Public Understand. Sci. **14** (2005) 331–352

Difficulties in evaluating public engagement initiatives: reflections on an evaluation of the UK *GM Nation?* public debate about transgenic crops

Gene Rowe, Tom Horlick-Jones, John Walls and Nick Pidgeon

In the realm of risk management, and policy-making more generally, "public engagement" is often advocated as an antidote to pathologies associated with traditional methods of policy-making, and associated deficit-model-driven communication strategies. The actual benefits of public engagement are, however, difficult to establish without thorough evaluation of specific engagement processes. Unfortunately, rigorous evaluation is difficult, and, perhaps for this reason, it has rarely been undertaken. In this paper we highlight a number of these difficulties in the light of our experiences in evaluating a major engagement initiative, namely the *GM Nation?* public debate on the possible commercialization of transgenic crops, which took place in Britain in 2003. The difficulties we identify seem likely to be relevant to many, if not most, engagement evaluations. They are concerned with both theoretical/normative (*how one should evaluate*) and practical (*how one does evaluate*) issues. We suggest a number of possible solutions to these evaluation difficulties.

1. Introduction: public engagement and the issue of evaluation

One evident trend in contemporary democratic societies is the growth in enthusiasm within policy circles for public "engagement" or "participation" as a means of approaching certain difficult issues like the management of risks. In "public engagement" the public is involved *in some limited manner* in the practices of policy-making (e.g. governmental or regulatory) bodies; though there seems confusion as to what extent of involvement is necessary for a particular event to qualify as a case of engagement per se (Rowe and Frewer, 2005). This model for policy-making contrasts with the predominant model in representative democracies in which the involvement of the public is limited to voting at election time, or via membership and influence of pressure groups. According to these arrangements, subsequent decisions are left to elected representatives in government, supported by a "neutral"

 ISSN 0963-6625 DOI: 10.1177/0963662505056611

administrative executive. Such decisions may draw upon advice from unelected experts (either individual advisers or expert committees) (Jasanoff, 1990).

"Engagement" is achieved through various means or mechanisms. In some cases, it is enacted through changing institutional forms—such as co-opting public members or stakeholders on to existing advisory committees. In many cases, however, involvement is achieved through one-off events rather than continuous processes. The use of referenda and consultation documents (on which interested parties can comment) are fairly traditional means of involving the public. More contemporary means include the use of activities such as citizens' juries and consensus conferences, in which members of the lay public, selected to act as representatives, are provided with balanced information, and required to debate an issue and propose recommendations (POST, 2001; Rowe and Frewer, 2000).

There are a variety of practical and ethical reasons for policy-making bodies to involve lay people in policy-making. Political theorists have invoked democratic theory, procedural justice, fairness and human rights, as providing the moral basis for involvement (e.g. Fiorino, 1990; National Research Council, 1996). It is also important to recognize that in many policy contexts there exists a high degree of scientific uncertainty combined with a plurality of value-based perspectives. In such cases decisions may be based to a significant extent upon the *values* of the involved experts, which, in themselves, have no greater inherent validity than those held by lay publics. There is also a need to contextualize scientific knowledge to take account of the specificity of the issue in question. In this regard, local lay knowledges may provide important insights in addressing a range of practical issues (e.g. Funtowicz and Ravetz, 1992; Horlick-Jones, 1998, 2004; Wynne, 1991).

According to this perspective, engagement provides an "antidote" to pathologies associated with a "deficit-model" approach to decision-making; namely one that regards public disagreement with official positions as arising simply from an ignorance of technical facts (House of Lords Select Committee on Science and Technology, 2000; Irwin and Wynne, 1996). In any case, making decisions without public support is liable to lead to a number of practical difficulties, such as confrontation, disruption, boycott, and public distrust. Indeed, a decline in trust in policymakers has been widely noted, and is regarded as having compromised the perceived legitimacy of governance in some areas of policy development (e.g. Frewer, 1999). A transition seems to have occurred from a position where information was seen as the key to resolving a knowledge deficit, and so resolving lay opposition, to one in which regaining trust in governments and regulators is seen as vital to solving a perceived legitimation (or trust) deficit (Walls et al., 2004).

The actual benefits of engagement are, however, difficult to establish without carefully evaluating specific real-world examples of this practice. Indeed, the evaluation of engagement exercises is important for all parties involved as it addresses a range of functions: financial (to ensure the proper use of public or institutional money), practical (to learn from past mistakes to allow exercises to be run better in future), ethical/moral (to establish fair representation and ensure that those involved are not deceived as to the impact of their contribution) and research-related (to increase our understanding of human and organizational behavior). As such, few would deny that evaluation *should* be done when possible.

However, there exist only a relatively small number of rigorously conducted evaluations within the academic literature. A recent review of these by Rowe and Frewer revealed a paucity of empirical work, and a tendency for papers to focus on the *results* of evaluations and their implications, with rather less concern for evaluation methodology and its difficulties (Rowe and Frewer, 2004). Arguably one reason for this is, simply, that evaluation is difficult and a range of theoretical and practical difficulties form a significant barrier to would-be evaluators.

A concrete discussion of the barriers to rigorous evaluation would therefore seem to be much needed. Such a discussion may benefit both an academic audience, in terms of identifying difficulties in the research process and ways to improve the collection, analysis, and interpretation of data, and a practitioner audience, in terms of identifying practical barriers to conducting and evaluating engagement processes. This paper aims to speak to both of these audiences. In this spirit we present reflections on an evaluation that we conducted of a major national engagement event in the UK—the *GM Nation?* public debate. We begin with a discussion of the context for the debate, then go on to a description of the debate process, how we became involved, and how we set about the evaluation. In the main part of the paper we identify both the theoretical and empirical/practical difficulties we faced, and ways in which we attempted to counter these. Our detailed evaluation of the debate has been reported elsewhere (e.g. Horlick-Jones et al., 2004, submitted; Pidgeon et al., 2005; Rowe et al., submitted).

2. The UK *GM Nation?* public debate: context and genesis

In the UK, as in much of Europe (and indeed, elsewhere in the world), there has been significant controversy surrounding the growing of genetically modified (GM) crops, and their use in food products. In the UK, this has been marked by a number of significant events, including the voluntary removal by retailers of at least one food product (GM tomato paste) from supermarket shelves, controversy over shipments of unlabeled GM soya to Europe by the multinational company Monsanto, and, in 1998–9, a widely aired disagreement amongst scientists about the safety of GM potatoes (the so-called Pusztai affair). Throughout this period Eurobarometer survey findings suggested that lay opposition to GM food was reaching a peak in many European countries including the UK (Gaskell and Bauer, 2001). The year 1998 also saw the British government initiate a scientific programme of nationwide farm-scale evaluations (FSEs) of selected GM herbicide-tolerant crops to evaluate the impacts of their cultivation regimes upon farmland biodiversity. These sites were subsequently targeted by anti-GM campaigners for attacks that attempted to destroy the crops, leading to a number of arrests. For more information on the controversies surrounding the growing use of GM crops, see, for example: Halford (2004); Myhr and Traavik (2003); Nielsen et al. (2003); Thorpe and Robinson (2004).

These events occurred during a period when the policy communities in the UK, and other parts of Europe, were just coming to terms with the events surrounding the bovine spongiform encephalopathy (BSE) or "mad cow" crisis of the mid-1990s. BSE marked a clear turning point in the way UK science policy, and with this risk assessment practice, was viewed (see "The Phillips Report": Lord Phillips et al., 2000). Indeed, a number of ensuing reports recommended that scientists and policymakers engage in dialogue with interested parties regarding risk, and regarding science and technology issues more widely (Cabinet Office, 2002; POST, 2001). The House of Lords Science and Technology Committee in particular, in its report on *Science and Society*, diagnosed an apparent "crisis of trust" in science policy-making in the UK, recommending that in order to regain public trust greater openness and transparency in the policy process should occur (House of Lords Select Committee on Science and Technology, 2000; see also Walls et al., 2004).

In this context of controversy, the *GM Nation?* public debate arose as a direct result of a recommendation to the British government by a new multi-stakeholder consultative body, the Agriculture and Environment Biotechnology Commission (AEBC). The AEBC had been established to provide the government with strategic advice concerning wider social and

political aspects of the regulation of agricultural transgenic species. In its report of 2001 entitled *Crops on Trial* (AEBC, 2001), the Commission considered the controversy generated by the government's farm-scale trials, and concluded that public policy on GM crops should "expose, respect and embrace the differences of view which exist, rather than bury them" (AEBC, 2001: 12). It went on to call for:

> an open and inclusive process of decision-making around whether the GM crops being grown in the FSEs should be commercialised, within a framework which extends to broader questions. (AEBC, 2001: 19)

The report also called for an improved understanding of the basis of public views on these matters, and for the future role of GM crops within UK agriculture to be considered in "a wider public debate involving a series of regional discussion meetings" (AEBC, 2001: 25).

In response the government asked the AEBC to advise on how best to implement such a debate. Subsequently, in her letter to the AEBC of 25 July 2002, the Secretary of State responsible for the debate, Margaret Beckett, confirmed that a public dialogue on GM would take place. The government, she stated, was committed to a "genuine, balanced discussion, and also to listening to what people say" (UK Government, 2002). At this stage she also set out the three-component form that the overall programme of dialogue would take. This included a "strand" that would review the science of GM and another one that would examine the economic implications of commercialization, both taking place in addition to the public debate. The intention was "to create a dialogue between all strands of opinion on GM issues." With the letter, the government published a detailed note responding to the AEBC advice that had been submitted on 26 April. The terms of reference for the overall program of dialogue were specified as follows:

- To identify, using methods which focus on grass roots opinion, the questions which the public has about GM issues, avoiding as far as possible the polarization that has characterized so much of the discussion to date;
- To develop, from this framing of the issues and through a wholly open process, the provision of comprehensive evidence-based information to the public on scientific, economic and other aspects of GM;
- To provide people with the opportunity to debate the issues openly and to reach their own informed judgments on this subject;
- To provide information to government on how questions raised by the public have shaped the course of the debate, including on the scientific, economic and other aspects of GM.

3. The design of the *GM Nation?* public debate process

GM Nation? took place during a six-week period from 3 June to 18 July 2003. It was overseen by an independent Steering Board drawn from members of the AEBC together with a number of co-opted individuals. Membership of the Steering Board included stakeholders from across the spectrum of opinion on GM agriculture. Much of the day-to-day implementation of the debate was carried out by an "arms length" agency of government, the Central Office of Information (COI), which acted as the main contractor to the Steering Board.

In November 2002, and prior to the main debate process, nine preliminary discussion groups (known as "Foundation Discussion Workshops") were convened in various locations

across the UK, eight of which comprised members of the general public who were not already actively engaged with the issues, while the ninth comprised those who were actively involved and interested in GM issues. Each workshop involved 16–20 participants. They were facilitated by two moderators, and lasted approximately three hours. The primary aim of these was to investigate how a cross-section of ordinary citizens would make sense of the GM issues. It was intended that these insights into lay framing of the issues would make it possible to design the debate process in such a way that lay perspectives could shape the terms of the engagement. The findings could also be used in the production of stimulus materials for use in subsequent stages of the debate.

The main "debate" process, conducted in the summer of 2003, comprised three principal engagement mechanisms.

- A series of open public meetings, which anybody could attend, organized into three levels or "Tiers". Tier 1 meetings (three in England and one each in Wales, Scotland and Northern Ireland) were conceived of as "national" high profile events, and were professionally facilitated. These attracted approximately 1000 participants in total. Tier 2 meetings, of which there were about 40, were typically hosted by a local authority or other major organization, often with the assistance of the main debate contractors. Tier 3 meetings, typically organized by local voluntary organizations, were highly variable in terms of their character and formality. The contractors *estimated* that over 600 of these events took place. In all some 20,000 individuals across the UK were estimated to have taken part in the various open meetings (PDSB, 2003).
- A dedicated interactive debate website, which contained a range of debate materials and interactive resources. This website recorded over 24,000 unique visitors during the course of the debate.
- A series of 10 closed discussions with ordinary members of the public, known as "narrow-but-deep" groups, which were conceived as a "control" on the character of the discussions produced by the self-selected participants in the open meetings. Here "narrow" refers to the limited scope of representation (only 77 members of the public took part, albeit being recruited to reflect a broad demographic cross-section of the UK population), and "deep" refers to the anticipated extended level of engagement and deliberation in these groups, in comparison to that typically possible at the open meetings. These groups met twice, with a gap of two weeks between which participants were invited to explore the GM issue individually, using official stimulus materials and any other information that they could access, and to keep diaries of their discoveries (including newspaper clippings, website downloads etc.), thoughts, relevant conversations and so on (PDSB, 2003: para 194).

The Steering Board's final conclusions and report were based upon a reading of several data streams: primarily qualitative, including rapporteurs' reports from meetings, analysis of the narrow-but-deep discussions, and open-ended feedback responses (letters and e-mails received); but also quantitative, from 13 standard attitude questions on a feedback questionnaire that was distributed in paper form but could also be completed on the debate website. The feedback questionnaire in particular proved pivotal, with a total of 36,553 responses obtained in almost equal proportions from paper copies distributed to meeting organizers and its equivalent website format. The 77 narrow-but-deep participants also completed this questionnaire, albeit twice: once at the commencement of their involvement and again at the beginning of their second meeting.

4. The evaluation process, and brief results

At the time the GM debate was proposed, several members of our research team had been monitoring the work of the AEBC for a period of about 18 months. In August 2002 we wrote to the AEBC, setting out our intention to follow the debate, and expressing an interest in having a close involvement in the process of evaluation. The importance of having a systematic evaluation of the process in order to learn lessons for future policy development had previously been noted by the AEBC in its initial guidance to government on the conduct of the debate. Though an appointment by competitive tendering would have been more satisfactory from the point of view of evaluating a publicly accountable process, the debate budget was insufficient to allow the Steering Board to fund its own systematic evaluation. We were subsequently invited (in September 2002) to present a detailed evaluation proposal to the debate's Steering Board, on the basis of which we were appointed as the debate's official evaluators. This quasi-contractual arrangement was achieved by means of a formal exchange of letters which were subsequently posted on the debate's website (www.gmnation.org.uk).

Our involvement was only formalized relatively late into the debate process, after much discussion had already taken place in the Steering Board on the purpose, nature, and implementation of the debate. Our first issue was to decide upon the theoretical basis for the evaluation. That is, against what benchmarks would we evaluate the event? We decided upon the use of several different sets of *evaluation criteria*, in order to reflect different perspectives on engagement quality and what this should entail: sponsor perspectives, participant perspectives, and a normative perspective informed by the academic literature (which proposed how engagement exercises *in general* should be conducted).

The sponsor perspective on what would make the exercise a "success" was not initially very clear, beyond the general aim of somehow involving the public in debate on the GM issue. After some prompting, the Steering Board eventually produced a list of objectives, which included:

- "to allow the public to frame the issues" (objective 1);
- "to focus on getting people at the grass roots level whose voice has not yet been heard to participate in the programme" (2);
- "to create new and effective opportunities for the deliberative debate about the issues" (3);
- "to create widespread awareness among the UK population of the programme of debate . . . and give widespread opportunities to register views" (5);
- "to calibrate [*sic*] the views of organisations who have already made their views known by contrasting their views with other participants in the debate" (8).

The document that details these objectives also included a subsequent section entitled "How will we know that the programme of debate has been successful?" (PDSB, 2003: Appendix B). This section identified a further *four* "indicators" of success. The *indicators* were:

- The extent of public awareness of the programme;
- The views of participants in the debate "about what they felt should be the criteria for success . . .";
- The extent to which "informed commentators" felt the exercise has been credible, innovative, balanced, and had moved the debate "beyond the polarisation that has so far characterised much of the discussion about GM crops";
- The extent to which the report from the debate "could reasonably be said to have had an impact on Government".

Indeed, the first Foundation Discussion Workshop had already taken place (14 November 2002) before the evaluators received a copy of the Steering Board's list of objectives and indicators of success. (At a meeting of the PDSB on 7 November there was a discussion of draft objectives with an action to the secretariat to circulate the final draft in the light of comments by the members. Though we were aware that draft objectives had been discussed, we did not receive the final draft until after 14 November.) This lack of clarity on the precise aims of the event was a major criticism of the *event* (not evaluation), noted in our evaluation and raised in the subsequent commentary of a House of Commons Select Committee (House of Commons Environment, Food and Rural Affairs Committee, 2003).

Of course, the participants might have different perspectives on the rationale for the event and how it ought to be judged than the sponsors. Hence, we attempted to tap into participants' own criteria for evaluating the event, focusing on those attending the main visible component of the debate, namely the Tier 1 meetings. In our questionnaires distributed to participants (described shortly) we included a number of open questions asking participants what they thought were the good and bad points from the event, and from analysis of responses we inferred their criteria.

Our third set of normative criteria was informed by the literature. Despite voluminous writings on public engagement, and copious assertions as to the necessary requirements for an exercise to be successful, there were relatively few frameworks to which we could turn. We chose that formulated by Rowe and Frewer (2000), which identified nine distinct criteria that a participation exercise should fulfil in order to be deemed successful. These criteria—distilled from a review of the relevant literature—are as follows:

- Representativeness: The public participants should comprise a broadly representative sample of the population of the affected public.
- Independence: The participation process should be conducted in an independent, unbiased way.
- Early Involvement: The public should be involved as early as possible in the process as soon as value judgments become salient.
- Influence: The output of the procedure should have a genuine impact on policy.
- Transparency: The process should be transparent so that the public can see what is going on and how decisions are being made.
- Resource Accessibility: Public participants should have access to the appropriate resources to enable them to successfully fulfil their brief.
- Task Definition: The nature and scope of the participation task should be clearly defined.
- Structured Decision Making: The participation exercise should use/provide appropriate mechanisms for structuring and displaying the decision-making process.
- Cost Effectiveness: The procedure should in some sense be cost effective.

Our choice of these criteria was influenced not only by the relative paucity of other normative frameworks, but also by the presence of one of the original authors in the evaluation team, and by the previous use of the framework in a past evaluation (Rowe et al., 2004), ensuring that there were extant instruments (questionnaires; a check-list) that could be used and adapted for the present evaluation. We did not adopt this framework wholesale, however, but adapted it to present circumstances. For example, we redefined "Structured Decision Making" as "Structured Dialogue," since participants were not intended to make any decision per se, and we omitted consideration of the Cost Effectiveness criterion, since this seemed difficult to establish and conceptually unclear. We comment upon

issues surrounding the choice and use of evaluation criteria in the later discussion of evaluation difficulties.

In order to conduct the evaluation we adopted a multi-method approach that used both qualitative and quantitative methods. This drew upon well-developed quantitative techniques and more recently developed qualitative approaches within the evaluation literature (e.g. Rossi et al., 1999; Shaw, 1999). Thus we used participant questionnaires, structured observation, in-depth interviews, media and document analysis and a major survey of public opinion.

Specifically, one member of the evaluation team attended each of the Foundation Discussion Workshops (save for that using "actively involved participants"), each of the Tier 1 meetings, several Tier 2 events, and two of the "narrow-but-deep" groups. Observers used an observation schedule that identified various activities and outputs that would enable us to comment upon how the events succeeded against our normative criteria, though observers also had the freedom to note broader impressions on how the events were progressing (i.e. they were not entirely constrained by the schedule). Participant questionnaires were also distributed to those attending the Foundation Discussion Workshops (all of these), the Tier 1 events, several Tier 2 events, and the "narrow-but-deep" groups. There were in fact two questionnaires—a short questionnaire, given to participants to complete immediately after certain events (e.g. the Foundation Discussion Workshops), and a long version that participants in the various events completed at home, and which they returned to us via postage paid envelopes (the reasons for these two forms will be explained later). The longer questionnaires included a number of open-ended questions at the start, asking, for example, what were the good and bad points about the events. Responses to these were used to infer the participants' own evaluation criteria. Various closed questions then addressed participants' perceptions of the events in terms of our normative criteria.

Evaluation team members also attended the various open meetings held by the debate Steering Board, conducted interviews with several Board members and other important stakeholders (at various times throughout and after the debate), and collected all documentation pertaining to meeting minutes, contractor reports, and so on (to support our subsequent analysis). We also commissioned a major public opinion survey to assess the views of a representative sample of the public on GM food (using the attitude items from the organizers' own feedback questionnaires), enabling us to consider how representative were the attitudes of debate participants. Furthermore, we monitored media reports—over the course of the debate and at the time of the release of the final report on the exercise—from all of the English national daily and Sunday newspapers; the main national news bulletins and current affairs programmes on the five terrestrial television channels and on BBC Radio; selected regional and local newspapers; and Internet news services and bulletin boards supported by selected newspaper and broadcasting organizations.

In summary, we found that, in spite of a number of positives (e.g. the Foundation Discussion Workshops and Tier 1 events were enjoyable for participants, and generally perceived to be fairly and competently run by the organizers), there were a number of significant problems with the debate. For example, the participants were not particularly *representative* of the wider public (one of our normative criteria, and also an implicit criterion of the sponsors and the participants themselves), and they tended to be more negative (or at least, less positive) about GM food than the UK public. The ultimate *influence* (another of our normative criteria) of the debate was also minor (as suspected might be the case by participants at the time), as became apparent later, when subsequent policy was set in accordance to other concerns and with no clear input from the debate results. Resources (especially in terms of time) were also insufficient. Precise details of all

the methods we used, and the results of our analysis, are reported elsewhere (e.g. Horlick-Jones et al., 2004, submitted; Pidgeon et al., 2005; Rowe et al., submitted).

5. Difficulties in conducting the evaluation

Here we present the main difficulties we encountered in the form of the various dilemmas we faced, how we did/did not counter them, and their wider theoretical and practical implications for other evaluations. We focus on the following main issues:

- whether evaluation should be an add-on extra or integral to the engagement process;
- selecting evaluation criteria;
- measurement issues;
- data quality;
- when is the evaluation exercise complete?
- relations with the debate organizers and other stakeholders;
- the resource demands of the evaluation exercise.

5.1. Evaluation: add-on or integral process?

As indicated previously, the issue of evaluation was not initially integral to the *GM Nation?* debate. Within the context of a cash-limited process, evaluation did not emerge as a priority. Indeed, the Chair of the Public Debate Steering Board stated at a public meeting that "there was no indication of any enthusiasm on the part of government to undertake such an evaluation itself" (Malcolm Grant, AEBC Meeting, 11–12 December 2003, Eden Project, Cornwall). A number of subsequent difficulties in the evaluation process undoubtedly stemmed from its "add-on" nature.

From the sponsors' perspective, lack of evaluation planning limited its options: it was not able to seek tenders to do the job, and was fortunate that we were available and eager to do the task without payment. The late presence of evaluators also raised concerns with the recruited contractors and sub-contractors, who clearly felt a degree of disquiet at suddenly finding out that their activities would be under scrutiny, leading to some difficult negotiations amongst the various parties that would not have been necessary had the evaluation component been designed-in at the beginning. For the contractors and sub-contractors, the problem must have been magnified by the presence of only fairly general working briefs and a lack of explicit guidelines as to how their activities were going to be judged.

From our perspective as evaluators, this relatively late involvement led to a degree of tension with the (justifiably) concerned contractors, and meant that the basis on which the evaluation would be conducted had to be rapidly revised *while the process was already underway*. Thus, although we had settled upon the normative framework we would use as one basis for the evaluation, the event was already underway (as noted, the first Foundation Discussion Workshop had already taken place) before we received a copy of the Steering Board's list of objectives and indicators of success, i.e., before we had the sponsors' evaluation criteria to work with.

An important lesson from this is that evaluation should be a fundamental part of the participation process. Preferably, proper contractual arrangements should be established, setting out, for example, the bounds of the evaluation and extent of evaluator access to relevant processes and information. This requirement stems not only from pragmatic and diplomatic considerations (in order to pre-empt potential difficulties with those running the event), but also from a need to ensure full accounting of the value of the entire process (i.e.

the validity of the evaluation). For example, if evaluators wished to establish the "fairness" of an exercise, they would arguably need access to the preliminary debates by the sponsors in which they set the terms of reference of the exercise, deciding such issues as who should be involved, how they should be recruited, what should be expected of them, and how their involvement should be operationalized. If accurate documentation of initial decisions were available (e.g. minutes of Steering Board meetings), this problem might be *partially* overcome—but if not, substantial important activities with bearing upon the ultimate effectiveness of the exercise might have already passed out of reach of the evaluators, to the detriment of the accuracy and comprehensiveness of the evaluation.

One other important issue worth considering is that some form of "evaluation" (or at least, informed commentary) might take place on a participation exercise *after* the event and irrespective of the desire of the sponsors. Such evaluations might be based on only partial information leading to a less-complete evaluation than otherwise, which would be to no-one's benefit. This again suggests a need for evaluation planning as soon as possible in the broader engagement process.

5.2. Selecting evaluation criteria

Perhaps the main difficulty in conducting an evaluation concerns the process of defining "effectiveness"—that is, selecting, or agreeing upon, suitable *evaluation criteria*. First, however, it is important to note that there is a school of thought that regards selecting evaluation criteria *prior* to an exercise taking place as problematic. The inductive research tradition seeks to first collect data and *then* to induce hypotheses/theories (and indeed, definitions). Here, stating what is meant by "effectiveness" *a priori* is seen as constraining the data one will collect, and so imposing a framework upon the data that might not be appropriate. This approach, which typically uses qualitative rather than quantitative research techniques, is particularly apt in new environments where little is previously known, where quantitative data are difficult to obtain and where hypotheses are difficult to generate (e.g. Joss, 1995). This position generally leads to "evaluations" that take the form of case studies, in which results are based upon the evaluators' subjective interpretations. However, such exercises create limitations as to the extent to which results may be replicated or generalized, and pose serious questions about the reliability and validity of their findings (e.g. Clarke, 1999).

We suggest that such inductive exercises should be termed *assessments* rather than *evaluations*. Although in some respects conducting *assessments* (as defined here) would seem appropriate in the complex participation environment, there are various difficulties involved in taking this approach, exaggerated by the often highly charged political nature of public participation. Perhaps the greatest of these problems is the different values and perspectives of those involved (from the sponsors and organizers to the various participants themselves) each of whom may have different rationales for involvement. Though differing perspectives are problematic for defining effectiveness *a priori*, this makes ascertaining effectiveness after the event a hugely fraught exercise, in which any party disagreeing with the assessment may (perhaps justifiably) question the conclusions. An analogy may be a game of football in which invisible goal posts are only revealed at the final whistle. If effectiveness is defined beforehand (the goal posts are evident) then there can be less cause for complaint.

In practice the exercise we conducted was rather more pragmatic in nature, with an emphasis on methodological plurality rather than being constrained by one or other of these evaluation paradigms. We carried out an *evaluation* in the sense that we were guided by

criteria identified in advance (or at the start) of the engagement process, however, our multi-method approach provided us with a capacity to *learn* from the process, and so an ability to arrive at emergent findings (discussed in Horlick-Jones et al., 2004; see also Bloor, 1978; Glaser and Strauss, 1967). Similar approaches have been advocated by Clarke, who concludes that pragmatic evaluators need to have regard to paradigmatic differences in research methods in order to make informed choices within a mixed-method evaluation design; and Patton, who uses the term "methodological appropriateness" to capture such an approach, which contrasts with one of paradigmatic orthodoxy (Clarke, 1999; Patton, 1990).

A clear definition of what it means for a participation exercise to be "effective" provides a benchmark against which performance may be compared. However, this leads to the question: is it possible to state a single definition of effectiveness for all participation exercises, or is each exercise unique, with specific aims and hence a need for a specific definition of effectiveness? In a review of evaluations in the academic literature, Rowe and Frewer (2004) found that answers to this question have varied, with many evaluations—at least implicitly—adopting a definition of effectiveness that is "universal" and apparently intended to be relevant for participation exercises generally (or at least a specified subgroup of these). It has been argued that, though any particular exercise may have very precise aims, these should be able to be phrased in terms of more general effectiveness criteria—for example, a specific aim to "effect policy in a certain way" may be cast as a general criterion "to have an impact on policy" (Rowe and Frewer, 2004).

Published definitions of effectiveness have varied on a number of dimensions. One concerns whether effectiveness is a *subjective* aspect (and if so, subjective according to *who*) or an *objective* aspect (that either incorporates no subjective opinions from the involved parties, or perhaps combines all of these in some manner). A second concerns whether effectiveness should relate to aspects of the quality of the exercise *process* (how the exercise was conducted) or aspects of the *outcome* (such as whether the exercise has a positive consequence), or indeed, both. Though "universal" definitions of effectiveness in the literature have varied, certain themes have tended to re-occur, such as the need for participants to be fairly representative of the affected stakeholders or public, and the need for exercises to have genuine impact upon the policy or behavior of the sponsors of the exercise (Rowe and Frewer, 2004).

In our evaluation we decided to utilize one of these "universal" frameworks, namely the one formulated by Rowe and Frewer (2000), which identified nine distinct criteria that a participation exercise should fulfil in order to be deemed successful. We did this in part because this framework had been used before and hence existing instruments were available (Rowe et al., 2004). However, given criticisms of the universal approach, and uncertainty as to the complete *relevance* and *comprehensiveness* of our chosen evaluation criteria, we attempted to elicit and use two other sets of criteria specific to the current event.

The first additional set of evaluation criteria was supplied by the organizers. At the time we submitted our proposal to carry out the independent evaluation, the Steering Board had still to clearly elucidate its objectives for the debate. Indeed, it was not clear that the Board had any definite intention to produce a written formalization of the aims of the process. That a formal statement of the aims for the debate was eventually produced was undoubtedly due in part to our evaluation proposal, in which we observed that, whilst it was likely that the debate would have multiple objectives, this created some ambiguity about the nature of the process and how it might be evaluated. We thus suggested that the Steering Board should give serious consideration to identifying criteria that could reasonably be thought to

be widely acceptable (Horlick-Jones et al., 2002). Eventually the Board agreed that the overall aim of the debate was:

> To promote an innovative and effective programme of public debate on issues around GM in agriculture and the environment, in the context of the possible commercial growing of GM crops in the UK. The public will frame the issues for debate. Through the debate, provide meaningful information to Government about the nature and spectrum of the public's views, particularly at grass roots level. (PDSB, 2003: 11)

The Steering Board subsequently agreed a set of nine specific objectives and four indicators, described earlier. Unfortunately, most of these objectives appeared to us to be conceptually unclear, not measurable in any rigorous or sensible manner (such as the objective to run a debate that was "innovative"), or were conceptually uninteresting (such as the objective "to provide a report to Government"). Furthermore, the indicators *did not* closely address the stipulated objectives, and indeed, sometimes implied other objectives. For example, ascertaining "the views of participants in the debate about what they felt should be the criteria for success . . ." did not speak to any particular objective. In our view "indicators" should have been closely associated with each objective, identifying how to ascertain or measure success—but this was not the case. As a consequence of these various difficulties, we were only able to comment upon the achievement of a few of these objectives, and to do so through a rather *ad hoc* analysis of our various pieces of data.

Our third set of criteria was elicited from debate participants, via a number of open questions in our participant questionnaires. In particular, we asked participants to indicate what they thought was good and bad about the debate (or at least, that component with which they were involved), and how it might be improved. By identifying themes in their qualitative answers we were able to comment on the criteria that participants themselves found relevant. For example, many participants assessed the quality of the debate according to whether they had or had not learnt anything, either about the issue (GM food and crops) or about the opinions of other people. This criterion was not specifically covered in our universal criteria (Rowe and Frewer, 2000), nor specifically identified in the organizers' criteria. This approach therefore also served to *validate* the other evaluation criteria that we used (which will be discussed later).

5.3. Measurement issues

Following stipulation of suitable evaluation criteria against which a participation exercise is to be judged, there is a need to develop *instruments* or *processes* to accurately measure performance on these criteria. Such instruments are themselves open to evaluation, although the appropriate criteria for this are perhaps less contentious than those for assessing participation effectiveness. These criteria include *reliability* and *validity*, to which one could also add the criterion of *usability*. By reliability we mean that use of the instrument is liable to give similar results on a number of different occasions or when used by a number of different evaluators. For example, a scale for measuring weight would be of little use if it gave different readings each time a particular person stood on it. Likewise, if a measure of "fairness" gave different results for a particular exercise when applied on different occasions or by different evaluators (e.g. rated as "fair" on one occasion, but "unfair" on another), then its lack of reliability would undermine its usefulness. Validity concerns whether the instrument measures what is intended to be measured. A scale measuring weight should give values related to weight and not, for example, height; an instrument measuring "fairness" should truly measure "fairness" and not some other quality (e.g. "enjoyment" of the event).

Usability is simply a pragmatic criterion that refers to the ability to actually deploy one's instrument when and where intended (Rossi et al., 1999: 252: "in addition to being reliable and . . . valid, a good outcome measure is one that is feasible to employ, given the constraints of time and budget"). As a case in point, we have found that lengthy questionnaires are often resisted by organizers and sponsors unwilling to over-burden participants: a questionnaire that could not be employed for this reason would fail the usability criterion.

Although criteria such as reliability and validity are usually discussed in relation to the development of *questionnaires* measuring psychological properties (such as IQ, job satisfaction, personality type, etc.), they are conceptually relevant elsewhere, such as for participant observation processes and document analysis. For example, an observational method that led two different observers to conclude two different things about the exercise being observed would not lead to confidence in the evaluative conclusions (in this case, the type of *reliability* established here is *inter-rater* reliability). There are various different forms of reliability and validity, which are explained in detail in most standard social science methodology textbooks (e.g. Bryman and Cramer, 1997; Oppenheim, 1992; Silverman, 1998).

Unfortunately, the nature of the public engagement domain frequently makes it difficult to establish instrument reliability and validity. In the first place, contention as to whether there is or is not a universally relevant set of evaluation criteria has undermined any concerted attempt by academics and practitioners to develop a coherent, structured, validated set of off-the-shelf instruments (as has been developed to measure intelligence, for example). Evaluators have therefore tended to develop their evaluative instruments *simultaneous to* the evaluation itself (e.g. see Rowe and Frewer (2004) for a review of instrument and process characteristics in published evaluation studies). In the second place, the complexity of the typical participation environment and lack of evaluator control (the nature and manner of data collection are often at the discretion of the organizers and sponsors, not evaluators), mean that data in a suitable form are rarely available to allow commentary on *instrument quality* (reliability, validity). For example, because of the contentious nature of evaluation, it may be difficult to convince sponsors to allow the distribution of lengthy questionnaires (needed to obtain *adequate quantities* of data), let alone presenting a particular questionnaire twice to enable its reliability to be assessed. Such complexity has been used as an excuse by researchers and evaluators to settle for conducting informal evaluations only: while this difficulty may mean that formal processes for establishing aspects such as reliability and validity are difficult, the concepts cannot be ignored. We contend that, at the very least, researchers should discuss reasons for believing that their measures of effectiveness are reliable and valid (and ideally, evidence for the quality of instruments should form the basis of a section in any study write-up) (Rowe and Frewer, 2004).

In the case of the *GM Nation?* debate, we experienced various problems in deploying the instruments we used (*usability* difficulties). At both the Foundation Discussion Workshops and the various Tier events, we were only able to hand out shortened versions of questionnaires to participants (which contained key questions), although we were also able to give participants longer questionnaires to complete at home and return to us in postage paid envelopes. Because response rates to the questionnaires taken away were lower than for those completed at the time, we gained less data than we might have had we been able to present the longer questionnaire for immediate, after-event completion. This did mean, however, that we were able to attain and compare *two* sets of data, and comment upon reliability of both questionnaires. We must emphasize that comparing responses from the

two questionnaires *does not* constitute a formal reliability check, in the sense that participants were not answering *exactly* the same questions (as an example, in the short questionnaires we were able to include just a single question about each of our normative evaluation criteria, whereas there were multiple questions per criteria in the long questionnaire). Nevertheless, this approach did allow us to confirm that participants had consistent interpretations as measured by the different items and questionnaires. For example, participants generally rated the "independence" of those running the events highly, as measured by single items in the short questionnaire, and several (slightly different) items in the longer questionnaire.

In terms of validity, we attempted to establish this through triangulation using a variety of measurement processes. In other words, we ascertained participant views via questionnaires and at the same time had observers assess the different events of the debate using an observation protocol based upon similar criteria. The nature of the data was not strictly comparable, but because we were able to come to roughly similar conclusions about performance of the events on the various criteria from both approaches this increased our confidence that both instruments (*questionnaires* and *observation processes*) were measuring what we hoped they would measure. Furthermore, we found that the participants, in responding to open questions about what was good and bad about the events in which they took part, often judged the events using similar criteria to our normative ones. For example, participants often criticized the Tier 1 events for mainly involving participants who were against GM foods and crops, meaning that they were unable to have a proper debate with people representing pro-GM positions. This suggests that the *representativeness* of participants was indeed important to them (one of our normative criteria)—and in this sense, that the normative criteria themselves (as opposed to the instruments used to measure them) were valid (relevant). Further, because these open responses tended to concur with a poor assessment of the Tier 1 events on this criterion attained through the other instruments (i.e. the *direction* of assessments largely corresponded to the direction of assessments attained in response to the closed questionnaire items and the observations), this lends weight to assertions that those other instruments were validly measuring the concepts intended. The validity of our normative criteria, and the reliability and validity of the instruments used to measure these, is discussed in more detail elsewhere (Rowe et al., submitted).

In summary, our discussion in this section has emphasized the difficulty in controlling the acquisition of appropriate data to enable the formal assessment of the quality (reliability, validity) of instruments intended to measure one's selected evaluation criteria. However, it also indicates a number of informal ways to make such instrument assessments (i.e. collecting data through different processes at different times), which may help increase evaluators' confidence in their evaluations.

5.4. Data quality

Here we consider a number of inter-connected barriers we encountered in seeking to collect a suitably high quality corpus of data during the evaluation exercise. We focus on the following issues:

- uneven data coverage;
- the logistics of data collection;
- disturbing the process that we were evaluating.

5.4.1. Uneven data collection Various time and logistic constraints are bound to impinge upon the collection of full data. Some of these constraints are more broadly applicable to

social research. For example, as noted, the participant questionnaires we developed needed to be relatively small so as not to over-burden participants. Other specific constraints applied to this particular evaluation, however. We had considerable difficulty in gaining information about the location and timing of local (Tier 2 and Tier 3) meetings, and hence we were only able to observe (and distribute questionnaires in) a relatively small sample of these. This particular engagement process was, however, atypical in its scope and ambition, to the extent that even the organizers were unaware of many of the events that were held (the Tier 3 events), and as *such these could have had no impact* upon the final report produced from the debate for the sponsors.

A further specific limitation to our observational work arose as a result of our need to be as unobtrusive as possible, which meant that only a single observer was acceptable to the contractors at each event—although the use of audio-taping increased the amount and precision of data gathered. However, we were not permitted observational access to the Foundation Discussion Workshop involving those who were actively involved and interested in GM issues, which was regarded by the contractors as particularly sensitive, and so in this case we had to rely upon tape recordings alone. We also were only able to negotiate access to two of the ten narrow-but-deep groups (a last-minute invitation to observe a further group proved impossible, given the existing commitments of the research team).

Naturally, limitations in data collection must impact upon the validity of an evaluation, and it is important that these limitations be reported and considered when reporting results. Our lack of a detailed formal contract to conduct the evaluation unfortunately meant that we were not in a position to make more strenuous requests for information access. We recognize that the Steering Board was in a difficult position, caught between access requests from the evaluators and the reservations expressed by their contractors (and sub-contractors). As we have already noted, early and formal establishment of the role, aims, and scope of evaluation is therefore crucial to minimize such difficulties.

5.4.2. The logistics of data collection Our experience of evaluating a large, multi-component engagement exercise like *GM Nation?* makes very clear the extent of practical, logistical, difficulties that may emerge during data collection. The experimental character of the debate gave rise to a rather untidy nexus of overlapping processes, with detailed implementation being contingent upon decisions taken at a sometimes quite advanced stage of the exercise. As the "public" phase of the debate approached, the pace of developments speeded up, and "backstage" activity often took on a rather frenzied nature. Given these circumstances, it was not possible for us to plan our data collection schedule and associated resource demands with any degree of certainty. In response, our attempts to maintain flexibility necessitated keeping empty diaries, and gaining the agreement of colleagues to provide data collection support at short notice.

When the "public" phase of the debate did arrive, attempting to follow and evaluate the Tier 3 public meetings proved especially challenging. These meetings were organized mostly by self-selecting local voluntary organizations. In a matter of a few weeks, it seems that literally hundreds of such meetings were organized throughout the UK, often in rather informal ways, and with scheduling taking place over short timescales. The debate executive appeared to have difficulty in keeping track of these developments, and at one stage, we found that a GM pressure group website offered the most up-to-date list of local events. In such circumstances, unsurprisingly, we were only able to gather data on a relatively small sample of these meetings. However, this does not appear to seriously undermine our evaluation since, as noted, the organizers and sponsors themselves were unaware of many of

these events, and hence they had *little if any* impact upon the summary report produced for government.

The other major logistical challenge we faced was posed by developments in the wider organizational and political context of the debate process. Whilst we did monitor a range of such developments—within government, the accompanying scientific and economic "strands" of the overall debate process, and shifting views of stakeholder organizations, media coverage of the debate, and a major poll of public opinion—it is perhaps difficult to conceive of *any* sufficiently well-resourced evaluation exercise that would be able *to fully capture* the enormity of the political history of such an engagement exercise.

5.4.3. Disturbing the process being evaluated In common with all social science research, we faced the potential difficulty of inadvertently influencing the debate process we were studying—by collecting data, and indeed, by virtue of our mere presence. Studying participants can rarely be achieved unobtrusively. Asking participants to complete a questionnaire is likely to focus their thinking about the debate and perhaps raise issues that they had not previously considered, some of which might subsequently affect their interpretation of the process. However, by presenting questionnaires at the end of an event, rather than at its beginning, one might hope that impact will be minimized. On the other hand, the presence of observers is difficult to hide (indeed, if observation takes place surreptitiously, it may raise ethical questions). Although some people may well "play to the camera" if they feel they are being observed, we suggest that the passive involvement of a single observer was unlikely to make a significant difference to participant behavior. Here it should be recalled that the meetings were also being recorded and/or observed by sponsors and organizers. Moreover, the kinds of issues being discussed were often contentious enough to attract most participants' full attention.

Nevertheless, in some circumstances, this possible influence can lead to evaluators being excluded. As noted above, this happened in the case of one particular Foundation Discussion Workshop, where our access was denied because the contractors felt that observer presence might unsettle a potentially tricky process still further. Although the inclusion of the evaluation in the original design of the debate might have reduced such access difficulties, such planning seems unlikely to have entirely eliminated them.

Additionally there arises the question as to whether observer influence is a bad thing. It is, of course, important to weigh up the (hopefully) minimal impact of evaluators on the exercise with the potentially significant gains from a comprehensive evaluation. It is also important to recognize that more immediate influences on the process may be beneficial rather than detrimental: indeed, it has been suggested that one important role of evaluation may be to allow "mid-course corrections"—identifying problems and reacting to them before it is too late (Chess, 2000). After all, in medical trials, evidence that a treatment's effects are significant may lead to a study being prematurely terminated in order that those in the control group also benefit from the new treatment. In the case of the *GM Nation?* debate our presence might have been regarded as beneficial. Indeed, we have noted that it appeared to stimulate the sponsors to think more clearly about the debate's objectives and produce a list of aims and indicators of success—although we found that these were not particularly well stated or coherent. There is little evidence that this list ultimately focused or changed the Board's thinking on the engagement events (having only been produced after the first public phase had already got underway!), or the processes then being delivered and designed to implement it.

In practice we sought to minimize such influences, being fairly strict about not providing the Steering Board or its contractors with feedback along the way. In our view,

such interventions would have been quite inappropriate, given our role. Indeed, we take the view that the adoption of such an interventionist role would have transformed our evaluation exercise into a quite distinct activity, and one that could certainly not have been claimed to be an "independent evaluation." Of course, "keeping quiet" about aspects of the debate that we recognized were flawed may well have resulted in various unsatisfactory outcomes for the debate. Nevertheless, to intervene in this way would have been to interfere, and so distort, an important experiment in engagement, and to impose our own "hunches" in advance of completing a rigorous evaluation.

5.5. When is the evaluation exercise complete?

One difficulty associated with conducting an evaluation that is of special relevance to practitioners, but with relevance to academic researchers, is identifying an end-point to the engagement exercise, bearing in mind that institutional and societal responses to a particular exercise may be manifest months or even years after it has officially finished. This is particularly problematic when measuring *outcomes* (or impacts) of exercises, as distinct from the quality of constituent *processes*. Hence, outcome measures may be difficult to ascertain in a timely manner, and in any case, the outcomes themselves may be due to some extent to other factors, such as the occurrence of simultaneous events, or externally mediated pressures influencing policy processes (Chess and Purcell, 1999).

Sponsors of engagement exercises that involve evaluation generally desire rapid appraisal so that this might be included in some report of the activities, and indeed, such evaluation may have a bearing on policy outcomes arising from the exercise. Clearly this requirement places evaluators in a contradictory situation, one characterized by Shaw (1999) as the "rigour or relevance" dilemma. This difficulty is especially pronounced for those carrying out independent evaluations such as ourselves in relation to *GM Nation?* Whilst we wished our evaluation to be rigorous in scholarly terms and as complete as possible, there was also a wish to retain a degree of relevance for the debate sponsors, who had, of course, provided us with unique access to the debate process.

Although the evaluation of outcomes is perhaps preferable to processes, because these will correspond more *directly* to the desired aims of the exercise, evaluation of exercise processes must often serve as a surrogate to outcomes. That is, if the exercise process is "good" (conducted well according to one's definition) then it would seem *more likely* that the outcomes will be good than if the process is "bad." In the case of *GM Nation?* our focus was therefore on the process, though we retain an interest in following the impact of the exercise.

It is important to recognize the sense of political urgency and importance association with the outcome of the *GM Nation?* debate. Indeed, with the publication of the Field Scale trials imminent, the Steering Board was under considerable pressure from government to deliver its final report according to a strict deadline (it appeared just ten weeks after the end of the debate). Elsewhere we have argued that the resulting sense of hurry resulted in an "over-hasty, under-resourced and methodologically worrying analysis of findings" (Horlick-Jones et al., 2004: 133). Although our own analysis was insufficiently complete to publish anything to coincide with the Steering Board's report, we were invited to submit evidence to a parliamentary inquiry into the conduct of the debate, and we managed to prepare a memorandum in time for the inquiry hearings in October 2003 (Horlick-Jones et al., 2003). It was not until February 2004 that we were ready to publish a substantial evaluation statement (Horlick-Jones et al., 2004). At the time of writing this paper, work on the substantial corpus of data we collected is still taking place.

5.6. Relations with the debate organizers and other stakeholders

The evaluation of public participation exercises is relatively rare. While a number of the problems previously discussed ensure that the *process* is a difficult one, and may deter researchers, practitioners, organizers and sponsors from conducting evaluations, there are further *strategic*, or perhaps, *political*, difficulties that may also militate against the conduct of evaluations. Indeed, as one commentator has noted, "it is almost inevitable that an evaluation has a political dimension to it" (Robson, 2002).

The meaning of the word "political" here needs a little discussion. In many ways, the practicalities of evaluation work are far closer to those of "action research" than those associated with "detached scholarly inquiry." Those actively engaged in the resolution of problem situations whilst conducting research commonly find themselves engaged in "quasi consultancy," necessitating a keen awareness of client/gatekeepers' needs and expectations, and the often messy and convoluted politics of the organizations in question and their relations with the outside world (Horlick-Jones and Rosenhead, 2002).

A number of potentially difficult issues arise from the structural nature of the relationship between evaluators and the sponsors (and their executive) of engagement processes, and one might anticipate finding:

- Evaluators taking up a great deal of time of busy sponsors/executives asking questions and making various other demands;
- Sponsor/executive anxiety that evaluators might interpret complex issues in inaccurate or stereotypical ways;
- A potential culture clash, with evaluators' ways of behaving, talking, and framing issues making sponsors/executives feel uncomfortable;
- The expert status of sponsors/executives being scrutinized by people who are members of their professional "tribe";
- A sense of vulnerability on the part of sponsors/executives, with possible mistakes and difficulties being carefully recorded.

Indeed, whilst conducting our evaluation of the *GM Nation?* debate, we encountered a number of difficulties of this kind. Most could be resolved fairly easily with patience and good will on both sides. On other occasions, concerns and discomfort persisted for an extended period, and they almost certainly had some negative impact on the overall quality of the evaluation, albeit a minor one.

Turning to the wider community of stakeholders, we found that our role as independent evaluators was predominantly accepted, and indeed valued. Of course, in highly charged policy areas like the ones related to transgenic crops, one would expect the evaluation findings to be contested by groups for which the conclusions were politically uncomfortable. In order to allow as much opportunity as possible for stakeholder groups to comment upon our evaluation, we carried out a consultation exercise, entailing posting the report on the University of East Anglia website, making copies of the report widely available, and holding a major day-long presentation and workshop in London to which key stakeholder groups in government, business and non-governmental organizations were invited. Perhaps surprisingly, we received relatively few detailed responses from this consultation exercise. We recognize that there was some discontent about our evaluation within the anti-GM lobby (and its supporters), however, we understand these criticisms to be political in nature rather than being based on a technical critique of our execution of the evaluation process that we set out to conduct. The tactic in these quarters seems to have been to largely ignore our work, which we regard to be a measure of its robust nature.

5.7. The resource demands of the evaluation exercise

Finally, we briefly consider the costs of carrying out the evaluation of the *GM Nation?* public debate. Whilst the core evaluation team comprised six individuals, we were fortunate in being able to gain support from a number of colleagues, and from scholars in two other universities. In total, over 20 individuals were involved in playing some part in the data collection. As noted above, work on the substantial corpus of data we collected is continuing—focusing on the dynamics of the "narrow-but-deep" process—a year or so after the publication of our major evaluation report in February 2004.

We estimate the direct cost of the evaluation process leading up to the February 2004 report (including our major nationwide survey of public opinion) as approximately £175,000. Clearly this is a non-trivial sum, and, in comparison with the cost to the UK government of sponsoring the debate, can be seen as amounting to the equivalent of some 15–20 percent of the debate's running costs. (The official cost of the debate, quoted in the Steering Board's final report (PDSB, 2003: 64), was £650,000.) However, the true costs, taking account of time put in by Steering Board members, as well as organization and publicity for Tier 2 and 3 events conducted by other parties, was undoubtedly far greater. This approximate percentage assumes a "true" figure closer to £1,000,000 for the debate process as a whole. Despite our evaluation being extensive in comparison to other such exercises, we encountered a number of shortcomings and difficulties (discussed in this paper), some of which (particularly the logistical ones) could have been addressed had we had access to more substantial resources. Our experience therefore offers a serious caution to would-be evaluators, and to potential sponsors of engagement exercises who intend to include an evaluation component within the overall exercise budget.

6. Conclusions

The evaluation of public engagement exercises is full of difficulties, ranging from the theoretical (what do we mean by effective public engagement?) and practical (how do we go about measuring effectiveness in the highly complex engagement environment?), to the political (how does one conduct evaluations in an environment full of competing parties with contrasting motives?). In this paper we have discussed a number of these difficulties in the context of a specific evaluation of a recent major engagement event in the UK.

One clear conclusion from this discussion is that there are limitations to any evaluation, particularly given the controversial issue of what we mean by engagement "effectiveness" and the general absence of well-validated instruments. For the unconvinced sponsor (e.g. one that is compelled, perhaps by statute, to conduct a participatory exercise) this realization might even come to be regarded as a boon. If the evaluation process is somehow flawed, then need it be conducted, or if conducted, need the results be heeded? Any criticism produced may thus be open to dispute by those against whom the criticism is leveled, or in the case of a positive assessment of an exercise, by those who disagree with the exercise outcomes. Ultimately, it is the sponsor that wields most power, and the evaluator, in order to gain a commission or to conduct research, may need to concede and conduct their evaluation in a way they would not ideally wish to do. Although unfettered evaluation activity may be undesirable, and possibly troublesome, for a sponsor, excessive sponsor interference in the evaluation process risks biasing the evaluation, and, if discovered by competing stakeholders, might lead to charges that undermine the whole exercise.

It is also important to note that it is difficult to stop some form of evaluation being conducted on any given engagement exercise (particularly after the event). Such evaluations

might be conducted by academics, perhaps sponsored by participants or other interested bodies. If conducted outside the control or influence of the main sponsor, then there may be bias in the evaluation towards the position of the other parties, and bias might also arise from incomplete evaluator knowledge of sponsor motives and other information related to the exercise. Such "unofficial" evaluations may therefore prove problematic. As such, it is probably best for the sponsor to provide for an evaluation to be conducted at the outset, and ensure that the process will be fair from *all* perspectives, particularly its own.

Finally, we wish to reiterate our belief that evaluation is a crucial process for engagement exercises, particularly in the current political climate in which the popularity of engagement as a policy tool is increasing, accompanied by a growth in the number of proposed mechanisms that might be used in implementing such processes. Rigorous evaluation is clearly not easy. However, through the use of sound research methodology, which recognizes the nature of this difficult research environment (including the strategic and political imperatives of those involved), good, insightful evaluations may be conducted. In this paper, and in others reporting the specifics of our evaluation, we have identified a number of ways in which research difficulties might be overcome—for example, using multiple qualitative and quantitative methods to enhance the likely validity of one's evaluation. One important practical consequence of further evaluation research and practice, we hope, will be an enhanced quality of public engagement, in terms of the suitable choice of mechanisms to use in any particular situation, and their appropriate implementation.

Acknowledgements

Work reported in this paper was partly supported by the Programme on Understanding Risk funded by the Leverhulme Trust (RSK990021) and partly supported through two grants from the Economic and Social Research Council including one from the Science in Society program (L144250037). We thank the members of the Debate Steering Board for their cooperation, and the debate secretariat, and staff at the Central Office of Information and CorrWilbourn, for their help. We are pleased to acknowledge the role of a number of colleagues who were involved in the GM debate evaluation; in particular Wouter Poortinga, Tim O'Riordan, Irene Lorenzoni, Karen Bickerstaff, Graham Murdock (of Loughborough University), and Joyce Tait and Ann Bruce (of INNOGEN, Edinburgh University).

References

AEBC (2001) *Crops on Trial*. London: Agriculture and Environment Biotechnology Commission.

Bloor, M. (1978) "On the Analysis of Observational Data; a Discussion of the Worth and Uses of Inductive Techniques and Respondent Validation," *Sociology* 12(3): 545–57.

Bryman, A. and Cramer, D. (1997) *Quantitative Data Analysis*. London: Routledge.

Cabinet Office (2002) *Risk: Improving Government's Ability to Handle Risk and Uncertainty*. London: Cabinet Office Strategy Unit.

Chess, C. (2000) "Evaluating Environmental Public Participation: Methodological Questions," *Journal of Environmental Planning and Management* 43(6): 769–84.

Chess, C. and Purcell, K. (1999) "Public Participation and the Environment: Do We Know What Works?," *Environmental Science and Technology* 33(16): 2685–92.

Clarke, A. (1999) *Evaluation Research: an Introduction to Principles, Methods and Practice*. London: SAGE.

Fiorino, D.J. (1990) "Citizen Participation and Environmental Risk: a Survey of Institutional Mechanisms," *Science, Technology, & Human Values* 15(2): 226–43.

Frewer, L.J. (1999) "Risk Perception, Social Trust, and Public Participation into Strategic Decision-making: Implications for Emerging Technologies," *Ambio* 28: 569–74.

Funtowicz, S. and Ravetz, J. (1992) "Risk Management as a Post-normal Science," *Risk Analysis* 12(1): 95–7.

Gaskell, G. and Bauer, M.W. (2001) *Biotechnology 1996–2000: The Years of Controversy*. London: The Science Museum.

Glaser, B. and Strauss, A. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. London: Weidenfeld & Nicholson.

Halford, N.G. (2004) "Prospects for Genetically Modified Crops," *Annals of Applied Biology* 145(1): 17–24.

Horlick-Jones, T. (1998) "Meaning and Contextualisation in Risk Assessment," *Reliability Engineering and System* Safety 59: 79–89.

Horlick-Jones, T. (2004) "Experts in Risk? . . . Do They Exist?," *Health, Risk & Society* 6(2): 107–14.

Horlick-Jones, T. and Rosenhead, J. (2002) "Investigating Risk, Organisations and Decision Support through Action Research," *Risk Management: an International Journal* 4(4): 45–63.

Horlick Jones, T., Pidgeon, N., Walls, J. and Rowe, G. (2002) *Proposal for the Evaluation of the UK Public Debate on the Possible Commercialisation of Genetically Modified Crops.* Paper submitted to the GM Public Debate Steering Board, October 2002.

Horlick-Jones, T., Walls, J., Rowe, G., Pidgeon, N., Poortinga, W. and O'Riordan, T. (2004) *A Deliberative Future? An Independent Evaluation of the GM Nation? Public Debate about the Possible Commercialisation of Transgenic Crops in Britain, 2003*. University of East Anglia, Programme on Understanding Risk, Working Paper 04-02.

Horlick-Jones, T., Walls, J., Rowe, G., Pidgeon, N., Poortinga, W. and O'Riordan, T. (submitted) "On Evaluating the *GM Nation?* Public Debate about the Commercialisation of Transgenic Crops in Britain."

Horlick-Jones, T., Walls, J., Rowe, G., Pidgeon, N., Poortinga, W., O'Riordan, T., Murdock, G., Tait, J. and Bruce, A. (2003) "Memorandum Submitted by the Understanding Risk Team and Collaborators," in House of Commons Environment, Food and Rural Affairs Committee *Conduct of the GM Public Debate*. Eighteenth Report of the Session 2002–03 HC 1220, Ev50–Ev56. London: The Stationery Office.

House of Commons Environment, Food and Rural Affairs Committee (2003) *Conduct of the GM Public Debate.* Eighteenth Report of the Session 2002–03 HC 1220. London: The Stationery Office.

House of Lords Select Committee on Science and Technology (2000) *Science and Society Third Report*. HMSO, HL Paper 38.

Irwin, A. and Wynne, B. (eds) (1996) *Misunderstanding Science? The Public Reconstruction of Science and Technology*. Cambridge: Cambridge University Press.

Jasanoff, S. (1990) *The Fifth Branch: Scientific Advisors as Policy Makers*. Cambridge, MA: Harvard University Press.

Joss, S. (1995) "Evaluating Consensus Conferences: Necessity or Luxury?," in S. Joss and J. Durant (eds) *Public Participation in Science: The Role of Consensus Conferences in Europe*, pp. 89–108. London: The Science Museum.

Lord Phillips of Worth Matravers, Bridgeman, J. and Ferguson-Smith, M. (2000) *The BSE Inquiry* [The Phillips Report]. London: The Stationery Office.

Myhr, A.I. and Traavik, T. (2003) "Genetically Modified (GM) Crops: Precautionary Science and Conflicts of Interests," *Journal of Agricultural and Environmental Ethics* 16(3): 227–47.

National Research Council (1996) *Understanding Risk: Informing Decisions in a Democratic Society*. Washington DC: National Academy Press.

Nielsen, C.P., Thierfelder K. and Robinson, S. (2003) "Consumer Preferences and Trade in Genetically Modified Foods," *Journal of Policy Modeling* 25(8): 777–94.

Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter.

Patton, M. (1990) *Qualitative Evaluation and Research Methods*, 2nd edn. London: Sage.

Pidgeon, N.F., Poortinga, W., Rowe, G., Horlick-Jones, T., Walls, J. and O'Riordan, T. (2005) "Using Surveys in Public Participation Processes for Risk Decision-making: the Case of the 2003 British *GM Nation?* Public Debate," *Risk Analysis* 25(2): 467–79.

POST (2001) *Open Channels: Public Dialogue in Science and Technology*. Parliamentary Office of Science and Technology, Report 152.

Public Debate Steering Board (PDSB) (2003) *GM Nation? The Findings of the Public Debate*. Department of Trade and Industry. URL: www.gmnation.org.uk.

Robson, C. (2002) *Real World Research*, 2nd edn. Oxford: Blackwell.

Rossi, P.H., Freeman, H.E. and Lipsey, M.W. (1999) *Evaluation: a Systematic Approach*, 6th edn. London: Sage.

Rowe, G. and Frewer, L.J. (2000) "Public Participation Methods: a Framework for Evaluation," *Science, Technology, & Human Values* 25(1): 3–29.

Rowe, G. and Frewer, L.J. (2004) "Evaluating Public Participation Exercises: a Research Agenda," *Science, Technology, & Human Values* 29(4): 512–56.

Rowe, G. and Frewer, L.J. (2005) "A Typology of Public Engagement Mechanisms," *Science, Technology, & Human Values* 30(2): 251–90.

Rowe, G., Horlick-Jones, T., Walls, J., Poortinga, W. and Pidgeon, N. (submitted) "Analysis of a Normative Framework for Evaluating Public Engagement Exercises: Reliability, Validity and Limitations."

Rowe, G., Marsh, R. and Frewer, L.J. (2004) "Evaluation of a Deliberative Conference," *Science, Technology, & Human Values* 29(1): 88–121.

Shaw, I. (1999) *Qualitative Evaluation*. London: Sage.

Silverman, D. (1998) *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction*. London: SAGE.

Thorpe, A. and Robinson, C. (2004) "When Goliaths Clash: US and EU Differences over the Labelling of Food Products Derived from Genetically Modified Organisms," *Agriculture and Human Values* 21(4): 287–98.

UK Government (2002) *UK Government Response to AEBC Advice Submitted in April 2002*. Press release, URL: http://www.aebc.gov.uk/aebc/reports/public_debate_advice.shtml.

Walls, J., Pidgeon, N., Weyman, A. and Horlick-Jones, T. (2004) "Critical Trust: Understanding Lay Perceptions of Health and Safety Risk Regulation," *Health, Risk & Society* 6(2): 133–50.

Wynne, B. (1991) "Knowledges in Context," *Science, Technology, & Human Values* 16(1): 111–21.

Authors

Gene Rowe is currently a senior scientist in the Consumer Science Group at the Institute of Food Research, Norwich (UK). His Ph.D., gained from the Bristol Business School at the University of the West of England, concerned the use of nominal groups to improve human judgment and decision-making. As well as a continuing interest in judgment and decision-making, his research activities, and publications, have also spanned topics from expert systems and forecasting to risk perception and public participation. Much of his recent work has focused on the issue of evaluating the effectiveness of public participation exercises. Correspondence: Gene Rowe, Institute of Food Research, Norwich Research Park, Colney, NR4 7UA, UK.

Tom Horlick-Jones is Senior Research Fellow in the School of Social Sciences at Cardiff University (Wales, UK), and was team leader of the GM debate evaluation project. He is an experienced researcher who has specialized in a range of issues concerned with risk and human behavior. His current research interests include risk, language and practical reasoning, risk-related and decision-making practices within government and business organizations, and problems concerning knowledge, expertise and disciplinarity.

John Walls is Senior Research Associate in the School of Environmental Sciences at the University of East Anglia (UK). His research interests include the changing governance of new technologies and environmental risks; public trust in regulatory institutions; and investigating influences on safety culture in organizations.

Nick Pidgeon is Professor of Environmental Sciences at the University of East Anglia (UK) and director of the *Understanding Risk* program. His research interests comprise the psychological and social aspects of risk perception and communication; human and organizational causes of major industrial accidents; and social science research methods, with a particular emphasis upon the use of qualitative and mixed-method approaches.