European Journal of Operational Research 229 (2013) 143–154

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Decision Support

Towards a new framework for evaluating systemic problem structuring methods

Gerald Midgley [a,b,c,d,e,*], Robert Y. Cavana [b], John Brocklesby [b], Jeff L. Foote [f,g], David R.R. Wood [f], Annabel Ahuriri-Driscoll [g]

[a] Centre for Systems Studies, Business School, University of Hull, Hull HU6 7RX, United Kingdom
[b] Victoria Business School, Victoria University of Wellington, PO Box 600, Wellington, New Zealand
[c] School of Innovation, Design and Engineering, Mälardalen University, Sweden
[d] School of Political and Social Sciences, University of Canterbury, New Zealand
[e] School of Agriculture and Food Sciences, University of Queensland, Australia
[f] Institute of Environmental Science and Research Ltd., PO Box 29-181, Christchurch, New Zealand
[g] School of Health Sciences, University of Canterbury, Christchurch, Private Bag 4800, New Zealand

ARTICLE INFO

Article history:
Received 1 December 2011
Accepted 28 January 2013
Available online 9 February 2013

Keywords:
Problem structuring methods
Soft operational research
Evaluation of methods
Participative methods
Systems methodology
Systems thinking

ABSTRACT

Operational researchers and social scientists often make significant claims for the value of systemic problem structuring and other participative methods. However, when they present evidence to support these claims, it is usually based on single case studies of intervention. There have been very few attempts at evaluating across methods and across interventions undertaken by different people. This is because, in any local intervention, contextual factors, the skills of the researcher and the purposes being pursued by stakeholders affect the perceived success or failure of a method. The use of standard criteria for comparing methods is therefore made problematic by the need to consider what is unique in each intervention. So, is it possible to develop a single evaluation approach that can support both locally meaningful evaluations *and* longer-term comparisons between methods? This paper outlines a methodological framework for the evaluation of systemic problem structuring methods that seeks to do just this.

1. Introduction

Participative methods facilitate the engagement of stakeholders and/or citizens in decision making to address complex organizational, social, environmental or technological issues. They are used by management researchers and practitioners (as well as other social scientists) in the context of interventions to stimulate deliberative dialogue and the development of change proposals (Beierle and Cayford, 2002; Rowe and Frewer, 2004).

A subset of the general class of participative methods is *problem structuring methods* (PSMs). A substantial number of these have been developed by operational researchers over the past 50 years, although the term 'problem structuring' itself was only introduced into the operational research (OR) lexicon a couple of decades ago (Rosenhead, 1989, 2006; Rosenhead and Mingers, 2001, 2004). A distinguishing feature of PSMs, compared with many other participative methods developed by social scientists, is the use of models as 'transitional objects' to structure stakeholder engagement (Eden and Sims, 1979; Eden and Ackermann, 2006) and provide a focus for dialogue (Franco, 2006). These models may use words, pictures and/or numbers to represent, for example, people's understandings of a problematic situation; the assumptions underpinning a particular stakeholder perspective; and/or the activities that might be needed to improve the situation. Usually, models are qualitative and are constructed collectively in a workshop, but sometimes they are brought in by a facilitator based on previous inputs from participants and are used to orientate engagement: "the model... plays a key role in driving the process of negotiation towards agreement through discussion and the development of a common understanding" (Franco, 2006, p. 766). However, a 'common understanding' does not necessarily imply consensus or agreement across the board: it may be an agreed understanding of the differences between people's perspectives and what accommodations are possible in the circumstances (Checkland and Scholes, 1990). Qualitative models have traditionally been produced on flip charts using marker pens, but computer-mediated modelling is increasing in popularity, and this can facilitate remotely distributed and/or anonymous stakeholder participation, bringing advantages compared with face-to-face, pen and paper modelling (Er and Ng, 1995; Fjermestad, 2004; Fan et al., 2007).

Some PSMs are explicitly *systemic* (Jackson, 2000; Midgley, 2000, 2003). They not only seek to enhance mutual understanding between stakeholders, but they also support participants in undertaking 'bigger picture' analyses, which may cast new light on the

* Corresponding author at: Centre for Systems Studies, Business School, University of Hull, Hull HU6 7RX, United Kingdom. Tel.: +44 (0)1482 463316.
E-mail address: g.r.midgley@hull.ac.uk (G. Midgley).

0377-2217
http://dx.doi.org/10.1016/j.ejor.2013.01.047

issue and potential solutions. Notably, systemic PSMs are used to broaden the perspectives of participants in order to facilitate the emergence of new framings, strategies and actions. Typical questions addressed by different systemic PSMs include:

- Whose viewpoints and what aspects of the issue should be included in analysis and decision making, and what should be excluded? (e.g., Ulrich, 1994; Midgley, 2000).
- What are people's different perspectives on the issue, and what values and assumptions underpin these perspectives? (e.g., Checkland and Scholes, 1990; Checkland and Poulter, 2006).
- What interactions within and across organisational, social and environmental phenomena could produce desirable or undesirable outcomes? (e.g., Vennix, 1996; Maani and Cavana, 2007).

We argue in this paper that a new framework is needed for the evaluation of systemic PSMs. However, given that so little has previously been written on this subject, we also draw upon the wider literature about evaluating participative methods (beyond problem structuring, systems thinking and OR).

2. Evidence for the value of systemic problem structuring and other participative methods

When claims are made for the success or failure of systemic problem structuring and other participative methods, the authors making those claims are usually required to justify them. Various reviews of the literature on the evaluation of participative methods suggest that most of the justifications provided by researchers are based on personal reflections alone (Entwistle et al., 1999; Connell, 2001; Rowe and Frewer, 2004; Sieber, 2006; White, 2006). Clearly, many researchers are highly experienced, so their reflections should not be dismissed out of hand. Nevertheless, unless they think broadly and from different perspectives about the criteria they use to evaluate their participative interventions, they may miss evidence that does not fit their current thinking about what is important (Romm, 1996; Midgley, 2011). We therefore suggest that there is a need for caution in accepting researcher reflections alone as reliable evidence of success or failure.

Most researchers undertaking evaluations of participative methods beyond personal reflections tend to conduct post-intervention debriefings or interviews with project participants. These evaluations are often based on explicit criteria reflecting the researcher's experience, a given theory, a literature review and/or stakeholder expectations generated through a consultative exercise (Beierle and Konisky, 2000; Rowe and Frewer, 2004). In some cases, formal evaluation instruments have been developed and applied (e.g., Duram and Brown, 1999; Rowe et al., 2004; Berry et al., 2006; Rouwette, 2011). Also a number of researchers advocate triangulation across two or more evaluation methods, such as interviews, focus groups, participant observations, surveys, literature reviews and document analyses (Duram and Brown, 1999; Buysse et al., 1999; Charnley and Engelbert, 2005; Rowe et al., 2005; Cole, 2006; McGurk et al., 2006; Franco, 2007; Rouwette, 2011).

What is clear from the literature, however, is that only a very small minority of studies (e.g., Valacich and Schwenk, 1995a; Halvorsen, 2001; Rouwette et al., 2011) seek to compare between methods or across case studies undertaken by different researchers. A particularly significant study was undertaken by Beierle and Cayford (2002), who quantitatively compared broad classes of methods using a standard set of variables applied to 239 case studies of public participation. They concluded that more intensive processes (such as mediation workshops) are better than less intensive processes (such as public meetings) at achieving a wide range of outcomes. We suggest that the use of systemic PSMs is relatively intensive compared with several of the other participative processes investigated by Beierle and Cayford (2002), so this gives us grounds to be cautiously optimistic. However, we cannot take this study as strong evidence because they did not specifically identify systemic PSMs as a category for comparison with other participative approaches.

Therefore, the overall picture is of many claims for the benefits of a diverse array of systemic problem structuring and other participative methods, with varying degrees of evidence provided by researchers to support these. Only a few studies have compared across methods, and even these have only been able to contrast broad classes of approach.

The key question is: what kind of evaluation is both necessary and possible? We have already argued that researcher reflections alone can be problematic, but are there methodological or practical reasons to prefer either locally focused evaluations (possibly with some learning across case studies, when this is feasible) or large-scale, quantitative comparisons between methods?

2.1. Different evaluation approaches

Rowe and Frewer (2004), reflecting on social science approaches to evaluating participative methods, classify them into three types. First there are *universal* evaluations: i.e., ones claiming to produce knowledge that is applicable across all types of participative method and intervention. According to Rowe and Frewer, to achieve universality, large-scale quantitative studies are needed. Nevertheless, to make comparisons possible, only variables of general relevance across all methods and interventions can reasonably be assessed. Next there are *local* evaluations: comparing between a subgroup of methods or intervention types. These require smaller scale studies and can incorporate more detailed questioning, as the variables to be examined may be relevant only to the subgroup of methods under study rather than to all possible methods. Some researchers working on local evaluations advocate a quasi-experimental approach, either testing methods in the laboratory or in controlled field conditions. Rowe and Frewer (2004) call the third and final type of evaluation, which the majority of researchers use, *specific*. This means focusing on only one method or intervention. The advantage of this is that the evaluation can be made locally relevant, drawing (for example) on information about the unique expectations of stakeholders to establish evaluation criteria. Rowe and Frewer argue that, while it is difficult (for practical reasons) to conduct truly universal evaluations, researchers should aim to achieve as much generality as possible, and should certainly do more than undertake evaluations with only a specific remit because generalising from these is highly problematic.

White (2006) argues that very similar distinctions have been made in the OR and group decision support literatures, and preferences for universality (to a greater or lesser extent) or specificity reflect the positivist and interpretivist paradigms respectively. Positivists are said to argue for objective, quantitative, comparative studies that are capable of revealing the generalisable advantages and disadvantages of different methods, although (like Rowe and Frewer, 2004) many are forced by the impracticality of undertaking truly universal studies to resort to more local quasi-experiments in either the laboratory or the field. Authors in this tradition include Nunamaker et al. (1991), Fjermestad and Hiltz (1998), Pinsonneault et al. (1999), Fjermestad (2004) and Joldersma and Roelofs (2004). In contrast, interpretivists (such as Eden, 1995; Eden and Ackermann, 1996; Shaw, 2003) argue that what matters most in an evaluation is what is achieved by the method *in a given context, judged from the perspectives of stakeholders*. It is therefore hardly surprising that most interpretivists are in favour of undertaking specific (single case study) evaluations. See Connell (2001), Bryant

and Darwin (2004), Phahlamohlaka and Friend (2004) and Sørensen et al. (2004) for examples.

Our own position on these debates is as follows. For both epistemological and methodological reasons, we do not accept that it is possible to generate universally applicable knowledge about methods. Our epistemological argument is that knowledge (or understanding) is always linked to the purposes and values of those producing or using it, and is dependent on the boundary judgements that they make (Churchman, 1970; Ulrich, 1994; Alrøe, 2000; Midgley, 2000). To claim that knowledge about systemic PSMs (or any other phenomenon for that matter) is universal is to ignore the purposes, values and boundary judgements that make the knowledge relevant and adequate for a particular context. This argument is consistent with the epistemological assumptions made by most of the creators of PSMs (Jackson, 2006).

We also have two methodological arguments following from our epistemological one. First, claiming universality for knowledge about systemic PSMs would suggest that this knowledge will remain stable over time. However, it is clear from the literature (e.g., Rosenhead and Mingers, 2004; Shaw et al., 2006; Franco et al., 2007) that new problem structuring methods are being produced on a regular basis, indicating that people are learning from previous practice and are also having to respond to an ever increasing number of unique practical situations. Given that this is a dynamic research environment, it would seem risky to assume that a standard set of variables will always be relevant. Undertaking a series of more limited comparisons between particular methods might be methodologically wiser than trying to set up a 'universal' study.

Our second methodological argument, following Eden (1995) and others, is that only seeking knowledge about the supposedly generic strengths and weaknesses of methods ignores legitimate questions that can be asked about the effectiveness of those methods in particular local circumstances. Given that operational researchers using systemic PSMs work most of the time in particular contexts with unique features, it would only meet a small fraction of the need for evaluation if we were to ignore non-generic questions, and this would be unacceptable to local stakeholders wanting to know what will best meet their particular needs.

There can also be problems with what Rowe and Frewer (2004) call 'local' evaluations (comparing more limited sets of methods in smaller scale research projects). Some have called for 'objective' local studies rather than the simple reporting of subjective impressions (Pinsonneault et al., 1999; Rowe and Frewer, 2004; Rowe et al., 2005). However, when the pursuit of objectivity involves a retreat into the laboratory to conduct controlled experiments (e.g., Valacich and Schwenk, 1995b; Montazemi et al., 1996; Delaney et al., 1997), then the validity of the comparison of methods has been questioned due to the artificiality of the situation (Eden, 1995; Er and Ng, 1995; Shaw, 2003; White, 2006). While we accept that laboratory experiments are valid when some technical questions are being investigated, such as whether computer mediation enables the capture of more participants' statements than use of a flip chart (Gallupe et al., 1992; Fjermestad, 2004; Fan et al., 2007), we suggest that questions relating to the performance of methods in the context of stakeholder disagreement and conflict are another matter entirely. In the laboratory, 'decisions' made by participants have no longer term consequences, so participants are unlikely to think or behave in the same way as they do when faced with disagreements and potential outcomes that really matter to them. If quasi-experiments are established in the field instead of the laboratory, then this raises other problems: McAllister (1999) argues that it is unethical to use a control when dealing with real community issues, and Duignan and Casswell (1989) simply point to the impracticality of finding two situations that are sufficiently alike to make a comparative study robust.

In making criticisms of attempts to take a controlled or quasi-experimental approach, some authors have advanced alternatives. Kelly and Van Vlaenderen (1995), McKay (1998), Jenkins and Bennett (1999), De Vreede and Dickson (2000), Gopal and Prasad (2000) and Allsop and Taket (2003) advocate 'emergent' methodologies: i.e., ones where criteria for evaluation emerge through engagement with stakeholders. Eden (1995) makes the important point that most interventions are complex, and researchers can rarely anticipate everything that will become important, so the evaluation approach needs to be able to respond to the unexpected.

However, does this mean that evaluations cannot legitimately generalise from single, specific case studies to other contexts that may be similar in at least some respects? It is widely accepted that the 'success' or 'failure' of a method in any particular case results from use of the method-in-context and cannot be attributed to the method alone (Checkland and Scholes, 1990; Buysse et al., 1999; McAllister, 1999; Murphy-Berman et al., 2000; Morgan, 2001; Margerum, 2002; Rowe and Frewer, 2000, 2004; Branch and Bradbury, 2006; McGurk et al., 2006; White, 2006; Warburton et al., 2007). Nevertheless, several researchers claim that *cross case study learning* is possible, with two or more research teams reflecting on similarities and differences between cases (e.g., McAllister, 1999; Yearley, 2006; White, 2006). Checkland (1981) argues that evaluating a systemic methodology depends on the long term accumulation of evidence from a diverse range of applications, giving progressively more confidence that the approach is useful across contexts: it is only through such an accumulation of evidence that the efficacy (does it work in the ways claimed?), effectiveness (is it the best approach for what is needed?) and efficiency (are maximum benefits gained at minimum cost?) of an approach can be reasonably assessed (also see Checkland et al., 1990; Zhang et al., 1997).

2.2. A pragmatic step sideways

It would appear from the literature that most researchers accept the logic of interpretivism and are more inclined to undertake specific, locally meaningful evaluations (and possibly learn across these) than attempt comparisons between methods using generic, quantitative measures (Mingers and Rosenhead, 2004; White, 2006). However, we have to ask whether this means that all forms of quantitative comparison are redundant. White (2006) argues that the debate in the problem structuring research community has become unhelpfully polarised, with many advocates on both sides taking 'purist' positions and spurning methods that could enhance their own evaluation practices. He therefore proposes a more pragmatic line: identifying important research questions and asking what evaluation methods might answer these most effectively. We agree that this is a useful step sideways from the either/or debate, but we nevertheless suggest that identifying effective evaluation methods to address particular research questions involves considering the *practicalities* of undertaking evaluations as well as the norms of what constitutes a valid or legitimate methodology. A difficult balance has to be struck between rigour and relevance (Shaw, 1999) because if the former is unquestioningly prioritised then there is good evidence that stakeholders will not co-operate (Rowe et al., 2005). Importantly, this balance has to be struck regardless of whether an emergent approach is being followed or whether a more traditional scientific study comparing methods is being undertaken.

In sympathy with White's (2006) pragmatic intent, we set out to propose an evaluation approach that supports locally meaningful evaluations *and* is capable of generating data for longer-term quantitative comparisons between methods without compromising local relevance. The overall framework is based in the tradition

of multi-method systemic intervention (e.g., Flood and Jackson, 1991; Jackson, 1991, 2000; Flood and Romm, 1996; Mingers and Brocklesby, 1997; Mingers and Gill, 1997; Midgley, 2000, 2003; Taket and White, 2000; Burns, 2007), but instruments can be employed as part of the emergent evaluation of methods that enable data gathering for both immediate local and longer-term comparative use. Below, we outline the rationale for our framework. We then discuss early work in developing and testing a questionnaire that can be used in the context of it.

3. A new evaluation framework

Our evaluation framework is represented in Fig. 1. An evaluation using it is primarily focused on the use of a particular *method* (or set of methods) in a *context* for particular *purposes*, giving rise to *outcomes*. The words in italics in the previous sentence represent what we regard as four necessary foci to evaluative inquiry, and they need to be interrelated in the context of a specific reflection on the use of a method. Exploration of these aspects may proceed in any direction around Fig. 1, and may loop back and forth according to the needs of those involved in the evaluation.

We note that it is possible to develop much more elaborate conceptual frameworks than ours, with strong utility for research (e.g., Champion and Wilson, 2010). However, if a framework is to be memorable in the context of practice, it needs to use relatively few high-level concepts organised in a visually appealing manner. Lower level concepts can be introduced under the higher level ones.

Other authors have proposed similar, but not identical, frameworks to ours. Buysse et al. (1999) and McAllister (1999) advocate the exploration of both purposes and context, but tend to take as given the nature of the method to be evaluated. Pinsonneault and Kraemer (1990), Flood (1995), McGurk et al. (2006) and Rouwette et al. (2009) ask researchers to reflect on the adequacy of their contextual analyses, their choices of methods or processes and their intervention outcomes. However, the purposes being pursued become implicit: whether or not these differ from the outcomes is not necessarily at issue. Warburton et al. (2007) propose reflection on context, purposes and methods, but they do not consider the implications of the researcher's role in the situation. This is an important issue for us (and is represented in Fig. 1 by the text in the lower parts of the four ellipses) because our experience is that the researcher becomes an *interactive part* of the situation in which he or she is seeking to intervene using systemic PSMs (also see Checkland, 1981), and his or her identity and relationships can significantly affect the trajectory of an intervention (Brocklesby, 1997; Mingers, 1997; Midgley et al., 2007).

In our approach, when looking at a single case study, there is no pretence that it is possible to evaluate a method independently from the purposes it is put to, its outcomes and the context in which it is applied. Nevertheless, we can still inquire about the *relationships between* the method, purposes, outcomes and context. Inquiry focused on an intervention can look, for example, at how satisfactorily the method addressed given purposes; what aspects of the context enabled or constrained its application; and whether it gave rise to anticipated or unanticipated outcomes. Some features of the context-purposes-methods-outcomes relationship may be apparent early on in an intervention, while others may only emerge as the inquiry unfolds. Hence the utility of an emergent approach for the evaluation of methods, which remains open to new understandings as inquiry deepens (e.g., Kelly and Van Vlaenderen, 1995; Jenkins and Bennett, 1999; Gopal and Prasad, 2000; Allsop and Taket, 2003).

Below, we examine the four aspects of evaluation (context, purposes, methods and outcomes) in turn, explaining why each of these is important to developing a rounded understanding of how a method has operated in a particular case study of practice.

3.1. Context

More has been written about context than the other aspects of evaluation, arguably because it is crucial to good practice to realise that the same method utilised by the same researcher can succeed

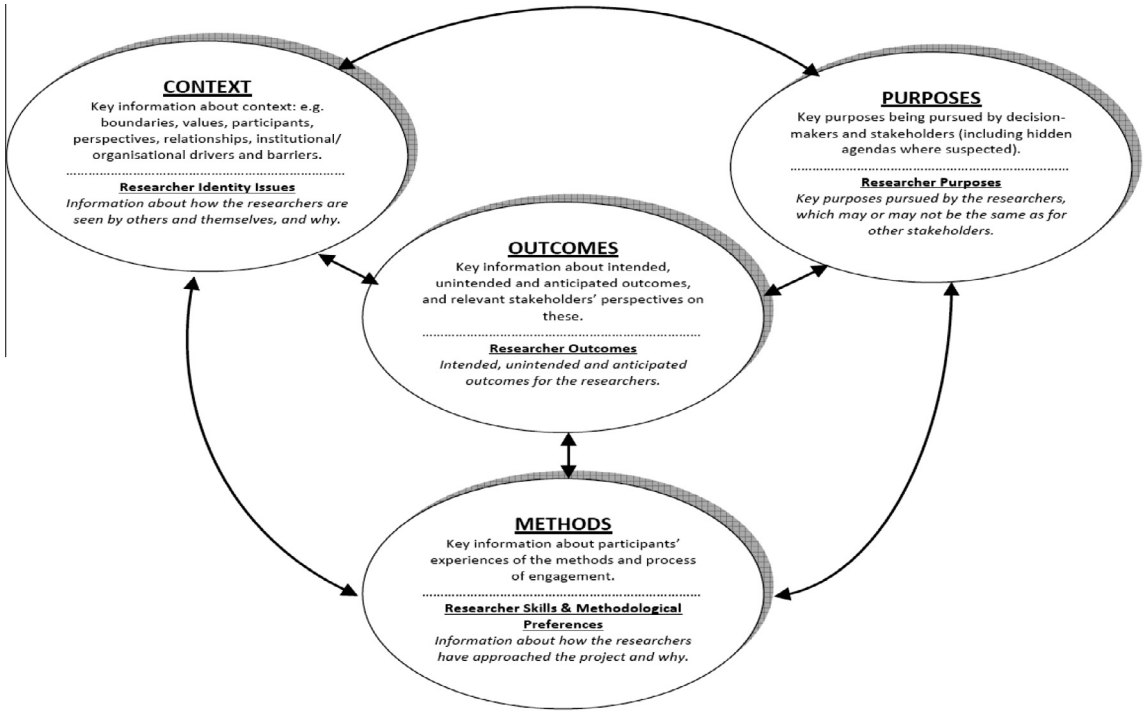

Fig. 1. Conceptual framework for the evaluation of systemic problem structuring methods.

or fail depending on the complexities and dynamics of the situation (e.g., Checkland and Scholes, 1990; Nunamaker et al., 1991; Buysse et al., 1999; McAllister, 1999; Murphy-Berman et al., 2000; Rowe and Frewer, 2000, 2004; Morgan, 2001; McGurk et al., 2006; White, 2006; Warburton et al., 2007; Champion and Wilson, 2010).

Relevant aspects of context identified by Jackson and Keys (1984) are the complexity of the issue being addressed using a systemic method and the relationships between the participants. In contrast, Margerum (2002) identifies potential contextual inhibitors of effective participation: a low level of commitment by key decision makers; parochialism (which can negatively affect inclusiveness); participants having inadequate skills and abilities; operational issues preventing the implementation of ideas; a lack of strategic thinking beyond the exercise at hand; poor leadership; and scarcity of resources. Ong (2000) discusses the facilitative effects of strong social capital, and Alberts (2007) documents the negative effects of participant inexperience and ignorance of technical issues. Branch and Bradbury (2006) claim that a key aspect of context is managerial attitude: especially the disclosure (or not) of relevant information; whether managers set agendas unilaterally or are open to power sharing; whether or not there is mutual respect in relationships; whether there is accountability to stakeholders; and whether or not people believe that a transparent decision making process will be used following stakeholder participation. McCartt and Rohrbaugh (1995) argue that a key aspect of managerial attitude is openness to change, and participative methods are often ineffective without it. Kelly and Van Vlaenderen (1995) and Brocklesby (2009) concentrate on stakeholder interactions, looking at how patterns of mistrust and miscommunication can become established and affect the use of participative methods. Related to this is the identity of the researcher: Midgley et al. (2007) discuss how identity issues can make a significant difference to the quality of relationships, and hence the success or failure of a method (this is represented in Fig. 1 by the lower half of the 'context' ellipse). Champion and Wilson (2010) provide a particularly useful set of contextual variables to be considered, based on a literature review and feedback from practitioners: organisational structure; influence of the external environment; length of history of the problem in focus; politics and personalities; perceived implementation difficulty; and the level of experience of stakeholders.

No doubt the list of possible aspects of context could be extended indefinitely (Gopal and Prasad, 2000), and different issues will be relevant in different situations, so we argue that it is more useful to give some methodological guidelines for exploring context in local situations than it is to provide a generic inventory of variables. We suggest that the following guidelines, derived from reflections on different systems paradigms (as represented by Jackson (1991), and others), can all contribute in different ways to *boundary critique* (the exploration of different possible boundaries, or frames, for a contextual analysis):

- Underpinning different boundary judgements may be quite different perspectives on the nature of the context (Churchman, 1970). Therefore, exploring diverse perspectives (e.g., as advocated by Checkland (1981)) may lead to the identification of alternative possible ways of bounding a contextual analysis (Ulrich, 1994).
- Establishing a boundary for analysis involves making a value judgement on what issues and stakeholders are important or peripheral (Ulrich, 1994). Therefore, undertaking an exploration of different stakeholders' values and priorities can be helpful. It is also useful to identify conflicts between people making different value judgements as well as processes of marginalisation that may constrain stakeholder participation or make the discussion of some phenomena taboo (e.g., Midgley, 2000).
- Identifying the presence of influential institutional or organisational systems may be important. Any such system can have its own agenda, rationality and momentum that may come to dominate an intervention (Douglas, 1986; Luhmann, 1986; Brocklesby, 2009), yet organisational systems still have to interact with others, and tensions can result (Paterson and Teubner, 1998). Thus, an institutional analysis can be a useful aspect of boundary critique.
- There may be socio-economic and ecological systems providing resources that can be used constructively by participants, or these systems may impose limits on what is achievable without incurring negative side-effects (Clayton and Radcliffe, 1996). Economic issues may point to concerns about social justice, which (if present) could influence people's perceptions of the effects of systemic PSMs: i.e., the use of a particular method may be seen as supportive of just or unjust social relationships (Jackson, 1991, 2006), so it can be useful to look at the effects of socio-economic systems as part of boundary critique. Taking explicit account of ecological systems can also enhance boundary critique by challenging a tendency to uncritically resort to boundaries defining exclusively human systems, thereby marginalising the ecological (Midgley, 1994). Pettigrew (1987) notes that wider systemic (e.g., socio-economic and ecological) contexts not only influence perceptions of methods and processes, but also the content of participants' deliberations.
- Within and across ecological, economic, social and organisational systems, there may be important causal pathways, and in particular feedback loops, that can point to systemic enablers of, or constraints on, an intervention (e.g., Forrester, 1969). Bateson (1970) argues that it is important not to 'cut' relevant feedback loops, and again this is a good principle to inform boundary critique: when we see interconnections stretching beyond people's usual understandings of context we can ask whether it is important to widen the boundaries of analysis to account for these.

Essentially then, a useful approach to exploring context may involve looking at different possible boundaries for analysis, concentrating in particular on different stakeholder perspectives; value judgements around the inclusion or exclusion of issues and stakeholders; processes of conflict and marginalisation; ecological, economic, social and institutional/organisational systems that may act as enablers or constraints; and causal relationships and feedback processes within and across those systems.

3.2. Purposes

The second aspect of our evaluation framework is concerned with exploring stakeholders' *purposes* in engaging with an intervention. Purposes are closely linked with values and motivations (McAllister, 1999), and they are important to an evaluation because particular methods are likely to appear more or less useful depending on the purposes being pursued. Different methods are generally good for different things (Flood and Jackson, 1991), and it is the perceived 'fit' between purpose and method that is important to evaluate: a disjunction may be responsible for an attribution of failure.

It is important to consider possible hidden agendas as well as explicitly articulated purposes. These may significantly affect the trajectory of an intervention (for instance through sabotage), and thereby the evaluation of the method used (Ho, 1997). It is also

useful to look out for mismatches between articulated purposes and ones attributed by others (both to individuals and organisations) because mismatches of this kind often signal mistrust or conflict that will be relevant to the performance and evaluation of methods (Kelly and Van Vlaenderen, 1995).

Whether or not there is mistrust or conflict, there will often be multiple purposes at play. If people come to an intervention with different purposes for engaging, then it is likely that different evaluation criteria will be important to them (McAllister, 1999; Murphy-Berman et al., 2000; Tuler et al., 2005; Rowe and Frewer, 2004; Masozera et al., 2006; White, 2006). While Rowe and Frewer (2004) say that an appropriate response is to *set aside* the purposes and preferred criteria of diverse stakeholders in favour of a single criterion of 'acceptability of the method to all parties', more nuanced findings will be generated by evaluating the method against multiple criteria of relevance to different stakeholders (Murphy-Berman et al., 2000).

Note here that the purposes of the researcher should not be excluded from consideration. There may be a good 'fit' between stakeholder and researcher purposes, but there may also be disjunctions. An example is when the researcher works in a university and brings a pre-defined academic agenda into the intervention, which may influence how systemic PSMs are chosen and used. Even when an academic researcher makes a significant effort to be responsive to stakeholders, there may still be mistrust stemming from *expectations* of divergent purposes (Adams and McCullough, 2003), and this may affect the evaluation of methods.

3.3. Methods

Earlier we mentioned that some authors have advocated taking account of the effects of stakeholder purposes and context, but they tend to take the nature of the method being evaluated for granted. It is important not to do this because different methods make different theoretical and methodological assumptions about (amongst other things) human relationships, knowledge, values and the nature of the situation that is the focus of the intervention (e.g., Jackson, 1991, 2000; Romm, 1996; Spash, 1997; Midgley, 2000). In an evaluation, we need to be able to account for if and how these assumptions have shaped the unfolding of the intervention.

There may also be elements of methods that people in some cultures (or sub-cultures) will find it easier to accept or work with than others. While culture may be conceived as an aspect of the context, it may also be reflected in the construction of a method, which is why a number of methodologists working outside the Western tradition have sought to establish systems/OR and other approaches developed from their own philosophical and cultural perspectives (e.g., Smith, 1999; Zhu, 2000; Shen and Midgley, 2007). Becoming aware of the cultural norms embodied in a method may be important to understanding its effects across cultural contexts.

The *process of application* of a method is important as well, not just the method as formally constructed (Keys, 1994). For instance, the same basic method may be enacted in quite different ways depending on the preferences and skills of the researcher/facilitator and the demands of the situation at hand. Compare, for example, two significantly different accounts of soft systems methodology (SSM): Checkland and Scholes (1990) discuss how the methods from SSM should be utilised in a flexible and iterative manner, while Li and Zheng (1995) insert some of the same methods into a 'general systems methodology'. In the latter case, it is clear that the methods of SSM are to be applied in a linear sequence. In many contexts, such a significant difference in the process of application of the same set of methods is bound to impact upon the way that set will be perceived.

Not only can the researcher's preferences and approach be important, but also the extent of his or her skills and experience may influence whether the use of a method is perceived as successful or not. Mingers (1997) describes these as the "intellectual resources" that the researcher brings into an intervention, and it is important to be able to distinguish whether problems encountered in the use of a method derive from the limitations of the method itself or from the inadequate resources of the researcher. Conversely, the evaluation may reveal that the researcher had exceptional skills that were used to good effect in securing a successful outcome.

3.4. Outcomes

In addition to collecting information about the assumptions embedded in, and the process of application of, the method being evaluated, it is most important to collect data on its *outcomes*. These can be viewed from the perspectives of those involved and affected (usually participants in workshops, but others might be relevant too, depending on the context). This is the crux of the evaluation of methods.

It is necessary to distinguish *outcome* criteria from *process* criteria (Chess and Purcell, 1999; Rowe and Frewer, 2004). Process criteria (e.g., did the process of applying the method give everyone a chance to speak, allow creative exploration, or enable a fair evaluation of options?) were discussed in the last section. In contrast, outcome criteria refer to whether, in a particular case, the method facilitated the achievement of specific goals (e.g., the production of a plan or the generation of a common vision). The difference between process and outcome criteria can get a little blurred when an explicit goal of an intervention is, for instance, to facilitate participatory engagement. Nevertheless, keeping the distinction explicit helps us avoid potentially major mistakes like focusing so much on process that we fail to notice that people's purposes for the intervention have not been achieved, or focusing so much on outcomes that we miss negative effects of the process on participants.

Outcomes may also be longer term in nature, and these are not always predictable or easy to measure (Duignan and Casswell, 1989). Indeed, making a causal link between an intervention and an outcome that emerges, say, 10 years down the line is often extremely difficult. Long-term follow up studies are needed if some kinds of outcomes (e.g., those concerned with sustainability) are to be properly assessed.

The usual means of measuring many short-term outcomes of a method (other than through personal reflections by the researcher) is by gathering feedback from participants following workshops, often giving them questionnaires to fill in as soon as the workshop is complete (e.g., Duram and Brown, 1999; Berry et al., 2006; Sykes and Goodwin, 2007; Rouwette, 2011). This is an approach that we have found valuable in our own practice, and we have developed a questionnaire with some sections that are changeable from intervention to intervention to reflect specific local needs. Other sections are relatively stable and are used repeatedly across a variety of local intervention contexts. Both types of section are useful for locally meaningful evaluations, but the latter (stable) sections can also yield quantitative data for use in longer-term, cross-method comparisons. More information about our questionnaire is provided below.

4. Developing an evaluation questionnaire

Our questionnaire is presented in Appendix A. Importantly, *it is not the only tool needed for evaluating systemic PSMs*: for instance, it cannot capture data on longer term outcomes. Nevertheless, it can make a useful contribution by gathering the viewpoints of

participants on process and short-term outcomes immediately after their involvement in a workshop. The questionnaire has the following sections:

1. A five-point scale for the quantitative assessment of usefulness, plus open questions about what people liked and disliked, and what could have been done differently. Additional open questions reflecting local contingencies can be added in here if and when required.
2. Fourteen questions with five-point Likert scales enabling the quantitative evaluation of whether certain things have been achieved. Both process and outcome questions are included here, and this is a set of questions that is not tailored to particular interventions (except occasional words where it is necessary to mention that the workshop is focused on water, housing, health, policing, etc.). The process we went through to derive this set of questions is discussed below.
3. Thirteen questions, again with five-point Likert scales, addressing potential negative attributes of (or things that can go wrong when using) systemic problem structuring methods. Once again this is an unchanging set of questions, and our process for deriving them is discussed below.
4. A set of open ended questions asking respondents to assess the process from their own cultural viewpoints. These questions are usually worded generally so they are relevant to multiple cultural perspectives, but specific questions relating to particular cultures can be added if required (for example, in New Zealand there often needs to be a specific focus on Māori perspectives).
5. Questions gathering basic demographic data (stakeholder category, gender, age, etc.).

4.1. The development process

Our questionnaire was first developed in the context of a research programme aiming to generate and evaluate new systemic problem structuring methods for use in promoting sustainable resource use (Winstanley et al., 2005; Hepi et al., 2008).

The adaptable parts of the questionnaire (Sections 1, 4 and 5 above) were relatively straight forward to design, although they required some iterative testing to get them right. The more difficult task was to produce Sections 2 and 3, which needed to yield data for meaningful use in both local evaluations and longer term comparisons between methods. Because of the latter, the questions had to be reasonably generic. Other authors suggest a number of different ways of producing generic evaluation criteria, and these have been summarised by Beierle and Konisky (2000) and Rowe and Frewer (2004). A combination of their thinking, plus an addition of our own, suggests that there are six distinct approaches: author-generated (resulting from personal experience); practice-based (deriving from explicit reflections on case studies); theory-based (evaluating according to the expectations one would have if one agreed with a particular theory); literature-based (deriving from a review of other authors' work); expert-based (drawing on the views of an advisory panel); and survey-based (finding out from potential participants, either through interviews or a mail survey, what their most widely held expectations are). Some authors have combined two or more of the above. We adopted an expert- and literature-based approach, with a couple of author-generated questions being added in as well. More details are provided below.

We started with a key question: what do we want to measure? One option was to focus only on criteria that we would expect to be meaningful for all systemic PSMs. This is the approach taken by Bjärås et al. (1991) and Beierle and Konisky (2000). However, while it is useful to identify 'common denominators' and assess methods against these, this does not help in evaluating the unique attributes of methods that might make them complementary rather than competing. To evaluate these, it is important to look at the set of possible common *and divergent* attributes that a range of systemic PSMs might exhibit.

We therefore set out to identify a number of methodologies and methods that could fairly represent the diversity of participative systems approaches. We established a panel of six internationally known writers on systems thinking, all of whom suggested candidate methods. We ended up with six systemic PSMs (soft systems methodology; interactive planning; causal loop diagramming; viable system diagnosis; critical systems heuristics; and strategic assumption surfacing and testing) that all claim to do different things. We then reviewed the literature on these, drawing out a set of attributes that could form the basis for questions to be asked of participants in workshops. We also asked the international panel to suggest their own evaluation criteria, and we added in a couple that were not apparent from the literature review but, in our experience, were important. This list was then sent back to the panel for peer review, resulting in some amendments. We ended up with a set of questions for field testing.

We note that our questions align well with five high level criteria suggested by Hjortsø (2004) for the evaluation of PSMs, except that our questions go into much more detail. Hjortsø's criteria are the extent to which the method is a good fit for the context, and whether it supports (i) mutual understanding; (ii) stakeholder involvement in decision making; (iii) the acceptance, transparency and accountability of decision making; and (iv) the collaborative management of complexity and conflict.

It is important to declare that we focused *only* on evaluation criteria relevant to systemic PSMs: we could not assume that criteria that have been used to assess other PSMs and participative methods would automatically be relevant. In making the decision to take this approach, we set aside another research opportunity that is available for others to pick up in future: looking at the literature on group processes and focusing evaluations quite specifically on how systemic PSMs enhance these. While there are certainly questions about group process in the questionnaire, these reflect the variables that our expert panel and literature review suggested were most relevant to the success or otherwise of systemic PSMs: they are not based on wider reading on group process.

It is generally accepted (e.g., Cavana et al., 2001) that a questionnaire to be employed in an experimental context should be tested for validity (does it measure what we think it does?) and reliability (does it give consistent results?). However, for an evaluation questionnaire to be employed in the field outside the context of experimental studies, *usability* is just as important, if not more so (Rowe and Frewer, 2004). Usability means asking whether people are actually prepared to complete the questionnaire and do so in a sensible manner. Rowe and Frewer (2004) note that, because compromises have to be made in questionnaire design to ensure usability (e.g., the questions need to be answerable in 5–10 minutes at the end of a gruelling day), usability is often inversely related to validity and reliability (both of which are enhanced by the generation of more rather than less data). This may be the case but, as Rowe et al. (2005) say, there is no point even beginning to consider validity and reliability if the instrument cannot be used in the first place. We agree with Rowe et al. (2005) that assessing usability has to be a first priority, although validity and reliability should not be ignored. At this point in our research, we have tested for usability but (for reasons to be explained in Section 5 of this paper) the more problematic task of evaluating validity and reliability has not yet been undertaken. This will be the subject of future research.

To check for usability, we field tested the questionnaire in five different interventions, each time making small amendments in

response to issues thrown up by the ways in which people approached the questions:

- facilitating consultation with land owners and community interest groups as part of a feasibility study for the construction of a new water storage dam (Winstanley et al., 2005).
- working with an Australian NGO and its stakeholders in exploring policy options to address the public injecting of illicit drugs (Midgley et al., 2005);
- facilitating workshops with the police and other stakeholders in the New Zealand criminal justice system to look at ethical issues associated with anticipated future developments of forensic DNA technologies (Baker et al., 2006);
- reviewing the process used by the New Zealand Ministry of Research, Science and Technology to develop 'roadmaps' for long-term investments in environment, energy, biotechnology and nanotechnology research (Baker and Midgley, 2007); and
- developing a new collaborative evaluation approach in partnership with regional council staff responsible for facilitating community engagement in sustainability initiatives (Hepi et al., 2008).

We also tested the questionnaire on interventions undertaken by people other than ourselves: a public meeting and a stakeholder forum convened in two different areas of New Zealand to discuss water shortages.

Following observations of participants completing the first version of the questionnaire, it was judged to be over-long. We shortened it, but then in later iterations found that omitting some of the questions led to important gaps in our data sets. We ended up finding a compromise between comprehensiveness and brevity. On our first iteration of field testing, we also undertook a basic analysis to check that there were no counter-intuitive answers (which might suggest the misinterpretation of a question); that there was no tendency for people to tick the same point on all the scales (indicating boredom or a lack of comprehension); and that similar questions generated similar answers. All these checks proved satisfactory. Having undertaken this series of field tests, we are now confident of the usability of our questionnaire. Nevertheless, we fully acknowledge that it would be possible to further test usability by interviewing participants on what they were thinking about when they answered the different questions.

4.2. Interpreting data generated through use of the questionnaire

Before closing this discussion of our questionnaire, it is important to note that the quantitative data generated through it, on process and short-term outcomes within the context of a particular systemic intervention (i.e., a single case study), always has to be interpreted in relation to the other aspects of our framework: context, purposes, longer-term outcomes and researcher skills and preferences. Failure to undertake analyses of these aspects could result in attributions to the method of results that might have had other origins.

It is important to note that qualitative information about the purposes of stakeholders, the context, the assumptions embedded in the method and the skills and methodological preferences of the researcher cannot easily be gathered using a standardised instrument like a questionnaire. Therefore, the questionnaire data needs to be considered in a reflective workshop covering all the relevant aspects of method, purposes, context and outcomes. Our normal practice is to bring the research team together with key stakeholders, and we use the concepts in our framework (Fig. 1) to structure a dialogue, recording people's viewpoints. By interpreting the findings from the questionnaire in relation to a participative, 'bigger picture', emergent analysis of the use of a method, it is possible to develop a more holistic and nuanced understanding of the performance of the method than if questionnaire data alone had been used.

Table 1 gives some additional, generic, high-level questions for use in a reflective workshop with stakeholders, going well beyond process and short-term outcome variables assessed through the questionnaire. However, we should note that these offer a guideline only, as questioning needs to be tailored to the specific context of the intervention that has been undertaken, taking account of the knowledge and perspectives of stakeholders. For example, it is unlikely that many stakeholders will have knowledge of the theoretical assumptions embedded in systemic PSMs, so this is something that needs to be considered by the researchers beforehand and then (if relevant) they can introduce information about assumptions into the workshop discussion. Also, in our experience, some stakeholders are puzzled by questions asking whether a 'method' has had any effect; if complete naivety about methods is anticipated, the questioning can ask about the effects of 'how the workshop was run' (but then it's important to ask about both specific modelling activities and how the workshop was facilitated, so that the effects of the method and the process of application can be distinguished).

Above, we have discussed how our framework can be employed in single case study evaluations of the use of a systemic PSM. However, in making longer term comparisons of methods using data from multiple case studies, we make the assumption that the more cases are included, the more likely it is that the effects of particular contexts, purposes, etc., will be evened out. Therefore, the qualitative information discussed in the workshops mentioned above, relevant primarily to single case studies, can mostly be set aside in favour of statistical analyses of the questionnaire findings.

5. Strengths and limitations of the evaluation framework and questionnaire

As we see it, this new framework for the evaluation of systemic PSMs has two significant strengths. First, by encouraging the exploration of the context-purposes-methods-outcomes relationship in a particular intervention, and by explicitly recognising that the researcher becomes part of the situation that he or she intervenes in, our framework offers a more nuanced (but still reasonably parsimonious) set of concepts and guidelines to work with than many others in the literature. Second, it incorporates a questionnaire that can support both locally meaningful evaluations *and* longer-term comparisons between methods, thereby giving us the potential to move beyond the either/or debate that has characterised the literature in recent years.

Nevertheless, it is important to clarify some of the framework's limitations. In our view, the first two of these are more or less inevitable, and have to be managed as part of the evaluation process, while the final four indicate the need for further research. Only the first limitation concerns our framework as a whole: the rest relate solely to the use of the questionnaire for longer-term comparisons between methods:

Within the context of a specific use of a method in a single intervention, there is scope for the researcher to avoid unwelcome conclusions, for example by exaggerating the effect of an aspect of context that was outside his or her control, thereby missing shortfalls in either the method or his or her own skill set. To help manage this, three methodological devices have been built into our framework to bring evidence of bad news to the attention of evaluators, making avoidance more difficult than it might be if the evaluators were basing their conclusions on personal reflections alone. First, the use of a questionnaire ensures that participant

Table 1
Generic questions for adaptation and use in reflective workshops bringing together researchers and stakeholders.[a]

Aspect of the framework (Fig. 1)	Questions
Context	What key perspectives, values and assumptions were participants bringing in, and how did these affect discussion? Were there significant processes of marginalization or exclusion of people and/or issues? What organizations, institutions, economic conditions and ecological factors influenced the perspectives that people came in with? Did people feel enabled or constrained by wider systems, and what effects did this have?
Researcher identity	How was the researcher seen by themselves and others, and why?
Purposes	What openly expressed and hidden purposes did different people have for participating? Which purposes were met and not met, and what were the effects? Were there conflicting purposes (or people thinking others had hidden agendas), and what were the effects?
Researcher purposes	What purposes did the researcher have? Was there any conflict between the participants' and researcher's purposes? Did the participants trust the expressed purposes of the researcher?
Methods	What theoretical assumptions made by the methods might have been influential? What cultural norms did the methods assume, and how did these relate to the culture(s) of the participants? Did the process facilitate effective participation? Did the process help people to think systemically? (Different theoretical understandings of what it means to think systemically, such as appreciating other perspectives and getting a 'bigger picture' understanding, have informed some of the questions in the questionnaire)
Researcher skills and preferences	What preferences (for methodologies, methods and processes of application) did the researcher have, and what were the effects? What other skills, resources and competencies did the researcher bring in (or not), and what were the effects?
Outcomes	What plans, actions or changes were achieved? Have longer-term outcomes been achieved, and can these be linked back to the use of the method? (This can only be asked in the context of a longer-term follow up) How do the outcomes relate to people's purposes? What outcomes (positive and negative) were anticipated or unanticipated?
Researcher outcomes	What outcomes were achieved by the researcher? What was the fit between the researcher's outcomes and the outcomes for the participants and those experiencing wider effects?

[a] Information from the questionnaire results can inform answers to some of these questions.

voices are available. In particular, the answers to the open ended questions are likely to include the participants' own theories about shortcomings. Second, by offering guidelines for exploring the context that draw upon multiple paradigmatic perspectives, the risk of 'paradigm blindness' (interpreting the context in the same paradigmatic terms as the method, thereby missing insights that would be apparent from other perspectives) is minimised (also see Romm, 1996; Midgley, 2011). Third, by explicitly focusing attention on the researcher's identity, purposes, outcomes, skills and preferences, the framework confronts evaluators with some of the questions that they are most likely to want to avoid. If desired, and if feasible, researchers can go one step further to minimise avoidance by including participants on the evaluation team (preferably ones that are themselves open to the possibility of receiving bad news).

The second limitation we are aware of, applying to longer term comparisons of methods using the questionnaire, comes from the observation that there is a strong movement advocating methodological pluralism or 'multi-methodology' (e.g., Flood and Jackson, 1991; Jackson, 1991, 2000; Flood and Romm, 1996; Mingers and Gill, 1997; Midgley, 2000). At its most flexible, a pluralist practice may involve the integration of several previously distinct methods into a new whole, perhaps also incorporating the design of novel elements (Midgley, 2000). It will be much easier to compare standard sets of methods (e.g., those associated with discrete systems methodologies) than it will be to compare mixed methods, drawn from different methodologies, that have not been widely applied. The irony here is that the more flexible and responsive that systemic problem structuring becomes, the more difficult it will be to evaluate methods over the longer term in a manner that can control for contextual effects. We certainly would not want to see our desire for improved evaluations of methods to result in the stultification of pluralist practice. Rather, we suggest that it may be wiser to accept that this limitation will restrict what can be asked of longer term comparisons between methods, but it will not make them redundant. It will still be possible to compare the sets of methods associated with well known and widely applied methodologies, giving us evidence of their strengths and weaknesses in relation to the set of attributes that a representative range of methods possesses. It will also be possible to compare pluralist practice in general with the use of particular discrete approaches. Finally, some mixed methods, if applied in several applications, can also be compared with other sets of methods. There are a number of relatively popular mixes in the literature that will no doubt qualify for evaluation. When comparisons between mixed methods using the questionnaire data look like they will be unreliable because the sample size is too small, it should nevertheless still be possible to facilitate cross-case study learning, where possible bringing together two or more research teams to reflect on their practice using our framework (Fig. 1).

The third limitation is that we have not yet tested the questionnaire for validity and reliability. Rowe et al. (2005) discuss the substantial difficulties in doing this in the field because participants are often reluctant to fill in two or more questionnaires asking similar things (the usual approach to testing for validity being to compare with another questionnaire constructed for similar purposes). Indeed, in this case, testing for validity will be difficult because there are so few instruments available in the public domain (e.g., Halvorsen, 2001), and those that exist are geared to evaluating forms of public participation other than the use of systemic PSMs. Also, checking reliability is even more troublesome than a validity test because it involves getting participants to fill in the same questionnaire on two separate occasions. Generally speaking, the researcher only has access to participants on the day of a workshop. Our intention is to do some validity and reliability

testing in due course when a good comparative instrument can be identified and the testing can be added to an intervention without difficulty.

The fourth limitation we have identified concerns the inability of standard metrics, such as those to be found in sections two and three of our questionnaire, to pick up novelty: they can only evaluate against already established criteria. This is arguably one of the most significant limitations in terms of conducting longer-term research based on multiple case studies: it appears that, after around 20 years of relative stability in the number of systemic PSMs that are widely used in practice, systems/OR practitioners are now producing a new generation of methodologies and methods (Rosenhead and Mingers, 2004; Shaw et al., 2006; Franco et al., 2007), and it is important that the questionnaire does not go out of date. Our solution to this problem, which will need to be enacted as part of a longer term international research program, will be to undertake a review of the questionnaire after a set period of data collection. This period will need to be long enough to allow sufficient data to be gathered on the application of well established approaches. Periodic reviews of the questionnaire followed by new data collection should enable a balance to be struck between stability (to facilitate robust comparisons) and change (to keep the longer term comparisons open to novelty).

The fifth limitation is that our questionnaire does not currently allow the comparison of systemic PSMs and non-participative modelling methods. Although we would ideally like to extend our research to include the latter, it may not be feasible to integrate questions about both types of method into a single instrument. Our field testing suggests that we have already hit the upper limit for the number of questions people are willing to answer, so feasibility would depend on reducing the number of questions about systemic PSMs in order to allow others to be included.

The sixth and final limitation we face is that no one group of researchers will be able to collect sufficient data on its own to enable the robust, longer term comparison of methods. International collaboration will therefore be essential, and we have made a start in moving towards this by establishing collaborative arrangements with over 80 systems/OR practitioners in 22 countries who are willing to test our evaluation framework and questionnaire in practice.

6. Conclusions

In this paper we have offered a new framework for evaluating systemic problem structuring methods, focusing on the context-purposes-methods-outcomes relationship. This framework can be used in an emergent mode, and it asks researchers to view themselves as active contributors to the success or failure of a method-in-context. We have also reported on the development of a questionnaire to gather data from participants that can be of use in reflecting on the strengths and weaknesses of methods. The same data may be useful for both evaluations of methods in single case studies *and* longer term comparisons between methods using information from multiple cases. However, undertaking longer term comparisons will require a new, international research program, which is currently under development.

Acknowledgements

We would like to acknowledge the support of colleagues at the Institute of Environmental Science and Research (ESR) Ltd., New Zealand, who helped to field test the questionnaire in the context of their projects: Virginia Baker, Jan Gregor, Wendy Gregory, Maria Hepi, Miria Lange, Johanna Veth and Ann Winstanley. We would also like to recognise the contribution made by our international panel of experts on systemic methods: John Brocklesby (Victoria University of Wellington, New Zealand), José Córdoba (Royal Holloway University, UK), Amanda Gregory (University of Hull, UK), John Mingers (University of Kent, UK), Leroy White (University of Bristol, UK) and Jennifer Wilby (University of Hull, UK). Furthermore, we would like to express our appreciation for the comments and feedback received following presentations at the annual conferences of the International Society for the Systems Sciences (Tokyo, Japan, 2007); the Australia and New Zealand Systems Society (Perth, Australia, 2008); and the Hellenic Society for Systemic Studies (Athens, Greece, 2011). Finally, we wish to acknowledge funding from various sources that enabled the research reported in this paper: the Foundation for Research, Science and Technology, New Zealand (Contracts C03X0304 and C03X0305); the Colonial Foundation Trust, Australia; and the Ministry for Research, Science and Technology, New Zealand. However, the interpretations and opinions expressed in this paper are the authors' own.

Appendix A. Supplementary material

The evaluation questionnaire associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ejor.2013.01.047.

References

Adams, R., McCullough, A., 2003. The urban practitioner and participation in research within a streetwork context. Community, Work & Family 6, 269–287.
Alberts, D.J., 2007. Stakeholders or subject matter experts, who should be consulted? Energy Policy 35, 2336–2346.
Allsop, J., Taket, A., 2003. Evaluating user involvement in primary healthcare. International Journal of Healthcare Technology & Management 5, 34–44.
Alrøe, H.F., 2000. Science as systems learning: some reflections on the cognitive and communicational aspects of science. Cybernetics and Human Knowing 7, 57–78.
Baker, V., Gregory, W., Midgley, G., Veth, J., 2006. Ethical Implications and Social Impacts of Forensic DNA Technologies and Applications: Summary Report. Institute of Environmental Science and Research (ESR) Ltd., Christchurch.
Baker, V., Midgley, G., 2007. Review of the MoRST Roadmaps Exercise: Final Report. Confidential ESR Client Report. Institute of Environmental Science and Research (ESR) Ltd., Wellington.
Bateson, G., 1970. Form, substance, and difference. General Semantics Bulletin 37, 5–13.
Beierle, T.C., Cayford, J., 2002. Democracy in Practice: Public Participation in Environmental Decisions. RFF Press, Washington, DC.
Beierle, T.C., Konisky, D.M., 2000. Values, conflict, and trust in participatory environmental planning. Journal of Policy Analysis and Management 19, 587–602.
Berry, H., Bowman, S.R., Hernandez, R., Pratt, C., 2006. Evaluation tool for community development coalitions. Journal of Extension 44, http://www.joe.org/joe/2006december/tt2.shtml (accessed: 30.03.07).
Bjärås, G., Haglund, B.J.A., Rifkin, S.B., 1991. A new approach to community participation assessment. Health Promotion International 6, 199–206.
Branch, K.M., Bradbury, J.A., 2006. Comparison of DOE and Army advisory boards: application of a conceptual framework for evaluating public participation in environmental risk decision making. Policy Studies Journal 34, 723–753.
Brocklesby, J., 1997. Becoming multimethodology literate: an assessment of the cognitive difficulties of working across paradigms. In: Mingers, J., Gill, A. (Eds.), Multimethodology: The Theory and Practice of Combining Management Science Methodologies. Wiley, Chichester.
Brocklesby, J., 2009. Ethics beyond the model: how social dynamics can interfere with ethical practice in operational research / management science. Omega, the International Journal of Management Science 37, 1073–1082.
Bryant, J.W., Darwin, J.A., 2004. Exploring inter-organisational relationships in the health service: an immersive drama approach. European Journal of Operational Research 152, 655–666.
Burns, D., 2007. Systemic Action Research: A Strategy for Whole System Change. Policy Press, Bristol.
Buysse, V., Wesley, P., Skinner, D., 1999. Community development approaches for early intervention. Topics in Early Childhood Special Education 19, 236–243.
Cavana, R.Y., Delahaye, B.L., Sekaran, U., 2001. Applied Business Research: Qualitative and Quantitative Methods. Wiley, Brisbane.
Champion, D., Wilson, J.M., 2010. The impact of contingency factors on validation of problem structuring methods. Journal of the Operational Research Society 61, 1420–1431.
Charnley, S., Engelbert, B., 2005. Evaluating public participation in environmental decision-making: EPA's superfund community involvement program. Journal of Environmental Management 77, 165–182.

Checkland, P., 1981. Systems Thinking. Systems Practice. Wiley, Chichester.
Checkland, P.B., Forbes, P., Martin, S., 1990. Techniques in soft systems practice. Part 3: Monitoring and control in conceptual models and in evaluation studies. Journal of Applied Systems Analysis 17, 29–37.
Checkland, P., Scholes, J., 1990. Soft Systems Methodology in Action. Wiley, Chichester.
Checkland, P., Poulter, J., 2006. Learning for Action. Wiley, Chichester.
Chess, C., Purcell, K., 1999. Public participation and the environment: do we know what works? Environmental Science & Technology 33, 2685–2692.
Churchman, C.W., 1970. Operations research as a profession. Management Science 17, B37–53.
Clayton, A.M.H., Radcliffe, N.J., 1996. Sustainability: A Systems Approach. Earthscan, London.
Cole, M., 2006. Evaluating the impact of community appraisals: some lessons from South-West England. Policy & Politics 34, 51–68.
Connell, N.A.D., 2001. Evaluating soft OR: some reflections on an apparently 'unsuccessful' implementation using a soft systems methodology (SSM) based approach. Journal of the Operational Research Society 52, 150–160.
Delaney, M.M., Foroughi, A., Perkins, W.C., 1997. An empirical study of the efficacy of a computerized negotiation support system (NSS). Decision Support Systems 20, 185–197.
De Vreede, G., Dickson, G., 2000. Using GSS to design organizational processes and information systems: An action research study on collaborative business engineering. Group Decision and Negotiation 9, 161–183.
Douglas, M., 1986. How Institutions Think. Routledge and Kegan Paul, London.
Duignan, P., Casswell, S., 1989. Evaluating community development programs for health promotion: problems illustrated by a New Zealand example. Community Health Studies 13, 74–81.
Duram, L.A., Brown, K.G., 1999. Assessing public participation in U.S. watershed planning initiatives. Society & Natural Resources 12, 455–467.
Eden, C., 1995. On evaluating the performance of 'wide-band' GDSS. European Journal of Operational Research 81, 302–311.
Eden, C., Ackermann, F., 1996. ''Horses for courses'': a stakeholder approach to the evaluation of GDSSs. Group Decision and Negotiation 5, 501–519.
Eden, C., Ackermann, F., 2006. Where next for problem structuring methods. Journal of the Operational Research Society 57, 766–768.
Eden, C., Sims, D., 1979. On the nature of problems in consulting practice. Omega 7, 119–127.
Entwistle, V., Buchan, H., Coulter, A., Jadad, A., 1999. Towards constructive innovation and rigorous evaluation: a new series on methods for promoting and evaluating participation. Health Expectations 2, 75–77.
Er, M.C., Ng, A.C., 1995. The anonymity and proximity factors in group decision support systems. Decision Support Systems 14, 75–83.
Fan, S., Shen, Q., Lin, G., 2007. Comparative study of idea generation between traditional value management workshops and GDSS-supported workshops. Journal of Construction Engineering and Management 133, 816–825.
Fjermestad, J., 2004. An analysis of communication mode in group support systems research. Decision Support Systems 37, 239–263.
Fjermestad, J., Hiltz, S., 1998. An assessment of group support systems experimental research: methodology and results. Journal of Management Information Systems 15, 7–149.
Flood, R.L., 1995. Solving Problem Solving. Wiley, Chichester.
Flood, R.L., Jackson, M.C. (Eds.), 1991. Critical Systems Thinking: Directed Readings. Wiley, Chichester.
Flood, R.L., Romm, N.R.A. (Eds.), 1996. Critical Systems Thinking: Current Research and Practice. Plenum, New York.
Forrester, J.W., 1969. Principles of Systems. Wright-Allen Press, Cambridge, MA.
Franco, L.A., 2006. Forms of conversation and problem structuring methods: a conceptual development. Journal of the Operational Research Society 57, 813–821.
Franco, L.A., 2007. Assessing the impact of problem structuring methods in multi-organizational settings: an empirical investigation. Journal of the Operational Research Society 58, 760–768.
Franco, L.A., Shaw, D., Westcombe, M., 2007. Taking problem structuring methods forward. Journal of the Operational Research Society 58, 545–546.
Gallupe, R.B., Dennis, A.R., Cooper, W.H., Valacich, J.S., Bastianutti, L.M., Nunamaker, J.F., 1992. Electronic brainstorming and group size. Academy of Management Journal 35, 350–369.
Gopal, A., Prasad, P., 2000. Understanding GDSS in symbolic context: shifting the focus from technology to interaction. MIS Quarterly 24, 509–546.
Halvorsen, K.E., 2001. Assessing public participation techniques for comfort, convenience, satisfaction, and deliberation. Environmental Management 28, 179–186.
Hepi, M., Foote, J., Ahuriri-Driscoll, A., 2008. Guidelines for Developing Resource Care Evaluation Criteria and Methods. Institute of Environmental Science and Research (ESR) Ltd., Christchurch.
Hjortsø, C.N., 2004. Enhancing public participation in natural resource management using soft OR: an application of strategic option development and analysis in tactical forest planning. European Journal of Operational Research 152, 667–683.
Ho, C.H., 1997. A Critical Process for the Evaluation of Methodology. Ph.D. Thesis, University of Hull.
Jackson, M.C., 1991. Systems Methodology for the Management Sciences. Plenum, New York.
Jackson, M.C., 2000. Systems Approaches to Management. Plenum/Kluwer, New York.
Jackson, M.C., 2006. Beyond problem structuring methods: reinventing the future of OR/MS. Journal of the Operational Research Society 57, 868–878.
Jackson, M.C., Keys, P., 1984. Towards a system of systems methodologies. Journal of the Operational Research Society 35, 473–486.
Jenkins, N.T., Bennett, M.I.J., 1999. Toward an empowerment zone evaluation. Economic Development Quarterly 13, 23–28.
Joldersma, C., Roelofs, E., 2004. The impact of soft OR-methods on problem structuring. European Journal of Operational Research 152, 696–708.
Kelly, K., Van Vlaenderen, H., 1995. Evaluating participation processes in community development. Evaluation & Program Planning 18, 371–383.
Keys, P., 1994. Understanding the Process of Operational Research: Collected Readings. Wiley, Chichester.
Li, X., Zheng, H., 1995. Study on general systems methodology. In: Midgley, G., Wilby, J. (Eds.), Systems Methodology: Possibilities for Cross-Cultural Learning and Integration. Centre for Systems Studies, Hull.
Luhmann, N., 1986. Ecological Communication. University of Chicago Press, Chicago.
Maani, K.E., Cavana, R.Y., 2007. Systems Thinking, System Dynamics, second ed. Pearson New Zealand, Auckland.
Margerum, R.D., 2002. Collaborative planning: building consensus and building a distinct model for practice. Journal of Planning Education & Research 21, 237–253.
Masozera, M.K., Alavalapati, J.R.R., Jacobson, S.K., Shrestha, R.K., 2006. Assessing the suitability of community-based management for the Nyungwe forest reserve, Rwanda. Forest Policy & Economics 8, 206–216.
McAllister, K., 1999. Understanding Participation: Monitoring and Evaluating Process, Outputs and Outcomes. Working Paper 2, IDRC, Ottawa.
McCartt, A.T., Rohrbaugh, J., 1995. Managerial openness to change and the introduction of GDSS: explaining initial success and failure in decision conferencing. Organization Science 6, 569–584.
McGurk, B., Sinclair, A.J., Diduck, A., 2006. An assessment of stakeholder advisory committees in forest management: case studies from Manitoba, Canada. Society & Natural Resources 19, 809–826.
McKay, J., 1998. Using cognitive mapping to achieve shared understanding in information requirements determination. Australian Computer Journal 30, 139–145.
Midgley, G., 1994. Ecology and the poverty of humanism: a critical systems perspective. Systems Research 11, 67–76.
Midgley, G., 2000. Systemic Intervention: Philosophy, Methodology, and Practice. Plenum/Kluwer, New York.
Midgley, G. (Ed.), 2003. Systems Thinking, vols. I to IV. Sage, London.
Midgley, G., 2011. Theoretical pluralism in systemic action research. Systemic Practice and Action Research 24, 1–15.
Midgley, G., Ahuriri-Driscoll, A., Baker, V., Foote, J., Hepi, M., Taimona, H., Rogers-Koroheke, M., Gregor, J., Gregory, W., Lange, M., Veth, J., Winstanley, A., Wood, D., 2007. Practitioner identity in systemic intervention: reflections on the promotion of environmental health through Māori community development. Systems Research and Behavioral Science 24, 233–247.
Midgley, G., Winstanley, A., Gregory, W., Foote, J., 2005. Scoping the Potential Uses of Systems Thinking in Developing Policy on Illicit Drugs. Drug Policy Modelling Project Research Memorandum #13. Turning Point, Melbourne.
Mingers, J.C., 1997. Towards critical pluralism. In: Mingers, J., Gill, A. (Eds.), Multimethodology: The Theory and Practice of Combining Management Science Methodologies. Wiley, Chichester.
Mingers, J., Brocklesby, J., 1997. Multimethodology: towards a framework for mixing methodologies. Omega, the International Journal of Management Science, 25, 489–509.
Mingers, J., Gill, A. (Eds.), 1997. Multimethodology: The Theory and Practice of Combining Management Science Methodologies. Wiley, Chichester.
Mingers, J., Rosenhead, J., 2004. Problem structuring methods in action. European Journal of Operational Research 152, 530–554.
Montazemi, A.R., Wang, F., Nainar, S.M.K., Bart, C.K., 1996. On the effectiveness of decisional guidance. Decision Support Systems 18, 181–198.
Morgan, L.M., 2001. Community participation in health: perpetual allure, persistent challenge. Health Policy & Planning 16, 221–230.
Murphy-Berman, V., Schnoes, C., Chambers, J.M., 2000. An early stage evaluation model for assessing the effectiveness of comprehensive community initiatives: three case studies in Nebraska. Evaluation & Program Planning 23, 157–163.
Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D.R., George, J.F., 1991. Electronic meeting systems to support group work. Communications of the ACM 34, 43–61.
Ong, B.N., 2000. Assessing the context for partnerships between communities and the National Health Service in England. Critical Public Health 10, 343–351.
Paterson, J., Teubner, G., 1998. Changing maps: empirical legal autopoiesis. Social and Legal Studies 7, 451–486.
Pettigrew, A.M., 1987. Context and action in the transformation of the firm. Journal of Management Studies 24, 649–670.
Phahlamohlaka, J., Friend, J., 2004. Community planning for rural education in South Africa. European Journal of Operational Research 152, 684–695.
Pinsonneault, A., Barki, H., Gallupe, R.B., Hoppen, M., 1999. Electronic brainstorm: the illusion of productivity. Information Systems Research 10, 110–132.
Pinsonneault, A., Kraemer, K.L., 1990. The effects of electronic meetings on group processes and outcomes: an assessment of the empirical research. European Journal of Operational Research 46, 143–161.

Romm, N.R.A., 1996. Inquiry-and-intervention in systems planning: probing methodological rationalities. World Futures 47, 25–36.
Rosenhead, J., 1989. Rational Analysis for a Problematic World. Wiley, Chichester.
Rosenhead, J., 2006. Past, present and future of problem structuring methods. Journal of the Operational Research Society 57, 759–765.
Rosenhead, J., Mingers, J., 2001. Rational Analysis for a Problematic World Revisited. Wiley, Chichester.
Rosenhead, J., Mingers, J., 2004. Problem structuring methods in action. European Journal of Operational Research 152, 530–554.
Rouwette, E.A.J.A., 2011. Facilitated modelling in strategy development: measuring the impact on communication, consensus and commitment. Journal of the Operational Research Society 62, 879–887.
Rouwette, E., Bastings, I., Blokker, H., 2011. A comparison of group model building and strategic options development and analysis. Group Decision and Negotiation 20, 781–803.
Rouwette, E.A.J.A., Vennix, J.A.M., Felling, A.J.A., 2009. On evaluating the performance of problem structuring methods: an attempt at formulating a conceptual model. Group Decision and Negotiation 18, 567–587.
Rowe, G., Frewer, L.J., 2000. Public participation methods: a framework for evaluation. Science, Technology & Human Values 25, 3–29.
Rowe, G., Frewer, L.J., 2004. Evaluating public participation exercises: a research agenda. Science, Technology & Human Values 29, 512–556.
Rowe, G., Horlick-Jones, T., Walls, J., Pidgeon, N., 2005. Difficulties in evaluating public engagement initiatives: reflections on an evaluation of the UK GM Nation? public debate about transgenic crops. Public Understanding of Science 14, 331–352.
Rowe, G., Marsh, R., Frewer, L.J., 2004. Evaluation of a deliberative conference. Science, Technology & Human Values 29, 88–121.
Shaw, D., 2003. Evaluating electronic workshops through analysing the 'brainstormed' ideas. Journal of the Operational Research Society 54, 692–705.
Shaw, D., Franco, A., Westcombe, M., 2006. Problem structuring methods: new directions in a problematic world. Journal of the Operational Research Society 57, 757–758.
Shaw, I., 1999. Qualitative Evaluation. Sage, London.
Shen, C-Y., Midgley, G., 2007. Toward a buddhist systems methodology 1: comparisons between buddhism and systems theory. Systemic Practice and Action Research 20, 167–194.
Sieber, R., 2006. Public participation geographic information systems: a literature review and framework. Annals of the Association of American Geographers 96, 491–507.
Smith, L.T. (Ed.), 1999. Decolonizing Methodologies: Research and Indigenous Peoples. Zed Books, London.
Sørensen, L., Vidal, R., Engström, E., 2004. Using soft OR in a small company – the case of Kirby. European Journal of Operational Research 152, 555–570.
Spash, C.L., 1997. Ethics and environmental attitudes with implications for economic valuation. Journal of Environmental Management 50, 403–416.
Sykes, C., Goodwin, W., 2007. Assessing patient, carer and public involvement in health care. Quality in Primary Care 15, 45–52.
Taket, A., White, L., 2000. Partnership and Participation: Decision-making in the Multi-Agency Setting. Wiley, Chichester.
Tuler, S., Webler, T., Finson, R., 2005. Competing perspectives on public involvement: planning for risk characterization and risk communication about radiological contamination from a national laboratory. Health, Risk & Society 7, 247–266.
Ulrich, W., 1994. Critical Heuristics of Social Planning: A New Approach to Practical Philosophy. Wiley, Chichester.
Valacich, J.S., Schwenk, C., 1995a. Devil's advocacy and dialectical inquiry effects on face-to-face and computer-mediated group decision making. Organizational Behavior and Human Decision Processes 63, 158–173.
Valacich, J.S., Schwenk, C., 1995b. Structuring conflict in individual, face-to-face, and computer-mediated group decision making: carping versus objective devil's advocacy. Decision Sciences 26, 369–393.
Vennix, J.A.M., 1996. Group Model Building. Wiley, Chichester.
Warburton, D., Wilson, R., Rainbow, E., 2007. Making a Difference: A Guide to Evaluating Public Participation in Central Government. Involve, London, http://www.involve.org.uk/evaluation (accessed: 30.05.07).
White, L., 2006. Evaluating problem-structuring methods: developing an approach to show the value and effectiveness of PSMs. Journal of the Operational Research Society 57, 842–855.
Winstanley, A., Baker, V., Foote, J., Gregor, J., Gregory, W., Hepi, M., Midgley, G., 2005. Water in the Waimea Basin: Community Values and Water Management Options. Institute of Environmental Science and Research (ESR) Ltd., Christchurch.
Yearley, S., 2006. Bridging the science-policy divide in urban air-quality management: evaluating ways to make models more robust through public engagement. Environment and Planning C 24, 701–714.
Zhang, J., Smith, R., Watson, R.B., 1997. Towards computer support of the soft systems methodology: an evaluation of the functionality and usability of an SSM toolkit. European Journal of Information Systems 6, 129–139.
Zhu, Z., 2000. Dealing with a differentiated whole: the philosophy of the WSR approach. Systemic Practice and Action Research 13, 21–57.