Analysis of a normative framework for evaluating public engagement exercises: reliability, validity and limitations

Gene Rowe, Tom Horlick-Jones, John Walls, Wouter Poortinga, Nick F. Pidgeon

Public Understand. Sci. **17** (2008) 419–441

Analysis of a normative framework for evaluating public engagement exercises: reliability, validity and limitations

Gene Rowe, Tom Horlick-Jones, John Walls, Wouter Poortinga and Nick F. Pidgeon

Over recent years, many policy-makers and academics have come to the view that involving the public in policy setting and decision-making (or "public engagement") is desirable. The theorized benefits of engagement (over traditional approaches) include the attainment of more satisfactory and easier decisions, greater trust in decision-makers, and the enhancement of public and organizational knowledge. Empirical support for these advantages is, however, scant. Engagement processes are rarely evaluated, and when they are, the quality of evidence is generally poor. The absence of standard effectiveness criteria, and instruments to measure performance against these, hinders evaluation, comparison, generalization and the accumulation of knowledge. In this paper *one* normative framework for evaluating engagement processes is considered. This framework was operationalized and used as part of the evaluation of a recent major UK public engagement initiative: the 2003 *GM Nation?* debate. The evaluation criteria and processes are described, and their validity and limitations are analyzed. Results suggest the chosen evaluation criteria have some validity, though they do not exhaustively cover all appropriate criteria by which engagement exercises ought to be evaluated. The paper concludes with suggestions on how to improve the framework.

1. Introduction: why public engagement?

There has been much debate in contemporary democratic societies about the best way to develop public policy, particularly in controversial domains such as the management of risks and the development of novel technologies (e.g. Dryzek, 2000; Kasperson et al., 1999). The traditional manner of dealing with policy dilemmas essentially involves responsible agencies and their expert advisors first determining policy and then communicating their solutions *to* the wider public (e.g. Jasanoff, 1990). This approach, however, has been compromised by a number of controversies (e.g. in the UK, the Bovine Spongiform Encephalopathy (BSE) crisis), which, it has been argued, have led to the diminution of trust in such agencies, and in scientists and other expert members of associated policy communities (e.g. Jensen, 2004; Walls et al., 2004).

 ISSN 0963-6625 DOI: 10.1177/0963662506075351

Partly in response to a perceived loss of trust in governments and expert bodies, a "novel" approach to policy-making has emerged, rooted in the idea of "public engagement," in which the public (including stakeholder communities) is more directly involved in policy development and decision-making. The *theorized* benefits of such enhanced involvement include: better quality decisions (achieved through including lay knowledge and values); easier decisions (through pre-empting public discontent); greater trust in decision-makers (achieved through demonstrating concern for public views); and enhanced public and organizational knowledge (through mutual learning) (discussed in Beierle and Cayford, 2002; Rowe and Frewer, 2005). Unfortunately, there is little empirical evidence to support these claims.

This paper considers how evidence of the benefits or otherwise of engagement might be obtained. It begins by considering the current paucity of evaluations and why this should be so. One set of normative evaluation criteria drawn from the literature is then described, followed by a discussion of how this was used in the evaluation of a major public engagement exercise in the UK: the 2003 *GM Nation?* public debate on the possible commercialization of transgenic crops. The focus of discussion here is on *the development of instruments to operationalize these criteria*, and on ascertaining the validity of the criteria and the quality of the instruments. The wider process of evaluation is discussed elsewhere, as are the results of our evaluation of the debate (Horlick-Jones et al., 2006; Pidgeon et al., 2005; Rowe et al., 2005).

2. The issue of evaluation

In a review of evaluations of public engagement exercises, Rowe and Frewer (2004) found little compelling *empirical* evidence for the advantages often assumed to be associated with public engagement. Indeed, there are very few cases of empirical evaluation in the academic literature at all. There are arguably two main reasons for this. The first is that public engagement is often seen as an end in its own right, as opposed to a means to an end. This viewpoint would seem especially cogent with respect to practitioners charged with conducting engagement according to regulatory requirements or organizational policy. From this perspective, the very act of engaging with the public indicates success, and *evaluation* itself becomes a superfluous concept. This perspective is highly unsatisfactory: consider, for example, a hypothetical exercise (perhaps a public meeting) that results in rancor and dissatisfaction as a consequence of being poorly facilitated, held at an inappropriate time and place, and addressing irrelevant questions (the answers to which are in any case subsequently ignored by sponsoring bodies). In what sense could such an exercise be considered a success? A second reason for a dearth of evidence on the quality of engagement is that evaluation is *difficult*. In the absence of a widely accepted framework for conducting evaluation, and importantly, instruments that may be applied to enable this, those conducting engagement exercises are unclear as to how evaluation should be done. It is perhaps of no surprise that when evaluations *are* conducted, they are often done in a rather informal and subjective manner, in which the evaluators (often the same people as those conducting the exercise) limit themselves to commenting upon apparent *positives* that emerge from the considered process.

The first step to conducting an evaluation is to define what is meant by a "successful" or "effective" public engagement exercise (Rowe and Frewer, 2004). Of course, this is also difficult, in the sense that "engagement" is not a simple concept. There are various reasons for conducting engagement (e.g. to ascertain public views and/or to provide an input into a decision-making process), and various methods for achieving this (Rowe and Frewer (2005) list over 100 different mechanisms). In this sense, there may be no *one* appropriate "universal" definition of what constitutes an "effective" exercise, although there may be a number of "local" definitions related to specific engagement purposes and/or types of engagement mechanism. In spite of a lack of consensus, Rowe and Frewer (2004) found that many of the (few) empirical evaluations that have been conducted appear, at least implicitly, to *assume* that there *are* certain universal characteristics applicable to successful engagement exercises—as indicated by the use of similar effectiveness criteria across evaluations. Hence, it is often stated or assumed that a good engagement exercise will have participants who are somehow *representative* of the relevant population of public/stakeholders (i.e., and that biased sampling is a sign of an unsuccessful exercise), and also that a good exercise will have clear *impacts* upon the policy setting or decision-making of the sponsoring organization (i.e., and that lack of impact indicates a poor exercise).

Rowe and Frewer (2004) argue that the second step in evaluation is to develop instruments to measure success according to one's stated effectiveness criteria. There is then a need to test and evaluate these instruments to ensure their appropriateness for this task; that is, to assess their reliability, validity, and usability. However, even in empirical evaluations that detail and justify the evaluation criteria used, instrument development is rarely discussed, and neither is the issue of *instrument quality* (a small counterexample to this tendency is the work by Halvorsen, 2001).

In this paper, one operationalized set of effectiveness criteria is described. Through the process of conducting an evaluation, sufficient data were accumulated to allow commentary upon the appropriateness and validity of the criteria, as well as of the quality of the developed instruments and their limitations.

3. A standard framework: the Rowe–Frewer criteria

One set of normative, "universal" evaluation criteria is that described by Rowe and Frewer (2000). These authors reviewed the academic literature on public engagement and identified a number of recurring themes concerning the necessary requirements for an engagement exercise to be successful. In their framework, these themes were translated into either "Acceptance Criteria," related to whether an exercise would likely be accepted by participants as fair, or "Process Criteria" related to the effective construction and implementation of a procedure (a distinction that bears some parallel to that made by Webler (1995) between "fairness" and "competence"). The nine criteria are as follows.

Acceptance Criteria:

- Representativeness: 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.

Process Criteria:

- Resource Accessibility: public participants should have access to the appropriate resources to enable them to successfully fulfill 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.

Rowe and Frewer (2000) used these criteria to subjectively evaluate a number of general engagement mechanisms. They subsequently developed a number of instruments and processes to enable a more "objective" (or at least, structured) analysis (e.g. Rowe, Marsh and Frewer (2004) developed an evaluator's checklist and a short questionnaire that they used to evaluate a deliberative conference). These instruments were used as a basis for developing a long and a short questionnaire to evaluate a number of the nine criteria in the setting of a major national public engagement exercise in the UK. The nature of the instruments is described after a brief overview of the exercise.

4. Applying the framework: the *GM Nation?* debate

The data used in this paper are drawn from an evaluation of the *GM Nation?* public debate, a major government-sponsored public engagement exercise that took place in the UK in 2003. One objective of this exercise was to gather information about public views on genetically modified (GM) food and crops in order to inform UK government decision-making regarding the potential future commercialization of the technology. The debate organizers sought to achieve this by means of a set of activities, including a large number of public meetings, an interactive website, and a small set of specially convened focus groups of pre-selected individuals (PDSB, 2003; Horlick-Jones et al., 2006).

The concern of this paper is with just one component of the exercise—arguably the main, publicly visible element. This comprised a series of six major public meetings (known as "Tier 1" meetings, since there were also more local "Tier 2" and "Tier 3" meetings), which anybody could attend. These meetings (three in England and one each in Wales, Scotland and Northern Ireland) were conceived of as "national" high profile, professionally facilitated events, and they attracted approximately 1000 participants in total. Generally, the meeting process was as follows. As participants arrived, a commissioned video (showing conversations between people addressing some of the main issues identified by participants in previous workshops) was played in the background. As participants were seated they had access to booklets that gave pro and con information on a range of the most salient GM-related issues identified in the previous workshops. Participants were given no time to read these as such: they were rapidly broken up into a large number of smaller groups, which were instructed to elect one person as moderator. There followed discussion within these groups, directed at their own whims, on the general debate topic. Finally, the participant moderators of the different groups presented their own summaries, one after the other, in plenary. The chair of the overall debate (generally a publicly known figure) would then wind down the debate and direct participants to complete the organizers' feedback questionnaire regarding views on GM foods and crops (comprising 13 standard attitude questions), as well as our own questionnaires (discussed below).

To conduct the evaluation, we adopted a multi-method approach using qualitative and quantitative methods (e.g. Rossi et al., 1999; Shaw, 1999). Specifically, we used participant questionnaires, structured observation, ethnographic techniques, in-depth interviews (with Steering Board members and other key stakeholders), media and document analysis, and a major survey of public opinion (Horlick-Jones et al., 2006; Pidgeon et al., 2005). Our evaluation referenced three distinct sets of evaluation criteria: first, those derived from the aims of the debate organizers; second, the Rowe–Frewer normative criteria (discussed above); and third, a set *derived from* an analysis of participants' responses to questionnaires we provided (i.e. not pre-defined), which indicated how *they* judged the success or otherwise of various aspects of the event. This diversity of evaluative benchmarks is important, as there is a school of thought that regards selecting evaluation criteria *prior* to an exercise taking place as problematic, risking the imposition of a potentially inappropriate framework upon the data (proponents of this view might argue for more inductive, qualitative, case study-based evaluations instead). Rowe et al. (2005) use the term *assessments* to describe evaluations that are not based upon pre-defined criteria (and to differentiate these from *evaluations* per se), and note that these have various limitations, e.g. in terms of result generalizability (see also Clarke, 1999). Our approach therefore involved methodological plurality rather than being constrained by one or other of these evaluation paradigms (cf. Patton, 1990): we carried out an *evaluation*, in being guided by criteria identified at the start of the process, though our multimethod approach provided us with a capacity to *learn* from the process, and so an ability to arrive at emergent findings (e.g. Bloor, 1978).

The rest of the paper focuses on our use of the normative criteria in evaluating the six "Tier 1" public meetings.

5. Evaluation using the normative criteria: design and implementation

Rowe and Frewer (2000) suggested that their nine evaluation criteria might not be appropriate in every situation. Rowe, Marsh and Frewer (2004), for example, omitted using the criterion of "Early Involvement" because the sponsors of the conference they evaluated considered it inappropriate. With regard to the *GM Nation?* debate, it was decided to omit consideration of the "Cost Effectiveness" criterion, because it did not seem sensible to ask participants to assess this aspect of the debate, and also because it seemed to us difficult to reconcile with the "Process Criteria" concept discussed in the original paper (i.e. this not being specifically related to the *quality of process enactment*). However, the other eight criteria were adopted in this evaluation, although "Resource Accessibility" was relabeled as "Resources," and "Structured Decision Making" as "Structured Dialogue" (as the exercise participants did not strictly have any decision-making requirements). In the rest of the paper, these are the criteria labels used.

Rowe, Marsh and Frewer (2004) produced a questionnaire with one question addressing each of the evaluation criteria. A "short questionnaire" was developed along similar lines, but with two questions addressing the "Resources" criterion (question wordings are shown in Table 2). The same 7-point scale was used for each question: Very Strongly Agree, Strongly Agree, Moderately Agree, Neither Agree nor Disagree, Moderately Disagree, Strongly Disagree, Very Strongly Disagree. The short questionnaire included a number of demographic and socioeconomic questions.

Use of such a short questionnaire is not ideal, however, as single items might be misinterpreted and hence not address the concepts intended. Better is to use several questions for each criterion, increasing the potential reliability of one's instrument. Therefore, in this study, a longer questionnaire was produced, which included a number of questions per criterion. Appropriate analysis enables identification of questions addressing a similar concept (i.e. with highly correlated responses) for inclusion in the measurement instrument. (The exact wording of these questions is shown in Table 3.) There were only single questions addressing the criteria "Task Definition," "Representativeness" and "Early Involvement," though two addressed "Resources," "Transparency" and "Influence," four addressed "Independence" and six addressed "Structured Dialogue." The variation in number of items reflects the ease with which we could think of potentially appropriate questions. The same 5-point scale was used for each question (Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree), plus there was a "don't know" option. Ideally the two questionnaires should have had similar rating scales, but differences arose from a combination of logistic and design tensions. That is, the original template questionnaire had a 7-point scale, but no "don't know" option. It was felt better to include such an option, but this led to the layout being cluttered—a problem resolved by reducing the scale to 5-point (which was felt sufficient to discriminate between people's opinions).

The long questionnaire items were placed within a larger questionnaire, which also asked demographic details. Following the approach of Rowe et al. (2004), there were also a number of open questions that asked participants what they thought were good and bad about the exercise. These occurred *prior to* the closed questions, to ensure that the nature and terminology of the closed questions did not bias the relatively unconstrained evaluations of respondents. In contrast, Rowe et al. conducted interviews after the event to gain this information. The aim of both approaches is to allow participants to express *in their own words* the criteria by which they regarded the exercise as a success or failure. The responses to these questions will be used later to speak to the validity and inclusiveness of the normative effectiveness criteria.

In addition to the two participant questionnaires, the meetings were evaluated through one member of the evaluation team, at each event, following an observation schedule. This schedule, which described the normative evaluation criteria and aspects to look for, sought to establish a uniform approach to data gathering. This was less detailed than the "evaluation checklist" used for this purpose by Rowe et al. (2004), so no detailed analysis is presented here: the observations will simply be discussed in the broad to indicate, later, whether there was general agreement or not between evaluator observations and participant questionnaire evaluations (more details on the observations are given in Horlick-Jones et al., 2006).

The use of two separate questionnaires reflects no clever design, but was a result of circumstance. It is preferable to present questionnaires to participants immediately after an event to increase response rate and attain immediate opinions. Logistically, this proved difficult: as the events lasted several hours, the event organizers were not keen on overtaxing participants by giving them extensive questionnaires at the end. As a compromise, we were allowed to present participants with the short questionnaire for immediate completion, and the longer questionnaire for participants to complete at home (to return to us via postage pre-paid envelopes). This ensured that we got at least *some* commentary from most participants.

In the next section, participants' responses to the questionnaires are described. This will be followed by commentary on the relevance and validity of the instruments, ascertained by reference to the answers to the open questions and comparison with the structured researcher observations.

6. Results: responses to the two questionnaires

The respondents

Table 1 summarizes the characteristics of respondents. The second column indicates the numbers attending the six events, according to the organizers' official website. These figures are estimates only—hence, the figures in the third column are also *estimates* (of the percentage of *participants* who responded to the first questionnaire). Otherwise, the third column indicates that there were 620 completed copies of the first questionnaire. The estimated response rates range from just under 50 percent to over 70 percent. The fourth column shows the number of respondents to the first questionnaire who also completed the second (returning these through the post), revealing that approximately two-thirds did so. The final column shows the

Table 1. Responses to questionnaires from the six Tier 1 conferences

Debate location	Total participants according to organizers*	Responses to Q1 (percent of participants)	Responses to Q2 (percent of Q1 respondents)	Extra responses to Q2 (total Q2 responses)***
Belfast	100	63 (63.0%)	35 (58.7%)	2 (37)
Birmingham	126	75 (59.5%)	54 (76.0%)	3 (57)
Glasgow	140	89 (63.6%)	56 (65.2%)	2 (58)
Harrogate	250	122 (48.8%)	88 (74.6%)	3 (91)
Swansea	180	130 (72.2%)	76 (62.3%)	5 (81)
Taunton	120–210**	141 (Unknown)	90 (68.1%)	6 (96)
Total	916–1006	620	399 (67.7%)	21 (420)

*Total participant numbers come from the organizers' website at: http://www.gmpublicdebate.org/ut_13/ut_13_25.htm
**The Taunton number is recorded as "90 in the morning; 120 in the afternoon": some of those attending in the morning did not stay until the afternoon, and new participants turned up at that time. The total is likely to be much closer to 120 than 210, and duplicate returns are possible (we have checked demographic information to confirm no repeats).
***Respondents who completed Questionnaire 2, but not Questionnaire 1.

number of extra responses received from people who only completed the second questionnaire, but did not complete the first at the event. The column also shows the total number of completed copies of Questionnaire 2 received, which totaled 420. This table confirms that we achieved good response rates to our questionnaires.

Averaging over all events reveals an almost equal gender split of respondents: for Questionnaire 1, there were 51.0 percent male and 49.0 percent female, and for Questionnaire 2 there were 48.4 percent male to 51.6 percent female. These figures are not greatly dissimilar to the national UK population (48.6 percent male to 51.4 percent female, according to the UK 2001 Census—see: http://www.statistics.gov.uk/default.asp). The mean age of respondents to both questionnaires was approximately 50 years, suggesting a slight bias towards more elderly members of the population. However, there was one notable bias in the nature of respondents, and this concerned their education level. About two-thirds of respondents had a degree (66.3 percent of Questionnaire 1 and 73.1 percent of Questionnaire 2), while approximately one-third of all respondents (29.4 percent of Questionnaire 1 and 33.3 percent of Questionnaire 2) claimed to have a higher degree (i.e. "postgraduate qualification (Masters or PhD)"). According to the UK 2001 Census (see: http://www.statistics.gov.uk/census2001/profiles/UK-A.asp), 19.6 percent of the population between 16 and 74 have qualifications at degree level or higher. This suggests that participants were significantly atypical of the national population with regard to education level. This issue will be returned to when discussing the Representativeness criterion.

The issue of reliability

Reliability is a necessary, though not necessarily sufficient, condition for instrument validity: a reliable instrument need not be valid (it could be measuring something different to that intended), though an unreliable instrument *cannot* be valid (as it is uncertain *what* is being measured). There are various ways to determine instrument reliability, though these are difficult to apply in highly complex natural situations in which "experimenters" have no control over the process in which they are trying to develop instruments. For example, one relevant concept is test–retest reliability, determined by applying a particular instrument on two separate occasions. If the instrument (e.g. questionnaire) is reliable, then similar results should emerge from the two applications (e.g. a scale would be reliable if it gave the same weight for the same person on different occasions). In public engagement contexts, it is difficult to persuade sponsors to allow repeated polling using an identical questionnaire. On this occasion, however, we were able to present two fairly similar questionnaires on different occasions—a shorter version completed after the conclusion of the meetings, and a longer version being completed some time later. The questionnaires were not identical, though the questions were highly similar, and the results from the two are—as will be described—very similar.

In the following analysis, data are aggregated across all six meetings. Although there were occasionally minor differences across these (i.e. respondents in some rated these significantly better or worse on certain criteria than did respondents in others), the general trends are fairly consistent, and space does not allow more fine-grained data interrogation.

Table 2 summarizes responses to the nine items in Questionnaire 1. The three "agree" and three "disagree" response categories have been conflated to simplify data presentation. Table 3 shows responses to the longer Questionnaire 2. Again, the different degrees of agreement/disagreement have been conflated to simplify presentation. The mean figures in

Table 2. Mean responses across the six conferences on Questionnaire 1

Criterion	Item*	Agree %	Neither agree nor disagree %	Disagree %	Mean response (SD) [*N*]
Representativeness	I think that the people taking part in this event are a fair cross-section of members of the public	33.9	9.6	56.5	4.51 (1.74) [616]
Independence	I feel that the people running the event were not promoting a specific view on the issues around GM	60.9	20.7	18.4	3.25 (1.38) [613]
Early Involvement	I think that this event has taken place too late to allow me and the other participants to influence Government policy on GM	79.5	9.4	11.1	2.33 (1.53) [614]
Influence	I think that feedback from this event will be taken seriously by the Government	15.9	12.3	71.8	5.25 (1.48) [616]
Transparency	I *don't* think there is any kind of "hidden agenda" behind this event	23.2	21.3	55.5	4.68 (1.60) [616]
Resources	The event provided me with all the information I wanted to enable me to contribute as I wished	33.5	15.8	50.7	4.50 (1.66) [614]
Resources	The event seemed to provide sufficient time for everyone who wanted to contribute to have their say	48.8	6.3	44.9	4.10 (1.67) [615]
Task Definition	It was clear to me what I was supposed to be doing throughout the event	71.2	9.2	19.6	3.19 (1.31) [618]
Structured Dialogue	The way the event was run allowed me to have my say	77.7	7.5	14.8	3.02 (1.34) [615]

*Very strongly agree = 1; strongly agree = 2; moderately agree = 3; neither agree nor disagree = 4; moderately disagree = 5; strongly disagree = 6; very strongly disagree = 7.

Table 3. Mean responses across the six conferences on Questionnaire 2

Criterion	Item*	Disagree %	Neither agree nor disagree %	Agree %	Mean response (SD) [*N*]
Representativeness 1	The people who attended the event were fairly typical of the sort of people who would be affected by GM issues	37.3	14.1	48.7	3.09 (1.34) [384]
Independence 1	The event was run in an unbiased way	20.1	22.4	57.5	3.44 (1.01) [397]
Independence 2	The facilitators were biased by the views of the people who commissioned this event	59.2	28.9	11.9	2.46 (0.91) [360]
Independence 3	There was too much control by the facilitator over the way the event was run	65.7	23.1	11.2	2.39 (0.89) [403]
Independence 4	The information that was given to participants was fair and balanced	33.4	24.4	42.2	3.04 (1.11) [386]
Early Involvement 1	The event has taken place too late in the policy-making process to be influential	13.1	9.6	77.3	4.11 (1.17) [375]
Influence 1	The people who commissioned this event will not take any action on the views and recommendations made by participants	17.1	23.9	59.0	3.72 (1.12) [327]
Influence 2	Feedback from this event will be influential on the future of GM food and crops in the UK	64.7	19.6	15.7	2.25 (1.10) [306]
Transparency 1	It was not clear how participants in the event were selected	24.7	22.3	53.0	3.43 (1.16) [373]
Transparency 2	It is not clear to me how the results of this event will be used	11.5	5.2	83.3	4.05 (0.97) [407]
Resources 1	There was not enough time to fully discuss all the relevant issues	19.0	12.5	68.5	3.82 (1.18) [416]
Resources 2	Participants had access to any information they wanted	58.4	19.4	22.2	2.45 (1.10) [391]
Task Definition 1	I was confused at times about what I had to do	64.0	12.2	23.8	2.54 (1.03) [386]
Structured Dialogue 1	All relevant issues were covered	66.6	9.6	23.8	2.38 (1.14) [407]

(continued)

Table 3. (continued)

Criterion	Item*	Disagree %	Neither agree nor disagree %	Agree %	Mean response (SD) [*N*]
Structured Dialogue 2	I didn't get the chance to say all that I wanted to say	36.7	16.6	46.7	3.17 (1.15) [409]
Structured Dialogue 3	I felt there was so much information that it was difficult to assess it all	55.1	14.7	30.2	2.74 (1.20) [408]
Structured Dialogue 4	The facilitator encouraged everyone to have their say, no matter how little or how much they knew about the subject	11.5	14.7	73.8	3.76 (0.906) [407]
Structured Dialogue 5	The event was well facilitated	12.8	23.8	63.4	3.59 (0.91) [404]
Structured Dialogue 6	The event was well organised and structured	28.3	22.5	49.2	3.20 (1.10) [409]

*Strongly disagree = 1; disagree = 2; neither agree nor disagree = 3; agree = 4; strongly agree = 5 (there was also a "don't know" option).

both tables give an indication of the magnitude of agreement/disagreement. The longer questionnaire used a 5-point scale and "don't know" option (to hopefully disentangle respondents with no opinion from those with an ambivalent opinion), as opposed to a 7-point scale in the short questionnaire. "Don't know" responses are not included in Table 3 or discussed in detail in the subsequent analysis.

Starting with the Representativeness criterion, it is clear in Table 2 that respondents tended to disagree that those taking part in the events were a fair cross-section of the population (over half "disagreed" to some extent with this statement, compared with only about one-third agreeing). Responses to the Representativeness item in Questionnaire 2 were somewhat more equivocal (see Table 3). Here, approximately one half agreed that those attending the events were fairly typical of those who would be affected by GM issues, while less than 40 percent disagreed (the mean response was only just over "3"—a slight positive evaluation here, given that the agree/disagree scale was reversed to that in Questionnaire 1).

Regarding the Independence criterion, respondents were much more positive. Over 60 percent of Questionnaire 1 respondents agreed that those running the debate *were not* promoting a specific view on the debate topic, compared to less than 20 percent who disagreed. There were four questions intended to address this criterion in Questionnaire 2 (note: the questions did not occur in the order presented in Table 3, but in a more random order ensuring that similar questions addressing the same criterion/issue were not adjacent). For each of these, responses were positive about the conferences. Thus, for two questions, the greater proportion *agreed* (than disagreed) with *positive statements* (the event was run in an unbiased way, and the information presented to participants was fair and balanced), and for the other two, the majority *disagreed* with *negative statements* (the facilitators were biased by the views of those commissioning the event, and there was too much control by facilitators over how the event was run).

Regarding "Early Involvement," respondents were far more negative about the conferences. Nearly 80 percent of Questionnaire 1 respondents agreed that the events had "taken place too late to allow [them] … to influence Government policy …" There was only one question addressing this concept in Questionnaire 2, and again the majority (nearly 80 percent) agreed that the event had taken place too late to be influential.

Respondent skepticism was also evident regarding the potential *influence* of the conferences: over 70 percent of Questionnaire 1 respondents disagreed that feedback from these would be taken seriously by the government. Questionnaire 2 confirmed this assessment: for each of the two questions intended to address the Influence criterion, the majority expressed negative sentiments, that is, nearly 60 percent agreed that those commissioning the events would not take action on recommendations arising from the conferences, and over 60 percent disagreed that feedback from the events would influence the future of GM foods and crops in the UK. Furthermore, a substantial proportion of respondents expressed uncertainty over these matters, with 82 out of 420 potential respondents (i.e. 19.5 percent) selecting the "don't know" option for the "Influence 1" item, and 100 (i.e. 23.8 percent) selecting the "don't know" option for "Influence 2."

Again, the majority held negative views with regard to the one item related to the Transparency criterion in Questionnaire 1 (disagreeing that there *wasn't* a "hidden agenda" to the conferences), and the two questions in Questionnaire 2 (in each case the majority tended to agree with negative statements regarding how participants were selected and how results would be used).

There were two aspects of resources (Resources criterion) that were assessed in Questionnaire 1, with mixed responses. Respondents generally felt that the conferences *did not* provide them with all the information they wanted, though they were more equivocal about whether they had sufficient time resources (slightly more agreed than disagreed). Resources were also assessed by two items in Questionnaire 2, again relating to time and information needs. Here, respondent assessments were negative for *both* respects, with the majority agreeing with a negative statement (there was not enough time) and disagreeing with a positive statement (participants had access to any information they wanted).

With regard to the Task Definition criterion, respondents to Questionnaire 1 were more positive, generally agreeing that they were clear about what they were supposed to be doing in the events (over 70 percent agreed and less than 20 percent disagreed). In Questionnaire 2 there was also just a single item intended to address this criterion, and the majority (nearly two-thirds) disagreed that they were confused about what they had to do in the events, i.e. reflecting a positive assessment of the meetings on this criterion.

The Structured Dialogue criterion was assessed by only a single item in Questionnaire 1 ("The way the event was run allowed me to have my say"), resulting in a positive assessment (over 70 percent agreed and less than 20 percent disagreed). In Questionnaire 2 this criterion was addressed by six different questions. A mixed message emerges from these. For two, assessments were generally negative: respondents generally did not feel that all relevant issues were covered and thought that they did not get the chance to say all that they wanted to say. However, for four questions, assessments were positive: the majority disagreed that there was too much information to handle, agreed that the facilitator encouraged everyone to have their say, agreed that the event was well facilitated, and agreed that it was well organized and structured.

In summary, respondents were generally negative about the meetings, which scored relatively poorly on the criteria of Representativeness, Early Involvement, Influence, and Transparency. However, respondents were more equivocal about the sufficiency of resources (Resources criterion), and were generally positive about the independence of the organizers (Independence criterion), how well defined were their tasks (Task Definition criterion), and the way in which the events were structured to allow them to have their say (Structured

Dialogue criterion). Significantly, responses to items intended to address the different criteria were similar across the two questionnaires. This does not allow us to conclude that either questionnaire is formally "reliable," but does suggest the items in each were assessing similar aspects, which should increase our confidence that the questionnaires are reliable measures of participant perceptions.

There are a number of other things that can be done with the data to assess reliability. First, considering Questionnaire 1, we conducted a Principal Components Analysis (PCA) on the data (using Varimax rotation). This is a method of data reduction that considers the extent to which different items measure the same thing (are correlated). This analysis revealed a two-component solution (i.e. two components with eigenvalues over 1.0). The first component accounted for 34.6 percent of the variance, and the second for 16.7 percent. Table 4 shows the loadings of the different items (described by the criteria they were intended to address), that is, the extent to which the different items loaded on (were correlated with) the two components.

The relationship between the items is better shown graphically. Figure 1 plots the items in the factor space, where the two axes indicate the two components. With only nine items we would not expect nine components to emerge; however, it is interesting that the items loading on the first component are those addressing Acceptance Criteria (the Early Involvement item loads negatively, as this was phrased so that "agreement" reflected a negative assessment—the reverse of the other items), and those on the second address Process Criteria (though these also include the Representativeness item, that might be expected to load onto the other scale). Rowe and Frewer (2000) divided their criteria into two in this manner, suggesting that an effective exercise should be both "acceptable" and have "good process," but that an exercise need not score well in both (implying independence). This analysis seems to bear out this distinction. The inter-item reliability of the five items on the "Process" scale here is 0.700 (Cronbach's alpha), and of the four items on the "Acceptance" scale is 0.712. These values indicate a fairly good (though not exceptional) degree of internal consistency. Reliability on the Process scale may be improved slightly (to 0.737) by omitting the Representativeness item (which we would have expected to load onto the other scale anyway), and may be improved slightly on the Acceptance scale (to 0.750) by omitting the Independence item.

A similar process considered the items in Questionnaire 2. Principal Components Analysis revealed a six-component solution (i.e. six components with an eigenvalue over 1.0,

Table 4. The rotated component matrix showing the relationships between the different items in Questionnaire 1

	Component	
	1	2
Influence	.832	.045
Early Involvement	–.770	–.068
Transparency	.761	.150
Independence	.491	.157
Structured Dialogue	.067	.757
Resources (time)	.270	.726
Task Definition	.243	.650
Resources (information)	.313	.648
Representativeness	–.169	.502

Extraction method: Principal Component Analysis.
Rotation method: Varimax with Kaiser Normalization.
For items, see Table 2.

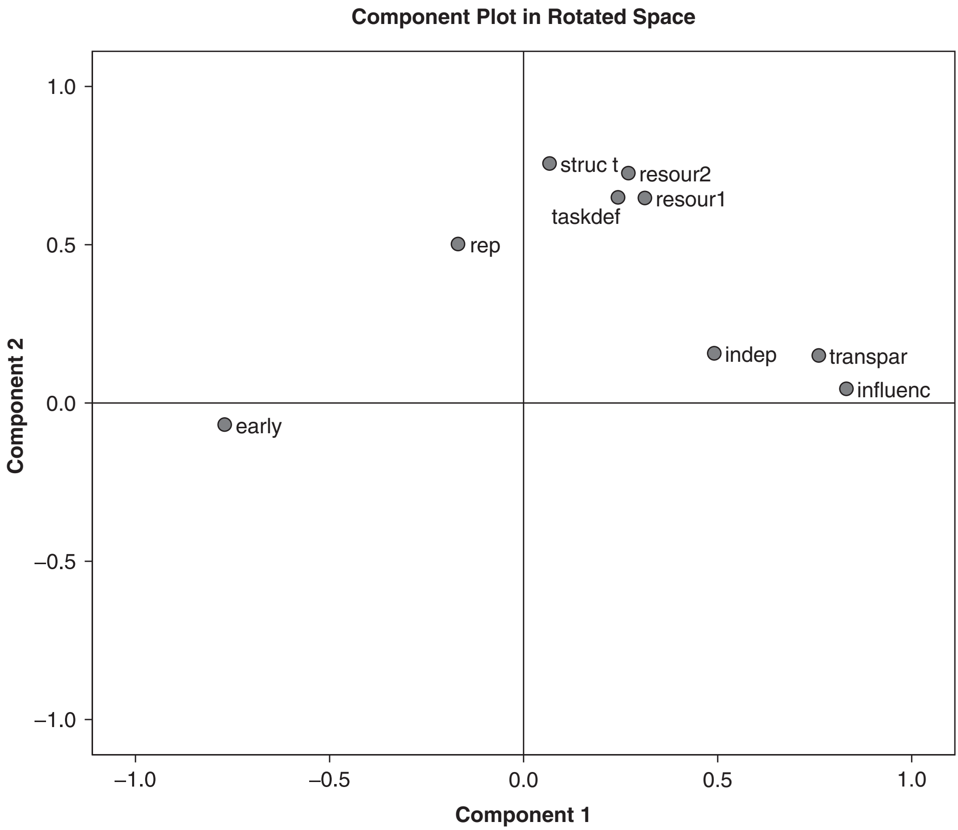

Figure 1. The items from Questionnaire 1 plotted in the factor space.
Note: rep = representativeness; indep = independence; influenc = influence; transpar = transparency; early = early involvement; taskdef = task definition; resour1 = resources (item 1); resour2 = resources (item 2); struct = structured dialogue.

the components accounting for 27.7 percent, 9.7 percent, 8.3 percent, 6.6 percent, 5.7 percent, and 5.4 percent variance respectively, or 63.3 percent of the total variance). Table 5 shows the rotated component matrix, revealing how the different items loaded on the different components. The items in the table are referred to by their shortened title: Table 3 reveals the specific wording of each. In this case, because of the number of components, a graphical representation would be confusing, so is not shown.

The first factor has good internal consistency (Cronbach's alpha = 0.803; not improved by deleting items). This comprises three of the items intended to address the Independence criterion, and three intended to address Structured Dialogue. It appears to address the quality of the facilitators/organizers, capturing *both* their fairness and competence. It might be best labeled "Organiser Behaviour." The second factor also has good internal consistency (Cronbach's alpha = 0.784; not improved by deleting items). This has five associated items, two intended to address the Resources criterion, two intended to address the Structured Dialogue criterion, and one intended to address the Independence criterion. This factor would seem best labeled "Resources," as the three "non-resources" items do in fact consider the issue of whether there was sufficient time (e.g. Structured Dialogue item 2 states: "I didn't get the chance to say all that I wanted to say," and Structured Dialogue item 1 states: "All relevant issues were covered")

Table 5. The rotated component matrix showing the relationships between the different items in Questionnaire 2

	Component					
	1	2	3	4	5	6
Independence 3	**–.801**	–.010	.118	–.022	.210	–.003
St. Dialogue 4	**.707**	.142	–.104	–.234	.076	.238
Independence 2	**–.700**	–.071	.257	–.063	.253	.072
Independence 1	**.641**	.270	–.160	.145	–.096	–.156
St. Dialogue 5	**.641**	.424	–.007	.243	.040	–.071
St. Dialogue 6	**.584**	.427	.012	.082	.109	–.238
Resources 1	–.094	**–.768**	.145	.041	–.017	–.044
St. Dialogue 1	.229	**.724**	–.031	–.171	–.191	.051
Resources 2	.123	**.657**	–.206	.119	–.141	–.052
St. Dialogue 2	–.415	**–.540**	–.011	.067	–.170	–.056
Independence 4	.268	**.442**	–.313	.379	–.123	–.283
Influence 1	–.141	–.061	**.765**	–.081	.091	.099
Transparency 2	–.011	–.233	**.725**	–.162	.138	–.007
E. Involvement 1	–.290	.159	**.656**	.307	–.050	.003
Influence 2	.076	.251	**–.619**	–.235	.241	–.017
St. Dialogue 3	.100	–.079	.070	**.851**	.090	.139
Transparency 1	–.174	–.106	.027	.070	**.831**	–.011
Rep'ness 1	–.001	.385	.092	–.010	–.251	**.678**
Task Definition 1	–.091	–.329	.037	.260	.268	**.643**

Extraction method: Principal Component Analysis.
Rotation method: Varimax with Kaiser Normalization.
For items, see Table 3.

or sufficient/fair information resources (Independence item 4 states: "The information that was given to participants was fair and balanced").

The third factor also has fairly good internal consistency (Cronbach's alpha = 0.714; not improved by deleting items). The four items loading on this suggest it be labeled the "Influence" factor. Two of the items were those intended to address this criterion, while the Early Involvement item essentially speaks to whether influence is likely given the timeliness of the engagement exercise. The fourth item ("Transparency 2") also speaks to the likely use of the meetings' outputs.

Only one item loaded strongly on the fourth factor, and this was "Structured Dialogue 3," which concerned the presence of too much information: we might tentatively label this the "Overload" factor. Similarly, a single item loaded on the fifth factor ("Transparency 1"), which concerned lack of information on how participants were selected. Tentatively, we call this the "Transparency" factor. Finally, two items loaded strongly on the sixth factor—those single items intended to address the criteria of Representativeness and Task Definition. It is unclear what the relationship between these is: one concerns whether participants were typical of those affected by the GM issue, the other whether respondents were confused by what they had to do in the event. We will not suggest a label for this factor.

So, what do these results tell us about what the longer questionnaire measures? First, we must not read too much into this analysis, as several of the criteria we hoped to address only had single items. The questionnaire needs to be expanded so that multiple items address all criteria, giving a greater chance to detect independent factors if these do exist. In the meantime, there appear to be three reliable scales, two of which address expected criteria (Influence and Resources), while a third hints at the idea of organizer Independence, though suggests that competence of performance also needs to be included in a wider criterion definition. Certainly, some of the items in the longer questionnaire, in retrospect, seem to address criteria other than those intended. With these limitations in mind, we can now consider the issue of validity—of the evaluation criteria themselves, and the instruments intended to operationalize these.

Validity of the instruments

The validity of the criteria and instruments may be judged by interrogating other data sources and comparing these with the findings from the structured questionnaires. In particular, the questionnaire responses can be compared with participant responses to the open questions in the long questionnaire, to reveal whether the participants assessed quality using similar or different criteria (and if different, this would suggest the normative criteria are not appropriate, at least from a participant perspective). Participant responses can also be compared to other evidence, including our *observations* of the events (noted earlier), as well as other evidence relating to the objective characteristics of the conferences (e.g. demographic details, which can indicate the extent to which participants truly were or were not "representative," and hence to what extent participant responses validly indicated actuality).

Comparison of closed questionnaire responses to open question responses In order to elicit participants' effectiveness criteria, the longer questionnaires included questions that asked: "what do you personally feel were the positive aspects of the event?" and "what do you feel were the negative aspects of the event?" The respondents' answers were transcribed and themes were identified. A list of positive and negative themes was produced (combining responses from all six events), and then respondents' answers were coded according to these. At this initial stage, all theme identification and coding was conducted by a single researcher (ideally, the data should be tested for inter-/intra-rater reliability). As such, we do not wish to attribute great meaning to the number of responses for each theme—the results are simply indicative of participants' unframed views (note: there were in total over 400 respondents to Questionnaire 2).

Most of the positive aspects related to perceived *outcomes*. The most frequent response/theme (54 instances) was that the event was positive because it provided an *opportunity* for respondents to exchange views/debate the issue. These responses emphasized the two-way nature of interaction as something of benefit in itself. Similarly, a large number of respondents (41) praised the event because it gave them and/or others a chance to air their views and opinions. A number of respondents also simply indicated that the most positive aspect of the debate was that it was happening at all (15), while a few commended the meetings as exercises in empowerment that allowed people to "feel their voices are heard" (6 responses).

The second most frequent theme (52 instances) elaborated a specific objective: these respondents felt the event was positive in providing an opportunity for themselves (and/or others) to *learn more* or be *educated about* the opinions of others. A number of other respondents phrased this learning issue slightly differently, in terms of learning information or facts about the topic itself (19 responses).

Respondents also found the event positive in providing personal comfort, in the sense of allowing them to meet others with similar views. A number suggested the event provided an opportunity to *confirm* their own beliefs, and find solidarity with others of the same view (25 respondents). Often, this was implicit in many responses, but we separately coded responses that simply stated that the result of the event (that people did not want GM) was the most positive thing (46 responses) (i.e. there was implicit satisfaction in announcing this result, which confirmed their views). Again, we separately coded responses that declared the event to be positive for providing an opportunity to personally *meet* other like-minded people (but also people with opposing views)—occasionally phrased as an "opportunity to network" (45 responses). These responses suggest there were many committed participants who wanted to advance their own position on the GM issue rather than learn from others (two respondents baldly stated that they had attended to "show support for the cause").

One positive that emerged from the event, in many respondents' eyes, was that it revealed human virtues. Thirty responses were coded as belonging to a theme related to an enlightening discovery of the intelligence, thoughtfulness, and passion of other participants, while a dozen respondents discussed as positive the perception that people were prepared to make sacrifices (e.g. in terms of time, money, and in spite of logistic difficulties) to attend/participate. A number (7) simply stated that they found the event personally enjoyable/stimulating/interesting.

Finally, a number of respondents praised the events for providing an increased profile of the issue via the media to the wider public (15 responses), while for others the positive was not just the increased profile, but the nature of the *message* that was being sent to the wider world (e.g. to the government), often with a *hope* of influence (20 responses).

Respondents also assessed the events positively with regard to various *process aspects*. Some suggested that the events allowed the collection of (many) diverse views, often implying inclusivity and a non-biased sampling (23 responses). Respondents also noted in various ways that the structure or environment of the events were such as to allow those attending to have their say (16 responses), without any sense of restrictions, such as through dominating individuals (11 responses). Respondents also suggested that there was an absence of pressure (from organizers/others) to conform, describing the events (or small group meetings within them) using terms such as "polite," "civil," "non-confrontational," and "adult" (19 responses). A number of respondents suggested that the events were either *fairly* facilitated (13 responses), or *fairly* run/organized (6 responses). Respondents further suggested that the information available was positive in some way (fair/balanced/accessible/intelligent/full; 13 responses). A few respondents indicated that their event was simply well run (9 respondents), and various specific aspects (sound system, available time, instructions) gained a very few positive comments.

There were considerably more (and more detailed) responses to the question on event negatives than positives. This is probably of no surprise, and should not necessarily be taken as a relative indication of how good or bad the event was: it is arguably easier to identify where things are wrong than right. In terms of indicating where changes ought to be made to an engagement process, however, negative opinions are likely to be more usefully informative than positive ones (and indeed, absence of comment on some aspect might even be taken as tacit approval).

A total of 65 responses were coded into a theme concerned with the *non-representativeness* of participants, with respondents critical that participants did *not* comprise the general public, or that there were too many participants from specific groups and too few from others (e.g. pro-GM groups). Perhaps relatedly, a number (23) criticized the events for being "too small," and not involving enough people to be truly considered part of a "national debate." Some criticized the "lobbying" activities of participants, using the events to distribute material and recruit members (10).

Many criticisms concerned the running of the events. By far the most frequent concern was about event publicity, or lack of this (98 respondents). The venues themselves were also criticized (28) for being difficult to get to, badly signposted, and distant. Respondents were also critical of the timing of the events, both in terms of the time of day in which they were held (when people were still at work—raised by 16 respondents), and in terms of when they occurred in the wider debate (being perceived by 11 respondents as either premature or too late in the process).

Resource issues were a concern to respondents. Many suggested that there was simply not enough time to allow the proper running of the events (55), or time available beforehand to consider and prepare material (17), or criticized the lack of information, generally in the sense of there being no experts to consult about the GM topic (48 respondents), or government or biotechnology representatives (6). The cramped nature of the venues (a problem of resource logistics) was criticized as leading to noise problems, with other sound problems related to the microphone or video (15 respondents in total). A number suggested the events were cheaply run (11)—not even offering participants a free coffee.

Specific process complaints centered on the small group activities. A score of respondents (20) noted that some self-selected groups were unbalanced, having no pro-GM people with whom to hold a proper debate, or were not well run as a consequence of having participant facilitators who were unskilled (or even biased) in their activities (several suggested a need for professional facilitators). The process for concluding the events (summarizing views of break-out groups in plenary) was also criticized: 21 respondents complained that this was over-long, repetitive, without time for all to report, and lacking emphasis on differences. Some suggested that the process in general did not lead to constructive views or consensus building, but rather, reinforced polarities and was divisive (7). Some expressed annoyance that they had essentially been deceived about the status of the event, which was "not a debate" and "really a workshop" (16). The behavior of some participants was criticized as vocal, ignorant, dishonest, and closed (19). The literature was criticized as bland, pro-GM, incomplete and inaccurate (17). And the events were seen as too constrained and fixed, not allowing them to consider some important issues and asking "the wrong questions" (12). Poor organization was noted, in a general sense, by a number of respondents (16).

A final set of concerns worth noting revolve around respondents' views about the rationale for the exercise. A considerable number (41) suggested that the events were a waste of time, without any possibility of *influence*, and some suggested that government decisions have already been made on the topic. A further set of respondents (18) described the events as "window-dressing," a "PR exercise," and just "going through the motions." Other respondents simply expressed uncertainty as to what, if anything, would come from the events (9).

In conclusion, respondents' answers to the open questions reveal their own criteria against which the events were judged. It is probably fair to say that all eight of the normative criteria were identified by at least some respondents (and recall, the open items appeared in the questionnaire before the potentially biasing closed questions), with certain criteria of major concern. For example, the process of the debate—in terms of both the *structured dialogue* and availability of *resources*—was described negatively by large numbers of respondents, and there was also clear concern about the *representativeness* (or lack thereof) of participants, and the unlikely *influence*. In this sense, the participants' responses suggest that the normative criteria used here are relevant, at least from the participants' perspectives. But were the criteria exhaustive, and were they the most important? From the participants' perspective, clearly they were not exhaustive. In particular, the absence of publicity for the events was the respondents' main concern, yet this does not appear as a specific item to be considered in the criteria (though might be linked to the Resources criterion, or Transparency criterion at a stretch). The issues of *learning* and having a *personal voice* (possibly linked to Influence) were also clearly important criteria for success. Indeed, a learning criterion was identified in Rowe et al. (2004) as an important additional criterion not covered by the criteria used here, and it is a criterion that has been identified by other authors (e.g. see Dahl, 1989 and his criterion of a democratic process). Future evaluations might learn from, and incorporate, the additional kinds of criteria identified here.

Comparison of questionnaire responses to other measures Responses to the questionnaires were equivocal about whether respondents thought participants were appropriately "representative" of the wider population (the majority of Questionnaire 1 respondents thought not; while Questionnaire 2 respondents were almost evenly split on this issue). Responses to the open questions reveal that this issue was important to many respondents, who generally regarded it as a *negative* feature of the events. From the demographic items in the questionnaires it was possible to address the Representativeness criterion more directly. As discussed, these indicated that respondents were highly educated compared to the general population (i.e. not representative in this respect). In addition to demographic and socioeconomic characteristics, participant representativeness might also be assessed according to the extent to which the participants' attitudes matched those of the wider public. Pidgeon et al. (2005) compared participant responses to the organizers' feedback questionnaire (which had 13 questions on the risks and benefits of GM foods and crops) to responses to similar questions posed to a nationally representative public sample ($N > 1000$) gained from an additional survey they conducted. (The respondents to the feedback questionnaires (over 36,000) included respondents who attended the Tier 1 conferences.) They found major differences between the two samples—in particular, their national survey sample rated GM foods and crops significantly more *positive* than did the *GM Nation?* sample. Our observers also noted concern being expressed by meeting participants about such issues (e.g. about participants being self-selected and unrepresentative). In other words, evidence from these different data sources seems to suggest that representativeness *was* an important criterion for effectiveness for participants, and one against which the meetings fared relatively poorly, both objectively and in terms of participant perceptions.

We turn now to findings from the observations. Regarding the Independence criterion, observers noted that event facilitators described themselves and their role in slightly different ways in each event. In all cases the relationship they had with the event organizers *was* briefly discussed, and the facilitators explained that they had been employed to act independently and objectively. There was, however, a small degree of unease over the independence of the events expressed by *some* participants, despite the "arm's length from government" characterization of the exercise. Questionnaire responses certainly revealed that a minority were concerned about this aspect, but the *majority* were generally content.

Regarding Transparency, observations suggested a number of inadequacies. In particular, the objectives of the debate were never mentioned, nor was there mention of how the participants could access the final report of the overall debate (when completed), nor was there discussion of how the data would be written up or presented. At a number of the events it was, however, made clear that participants could access the transcripts of the event in the following week on the *GM Nation?* website, the address for which was provided in the introduction. There was no mention of any other form of feedback mechanism aside from a statement about the organizers guaranteeing to relay participants' views to government (although this tended to be received with skepticism). In summary, our observations suggest that "Transparency" was not particularly high, with important information about the nature and purpose of the events unstated. This conclusion corresponds with the rather negative assessments of event transparency revealed through the questionnaires.

Observers noted a variety of resource difficulties (Resources criterion). The style of facilitation was directed at setting the scene for participants, with the facilitator taking no role in the break-out group discussions (i.e. participants were left to their own devices). Participants rarely referenced the booklet or discussion guide information, and if they did engage with this material, it was only in a fleeting fashion. Despite the provision of written information, observers noted numerous occasions where participants raised questions for which relevant material was available in the booklets (but remained undiscovered). Observers also noted many occasions where people expressed a wish for more up-to-date scientific information. Unfortunately participants had no opportunity to examine the material in advance of the meetings (some participants expressed regret at this), and nor was time set aside for this once participants started the group process. Time resource constraints appeared to limit the ability of the round-table discussions to fully talk through the issues, and only five to ten minutes were provided for each table's spokesperson to report back on the previous hour's conversation. These observations largely parallel participant concerns and negative ratings from the open and closed questions in our evaluation questionnaires, respectively.

Regarding Task Definition, observers noted that facilitators gave attendees background information about the GM debate in all events—these being presented as a way to feed public opinion back to government. At some events a member of the Steering Board spoke briefly about the debate. However, a number of participants expressed confusion with the process, particularly why it was organized in the way it was (many seemed to expect an old-style public meeting with presentations from speakers and questions from the floor), and what the government intended to do with the information when a public debate of sorts had been in progress for some years. At most events, facilitators asked the groups to focus on a number of broad topics (usually the risks/benefits of GM and whether GM crops should be commercialized)—though people generally did not stick to these specific areas. No reasons were given to explain this choice of topics, save at one meeting where a facilitator mentioned that these had been derived from 54 questions identified by the Steering Board. In summary, there appears to have been inadequacies in the presentation of the task and its processes: from our observations we would generally rate the events more poorly on the Task Definition criterion than did participants via the questionnaires. We suggest that the Task Definition questionnaire items (asking participants whether they were "confused" about what they had to do) were not subtle enough to address the multiple complexities intended to be addressed by this criterion, and ought to be further developed.

Regarding the Structured Dialogue criterion, our concern was whether the dialogue that took place between participants was conducted in a manner that pre-empted various biases that often arise in group processes (e.g. leading to misrepresentation of participant opinion). Observers reported a number of difficulties, though their conclusions were by no means entirely negative. Whilst there were some groups in which individuals tended to dominate the discussion, participants were generally able to have their say, and conversations were generally good natured and respectful. Observers could detect no clear signs of frustration with the group discussions, and there was little evidence of disruptive behavior. Across all of the groups there was at best tentative articulation of potential benefits of GM crops, though observers suggested that this *did not* reflect the active suppression of such views, but rather the overwhelming predominance of participants who had an "anti" view of GM (see discussion of Representativeness). There were very few suggestions that the spokesperson for each group had *not* captured the opinions and texture of the group discussions. On the negative side, the process of getting each table's spokesperson to provide a discussion summary, whilst under time pressure, resulted in a loss of information subtlety, which served to reduce much of the feedback discussion to rather predictable (and repetitive) "sound-bites." Our overall evaluation is therefore somewhat equivocal: there were some positive features, and some negative, and this was reflected in the responses to the two questionnaires and the open questions.

The criteria of Early Involvement and Influence could not be considered by the event observers. Certainly, Influence might ultimately be assessed objectively, though impacts may be difficult to establish, being temporally distant and possibly intangible (e.g. subtle changes in government stance towards using such mechanisms). Nevertheless, the UK government position that subsequently emerged has been to allow GM crops to be grown commercially after a case-by-case analysis, and this seems to have been largely influenced by legal commitments arising from membership of the European Union. From this perspective, the views of participants do not seem to have had a great or immediate effect on policy—as suspected by a considerable number of the skeptical respondents.

7. Conclusion: merits and limits of using the normative framework

In this paper, one normative framework for the evaluation of engagement exercises has been operationalized in the context of a major UK public engagement process. The focus has been on the *qualities of the evaluation framework* and its measurement *instruments/processes*—not on the event itself (reported elsewhere, e.g. Horlick-Jones et al., 2006). "Quality" here essentially means the validity and completeness of the set of evaluation criteria, and the reliability and validity of the developed measures. It is not possible to conclude that the developed instruments (the two questionnaires) are *formally* reliable and valid, as lack of control over the engagement process severely limited the collection of necessary data. Such limitations are typical in the majority of engagement evaluations, and undoubtedly one reason for the paucity of rigorous past evaluations (e.g. Rowe and Frewer, 2004). However, a more informal analysis does suggest that the questionnaires (within limits) do have a degree of reliability (analysis of the two questionnaires led to largely similar conclusions), and are of acceptable validity, in the sense that other data sources provided corresponding assessments of the effectiveness of the events. The criteria themselves (as opposed to their measures) seem to have some validity, in the sense that unprompted participant evaluations (responses to open questions) tended to identify the same criteria (perhaps using different phrasing), though this latter method also implicated other criteria that ought to be considered.

It is also important to consider the *inadequacies* of the framework and the developed measurement instruments/processes. As already noted, it was decided not to consider the Cost Effectiveness criterion from Rowe and Frewer (2000), as this did not seem something that could sensibly be addressed by surveying participants. Furthermore, this criterion is arguably different from others in the Process Criteria class, and we suggest its nature, and how it ought to be measured, require more thought. Of the other Process Criteria, it also seems that Structured Dialogue (originally, Structured Decision Making) is rather vaguely expressed (see earlier definition) and might be more fully explicated (which would aid the development of appropriate questionnaire items and observation schedules). For example, it might be better phrased as: "the exercise should be so structured as to ensure that all participants have the opportunity to freely express their opinions, and have full opportunity to discuss any areas of disagreement. The process should also ensure that the final summary of group opinion adequately reflects the extent of agreement/disagreement, and conclusions reached." Indeed, it might be that Structured Dialogue needs to be restated as two criteria (one dealing with quality of deliberation, and the other with how information is collated)—or perhaps even more. The development of a longer questionnaire, and use of factor analysis techniques, might aid in establishing the relationship between different facets related to this broad criterion. As well as Structured Dialogue, there are other areas of vagueness in the original formulation of the nine criteria. For example, in the Representativeness criterion, the term "affected public" is not defined, but is rather left to be decided by the evaluator; while the Transparency definition does not make it completely clear whether the "public" alluded to are those inside or outside the engagement process (or both). Definitional fuzziness is probably appropriate in some cases; for others, increased clarity would seem necessary.

While the participant questionnaires did allow some valuable insight into public perceptions of dialogue quality, the Structured Dialogue criterion, and the other process criteria, might best be ascertained through the observation process—at least in part. For example, regarding the Resources criterion, the main problem was not so much the absence of appropriate resources (the perception of many participants), but rather time constraints and poor organization, which effectively undermined the utility of the information that was provided. Likewise, there were some inconsistencies between participant responses to our questions intended to address the Task Definition and Transparency criteria and our observations. This may simply be due to poorly phrased or too-general questions—for example, the lack of subtlety in the Task Definition questions; and the fact that the separate Transparency questions may have actually addressed different independent constructs (indicated by factor analysis). Alternatively, this discrepancy may have been due to the external observers' knowledge of key issues that arguably ought to have been communicated to participants (e.g. how opinions from the events were used) but that were not. The issue here is whether we can expect participants to comment on things that are *not* present or that cannot be seen.

On the other hand, collecting participant views regarding the Acceptance Criteria, related to the idea of fairness, seems the most appropriate assessment method (*perceptions* being paramount). Even here, however, a full evaluation would seem to require interrogation of other evidence. For example, a highly unrepresentative sample might perceive representation as being fair, being unaware that other people hold different views. It is also possible that people might appreciate representational inadequacy, yet deny this for political reasons so as to undermine any potential criticism about this aspect. And skepticism about potential influence may also be misplaced, since subsequent influence might well emerge.

Rowe et al. (2004) additionally asked respondents to indicate their own evaluation criteria in order to "validate" their normative criteria. A similar approach was used here, and likewise found to be useful. Respondents to open questions often raised issues related to the normative criteria, though perhaps phrased differently. However, as in that previous study, respondents also used a *learning* criterion to judge exercise quality (i.e. the exercise was good or bad depending, to a degree, on whether they had learnt anything). It would seem sensible to add this criterion to those in the original framework.

To conclude, we found that the normative framework used to evaluate the engagement events discussed here generally formed a useful and "valid" basis for evaluation, though the identified criteria were not exhaustive, and a number of areas in the evaluation scheme require further thought. We hope that the results from this analysis will be useful for practitioners, helping to inform the design and conduct of better engagement exercises, and that the process described here may help inform the activities of other evaluators. Future research might consider revisions to the current evaluation scheme, and the relevance of the stipulated criteria in other engagement contexts using different methods.

Acknowledgements

This work was supported by the Programme on Understanding Risk funded by a grant from the Leverhulme Trust.

References

Beierle, T.C. and Cayford, J. (2002) *Democracy in Practice: Public Participation in Environmental Decisions*. Washington DC: Resources for the Future.

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.

Clarke, A. (1999) *Evaluation Research: An Introduction to Principles, Methods and Practice*. London: SAGE.
Dahl, R. (1989) *Democracy and its Critics*. New Haven: Yale University Press.
Dryzek, J. (2000) *Deliberative Democracy and Beyond*. Oxford: Oxford University Press.
Halvorsen, K.E. (2001) "Assessing Public Participation Techniques for Comfort, Convenience, Satisfaction and Deliberation," *Environmental Management* 28(2): 179–86.
Horlick-Jones, T., Walls, J., Rowe, G., Pidgeon, N., Poortinga, W. and O'Riordan, T. (2006) "On Evaluating the *GM Nation?* Public Debate about the Commercialisation of Transgenic Crops in Britain," *New Genetics and Society* 25(3): 265–88.
Jasanoff, S. (1990) *The 5th Branch: Science Advisers as Policy-makers*. Cambridge, MA: Harvard University Press.
Jensen, K.K. (2004) "BSE in the UK: Why the Risk Communication Strategy Failed," *Journal of Agricultural and Environmental Ethics* 17(4–5): 405–23.
Kasperson, R.E., Golding, D. and Kasperson, J.X. (1999) "Risk, Trust, and Democratic Theory," in G. Cvetkovich and R.E. Lofstedt (eds) *Social Trust and the Management of Risk*, pp. 22–41. Glasgow: Earthscan Publications.
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.
Public Debate Steering Board (PDSB) (2003) *GM Nation? The Findings of the Public Debate*. London: Department of Trade and Industry. URL: www.gmnation.org.uk
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 and Human Values* 25(1): 3–29.
Rowe, G. and Frewer, L.J. (2004) "Evaluating Public Participation Exercises: A Research Agenda," *Science, Technology and Human Values* 29(4): 512–56.
Rowe, G. and Frewer, L.J. (2005) "A Typology of Public Engagement Mechanisms," *Science, Technology and Human Values* 30(2): 251–90.
Rowe, G., Marsh, R. and Frewer, L.J. (2004) "Evaluation of a Deliberative Conference," *Science, Technology and Human Values* 29(1): 88–121.
Rowe, G., Horlick-Jones, T., Walls, J. and Pidgeon, N. (2005) "Difficulties in Evaluating Public Engagement Initiatives: Reflections on the Evaluation of the UK *GM Nation?* Public Debate about Transgenic Crops," *Public Understanding of Science* 14(4): 331–52.
Shaw, I. (1999) *Qualitative Evaluation*. London: SAGE.
Walls, J., Pidgeon, N., Weyman, A. and Horlick-Jones, T. (2004) "Critical Trust: Understanding Lay Perceptions of Health and Safety Risk Regulation," *Health, Risk and Society* 6: 133–50.
Webler, T. (1995) "'Right' Discourse in Citizen Participation: An Evaluative Yardstick," in O. Renn, T. Webler and P. Wiedemann (eds) *Fairness and Competence in Citizen Participation: Evaluating Models for Environmental Discourse*, pp. 35–86. Dordrecht: Kluwer Academic Publishers.

Authors

Gene Rowe is currently Head of Consumer Science 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 (UWE), 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:** Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK; e-mail: gene.rowe@bbsrc.ac.uk

Tom Horlick-Jones is an independent researcher and consultant, currently based at the School of Social Sciences at Cardiff University. He was team leader of the *GM Nation?* public debate evaluation project. Over a period of some twenty years he has specialized in issues concerned with applied and conceptual aspects of risk, organizations and decision-making. His research is much concerned with the roles of talk, practical reasoning and knowledge in these areas. His

publications include *Natural Risk and Civil Protection* (co-editor; Spon, 1995), *Social Amplification of Risk: the Media and the Public* (co-author; HSE Books, 2001) and *The GM Debate: Risk, Politics and Public Engagement* (with Gene Rowe, John Walls et al.; Routledge, 2007).

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.

Wouter Poortinga (M.Sc. Groningen, 1997; Ph.D. University of East Anglia, 2004) is an environmental psychologist, currently holding an RCUK Academic Fellowship in Health and Risk Communication at the Welsh School of Architecture and the School of Psychology, Cardiff University. His research has mainly focused upon how people respond to environmental and technological risks. Wouter's wider research interests are in studying the environmental and psychological basis of people's health, well-being and quality of life.

Nick Pidgeon is Professor in the School of Psychology at the University of Cardiff (UK), and was 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-methods approaches.