesize the best knowledge available, which they provide to decision-makers who use it to balance competing interests (the "linear model", Pielke 2007). Terms such as "science-based policy" and "evidence-based policy" used among others by the European Commission or "science-based trade discipline" backing the World Trade Organization (WTO) practices (Bonneuil and Levidow 2011) suggest a sequential temporality where knowledge production is separated from, and precedes, decision-making. This temporality is reflected in the distinction between "risk assessment" and "risk management", assuming that the resolution of uncertainty is a policy issue, and therefore a prerogative of policy-makers (Silbergeld 1991). However, scientific advisors are also confronted with uncertainty and have to make choices about the relevance of the available knowledge in the very process of risk assessment (Wickson and Wynne 2012).

Thus, in reality, the separation of temporalities and responsibilities between scientific knowledge producers and decision-makers is much less straight. Experts are not outside the decision process, even if they might want to be, they are one of its stakeholders.

*Email: laura.maxim@cnrs.fr

The science and technology studies (STS) literature analyzed the scientific advisory activity as the source of authority and hence of socially legitimate power, and insisted that there is no way to separate science from values in the policy arena. Any such separation is artificial, temporary and serves the interests of the group who draw that line (Cozzens and Woodhouse 1994).

A number of authors have analyzed the influence of values in science for policy and, based on empirical criticism of the value-free model (Longino 1990), suggested accounting for these in scientific advisory activities (Sarewitz 2004; Wickson and Wynne 2012; Elliott and Resnik 2014; Elliott 2017) through transparent communication (Krimsky 2000; Van der Sluijs, Van Est, and Riphagen 2010; Wickson and Wynne 2012; Elliott and Resnik 2014) and a normative distinction between acceptable and non-acceptable value influences (Douglas 2009; Elliott 2017). Values were found to be essential at many steps of the technical thinking, and to have a crucial role when uncertainty leaves open questions about, for example, the level of evidence to be considered "enough" for risk assessment or the appreciation of data quality – when a multitude of diverse and multidisciplinary studies are available. Douglas (2009) distinguished between the direct and indirect roles of values in scientific reasoning and articulated several constraints on these roles, allowing a differentiation between "acceptable science and politicized science, between sound and junk science" (15). In the same vein, Elliott (2017, 10) established three conditions "for bringing values into science in an appropriate fashion": transparency, representativeness for major social and ethical priorities, and engagement with stakeholders.

Accounting for the particularities of scientific advice as compared with academic science, Jasanoff (1990) argued that advisory committees are platforms for negotiating scientific and political conflicts using the codified language of technical choices. How uncertainty is dealt with in a particular context and by a particular group of scientists determines what is considered to be "good science", which has the political function of answering the question: is the risk severe enough to warrant immediate regulatory action (Jasanoff 1990)?

Failing to acknowledge value influences was found to distort the democratic political process, as trade-offs between the interests of those concerned were hidden under the cover of scientific arguments. While potentially influencing scientists in subconscious ways (Elliott and Resnik 2014), values would still be compatible with objectivity (Douglas 2009; Elliott and Resnik 2014) when ethical standards were respected ("honest science" for Krimsky 2000, "honest broker" for Pielke 2007). According to Sarewitz (2004), policy should intervene ahead of technical input, and should set the value bases of disputes underlying environmental controversies before science comes into play to assess environmental problems.

In theory, scientific advisors are not supposed to operate trade-offs between the interests of the various stakeholders concerned by a risk. In practice, however, consciously or otherwise, they may do so.

But how do expert knowledge and its socio-political background merge so closely that they become inseparable? By which mechanisms do they merge? Is it because of this merger that paradoxical situations appear, such as different and even opposite conclusions from expert groups, even if the scientific knowledge base is the same?

Such disparities between different expert groups are particularly prominent for the case of Bisphenol-A (BPA), one of the most controversial chemicals over the last decade because of its suspected endocrine-disrupting properties. The in-depth analysis of several expert reports on BPA by Beronius *et al.* (2010) showed that conclusions about health

risks of BPA vary dramatically – from "there is no risk to any part of the population" to "there is risk to the entire population."

Several EU-level expert committees assessed BPA from 2000 to 2014, but none concluded on risks from this substance. The exposure of the European population was always considered to be above the Tolerable Daily Intake (TDI), set at 0.05 mg BPA/kg body weight (bw)/day until 2010 (EFSA 2010) and temporarily at 5 μg/kg bw/day in 2014 (EFSA 2014). This contradicts France's risk assessment, where a report from the French Agency for Food, Environmental and Occupational Health & Safety (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail [ANSES]) outlined scenarios in which there is risk of four critical effects for fetuses of pregnant women exposed to BPA. These effects are due to exposure of pregnant women, can appear after birth through childhood or adulthood, and include increased susceptibility of the mammary gland to cancer; effects on metabolism and obesity; neurological effects, particularly on learning and memory; and effects on the female reproductive system (ANSES 2013). Similarly, different EU experts have disagreed about the developmental neurotoxicity of BPA. Three countries (Denmark, Sweden, and Norway) have determined that certain studies of low-dose effects of BPA were sufficiently reliable for regulatory use; EU RAR (2008) determined the same set of studies to be inconclusive.

As highlighted by Beronius *et al.* (2010), such differences are not simply the result of time, i.e. of some trend from "no risk" to "risk" because of accumulating toxicological knowledge. These authors attribute the differences to the diverging ways in which experts considered low-dose effects and uncertainties surrounding the significance of these data for health risk assessment. Still, the question remains: why did experts consider low-dose effects and uncertainties differently?

The STS literature extensively shows that the quality of science that is used to set policy depends as much on the robustness of available scientific knowledge as it does on the procedural rules operating in advisory organizations (Demortain 2009; Lentsch and Weingart 2011; Maxim and Van der Sluijs 2011). Used outside its sites of production in academia, scientific knowledge changes nature when mobilized in a different institutional framework. It is through the procedures that reign the activity of scientific advising (e.g. multiple experts from different disciplines chosen by, or in interaction with, policy-makers, consensus as a general rule for producing results, close interaction between scientists from academia and experts from other professional backgrounds including health agencies, use of data from different sources including industry) that science and policy merge to result in knowledge of a very different nature compared to academic science. Procedural arrangements, legal contexts and framing of the questions to address, produce a particular, situated kind of scientific knowledge (Jasanoff 1995), responding to the specific needs of political bodies. In the European Union, the literature shows that expert groups are created not only for gathering knowledge, but also for anticipating reactions to Community initiatives and to secure the Commission's projects using scientific legitimacy (Robert 2010).

In the WTO arena, for the case described by Bonneuil and Levidow (2011), the selection criteria for choosing the experts privileged the anticipation of their potential views more than the appreciation of their competence or of potentially conflicting interests.

In health agencies, the question to be addressed by an expert group, the resources allowed, and the selection of the participating experts are set by policy-makers in interaction with the agency. There are currently no clear-cut criteria for assessing competence during the selection process. However, the literature supports the idea of a distinctive influence of

the disciplinary backgrounds and academic experience of the experts. For the case of BPA, Beronius *et al.* (2010) observed that the majority of the authors of the Chapel Hill assessment, who concluded that there was a risk for the entire population, have published scientific articles on BPA. This was not the usual case for the expert groups having reached the opposite "no risk" conclusion. Similarly, Maxim and Van der Sluijs (2014) showed that the discipline could influence expert judgment about the quality of a scientific article on BPA. What is considered a "good paper" can be different between specialists in endocrinology and scientists trained in other disciplines relevant for chemical risk assessment. The appreciation of the usefulness of scientific articles and reports for the advisory activity on this substance thus depends on the share of the two disciplines in an expert group. Furthermore, specific references govern the quality assessment of the existing scientific knowledge in toxicology as regulatory science, such as respect for the Organisation for Economic Cooperation and Development (OECD) standardized protocols, which can come into conflict with academic references for scientific quality (Demortain 2013; Maxim and Van der Sluijs 2014).

In this paper, we test the hypothesis that, in addition to the institutionalized procedural settings addressed earlier, each expert group's working procedures will influence its results. In certain cases, such influences can be decisive for orienting the results either on the "risk" or "no risk" side, which could explain differences between expert groups, as found by Beronius *et al.* (2010).

We make the distinction between procedural settings and working procedures. *Procedural settings* are rules for advisory activity, which are institutionalized through recommendations, guidelines, and formal or informal constraints on the advisory work. Examples of procedural settings are the processes of expert selection (including formal and informal criteria), the processes of formulating the questions to be asked of the experts, the time-frame and duration of the work, the processes of communicating and accounting for conflicts of interest in the routine advisory work, the rules for organizing the collective work of the experts (e.g. consensus rule and minority opinions), and the institutionalized methods for organizing the available evidence (e.g. systematic reviews, weight of evidence, uncertainty assessment, and communication). Procedural settings can be informal and result from the experience accumulated in a specific health agency, which becomes a kind of informal 'jurisprudence' for its routine work (e.g. giving priority to studies following OECD guidelines).

Procedural settings are outside the influence of the experts or, at most, can be challenged only by concealing an important personal investment. However, in some cases, working procedures are setting a specific use of procedural settings (e.g. minority opinions can be institutionalized in a health agency, but their use in the routine work of particular expert committees can be informally discouraged).

Under "working procedures", we include here the set of informal rules that experts use within a group in order to function together, producing common and coherent results. Working procedures are put in place by the experts themselves, in interaction with agency employees, for the organization of those aspects of the work which are not institutionally formalized through procedural settings, or for which procedural settings are flexible enough to be adapted to each particular advisory activity (e.g. method employed for assessing the quality of the studies, criteria for validating/invalidating a study, procedure for extracting data from a study…). For example, in order to combine different available studies, either institutionalized methods such as weight of evidence can be used, or original methods can be developed by each expert committee (e.g. the method employed by ANSES for declaring effects "recognized", "controversial," or

"suspected", see Section 4). For the reporting of assumptions and other uncertainties – either institutionalized methods such as probabilistic analyses may be available, or specific methods can be employed (e.g. reporting assumptions throughout the text in qualitative ways, without preexisting procedural rules for doing so).

While procedural settings are an expression of institutionalized political choices (e.g. favoring consensus, choosing experts through a process involving both technical personnel and the higher hierarchical levels of health agencies), working procedures are the expression of the very specific choices of the experts themselves and of the agency employees working with them. Depending on the experts, working procedures are a variable mixture of both technical content and underlying personal values (e.g. concern for the protection of public health, concern for potential economic impacts of risk assessments). Working procedures can be either formalized in writing (e.g. choices about how to combine different studies, as in ANSES [2013]) or informal (e.g. evaluation of the quality of the studies available can be done in different ways: either studies can be distributed among experts in such a way that each study is assessed by one expert, or two experts from similar disciplines evaluate each study, or two experts from different disciplinary or theoretical orientations can evaluate the same studies, or the most problematic studies as previously selected by agency employees, etc.).

Each group establishes its own working procedures, based on previous experience of the health agency and on discussion and negotiation with and between the experts involved, both from the agency and external to it. Through the work of a group, experts discuss the setting of rules to be used when these are not already institutionalized. For this reason, the final working procedures depend on the individuals who comprise that group, and can differ from one group to another. Such differences could be minor and have marginal influences on the results produced by the group, or could be significant and produce important effects on the results, as compared to other groups. In other words, different experts – selected for responding to objectives specific to certain political arenas – could use the same scientific knowledge in different ways, which could lead to different results.

For testing our hypothesis, we used as a case study the report produced by a European Food Safety Authority (EFSA) expert group in 2010, and feedback from six scientists (see Section 2.3. for the selection procedure). The EFSA 2010 report was produced in response to the European Commission's demand to assess the scientific literature produced between 2007 and 2010 in terms of relevance for the risk assessment of BPA.[1] Based on a literature review, the ultimate conclusion was that there was no critical effect identified below 5 mg/kg bw/day and so no change was needed in the existing TDI.

This EFSA report was highly controversial. The Health and Environment Alliance (HEAL) declared itself "shocked" by EFSA's appreciation of critical effects and of the level of evidence that could be established based on the available knowledge. HEAL considered that "BPA may play a major role in major chronic diseases, such as breast cancer and diabetes" (Chemtrust *et al*. 2010). Similarly, Chemtrust estimated that the level of evidence was sufficient to drive political action. Réseau Environnement Santé (RES), a French NGO, was deeply concerned about the way in which studies were excluded from the literature review and the subsequent quality assessment, stating that the EFSA group reached their conclusions by rejecting 95% of the literature (RES 2010). RES also stressed effects that they felt were critical, but that had not been considered so by the EFSA group: breast and prostate cancer, reproductive toxicity, diabetes, obesity,

and behavioral disorders. These reactions were prolonging an initiative taken three months before the release of the EFSA report, when 41 non-profit organizations and 19 scientists from 19 countries sent a letter to the chair of that EFSA group.[2] This letter highlighted BPA effects on human health that the signatories considered critical and concluded that available knowledge was enough to drive political action to further restrict BPA exposure.

What was under dispute was the working procedures of the EFSA's expert committee, in particular identification of critical effects, identification of vulnerable groups, method used to validate or reject studies, and the relevance of the available studies for justifying any action.

We have developed a methodological framework for identifying and analyzing the range of procedural influences on the results of expert groups engaged in literature reviews for hazard characterization in chemical risk assessment. Section 2 presents this methodological proposal, Section 3 the results from applying this method to the EFSA (2010) report, and Section 4 discusses them and concludes.

2. Methods

2.1. Choice of case study

Focusing on the EFSA (2010) report was a methodological choice, as political stakes around this particular report are no longer topical (mid-2017). Indeed, our objective was not to contribute to criticism of a particular health agency such as ANSES or EFSA, or of a particular report, but to provide a generic framework for understanding and communicating procedural choices in scientific advisory activities on chemical hazard assessment. BPA and the EFSA (2010) report are just case studies.

2.2. Criteria for describing working procedures

The first step was to identify the relevant criteria able to describe the expert group's working procedures. EFSA's report aimed to provide an extensive literature review, therefore we developed the typology of such criteria (Figure 1) iteratively, starting from the existing literature on systematic reviews (Centre for Reviews and Dissemination 2008). In addition, we followed the main steps of the process of knowledge production of a literature review for hazard characterization. We also analyzed criticisms from academic scientists or NGOs of the EFSA report to help set our typology. Finally, to iteratively improve our initial typology, we incorporated feedback from six academic and health agency scientists – a sample size in line with the current literature on expert elicitation (Knol *et al.* 2010; Grigore *et al.* 2016) that recommends 6–12 experts.

Our criteria are assembled into six different classes (Figure 1) that fall into two general categories: Protocol for literature review, and Reporting and interpretation of the existing knowledge. These two categories include 23 criteria that are relevant to

- methodological aspects of systematic reviews;
- communication (such as results reporting);
- normative aspects of reviews, directly related to expert judgment (such as synthesis and interpretation of the review);
- the larger epistemological context of the science available on the studied hazards.

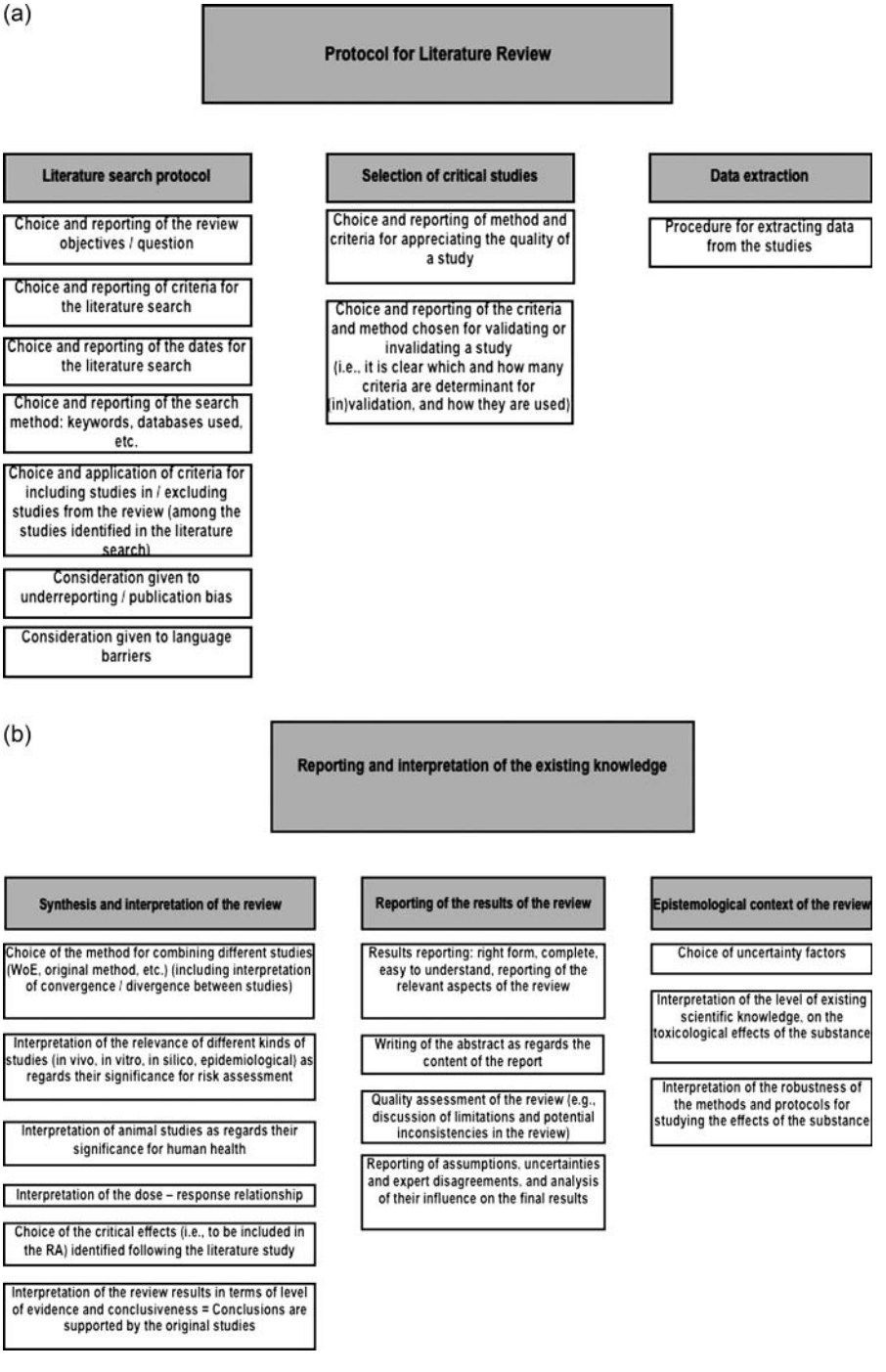

Figure 1. Criteria describing working procedures for chemical hazard characterization in expert groups.

2.3. Choice of respondents

We tested the validity of our typology with six scientists. Respondents were selected from among scientists who were specialists in BPA, endocrine disrupters or chemical risk assessment, and were knowledgeable about chemical risk assessment procedures and scientific advisory activities.

Following this process, we contacted 20 scientists by email. Six agreed to participate. Four respondents were employed by safety agencies in Europe and two were academics but were also acting as scientific advisors. All of them were competent for assessing the EFSA report, as they either published on BPA and/or were involved in expert activities on this substance or on endocrine disrupters.

Five had backgrounds and professional experience in toxicology and one in ecotoxicology. All were specialists in chemical risk assessment.

2.4. Elicitation protocol

We interviewed the respondents in 2012 and 2013. To prepare the interviews, we sent the report by email before the meeting to allow in-depth reading, and provided a printed copy during the meeting. In addition, we pasted relevant text from the study below each question.

To assess our respondents' judgment about the working procedures used by EFSA – specifically, how well they incorporated best scientific knowledge and practices – we presented each respondent with a question related to each of the criteria included in our typology.[3] For example, the first question of our protocol was: "Were the review objectives and question(s) chosen and reported in accordance with the best scientific practices?" The text from the report that refers to the review objectives and questions was copied below the question. The respondent was invited to answer using a Likert scale (Table 1) and to explain their response (Section 3).

Interviews were recorded and transcribed. We used the transcriptions to analyze the results (Section 3).

2.5. Graphical representation of respondents' assessments

To facilitate understanding and focus on the most important results, we have defined *controversial criteria* as those for which:

Table 1. Scale used for expert elicitation.

Answer	On a scale from 1 to 6, the answer corresponds to the score
Agree strongly	6
Agree moderately	5
Agree slightly	4
Disagree slightly	3
Disagree moderately	2
Disagree strongly	1
I cannot answer	CA
Not applicable	NA

- at least one respondent gave a score of 3 or less, or
- there is a difference of at least two points between any two scores.

The graphical representation of the results (see Figure 1) was built using an Excel file[4] and represents the controversial criteria. For each such criterion, it shows the range of scores as a line, the interquartile range of the scores as a rectangle, the median of the scores as an x, and three colored areas: red (for scores and medians < 3), orange (for scores and medians between 3 and 4), and green (for scores and medians > 4).

We consider the median of the scores assigned by all experts to be the "aggregated quality" for a single criterion, i.e. the quality of that criterion based on the aggregation of opinions of all experts. This is an indicator of majority views on aspects of working procedures. We divided aggregated quality into three levels:

- *High aggregated quality*: high scores with a median in the green area (>4);
- *Average aggregated quality*: moderate scores with a median in the orange area (ranging from 3 to 4);
- *Low aggregated quality*: low scores with a median in the red area (<3).

The figure provides two indicators of *heterogeneity of the responses*:

- the range between the minimum and the maximum score in the group of responding experts;
- the interquartile range.

3. Assessing the influence of working procedures in the EFSA 2010 report

The graphical representation of the six scientists' responses (Figure 2) shows only the 15 controversial criteria out of the total set of 23 criteria. The remaining criteria were not controversial according to our definition – they received scores of 5 or 6 representing medium or high agreement:

- Choice of the review objectives/question;
- Choice of the dates for the literature search;
- Procedure for extracting data from the studies;
- Interpretation of the relevance of different kinds of studies (*in vivo*, *in silico*, epidemiological) as regards their significance for the risk assessment;
- Results reporting;
- Concordance between abstract and the content of the report.

The respondents felt that the remaining two criteria were not appropriate for assessing this particular EFSA report: Consideration given to underreporting/publication bias, and consideration given to language barriers.

3.1. Criteria with median in the red area

The medians of three of the criteria represented in Figure 1 fall in the "red area", indicating general disagreement, among the respondents, with the working procedures of EFSA referring to this criterion.

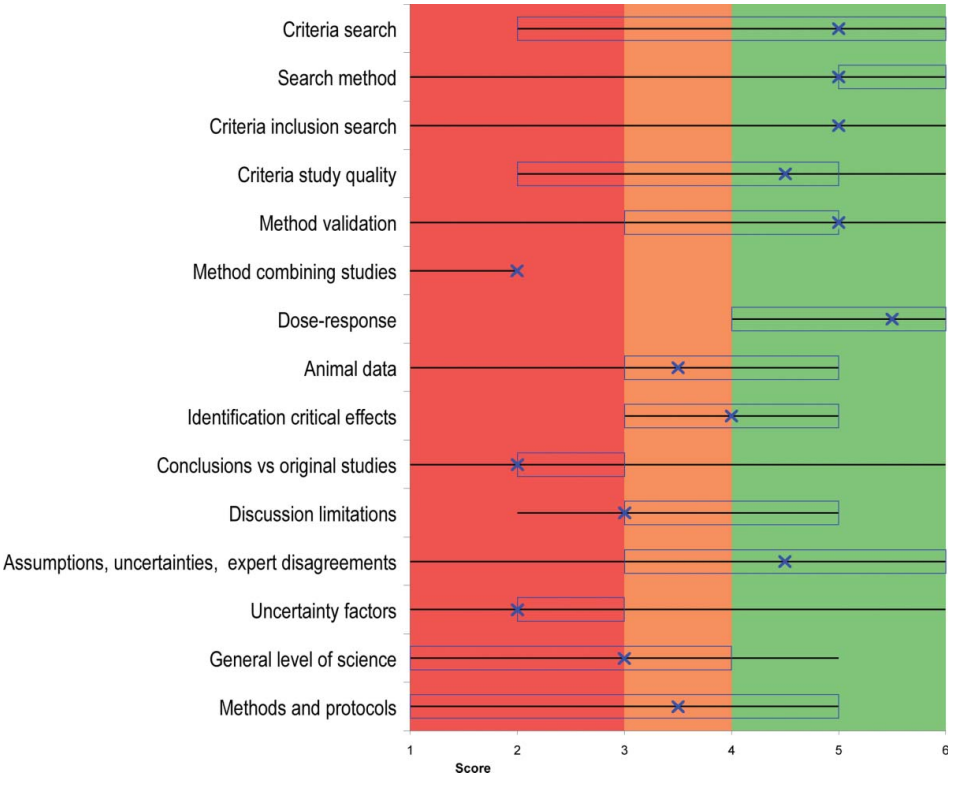

Figure 2. Assessment of working procedures used for producing the EFSA 2010 report, based on feedback from six scientists. See online colour version for full interpretation.

3.1.1. Choice of the method for combining different studies (short name: method combining studies)

All respondents scored this criterion below 3 – this showed that all disagreed with the EFSA expert group's method of combining the different studies. EFSA assessed the quality of each study individually, and made a separate decision about whether to include each in the risk assessment. The result was that all studies that showed effects below 5 mg/kg bw/day were rejected. The criticism of this method was that looking at each study in isolation ignored the larger context provided in the scientific literature.

No single study is perfect, and so the respondents argued that it is important to understand whether the different studies of acceptable (if imperfect) quality go in the same direction or whether they contradict one another. In addition, the respondents were concerned that EFSA did not explicitly describe the method used to drive their conclusions based on the original studies.

3.1.2. Interpretation of the review results in terms of level of evidence and conclusiveness, i.e. conclusions are supported by the original studies (short name: conclusions vs. original studies)

The respondents varied dramatically in their views of how accurately EFSA's expert committee interpreted the original studies – scores ranged from 1 to 6. Most respondents

felt that the EFSA experts' interpretation was incorrect, as it did not reflect the larger picture that arises when the different studies are brought together. The most critical respondents felt that the uncertainty around the low dose studies, despite their imperfections, should have led to expressions of doubt about the current TDI. Some respondents proposed an additional safety factor to incorporate these uncertainties.

The respondents who gave maximal scores for this criterion did not explain their responses.

3.1.3. Choice of uncertainty factors (short name: uncertainty factors)

Respondents also gave extremely varied scores for this criterion – ranging from 1 to 6. The median, though, is 2 so the aggregated score of the report based on this criterion is weak. The respondents who gave low scores for this criterion recommended an additional safety factor to account for the uncertainty related to studies that show effects at doses lower than the official no observed effect level (NOEL).

Two respondents recognized that the chosen uncertainty factor of 100 is standard, but noted that it is nevertheless arbitrary, and that its efficacy has not been demonstrated. On the contrary, some studies seemed to show that this uncertainty factor is not protective in certain situations.

Another respondent provided a maximal score of 6, based on the same argument – 100 is a standard uncertainty factor. This respondent generally framed their responses by referencing an institutional risk assessment perspective, in which standard norms in use (OECD testing guidelines, normalized safety factors, etc.) are expected to guarantee scientific quality.

3.2. Criteria showing important heterogeneity

In addition to the three criteria with the median in the red area, nine others showed important heterogeneity (differences as high as 4 or 5 between two respondents). This indicates that some of the experts are in disagreement with EFSA's working procedures, whereas others do agree with them.

3.2.1. Choice and reporting of criteria for the literature search (short name: criteria search)

The most critical respondents on this criterion felt that the EFSA report should have more precisely and explicitly described the literature search criteria and the procedure for including or excluding studies from the literature study (for example: Were full texts available for all papers? Were any papers excluded because of inaccessibility?). The respondents indicated that they could not tell whether papers were excluded from the study based on the selection criteria, or because they were not evaluated in the first place.

One of the respondents indicated having a friend in the EFSA committee, and, based on exchanges with this friend, our respondent determined that all studies had been considered. The interviewee therefore assumed, based on this "inside" knowledge of the EFSA expert group's work, that all the studies found in the literature had been included. However, the score was tempered, to recognize that the report is not clear about the procedure used.

Finally, another respondent gave a high score based on the assumption that an institution as prestigious as EFSA could not have made a mistake as basic as ignoring available studies.

3.2.2. Choice and reporting of the search method: keywords, databases used, etc. (short name: search method)

The most critical respondents noted that the method used for the literature search – and in particular the key words, the identified articles, and the papers retained – were not described in the report and therefore cannot be verified by an external reader.

One of the respondents who gave a high score to this criterion suggested that a prestigious institution such as EFSA probably used the most recent technologies, which made it possible for them to easily and exhaustively identify relevant studies in the scientific literature.

3.2.3. Choice and application of criteria for including or excluding studies in/from the review (short name: criteria inclusion search)

The respondent most critical of this criterion felt that the EFSA report did not appropriately consider studies that used non-oral exposure or those that show a statistically significant effect without showing a dose–response relationship. This respondent felt that the EFSA group's insistence on rejecting knowledge provided by these two kinds of studies was unjustified, and that more moderate language that recognized a certain degree of uncertainty about what these studies meant would have been more desirable.

Another respondent argued that there was a problem with the word "several" being used in reference to the number of doses considered relevant for including studies. This respondent argued that a more precise description was needed, which led him to temper his score (5).

3.2.4. Choice and reporting of method and criteria for appreciating the quality of a study (short name: criteria study quality)

Scores for this criterion ranged from 2 to 6. The most critical respondent felt that EFSA's criteria such as "absence of clear dose response curves" are not scientific; rather, they are regulatory.

Another respondent qualified the EFSA expert group's work as a search for even the smallest default that would make it possible to reject a study, instead of using a constructive approach aimed at getting the best information from the studies available. This strategy could only lead to rejection of all available studies, because there is no such thing as a perfect study. Furthermore, this approach loses the overall perspective provided by looking at the available scientific literature together. This respondent also stressed that experts generally have little time to devote to EFSA, which would have prevented them from looking at each study in detail. Therefore, some experts just adopt the comments expressed by their colleagues during the group meetings, and repeat them when they have to provide their own advice. Some experts propagate arguments that they consider "mainstream", without attentively reviewing the original studies themselves and forming their own opinions. This respondent called this the "parrot effect", and noted that the propagation of "parroted" ideas can have a negative influence on the scientific quality of expertise reports, and on science itself.

Another respondent, who was knowledgeable about the working procedures in health agencies in general, and at EFSA in particular, also noted that studies are not often

assessed by the whole group. Given the great volume of studies that must be assessed, each one is usually reviewed by one or two experts, who are responsible for reporting that study. Colleagues in the working group could react during the expert group meetings if they read the article. But, if they did not – and in most cases, experts do not all review all of the studies – the experts trust their colleague(s) and endorse their work on behalf of the group. This can give the impression that the whole group has reviewed a study, when the group opinion is really based on the views of just one or two of the experts. This criticism extends beyond the study quality criteria and refers to expert group working procedures that weaken the evaluation of study quality.

Another respondent expressed criticism over the typical preference of expert groups for OECD guideline studies. This respondent felt that if all scientific studies followed OECD guidelines, science would never move forward. Furthermore, OECD guidelines do not include the latest scientific advances and, on the contrary, are very old. Indeed, it takes a lot of time to renew an OECD guideline or propose a new one. Referring to OECD guidelines as a guarantee of scientific quality is more a procedural habit than an application of best scientific practices.

Finally, a respondent criticized the argument that the inherent design of cross-sectional studies makes it impossible to establish a causal relationship – this is true for all studies and would not be a justified argument for criticizing epidemiological cross-sectional studies. Indeed, causal relationships are often very difficult to demonstrate with a high degree of certainty with any kind of study, and therefore the lack of a causal relationship cannot become an argument for rejecting any particular kind of study. Cross-sectional data could contribute, together with other studies, to the global image of the trends of the risk.

3.2.5. Choice and reporting of the criteria and method chosen for validating or invalidating a study (short name: method validation)

Two respondents suggested that EFSA's experts invalidated available studies more on the basis of a "belief" than on science. According to these respondents, the report shows that the experts refused to believe the results communicated in the studies, and did not provide clear arguments for rejecting them. One of the respondents even suggested a mechanism for blinding the review process for controversial substances such as BPA by excluding the name of the substance and the measurement unit for the doses from the information communicated to reviewers on a particular study. In this way, the reviewer would not be in a situation to say that they simply cannot believe that such low doses can produce such effects.

Another respondent was critical of the fact that the report does not communicate the method used to validate or invalidate a study. This raises important questions: Which quality criteria were considered the most important for rejecting a study? How many such criteria were needed to exclude a study from the risk assessment? Should some information contained in a study be considered important enough to be used in the hazard characterization, even if the study as a whole was not qualitative enough? Even if they are not of the best quality, should some studies be considered as emerging signals, or hints, of potentially real effects?

Finally, three other experts suggested that even if the report does not communicate the method used to validate or invalidate a study, EFSA's experts probably did their best. However, besides their confidence in EFSA's work in general, these respondents did not give other reasons for their scores.

3.2.6. *Interpretation of animal studies as regards their significance for human health (short name: animal data)*

Scores for this criterion ranged from 1 to 5. The most critical respondents felt that extrapolation from animals to humans is at the core of chemical risk assessment and is not an appropriate reason to consider that animal studies' relevance for human health "cannot be assessed" (EFSA 2010, 1). In addition, assessing risk to human health is the only ethical reason to sacrifice animals during *in vivo* studies. Even if animal studies were not considered of high enough quality, saying that it is not possible to draw conclusions based on animal studies seems excessive.

3.2.7. *Reporting of assumptions, uncertainties and expert disagreements, and analysis of their influence on the final results (short name: assumptions, uncertainties, expert disagreements)*

Several respondents criticized the lack of consideration of the EFSA's expert group for uncertainty about studies that show effects at low doses. While one respondent felt that simply expressing minority opinions is enough to document disagreements between experts inside a group, another felt that the mechanism is too rigid and that only an expert with a very strong character would be willing to express a minority opinion. According to this respondent, even when they disagree, many experts may "give up" when their opinion differs from that of the majority, either through opportunism or by willingness to remain in line with the rest of the group.

3.2.8. *Level of existing scientific knowledge on the toxicological effects of the substance (short name: general level of science)*

Respondents varied in their assessments of the current level of scientific knowledge on BPA and of its potential to provide a good understanding of BPA toxicology. Scores ranged from 1 to 5, with a median of 3. In agreement with EFSA's working group, for some respondents, the ongoing questions about BPA is a clear indication of a high level of epistemic uncertainty; for others, there is enough scientific understanding available to draw conclusions about the risks of BPA (score 5).

3.2.9. *Robustness of the methods and protocols for studying the effects of the substance (short name: methods and protocols)*

As EFSA's group considered, most of the respondents felt that the scientific methods used to study BPA are neither homogenous nor, generally, robust. If heterogeneity is not necessarily a weakness of BPA toxicology, the lack of methodological robustness is a significant problem. It is difficult to interpret a wide range of different approaches at the same time, which is why one respondent prefers to rely on OECD studies – at least all OECD studies point in the same direction. However, other respondents insisted that recent studies have found that the scientific underpinnings of some OECD guidelines can be significantly improved, in particular for those intended to study neurobehavioral toxicity. A respondent noted that academic teams should better communicate their studies, as some problems with using them arise less from the protocol itself than from how it is reported. Testing companies, on the other hand, may not have the best working procedures even if they follow OECD guidelines. This might lead them to

inappropriately deal with scientific aspects that are not included in the guidelines, for example, failing to correctly manage the confounding impact of noise level during neurobehavioral *in vivo* testing.

4. Discussion and conclusion

The quality of expert hazard and risk assessments is influenced not only by the level and quality of the scientific knowledge available, but also by the procedural settings in health agencies and by working procedures in place in expert committees. Using a different protocol for literature search, criteria for assessing the quality of the studies, or method for drawing conclusions from the whole range of available studies, can lead to different results among expert groups and ultimately fuel controversies.

Previous criticism of the EFSA (2010) report in the light of systematic review techniques (Whaley 2013) concluded that it did not clearly state its objectives, did not pre-publish its protocol, did not consistently identify methods for locating data, incompletely stated the criteria for data analysis, did not transparently and consistently report criteria for assessing study quality, and lacked clarity in the synthesis of its results.

A different working procedure used at ANSES to select the relevant studies and identify critical effects of BPA led to opposite results as compared to EFSA 2010. At the opposite of EFSA's experts, who assessed the quality of each individual study and decided to validate only one study and reject all the others, ANSES' experts considered all the available information regarding a health effect and combined different studies according to their quality and similarity of results. When the results of multiple high-quality studies undertaken by different scientific teams converged, the effect was considered to be "recognized." When they diverged, the effect was considered to be "controversial." When studies having non-major methodological limitations converged, the effect was considered to be "suspected" and when they diverged, the effect was considered to be "controversial." The effects used for characterizing BPA hazards in the calculation of risk were "recognized" effects in animals (any recognized effect in humans being identified) and "suspected" effects in humans. Based on this working procedure, ANSES identified four critical effects of BPA (increased susceptibility of the mammary gland to cancer; effects on metabolism and obesity; neurological effects; and effects on the female reproductive system) in 2013 (ANSES 2013), whereas EFSA identified only one (changes in kidney and liver weight) just one year later, in 2014 (EFSA 2014).

Following political crises, such as the outbreak of Bovine Spongiform Encephalopathy, health agencies have been created in Europe as independent bodies for displaying separation between objective, science-based risk assessment and political, value-laden risk management (Demortain 2009). However, if the political influence is not as explicit as it used to be, it is always embodied in the fabric of the scientific advisory activity, where it enters through the mean of procedural framing of experts' selection and work, which we divided into procedural settings and working procedures.

Our distinction between procedural settings and working procedures is helpful for reflecting on the relative roles of scientific advisors and policy-makers. Scientific advisors cannot influence the whole context of their own work, procedural settings being institutionalized and usually beyond their scope. For example, the experts themselves rarely know how and why they have been chosen, or why the question submitted to them has been formulated in the way it is. However, each individual in a group can influence the formulation and communication of working procedures, although only to a limited extent, since each scientist's individual influence is mitigated by that of colleagues in the group.

One of the most direct influences of policy on scientific advisory activity is exerted through the choice of the experts included, and we included this in procedural settings. Criteria for selecting experts are flexible, so potentially influenced by the hierarchical politico-administrative layers in health agencies. Our paper does not account for the influence of the criteria applied for selecting the experts, but these criteria can be an important leverage for political influence inside expert groups (Michaels *et al.* 2002).

Furthermore, the influence of policy exerts through the means of the questions asked to the experts, that they are not supposed to challenge or modify, and through the expectation that the expert group will produce consensus opinions. Our expert elicitation (Figure 2) showed that experts' judgments are intrinsically heterogeneous. In expert groups, consensus-based procedures can favor strong personalities who take the lead in collective discussions – important and scientifically robust minority opinions can be lost in the process of reaching a common position (Van der Sluijs, Van Est, and Riphagen 2010; Maxim and Van der Sluijs 2014).

Established during the work of expert groups and through their interactions with employees of health agencies, working procedures are set for the organization of those aspects that are not included in procedural settings and remain barely accessible, or even invisible, to anyone not participating in the work of that expert group. While our empirical setting cannot allow us to conclude irrefutably that different groups will produce different results, it nevertheless provides evidence on the influence of working procedures on the advisory work and suggests that different experts can establish different working procedures for getting results based on the same scientific knowledge.

Variance in procedural settings and working procedures is not random, but the result of socio-political influences embodied in the scientific advisory processes. As a more general pattern, as long as the influence of the procedural settings and working procedures on the results is not recognized, the results produced by experts groups will not be comparable. They will always depend on the informal exchanges between experts, on their agreement to work together in a very particular way and on the specific institutional context of their work.

The methodological framework developed here allows understanding of the very subtle technical levers of socio-political and value influences on the results of risk assessments. The literature has already shown that choices of particular technical details (e.g. how to deal with non-monotonic dose–response relationships) decisively influence the results of risk assessments (Jasanoff 1990, 1995; Krimsky 2000; Pielke 2007; Douglas 2009; Elliott 2017), but no systematic and reproducible framework had previously been developed for displaying these choices in their entirety for a specific aspect of risk assessment (here, chemical hazard characterization).

According to Pielke (2007, 17), scientists can overcome the dilemma of socio-political influences in advisory activity by opting for a role of "honest brokers of policy alternatives", engaging in decision-making by clarifying and eventually expanding the choices available to decision-makers. For Krimsky (2000), honest scientists should disclose financial interests and other social biases influencing their advisory work.

For hazard characterization in chemical risk assessments, one option to address the entanglement between science, values and policy, and its procedural expression, would be that advisory panels communicate transparently on their choices, but also on the institutional constraints on their work, in accordance with a set of common criteria such as those proposed in our typology. Some of these are already communicated in some reports, but not in a systematic way and with no comparable framework shared by different expert groups. Though the EFSA (2010) report was used as a case study here,

our classification was developed to be generic enough that it can be applied to any other literature reviews for chemical hazard characterization. Similarly, various guidelines have been proposed lately for facilitating transparent reporting in politically sensitive areas of scientific research and advisory activities, e.g. CONSORT – for randomized controlled trials (Schulz *et al*. 2010); GRADE – for the quality of evidence, the strength of recommendations about therapeutic and diagnostic interventions, and clinical management strategies (Brozek *et al.* 2011); STROBE – for observational studies in epidemiology (von Elm *et al.* 2007); NUSAP – for uncertainty in model-based environmental assessment (Van der Sluijs *et al.* 2005), or the RIVM guidance for uncertainty assessment and communication (Van der Sluijs *et al.* 2008).

A systematic communication of working procedures does not cover all the aspects of procedural influences because it cannot include procedural settings that are outside the scope of the experts themselves. However, it has the merit of accountability, as relevant here as for any other domain of democratic political life. Even if transparency may not "solve" the differences between expert groups, common criteria for describing working procedures would reveal some of the reasons for such differences.

As a more general pattern, a lack of recognition of the influence of procedural settings and working procedures creates suspicion, fuels controversies, wastes public money by necessitating multiple expert groups, and delays decision-making (which, in some cases, might be a political objective). A more systematic way of reporting could make the advisory work more effective by displaying the experts' choices and assumptions, and their combination with the institutional design.

Notes

1. In the first part of this report, EFSA evaluated a dietary developmental neurotoxicity study in rats (Stump 2009). We have only focused on the second part of the report, which assesses the 2007–2010 scientific literature.
2. http://www.wecf.eu/english/articles/2010/06/bpa-call.php
3. Elicitation protocol can be provided on demand.
4. This Excel file can be provided on demand.

Acknowledgements

This work has been funded by the French Ministry of Ecology 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). I am grateful to two reviewers who significantly helped me improve the paper, to Céline Vaslin for help with the figures, and to Sharilynn Wardrop and Jean Morris for stylistic and linguistic improvements.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

Ministère de l'Ecologie, du Développement Durable, des Transports et du Logement (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), grant number [11-MRES-PNRPE-4-CVS-30].

ORCID

Laura Maxim http://orcid.org/0000-0001-9641-6649

References

ANSES (Agence Nationale de Sécurité Sanitaire, Alimentation, Environnement, Travail). 2013. "Évaluation Des Risques du Bisphénol A (BPA) Pour La Santé Humaine. Tome 1." [Risk assessment of Bisphenol A (BPA) for Human Health. Volume 1]. 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.

Beronius, A., C. Ruden, H. Hakansson, and A. Hanberg. 2010. "Risk to All or None? A Comparative Analysis of Controversies in the Health Risk Assessment of Bisphenol A." *Reproductive Toxicology* 29 (2): 132–146. doi:10.1016/j.reprotox.2009.11.007

Bonneuil, C., and L. Levidow. 2011. "How Does the WTO Know? The Mobilization and Staging of Scientific Expertise in the GMO Trade Dispute." *Social Studies of Science* 42 (1): 75–100. doi:10.1177/0306312711430151

Brozek, J.L., E.A. Akl, E. Compalati, J. Kreis, L. Terraciano, A. Fiocchi, E. Ueffing, et al. 2011. "Grading Quality of Evidence and Strength of Recommendations in Clinical Guidelines. Part 3 of 3. The GRADE Approach to Developing Recommendations." *Allergy* 66: 588–595. doi:10.1111/j.1398-9995.2010.02530.x

Centre for Reviews and Dissemination. 2008. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care*. York, UK: University of York. Accessed June 4, 2014. www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf.

Chemtrust, HEAL, WWF, and Breast Cancer UK. 2010. "European Food Panel Fails to Protect Citizen's Health from Plastic Component, BPA." WECF press release, September 30, 2010. http://www.wecf.eu/english/articles/2010/10/European-FoodPanel.php.

Cozzens, Suzan E., and Edward J. Woodhouse. 1994. "Science, Government and the Politics of Knowledge." In *Handbook of Science and Technology Studies*, edited by Sheila Jasanoff, Gerald E. Markle, James C. Petersen, and Trevor Pinch, 533–553. Thousand Oaks: Sage Publications.

Demortain, D. 2009. "Standards of Scientific Advice: Risk Analysis and the Formation of the European Food Safety Authority." In *Scientific Advice to Policy Making: International Comparison*, edited by Julius Lentsch, and Peter Weingart, 141–160. Opladen: Verlag Barbara Budrich.

Demortain, D. 2013. "Regulatory Toxicology in Controversy." *Science, Technology and Human Values* 38 (6): 727–748. doi:10.1177/0162243913490201.

Douglas, H.E. 2009. *Science, Policy and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

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 Journal* 8 (9): 1829. doi:10.2903/j.efsa.2010.1829.

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." EFSA. http://www.efsa.europa.eu/fr/consultations/call/140117.pdf.

Elliott, K.C. 2017. *A Tapestry of Values: An Introduction to Value in Science*. New York: Oxford University Press.

Elliott, K.C., and D.B. Resnik. 2014. "Science, Policy and the Transparency of Values." *Environmental Health Perspectives* 112 (7): 647–650. doi:10.1289/ehp.1408107

EU RAR (European Union Risk Assessment Report). 2008. *Updates Risk Assessment of 4,4′-Isopropylidenediphenol (Bisphenol-A). Final Approved Version Awaiting Publication. Section 4.1.2.9 Reproductive Toxicity. April 2008*, 84–151. Luxembourg: Office for Official Publications of the European Communities.

Grigore, B., J. Peters, C. Hyde, and K. Stein. 2016. "A Comparison of Two Methods for Expert Elicitation in Health Technology Assessments." *BMC Medical Research Methodology* 16: 85–96. doi:10.1186/s12874-016-0186-3.

Jasanoff, S. 1990. *The Fifth Branch: Science Advisers as Policymakers*. Cambridge, MA: Harvard University Press.

Jasanoff, S. 1995. *Science at the Bar: Law, Science and Technology in America*. Cambridge, MA: Harvard University Press.

Knol, A.B., P. Slottje, J. Van der Sluijs, and E. Lebret. 2010. "The Use of Expert Elicitation in Environmental Health Impact Assessment: A Seven Step Procedure." *Environmental Health* 9: 19. doi:10.1186/1476-069X-9-19.

Krimsky, S. 2000. *Hormonal Chaos: The Scientific and Social Origins of the Environmental Endocrine Hypothesis*. Baltimore, MD: The John Hopkins University Press.

Lentsch, J., and P. Weingart. 2011. *The Politics of Scientific Advice*. New York: Cambridge University Press.

Longino, H.E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Maxim, L., and J. Van der Sluijs. 2011. "Quality in Environmental Science for Policy: Assessing Uncertainty as Component of Policy Analysis." *Environmental Science and Policy* 14 (4): 482–492. doi:10.1016/j.envsci.2011.01.003

Maxim, L., and J. Van der Sluijs. 2014. "Qualichem In Vivo: A Tool for Assessing the Quality of In Vivo Studies and Its Application for Bisphenol A." *PloS One*. doi:10.1371/journal.pone.0087738.

Michaels, D., E. Bongham, L. Boden, R. Clapp, L. R. Goldman, P. Hoppin, S. Krimsky, et al. 2002. "Advice Without Dissent." *Science (New York, N.Y.)* 298 (5594): 703.

Pielke, R.A. Jr. 2007. *The Honest Broker: Making Sense of Science in Policy and Politics*. New York: Cambridge University Press.

RES (Réseau Environnement Santé). 2010. "Avis de l'EFSA Sur Le Bisphenol A: Une Décision Ubuesque." Press release, September 30, 2010, Accessed June 4, 2014. http://reseau-environnement-sante.fr/2010/10/01/espace-presse/presse/communique-de-presse-30-septembre-2010-avis-de-lefsa-sur-le-bisphenol-a-une-decision-ubuesque/.

Robert, C. 2010. "Introduction. Les Groupes D'experts Dans Le Gouvernement De l'Union Européenne. Bilans et Perspectives De Recherche." [Introduction. Expert Groups in the Government of the European Union. Summary and Research Perspectives]. In *Les Groupes D'experts Dans Le Gouvernement De l'Union Européenne* [Expert Groups in the Government of the European Union], edited by Cécile Robert, 7–38. Paris: L'Harmattan.

Sarewitz, D. 2004. "How Science Makes Environmental Controversies Worse." *Environmental Science and Policy* 7: 385–403. doi:10.1016/j.envsci.2004.06.001

Schulz, K.F., D.G. Altman, D. Moher, and CONSORT group. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials." *PLOS Med* 7 (3): e1000251. doi:10.1136/bmj.c332.

Silbergeld, Ellen K. 1991. "Risk Assessment and Risk Management: An Uneasy Divorce." In *Acceptable Evidence: Science and Values in Risk Management*, edited by Deborah G. Mayo, and Rachelle D. Hollander, 99–114. New York: Oxford University Press.

Stump, D.G. 2009. *A Dietary Developmental Neurotoxicity Study of Bisphenol A in Rats*, 4796. (Study no. WIL-186056). Ashland, OH: WIL Laboratories.

Van der Sluijs, J.P., M. Craye, S. Funtowicz, P. Kloprogge, J. Ravetz, and J. Risbey. 2005. "Combining Quantitative and Qualitative Measures of Uncertainty in Model Based Environmental Assessment: The NUSAP System." *Risk Analysis* 25 (2): 481–492. doi:10.1111/j.1539-6924.2005.00604.x

Van der Sluijs, J.P., A.C. Petersen, P.H.M. Janssen, J.S. Risbey, and J.R. Ravetz. 2008. "Exploring the Quality of Evidence for Complex and Contested Policy Decisions." *Environmental Research Letters* 3: 024008. doi:10.1088/1748-9326/3/2/024008

Van der Sluijs, J.P., R. Van Est, and M. Riphagen. 2010. "Beyond Consensus: Reflections from a Democratic Perspective on the Interaction Between Climate Politics and Science." *Current Opinion in Environmental Sustainability* 2 (5–6): 409–415. doi:10.1016/j.cosust.2010.10.003.

von Elm, E., D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, and J.P. Vandenbroucke, for the STROBE Initiative. 2007. "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration." *Annals of Internal Medicine* 147 (8): 573–577. doi:10.7326/0003-4819-147-8-200710160-00010

Whaley, P. 2013. "Systematic Review and the Future of Evidence in Chemicals Policy." Policy from Science Project. http://policyfromscience.com/wp-content/uploads/2013/11/PFS-Report-Electronic-Release-Version.pdf.

Wickson, F., and B. Wynne. 2012. "The Anglerfish Deception." *EMBO Reports* 13 (2): 100–105. doi:10.1038/embor.2011.254.