Supporting the Construction of Qualitative Knowledge models
Bessa Machado, V.

Citation for published version (APA):
Bessa Machado, V. (2004). Supporting the Construction of Qualitative Knowledge models
This chapter documents two studies on model-building environments. In the first study we examine the manner in which subjects use MOBUM to solve model-building tasks. Analyses of the actions of the subjects (novices and experts), verbal communications and interviews were conducted in order to pinpoint to what extent MOBUM is effective in supporting modelling tasks and how easy it is to use. The second study compared novices users of MOBUM and HOMER in performing modelling tasks. The study addresses questions of usability, preferences of the participants for one system over the other, the help functionality of the engine (its usefulness for successful completion of tasks) and the perspectives of participants on the use of such systems.

6.1 Introduction

This chapter describes and discusses the results of two studies performed using our two model building environments. In the first study, MOBUM, a model building environment designed and implemented within the scope of this research, is used. MOBUM was developed based on the outcomes of the experiment with HOMER (see Chapter 3) with the goal of overcoming the problems found therein. MOBUM assists users in developing and understanding qualitative modelling concepts. The research questions addressed in this study are: a) to what extent are the improvements and features of MOBUM effective in supporting the model building process? and b) Is MOBUM an easy to use model building environment? What are the shortcomings of the current implementation?

The second study is a comparison between the two systems presented in the scope of this research: MOBUM and HOMER. The main research question in this study addresses the ease of use and the preferences of the users for either model building environment. We expect that the participants will like MOBUM better due to improvements that lead to an easier to use interface. Following this expectation, we also expect that by being easier to use, participants are going to perform better in their modelling tasks and therefore produce better results.

The chapter is organised as follows. The first study is presented in Sections 6.2 and 6.3. The second study is described in Section 6.4. The chapter ends with the conclusions and discussion in Section 6.5.
6.2 MOBUM Evaluation

This section presents the experiment for evaluating MOBUM. The aim is to assert to what extent the features of MOBUM effectively support users in building models. For that purpose, three main aspects are considered in this experiment: the progress of the participants during their model building activities, their interaction with the user interface and their qualitative evaluation of the system during an interview.

6.2.1 Method

Participants

A total of 7 participants were recruited to take part in the experiment, three of whom were considered novices in model building. One of the three novices never had had any contact with qualitative reasoning and modelling in general. The other two had some basic knowledge about qualitative reasoning but had never experimented with any modelling activity. The remaining four participants were considered to be experts. Experts, in this context, are considered to be users that have built qualitative models before. The purpose of the novice/expert distinction was not to compare the performance of the two, but rather to ensure that an adequate range of users was covered. For this purpose, the participants were given tasks which corresponded to their capabilities. In fact, with this experimental setup the help system as well as the user interface could be broadly tested and evaluated.

Design and Data Gathered

The assignments were designed to match the participants levels of expertise as much as possible to allow everyone to adequately experience the tool as well as giving us the opportunity to effectively evaluate the embedded help system. Expert users were therefore given tasks that were difficult to complete and for which we would expect them to require some additional help in order to be able to successfully complete the assignment. The same applied for the tasks given to the novice users. Although they were not as difficult as the ones given to the experts, the probability was high that a beginner would need some help.

The expert subjects were asked to construct a simulation model of an U-tube system from scratch. The task was essentially the same as the one used in the Homer Experiment (see Appendix B for a complete specification of the assignment). The subjects received documentation containing the assignment, a short explanation of the employed qualitative modelling terms as well as a brief introduction to the MOBUM Environment.

For novices, three sequential assignments were planned. As soon as the subjects completed one assignment the following was presented to them. A brief introduction to the modelling task, "Garfield's weight" (see Appendix C for details) was given beforehand. The main task for the participants was to determine the effect of food intake and physical exercise on the weight of Garfield. The participants were also handed documentation containing a general introduction to building qualitative models which they were allowed to read for 7 minutes. Each session lasted one hour.
Additionally, in both situations (novices and experts) a drawing, illustrating the situation the participants should model, was available in the SketchPad, the drawing tool of MOBUM. So, for the experts, a drawing of the U-Tube system was available and for novices a drawing of the Garfield system. The participants were not requested to use the drawing as a reference. It was explained that the drawing depicted only a general overview of the system they should model, see Figure 6.1 for illustration.

![Figure 6.1](image)
The drawing available for the novices in the experiment.

All computer actions as well as verbal data for each of the sessions were recorded on video in order to facilitate analysis. Two types of data were collected in order to evaluate MOBUM: 1) screen information and 2) the verbal utterances of the participants. Participants were asked to think aloud as much as possible. An added value of the "think aloud" protocol is that the user provides us with valuable information regarding the reasoning underlying the actions taken during the model building task.

In order to evaluate the user interface and its usability, the heuristic evaluation method was used [83], as was also done previously in the evaluation of the user interface of Homer in Chapter 3. A usability problem occurs whenever a participant sets out to accomplish a task but fails to do so due to an unexpected behaviour of the model building environment.

In order to measure the usefulness of the help system, we observed at which moments a help agent was requested and if the given feedback was sufficient for clearing the doubts of the subjects about the problem at hand. Additionally, the questions posed by the participants to the experiment leader were analysed to verify whether they were in principle covered by the implemented help system in which case they could just as well have been solved by the help agents. While the participant completed each task, the experimenter noted the number of times an agent was used and the number of different approaches
attempted within the help system (that is: the type of agent-based help facility that was consulted). In order to measure the participants' performance, the models they created were compared to existing models created by experts.

6.3 Results of the MOBUM Evaluation

This section describes the results of the MOBUM evaluation following the approach taken in Chapter 3. Firstly we describe confusions and misconceptions detected during model building activities (Section 6.3.1). Every description is followed by an explanation, examples and, in some instances, protocol fragments. This part also includes a discussion about the usage of agents. Secondly, the feedback of the participants on the user interface is presented (Section 6.3.2). This feedback is analysed within the framework of the Heuristic Evaluation method. Notice that in the following two sections the results are divided into two categories according to the level of the participants, i.e. either novice or expert. Finally, the qualitative evaluations of the participants of the model building environment are presented, Section 6.3.3.

6.3.1 Model building activities

Novices

Similar to what was done in Chapter 3, the problems encountered by the participants during the model building activities are classified into types. The types used in Chapter 3 are summarised in Table 6.1. Notice that below in this section a few additional problem types are defined (notably 16, 17, and 18) in order to describe the MOBUM evaluation results.

Understanding and using the concept of an isa hierarchy (Type 1). One participant failed to understand that while creating the entities and thereby specifying its super-type, the is-a hierarchy is automatically defined. The participant defined an explicit is-a relation between two entities (Garfield is-a cat) in the Structure Builder as well as in the Scenario Builder. This is possibly due to the fact that in the Structure builder the user creates the hierarchy of entities, the attributes, as well as relations between those entities.

Understanding the relation between a Quantity and its Quantity Space (Type 3). One of the participants did not realise that every quantity space has its own specific name and therefore named the quantity space with a quantity name. The name of the values in the quantity space, however, were all correct (E.g. Fat, normal, skinny).

P5: Garfield has another 'Quantity Space', Weight!

Understanding a Model Fragment Type and its implications (Type 4). When adding a new item to a model fragment, the user interface requests the user to
### 6.3. RESULTS OF THE MOBUM EVALUATION

<table>
<thead>
<tr>
<th>NR</th>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Understanding and using the concept <em>isa</em> hierarchy</td>
<td>The user's awareness of the fact that entities should be hierarchically organised.</td>
</tr>
<tr>
<td>2</td>
<td>Deciding upon which quantities to define</td>
<td>One of the principal tasks when modelling a system consists of defining the quantities that describe the behaviour of its entities.</td>
</tr>
<tr>
<td>3</td>
<td>Understanding the relation between a Quantity and its Quantity Space</td>
<td>The understanding of the two concepts quantity and quantity space and how they are used within the ontology.</td>
</tr>
<tr>
<td>4</td>
<td>Understanding a Model Fragment Type and its implications</td>
<td>The understanding the different types of model fragments as well as how the knowledge they represent must be organised.</td>
</tr>
<tr>
<td>5</td>
<td>Deciding upon structural decomposition</td>
<td>This category is concerned with the way that entities are used in different contexts as opposed to focussing on the entities themselves.</td>
</tr>
<tr>
<td>6</td>
<td>Deciding upon which behaviour to assign to which entity</td>
<td>Relating quantities to the right entities.</td>
</tr>
<tr>
<td>7</td>
<td>Understanding and using the notion of configuration</td>
<td>The understanding of the configuration concept as well as knowing where and how to use it.</td>
</tr>
<tr>
<td>8</td>
<td>The difference between an Attribute and a Quantity</td>
<td>The understanding the difference between describing an entity's static features, conceptualised as attributes in the vocabulary of GARP, and its dynamic behaviour, represented by quantities in the same ontology.</td>
</tr>
<tr>
<td>9</td>
<td>Understanding Quantity Space definitions</td>
<td>Understanding the rules for defining quantity spaces.</td>
</tr>
<tr>
<td>10</td>
<td>Understanding and using dependencies</td>
<td>Understanding of the different types of dependencies and their correct use in model fragments and scenarios.</td>
</tr>
<tr>
<td>11</td>
<td>Understanding the organisation of Model Fragments (and Scenarios)</td>
<td>Understanding of the different types of knowledge and how they can be specified within model fragments and scenarios.</td>
</tr>
<tr>
<td>12</td>
<td>Understanding the differences between generic knowledge (i.e. types) and instantiated knowledge (i.e. user-defined names) and knowing when to apply either</td>
<td>Understanding the difference between: 1) Generic knowledge, which in our ontology includes the entities of the <em>isa</em>-hierarchy, configurations, quantities and quantity spaces. 2) Instantiated knowledge, categorising instances and occurrences of generic knowledge used in model fragments and scenarios in order to express a specific situation. Different instances of the same concept are identified by user-defined names.</td>
</tr>
<tr>
<td>13</td>
<td>Not knowing how to specify values</td>
<td>Unawareness of the role of defining initial values especially in scenarios.</td>
</tr>
<tr>
<td>14</td>
<td>Deciding upon which values to include in a quantity space</td>
<td>Optimal definition of values for expressing the change of behaviour of quantities.</td>
</tr>
<tr>
<td>15</td>
<td>Attribute-Configuration confusion</td>
<td>Conceptual misunderstanding as to the correct use of attributes and configurations.</td>
</tr>
</tbody>
</table>

Table 6.1
Identified Types of Problems.

choose whether that item will be a conditional or a consequential knowledge item. While creating a model fragment, two participants pointed out that they did not understand the notions of *Conditions* and *Consequences*. 
The difference between an Attribute and a Quantity (Type 8). One of the participants confounded the 'attribute' concept with that of a quantity. Consequently, the participant defined all the quantities in the model using attributes (e.g., Weight).

Understanding Quantity Space definitions (Type 9). The difference between the notions of 'points' and 'intervals' was unclear to one of the participants.

P1: I am not sure when it should be a 'point' or an 'interval'.

Another participant was intrigued by the existence of the default quantity space 'MZP'. Remarkably, the same participant defined quantity spaces upsidedown.

Understanding and using dependencies (Type 10). One of the participants wanted to add an influence to a Scenario. Another participant requested help of the experiment leader to clarify the difference between an influence and a proportionality.

Not understanding derivatives (Type 16). This type refers to misunderstanding the meaning of a derivative's values. A participant was in doubt about the derivative option when adding a quantity to a model fragment. At that point, the derivative of the quantity could have been specified but the participant did not know yet what it meant.

Confusion between entities and quantities in the Structure Builder (Type 17). Two subjects defined what they intended to be quantities as entities in the Structure Builder (e.g., "weight" and "amount of food" were defined as entities).

Figure 6.2 gives an overview of the amount of misconceptions/problems participants had, categorised according to the identified types. In total the three novices encountered 24 difficulties.

Figure 6.2
Number of Problems categorised according to the above mentioned types (Novices).
6.3. RESULTS OF THE MOBUM EVALUATION

Running a simulation. Within the novice group, only one participant (P1) did not complete the creation of a scenario following the first assignment and therefore was not able to run the model using the simulator. The other novices integrally finished the first assignment and simulated their models. The last of the three participants finished the first assignment in 40 minutes and the second assignment in 8 minutes. Interestingly, this participant decided to manipulate initial values in the scenario and recheck the results by rerunning the simulation.

Help and support on model building activities. Without exception all novices found the agents useful and essential. The agent-based help facility was essential in aiding the participants in solving conceptual problems. For example, two participants (P1, P2) wrongly specified the quantities in their models as entities using the Structure Builder (Type 17). When specifying a model fragment, one of them realised that it was not possible to define dependencies between 'entities' but only between 'quantities'. So, the participant backtracked and consulted the agent to understand what had been done wrongly. In doing so, the participant learned what the mistake was. The following protocol was extracted from the interview with the participant.

P1: I had never built models before. It is a complex activity and I had to learn it now...
   I had to develop the concepts while I built the model (e.g., Garfield is an entity; weight is a quantity) Now it all makes more sense...
   The feedback and the suggestions are positive features and are very useful.

As already pointed out, this participant (P1) had no knowledge about (qualitative) modelling and consequently also no understanding about 'points' and 'intervals' in a quantity space. But during the process of creating a quantity space, the participant learned about them. It took the participant 15 minutes to specify the first quantity space, 2 minutes for the second and 30 seconds for the third. In another case (Type 16), after consulting the agents, the participant found the explanation about derivatives and understood their meaning. Later, the participant returned and used the concepts correctly.

Table 6.2 summarises the types of problems novices encountered during their model building activities. We verified who solved the problems (agents or experiment leader) and checked whether the problems solved by the experiment leader could also have been solved by consulting the agents.

Discussion and Conclusions. It is encouraging to see that almost all the problems encountered by the participants were or could have been solved by consulting the agent-based help facility. Only the order of the values of a quantity space was not covered by the agent-based help facility. A mechanism to verify this order as well as an explanation about the reason behind it should be included in the agents. Additionally, not too many types of errors occurred and it is interesting to notice that most of the ones that did occur were similar to problems participants had in the HOMER experiment, only fewer. In the HOMER experiment, 15 types of problems were encountered by four participants, see Table 6.1, while in the MOBUM experiment 8 types were encountered. Among these 8
problem types two new types emerged, Type 16 and Type 17. These problem types were new types which had not been observed in the HOMER Experiment.

Considering the four categories defined in Chapter 3, all the problem types found in the MOBUM experiment fall into the category of model building Concepts. Figure 6.3 illustrates the average number of errors that participants had in the HOMER experiment compared with the average of the number of errors encountered by the novices in MOBUM. Certainly, the assignment given to the novices cannot be compared to the assignment given to the participants in the HOMER experiment. Thus, we did not expect to find errors concerning structuring the model and scoping the model in the MOBUM experiment since that assignment did not enforce such decisions. Additionally, the drawing which was given to the participants in the SketchPad of MOBUM may have been of help for the participants because it already included the main model ingredients, the entity Garfield and the quantities. Moreover, the SketchPad information is also used by the agents in order to support the participants. On the other hand, problems concerning the categories of model building concepts and representing the model were expected. The novices did not have conceptual knowledge and also did not know how to represent that knowledge. Therefore, looking at the number of problems encountered (only in the model building concepts category) and how these problems were solved, or could have been solved, suggests that MOBUM indeed supports the model building process.

Experts

The results of the analysis of problems encountered by the expert users during their model building activities are also classified into Types. The enumeration of the types follows the same scheme as presented in Table 6.1 and when a problem encountered does not belong to any of the already defined types a new type is created. Similarly, the enumeration of each problem is followed by an example and a protocol transcription when appropriate.

Understanding and using relations (Type 7). One of the participants had a persistent problem with creating structural relations and was constantly creating and deleting model ingredients of this sort. It seems that the participant was not sure where the structural relation should be specified and was constantly trying new configurations

<table>
<thead>
<tr>
<th>Type</th>
<th>During the experiment solved by</th>
<th>Answer available via the agent-based help facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Experiment Leader</td>
<td>✔</td>
</tr>
<tr>
<td>3</td>
<td>Participant and agents</td>
<td>✔</td>
</tr>
<tr>
<td>4</td>
<td>Participant and agents</td>
<td>✔</td>
</tr>
<tr>
<td>8</td>
<td>Experiment Leader</td>
<td>✔</td>
</tr>
<tr>
<td>9</td>
<td>Concepts: point/interval: agent</td>
<td>□</td>
</tr>
<tr>
<td>10</td>
<td>QS upsidedown: Experiment Leader</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>Agents</td>
<td>✔</td>
</tr>
<tr>
<td>17</td>
<td>Agents</td>
<td>✔</td>
</tr>
</tbody>
</table>

Table 6.2
Summary of problems with model activities (Types).
6.3. RESULTS OF THE MOBUM EVALUATION

within the Structure Builder. Later, during a model fragment creation, the same participant requested help for specifying a relation, which was strange, as the person had already created such a relation using the Structure Builder. It seems that there was a conceptual confusion with regard to using relations.

Understanding and using dependencies (Type 10). Two participants demonstrated their confusion about dependencies.

P3: In fact, I would like to know what exactly is a proportionality...
P4: I am wondering where I need (to specify) the direction of the 'flow'

Another participant requested help regarding the specification of inequalities between a quantity and its value and in another instance this person inquired about creating an inequality featuring three quantities (Flow = P1 - P2). Yet another participant was puzzled when creating an inequality using 3 quantities. The problem of this participant was partially attributable to the tool and partially to the fact that the participant was unacquainted with this particular modelling issue.

Understanding model fragment organisation (Type 11). Although no help was requested, participant P3 demonstrated difficulties in understanding the differences between conditions and consequences.

Running a model without a scenario (Type 18). Three of the experts asked whether it was possible to run a simulation with their model at a time when they had not yet created a scenario. They had to be told that it was not possible because a scenario is essential for running a simulation.

The types identified above were compiled and categorised based on the questions that the participants asked to the experiment leader as well as on observations made by the experiment leader after having concluded the experiment. Figure 6.4 shows the occur-
rence of problems classified according to their type. In total the experts encountered 10 difficulties.

![Figure 6.4](image)

Number of problems classified according to the identified types (Experts).

**Running a simulation.** Table 6.3 gives an overview of the models produced by the participants. Although there were some missing model parts, all participants were able to finish their assignment. Two participants were able to actually run a simulation and get a behaviour prediction. The other two were able to run their model, but their simulation did not result in any states due to missing model ingredients in their models.

Using the simulation inspection tools embedded in MOBUM, these participants tried to verify, for instance, which model fragments fired during the simulation. In fact, all created model fragments should have fired during the simulation. So, they went on debugging their models in order to solve the problem. However, due to the short time left (each session lasted one hour), the debugging session was not successfully completed.

<table>
<thead>
<tr>
<th>Finish the Assignment</th>
<th>Run Simulation</th>
<th>Missing Parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3</td>
<td>✓</td>
<td>no</td>
</tr>
<tr>
<td>P4</td>
<td>✓</td>
<td>no</td>
</tr>
<tr>
<td>P6</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P7</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 6.3
State of the models created by the expert users.

**Help and support on model building activities.** The experts did not seem to use the agents to solve problems. When the experts got stuck, they consulted the experiment leader. However, the participants might as well have consulted the agents, as their problems could have been dealt with using the agent-based help facility, see Table 6.4. They only used the agents to assess the help potential by trying the help in different situations. However, when trying the agents they got inspired with the advices given.

Another support feature frequently consulted was the SketchPad, the drawing tool of MOBUM, which had the U-Tube drawing on it. The participants were consulting the drawing in order to verify if their model included all the details presented in the drawing.
6.3. RESULTS OF THE MOBUM EVALUATION

<table>
<thead>
<tr>
<th>Type</th>
<th>During the experiment solved by</th>
<th>Answer available via the agent-based help facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Experiment Leader</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>Experiment Leader</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>Experiment Leader</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>Experiment Leader</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 6.4
Summary of problems with Model Activities (Types).

Discussion and conclusions. Only a few types of problems were detected with the experts. Three out of the four types appeared also in the experiment with HOMER (Type 7, 10, and 11). It is interesting to notice that two of these types of problems, Understanding and using dependencies and Understanding model fragment organisation, were the main problems encountered during the experiment with Homer. The new type, Type 18, that appeared in the MOBUM experiment was Running a model without a scenario.

Categorising the encountered problems according to the four types as defined in Chapter 3 most of the problems belong to the class of model building Concepts and just one problem to the Representing the Model category. Figure 6.5 shows a comparison between the problems encountered in the MOBUM and HOMER experiments. Remarkably, in HOMER a total of 67 problems were observed while in MOBUM only 10 problems were observed. These results indicate that the features implemented in MOBUM, e.g. agent-based help facilities and SketchPad, helped in overcoming problems and therefore better supported modelers in building their models.

![Figure 6.5](image)

MOBUM experiment results (experts) versus HOMER experiment results.

Figure 6.6 summarises the usage of the agents by our novice and expert participants. As expected, the novices requested help in all the builders and the requested help was of different kinds. Experts on the other hand needed help mainly in the context of model fragments and they accessed the local agent most frequently. This may be explained by the fact that creating a model fragment involves manipulation of all the single model ingredients created previously, as well as determining relations between them. Apparently,
the local agent could serve as a kind of memory to the users by suggesting possible actions within the builder.

6.3.2 Results: User Interface Evaluation

As was also done in the experiment with HOMER, Chapter 3, the problems encountered while using the tool were catalogued according to the heuristic evaluation method [83]. This section provides an overview of the results found during the evaluation sessions, including a list of events where the identified heuristics were violated. The results for the novices and experts are presented separately as before.

Novices

General. The system allows users to select objects with which the participant may do nothing afterwards. The participants got confused because they thought that all that could be selected could be subsequently manipulated (which was not the case) (Violation of: Error Prevention).

Main window. Two of the subjects did not find the way to access the Model Fragment Builder. They sought their answer trying out all menu bar options. The experiment leader had to tell them that it was only possible to access the Model Fragment Builder through the main toolbar buttons (Violation of: Visibility of system status).
6.3. RESULTS OF THE MOBUM EVALUATION

Model fragment library window. Double clicking on the type of a Model Fragment was expected to initiate a new Model Fragment by one of the subjects (Violation of: Consistency and Standards).

Creating a new quantity space. In the Quantity Space Builder dialog, the 'Below' and 'Above' buttons were confusing when there was no value (initial situation). Two subjects had a problem with using these buttons under those conditions (Violation of: Visibility of system status). It was also noticed by the actions of one of the participants that the task of changing a point to an interval was missing in this window. In fact, this task is not supported by the system. In this case, the way out is to delete the created value and create a new one again. (Violation of: User control and freedom).

Scenario builder. At two instances, one of the subjects was confused by the Scenario Builder because it looked too similar to the Model Fragment builder. Consequently, the subject did not know for sure whether he was working on a scenario or on a model fragment (Violation of: Visibility of system status). A missing functionality detected by the actions of one of the subjects was the possibility to just copy and paste objects from a model fragment to a scenario (Violation of: User control and freedom).

Model Fragment Builder. One of the subjects intended to change the quantity space of a Quantity that belonged to a Model Fragment. However, the system does not allow the modification of a quantity if it is used elsewhere in the model (Violation of: User control and freedom).

Experts

General. User selections are used in MOBUM to determine the intended modelling actions. Multiple selections can therefore not be used to move several items as is normally done in mainstream graphical user interfaces. (Violation of: Consistency and Standards).

After giving an error message, the system aborts the task that is currently being worked on and the user is forced to start over (Violation of: Error prevention).

The system allows users to select objects with which the participant may do nothing afterwards. After having selected an object, the system allows the user to do nothing with it by blocking every action which will lead to inconsistencies. It would be more sensible to not let the user select the object in the first place. (Violation of: Visibility of system status).

Model fragment library. Double clicking a Model Fragment does not open the model fragment in the builder in order to edit its content (Violation of: Consistency and Standards).
Quantity Space builder. The value "zero" in the 'New QS' was not visible to one of the participants. (Violation of: Visibility of system status). The participant did not notice the default presence of the value "zero" in the 'New QS' dialog. He consequently added the value 'zero'.

Scenario builder. Two of the subjects complained about the absence of the copy and paste mechanism. The participants did not want to specify everything again like they had done in the model fragment builder.

P4: Can I copy and paste?
Is there a way to include these (all quantities) automatically?
P6: I am not going to model everything again? Right?

Discussion and conclusions. Figure 6.7 presents an overview of the heuristics violated in the MOBUM study, for novices as well as experts. The heuristic violations from the experiment with HOMER are also included for comparative purposes. Notice that the number of errors are averaged over the subjects. Most of the occurrences of errors and complaints are related to the Visibility of the System Status. This was true for both novices and experts which used MOBUM (but also in the HOMER experiment). Even when considering both, novices and experts, the users in MOBUM encountered much less usability problems than the users of HOMER. Additionally, some of the errors encountered were due to the existence of simple bugs, e.g. the double-clicking on a model fragment name in the Model Fragment Library and allowing the selection of objects that the user could not do anything with. These bugs should be fixed in a next version of the software.

Special attention should be given to redesigning the window for creating a new quantity space (i.e. the 'New QS Dialog'). Novices as well as experts experienced problems using this editor. Another point of disappointment for most of the participants, was the missing functionality of a 'copy' and 'paste' mechanism, especially while creating scenarios. Participants thought that it would be possible to copy and paste objects that had already been created thereby substantially reducing their efforts.

![Figure 6.7](image)

Heuristics violated in MOBUM (novices and experts) and HOMER.
6.3. RESULTS OF THE MOBUM EVALUATION

6.3.3 Reflection of the Users

This section provides an overview of the results obtained from the evaluation sessions including the analysis of interviews and a list of useful comments and suggestions made by the participants concerning the overall design of the tool. Phrases such as "What do you think about the environment?", "What do you like about the tool?", "What did you not like?", "What are your suggestions for improvement?" were frequently used during the interview to elicit more information from users.

What did you like about MOBUM? Table 6.5 shows the answers of the participants when asked to describe which features they liked most about the tool. Except for one participant (P6), all participants explicitly mentioned that they liked the agent-based help facilities and were positive about its usefulness. Only one of the participants did not comment on the 'visual appearance' of MOBUM (P2), all the others made positive comments about the overall user interface design (icons, graphics). Two participants (P3, P7) appreciated the compositional nature of the Structure Builder. One participant compared the Structure Builder to an ontology builder.

<table>
<thead>
<tr>
<th>Overall comments on good features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P1</strong> (i) The feedback and suggestions are a good feature, very helpful.</td>
</tr>
<tr>
<td>(ii) The help is essential.</td>
</tr>
<tr>
<td>(iii) Visually, diagrammatic, is very good.</td>
</tr>
<tr>
<td><strong>P2</strong> (i) It was not so difficult to build the model.</td>
</tr>
<tr>
<td><strong>P5</strong> (i) The agents are really useful. Especially, the local agent, which gives feedback in your model.</td>
</tr>
<tr>
<td>(ii) Visually, diagrammatic, is very good. The graphics are nice.</td>
</tr>
<tr>
<td><strong>P3</strong> (i) I like the agents. The overall approach to build models is quite ok.</td>
</tr>
<tr>
<td>(ii) I like the way it looks; The builders are self explanatory.</td>
</tr>
<tr>
<td><strong>P4</strong> (i) I liked the agents, especially the 'Cross builder' because it gives advises about the content.</td>
</tr>
<tr>
<td>(ii) The icons in the environment are clear. For example, in the Structure Builder, it looks like an ontology builder.</td>
</tr>
<tr>
<td><strong>P6</strong> (i) I liked it!</td>
</tr>
<tr>
<td><strong>P7</strong> (i) The agents put me on the trail.</td>
</tr>
<tr>
<td>(ii) I liked the 'Sketch Pad' .</td>
</tr>
<tr>
<td>(iii) The different ways to initiate the creation of things (menu + toolbar).</td>
</tr>
<tr>
<td>(iv) The interface moving things around.</td>
</tr>
<tr>
<td>(v) I think the Structure Builder is an improvement.</td>
</tr>
</tbody>
</table>

Table 6.5
General impressions about MOBUM (novices: P1, P2, and P5, experts: P3, P4, P6, and P7).

Comments and suggestions for improvements  During the interviews the participants were asked to suggest the features which in their view should be improved. They were also asked to comment on the overall design of MOBUM. The answers, summarised in Table 6.6, indicate that the comments/suggestions could be roughly grouped into two
broad categories, one concerning the features of the agents, the other concerning the user interface design, considering both its presentation and functionality.

The first group of general comments/suggestions points out the need for changes in some of the features of the agents. Two suggestions referred to the detailed textual information inside an agent dialogue (the advice). One problem with displaying advice was that all further suggestions for possible model building actions were based on the content of the model at that specific instant. The amount of information given was sometimes excessive. Indeed, two participants (P3, P4) acknowledged during the interview, that the advice highlighted too many issues for a user to process. P3 suggested that the advice should have an order of relevance and stressed the need for having a more direct feedback in the form of e.g. "Now, build your scenario". Similar suggestions were given by a novice (P5) who asked for more guidance on the sequence of model building tasks. Two participants (P1, P2) suggested changes concerning the 'static agents', How to and What is. In their view, the textual information contained 'too many similar words'. Also a participant suggested that the static agents could be made part of the main application (i.e. more global) instead of being available within each builder in order to avoid confusion in the builders.

One participant, P5, would like to have an overview of what the agents could do. The participant had only used some of the agents because he did not know in which way they could have helped him.

Although, within the static agents it is possible to navigate through the contents via hyper-links, one of the participants missed the functionality of searching for keywords.

The second group of comments/suggestions refers to the design of the user interface which also reflects the way in which model concepts are created and manipulated. One participant (P5) was skeptical about the model building tasks and the design of the tools. The participant thought both, tasks and tools, were extremely complex. One point cited was the quantity space concept attached to a quantity. For the participant, a quantity is a concept (e.g. 'weight') that has a set of possible values (skinny, normal and fat) and therefore it is not necessary to use another concept (the quantity space) to represent this set of values. Missing functionalities were noticed especially within the model fragment and scenario builders. They were pointed out by two of the participants, P3 and P6. Firstly, they suggested that a hide/show functionality should be implemented in order to aid one in focusing on just a subset of a model fragment’s aspects, thereby avoiding the risk of a cluttered visual representation of the complete model fragment. Secondly they suggested the implementation of a 'copy' and 'paste' mechanism. In fact, most participants were frustrated by the way in which a scenario had to be built, as they needed to create knowledge which was already obvious from the contents of a previously created model fragment. Two contradictory comments were made about the Structure Builder. While a novice (P2) thought that it was confusing to specify entities, relations and attributes inside the same builder, an expert, P7, commented that the Structure Builder was like an ontology builder. P4 complained about the way in which relations are built. Instead of selecting origin and destination and subsequently naming the relation, the participant would rather have preferred to draw a line between the entities involved and then to give it a name. The same participant also did not appreciate the fact of being obliged to specify
### 6.4. **COMPARISON STUDY BETWEEN MOBUM AND HOMER**

<table>
<thead>
<tr>
<th>Comments or suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P1</strong></td>
</tr>
<tr>
<td>(i) In the Builder's Help there is too many similar terms: &quot;The QS Builder is the place to create Quantity Spaces...&quot;</td>
</tr>
<tr>
<td><strong>P2</strong></td>
</tr>
<tr>
<td>(i) I suggest not to put all agents in all builders. The 'what is' and 'how to' could be placed as general help.</td>
</tr>
<tr>
<td>(ii) The Structure Builder is ambiguous because you have the <em>is-a</em> relation and also other relations between entities.</td>
</tr>
<tr>
<td><strong>P5</strong></td>
</tr>
<tr>
<td>(i) Extremely complex both the task and the tool. For instance, the idea of having the Quantity Space concept attached to an object (quantity).</td>
</tr>
<tr>
<td>(ii) It should have more guidance towards sequence in which the tasks should be performed. Like when booking a flight, you have 5 steps, now you are in 'step 1' go to 'step 2'.</td>
</tr>
<tr>
<td>(iii) Maybe it could have an explanation about what the agents do. I did not use others (only 'what is' and 'local') because I did not know what they could do for me.</td>
</tr>
<tr>
<td><strong>P3</strong></td>
</tr>
<tr>
<td>(i) Sometimes the agents gave too much advice with complex sentences. I think the advises should have an order of relevance. The agent does not say when you are finished. I would like to have something like: Now, build your scenario. The agents advises are not conclusive instead they are open feedback.</td>
</tr>
<tr>
<td>(ii) A help about 'calculus' should be more explicit.</td>
</tr>
<tr>
<td>(iii) The system should give precise feedback, for instance, concerning the order of the relations (LHS and RHS).</td>
</tr>
<tr>
<td>(iv) The Model Fragment Builder gets cluttered when a number of model ingredients are included.</td>
</tr>
<tr>
<td><strong>P4</strong></td>
</tr>
<tr>
<td>(i) The agents are sometimes a bit too much.</td>
</tr>
<tr>
<td>(ii) Defining structural relations were not straightforward, it was complicated. I prefer to draw a line between the entities and after that to give a name. Also, I would rather prefer to specify a relation in a Model Fragment without being obliged to create it beforehand in the other builder (Structure Builder).</td>
</tr>
<tr>
<td><strong>P6</strong></td>
</tr>
<tr>
<td>(i) Debugging the model is difficult.</td>
</tr>
<tr>
<td>(ii) Help: different levels of the problem that I have, you can not search for keywords.</td>
</tr>
<tr>
<td>(iii) The way of building a scenario: copy and past is missing.</td>
</tr>
<tr>
<td>(iv) The representation of the quantity on the screen is too much, not needed. I miss the 'hide/show' functionality.</td>
</tr>
<tr>
<td>(v) The notion of conditions and consequences confuses me. Two parts in one editor.</td>
</tr>
<tr>
<td><strong>P7</strong></td>
</tr>
<tr>
<td>(i) I did not like the extra steps: deciding what to include from a conditional model fragment.</td>
</tr>
</tbody>
</table>

**Table 6.6**

Comments and suggestions for improvement.

The building blocks before using them in the model constructs.

### 6.4 Comparison Study between MOBUM and HOMER

This section describes the comparison study between MOBUM and HOMER using novice model builders. The main goal of this comparison study was to see whether differences
could be found in the preferences of the participants for MOBUM or HOMER, and thus which one is more easy to use.

We anticipated that participants would prefer MOBUM above HOMER mainly due to two reasons. Firstly, because of the agent-based help facility implemented in MOBUM, participants would be able to find answers to questions such as "How to do this?" or "What does it mean?", whereas in HOMER no help is provided in this way. Secondly, MOBUM was expected to be easier to use, because the user interface of MOBUM looks more familiar to computer users. This is mainly attributable to the more common way of interacting with the user interface, e.g. the main options within the builders are available in the menu as well as in the toolbar following the general trend of mainstream desktop applications.

6.4.1 Method

Two situations with alternative sequences in which both programs were presented to the user are compared within this study. The situation in which the participant uses MOBUM first and subsequently HOMER is labelled MOBUM-HOMER for short. The inverse situation is labelled HOMER-MOBUM.

Participants

Twenty-eight first-year Psychology students were asked to participate in this study. Since the participants were first-year university students, we could be quite certain that they had little knowledge about building qualitative models. In addition, we ensured that none of the participants had had any previous contact with the two software systems being evaluated.

Design and Data Gathered

The participants were randomly divided into two groups of 14 students. One of the groups started working with MOBUM for one hour and then changed to HOMER using it for thirty more minutes. For the other group the order of the two programs was reversed, i.e. one hour using HOMER and subsequently 30 minutes using MOBUM.

Participants were asked to complete tasks which were tailored to match their level of competence to allow them to adequately experience the functionality of both systems. While working on the various assignments we kept in mind that the time available for performing the tasks was limited. Furthermore, the main goal of the assignments was to provide the participants with the possibility of using all the functionality of the systems so as to be better equipped to answer the questionnaires. The safest way of guaranteeing that the participants used all of the functionality of the system was to initially give them simple model building tasks and later guide them towards more complex aspects of model building.

At the beginning of each session, the experiment leader explained the experiment to the participant as well as the goals emphasising that we were not evaluating their model
building nor their computer skills but the quality and usability of the systems. They were then asked to complete the first questionnaire. The first questionnaire (A) was about the general attitude towards the use of computers. It gathers general information about the participants, such as, weekly time spent using computers, attitude towards computer tasks and preference of using a computer over watching TV.

After the introduction to the experiment, the participants were given eight minutes to read a short introduction about qualitative modelling. This introduction was accessible to them for the duration of the whole experiment. Subsequently, the assignment was handed out. Each participant was left alone in a room with semi-transparent/one-way glasses so that the experiment leader could monitor the actions of the participant without intervening. In the case the participants had problems they were allowed to call for help from the experiment leader.

The assignment consisted of building a Garfield model using each one of the two model building systems. Thus, first one hour with MOBUM and then half an hour with HOMER, or the other way around. The participants were then asked to complete four questionnaires. The first two questionnaires referred to their experiences with the systems: one referring to MOBUM (QM), the other to HOMER (QH). The third questionnaire (comp) referred to a direct comparison between the two systems and the fourth questionnaire (E) was concerned with a general evaluation of the experiment. Table 6.7 gives an overview of the sequence of tasks, the time to complete the tasks and questionnaires. Additional data was gathered by counting the model ingredients produced by the participants with the goal of verifying their productivity when working with the systems. Appendix C.3 presents the complete questionnaires.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Quest</th>
<th>Tasks</th>
<th>Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobum-Homer</td>
<td>A</td>
<td>8 min.</td>
<td>Mobum</td>
</tr>
<tr>
<td>Homer-Mobum</td>
<td>A</td>
<td>reading introduction</td>
<td>Homer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60 min.</td>
<td>Mobum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30 min.</td>
<td>HOMER</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QH</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Comp</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E</td>
</tr>
</tbody>
</table>

Table 6.7
Sequence of the questionnaires and tools in the experiment.

The QM and QH questionnaires, collected the opinions of the participants about the tools, MOBUM and HOMER, on the (ranked) issues of: overall user interface design, the ease of use of the system, the helpfulness of the help system provided by the tool (see Table 6.8). The questions were measured using a 5-point scale ranging from 1 ("I disagree") to 5 ("I agree"). Only one option could be selected.

The Comp questionnaire forced the participants to make a direct comparison between the two systems. In this questionnaire a small adjustment was made in order to account for the order in which both programs were used, i.e. either MOBUM-HOMER or HOMER-MOBUM. In the case the participant worked first with MOBUM the question in this questionnaire takes the following format: MOBUM is easier to user than HOMER. Had the participant used HOMER first the question would be the same but then formulated with the order of the program names inverted: HOMER is easier to user than MOBUM.

Finally, the last questionnaire was concerned with a general evaluation of the experiment, the assignment and the software packages. It includes issues such as the clarity
CHAPTER 6. MOBUM EXPERIMENT

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The various screens of [MOBUM][HOMER] are insightful and understandable</td>
</tr>
<tr>
<td>2</td>
<td>The program [MOBUM][HOMER] is easy to use</td>
</tr>
<tr>
<td>3</td>
<td>The use of colour in [MOBUM][HOMER] is insightful and understandable</td>
</tr>
<tr>
<td>4</td>
<td>The error messages that [MOBUM][HOMER] presents are good</td>
</tr>
<tr>
<td>5</td>
<td>The help and support that [MOBUM][HOMER] offers is good</td>
</tr>
<tr>
<td>6</td>
<td>The icons used in [MOBUM][HOMER] are insightful and understandable</td>
</tr>
<tr>
<td>7</td>
<td>The navigation between the different screens of [MOBUM][HOMER] is user friendly</td>
</tr>
</tbody>
</table>

Table 6.8
Questions of QM and QH questionnaires.

of the tasks in the assignment and the suitability of the systems. See Table 6.9 for the complete list.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The assignment in the experiment was clearly formulated</td>
</tr>
<tr>
<td>2</td>
<td>Software systems, such as the ones used in the experiment, make the users to think structured about...</td>
</tr>
<tr>
<td>3</td>
<td>Simulation models, such as the ones in the experiment, ...</td>
</tr>
<tr>
<td>4</td>
<td>The error messages that [MOBUM][HOMER] presents are good</td>
</tr>
<tr>
<td>5</td>
<td>Software systems, such as the ones used in the experiment, should be better used in the following levels (More than one option can be selected)</td>
</tr>
</tbody>
</table>

Table 6.9
General questions about the experiment.

6.4.2 Results

Appreciation of each tool. A summary of the results of the QH and QM questionnaires is shown in Table 6.10. Overall there was a strong preference for MOBUM over HOMER. Consequently, a significant difference between the two conditions: the evaluation of MOBUM, mean 3.6 with std. deviation 1.1, and the evaluation of HOMER, mean 2.4 with std. deviation 1.2 (1 = Strongly disagree, 5 = Strongly agree). See also Figure 6.8 and Table 6.10 for the details.

The result of the fifth question (*The help and support that [MOBUM/HOMER] offers are good.*) is quite interesting and tells us that there has been an improvement by MOBUM when compared to HOMER. QM scored with a mean of 3.5 while QH’s mean was 1.8.

The internal consistency based on the average inter-item correlation, indicates that the results of both questionnaires are sufficiently reliable. Coefficients of reliability of .83 and .76 (Cronbach’s $\alpha$) were obtained for the QH and QM questionnaires respectively.

A variance analysis was performed to find out if the order in which the experiments were performed, i.e. MOBUM first and then HOMER or vice-versa, would have an influence on the results of the experiment. The seven responses in each condition for each participant were averaged so that each participant contributed one data point in each condition for the analysis. The ANOVA test of the two conditions show that the two groups do not differ significantly in their responses (F(1, 26) = 2.592, p = .119). From this analysis we can therefore conclude that the data obtained in these two situations did not differ significantly and as such no measurable effect on the appreciation of any of the tools or the
6.4. COMPARISON STUDY BETWEEN MOBUM AND HOMER

<table>
<thead>
<tr>
<th>QH-Question 1</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>QH-Question 2</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>2.4</td>
<td>1.4</td>
</tr>
<tr>
<td>QH-Question 3</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>QH-Question 4</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.0</td>
<td>1.2</td>
</tr>
<tr>
<td>QH-Question 5</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>2.5</td>
<td>1.2</td>
</tr>
<tr>
<td>QH-Question 6</td>
<td>28</td>
<td>1.00</td>
<td>3.00</td>
<td>1.8</td>
<td>1.0</td>
</tr>
<tr>
<td>QH-Question 7</td>
<td>28</td>
<td>1.00</td>
<td>4.00</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>QH-Total</td>
<td>28</td>
<td></td>
<td></td>
<td>2.4</td>
<td>1.2</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QM-Question 1</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>QM-Question 2</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.6</td>
<td>1.1</td>
</tr>
<tr>
<td>QM-Question 3</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.3</td>
<td>1.3</td>
</tr>
<tr>
<td>QM-Question 4</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.8</td>
<td>1.1</td>
</tr>
<tr>
<td>QM-Question 5</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.1</td>
<td>.9</td>
</tr>
<tr>
<td>QM-Question 6</td>
<td>28</td>
<td>2.00</td>
<td>5.00</td>
<td>3.5</td>
<td>1.3</td>
</tr>
<tr>
<td>QM-Question 7</td>
<td>28</td>
<td>1.00</td>
<td>5.00</td>
<td>3.8</td>
<td>1.0</td>
</tr>
<tr>
<td>QM-Total</td>
<td>28</td>
<td></td>
<td></td>
<td>3.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.10
Answers on the general impressions about HOMER and MOBUM. The mean score of HOMER based on QH is 2.4 whereas MOBUM scores 3.6 using QM.

The score in general was detected. Table 6.11 summarises the results of the ANOVA variance analysis.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>Mobum-Homer</td>
<td>14</td>
<td>39.2143</td>
<td>9.25019</td>
<td>2.47222</td>
<td>33.8734</td>
<td>44.5552</td>
<td>22.00</td>
</tr>
<tr>
<td>Homer-Mobum</td>
<td>14</td>
<td>44.5000</td>
<td>8.08370</td>
<td>2.16046</td>
<td>39.8326</td>
<td>49.1674</td>
<td>33.00</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>41.8571</td>
<td>8.53895</td>
<td>1.68930</td>
<td>38.3910</td>
<td>45.3233</td>
<td>22.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Between Groups</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Groups</td>
<td>195.571</td>
<td>26</td>
<td>7.5456</td>
<td></td>
<td>.119</td>
</tr>
<tr>
<td>Total</td>
<td>2157.429</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.11
ANOVA analysis demonstrating that the order in which the programs were presented to the participants had no influence on the overall results of QM and QH.

Additionally, there is a significant preference for MOBUM over HOMER when only the first tool is measured in both of the situations described above (z=2.7, p=0.007). See table 6.12 for details. The same happens when only the second tool is measured (z=3.6, p < 0.0005). See table 6.13 for details.

Again, the reliability of the results of the QH and QM questionnaires in their roles as evaluators of the first tool, MOBUM =.71 and HOMER =.75 (Cronbachs α) as well as in their roles as evaluators of the second tool, MOBUM =.80 and HOMER =.88 (Cronbachs α).
The means for the QM and QH questionnaires.

<table>
<thead>
<tr>
<th>The order of the tools</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first evaluated tool</td>
<td>Mobum-Homer</td>
<td>14</td>
<td>18.64</td>
</tr>
<tr>
<td></td>
<td>Homer-Mobum</td>
<td>14</td>
<td>10.36</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>The first evaluated tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mann-Whitney U</td>
<td>40.000</td>
</tr>
<tr>
<td>Wilcoxon W</td>
<td>145.000</td>
</tr>
<tr>
<td>Z</td>
<td>-2.674</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.007</td>
</tr>
<tr>
<td>Exact Sig. [2*(1-tailed Sig.)]</td>
<td>.007(a)</td>
</tr>
</tbody>
</table>

Table 6.12
The Mann-Whitney (non-parametric) test for the first QM and QH questionnaires, that is MOBUM in the MOBUM-HOMER condition and HOMER in the HOMER-MOBUM condition.

α) was verified. Thus, even when the questionnaires are used to evaluate only the first or only the second tool, the significant preference towards MOBUM is apparent.

Comparison Questionnaire  This questionnaire asks the participants to compare the first tool used with the second one. The results of these questionnaires are supplemental to QH and QM and strengthen the results obtained from their analysis. Notice that since this questionnaire is used to compare the first tool to the second one, the questions will differ according to the order in which the programs were presented to the participant (MOBUM-HOMER and HOMER-MOBUM). For simplicity, we considered the scale ranging from "I strongly agree" to "I strongly disagree" to be symmetric, that is to say, for instance, that when the answer to the question The screens in MOBUM are more insightful and understandable than the screens in HOMER" is "I agree", the assumption is made that this answer is equivalent to answering "I disagree" to the question: The screens in HOMER
6.4. COMPARISON STUDY BETWEEN MOBUM AND HOMER

<table>
<thead>
<tr>
<th>The order of the tools</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>The second evaluated tool</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobum-Homer</td>
<td>14</td>
<td>8.96</td>
<td>125.50</td>
</tr>
<tr>
<td>Homer-Mobum</td>
<td>14</td>
<td>20.04</td>
<td>280.50</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.13
The Mann-Whitney (non-parametric) test for the second QM and QH questionnaires, that is MOBUM in the MOBUM-HOMER condition and HOMER.

are more insightful and understandable than the screens in MOBUM. Thus, the disliking of one tool is viewed as being equivalent to the liking of the other tool. There are arguments against this reasoning as the two statements are clearly not interchangeable. As such, this questionnaire should only be viewed as supportive of the preceding two and not as an entity on its own.

The results of this comparative questionnaire support the previous findings where there is a significant preference of MOBUM over HOMER (z=4.4, p<0.0005). The reliability factor is high (Cronbach's α=9). See table 6.14 for details.

<table>
<thead>
<tr>
<th>The order of the tools</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobum-Homer</td>
<td>14</td>
<td>21.32</td>
<td>298.50</td>
</tr>
<tr>
<td>Homer-Mobum</td>
<td>14</td>
<td>7.68</td>
<td>107.50</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.14
The Mann-Whitney of the Comp questionnaire.
**General attitude questionnaire** The first questionnaire in the experiment is about the general attitude of the participants towards computers. The results of this questionnaire are presented here just as additional information about the profiles of the participants.

Table 6.15 shows that most participants think they spend a normal amount of time using computers. The hours per week spent on working with the computer has a modal of 30 minutes and a mean of about 10 hours.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Valid</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very rarely</td>
<td>1</td>
<td>3.6</td>
<td>3.6</td>
<td>3.6</td>
</tr>
<tr>
<td>rarely</td>
<td>6</td>
<td>21.4</td>
<td>21.4</td>
<td>25.0</td>
</tr>
<tr>
<td>normal</td>
<td>9</td>
<td>32.1</td>
<td>32.1</td>
<td>57.1</td>
</tr>
<tr>
<td>frequently</td>
<td>9</td>
<td>32.1</td>
<td>32.1</td>
<td>89.3</td>
</tr>
<tr>
<td>very frequently</td>
<td>3</td>
<td>10.2</td>
<td>10.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.15
The computer usage of the subjects in the experiment..

About 70% of the subjects say they have very little programming experience, 21% said they had little experience and only 7% said to have some experience, see Table 6.16. None of the subject claimed to have much or very much experience with programming.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very little</td>
<td>20</td>
<td>71.4</td>
<td>71.4</td>
</tr>
<tr>
<td>little</td>
<td>6</td>
<td>21.4</td>
<td>21.4</td>
</tr>
<tr>
<td>normal</td>
<td>2</td>
<td>7.1</td>
<td>7.1</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 6.16
The programming experience of the subjects..

93% of the subjects affirmed to have private access to computers, only about 7% stated not to have private access to computers, Table 6.17. The Dutch average for 2001 was about 74%\(^1\) having private access to computers.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>26</td>
<td>92.9</td>
<td>92.9</td>
</tr>
<tr>
<td>no</td>
<td>2</td>
<td>7.1</td>
<td>7.1</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 6.17
The access to private computers.

The majority of the subjects were quite positive towards working with a computer. The average was above the middle of the scale (scaling from 1 to 5, with a middle of 3), as Table 6.18 confirms.

Finally, it seems that the subjects slightly prefer watching TV to working with a PC (see Table 6.19).

\(^1\)This figure is taken from the CBS database. The figure was taken on November 2001
6.4. COMPARISON STUDY BETWEEN MOBUM AND HOMER

<table>
<thead>
<tr>
<th>Valid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>disappointing</td>
<td>2</td>
<td>7.1</td>
<td>7.1</td>
<td>7.1</td>
</tr>
<tr>
<td>no opinion</td>
<td>7</td>
<td>25.0</td>
<td>25.0</td>
<td>32.1</td>
</tr>
<tr>
<td>nice</td>
<td>18</td>
<td>64.3</td>
<td>64.3</td>
<td>96.4</td>
</tr>
<tr>
<td>very nice</td>
<td>1</td>
<td>3.6</td>
<td>3.6</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.18
Attitude towards working with a computer.

<table>
<thead>
<tr>
<th>Valid</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>disagree</td>
<td>8</td>
<td>28.6</td>
<td>28.6</td>
<td>28.6</td>
</tr>
<tr>
<td>partially disagree</td>
<td>6</td>
<td>21.4</td>
<td>21.4</td>
<td>50.0</td>
</tr>
<tr>
<td>neutral</td>
<td>8</td>
<td>28.6</td>
<td>28.6</td>
<td>78.6</td>
</tr>
<tr>
<td>partially agree</td>
<td>4</td>
<td>14.3</td>
<td>14.3</td>
<td>92.9</td>
</tr>
<tr>
<td>partially agree</td>
<td>2</td>
<td>7.1</td>
<td>7.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.19
Preference of watching TV compared to working with a computer (dis)agreement with the statement: "I prefer working with a computer above watching TV".

General evaluation of the experiment
This final questionnaire was a general evaluation of the experiment and the presented software. It included questions about the clarity of the assignment, the use of the software in general, the use of the software in educational contexts, and building models in educational contexts.

Figure 6.9 shows the results of this questionnaire. Two questions in this questionnaire Q1: "Understandability of the assignment" and Q2: "The software promotes structured thinking about the behaviour of the system" had a positive trend; most subjects were mildly positive about them. Only one subject in Q1 thought the assignment was unclear. In Q2, the answers of the participants are more regularly distributed.

The answers towards Q3 were neutral ("Building models helps understanding the behaviour of a system"). Most positive answers have been given to Q4 ("Interactive learning, by articulating and organising knowledge in a computer tool, is preferred over gaining knowledge by reading and listening").

Measuring the Productivity
A method for measuring the productivity [54] was developed in order to evaluate and score the models created by the participants. This is done in a straightforward manner by enumerating the model parts, categorising them and extracting an overall figure of merit for the productivity of the participant. These model parts are the building blocks (the entity hierarchy, relations, attributes, quantities and quantity spaces) and the specification of these building blocks within model constructs (model fragments and scenarios). The scoring is based on three categories:

2A more complex method for measuring complexity of Garp Models has been developed by [77]
Results of the general evaluation.

C Correct model parts

E Erroneous model parts

S Superfluous model parts

The values for these separate categories were measured for each model, either using MOBUM or HOMER, and a total score was calculated as:

\[ P = C - (E + S), \]

where \( P \) is the total productivity. For each participant two models, one created with MOBUM and one created with HOMER, were scored using the above method.

The results after scoring the models produced with MOBUM and HOMER using the above-mentioned method shows that there is no significant difference between the models produced with one or the other (the T-test gives \( t=1.93, \text{df}=27, p=0.063 \), see Table 6.20). Analysing the correct model parts only, the difference however is significant (\( t=2.049, \text{df}=27, p=0.05 \), see Table 6.21). Although there is a 'small' significant difference between the two groups, no strong conclusions can be drawn due to high variance among the subjects (MOBUM: mean=18.6 st. with deviation=16.6, and HOMER: mean=13.5 st. with deviation=11.0). The data shows an extremely high variance of the score, and thus an extremely high standard deviation. The created models vary highly between the individual subjects and even between the conditions of individual subjects.

Summary of the Results

The results of the comparison study show a strong preference for MOBUM over HOMER. For all questions in Questionnaires QM and QH, MOBUM scored higher than HOMER.
6.4. COMPARISON STUDY BETWEEN MOBUM AND HOMER

<table>
<thead>
<tr>
<th>Pair</th>
<th>Total Score Mobum</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18.6429</td>
<td>28</td>
<td>16.63823</td>
<td>3.14433</td>
</tr>
<tr>
<td></td>
<td>13.5357</td>
<td>28</td>
<td>10.98309</td>
<td>2.07561</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pair</th>
<th>Total Score Mobum * Total Score Homer</th>
<th>N</th>
<th>Correlation</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>.556</td>
<td>.002</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the Difference</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1071</td>
<td>13.93726</td>
<td>2.63390</td>
<td>-.2972</td>
<td>10.5115</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.20
The T-Test for comparing the total of results from MOBUM with the total of results from HOMER.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Correct scores for mobum</th>
<th>N</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.0357</td>
<td>28</td>
<td>18.08310</td>
<td>3.41739</td>
</tr>
<tr>
<td></td>
<td>15.1429</td>
<td>28</td>
<td>11.16779</td>
<td>2.11051</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pair</th>
<th>Correct scores for mobum * Correct scores for homer</th>
<th>N</th>
<th>Correlation</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>.545</td>
<td>.003</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
<th>95% Confidence Interval of the Difference</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8929</td>
<td>15.21778</td>
<td>2.87589</td>
<td>-.0080</td>
<td>11.7937</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.21
The T-Test for comparing only correct model parts.

The results are weak for the measure of the productivity of both tools. We expected that by being more easy to use and giving more guidance in terms of the help engine, the
difference in productivity would be higher. However, there was a high variance among the participants and therefore no strong conclusions can be drawn with respect to this issue.

6.5 Discussion and Conclusions

Qualitative analysis of system behaviour is an important aspect of science teaching. MOBUM is a tool that supports users in the process of building qualitative models in a learning-by-doing context, and demonstrates that users can learn satisfactorily about modelling. The main conclusions that can be drawn from the research reported in this chapter are the following:

1. Besides only allowing the building of syntactically correct models, MOBUM provides model building support from a content point of view. The help system implemented in MOBUM has knowledge about the model construction process and the actual status of the model. This help system coaches users in building models beyond the syntactic level. The results of our first study have shown that the help system was essential in enabling the participants to accomplish their tasks. In fact, most of the problems of the participants were solved by interacting with the implemented agent-base help facility. Only in a few instances did the experiment leader intervene or had to be called in order to provide help.

Analysing the usage of the agents by the novices as well as by the experts participants, we conclude that the usage of the help system by novices was more distributed within builders and agents. The novices did not possess conceptual knowledge and as such had to acquire it in order to understand their task. Therefore, as expected, we can see that the What is agent was regularly used in all builders. Guidance in acquiring model building skills involves coaching the user in:

- understanding the ontology for building qualitative models
- manipulating tools with which they can represent the model
- not allowing them to construct erroneous models (only allowing syntactically correct models by providing only correct primitives to build models)
- learning the way that model parts interact with each other.

Experts' usage of the agents was more centred around the Model Fragment Builder and the most used agent was the Local agent, which provides feedback with regard to the actual state of the model. The confusions and misconceptions that arose could be solved by consulting the agents and that was one of the main goals of the help system: to decrease the cognitive load. As one of the participants mentioned: "the agents put me on track".

2. The user interface of MOBUM was clearly appreciated by all participants in both experiments. The comparison study, between MOBUM and HOMER, documented in this chapter strengthens the positive attitude towards the user interface
of MOBUM. In both situations, HOMER-MOBUM and MOBUM-HOMER, there was a strong preference for MOBUM over HOMER.

The development of the MOBUM system incorporated a number of design decisions. These decisions were mainly based on the aims of minimising the cognitive load of the user while preserving the generality of a domain independent approach. As a result, the agent-based help facility is applicable to any model constructed with the system. The generality of the help system is one of the main features of MOBUM, and hence every extension to the help system should be carefully considered so as to maintain this generality. A number of suggestions for improvements can be pointed out. They are summarised below and can be a source for future developments.

Help System Possible improvements to the Help system and agents are the following. Firstly, making the language within the static help, i.e. How to and what is, more clear and insightful. The text to explain concepts and how to perform tasks can be more elaborate and should avoid repeating words, as suggested by one participant. Secondly, allowing search for keywords and also a mechanism to directly access the desired information should be implemented. For instance, the user could select an object on the screen and subsequently call for the What is agent. Thirdly, the use of priorities within the advices could be added by ranking the rules. Then not too many advices would be available at the same time and the agent could control which advice had already been given so as to not to show it next time.

The first and second items do not require any significant change to the present system. Therefore, this can be easily done without substantial adjustment and would increase user satisfaction with the help system. The third item requires modifications of the reasoning engine. However, these modifications are domain independent.

General User Interface Design Some problems with the design of the user interface have been pointed out and deserve further attention. At least, four main points should be addressed. The most important one concerns the missing functionalities of 'copy'/'paste' and 'hide/show', which were a major point of dissatisfaction with the current system. Especially when creating a Scenario, this was a cause of frustration and when a model fragment consisted of a large number of ingredients the 'hide/show' option was needed. Secondly, a re-implementation of the editor for building a new quantity space is needed, because it currently lacks important functionality. Thirdly, the way of including a model fragment as a condition must be changed. This third point can be easily resolved by adding the 'copy' and 'paste' functionality, so that the user could select only the parts that he wants and copy them into another model fragment. The fourth point concerns the concept of a 'quantity space'. This is a difficult concept to grasp for the novices. It is not the concept of a quantity having a set of possible values that is difficult to understand but the idea of having a name to refer to the list of possible values. Depending on the audience, this concept should be better addressed or camouflaged.