Group model building effectiveness: a review of assessment studies†

Etiënne A. J. A. Rouwette,* Jac A. M. Vennix and Theo van Mullekom

Abstract
Over the last decades system dynamicists have experimented with approaches to achieve more involvement of their clients in the model building process. As a result the number of reports in the literature on the use of system dynamics as an organizational intervention tool has increased dramatically. From the literature we have identified 107 cases that provide details on the modeling process and the assessment of results. The cases show a wide variety in the way the interventions are reported and assessed. From a research point of view this is clearly an undesirable state of affairs. This article reports on a meta analysis of findings of these studies and tries to provide an overview of outcome studies on system dynamics interventions. It attempts to draw some preliminary conclusions on the effectiveness of system dynamics modeling interventions, to clarify differences in definitions of outcomes, and to provide guidelines for more standardized assessments and reports. Rather than remaining in the stage of single case descriptions, the latter will enable the accumulation of research results in the future, a prerequisite for institutional learning within the system dynamics community. Copyright © 2002 John Wiley & Sons, Ltd.


Background
Since the inception of the field, implementation of results and system improvement have been primary goals of system dynamics interventions. Client involvement has often been considered a useful way to create model ownership and to ensure implementation. Over the last four decades, more and more people have started to use system dynamics, and more and more practitioners and consultants have started to involve clients in system dynamics model building projects. This has been shown by, amongst often effects, the number of published case studies, which seems to have increased exponentially over recent decades. On the one hand, this is a encouraging development, because it generates more insights into what is done in the field. On the other hand, this state of affairs is also very undesirable, because we do not seem to be able to go beyond single case descriptions and to come up with more general conclusions on the effectiveness of system dynamics interventions. So far, the system dynamics community has not produced systematic answers to important questions such as:
His research interests focus on problem structuring methodology and empirical assessment of problem structuring interventions in organizations. More in particular on how these methodologies (primarily group model building and scenario construction) can assist in increasing the learning potential of organizations.

Theo van Mullekom is a junior researcher at the Methodology Section, Faculty of Management Sciences at Nijmegen University. His research centers on using scenario building as an intervention tool for solving complex problems in organizations.

- What practices and scripts do practitioners use in systems thinking interventions?
- Are there different types of system dynamics interventions and, if so, what kinds of situations demand which type of intervention?
- Last but not least: when is a system dynamics intervention (un)successful and what are the determining factors of success?

Although individual practitioners and consultants may have learned a lot from conducting these interventions, our guess is that the system dynamics community as a whole has not. Apart from some scattered discussions on a number of related topics (e.g., eliciting knowledge from teams, approaches to client involvement, the use of management flight simulators in interventions, the appropriate size of models, the role of qualitative models), we have primarily produced a huge number of, largely incomparable, case descriptions of (un)successful model building projects in organizations. And, although a well-conducted case study may provide in-depth insights into a particular intervention process (for examples see Lane 1993; Vennix 1995), and may help to generate hypotheses for research, these case studies are limited because they do not allow the production of more general conclusions. This is primarily due to the fact that each of these cases emphasizes different elements and employs unique ways to establish the effectiveness of the intervention. Stated differently, we do not have a research program in which we systematically collect identical data over cases, in order to generate and test hypotheses about the effectiveness of our system dynamics interventions.

In an effort to bring this research program one step closer, this paper focuses on these case descriptions to which, for reasons of simplicity, we will refer to as group model building interventions. Using these case descriptions as our primary data, a meta-analysis was conducted with three purposes in mind: first, to provide a review of these case descriptions and to report on their most important characteristics; second, to generate some preliminary insights with regard to the effectiveness of group model building interventions; third, to produce more identical professional standards to (a) report on group model building interventions and their characteristics in case descriptions, and (b) assess the effectiveness of group model building interventions. Once established, this analysis will make it possible to build a database from which more general conclusions can be drawn in the future, which will in turn foster the learning process for system dynamics practitioners and consultants.

The remainder of this article is organized as follows. First, the case selection process and the construction of the database is discussed. Next, a number of preliminary results will be presented on both case characteristics (e.g., type of problem tackled, number of people involved, duration of intervention) and evaluation results (e.g., Did learning occur? Were policies changed? Did system improvement occur?). An attempt is then made to find patterns in the cases by looking at specific combinations of the context of the intervention,
the mechanism that was used in the intervention, and the outcomes that were produced. In a final section, more rigorous guidelines for reports on group model building interventions will be discussed as a first step towards a systematic research program.

Method: selection of cases and construction of the database

Selection of cases

A literature search was conducted for publications on group model building in the System Dynamics Conference Proceedings (1981 to 1999), *System Dynamics Review* (1985 to Winter 1999), and publications on system dynamics by Productivity Press. We also placed a request for additional publications to the system dynamics discussion list. Each source was reviewed by one of the authors, who subsequently selected relevant publications. Publications were deemed relevant if (a) they described a system dynamics modeling project involving a client team in at least the stage of conceptualization, and (b) any empirical results on its effectiveness were described. In total, 86 publications describing 107 cases were catalogued (see Appendix 1).

The procedure for gathering cases may have produced a bias, either because cases were erroneously excluded, or because cases were reported in other sources than the ones used here. In addition, we have to point out that an additional bias may have been introduced because, for a variety of reasons, not all group model building interventions are published. For example the client may prevent publication because of proprietary rights, something one would expect to happen more often in profit organizations. Two factors plead against this potential bias: first, the emergence of a special consultancy track at System Dynamics Conferences, in which mainly cases in profit organisations are discussed; second, the fact that a total of 60 percent of the cases in this meta-evaluation were conducted in a profit setting.

Another reason for not publishing a case may be that the person conducting the intervention is reluctant to publish because the intervention was largely unsuccessful. We have indeed found few descriptions of unsuccessful interventions. It could thus be that unsuccessful cases are under-reported. This would be an unfortunate state of affairs, because unsuccessful interventions may be the ones we can learn most from. On the other hand, we do have to point out that it may also be that the overall positive character of cases is not surprising, since the group modeling interventions employed here have been developed in a design–test–redesign fashion in a number of iterations. Early publications (e.g. Randers 1977) report on problems on the basis of which future interventions were adapted.

In short, we do not know to what extent systematic biases exist and we thus have to be careful when drawing conclusions. On the other hand, we do have
107 case descriptions, which gives us the opportunity to try to go beyond the insights generated in the single case descriptions, as a first step towards a more rigorous research program in this area.

Data gathered

The data that were gathered on each modeling project (case), and recorded in a database, can be ordered in five general categories: background characteristics, organizational characteristics; problem to be modeled, characteristics of the intervention; and assessment of modeling impacts. Each of these will be discussed in more detail below.

BACKGROUND CHARACTERISTICS Background characteristics of each publication that were recorded are: the authors; title; source (journal or proceedings); and date of publication.

ORGANIZATIONAL CHARACTERISTICS With regard to organization, we stored the following characteristics: sort (profit, non-profit or governmental); sector (e.g., energy or financial services); name and size of the client organization; geographical location; and name of the consultant organization.

PROBLEM CHARACTERISTICS Problem elements that were deemed important are motive (for starting the intervention), the research question(s) on which the modeling effort focused, and its type (i.e. exploratory, descriptive, explanatory, or prescriptive). Furthermore, we recorded the importance of the problem as judged by the participants, and whether or not implementation of results was expected from the outset of the project.3

INTERVENTION CHARACTERISTICS First, we recorded a more or less open description of the different techniques employed for building the model (e.g. NGT, Hexagons). The model itself was characterized as qualitative or quantitative, by size, and whether a preliminary model was used or not. We also recorded which phases were followed and in what phases the client actually participated. In addition, we recorded the number and functions of participants involved. The database also contains the sources of information for building the model (as well as persons or groups, these could be documents, real-life situations, or models—system dynamics or otherwise), the software used, other materials, and the total time span of the intervention (i.e., from initiation of contact to close of project).

ASSESSMENT OF IMPACT OF INTERVENTION Naturally, the most elaborate category in the database is the assessment of modeling results. Both the research design
of the evaluation (e.g., experiment, survey, case study) and the data collection methods (i.e., individual/group interviews, content analysis, questionnaires, observations) were recorded. We also recorded: (a) the subjects (number and function of persons involved in the evaluation); (b) the researchers conducting the evaluation; as well as (c) the time span covered. Record was also kept of the way in which the client received feedback on the results of the modeling project (e.g. written report, oral presentation).

Finally, an entry includes the outcome variables that were reported. Here quotations from the original text were used as much as possible. The quotes from the original text were summarized by employing the keywords in Table 1. The keywords in the table are used as a representation of the most important results of group model building (see, for example, Randers 1977; Richmond 1987; 1997; Lane 1992; Huz et al. 1997a,b; Vennix 1996). Considering for example the individual level, we can identify the following four potential outcomes. First, several authors consider clients’ reactions to the model or other elements of the intervention as an important outcome, e.g., ownership of the model or trust in the modeler (Lane 1992). Second, most modelers see learning as an important outcome; clients are encouraged to take a broader perspective on the problem modeled. Third, commitment to results and the resulting changes in behavior are also widely agreed on as important potential outcomes of client involvement. Similarly, at the group level potential outcomes relate to: better exchange of viewpoints; more alignment; and shared language.

At the organizational level outcomes can be system changes (are things done differently after the intervention?) or results of system changes (system efficiency or effectiveness). Finally, in the category “method” we recorded outcomes related to the further use of the system dynamics methodology and the intervention’s efficiency compared to traditional methods (e.g., meetings).

An entry was made in the categories given in Table 1 only when authors reported on this aspect. If for example nothing is said about insight, this category is left open. In other words [no insight] is only recorded if the author explicitly states that no learning has occurred. The decision to use a dichotomized score follows from the general nature of the studies collected. A large number of studies revert to rich descriptions of the phases and conclusions reached in the modeling process. Only a small subset use quantitative measures of results such as Likert-type questionnaires. If available, these quantitative results were recorded separately after each keyword. In the analysis of results, qualitative and quantitative results will be contrasted.

To fill the database, the following coding procedure was used. First, one author read a complete publication and entered his scores in the relevant fields described above. Second, 36 publications were reread by another author and entries were checked. A comparison of the entries of different authors revealed no significant differences.

Now that this section has described the selection of cases and the construction of the database, the next sections will focus on the results of
Table 1. Keywords used for scoring the outcomes of group model building cases

<table>
<thead>
<tr>
<th>Category</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>[positive reaction] or [negative reaction]: personal evaluations of the intervention or model (e.g., ownership, discomfort, trust) [insight] or [no insight]: learning [commitment] or [no commitment]: a decision or commitment to results [behavior] or [no behavior]: changes in individual behavior or implementation of conclusions</td>
</tr>
<tr>
<td>Group</td>
<td>[communication] or [no communication]: exchange of viewpoints [consensus] or [no consensus]: a shared view of the problem or actions [shared language] or [no shared language]: understanding of other participants</td>
</tr>
<tr>
<td>Organization</td>
<td>[system changes] or [no system changes]: organizational or physical changes (e.g., production lines, personnel policies) [positive results] or [negative results]: results of system changes (e.g., for profit or morale)</td>
</tr>
<tr>
<td>Method</td>
<td>[further use] or [no further use]: further use of system dynamics methods [efficiency] or [no efficiency]: efficiency of the method compared to traditional methods (e.g., a meeting)</td>
</tr>
</tbody>
</table>

our meta analysis, starting with a number of general characteristics of the cases.

**General characteristics of the cases**

**Background characteristics**

The first empirical study into the effects of client involvement in system dynamics modeling dates back to 1961. From our literature review it appears that only three other cases were published before 1970. In the 1970s, a total of four cases appeared, growing to 16 in the 1980s. From 1990 to 1999 between three and 13 cases appeared in print each year, indicating a fast growth in publications on group model building interventions.

**Organizational characteristics**

About 60 percent of all group modeling studies were conducted in profit organizations, 20 percent in non-profit settings, and 17 percent in governmental institutions. The size of client companies ranges from a few employees to (divisions of) large multinationals with revenues in the hundreds of millions of dollars. Table 2 provides more detail on the organizations where the interventions took place.

**Problem characteristics**

Problems that are not perceived as particularly important by participants are mostly modeled in a training or demonstration exercise and in most cases...
Table 2. Background of client organizations

<table>
<thead>
<tr>
<th>Sort of organisation</th>
<th>Number of organisations</th>
<th>Field</th>
<th>Number of organisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profit</td>
<td>65</td>
<td>Production</td>
<td>37 (oil 14; electronics 5; chemical 4; transport and vehicles 5; other 9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Services</td>
<td>25 (insurance 9; software 7; finance 3; other 6)</td>
</tr>
<tr>
<td>Non-profit</td>
<td>21</td>
<td>Distribution</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>University</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary school</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Energy</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Defense research</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other research</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K-12</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Charity</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Broadcasting</td>
<td>1</td>
</tr>
<tr>
<td>Governmental</td>
<td>18</td>
<td>National</td>
<td>9 (healthcare 1; transport 4; defence 2; development issues 1, forestry 1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>State</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>County</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>City</td>
<td>1</td>
</tr>
<tr>
<td>Mixed</td>
<td>3</td>
<td>Inter-organizational co-operations</td>
<td></td>
</tr>
</tbody>
</table>

There is no expectation of implementation of results. These projects are either initiated by champions working within an organization (who have recently come into contact with system dynamics tools) or by internal consultants suggesting these as tools for continuous improvement. One case started out as quite urgent, addressing a matter of considerable importance to the client organization, but became less urgent as a more serious matter (a merger) developed in another place in the client organization (Verburgh 1994).

A modeling project was listed as prescriptive if its purpose from the outset was to identify actions to steer system behavior in the preferred direction. There is a subset of group modeling cases that are focused on implementation, but are not expected to come up with concrete actions to alleviate the problem. These projects:

- aim at discovering relevant changes in the organizational environment (scenario studies by Genta et al. 1994; Morecroft and Van der Heijden 1992);
- explore policy impacts (Rohrbaugh et al. 1997; Akkermans 1995a,b: case banking; Royston et al. 1999; Delauzun and Mollona 1999; Morecroft, Lane and Viita 1991; Covert-Weiss, Clark and Odence 1998); or
- are intended as a pilot for assessing the fit of the method to a specific problem field (Cavana et al. 1999).
As can be expected, all projects aim to explain situations or developments over time. System dynamics is sometimes credited for its free format, in which models are usually started with a “blank paper” instead of pre-fixed notions on elements that have to be included (Coyle 1998). However, only five studies can be said to have an explicit exploratory orientation from the outset. Most studies start from specific hypotheses on the causes of problematic behavior and add additional structure when needed.

**Intervention characteristics**

**Elements and scripts** There exist a wide variety of group modeling techniques. The empirical studies gathered here are clear reflections of this. In building a model, participants perform three types of cognitive tasks (Vennix et al. 1992): elicitation of information, exploring courses of action or convergent tasks, and evaluation. From the modeling studies it appears that individual techniques such as interviews, cognitive mapping, nominal group technique, or workbooks are used to support the elicitation phase. Alternatively, elicitation of information is done in small subgroups (Andersen and Richardson 1997). The elicitation phase may be started after there is agreement on the problem to be addressed, but it can also consist of discussing and adaptation of a preliminary model. This pre-made model can be qualitative (e.g., an archetype in causal loop format) or quantitative, ranging from several variables and loops to a model of substantial size (e.g., Verburgh 1994). In the latter case, the group modeling part is limited to commenting on an already existing structure, and the difference from using a management flight simulator becomes small. A preliminary model was used in about one quarter (23) of the studies. In 20 studies the preliminary model was quantitative, in three qualitative. Contrary to what might be expected, there is no difference in learning or commitment between projects starting from scratch or using a preliminary model.

When it comes to convergent tasks, participants are asked to choose between alternative problem formulations, model structures, and policy options. These require the input and confrontation of opinions of the group of participants as a whole (Andersen and Richardson 1997; Vennix 1996). This phase mostly takes the form of a face-to-face discussion, although the Delphi method (Vennix et al. 1990; Hendrikx 1998) and GroupSystems (Rouwette, Vennix and Thijssen 2000) are also used.

The evaluation phase also requires the group as a whole to discuss and agree on issues, although individuals and subgroups are used to prioritize issues (Andersen and Richardson 1997).

In summary, a wide variety of approaches are used to involve the client in the model building process. Since descriptions of the interventions are far from complete, it is not clear how each modeler scripts his/her intervention. And only when “... we can comprehensively classify the diversity of our
practice, we can begin to probe more deeply into what matters in that practice” (Andersen, Richardson and Vennix 1997: 191–192).

**Number of participants and time duration** Projects also differ considerably with regard to the number of people involved and the time duration of the project. The number of participants involved in face-to-face interaction is mostly between five and 12, and seldom larger than around 20. If more people are involved, most work is done in subgroups that meet at regular intervals to present findings to each other. In some cases, groups as large as 30 to 160 participants work in subgroups using tools such as hexagon brainstorming, GroupSystems, and management flight simulators.

Generally participants are line managers. Members of staff or other experts participate in some projects, and a minority of projects are done with students.

The time between start of the project and handover of final results varies from two days to five years. Of the 66 studies providing detail on duration, about one half were completed within three months, and two out of three in six months. Most projects take the form of two to four workshops, with intermediate feedback and reports, for example in the form of a workbook. Workshops may be an intensive full-day meeting or consist of two to three hours of model building. The hours the client is involved in building the model are specified in only a few studies. Time investment seems to vary between 12 and 25 hours. An exception is the study by Hines and Johnson (1994), who involve participants in 12 full-day sessions over 12 weeks.

**Involvement in stages of the model building process** This raises the issue of the phases of model building a client is involved in. In 85 studies a fully quantified model was used, while in 22 studies a qualitative model was built. Although some papers discuss the elicitation of variables and the construction of a conceptual model, the literature offers few guidelines for the involvement of the client in the quantitative stages of modeling. Morecroft (1992: 13) proposes to use “friendly algebra” and Ford and Sterman (1998) use a stepwise approach to have experts estimate graph functions (see also Andersen and Richardson 1997; Richmond 1987; 1997).

Of a total of 85 quantitative model building projects, 69 focus on conclusions that can be implemented; 56 of these explicitly mention client involvement in the formalization phase. In three projects the formal model was discussed with the client. Participants in 12 studies contributed to the formalization phase by estimating parameters, by sketching variables over time, or by other forms of data gathering. In a small number of cases, members of the client organization possessed modeling expertise themselves and in effect built the complete model.

In addition to the client, other information sources such as documents and observation of real life situations were used in a couple of cases.
Size of the model. There is some discussion in the literature about the appropriate lower and upper bounds on the size of system dynamics models. Senge (1987: 875) discusses the benefits of very simple formal models “involving only one stock variable and virtually no significant feedback loops” in the direct interaction with clients. Lyneis (1999: 45) feels that a model “would probably need a minimum of several feedback loops and 20–30 equations” but also states that small Pugh–Roberts models contain 200–400 equations. He assumes that the appropriate size partly depends on the experience of the modeling team. With more experience, insight can be gained more quickly from large models. In addition, the different lower bounds mentioned by different authors may be due to the fact that Senge refers to models used in direct interaction with clients, while Lyneis seems to discuss the use of models for giving insight to an experienced modeling team. As an upper bound, Morecroft (1985: 16) suggests 100–200 equations and Hines and Johnson (1994) 200–400 equations. From our database it becomes clear that the size of models built falls somewhere between an upper range of several thousand (two models), and a lower range of five to 19 variables (six models). Most models are either 20 to 49 variables (23 studies) or 50 to 199 (18 studies). Another 10 models are anywhere between 200 and 1000 variables.

As the previous sections reveal, there is a wide variety in case characteristics, i.e., background, type of organization and type of problem. Also interventions differ widely with regard to such elements as number of participants, time investment, size of the model, as well as the way the model building process with the client is scripted. Although it is clear that with regard to the latter element the level of detail of descriptions of the process differs widely between cases. This should clearly be altered in the near future. In the next section, on the effects of group model building interventions, a similar picture will emerge.

Results of group model building interventions in organizations

Before describing the results of the studies collected in the database, we have to address three issues.

First, the authors of most of these studies did not set out to assess their modeling projects on all aspects contained in this review (see Table 1). Second, the limitations of the data have been discussed earlier but need to be emphasized again. Most projects (88 out of the 107) can be described as qualitative case studies. All use observation for data collection; six studies include individual assessment interviews; and two use a group interview. Only 19 studies use a quantitative estimation of results (see Appendix 2), either through a posttest survey (14 studies) or through questionnaires employed at two points in time (five studies). Three objective studies employ a

A third issue concerns the effect of measurement type (qualitative versus quantitative) on outcomes. Weil (1980) and Doyle (1997) for example, warn against the biases introduced by specific evaluation procedures. A useful way of identifying possible biases is to look at differences between formal quantitative assessments using questionnaires and qualitative assessments using interviews or observations. If different research methods arrive at different conclusions, this may be caused by inconsistencies between operationalization or measurement of concepts. These inconsistencies should be identified and studied in more detail in order to arrive at valid measurements.

Comparison of questionnaires and interviews/observations

Table 3 depicts the positive outcomes of the 107 modeling projects for three situations: overall (all studies) and by type of measurement (interviews/observations or questionnaires). For each of these three situations, the second column contains the frequency of studies reporting positively on the outcome.

Before turning to the differences between measurement types, we will take a look at overall outcomes of all studies in the second and third columns. The total measurements indicate that the proportion of positive outcomes is quite high, from 0.85 for shared language, to 1.00 for reaction and results (third column). Looking at the total group of studies, we find significantly more positive than negative results for all outcome variables (using a nonparametric binomial test with test proportion 0.50).6

Table 3. Assessment of group model building results in total and by measurement method (figures in brackets indicate the fraction of positive outcomes of measurements in that category). Phi coefficients for reaction, results and further use are incomputable as they are constants. Phi coefficients for reaction, results and further use are incomputable as they are constants.

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>All studies (n = 107)</th>
<th>Interviews/observations (n = 88)</th>
<th>Questionnaires (n = 10)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Positive</td>
<td>Total</td>
<td>Positive</td>
</tr>
<tr>
<td>Reaction</td>
<td>29</td>
<td>29 (1.00)</td>
<td>19</td>
<td>19 (1.00)</td>
</tr>
<tr>
<td>Insight</td>
<td>101</td>
<td>96 (0.95)</td>
<td>82</td>
<td>78 (0.95)</td>
</tr>
<tr>
<td>Commitment</td>
<td>35</td>
<td>31 (0.89)</td>
<td>23</td>
<td>19 (0.83)</td>
</tr>
<tr>
<td>Behavior</td>
<td>30</td>
<td>29 (0.97)</td>
<td>26</td>
<td>25 (0.96)</td>
</tr>
<tr>
<td>Communication</td>
<td>41</td>
<td>40 (0.98)</td>
<td>29</td>
<td>28 (0.97)</td>
</tr>
<tr>
<td>Consensus</td>
<td>53</td>
<td>49 (0.92)</td>
<td>38</td>
<td>36 (0.95)</td>
</tr>
<tr>
<td>Shared language</td>
<td>13</td>
<td>11 (0.85)</td>
<td>12</td>
<td>10 (0.83)</td>
</tr>
<tr>
<td>System changes</td>
<td>46</td>
<td>42 (0.91)</td>
<td>44</td>
<td>40 (0.91)</td>
</tr>
<tr>
<td>Results</td>
<td>24</td>
<td>24 (1.00)</td>
<td>23</td>
<td>23 (1.00)</td>
</tr>
<tr>
<td>Further use method</td>
<td>41</td>
<td>41 (1.00)</td>
<td>39</td>
<td>39 (1.00)</td>
</tr>
<tr>
<td>Efficiency</td>
<td>34</td>
<td>32 (0.94)</td>
<td>24</td>
<td>22 (0.92)</td>
</tr>
</tbody>
</table>
The difference between interviews/observations and questionnaires is small compared to the number of studies in each category. The difference percentage varies from 0 (reaction and results) to 17 (commitment). If we regard the studies as a representative sample of a larger population of group model building studies, we can test the statistical significance of these differences using the phi coefficient (which in a two-by-two table is equivalent to Cramer’s V). A phi coefficient close to zero should be interpreted as support for our expectation that outcomes do not depend on measurement type. A phi coefficient of 0.25 can be interpreted as a weak association (Nijdam and van Buuren, 1994: 125). For reaction, insight, behavior, system changes, results of system changes, further use of method and efficiency, outcomes do not seem to depend on measurement type. For consensus, communication and shared language, we find a very weak dependence on measurement type and, for commitment, a weak relationship with measurement type. In the following sections we will describe the results at the four outcome levels in more detail.

Outcomes at the individual level

Reaction As shown in Table 3, in only about one quarter of studies (29) can a statement on reactions be found. Statements such as improvement of work climate, more interesting work, belief in or acceptance of modeling results, enthusiasm, satisfaction or credibility are all coded as positive reactions. All reactions to the group model building interventions are positive.

Insight A total of 101 out of 107 cases report on insight into the problem gained during modeling. In 96 cases the result is positive, indicating that group model building resulted in an increase in insight. In five cases no insight was gained. Two of the five cases in which no insight resulted are projects in which models are built with students. In one case, the aggregation level of the model did not correspond to the mental models of students; the model was too abstract (Ginsberg and Morecroft 1995). In the modeling course facilitated by Rouwette, Vennix and Thijssen (2000), participants gained only moderate insight into the problem, and no insight in each other’s assumptions. This is explained by the focus on document analysis for data gathering and the lack of discussion between students about the problem. In the three cases in which models around real-life problems did not lead to insight, the model was too big to understand (Fey 1978) or the issue was politically sensitive and too broad to achieve focus (Akkermans 1995a,b: case software services (b)). Cavanaugh et al. (1999) report on a study in which two subgroups modeled drivers of health-sector developments in New Zealand. The majority of the participants, clinicians and an environmental scientist, conclude not to have gained insight from the resulting abstract model. Broadly stated, these five studies share a mismatch between the level of abstraction of the system dynamics model and
the clients’ mental models, and the modeling techniques used do not match the project’s circumstances (unstructured discussion on a political sensitive issue, individual data collection hampering learning about others’ opinions).

In conclusion, if the level of abstraction is adequate and techniques are matched to the objectives of the study, group model building studies generally result in increases in insight. On the basis of the data gathered, the amount of increase is difficult to determine. The issue is further confused by the difficulty of establishing what counts as a “large” or “sufficient” increase in insight. In cases aimed at finding solutions that can be implemented the increase in insight is probably best considered in relation to behavioral and systemic changes: if the clients succeeded in finding a solution to their problem, we assume that the insight gained was adequate and sufficient. This will be further discussed under system changes.

**Commitment** In only 35 cases is an influence on commitment reported. In the majority of these cases (31) commitment to the results of the modeling effort is created. (These include all 12 cases in which commitment is measured in a quantitative way.) In four instances clients indicate that they do not feel committed to the study’s outcomes. Among the studies where authors report that the project did not lead to commitment, we find the political issue again (Akkermans 1995a,b: case software services (b)). In two other cases, management agreed on the analysis of the problem, but decided not to back up conclusions nevertheless (Watts and Wolstenholme 1990; Raynolds and Raynolds 1992). In the study by Campbell and McGrath (1999: case CSC problem) their direct client is reluctant to compose a formal report and present the recommendations to management.

At first sight, it may seem disappointing that only 35 out of 107 cases report on commitment. However, this may be due in part to the different interpretations of commitment. In one sense, commitment is taken to be the intention to implement results or changes in behavior. Statements such as “the client agreed on implementing result X” fall under this category of commitment. We then have to take into consideration that only 84 studies focus from the start of the project on results that can be implemented. In 18 cases models are built for training or educational purposes, in which no implementation of conclusions is expected. In addition, even if clients in the modeling process are managers working on their own problem, the conclusion of the project does not necessarily imply a change in behavior. The conclusions of the modeling process may indicate that no changes in management practices are needed, or that adaptations on other (“systemic”) levels of the organization are necessary. Also, some of the reports are completed immediately after the project, which may be too early to reach conclusions about implementation.

In summary, we find a low number of measurements of commitment that are generally positive, but this might be said to rest on different interpretations of
the concept. From the 35 studies reporting on commitment, 31 state a positive effect on commitment. It thus seems that, in general, group modeling may have a positive effect on commitment, but the exact definition of the concept should be clarified in the future.

**Behavior** For behavioral changes, results are comparable to commitment. Thirty studies report on this topic. In 29 studies projects are followed by changes in behavior, and in one instance modelers report not to have affected behavior. In this last study (Sancar 1987: case Door County) participants in general agreed to the statement “the diagrams imply solutions”, but the management (who did not participate fully in building the model) did not implement conclusions because, according to them, the model did not include all relevant aspects.

In conclusion, 29 out of 84 projects, aiming at solutions that can be implemented, result in behavioral change. Although only in few projects do clients state that they are not committed to results or they do not rush to implement conclusions, the effect of group model building on individual behavior seems to be surprisingly small. Possible reasons for the gap between the number of studies focused on implementation and the reports containing behavioral changes may be similar to those mentioned above: results at other organizational levels may be aimed for. In addition, a number of reports are written immediately after the project or while it is still in its concluding phase, which may be too soon after the project to assess any behavioral changes.

**Outcomes at the group level**

**Communication** Frequently system dynamics is considered as a tool to improve communication within a team. The results for communication are as follows: 41 out of 107 cases report an influence on communication, 40 of which indicate an increase in the quality of communication. In one instance quality of communication is reported not to have increased (Akkermans 1995a,b: case software services (b)). There is a weak dependence of outcomes on measurement type, but most striking about this result are the low number of reports on communication and the overall positive outcome.

In this case the low number of reports cannot be explained by measurement in the short term. Applying a method such as model building must affect communication immediately, which makes it unlikely that an assessment study fails to note its impact because it is limited to short-term outcomes. An explanation for the low number of reports may be that in applying a new tool for decision making, changes in communication are unavoidable and therefore are not detailed, in order to avoid “stating the obvious”. The fact that in 19 cases consensus does result, although nothing is reported on communication, points in the same direction. After all, exchange of viewpoints, or communication, is a necessary condition for consensus to emerge (Schep 1991).
Although the number of cases that report on quality of communication is low, it seems that in general group model building leads to an increase in the quality of communication between participants.

**Shared Language** In many cases persons from different departments in a large organization, e.g., finance or research and development, can be said to use different languages to describe organizational problems. An effect on shared language is reported in only a minority of studies: 13 out of the total of 107 studies. In 11 cases this effect is positive and in two cases it is explicitly mentioned that no shared language resulted. One of the two cases in which no shared language is reported, is the politically sensitive issue described by Akkermans (1995a,b: case software services (b)). The other case (Zazara and Fisher 1996) reports on the development of cross-curricular models with teachers of pre-college students. After three weeks of training, 70 per cent of teachers use modeling in their classes, but each of them uses models specific to their own discipline. The interdisciplinary models were not used.

The low number of reports and the low significance of results make it impossible to draw conclusions. In the system dynamics literature, the expectation can be found that system dynamics serves as a uniform platform for communication (e.g., Richmond 1987: 132). On the basis of the studies gathered here, this expectation can be neither confirmed nor disconfirmed.

**Consensus or Mental Model Alignment** Alignment of problem visions is often considered a prerequisite for shared action (Huz 1999). In 53 out of 107 studies an influence on consensus is reported. In 49 cases a consensus view has been created, but in four instances clients indicate that there is no consensus on the conclusions of the modeling project. Two of the studies reporting no consensus have been discussed above (Akkermans 1995a,b: case software services (b); Ginsberg and Morecroft 1995). Verburgh’s (1994) objective assessment of mental model alignment shows no significant increase between pretest and posttest. Akkermans and van Schaik (1998) find an increase in subjective self-assessments of consensus, but a decrease in objective consensus (i.e., an increase in the variance of problem elements). This again presents a picture of a limited number of reports of an overall positive character.

We feel that the studies gathered here indicate the lack of a clear definition of consensus, which is again supported by the differences found between measurement types. Consensus could refer to agreement on a problem definition, on the actions for alleviating the problem, or on both. If the first definition is used, agreement on a model representing the problem would already constitute consensus. If consensus on actions is referred to, the concept is close to commitment. In addition, the concept of consensus implies “complete agreement” for some, while for others an increase in convergence of ideas denotes consensus as well (Scheper 1991). As mentioned above, one
study using an objective assessment of mental model alignment shows no
significant effect (Verburgh 1994). Huz et al. (1997a,b) report more alignment
in perceptions of systems goals, but no significant increase in perceptions of
strategies for change.

We therefore conclude that the studies collected provide some support for the
influence of group model building on consensus (in only four out of 53 reports
is the result negative). A more careful consideration of the exact definition
used, especially the difference between consensus on problem analysis and
consensus on actions, might enable a more accurate assessment of the effect on
consensus.

Outcomes at the organization level

SYSTEM CHANGES Of the 107 cases, only 46 report on system changes. Of these,
42 out of 46 report implemented changes at the system level. In four cases
modeling conclusions do not lead to changes at the system level. In two
instances the model suggested changes in the reward system in the client
organization, which the management did not rush to implement (Roberts,
Abrams and Weil 1978; Akkermans 1995a,b: case software services (a)).
The third study is the political sensitive issue mentioned above (Akkermans
1995a,b: case software services (b)). In the study by Campbell and McGrath
(1999 cases CSC problem), discussed under commitment, the direct client does
not want to pass on the results of the modeling project to higher management.

RESULTS With regard to results of system changes, all 24 studies report positive
results. These results should be compared with the number of studies that set
out to find a solution that can be implemented. We found 84 projects focused
on implementation, which suggests that in half (42) of the relevant cases
changes are implemented. More than half (24) of these changes led to positive
results. As a considerable number of the reports collected here were written
immediately after the project, limiting measurement to short-term outcomes,
this number may be a low estimate.

The question of system changes connects to the issue on what constitutes a
“sufficient” increase in insight: in at least half of the relevant projects (those
focused on implementation), learning about the problem took place to such
an extent that new solutions were implemented. In conclusion, group model
building changes problem understanding to such an extent that new solutions
are implemented in about half of the projects concerning real-world problems.

Outcomes with regard to method

FURTHER USE OF MODELING When it comes to further use of the method, in 41 out
of 107 cases system dynamics modeling continues to be used after the initial
project is over, which again suggests that group model building has an impact on client organizations.

**Efficiency** In 34 out of 107 cases system dynamics modeling is considered more efficient than traditional methods used for tackling similar problems.

This concludes the discussion on results for the overall group. In the following section we will try to determine if there are subsets of studies that use particular elements of modeling and produce specific results.

**Context—mechanism—outcome configurations**

The description of the results in the previous sections produces a somewhat blurred picture. No clear patterns seem to emerge from the data. Pawson and Tilley (1997) point out that, in order to find significant patterns, it is important to assess meaningful differences between studies. As discussed in the section on evaluation, differences between contexts and mechanisms will inevitably lead to differences in outcomes. If an organization is, for instance, characterized by political struggles, the process of model construction will very likely be affected, if only because participants are less likely to exchange information openly. The combination of a highly political context and a model that contains only a subset of the available information is in turn less likely to lead to a result that can be implemented. Naturally, many different context—mechanism—outcome configurations could be created. A logical choice is to connect to two discussions within the system dynamics literature mentioned earlier:

- qualitative versus quantitative models, which is in turn linked to discussions on problem tangibility and availability of data;
- small versus large models.

We decided to construct three different kinds of contexts and four mechanisms. The three contexts are: not demonstration/training situation; situations involving conflict and intangibility of problem; and situations that can be described as data rich with a more tangible problem. With regard to mechanism, we chose to place modeling efforts not aimed at implementation in a separate category, i.e. demonstration/training (column 1 of Table 4), in order to arrive at valid conclusions on implementation. Three other situations to be compared are then: qualitative models (column 2); small quantitative models (column 3); and large quantitative models (column 4). A total of 82 studies provided enough data for them to be placed in one of these four categories.

In Table 4 results are tabulated for the most important goals of group modeling. We will not discuss reactions or shared language in comparing
Table 4. Context–mechanism–outcome configurations. Following each result, the proportion of positive outcomes is given as a percentage of measurements of that outcome. The total number of measurements is indicated in parentheses.

<table>
<thead>
<tr>
<th>Context</th>
<th>Mechanism</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demonstration/training</td>
<td>18 studies, diverse models no</td>
<td>Insight 0.88 (17)</td>
</tr>
<tr>
<td></td>
<td>implementation intended</td>
<td>Commitment 1.00 (6)</td>
</tr>
<tr>
<td></td>
<td>(students or experts)</td>
<td>Behavior 1.00 (1)</td>
</tr>
<tr>
<td></td>
<td>Approximately four</td>
<td>Communication 1.00 (13)</td>
</tr>
<tr>
<td></td>
<td>participants</td>
<td>System changes none</td>
</tr>
<tr>
<td></td>
<td>4 sessions of 2–3 hours each</td>
<td>Results none</td>
</tr>
<tr>
<td>Conflict/intangible</td>
<td>15 studies, qualitative models</td>
<td>Insight 0.87 (15)</td>
</tr>
<tr>
<td></td>
<td>Average 7 participants</td>
<td>Commitment 0.71 (7)</td>
</tr>
<tr>
<td></td>
<td>3–5 sessions of 3–8 hours each</td>
<td>Behavior 1.00 (3)</td>
</tr>
<tr>
<td>Data-rich/tangible</td>
<td>19 studies, small quantitative</td>
<td>Communication 0.86 (7)</td>
</tr>
<tr>
<td></td>
<td>models</td>
<td>System changes 0.71 (10)</td>
</tr>
<tr>
<td></td>
<td>Average 7 participants</td>
<td>Results 1.00 (3)</td>
</tr>
<tr>
<td></td>
<td>5 sessions of 3–8 hours each,</td>
<td>System changes 1.00 (8)</td>
</tr>
<tr>
<td></td>
<td>up to several months</td>
<td>Results 1.00 (3)</td>
</tr>
<tr>
<td></td>
<td>30 studies, large</td>
<td></td>
</tr>
<tr>
<td></td>
<td>quantitative models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average 22 participants</td>
<td></td>
</tr>
<tr>
<td></td>
<td>About one year</td>
<td></td>
</tr>
</tbody>
</table>

Results because of the low number of studies that included measurements on these outcomes. In addition, we will concentrate on outcomes that are relevant to the problem addressed in the project and disregard outcomes related to the method. In order to make a comparison between subgroups easier, the proportion of positive outcomes of measurements in each subgroup is indicated. For example, out of 18 demonstration/training projects, 17 measure insight, of which 0.88 report a positive effect. In the following we will look at context, mechanism and outcome in turn.

**Context**

As stated, the relevant studies for comparing the effects of different model types are the projects that focus on implementation in columns 2, 3 and 4 of Table 4. These subgroups do not differ with regard to organizational background; sort, sector and size of client organizations are comparable. No difference can be found in problem importance either.

In each subgroup a comparable number of studies (53 to 60 percent) are considered important by participants. Several authors feel that the choice for qualitative versus quantitative models depends on problem characteristics such as scope, tangibility, data availability and conflict between stakeholders (Akkermans 1995; Coyle 2000). However, it is not very clear how these should be operationalized, as many problems entail soft, intangible factors and some
degree of conflict. These problem characteristics seem to boil down to the extent to which stakeholders have information on the problem and are in dispute over the problem. In the research by Hickson et al. (1986) these factors form the analytical and social dimensions of problem complexity. Hickson et al. (1986) provide further detail on how to operationalize complexity. Several elements of complexity can be found in the data in this study, i.e. rarity or frequency with which similar matters occur, diffusion of consequences and interests involved. These factors are included in the database entry on the motive for starting the intervention. These indicators show that qualitative models are more likely to be applied in situations that are new to the people involved, have widespread consequences and involve a broad range of interests. Quantitative models are more often applied to production, distribution or human resource problems that organizations are likely to have encountered in some form before, involve fewer parties and are more restricted in the consequences. Basing ourselves on data on frequency, diffusion of consequences and interests involved, it seems that the decision whether or not to quantify does indeed depend on problem characteristics.

Mechanism

The general process of modeling in each of the subgroups can be described as follows. In the first subgroup (demonstration models) participants are not expected to implement the results. In these projects diverse models are constructed, of different sizes and qualitative as well as quantitative. The group of participants is on average small and usually the time investment is limited.

In the second subgroup qualitative models are used for finding results that can be implemented. Models are made with a small group of managers in a limited span of time.

The cases in the third subgroup result in small quantitative models of a maximum of 50 equations. These models are built with the participation of a small group of managers. Time investment is limited, although five projects are measured in months. In most cases the quantitative modeling is not done in front of the group of participants, but rather by the consultants separately and then discussed with the clients.

The largest group model building projects are those in which models of 50 and more equations are built. The scope of these projects is large with respect to both number of participants and time investment. On average, building these models requires a client organization’s commitment for about one year and about 22 participants in model construction. In some cases, a very large number of people are involved, e.g. the “team of teams”, which consists of 160 people (Graham and Walker 1998). Not surprisingly, these projects also take the longest to complete.
Because of the low number of cases in each subgroup, a small change in outcomes found has a large proportional impact. We will therefore only consider a difference of percentages larger than or equal to 20.

The only outcomes that differentiate between subgroups 2, 3 and 4 are commitment, consensus and system changes. Qualitative models seem to be less likely to lead to commitment, consensus or system changes than (small or large) quantitative models. The likelihood of positive results of system changes is equal for all three types of models. In addition, qualitative models do not seem to differ substantially from quantitative models with respect to their capacity for generating insight, behavioral changes or communication. These outcomes should be considered in terms of the different contexts in which models are used. Qualitative modeling appears to be an intervention requiring relatively few organizational resources (time and participants) and is instrumental in clarifying intangible matters. Quantification adds substantial benefits in situations where the problem is more structured and justifies a larger investment in terms of participants’ time; it is thus more effective in producing consensus and system changes.

The diversity in case descriptions and in the way projects are evaluated by practitioners and consultants prevent more detailed conclusions. It is difficult to find more patterns in a database that contains many “missing values”. Surely this is an undesirable state of affairs. The system dynamics community deprives itself of critical learning opportunities. A large number of systems thinking interventions are conducted each year. As stated, between 1990 and 1999 from three to 13 case descriptions were published each year. And we may assume that many more interventions were conducted that were not published. Case studies will no doubt be conducted in the future and, in order to get more out of these case studies, a first step is to make sure that each practitioner employs the same “score sheet” when describing and assessing a particular intervention. From this we can start to build a (quantitative) database, which allows us to conduct formal analyses to test hypotheses in a more rigorous way. The next section will discuss the basic layout of such a score sheet, and may help different practitioners to describe and assess their interventions in similar ways.

Towards rigorous guidelines to report group model building interventions

In order to capture all relevant variables in our score sheet, we take Pawson and Tilley’s (1997) framework of context—mechanism—outcome configurations as our point of departure. Context refers to factors that are more or less independent of the consultant’s activities, e.g., type of
Mechanism, on the other hand, refers to all those things that are done during an intervention. Andersen et al. (1997) make a distinction between three stages in group model building interventions: pre-meeting activities; the actual meetings; and follow-up activities (see also Richmond 1997). Outcome refers to the effects of the intervention, be it for individuals, groups or the larger organization.

As may be anticipated it will not always be clear where a particular factor will have to be located. For example, team composition or level of top management support may be a contextual factor as far as it is not under the consultant’s control, but it may also be placed under mechanism as it may be part of the negotiation process prior to the actual systems intervention meetings. For analysis purposes, when the database is finished, this is of course a trivial question.

The reader may wonder why such an extensive record needs to be kept of group model building cases, and whether all entries need to be included. The entries proposed in the following store the context, intervention and outcome elements that have been considered most relevant in the literature. Since we have only few hypotheses about how these factors are related, it is difficult to decide beforehand which factors to leave out and we have to revert to an exploratory approach. We intend to publish the full database on the Internet and start a discussion in the wider system dynamics community on the relevance of each entry.

**Context characteristics**

Context variables concern such variables as geography, characteristics of the organization, and characteristics of the problem.

With regard to geography the database should have an entry specifying the continent, the country (and possibly the state/city) in which the intervention took place.

Since it will be difficult to record such things as organizational culture without using elaborate questionnaires, we propose to limit organizational characteristics for the time being to structure, type and size. Structure refers to the way things are organized, for example, functional, team-based or network organization. Type can be subdivided into:

- Profit (production, services, distribution).
- Non-profit (e.g., teaching, broadcasting).
- Governmental (national, state, county, city), as well as type (e.g., defence, finance, economic affairs).

Size can be characterized by:
• Number of people employed.
• Estimated returns.

With respect to problem characteristics, we distinguish an analytical and a social dimension (see also Hickson et al. 1986). Both have a number of subdimensions, each of which will add to the complexity of the problem:

• Analytical
  — Rarity or uniqueness of the situation (how often was this or a similar situation confronted before?).
  — Consequentiality (how radical, serious, long term or widespread are the consequences of the problem expected to be?).
  — Precursiveness (to what extent does the decision set parameters for subsequent decisions?).
  — Number and diversity of interests involved.
  — Openness to alternatives (had the decision already been made?).

• Social:
  — Pressure of influence (how much influence was exerted?).
  — Intervention (how much external influence was exerted?).
  — Imbalance (to what extent was the pressure uneven between units?).
  — Contention of objectives (did units exert influence in opposite directions?).

Mechanism

Mechanism refers to the process of the intervention itself. Up to now interventions have largely been treated as black boxes. But a group model building intervention is a complex process containing many different elements and processes. It is just impossible to draw valid conclusions on effectiveness, unless we open up the black box and record in more detail what exactly happens during group model building interventions. Below we present a list of items that should be recorded prior to, during and after an intervention.

Pre-project activities These concern all activities that take place before actual meetings with a management team are conducted. A number of things should be recorded here:

• Initiation of contact:
  — Who initiated contact: system dynamicist or client organization?

• Expectations and goals of project:
  — What were the initial expectations of the client?
— What were the goals of the project (getting more insight, making decisions, implementing results, testing proposed strategies/policies, cf. Richmond, 1997).
— What was the level of top management support?
• Type of question addressed: exploratory, explanatory, descriptive, or prescriptive.
• Composition of participating management team:
  — Who decided on the composition of the team (client, consultant or both)?
  — Size of the team.
  — Composition and heterogeneity of team; characteristics of team members, i.e. their functions.
  — Was there an official gatekeeper?
• Modeling team:
  — Number of persons?
  — Their roles: facilitator, recorder, content coach, process coach (cf. Richardson and Andersen, 1995)?

THE MODEL BUILDING MEETINGS  Here we have the following elements: the actual meetings and time investment, modeling procedure, aspects of facilitation, and meeting logistics:

• Meetings and time investment:
  — Number of meetings and average duration.
  — Total time investment by participants.
  — In what stages of model building did the client participate, and to what extent (marginal to fully)?
  — How much (and what kind of) work was done off-site and how much with the group?
  — Total time investment by modelers.
  — Total time span of the intervention (from initial contact to project close).
• The model and modeling procedure:
  — Introduction to system dynamics:
    – Was an introduction to system dynamics given? And if so what type of introduction?
    – Were management flight simulators used to introduce people to system dynamics thinking?
  — Modeling process:
    – What type and process of modeling was used (flow diagrams or causal loop diagrams, quantitative modeling, simulations and how policies were assessed)?
    – What software for modeling was used?
  — Model:
– Size of model.
– Dynamic complexity of model (number of feedback loops).

The system dynamics literature contains several papers on knowledge elicitation. Important elements in this literature are the use of group process techniques and scripts used to elicit knowledge (cf. Andersen and Richardson, 1997)

- Elicitation of mental models:
  - Sources of information:
    - What other sources of information (apart from the client’s mental models) were used to build the model?
  - Process of eliciting knowledge:
    - Were pre-meeting interviews held?
    - Was a preliminary model used or did the meetings start from scratch?
    - Were specific group process techniques used, and if so which (NGT, brainstorming, Delphi etc.)?
    - Were particular scripts used to elicit knowledge, and if so which (e.g. feedback elicitation script)?
    - Were questionnaires/workbooks used?

An important aspect of the process of group model building is facilitation. Here it is more difficult to record “objective” characteristics. We will have to rely on participants’ opinions on the role of the facilitator.

- Facilitation aspects:
  - Degree of neutrality of the facilitator.
  - Was the facilitator seen as skilled by the group?

Finally there are a number of practical things that should be recorded:

- Logistics:
  - Were meetings held away from the office?
  - Room design and layout.
  - What materials were used: whiteboard, overhead, computer, hexagons etc.

Several activities after the meeting are worth noting as well:

- Follow-up activities:
  - Was there an official report?
  - Did any follow-up activities take place, and if so which?
Outcome: assessment of effects

As this article has revealed, evaluations of the success of an intervention range from consultants’ observations, through interviews with participants, to more sophisticated questionnaires asking for clients’ opinions. Although the most desirable state of affairs would maybe to have “objective measurements” of changes as a result of the intervention, this situation cannot always be accomplished. However, the least we should do is to conduct the same measurements, in all interventions, of participants’ opinions of the process and its outcomes.

Since we are dealing here with data collection over a number of cases, it is important to employ identical ways of gathering evaluation data. Huz et al. (1997a: 151) distinguish three levels and 10 domains of measurement and evaluation to assess the impact of a systems thinking intervention. The three levels are: reflections of the modeling team, participants’ self reports, and measurable system change and “bottom line” results. The domains include things like: shifts in participants’ goal structure or change strategies, shifts in understanding how the system functions or changes in system wide policies. More generally, it seems that we have to distinguish between:

- the source of data (e.g. modeling team, participants);
- how the data are collected (e.g. interviews, questionnaires, (in)formal observations);
- when data are collected (i.e. pre, post, during intervention);
- what data are collected.

To start with the last issue, data can be collected on outcome variables (e.g. shifts in participants’ perceptions of the problem, system changes) and process variables (e.g. participants’ satisfaction with the process, participants’ perceptions of the value of different elements of the intervention, for example the role of the facilitator). These data can either be generated by the modeling team and or be elicited from the participants (cf. Huz et al., 1997a,b). To keep the information generated as identical as possible, standardized questionnaires and observation protocols seem most reliable.

Let us first take a look at the outcome effects of the intervention. As the cases in this article suggest, the outcomes of an intervention may vary depending on the type of organization and the type of problem. Based on the cases in this paper and previous divisions that have been made (cf. Andersen et al., 1997), the effects of systems thinking interventions seem to occur at four levels: individual; group; organization; and method. The relevant outcomes on each level are depicted in Table 1.
On a number of these outcome elements questionnaires have been used in the past (cf. Vennix and Gubbels, 1994; Huz et al., 1997a,b). These have become more and more standardized, but elements will need to be added to make the questionnaire consistent with the above framework. This is also true with regard to an observation score sheet for the modeling team. A standardized score sheet to record observations made by the modeler/modeling team is currently in preparation. In the near future these intervention score sheets (the modeling team assessment sheet and the questionnaire) will be put on the web, with appropriate instructions, so that model-builders can use them. As filled-out score sheets and questionnaires are returned, the database can gradually be filled, analyses can be performed in the future and our institutional learning process with regard to the effects of systems thinking interventions will gradually take shape.

Table 5 presents the context—mechanism—outcome variables discussed in the above and the data collection approaches that can be used for measuring each variable.

Contextual as well as mechanism elements can easily be assessed by the project team. Additional data on meeting elements and their effects is provided by session observations and participant reports on the meeting process. Outcomes at the individual and group level can be observed by the modeling team during the sessions, as well as obtained from participant reports (interviews or questionnaires). Participant reports also cover organizational impacts (system changes and results) and outcomes related to the method. Documents and databases used in the organization data are an important information source for background data, actual system changes and results.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Project team Reflections</th>
<th>Session observations</th>
<th>Participant self-reports</th>
<th>Documents/ databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Problem</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanism</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-project activities</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model building meetings</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Method</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and discussion

Case descriptions and effectiveness studies into system dynamics interventions have been published almost since the inception of the field. Many case studies show how clients were involved in model construction. There is also a growing tradition of research into microworlds (Andersen, Richardson and Vennix 1997; Größler 2000). However, some of the most fundamental intuitions about the outcomes of system dynamics interventions, i.e. gains in insight and system performance, are very difficult to test. The analysis presented here draws on published research into group model building and attempts to draw some preliminary conclusions with respect to its effectiveness. We started this article with three goals in mind:

- to provide an overview of outcome studies on system dynamics interventions;
- to draw preliminary conclusions on the effectiveness of system dynamics interventions and to assess differences in definitions of outcomes;
- to contribute to more rigorous assessment and reports.

Overview of cases

First, this study provides an overview of publications on group model building. A much larger number of studies has been found than we expected or than have been considered in previous multiple case studies. Although quantitative assessment of outcomes seems largely limited to the last decade, there are many qualitative case reports to be found in the literature. This allows us to compare systematically in-depth qualitative studies with surveys and field experiments, with regard to the outcome variables reported. In this way a number of robust results, differences in operationalization of concepts and possible improvements in measurement could be identified.

Preliminary insights and definitions of outcomes

Second, although this meta analysis is limited, both because not all model building interventions will be included and because of the wide variety in case study reports and the way assessments are made, some preliminary insights may be drawn from this analysis.

Learning about the problem seems to be a robust outcome of group model building. A large number of studies identify an increase in insight. Although it seems rather straightforward to expect modeling to lead to learning about a problem, research into microworlds frequently fails to find an effect (Andersen, Richardson and Vennix 1997). Our database shows that the majority of both interviews/observations and questionnaires report learning effects. These measurements include subjective self-reports as well as more objective
pretest–posttest comparisons. This seems to point to a real difference between participation in model construction and operating a finished model, as in microworlds. The results here are in line with Sterman’s (1994: 320) assertion: “To learn […] participants must become modelers, not merely players in a simulation. In practice, effective learning from models occurs best, and perhaps only, when the decision makers participate actively in the development of the model.” (See also Vennix and Gubbels 1994.)

Fewer studies, although still a substantial number, report on system changes and improvements. Half of the studies focused on finding solutions that can be implemented and result in changes in organizational policies. And, in turn, about half of these changes resulted in system improvements.

In general, commitment and consensus are found to increase after participation in modeling, although research can benefit from careful definition and operationalization of both concepts. Commitment and consensus are defined and operationalized in different ways. Definitions employed in questionnaires diverge from those in case studies.

A further insight relates to the discussion on types of models. Repeatedly, system dynamicists express their appreciation for one type of model over another, i.e., quantitative versus qualitative (see the discussion by Coyle 2000) and small versus large models (see Lyneis 1999). In this analysis we found a number of differences between modeling projects focused on finding implementable conclusions, depending on whether they employed qualitative models, small quantitative models or large quantitative models. It seems that qualitative models are less likely to lead to commitment, consensus or system changes than (small or large) quantitative models. However, for fostering insight and quality of communication, the three model types do not differ.

The idea that qualitative models are deliberately chosen in contexts where quantification is difficult is supported by this analysis. Qualitative models are often used in situations that are new, that have wide-ranging consequences and that have many stakeholders. Constructing this type of model requires fewer resources than fully quantified models. It seems that quantification is particularly helpful when conceptualization is less difficult and the time and money for testing the effects of alternative interventions is warranted.

Uniformity in reports

The third, and maybe most important, contribution of this article is that it presents more rigorous guidelines with which (a) to report on the characteristics of the intervention and (b) to assess the effectiveness of model building interventions. Doyle, Radzicki and Trees (1998) describe several limitations of individual case studies. However, it seems likely that reports on individual case studies will continue to be a major source of information on group model building in the years to come. As stated, each of these case studies may provide in-depth insights into a particular intervention. However, for a
systematic research program on the effectiveness of group model building.
this is not sufficient. In order to get the most out of these case studies and to
enhance our own learning process as consultants and practitioners as much
as possible, we should at least make sure that we create uniformity in the
way individual interventions are reported and assessed. There is great value
in working towards a standard measurement procedure to assess group model
building interventions. In this article we have outlined a standardized score
sheet for the description and assessment of cases. As stated, this standard score
sheet (including guidelines on how to use it and standard questionnaires) will
be posted on the web in the near future. We can only urge practitioners to use
this standard score sheet, not only in successful cases, but also particularly in
cases that do not seem to be successful. In that way you will not only help to
make the research program of the system dynamics community more scientific,
but you will also contribute to our institutional learning process, which will
advance the effectiveness of our interventions in the future.

Notes

1. We have chosen this label, because in our survey we want to focus on
those interventions in which a system dynamics model is built with a client
group. We therefore want to exclude system thinking interventions in which
for instance only flight simulators are used.
2. If a modeling project was reported in more than one publication, the
information from all publications found was combined. Similarly if a
publication was a multiple case description, cases were split into single
entries in the database. One study that employed an experimental design
was coded as a single entry, as data were aggregated for all cases
involved.
3. The types of research questions are borrowed from a typology of fundamental
and applied research (Swanborn 1987). An open question aimed at
uncovering elements related to the subject of the study was marked as
“exploratory”. If the aim was to identify facts or delineate a state of affairs,
the modeling question was categorized as “descriptive”. If the model was
used to identify the causes or reasons for a situation or development, it was
termed “explanatory”. A “prescriptive” focus was one in which a concrete
action to bring about change was intended.
4. The designs are (Swanborn, 1987; Cook and Campbell, 1979):
- an experiment, using pretests and posttests, a control group and random
  assignment of individual subjects to the experimental or control group;
- a field experiment, using pretests and posttests, a control group but no
  randomization;
- a one-group pretest–posttest, identical to the above but without a control
group;
5. In a posttest survey participants are asked to indicate their level of learning, resulting in a self-report or subjective estimation. Measurement of knowledge on the problem at two points in time allows for a more objective estimation of changes in, for example, insight and consensus. (However, measurement at one point should not be equated with a subjective estimation, and neither is measurement at two points in time of necessity objective. For example, subjective self-reports on knowledge at two points in time could be used to estimate the degree of learning.)

6. Because of the small number of measurements we used a significance threshold of 0.01. For shared language significance is 0.022 (exact, two-tailed); for all other variables significances are 0.000. With the exception of use of shared language, we interpret this outcome as support for the positive effect of group model building on the outcomes in Table 4.

7. With regard to shared language only one quantitative assessment is included in the database, so we will not take this difference between measurement methods into consideration in the discussion of results.

8. There are a couple of projects that do not follow the general pattern with regard to the number of participants or time investment. Educational projects usually involve a small group of participants and a limited time investment. Exceptions are Rouwette, Vennix and Thijssen (2000), Knops (2000) and Ginsberg and Morecroft (1995), who involved groups of 30, 32 and 70 students respectively. Vennix (1996), Guthrie and Patton (1998) and Eriksen and Nielsen (1985) take a time investment that is much longer than average. Qualitative models that, in contrast to the general pattern, involved a large group of participants are described by Genta et al. (1994, using scenarios) and Vennix et al. (1990, using Delphi).

References


Lane DC. 1993. From discussion to dialogue: how an interactive modeling approach was used with managers to resolve conflict and generate meaning. *Proceedings of the 1993 International System Dynamics Conference*. Cancun, Mexico, 231–234.


Appendix 1. Case-study literature


Henderson SM, Wolstenholme EF. 1990. The application of a dynamic methodology to assess the benefit of a battlefield information system. *Proceedings of the


Lane DC. 1993. From discussion to dialogue: how an interactive modeling approach was used with managers to resolve conflict and generate meaning. *Proceedings of the 1993 International System Dynamics Conference*. Cancun, Mexico, 231–234.


### Appendix 2: Studies employing quantitative assessments

<table>
<thead>
<tr>
<th>Study</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akkermans (1995, case software services a)</td>
<td>Insight; Commitment; Communication; Consensus; No system changes</td>
</tr>
<tr>
<td>Akkermans and Van Schaik (1998)</td>
<td>Insight; No consensus</td>
</tr>
<tr>
<td>Bentham and De Visscher (1994)</td>
<td>Positive reaction; Insight; Behavior; Consensus; System changes; Results; Further use</td>
</tr>
<tr>
<td>Berkvens and Neomagus (1997)</td>
<td>Insight; Commitment; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Cavaleri and Sterman (1997)</td>
<td>Insight; Behavior; System changes</td>
</tr>
<tr>
<td>Draper and Swanson (1990)</td>
<td>Reaction; Insight; Commitment; Efficiency</td>
</tr>
<tr>
<td>Hendrikx (1998)</td>
<td>Insight; Commitment; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Huz (1999)</td>
<td>Positive reaction; Insight; Behavior; Communication; Consensus System changes; Results</td>
</tr>
<tr>
<td>Kelly (1998)</td>
<td>Reaction; Insight; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Knops (2000)</td>
<td>Reaction; Insight; Commitment; Behavior; Communication; Consensus; Shared language; Efficiency</td>
</tr>
<tr>
<td>Rouwette, Vennix and Thijssen (2000)</td>
<td>Positive reaction; Insight; Commitment; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Sancar (1987, case Door County)</td>
<td>Positive reaction; Insight; Commitment; No Behavior; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Sancar (1987, case Janesville)</td>
<td>Positive reaction; Insight; Commitment; Communication; Consensus</td>
</tr>
<tr>
<td>Vennix (1995)</td>
<td>Insight; Commitment; Behavior; Communication; Consensus; System changes; Efficiency</td>
</tr>
<tr>
<td>Vennix, Gubbels, Post and Poppen (1990)</td>
<td>Insight</td>
</tr>
<tr>
<td>Vennix, Scheper and Willems (1993, case Nostradamus)</td>
<td>Insight; Commitment; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Vennix, Scheper and Willems (1993, case Dutch river system)</td>
<td>Insight; Commitment; Communication; Consensus; Efficiency</td>
</tr>
<tr>
<td>Verburgh (1994)</td>
<td>Insight; No consensus; Efficiency</td>
</tr>
<tr>
<td>Wallace and Sancar (1988)</td>
<td>Positive reaction; Insight; Commitment</td>
</tr>
</tbody>
</table>

Concepts in italics were measured qualitatively.