Chapter 3: Evaluation criteria for foresighting methods

3.1 Introduction

This chapter reviews the literature on evaluation of futures studies and related areas, using the literature to explore the effectiveness of methods of scenario planning and qualitative futures research. The initial focus is on the extensive literature of scenario planning, later extending into related areas: other qualitative futures studies methods, and the literatures of corporate and social planning, evaluation, and organizational development. Further findings are drawn from relevant areas of the literatures of social psychology, sociology, and systems theory. From the combined findings, the chapter then develops a set of criteria for evaluating the effectiveness of the type of process being developed in this study. The chapter finds that the most effective scenario planning method will meet a range of criteria of several types: (1) process criteria and (2) outcome criteria. It concludes that the most effective foresighting method will be participative, will include a thoroughly canvassed range of possibilities, will be presented in a visual format, and will embody a theory of the future. The chapter concludes with a discussion of the shortcomings of the evidence on this issue, contradictory findings in the literature, and any outstanding empirical research questions.

3.2 The particular difficulties of evaluating futures studies methods

For social research methods in general it is not conceptually difficult to develop criteria for evaluating the method, but in practice this does not seem to have been done – partly because almost no social research methods appear to have been developed deliberately from the outset. Discussion of the evaluation of research methods has generally been couched in terms of their (a) accuracy, (b) usefulness to their consumers, and (c) ability of their performers to carry them out well. With a normal research method it would be possible to evaluate the method as a whole, even if this has not been done in practice. However in the case of a futures studies method, two new conceptual problems occur: (a) if accuracy of prediction is a criterion, this can be assessed only after the target period for the prediction; (b) self-altering prophecies are likely to apply, in which the foresighting exercise itself causes the eventual outcome to change.
Initially I was hoping to be able to apply a standard technique for evaluating social science methods, particularly qualitative methods, but an extensive literature search failed to reveal the existence of any such standard technique. For social research methods in general, it is not conceptually difficult to develop criteria for evaluating a method, but in practice this does not seem to have been done – partly because (as discussed in chapter 6, section 6.2.1) almost no social research methods appear to have been developed deliberately from the outset.

Discussion of the evaluation of research methods is generally (e.g. Patton, 1990; Bernard, 1994) couched in terms of their (a) accuracy, (b) usefulness to their consumers, and (c) ability of their performers to carry them out well. With a normal research method, it would be possible to evaluate the method as a whole – even though this seems not to have been done in practice. However in the case of a futures studies method, two additional conceptual problems occur:

(1) If realization of the foresighting is one of the criteria, this can be assessed only after the target period for the foresighting. Though a 10-year or 20-year delay is not a problem in principle, it was a problem in this case because of the need to finish this thesis on time.

(2) Self-altering prophecies such as the “Oedipus effect” and the “bandwagon effect” may apply, in which the foresighting exercise itself causes the eventual outcome to change (Popper, 1956; Henshel, 1978).

In this case, to solve problem 1, I considered using hindcasting – that is, predicting the present using only data available in the past (Salant, 1971; Harrald and Mazzuchi, 1993) – but realized that this would preclude the use of stakeholder participation, and would have to rely solely on documented input. Though one of the cases is based on secondary data, there would still exist conceptual problems related to the unidirectionality of hindsight (Einhorn and Hogarth, 1980; Hawkins and Hastie, 1990). My partial solution was to include some short-term effects, of around one to two years, so that these effects could be assessed in the follow-up phase of the fieldwork.

Problem 2, on closer examination, turned out to be soluble, because there is no reason why a thorough foresighting study should not include assessment of its own potential self-altering effects: in other words, scenarios of responses made to another scenario. Also, studies of expectancy effects have found them to be fairly rare, with earlier conclusions of studies such
as the Hawthorne Experiments (Roethlisberger and Dixon, 1939) now being questioned (Draper, 2004).

### 3.2.1 Separating evaluation of the method from evaluation of the execution

The primary purpose of this thesis is to develop a “good” method of foresighting. To demonstrate the goodness of the method, its execution was compared against the criteria developed later in this chapter. However, the criteria alone may not be sufficient, because of the problem of additionality (whether the same result would have occurred without the intervention; see Georghiou 2003, p5). Thus the effectiveness of a particular case study is determined by several factors, including:

1. The quality of the method itself;
2. The quality of the way in which the method was applied by its organizer/s; and
3. The quality of the way in which the processes and outcomes were used by the entity in the case study.

Thus it might be possible that the Process I developed met the criteria inherently, but my execution of it was flawed, perhaps due to my lack of experience in applying this new process. Another possibility would be that an entity whose future the Process helped to anticipate did not regard its execution as effective – perhaps because the entity did not want to face unpleasant possibilities.

As the goal of the present study was to assess the quality of the method itself, it was important to disentangle the above three factors. In view of the findings of Skumanich and Silbernagel (1997) in their interviews with foresighting organizations, and the gloomy prognosis of Hackman and Wageman (1995) concerning the possibilities of evaluating the effectiveness of Total Quality Management, this is far from simple, and possibly cannot be established at an incontrovertible level. The best possible indication may turn out to be preceded by a statement such as “On the balance of evidence, it seems that....”

A further complication is that perhaps the quality of any process may not be relevant to its success. It might be, for example, that any process in which a group of staff of an organization met for several days to discuss the organization’s future would be about equally successful, regardless of how much thought had been devoted to designing the process. The futures literature described below does not address this additionality issue.
3.2.2 Process and outcome criteria

Patton (1990, pp94ff) distinguishes between the evaluation of outcomes and the evaluation of processes. He notes that with some methods of community and organizational development, “what we do is less important than how we do it.” Where, as in these cases, the process is more important than the outcome, it is important to evaluate the process, as it is to some extent an end in itself. But even for interventions that emphasize results and outcomes, a process is undertaken to produce results. To that extent, some process evaluation is always useful. As this criterion certainly applies in the case of futures studies methods, it will be useful to divide the criteria for the Process being developed into process criteria and outcome criteria.

Useful as is the separation of process and outcome criteria, some further division is helpful in the present case. Any management intervention has a number of chronological stages – often divided into five, in the literature of evaluation, specifically program theory and logic modelling (Chen, 1994; Funnell, 1997; Weiss, 1998):

1. inputs
2. process
3. outputs
4. direct outcomes (short term and directly identifiable as effects)
5. broad outcomes (often labelled “impacts”).

The innovation introduced by program theory was that when an intervention is divided into the above stages, each stage can be evaluated separately, using the most appropriate method for that stage, such as financial data (for inputs), or participant evaluations (for processes). The flexibility of this approach is so advantageous that the criteria listed below are classified into these five stages, as follows:

1. Input and planning – e.g. preparation for using the Process with a particular entity;
2. Activity – producing the outputs – e.g. scenario-building;
3. Output – e.g. written scenarios;
4. Direct outcomes – e.g. effects on thinking of participants;
5. Broad outcomes – e.g. non-immediate effects on the entity being studied.

3.3 Evaluation criteria for futures methods

This section reviews the relevant literature seeking criteria for the evaluation of futures studies methods (focusing mainly on scenario planning), but also with an eye to other relevant futures studies methods, as well as evaluation of the varieties of large-group interventions used in
organizational development (Bunker and Alban, 1997; Holman and Devane, 1999), and the evaluation of qualitative research methods in general. Thus this literature review is divided into three increasingly generalized groups:

- evaluation criteria for futures methods in general;
- evaluation criteria for organizational interventions;
- evaluation criteria for qualitative research methods.

Each of these areas is now addressed in turn, with each group listed in chronological order of publication. Three sources were used to find suitable criteria:

1. A review of the methodological literature, covering futures methods thoroughly, and organizational development and qualitative research to a lesser extent (covered in sections 3.3.1 to 3.3.3 below).
2. From Appendix 1: review of a number scenarios for 2000, comparing the scenarios with outcomes. A summary is in section 3.3.4 below.
3. Criteria emerging from chapter 2, not found in the above two categories: in section 3.3.5.

Following that review, a summary is made of the criteria that were most relevant to the current study, and were mentioned most frequently.

3.3.1 Evaluation criteria in the methodological futures literature

A search of the literature of futures studies found no detailed studies focusing in detail on the evaluation of futures studies methods: i.e. nothing as detailed as this chapter. However, that search did find around 40 journal articles and reports that, either explicitly or in passing, described the characteristics of desirable futures studies methods.

Criteria preceded by a ➔ symbol and followed by a reference number in square brackets were included in the database that forms the basis of the following section (3.4). As will become evident, few papers address this issue squarely, but a number of papers mention criteria *en passant*: often only a few words per criterion. Some criteria were expressed so briefly that they were ambiguous and could not be included here.

Hoos (1974), using ➔ veracity and accuracy as possible criteria [1], is skeptical about the possibility of “good” futures research (as she labels it, using quotation marks). She seems to have been referring mainly to Delphi, rather than to scenario planning, little on which had been published by that time.
Zentner (1975, reviewed by Wilson, 1978, p233) names three criteria for testing scenarios (in the eyes of “planners”): credibility [2], utility [3], and intelligibility [4] (“scenarios must be set forth in a manner easy to understand and use”). As Wilson points out “there can scarcely be any quarrel with these criteria.” Wilson (1978) adds a further criterion: relevance to the interests of the organization in question and the issues being addressed [5], organized around “key branching points.” Wilson also discusses the purpose and uses of scenarios, which is considered in section 3.4 below.

Amara (1981), in one of the few papers directly addressing the quality of futures studies work, discusses the objection that validation criteria are nonexistent in futures studies. He initially argues against successful prediction [6] as a criterion, because “the futurist’s purpose is not to predict but to generate images or perceptions and to analyze them so that we can increase the probability of producing futures more to our liking” (Amara, 1981, p66). Instead, he focuses on plausibility, and states that the most useful indicator is whether the forecast meets criteria of proximate validity and internal validity. Proximate validity [7] amounts to approximate prediction early in the period of a forecast (“forecast” being a term which Amara uses in an unusually broad sense), while internal validity [8] is a measure of how well it “hangs together” and its integration of elements. This is puzzling: Amara first dismisses prediction as a criterion, but two paragraphs later he is setting forth “proximate validity” as a validation criterion – but what is proximate validity if not a form of prediction?

Amara (1981) distinguishes three different purposes of futures studies: the possible (“what can be”), the probable (“what may be”), and the preferable (“what should be”). Depending on the purposes of a particular study (though these are often combined in individual studies) he offers three sets of criteria, under three broad headings: conceptual explicitness, analytical clarity, and utilitarian objectives:

**Conceptual explicitness**
- Are the futurist’s premises explicit? [9]
- Are the purposes of the activity explicit? [10]
- Are the futurist’s values explicit? [11]

**Analytical clarity of the product**
- Are the futurist’s methods explicit? [12]
- Does the product include a description of change processes? (i.e. is there a trail that can be followed?) [13]
- Is there time to act? (i.e. is the product really describing the present, not the future?) [14]
Utilitarian objectives

- Is the product clear and specific? [15]
- Is the product credible? [16]
- Does the product change perceptions and guide actions? [17]

Amara (1981) also discusses the criterion of plausibility, which he defines as having three components: 
- “general conformance with basic physical and behavioral principles” [18],
- internal consistency [19], and
- reasonableness [20]. He admits, however, that many innovations when first developed did not meet these criteria. He concludes by matching the criteria with his three previously stated purposes of futures studies, stating that
  - When the goal is the possible, the criterion is plausibility [21].
  - When the goal is the probable, the criterion is reproducibility (analogous to reliability) [22].
  - When the goal is the preferable, the criterion is explicitness of values and likely impacts [23].

Amara’s focus is largely on the “product” or output, not on the process; the emphasis on “scenario learning” did not develop until some ten years later. Thus Amara includes no criteria relating to participation or enlightenment resulting from involvement with the process itself.

Wack (1985a) writes of the uses of scenarios as “meeting the deepest concerns of managers” through
- “the reperception of reality and the discovering of strategic openings [24] that follow the
- breaking of the managers’ assumptions (many of which are so taken for granted that the manager is no longer is aware of them) [25].”

Adelson (1989, p30) focuses on the outputs of futures research, stating that “any result... is likely to matter in one or another of the following ways:”
  - influencing dialogue [26]
  - dispelling misconceptions about the past [27]
  - interpreting the meaning of the present [28]
  - articulating aspirations [29]
  - reframing issues [30]
  - enhancing risk appreciation [31]
  - formulating alternatives [32]
  - multiplying perspectives [33]
  - highlighting residual doubt [34].
Dror (1989), in a paper on policy reasoning for forecasting, presents 22 abstract ideas for policy reasoning for evaluating forecasting and planning. Dror takes very much a macro-historical and large-scale viewpoint, and several of his 22 ideas are relevant to futures inquiry:

- Thinking-in-history serves as a main mode (i.e. the need to pay attention to long-term trends) [35]
- Passing windows of opportunity and various surprise domains and possibilities need intense attention [36]
- Interactions with dynamic and in part responsive environments are carefully considered, with attention to required net advantages [37]
- Critical choices are identified and concentrated on [38]
- Innovation, including iconoclasm, is central [39].

Amara (1991) repeats his 1981 criteria almost verbatim, indicating that after a further 10 years’ work in the futures field, he found no reason to change his position on the criteria for quality in futures work – though he did change his position on some other issues, such as moving away from the “hard” methods such as cross-impact analysis toward a preference for more qualitative approaches.

Nanus (1992), focusing on corporate visioning, writes “the best vision is the one that works across all the scenarios” (Nanus 1992, p108). This is the criterion of robustness [40]. Later (p.121) he lists the properties of a “good vision”:

- Future oriented [41]
- Utopian: likely to lead to a clearly better future [42]
- Appropriate: fitting in with the organization [43]
- Set standards of excellence and reflect high ideals [44]
- Clarify purpose and direction [45]
- Likely to inspire enthusiasm [46]
- Reflect the uniqueness of the organization [47].

The focus of the above criteria is a little different from those reported earlier because Nanus is discussing a single vision (Amara’s “preferable future”) rather than alternative futures.

Bezold (1993, p.1-3) quotes James Robertson (1989) as saying “thinking about the future is only useful and interesting if it affects what we do and how we live today” [48]. Though Bezold notes that this contains the “essence of futurism,” in some senses Robertson’s position is an extreme one. For example, Shell Oil created a scenario of the break-up of the Soviet
Union in the mid-1980s, but did not begin to act on this idea until the break-up began several years later (van der Heijden, 1996). Because the scenario was not used immediately, Robertson’s position would be that this was not “useful and interesting” – but that was certainly not the experience of Shell, which van der Heijden (1996) reported to have made significant gains due to its prescience. Because of this narrowness, Robertson’s criterion is one that I did not use in evaluation of the Process.

Bezold’s own position follows the mainstream of alternative-futures thinking: that the methods work by

- “stimulating the imagination, encouraging creativity, [49]
- identifying threats and opportunities, and [50]
- allowing us to relate possible future choices and consequences to our values” [51]

(Bezold, 1993, p.1-4; my lineation).

Schoemaker (1993) does not explicitly address quality criteria, but notes several times that “scenarios do not aim to predict the future, but rather bound it” [52]. He reports the results of four experiments, which found that the psychological benefit of scenario planning appeared to lie in the use of one set of psychological biases (such as conjunction fallacies) to counteract another (such as overconfidence) [53].

Schoemaker (1995) reported another experiment that used a similar approach; this seems to be the only positivist assessment of the effect of scenario planning. After noting one of the major problems with human judgement identified in the research of Kahneman, Slovic, and Tversky (1982) – specifically the optimism bias – Schoemaker (1995) had some MBA students engage in scenario planning. A post-test found that the students, after engaging in scenario planning, had broadened their numerical estimates of some sales figures, and were less optimistic than they initially had been. Implication: scenario planning is effective in reducing (unrealistic) optimism [54]. One might comment, though, if reducing unrealistic optimism is to be the only purpose of scenario planning, perhaps a course in basic probability might have been more effective still. Schoemaker did not address this issue.

Rhyne (1995, pp539-40), after his work at Patterns and Systems International, “came to feel that a description of a future context should possess all (or anyhow, most) of [these] attributes, if it is to induce similar images of that context in the minds of its users:”

- Scenarios, not snapshots (an unfolding tale, rather than an image) [55]
Wide-ranging sets of scenarios (including both barely the credible and the almost obvious) [56]
Scenarios should reflect patterns of circumstances rather than events [57]
A scenario should be self-consistent at each time point [58]
Sequential consistency, along each scenario line [59]
Scenarios should specific enough to be comparable with reality, as it unfolds (to help determine subsequent trends) [60]
Alternatives should be described even handedly, with no overt preference for one scenario [61].

Rhyne’s emphasis is on the importance of narrative. Note his implicit assumption that the scenarios being discussed are imposed (in Amara’s terminology: the possible, rather than the preferable). Thus the above criteria apply to scenarios of “what the world may be like” rather than scenarios of “how we might change our environment.”

Van der Heijden (1996) lists five objectives for scenario planning
- the development of robust strategies [62]
- better understanding of the future [63]
- better perception of patterns and change [64]
- transmission of management ideas via scenarios [65]
- improved leadership [66].

Elsewhere, van der Heijden (1996, p41) describes another function of scenario planning as to “create a more adaptive organization, which recognizes change and uncertainty and uses it to its advantage” [67]. His emphasis is on the value of scenarios for “strategic conversation” – discussions between managers in which they make use of alternative-futures thinking. For this to occur, the scenarios must possess simplicity [68], evocativeness [69], plausibility [70], relevance [71], and a memorable label [72].

Elkington and Trisoglio (1996; reinforced by Bazerman and Watkins, 2004) discuss the Brent Spar episode, in which Shell Oil, despite its foresighting focus, seriously miscalculated the effect of public opinion on its plans to scuttle an unwanted oil rig in the Atlantic Ocean. Trisoglio analysed the scenarios produced by Shell up to 1995 and found on mapping them to three psychological archetypes derived from cultural theory that all the scenarios fell into the “individualist” or “hierarchist” groups, while the emerging “egalitarian” perspective (in line with the thinking of many environmentalists) was ignored. Elkington and Trisoglio also note that Shell’s scenarios focused strongly on Shell as an organization, and not much on its
broader environment. Though these writers do not explicitly state evaluation criteria, a criterion clearly emerging from the paper is that the conceptual space of scenarios should be broadly cast, in terms of potential stakeholders [73].

Slaughter warns against extrapolative futures: “An issues-based ‘future of...’ approach tends to enlarge or exaggerate aspects of the present world. In many cases an underlying assumption remains that of a basically static frame of reference” (Slaughter, 1996c:148). Thus Slaughter is urging futurists to probe behind the surface issues [74].

Ogilvy (1996) argues that the purpose of futures studies is not successful prediction [75] nor minimizing risk [76], but the articulation of risk “so that we know what is at stake in our daily decisions” [77] (Ogilvy, 1996, p32).

Skumanich and Silbernagel (1997) of the Battelle Institute reviewed foresighting activity around the world for the US government Department of Environment (DOE), which was considering the use of foresighting. Of all the works I found, this is the most detailed treatment of the evaluation of foresighting, though its focus was simply on seeking one best method. Skumanich and Silbernagel, in this review, drew four main conclusions:

- Since the future is essentially unpredictable, the emphasis in foresighting should not be on making accurate predictions but on the “ability to imagine a range of possible futures” [78].
- The process is as important as the outcome, in its ability to aid people to think more flexibly together [79].
- A variety of foresighting methods can be used to achieve a given purpose [80].
- Foresighting activities create impacts in a variety of ways, which are mostly near-impossible to measure; consequently the commonest yardstick of effectiveness is “high level buy-in” and public legitimization [81].

Other relevant points made by Skumanich and Silbernagel are that:

- Foresighting is often not a passive activity, in that it facilitates bringing desired and likely futures into closer alignment [82].
- Royal Dutch/Shell believed that because the future is uncertain, evaluation of the accuracy [83] and even the usefulness [84] of scenarios is not feasible or even helpful. Shell regarded scenarios as successful when they are used at a point when a decision is made [85]. (The apparent paradox between the last two items
might be resolved by regarding usefulness as perceived usefulness at the time the scenarios are developed.)

- Rather than focusing on “foresighting” Shell preferred to emphasize long-term perception, to develop a greater understanding of the present [86].
- In view of the difficulty of measuring the effectiveness of foresighting, organizations studied by Skumanich and Silbernagel tended to use visible metrics as their indicators of success – for example, the number of copies of reports requested [87].

Schoemaker (1998) takes a negative approach, listing 20 common pitfalls in scenario planning, dividing them into pitfalls of process and pitfalls of content:

**Process pitfalls:**
- failing to gain top management support early on [88]
- lack of diverse inputs [89]
- poor balance of line and staff people [90]
- unrealistic goals and expectations [91]
- confusion about roles [92]
- failure to develop a clear road map (for the process itself, not its output) [93]
- developing too many scenarios [94]
- insufficient time for developing scenarios [95]
- failing to link into the planning process [96]
- not tracking the scenarios via signposts [97].

**Content pitfalls:**
- inappropriate time frame and scope [98]
- too limited a range of outcomes [99]
- too much focus on current trends [100]
- lack of diversity of viewpoints [101]
- internal inconsistencies in the scenarios [102]
- insufficient focus on drivers [103]
- not breaking out of the paradigm [104]
- failing to tell a dynamic story [105]
- failure to connect with managerial concerns [106]
- failure to stimulate new strategic options [107].
By reversing these negatives, the 20 pitfalls can be converted into 20 criteria for evaluation. Note Schoemaker’s implicit assumption that the scenarios being developed will focus on the future of a specific business.

Ringland et al (1999), reviewing 20 scenario projects, found three main classes of defects noted by Cohen and Gooch (1991): failure to learn, failure to anticipate, and failure to adapt. This view is confirmed by the findings of Tuchman (1984) and Durschmied (1999). The first and last of the three classes are not directly relevant to the current inquiry, but criteria can be drawn from their recommended solutions to the failures of anticipation:

- Do not focus too much on the present and immediate past [108]
- “Ask the right questions” (to do this, the authors suggest looking for “an empty space we cannot explain”) [109]
- Avoid the dangers of over-dramatization of one viewpoint [110]
- Note the unreliability of experts with vested interests and limited focus [111]
- Allow enough time to do a high-quality analysis [112]
- Beware of untested assumptions about the present [113]
- Note that it can take many years for ideas to become accepted [114]
- Pressures for organizational conformity can overwhelm some scenarios [115].

Greeuw et al (2000, p13) in an assessment of recent European and global scenario studies, distinguished three classes of quality criteria for scenario studies:

- Methodological quality: the quality of the development process and the methods and approaches adopted [116]
- Analytical quality: the quality of the scenarios themselves (e.g. thoroughness) [117]
- Usability: whether the recommendations are concrete, challenging, realistic, and practical [118].

They concluded that the following elements are crucial in scenario planning:

- “Participatory development [119]
- two-way integration of scales, [i.e. include cross-impacts] [120]
- integration of surprises, resulting in more peripheral scenarios, [implying that many scenarios in an ensemble were too similar to one another] [121]
- balanced integration of environmental, social, economic and institutional processes [122]
- integration of various scenario methods [123]
- explicit inclusion of a wide variety of perspectives [124]
> translation of long-term policy recommendations to short-term policy agenda” [125] (Greeuw et al, 2000, p91).

Oughton and Reed (2000) carried out a study in a completely different area which was nevertheless highly relevant to scenario building. In a study of the effect of hypermedia knowledge and learning style on student-centred concept maps about hypermedia, they had students prepare concept maps about hypermedia, then scored those maps for complexity, and found that the students with the most knowledge of the topic also had the most complex maps. How might this relate to scenario planning? The parallel would be to evaluate completed scenarios using similar methods to the concept map scoring. For this, the scenarios might be expressed as concept maps. The criterion would be that successful scenario planning causes those involved to > create more detailed mental maps of the future [126].

Chermack et al (2001, pp23-28) list two types of characteristics of effective scenarios, drawn from the literature: as tools for > “organizational learning” [127] and as > “creating future memory” [128]. They also point out that the evaluation component is “nearly absent from the literature of scenario planning.” The only citation they found that empirically evaluated the effects of scenario planning in any form was Schoemaker (1995), as described above.

Davies et al (2001) reviewed a set of 13 scenario studies with a British focus, in the course of which they created a set of criteria for the best practice in uncertainty management, divided into two groups: sources and analysis:

Sources:
> Explicit criteria for selection and use [129]
> List sources of data used [130]
> Combination of research types [131]
> Review of previous work conducted [132]
> Statement of uncertainty in inputs [133].

Analysis:
> Method should be documented [134]
> Statements of purpose of outputs [135]
> Statement of uncertainty in outputs [136]
> Dimensions and spread of uncertainty illustrated [137]
> Scenarios should be of sufficient breadth [138]
> Tests for self-consistency and coherence [139]
Inclusion of discontinuities and wild-cards [140]  
Equivalent effort across scenarios [141]  
Inclusion of scenario narratives [142].

The above amounts to a checklist of desirable properties for scenario reports. Given the secondary analysis approach they used, the above criteria are (not surprisingly) all related to outputs, not to processes. These writers regarded robustness as highly important:

“We can...explore how different situations might affect postulated plans. Undertaking this robustness analysis is perhaps the most valuable outcome and use of futures studies” [143] (Davies et al, 2001, p3).

They also commented (2001, p5) that the possibility space from which scenarios are drawn is more interesting than the scenarios themselves [144].

Phelps et al (2001) used an impact-related set of criteria for evaluating scenario work, attempting to correlate the use of scenario planning to business performance. They investigated the apparent effect of scenario planning on the financial performance [145] and customer orientation [146] of a sample of British companies: 22 water companies (of which 5 had used scenario planning) and 25 IT consultancies (of which 11 had used scenario planning). For both groups, the finding was a weak positive relationship between financial returns and scenario planning, though in the case of the water industry this may have been at the expense of quality levels – not a sign of long-term thinking.

Rubin (2001, p4) noted that “the scenario method is often seen...as a helping tool in decision-making and strategy work, and the need of its use lies in the present. So the scenario method is used mainly to create instrumental images of the future [147] – knowing about the future per se is not a goal” [148]. She adds: “Some people are interested in what the future will bring and they use scenario methods to explore the future alternatives and possibilities to know more about them [149]. Some other people want to make some specific future, which they value as good and essential... and search for knowledge as how to contribute to make that future happen [150]. And still some others want to know what kind of strategies would be the best and most reliable under different boundary conditions in the future” [151].

Anastasi (2003, p31) noted that “the real value in scenarios lies not in the insight they provide to the core team [152] but in the quality of strategy derived by decision makers” [153].
Harries (2003, p806) suggests that “It is not that one of the handful of scenarios will turn out to have been the correct forecast [154] but rather between them they should reveal the underlying working of and interactions between uncertainties” [155]. Harries also notes that “although the proponents of scenario planning emphasize the importance of not predicting but being prepared for a range of possibilities [156], the fact is that case studies often uphold foresight as an example of successful application of scenario planning” – and cites examples given by Ringland (1998). In other words, despite their disavowal of prediction as a purpose of futures work, leading writers in this field (such as Wack, Ringland, and van der Heijden) justify the work by listing successful predictions – such as Shell’s frequently-cited prediction of the downfall of the Soviet Union.

Cuhls (2003a), discussing the large-scale German foresighting project FUTUR, has no solid conclusions to offer on its evaluation, except that the project will have to be evaluated by outsiders [157] (in this case, non-Germans). She raises a list of key questions:

- “What are the targets that can be evaluated when every participant has his or her own reason to participate?” (This in turn raises the question of the motives of participants, versus the purposes of the organizers of the project.)
- “What is the ‘success’ of foresight? Can it only be measured by the resulting lead visions” (i.e. the recorded output)?
- “Does it also matter if the stakeholders in the system become motivated to act in an intended way? Or if foresight supports the communication processes? Or if cooperation partners find each other?”

Cuhls (2003b) notes that the major outcomes of the FUTUR project were to be “lead visions” (similar to scenarios) which should reflect the demand for research, and result in publicly funded projects. Bearing in mind the above questions, Cuhls then reviewed the desired characteristics of the lead visions, which were to:

- Include precise objectives [158]
- Include a new quality of problem-solving [159]
- Be interdisciplinary and integrate multiple perspectives [160]
- Start from a societal need and build steps to meet that need [161]
- Be understandable by the public [162]
- Have a high economic relevance [163].

These characteristics (as often proved to be the case with the evaluation of specific projects) are a mixture of different levels, and do not necessarily translate into criteria usable in other
In order to achieve the above goals, Cuhls (2003b) then specified seven process-oriented criteria:

- The process should be open and independent of other initiatives [164]
- The process should be result-oriented [165]
- To achieve interdisciplinary lead visions, interdisciplinary teams are needed [166]
- Participation of non-experts [167]
- Combination of different methods [168]
- Thus it is important to ensure sustainability of the process [169]
- Importance of reflexive learning, flexibility, and adaptation to experience [170].

The FUTUR project was summatively evaluated by an international team led by Georghiou (2003) who presented a generalized evaluation framework with three basic foci:

- Accountability: efficiency in use of public funds [171]
- Justification: whether the efforts of foresight justify its continuation [172]
- Learning: how foresight can be done better [173].

Georghiou (2003) introduced the concept of generational evaluation, in which different criteria are applied to a project at different stages of its development. Georghiou was also the only writer in this review who mentioned the problem of additionality, posing these questions for evaluation (2003, p5):

- Would foresight have happened without the policy intervention? [174]
- Is foresight done differently or better because of the intervention? [175]
- Are the resulting actions better because of foresight? [176]
- Have persistent changes been achieved (e.g. foresight culture)? [177]

For the FUTUR project, Georghiou (2003, p7) organized what was largely a process evaluation, focusing on the objectives of FUTUR: the different instruments and methods, with regard to their effectiveness, efficiency, and interplay, and the process in general. This is a macro-scale evaluation model, and the kind of criteria being sought for the present study would all fall into Georghiou’s “effectiveness” group.

In a separate evaluation, of a British foresight program, Georghiou (2003, p9) used an indicator-driven framework, separating process from impact and focusing on the evaluation of impacts on five main stakeholder groups. Indicators were collected to show evidence of participation levels [178], participant satisfaction [179], changes in industry behaviour
[180], and some other less transferable criteria. In contrast to the FUTUR study, this was an outcome-driven evaluation.

Cairns et al (2004, p234) commented that “scenario planning... promotes dissenting opinion, encourages divergent thinking, and does not apply selection and exclusion in search of a causa finalis” [181], later mentioning “maintenance of divergence of opinion, belief, perception within an overall unifying frame of limits of possibility.”

Niewöhner et al (2004) performed a participative scenario development process in Germany, using von Reibnitz’s method with members of the general public and experts on biomedicine. This paper was the only evaluation by participants that I found in the futures literature. Niewöhner et al (2004) had their participants complete a questionnaire on satisfaction with the process [182], using 5-point Likert scales. With this type of scale, according to Anderson and Fornell, 2000) the “normal” result in customer satisfaction research is around 70% of the maximum (i.e. around 3.5 out of 5, where 5 indicates the highest satisfaction level; for the 1999 Bain survey of management tools (Rigby, 2001a) the average was 3.76). In the 2004 Niewöhner study, the only two items of the nine that differed significantly from that “normal” level were “if you stray from the original structure of the scenario method, it will tell in the quality of the output” (which attained below-average agreement) and “Proceeding according to the scenario method can only be seen as a guideline. To get decent results, deviations from the standard procedure have to be tolerated” (above-average agreement). As these two statements are almost mirror images, the combined implication is a plea for flexibility within a process [183]. This may be a reflection on the application of von Reibnitz’s method, which is highly process-oriented (Schwab, Cerutti, and von Reibnitz, 1993; von Reibnitz, 1992).

O’Brien (2004, p720) describes “good scenarios” as having three characteristics:

- They are multidimensional, capturing a broad range of uncertain factors [184]
- They challenge participants’ implicit assumptions about what will not change in their current world and help move their audience beyond it [185]
- Engaging scenario titles and narratives are more likely to capture the readers’ imaginations, and thus influence their understanding of how the future may develop [186].
Postma and Liebl (2005, p5) list four early and three more recent functions of scenario work:

- evaluation and selection of strategies [187]
- integration of various kinds of future-oriented data [188]
- exploration of the future [189], and
- identification of future possibilities” [190].

“More recently, scenarios aim at

- making managers aware of environmental uncertainties [191]
- stretching manages’ mental models [192], and
- triggering and accelerating processes of organizational learning” [193].

### 3.3.2 Evaluation criteria for organizational interventions

Though interventions used for organizational development (OD) do not always have a lot in common with the methods of futures studies, because I was looking toward a participative method, it seemed that the process-related criteria for OD interventions may be relevant for any kind of participative process. Thus the literature search for criteria was extended to include the OD literature.

Considering the large volume of work on methods for organizational intervention – here defining it very broadly, including everything from action research to scenario planning, as well as the more standard OD methods – surprisingly little has been written on the evaluation of this type of work. Typically, the question is ignored, or shrugged off as being too difficult to assess. Nevertheless, organizations spend a substantial amount of money on OD (Rigby, 2001a). In May 2004 I made a query on the online mailing list of the International Association of Facilitators1, asking about methods used to evaluate the effectiveness of facilitated projects. From the 820 subscribers, I received only two responses, neither of which revealed any accepted evaluation method or any relevant literature on this issue. Both replies stated that if

- the intervention agents were invited by the same clients to do subsequent work [194], that was an indicator of success.

Greenwood and Levin (1998, pp77-80) mention the criteria of

- workability [195] and
- internal [196] and
- external credibility [197]. Internal credibility is with clients; external, with readers of reports.

---

1. GRP-FACL, available online at [www.albany.edu/cpr/gf](http://www.albany.edu/cpr/gf)
Reason and Bradbury, in the summary chapter of their handbook of action research (2001b:447ff) discuss the evaluation of action research, employing five criteria:

- quality as relational praxis (i.e. inclusion of wide range of participants) [198]
- quality as reflexive-practical outcome (useful consequences for those involved) [199]
- quality as plurality of knowing (e.g. appropriateness for participants’ needs) [200]
- quality as engaging in significant work (the task was seen as important by all concerned) [201]
- quality as enduring consequences [202].

Because the development of the Process used action research, these criteria are clearly relevant. However, because they are so general, they are (except the last) too broad to be directly applicable as specific criteria. Indicators would need to be derived from them.

Harries (2003, p806) takes a more indirect approach, evaluating scenario-based decision making. She discusses three approaches to such evaluation: case studies, empirical evaluations, and theoretical evaluations – and finds problems with all of these. Case studies are not comprehensive: empirical studies find it near-impossible to take all factors into account; and theoretical evaluations, with their lack of empirical data, are likely to make unwarranted assumptions. Though Harries offers no solution, she does note that “given all the potential mechanisms by which it might be advantageous, scenario planning seems plausibly so.” However, this is a one-sided argument, as she failed to canvass any disadvantages of scenario planning.

I searched the literature of Total Quality Management, because of its emphasis on process, seeking criteria that might be useful in the present situation. Again, I found almost nothing relevant, because this literature focused on its substantive level, rather than the meta-level of assessing the quality of the quality management. One of the few in the latter category was by Hackman and Wageman (1995), who in a broad overview assessed TQM according to three classes of criteria: implementation, process, and outcomes. Pointing out the problems of measuring outcomes (obvious methodological problems which would also apply to foresighting, but which were not mentioned in the foresighting literature discussed above, except by Georghiou, 2003) they concluded somewhat pessimistically that the success of TQM may be “a prediction that can never be definitively confirmed in empirical research” (Hackman and Wageman, 1995, p325). However, that paper was framed in a positivist paradigm, and there may be an alternative way out of this impasse, as discussed below.
3.3.3 Evaluation criteria for qualitative research methods

Because the Process being developed was necessarily qualitative (as explained in Chapter 6), it was useful to study the literature on the evaluation criteria for qualitative research methods, to determine to what extent those criteria were relevant for the current Process.

Though there exists an extensive literature on qualitative evaluation methods (e.g. Denzin and Lincoln, 2000; Guba and Lincoln, 1989; Patton 2002), its coverage of meta-evaluation methods themselves is far from comprehensive. In defence of that situation, Seale (2002) points out that evaluation of evaluation of evaluation (and so on) can become a never-ending spiral. In the literature of qualitative research, chapter 8 in Guba and Lincoln’s 1989 book has perhaps the fullest coverage of meta-evaluation. Though most of the criteria listed there are not directly applicable to the evaluation of futures studies methods, they can generally be adapted to that purpose with a minimal change of wording.

Lincoln and Guba (1985) offered four criteria for assessing the value of research using their method of naturalistic inquiry, which is situated in the constructivist paradigm:

- **Credibility** (parallel to internal validity) focuses on establishing a match between the realities of the researcher and the “members.”
- **Transferability** (parallel to generalizability) depends on the degree of matching of salient conditions, and is established mainly through “thick description.” It is not the senders (original researchers) who determine transferability, but the receivers – those who want to establish the possibility of a transfer.
- **Dependability** (parallel to reliability) relates to the stability of the data over time.
- **Confirmability** (parallel to objectivity) involves establishing that the data is not a subjective creation of the researcher; this can be accomplished by tracking the data back to its origin.

Following the publication of Lincoln and Guba’s 1985 book, a reader pointed out to them that (to quote Seale, 2002, p105):

“their criteria depended on a contradictory philosophical position, since their belief in ‘multiple constructed realities’ (Lincoln and Guba, 1985, p294) rather than a ‘single tangible reality’ (Lincoln and Guba, 1985, p295) which lies at the heart of the constructivist paradigm, is not consistent with the idea that criteria for judging the trustworthiness of an account are possible. Relativism does not sit well with attempts to establish truth, even if the term is placed in inverted commas.”
Further, Lincoln and Guba’s above 1985 criteria, useful as they are, apply only to the “research” – that is, the outputs of the research process. However, as Patton (1990, p.95) notes, the findings of a process cannot be evaluated comprehensively without also evaluating the process itself.

Bearing that reader’s criticism in mind, Guba and Lincoln’s next book (1989) reclassified all the above four criteria under a broad heading of trustworthiness (with parallels to quantitative research) but added another broad heading: authenticity. This is achieved when researchers can demonstrate that they have represented a wide range of viewpoints among “members” (those among whom the research was done). Authenticity has five components:

- **Fairness** (representing a wide range of realities),
- **Ontological authenticity** (well-informed understanding among members),
- **Educative authenticity** (mutual understanding between members),
- **Catalytic authenticity** (stimulating action), and
- **Tactical authenticity** (empowering members to act).

Adapting these two sets of criteria to a scenario evaluation context, they become:

1. **Trustworthiness**
   - **Credibility**: are the scenarios plausible? [203]
   - **Transferability**: will the scenarios apply in other entities or places, in the face of changes in the entity’s environment? [204]
   - **Dependability**: because, like the reliability concept that it parallels, this is time-based, this may refer to the ability of scenarios to transcend their time [205], such as freedom from the “fittingness” noted by Michael (1985).
   - **Confirmability**: is enough data available to audit the way in which the scenarios were produced? [206]

2. **Authenticity**
   - **Fairness**: was a wide range of stakeholders involved? [207]
   - **Ontological authenticity**: was their understanding enlarged by the scenario development? [208]
   - **Educative authenticity**: did they better understand the positions of other stakeholders? [209]
   - **Catalytic authenticity**: were they stimulated to act? [210]
   - **Tactical authenticity**: did they feel empowered to act? [211]
Note that the authenticity criteria relate to the process stage (of the five-stage model described above), while the trustworthiness criteria relate mainly to the output stage – though transferability also relates to the broad outcomes stage.

A general approach to evaluation is to evaluate something by comparing it with a set of standards. Thus Hastie (2001, pp657-8) points out that “the standards used to evaluate the quality of decisions usually involve comparisons between behavior and the prescriptions of rational, normative models, which often take the form of tests for the coherence of expectations, values, and preferences of the achievement of ideal optimal outcomes.” Most of the papers described in this section take this approach: describing a set of standards that a method must meet. With the ensuing emphasis on methodology, they fall into the trap described by Guba and Lincoln (1989): that they are purely methodological criteria, while “outcome, product, and negotiation criteria are equally important in judging a given inquiry.” Specifically, adherence to a method “cannot guarantee that stakeholders benefit” (Guba and Lincoln, 1989, p237) – which statement suggests that, for Guba and Lincoln, stakeholder benefit is the ultimate criterion.

Little (1970) specified six conditions for mathematical models, which, as Daellenbach (2003, p86) points out, can be applied to any symbolic model. Namely, a model should be...

- **Simple** – so that it can be understood by those who may need to use it or act on it
- **Complete** – including all factors that may impinge on the system’s performance
- **Flexible** – easy to manipulate and communicate with; this applies particularly to computerized mathematical models
- **Adaptive** – so that changes to the problem situation do not invalidate the model
- **Appropriate** – for the situation studied, specially in terms of its scale and complexity
- **Productive** – of information that is relevant for decision making.

Little (1970) acknowledged that some of the above characteristics will often be contradictory: thus a model that is complete may not be simple. Most of Little’s criteria may be applied, in a general way, to the output produced by the nascent Process.

This concludes the literature review of criteria that might be used in the evaluation of scenario planning and related methods. Because these criteria were harvested out of context from a
wide range of papers, not all of them are directly comparable, or usable in a different context – so before they can be directly applied to the present Process, they need to be sifted for relevance. Some factors to take into account are that...

- Many of the above criteria focus on evaluating the uses of foresighting methods rather than evaluating the methods themselves: an area beyond the scope of the present study.
- Though it is reasonable to suppose that different purposes might require different evaluation criteria, only Amara (1981) and Guba and Lincoln (1989) explicitly noted this.
- Some of the criteria sets (e.g. those of Cuhls, 2003a and Georghiou, 2003) derived criteria for specific studies, while others (such as Greeuw et al, 2000, and Davies et al, 2001) derived criteria for small groups of studies, without seeking broader application.
- With the exception of Amara’s 1981 paper, none of the above papers had the evaluation of futures studies methods as its primary focus. However the documents by Guba and Lincoln (1989), Skumanich and Silbernagel (1997), Greeuw et al (2000), Davies et al (2001), and Georghiou (2003) all include at least several pages’ discussion of evaluation criteria – perhaps an indicator of depth of thought. For that reason, comments made in those articles and books were given greater weight in the selection of criteria for the current Process.

Despite those reservations, the 218 criteria drawn from the above 42 studies practically exhaust the literature in this area, so they will be used as the primary basis for further development. However, two further groups of criteria were drawn from other sources, as follows.

### 3.3.4 Evaluation criteria from the study of scenarios for 2000

Appendix 1 reports the result of a study of scenarios anticipating the year 2000. On comparing the outcomes with the anticipations, a number of problems with the scenario methods were found, and some lessons for later practice were drawn from these, as follows.
TABLE 3.1 PROBLEMS WITH SCENARIOS FOR 2000, AND POSSIBLE SOLUTIONS

<table>
<thead>
<tr>
<th>Problem</th>
<th>A solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuitive and normative methods were more effective than the Critical Uncertainties method and econometric modelling.</td>
<td>➢ Use intuitive and/or normative methods [219]</td>
</tr>
<tr>
<td>Large-scale scenario exercises (involving more than 5 person-years' work) were no more effective than smaller-scale exercises.</td>
<td>➢ When large-scale resources are available, commission several smaller projects rather than one huge project [220]</td>
</tr>
<tr>
<td>Scenarios produced solely by famous people or experts were less effective than those involving the general public.</td>
<td>➢ Include members of the public as participants [221]</td>
</tr>
<tr>
<td>Technological change is over-predicted, but social change tends to be overlooked.</td>
<td>➢ Consider both technological and social aspects together, and how they might affect each other [222]</td>
</tr>
<tr>
<td>Scenarios were often not distinct from each other.</td>
<td>➢ Use a morphological approach to scenario construction, rather than a dimensional one, because morphological classification is inherently mutually exclusive [223]</td>
</tr>
<tr>
<td>End-state scenarios were difficult to envisage; Lack of detail at human scale</td>
<td>➢ Increase vividness by providing more context and more detailed paths [224]</td>
</tr>
<tr>
<td>Restricted focus and overly narrow boundaries</td>
<td>➢ Begin a project by questioning boundaries, reviewing likely stakeholders [225]</td>
</tr>
</tbody>
</table>

3.3.5 Additional criteria derived from chapter 2

Several other criteria not specifically mentioned by any writer above were added, because a broad interpretation of the futures literature summarized in chapter 2 suggested that these are desirable characteristics of a futures process. These additional criteria are shown in the sequence of the five-stage model.

Stage 1: Input and planning

➢ If the reason for carrying out a futures study is related to a specific problem, the time horizon of the study should extend at least several years beyond the expected occurrence date the problem. This is for two reasons: (1) planned events tend to be delayed, rather than brought forward, and (2) because anticipations of the future can affect the present, any future anticipated at the time the expected problem occurs may affect the handling of that problem [226].
Stage 2: Activity

- A case should be developed over a number of sessions, interspersed by pauses for reflection, data-finding, and reconsideration, as is the practice in action research (Greenwood and Lewin, 1998) – thus not (unlike Mercer, 1997) in a single day [227].
- The Process should be relatively quick and simple to apply [228]. A case should be developed over a period short enough that it is unlikely to be overtaken by events: in the order of weeks, but not (unlike Shell International, 2003) 18 months.
- Since the futures for a case will be deeply grounded in the continuing influence of the past, there is a need to look about as far back in time as the scenarios look ahead [229]. For example, in order to foresee 2000 from 1966, Kahn and Wiener (1967) looked back to 1900.
- With new data constantly coming to hand, it would be wasteful to repeat the entire Process prematurely. Thus an added criterion is that scenarios produced by the Process should be readily updatable as situations changed [230].

Stage 3: Output

- In order to accomplish item [230] above, output should be presented in a format conducive to “drilling down” to different levels of detail, presenting findings in a set of systems, super-systems, and sub-systems (after Koestler, 1967; further discussed in chapter 4) [231].

Stage 5: Broad outcomes

- The Process should result in some action for change [232], similar to Reason and Bradbury’s (2001b) “quality as enduring consequences.” In the absence of any change (though perhaps delayed) in the plans, decisions, or behaviour of participants or their entity, the Process must have failed.
- Some participants should use learning gained during the Process in their relationship with the entity (parallel to the third level of Kirkpatrick, 1994 – described further below) [233].
- Their use of that learning should contribute to the achievement of the entity’s goals (parallel to the fourth level of Kirkpatrick, 1994) [234].

3.4 Grouping and critique of criteria

The criteria found in the literature search were categorized according to (1) the frequency of mentioning, after very similar statements had been combined, (2) the extent of agreement among writers, and (3) the circumstances in which they are applied.
### 3.4.1 Grouping

The following (multi-page) table summarizes all criteria listed above, noting some that are too vague to be usefully applicable. Criteria that were almost identical were combined into separate groups, as shown in the following table, which lists them in descending order of frequency within the five-stage model. Groups are numbered (1.1. to 5.4) for later reference.

**Table 3.2 Frequency of Evaluation Criteria in the Literature**

<table>
<thead>
<tr>
<th>Criteria group</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1. Input and planning</strong> (31 references)</td>
<td></td>
</tr>
<tr>
<td>1.1 Purpose of activity and output should be explicit&lt;br&gt;[9, 10, 11, 12, 23, 93, 129, 133, 134, 135, 158]</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Include a wide range of participants, including non-experts&lt;br&gt;[90, 111, 119, 166, 167, 178, 198, 207, 221]</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Allow enough time, with several sessions [95, 98, 112, 227]</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Process should be quick, simple, efficient, nimble [171, 220, 228]</td>
<td>3</td>
</tr>
<tr>
<td>Other input criteria (each mentioned once only)&lt;br&gt;Gain top management support early on [88]&lt;br&gt;Realistic goals and expectations [91] – vague&lt;br&gt;Review previous foresight work first [132]&lt;br&gt;Process should be open and independent of other initiatives [164]</td>
<td>4 (1 too vague to be usable)</td>
</tr>
<tr>
<td><strong>Stage 2. Activity</strong> (77 references)</td>
<td></td>
</tr>
<tr>
<td>2.1 Challenge participants’ assumptions and focus; reframe perceptions and re-perceive reality [24, 25, 36, 28, 30, 33, 39, 53, 104, 109, 113, 115, 185, 208, 209, 225, 226]</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Include diverse perspectives, broad range of scenarios&lt;br&gt;[56, 73, 101, 124, 138, 144, 149, 160, 184, 189, 190]</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Look toward improvement, be normative [29, 42, 44, 82, 147, 150, 161, 219]</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Integrate various methods and approaches [80, 89, 122, 123, 131, 168, 188, 214]</td>
<td>8</td>
</tr>
<tr>
<td>2.5 Probe behind issues, considering values and drivers [51, 74, 103, 155]</td>
<td>4</td>
</tr>
<tr>
<td>2.6 Participants are satisfied with process [179, 182, 200, 201]</td>
<td>4</td>
</tr>
<tr>
<td>2.7 Do not focus narrowly on the present and short-term [14, 41, 100, 108]</td>
<td>4</td>
</tr>
<tr>
<td>2.8 Include discontinuities, with some peripheral scenarios [77, 121, 140]</td>
<td>3</td>
</tr>
<tr>
<td>2.9 Even-handed treatment of scenarios [61, 110, 141]</td>
<td>3</td>
</tr>
<tr>
<td>2.10 Consider the continuing effect of the past [27, 229]</td>
<td>2</td>
</tr>
<tr>
<td>Other activity criteria (each mentioned once only)&lt;br&gt;Pay attention to long-term trends [33]&lt;br&gt;Interactions with dynamic environments considered [37] – vague&lt;br&gt;Identify and concentrate on critical choices [38]&lt;br&gt;Participants should have no confusion about roles [92]&lt;br&gt;Do not develop too many scenarios [94]&lt;br&gt;High methodological quality [116] – vague&lt;br&gt;Include cross-impacts – consider interactions of variables [120]&lt;br&gt;Explore what strategies would be best under different conditions [151]&lt;br&gt;Process should be result-oriented [165] – vague&lt;br&gt;Process should be flexible [183]&lt;br&gt;Appropriate scale and complexity [217] – vague&lt;br&gt;2.11 Consider how technical and social aspects affect each other [222]&lt;br&gt;2.12 Use morphological approach to scenario construction [223]</td>
<td>13 (4 too vague to be usable)</td>
</tr>
<tr>
<td>Criteria group</td>
<td>Items</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
</tr>
<tr>
<td>Stage 3. Output (48 references)</td>
<td></td>
</tr>
<tr>
<td>3.1 Plausibility, credibility, reasonableness – among both participants and external audiences [2, 7, 16, 18, 20, 21, 54, 70, 81, 196, 197, 203]</td>
<td>12</td>
</tr>
<tr>
<td>3.2 Include narratives and describe change processes [13, 55, 69, 99, 142, 186, 224]</td>
<td>8</td>
</tr>
<tr>
<td>3.3 Anticipations should be trackable and confirmable [22, 52, 60, 97, 205, 206]</td>
<td>6</td>
</tr>
<tr>
<td>3.4 Consistency within and between scenarios [8, 19, 59, 102, 139]</td>
<td>5</td>
</tr>
<tr>
<td>3.5 Output should be readily understandable by all involved [4, 15, 68, 162, 213]</td>
<td>5</td>
</tr>
<tr>
<td>3.6 Futures identified should be relevant to the entity [5, 43, 47]</td>
<td>3</td>
</tr>
<tr>
<td>Other output criteria (each mentioned once only)</td>
<td></td>
</tr>
<tr>
<td>Identify threats and opportunities [50]</td>
<td>9</td>
</tr>
<tr>
<td>Reflect patterns of circumstances rather than events [57]</td>
<td></td>
</tr>
<tr>
<td>Self-consistency at each time point [58]</td>
<td></td>
</tr>
<tr>
<td>Memorable name for each scenario [72]</td>
<td></td>
</tr>
<tr>
<td>Analytical quality of scenarios [117] – vague</td>
<td></td>
</tr>
<tr>
<td>Reports should list sources of data used [130]</td>
<td></td>
</tr>
<tr>
<td>Not to create insight among the core team [152]</td>
<td></td>
</tr>
<tr>
<td>“Include new quality of problem-solving” [159] – vague</td>
<td></td>
</tr>
<tr>
<td>3.7 Present the output in a format conducive to drilling down [231]</td>
<td></td>
</tr>
<tr>
<td>Stage 4. Direct outcomes (37 references)</td>
<td></td>
</tr>
<tr>
<td>4.1 Output should be directly usable [3, 32, 48, 71, 84(-), 96, 106, 118, 125, 163, 199, 212, 218, 233]</td>
<td>14</td>
</tr>
<tr>
<td>(1 disagreed)</td>
<td></td>
</tr>
<tr>
<td>4.2 Create future memory and more detailed mental maps, deeper understanding, and perception of patterns [17, 31, 49, 63, 64, 86, 107, 126, 128, 156, 192]</td>
<td>11</td>
</tr>
<tr>
<td>4.3 Participants feel empowered to act [46, 210, 211]</td>
<td>3</td>
</tr>
<tr>
<td>4.4 Extent of uncertainty is clarified [34, 136, 137]</td>
<td>3</td>
</tr>
<tr>
<td>Other direct outcomes criteria (each mentioned once only)</td>
<td></td>
</tr>
<tr>
<td>Quality of strategy derived [153] – vague</td>
<td>6</td>
</tr>
<tr>
<td>Organize evaluation by outsiders [157]</td>
<td></td>
</tr>
<tr>
<td>Would foresight have happened without this intervention? [174]</td>
<td></td>
</tr>
<tr>
<td>Evaluation and selection of strategies [187] – vague</td>
<td></td>
</tr>
<tr>
<td>Workability [195] – vague</td>
<td></td>
</tr>
<tr>
<td>Can apply output in other situations [204]</td>
<td></td>
</tr>
<tr>
<td>Stage 5. Broad outcomes (41 references)</td>
<td></td>
</tr>
<tr>
<td>5.1 Become more future-oriented, open to divergent thinking, adaptable to change [26, 45, 65(-), 67, 79, 127, 169, 170, 172, 173, 177, 181, 191, 193, 215, 230]</td>
<td>16</td>
</tr>
<tr>
<td>(1 disagreed)</td>
<td></td>
</tr>
<tr>
<td>5.2 Not prediction, or knowing about the future [1(-), 6, 75, 78, 83, 148, 154]</td>
<td>7</td>
</tr>
<tr>
<td>(1 disagreed)</td>
<td></td>
</tr>
<tr>
<td>5.3 Action for change [175, 176, 180, 202, 232, 234]</td>
<td>6</td>
</tr>
<tr>
<td>5.4 Enable robust decisions [40, 63, 76(-), 143, 216]</td>
<td>5</td>
</tr>
<tr>
<td>(1 disagreed)</td>
<td></td>
</tr>
<tr>
<td>Other indicators of outcomes (each mentioned once only)</td>
<td></td>
</tr>
<tr>
<td>Improved leadership [66] – vague</td>
<td>7</td>
</tr>
<tr>
<td>Scenarios successful when used in making decisions [85]</td>
<td></td>
</tr>
<tr>
<td>Usage indicators, eg. copies of reports requested [87]</td>
<td></td>
</tr>
<tr>
<td>It can take many years for ideas to be accepted [114] – vague</td>
<td></td>
</tr>
<tr>
<td>Positive effect on financial performance [145]</td>
<td></td>
</tr>
<tr>
<td>Increased customer satisfaction [146]</td>
<td></td>
</tr>
<tr>
<td>Organizers asked to do subsequent work [194]</td>
<td></td>
</tr>
<tr>
<td>Total references to criteria</td>
<td>234</td>
</tr>
</tbody>
</table>

It is clear from the above table that negligible overt disagreement existed among the writers studied: only 4 of the 28 grouped sets of criteria contained explicit dissent, and each of those
had only a single dissenter. A more serious issue (discussed below) is that two criteria groups in Stage 2 (Look toward improvement and Even-handed treatment) are likely to be contradictory in practice. On re-examining the 28 criteria that were mentioned more than once (in the above grouping), these between them accounted for 195 of the 234 references in the above table: 83 percent of them. This level of agreement provides further confidence in the consistency of the writers in this area.

As well as the 28 grouped sets of criteria, 39 criteria were mentioned only once, and were unable to be sensibly grouped. After discarding 12 that were too vague for their accomplishment to be unambiguously ascertainable, a total of 55 distinct criteria remained. As it was not feasible to evaluate a Process using so many criteria, it was necessary to reduce them to a usable set. This was done by including all 28 criteria mentioned more than once in this chapter’s literature review, and adding the broad supplementary criteria [222, 223, and 231] that were found only once in the literature: a total of 31 criteria. The selected criteria were then divided into those that could be evaluated on the design of a method, and those that could be evaluated only on application of the method. However, before that took place, several criteria needed closer scrutiny.

3.4.2 Critiques of selected criteria

Some of the above criteria that were frequently mentioned (but not explored in detail by other writers mentioned above) warrant further consideration. These are robustness, prediction, consistency, and plausibility. The possible contradiction between normative scenarios and even-handedness is also addressed here.

Critique 1: robustness

This criterion (number 5.4 in Table 3.2), mentioned by five separate writers, was highly regarded. A robust decision is generally defined as one that is likely to be successful in any scenario. For example Davies et al (2001, p3) state that “Undertaking this robustness analysis and examining our sensitivities to future uncertainties is perhaps the most valuable outcome and use of Futures Studies.” However, I contend that robustness is not a characteristic of scenarios themselves, but of the way in which they are applied. Though this thesis develops a Process, it cannot develop a generalized way of applying it, because that will depend on specific circumstances.

2. Interestingly, all disagreement concerned the outcome stages: the results of futures work, not the work itself.
Further, robustness is not a universally accepted criterion in business. Robust decisions may be appropriate in some situations, such as those with high capital requirements and long amortization periods – for example, Shell constructing an oil refinery in an unstable country, as mentioned by van der Heijden (1996). The robust decision in this case is not to build the oil refinery unless profitability can be foreseen in all scenarios. But in a highly competitive service industry, robustness (which is essentially risk-avoidance) can be an over-conservative criterion that may lead to decisions not being taken and to eventual business failure. Robustness was therefore not used as a criterion in evaluating this Process.

Critique 2: prediction

Many writings on scenario methods disavow prediction as a purpose, generally on the ground that prediction is impossible. For example, Skumanich and Silbernagel (1997, p3) state: “A basic philosophical orientation of the programs interviewed is that the future is essentially unpredictable. As a result, the emphasis in these programs is not on making ‘accurate’ predictions of the future but on the ability to imagine a range of possible futures.”

But what, precisely, do these writers mean by prediction? Different writers seem to use this term in different senses, and many writers do not make those senses clear. For example, as noted above, Amara (1981) appears self-contradictory on this issue: he denies that successful prediction is a criterion in futures work, but two paragraphs later recommends “proximate validity,” which amounts to an interim prediction, part-way through the future being studied. Further, of all the writers on criteria listed in section 3.3 above, only Harries (2003) notes the different potentials of prediction of a single scenario and an ensemble of scenarios.

The most comprehensive work in English on prediction appears to be Rescher’s 1998 book, *Predicting the Future*, of which the author states, “As best I can tell, this is the first book on the theory of prediction-in-general since Cicero’s *De Divinatione.*” (Rescher, 1998, p.xiii). Rescher (1998, pp38-43) declares that a prediction must meet these conditions:

1. it is resolvable in principle,
2. it rests on some evidential basis, and
3. there must be a predictor who takes responsibility for the prediction.
4. it states what the future will be, not what it might be – thus multiple scenarios cannot be predictions.

He defines a forecast as a class of prediction: one that is concrete, verifiable, and not probabilistic. Others, however, might disagree. For example, on Rescher’s criteria, a statement that “The probability of rain tomorrow is 80%,” being probabilistic, is not a forecast— a statement that would surprise meteorologists.

Since prediction is a problematic concept, it is useful to further analyse it, beginning with Rescher’s criteria for a forecast: an unconditional statement that

(a) an event will occur
(b) at a certain time
(c) in a certain situation.

Note that the three statements are linked by an implicit Boolean AND: if any one of the three conditions is not met, the entire prediction fails. An example of such a prediction is “it will rain tomorrow in Adelaide.” Thus the event is rain falling, the time is “tomorrow,” and the situation is Adelaide. (Not all situations are place-specific: some involve actors rather than places. 4)

Since one difference between a forecast and a prediction, according to Rescher, is that the former is concrete, it follows that, depending on the precise definition of concreteness, there must be a point at which a statement is not concrete enough to be considered a forecast, and therefore falls into the more general class of a prediction. Consider these three statements:

(a) Between 9:00am on the 13th of May and 9:00am on the 14th of May 2004, at least 0.5 mm of rainfall will be recorded at the meteorological office at Adelaide, South Australia
(b) The rainfall in Adelaide in 2005 will be between 400 mm and 700 mm (bearing in mind that this condition has been met in 68 of the last 100 years).
(c) At least one shower of heavy rain will be recorded in Adelaide in 2005.
(d) Some precipitation will occur in South Australia between 2005 and 2010.

At what point is enough concreteness lost that the statement is downgraded from forecast to prediction? I submit that concreteness is a characteristic of the viewpoint as much as of the view, and that statement (d), though not at all concrete today, would have seemed highly concrete if made in, say, Egypt, a thousand years ago. If it is thus agreed that concreteness is relative, it follows that there is no clear distinction between a forecast and a prediction. Note that in the above sequence of statements, concreteness is progressively lost through implied

4. As “situation” was later found ambiguous, it was subdivided into location and context. See chapter 7, section 7.1.
Boolean OR statements. The sequence makes it clear that the less specific a prediction, the more likely it will turn out to be true, because of the implied ORs separating quantities, times, and places. It thus follows that any prediction can be accurate – as long as it is loose enough, and that it is therefore meaningless to discuss prediction without at the same time considering the degree of specificity.

For example, Popper (1982, chapter 2) attempts to demonstrate that the future cannot be predicted because, even if we had the knowledge, it would take all our time to predict the future in detail. It seems that by “prediction” in this case, Popper is referring to a moment-by-moment prediction, including the details of all surroundings, such as precisely where Popper’s cat decides to sit on his writing pad. However, by the same argument, one cannot describe even the present: because humans can think faster than they can speak, I could never catch up when describing my own thoughts – as Sterne (1759) makes clear in *Tristram Shandy*. However, Popper’s problem with prediction could be sidestepped by selecting a scale of prediction that is most appropriate in a particular situation.

Though Popper (1982) appears to regard prediction as complete precision, to define prediction not as binary but as scalable will help to resolve the problems that writers such as Amara have grappled with. “Scenarios are not prediction,” they say in one breath, but in the next cite “plausibility” as a criterion. And what is plausibility, if not an implicit prediction based on the future being similar to expectations derived from the past? The use of a continuum for prediction resolves Amara’s paradox and allows prediction-disavowing futurists such as Ogilvy (1996) to be consistent with the minority such as Harries (2003) and Bell and Olick (1989), who claim that some degree of prediction is inherent in all scenarios.

Accordingly, a central tenet of this thesis is that prediction is not something that either exists or does not exist: rather, it is inherent in any statement expressed in the future tense, and even in some statements expressed in the present tense. For example the statement “I am going to the city tomorrow” is (in English) synonymous with “I shall go to the city tomorrow.” Even to say “I travel to the city via Main North Road” implies a habit that will continue into the future. I thus propose that a valid criterion for assessing scenarios is that a loose level of prediction applies. It is not necessary that any one scenario in an ensemble should turn out to be a “correct” prediction, but rather that at least one scenario in the ensemble should be a prediction that turns out to be closer to the final outcome than to the time when the scenarios were developed. However, such a comparison will make sense only in a specific situation. That is because, even though it is possible to conceptually place predictions on a scale (much
as McBurney and Parsons (2002) did for the concept of plausibility) in practice the elements of prediction (event, time, and situation) can interact in so many ways that predictions such as “2010 will be more similar to 2000 than to 2020” are meaningless without tighter specification of the context. To demonstrate with a thought experiment (as in Horowitz and Massey, 1991), subtract 100 years from each date in that last prediction, and consider World War I. To a European in 1920, that prediction would seem true, because 1920 was very different from 1900 and 1910. However, to people in countries where the 1910-1920 decade saw less change than 1900-1910, the opposite would have applied.

The outcome of the above discussion is that, despite six writers decrying prediction as a criterion for evaluating futures methods, I contend that a broad degree of prediction is a useful criterion, but to avoid confusion, the term “prediction” in this sense is replaced by “anticipation.” Broadly correct prediction is referred to as “foresightfulness.”

Critique 3: consistency

Four writers reviewed above (Amara, Davies et al, Rhyne, and Schoemaker) mentioned consistency as a criterion of desirable futures methods (criterion 3.4 in Table 3.2). They refer to internal consistency: that a scenario should not include incompatible elements. (However, consistency between scenarios is regarded as undesirable.) For Amara, consistency is a component of plausibility, in that a scenario cannot be plausible if elements of it appear inconsistent. Though clearly there are degrees of inconsistency, Liebl (2002, p175) notes that “it is striking how often situations occur that were dismissed as ‘logically impossible’ or ‘inconsistent’ during the process of scenario building.”

Concurring with Amara that consistency is related to plausibility, I contend that any scenario that is too consistent is ipso facto implausible. The world, as we experience it, is not fully consistent. Humans resolve apparent inconsistency using two powerful defence mechanisms: attribution in hindsight, and the construction of intervening variables such as “God’s will.” If the world is not consistent, why should a scenario be so? In fact, the deliberate inclusion of inconsistencies in scenarios may well provoke a deeper level of analysis, in which inconsistencies are taken as paradoxes to be resolved. Thus consistency of output was not used as a criterion for evaluating this Process.

Critique 4: plausibility

In the above literature review, plausibility was one of the most commonly named characteristics of good scenarios, particularly by writers in the “scenario learning” group – the Shell
school, whose scenario work is done for specific firms. By plausibility, they mean that managers should accept scenarios as being possible outcomes. Among all the works consulted, only Schultz (2002) argues against plausibility, citing what has become known as Dator’s Second Law of the Future that “any useful statement about the future should seem ridiculous” – in other words, implausible.

However, none of the papers and books reviewed above examined plausibility in detail. Because of the frequency with which plausibility is mentioned, and its importance to the argument, it will be worthwhile to analyse it further. I argue here that plausibility is (1) a continuum, not a binary characteristic, and (2) is essentially subjective.

McBurney and Parsons (2002) addressed the concept of plausibility, placing it on a 5-point ordinal scale of certainty as to the truth of a claim, thus:

0. open: no arguments in favour of the claim have been presented;
1. supported: an argument in favour has been presented;
2. plausible: a consistent argument has been presented;
3. probable: a consistent argument has been presented, and no undercuts or rebuttals have occurred;
4. accepted: a consistent argument has been presented, and every counter-argument to it has been attacked.

The above scale can be regarded as the positive half of a symmetrical scale: the other half of which would be (moving back beyond Open)... 

-1. unsupported: a counter-argument has been presented;
-2. implausible: a consistent counter-argument has been presented;
-3. improbable: a consistent counter-argument has been presented, and no undercuts or rebuttals to it have occurred;
-4. rejected: a consistent counter-argument has been presented, and every counter-argument to that has been attacked.

Though there is no intrinsic reason why an Open...Accepted scale is better or worse than a Rejected...Open...Accepted scale, the latter has the advantage of being alignable with a probability estimate, where Rejected corresponds to 0% probability of truth, and Accepted corresponds to 100%. In these terms, the concept of plausibility can be defined as a minimum probability threshold: for example, we might choose to define “plausible” as “subjective probability greater than 10%.” The actual figure may be arguable, and it may vary in different
settings, but what I am trying to establish is that it is reasonable to regard plausibility as a range on a scale: somewhere between a minimum threshold and almost total certainty. None of the documents covered in Table 3.2 mentioned the possibility that plausibility might be a continuum rather than a logical variable – that there might exist degrees of plausibility. However, I found support for this position in the writings of Peirce, who observes that plausibility ranges “from a mere expression of it in the interrogative mood, as a question meriting attention and reply, up through all appraisals of Plausibility, to uncontrollable inclination to believe.” (Peirce, 1908, volume 6, p469, cited in Rescher, 1995, p311).

The second point (which Peirce does not address in that discussion of hypothesizing) is that plausibility is not a property of a scenario, but a property of an individual’s view of that scenario. What is plausible to one person may be implausible to another. Different people may have different thresholds of belief, depending on their knowledge of the mechanisms involved, and the strength of their attitude toward the scenario. What seems like implausible magic to the uninformed may be simple technology to the better-educated: an example is the introduction of mirrors to a tribe in Papua New Guinea (Carpenter, 1972, p112).

If plausibility is subjective, it can change. Thus a scenario which is not plausible at the time it is created can become extremely plausible in hindsight. Therefore perhaps a suitable criterion for plausibility is that it should not be evaluated too early after the fact: at least a substantial part of the way into the period to which the scenarios apply.

A danger with scenario-building is that if plausibility is a major criterion, the ensuing scenarios will not extend participants’ thinking about the future – which is another important criterion for scenarios. For example, van Notten, Sleegers, and van Asselt (2005) found that participants in scenario-building did not view scenarios as plausible when they included negative outcomes for the organization being studied (though it seems possible in this case that plausibility and possibility were not clearly distinguished). Also, setting too high a value on plausibility can cause scenario-builders to not explore trends that in retrospect become obvious. This is demonstrated both in Appendix 1, which finds that scenarios for 2000 predicted their own times better than they predicted 2000, and in the legal services case, described in Chapter 8, in which, only a few months after completion of the study, a radical and unexpected change occurred – one that I would have regarded as highly implausible, had it occurred to me beforehand.
As noted in Chapter 2, a common problem with scenario planning is that it tends to contain uncertainty rather than confront it, and that there is a need for futures studies methods that can anticipate surprise. However, plausibility is almost the opposite of surprise. The only writer I found who mentioned this tension was Hirschhorn (1980), who notes that “a good scenario is one that optimally combines the two [viz. plausibility and surprise]. Such a scenario... produces surprises by developing sequential arguments that are inherently plausible.”

Bearing in mind the “scenario learning” function, one might argue that a key purpose of scenario building is to change participants’ views about what is plausible: in other words, to align plausibility more with hindsight. In view of the “failures to anticipate” described by Cohen and Gooch (1991), and the predominance of Type I over Type II errors in anticipation (as demonstrated in Appendix 1), participants’ views of plausibility will generally need to be enlarged rather than restricted – particularly in the case of social change.

Another aspect of plausibility is timing. A projection that is implausible for tomorrow (holidays on the moon, perhaps) may become quite plausible when the horizon is extended. In the case of technology, where Type II errors predominate, innovations are seen as too plausible – not so much that they cannot occur, but unrealistically likely to happen in the time-frame envisaged. It may be possible to overcome this problem by using a method such as technology roadmapping (Phaal, Farrukh, and Probert, 2004): listing a chain of events necessary for a particular outcome to occur, and estimating time delays for each element of that chain.

Thus the criterion adopted for plausibility in this thesis is not the simplistic notion that “scenarios should be plausible” but Hirschhorn’s (1980) suggestion that scenarios should attain an optimal level of plausibility: neither too high nor too low. This is supported by the findings of Kahneman, Slovic, and Tversky (1982), whose studies of subjective probability found that people overestimate the probability of both highly likely and highly unlikely events. Criterion 3.1 is thus revised: that plausibility should converge during the process. Events and states that initially seemed totally implausible should become more plausible (in a specified time horizon), while events and states that initially seemed certain or imminent should become less plausible. This would foster appreciation of a wide range of possible futures.

Critique 5: normative scenarios versus even-handedness

Though two criteria groups in Table 3.2 (2.3 Look toward improvement and 2.9 Even-handed treatment of scenarios) may not at first appear contradictory, in practice this is likely to occur. Some writers on futures (such as Ogilvy, 1996, and Slaughter, 2004) believe that an important
function of futures work is not simply to outline possible futures, but to work towards a “better” future – however defined. Other writers (such as Simpson, 1992; Rhyne, 1995; Ringland et al, 1999; Davies et al, 2001) point out that to include a preferred scenario in an ensemble is likely to cause others to be neglected – hence their advocacy of even-handedness in the treatment of scenarios. Since this Process was designed to be one of inquiry rather than of planning, the normative criterion was refocused as “Participants have a clearer perception of their desired futures.”

### 3.4.3 Outcomes of critiques

The following table summarizes the outcomes of the above five critiques.

<table>
<thead>
<tr>
<th>Tentative criteria</th>
<th>Outcome of critique</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4 Robustness</td>
<td>Not used as a criterion</td>
</tr>
<tr>
<td>5.2 Prediction</td>
<td>Rather than “no prediction”, the criterion used is “anticipation of broad situation” or “foresightfulness”</td>
</tr>
<tr>
<td>3.4 Consistency</td>
<td>Not used as a criterion</td>
</tr>
<tr>
<td>3.1 Plausibility</td>
<td>Convergent, rather than maximal. Thus criterion 3.1 in Table 3.2 was combined with criterion 4.4 (“Explore possibilities; clarify extent of uncertainty”) as “Boundaries of uncertainty and plausibility are clarified.”</td>
</tr>
<tr>
<td>2.3, 2.9 Normativity vs even-handedness</td>
<td>Refocused, toward clarification of desires rather than scenarios that would fulfil them</td>
</tr>
</tbody>
</table>

Thus two of the criteria from Table 3.2 were dropped, and four others were modified.

### 3.5 The criteria selected

The next step in the process was to divide the selected criteria into design-related criteria (verifiable in the design of a process) and executional criteria (the achievement of which depends on the execution of that process in a particular environment).

Having settled on 29 criteria to be used (the 31 selected from Table 3.2, with changes detailed in Table 3.3), the next stage was to separate criteria that could be designed into the Process from those that needed to be assessed on execution. The following table lists the final criteria, noting the relevant category for each. Four criteria fell into both categories. Where possible, these criteria were incorporated in the design of the Process, but some also needed to be assessed following execution. Though the numbering (at left) matches that in Table 3.2, the
wording of some criteria has been sharpened, to enable accomplishment to be more readily assessed. The table shows the point at which each criterion should be applied:

- Design – a design characteristic of the Process itself, to be assessed at the design stage of the process. The criteria are numbered, beginning with D, for comparison with the Process design at the end of chapter 5.

- The other possibilities are characteristics of the execution of the Process, best assessed either during the execution of a case (generally toward the end of the casework) or on follow-up, some time after the casework is completed, the latter denoted by an asterisk. These criteria are numbered beginning with E, for comparison with the operation of the Process following the case studies, in chapter 9.

- Three criteria need to be assessed both at design stage and execution stage. These are numbered beginning with DE, and their accomplishment is assessed both at the end of chapter 5 and in chapter 8.
TABLE 3.4 ASSESSMENT FRAMEWORK FOR CRITERIA

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Design</th>
<th>Execution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1. Input and planning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>The purpose of the futures work is made explicit to all involved</td>
<td>E1</td>
</tr>
<tr>
<td>1.2</td>
<td>Workshop participants include all major stakeholder groups, covering all likely impinging systems</td>
<td>DE1</td>
</tr>
<tr>
<td>1.3</td>
<td>Enable a reflective process but with efficient use of time</td>
<td>D1</td>
</tr>
<tr>
<td>1.4</td>
<td>Process should be quick, efficient, and nimble</td>
<td>D2</td>
</tr>
<tr>
<td><strong>Stage 2. Activity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Participants’ initial assumptions are challenged, focus broadened, and their perceptions reframed</td>
<td>E2</td>
</tr>
<tr>
<td>2.2</td>
<td>Participants gain more detailed perceptions of future possibilities</td>
<td>DE2</td>
</tr>
<tr>
<td>2.3</td>
<td>Integrate a wide range of methods, approaches, and data types</td>
<td>D3</td>
</tr>
<tr>
<td>2.4</td>
<td>Include wide diversity of viewpoints and range of scenarios</td>
<td>D4</td>
</tr>
<tr>
<td>2.5</td>
<td>The boundaries of uncertainty and plausibility are clarified</td>
<td>E3</td>
</tr>
<tr>
<td>2.6</td>
<td>Probe behind the issues, focusing on underlying drivers</td>
<td>D5</td>
</tr>
<tr>
<td>2.7</td>
<td>Participants satisfied with process, feel activity was worthwhile</td>
<td>E4</td>
</tr>
<tr>
<td>2.8</td>
<td>Extend the focus beyond the short-term future</td>
<td>D6</td>
</tr>
<tr>
<td>2.9</td>
<td>Include a means of anticipating discontinuities</td>
<td>D7</td>
</tr>
<tr>
<td>2.10</td>
<td>Each possibility is explored with equal attention, not neglecting any that seem awkward or inconvenient</td>
<td>E5</td>
</tr>
<tr>
<td>2.11</td>
<td>Look back into the past to see the future emerging</td>
<td>D8</td>
</tr>
<tr>
<td>2.12</td>
<td>Consider how technological and social aspects may interact</td>
<td>D9</td>
</tr>
<tr>
<td>2.13</td>
<td>Use a morphological approach to ensure comprehensiveness</td>
<td>D10</td>
</tr>
<tr>
<td><strong>Stage 3. Output</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Boundaries of uncertainty and plausibility are clarified</td>
<td>E3</td>
</tr>
<tr>
<td>3.2</td>
<td>Include narratives to describe change processes</td>
<td>D11</td>
</tr>
<tr>
<td>3.4</td>
<td>Output should be readily understandable by all concerned</td>
<td>D12</td>
</tr>
<tr>
<td>3.5</td>
<td>Anticipations are expressed specifically enough that they can be tracked and confirmed</td>
<td>DE3</td>
</tr>
<tr>
<td>3.6</td>
<td>Focus on the situation of the entity in its changing environment</td>
<td>D13</td>
</tr>
<tr>
<td>3.7</td>
<td>Output in format conducive to re-analysis and expansion of detail as needed</td>
<td>D14</td>
</tr>
<tr>
<td><strong>Stage 4. Direct outcomes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Output is directly usable by the entity</td>
<td>E6</td>
</tr>
<tr>
<td>4.2</td>
<td>Participants gain more detailed perceptions of future possibilities, creating “future memory,” to help prepare for later action</td>
<td>DE2*</td>
</tr>
<tr>
<td>4.3</td>
<td>Participants feel empowered and stimulated to act</td>
<td>E7</td>
</tr>
<tr>
<td><strong>Stage 5. Broad outcomes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>The entity becomes more future-oriented as a result of the Process, open to divergent thinking, and adaptable to change</td>
<td>E8</td>
</tr>
<tr>
<td>5.2</td>
<td>The broad future situation is successfully anticipated</td>
<td>E9</td>
</tr>
<tr>
<td>5.3</td>
<td>The Process results in action for change: in the entity or in participants’ behaviour</td>
<td>E10</td>
</tr>
</tbody>
</table>

* Outcome criterion 4.2 proved so similar to activity criterion 2.2 that these were combined.

In other words, the design of a new futures Process should fulfil the design criteria in the above table, while the execution of the Process should fulfil the execution criteria, measured during and/or after the casework.
3.6 **Ongoing development of criteria during the inquiry process**

In research that uses the positivist paradigm, when a set of outcomes was being evaluated against a set of criteria, the following standard sequence would be used:

1. All the criteria would be set out.
2. All the data would be collected (often through a survey).
3. Finally, the outcomes would be evaluated against the original criteria.

Though the above sequence seems logical enough, for a linked series of studies (such as a set of experiments seeking to perfect a method), it may be more appropriate to modify the criteria between studies. Such an approach amounts to action research, as originally envisaged by Lewin (1946). However, regardless of whether the research approach is qualitative or quantitative, failing to update the criteria between waves would entail losing much of the value of the iterative process, as Greenhalgh and Taylor (1997) point out.

In this study, following the standard action research paradigm of multiple cycles interspersed by reflection, I began with a somewhat different set of criteria from that shown above, but after working through several case studies and reading additional literature, modified the criteria in the light of the findings and the literature. This meant that the final criteria were different from the initial criteria for which data were collected. Fortunately, I had anticipated such changes by designing the evaluation questionnaires to be largely open-ended, allowing for later recoding. New criteria were added, some were changed in emphasis, some were made more specific, and some were removed. This explains the minor changes in wording between tables 3.2 and 3.4. With the benefit of hindsight, after the reflection on the last case, I returned to the data, and compared each case against the final criteria (as listed above in Table 3.4). The results are reported in chapter 9.

3.7 **Review of this chapter**

This chapter has scoured the futures literature, and also searched the related literatures of organizational development and qualitative research methods for criteria by which a futures studies method might be evaluated. More than 230 criteria were examined, and placed into a five-stage model commonly used in program logic modelling. Some criteria were combined, some were redefined, and several were rejected as irrelevant to the current purpose. The criteria that survived this process were grouped (with a little overlapping) into 18 design
The design criteria are used in chapter 5 to help develop the basis of the Process. The execution criteria were assessed in the two data-collection exercises (survey and group discussion) carried out for each case: one during and towards the end of the Process activity, and the other during follow-up around a year later. Findings from the surveys and group discussions in each case (covered in chapter 8) were used to improve the process in next case studied; these are discussed in chapter 9.

### 3.7.1 Limitations of the evidence presented on this issue

Though the approach taken in gathering criteria was as rigorous as was feasible, there are some likely shortcomings:

**Shortcoming 1.** It is probable that I failed to find some of the literature on the evaluation of futures methods. However, it is unlikely that a highly important document was omitted, because

(a) if it were an older one, a reference would have been listed in comprehensive works such as Skumanich and Silbernagel (1997), Greeuw et al (2000), and Davies et al (2001).

(b) If it were more recent than the late 1990s, and published in a major futures journal, an evaluation journal, or a widely-reviewed book, I should have seen it while keeping abreast of new publications in those fields.

Another limitation is that my search was largely limited to the literature in English – or with English abstracts, as on the OECD Future Trends database (OECD, 2001). There may be important documents in other languages (most likely French, German, Italian, Russian, and Japanese) that I did not find. It is less likely that there are other documents in Dutch, Finnish, and other Scandinavian languages, despite the high volume of futures work in these countries, because major work from there is often published in English.

**Shortcoming 2.** This eclectic approach could be criticized for not sufficiently taking into account the subtle interactions between criteria. However, none of the 50 documents examined covering the evaluation of futures methods made reference to such interaction. The closest to this was the approach of Amara (1981), who delineated different criteria depending on the purpose of a study, whether it was considering probable, possible, or preferable.
futures. By “probable futures” Amara was referring mainly to forecasting and quantitative methods, which (as explained in Chapter 2) are outside the scope of the present study.

**Shortcoming 3.** A further criticism is that only 31 criteria were used, and the others (each found once in the literature) were not incorporated. However, most of the others were either “motherhood” statements (such as “high methodological quality”), or would normally be incorporated as a matter of course (such as “gain top management support”), or embodied specific assumptions which might not apply to a new Process (such as memorable names for scenarios), or referred to a subsequent planning stage rather than the inquiry stage on which the Process was to focus.

**Shortcoming 4.** Six of the seven writers who mentioned prediction disagreed with its use as a criterion, stating that prediction is not a purpose of futures studies. However, by reconceptualizing prediction, from a binary attribute to a continuum of anticipation, I have tried to demonstrate that futures are not totally unpredictable, and that if an ensemble of scenarios cannot produce even one scenario that turns out closer to its anticipated future than to the circumstances of its creation, that ensemble will not have been very useful.

**Shortcoming 5.** Another controversial aspect may be the rejection of plausibility and credibility as criteria. However, this is because the literature review in chapter 2 found evidence that one useful role of a futures exercise would be to optimize (rather than to maximize) plausibility: not to make any future seem either too plausible or too implausible.

**Shortcoming 6.** A conceptual problem in establishing the achievement of these criteria in practice is the difficulty of disentangling effects due to the Process from (a) effects due to its execution, (b) effects due to the use that the entity makes of the findings, and (c) whether those effects would have happened in any case – or as a result of the use of any method which encouraged people to think about their future in detail. The solutions applied are covered in chapter 6; briefly, an abductive and formative approach was used.