CHAPTER 6 EVALUATING MAPTUTOR

6.1 Introduction

This chapter describes a series of preliminary experiments carried out to appraise MapTutor. It also attempts to explain some internal workings of MapTutor by means of real-world events generated by the subjects who took part in the experimental sessions. Given the limited amount of time available for the experiments, the ultimate question about the program's educational effectiveness — i.e., *Does* MapTutor *really teach how to map?* — could not be answered. Nonetheless, this chapter also proposes a systematic large-scale evaluation which could lead to an answer to this fundamental question.

6.2 Preliminary Evaluation

The objectives of the preliminary series of experiments conducted to appraise MapTutor were to gather information which could be used to (1) assess the extent to which the program is able to diagnose correctly; (2) evaluate the impact of the program's feedback; and (3) to attempt to validate some design decisions presented in this book.

6.2.1 Data Collection

Data collection was carried out by means of multiple measures which complement the weaknesses of each other. These measures are described below.

The Use of Reports. MapTutor's end-of-session reports (see **Section 5.8**), created automatically by the program at the end of each experimental session, were used. As seen in **Section 5.8**, such a report contains general information about the session as well as a list of all moves (i.e., drawing tasks) the learner carried out in the map pane together with each task's evaluation (if appropriate). Special attention was given to this latter information, because by using it, it is possible to understand better both the internal workings of the program and some learners' behaviours. Furthermore, the validity of the reports themselves as a research tool was evaluated in a practical situation.

MapTutor had also the capability of generating another type of report by using data gathered from evaluation questions asked during the session. However, this would have required the use of on-line questionnaires whenever (1) the diagnostic

was anything but CORRECT_LINK, and (2) corrective feedback was provided. See **Figure 6–1** and **Figure 6–2** for examples of each of these types of questions. This idea was abandoned and not used during the experimental studies described here, because it would be too annoying for the subjects[1].

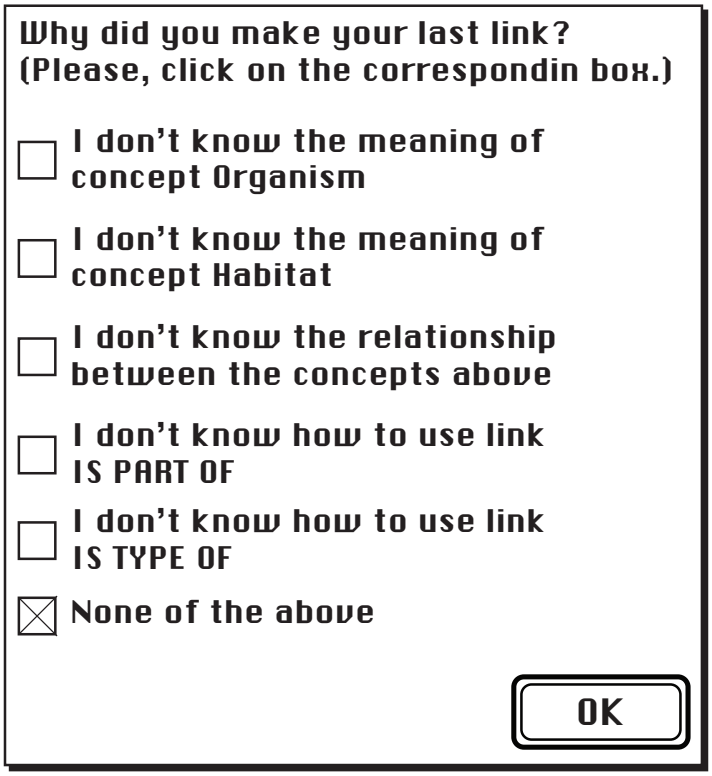

FIGURE 6–1: EXAMPLE OF AN ON-LINE DIAGNOSTIC EVALUATION QUESTION

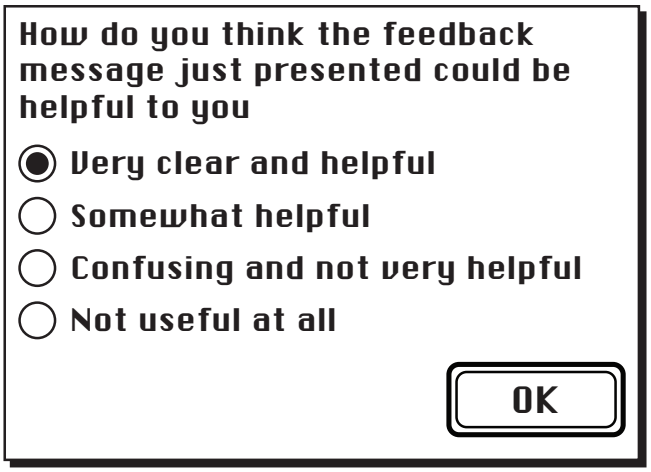

FIGURE 6–2: EXAMPLE OF AN ON-LINE FEEDBACK EVALUATION QUESTION

[1] This decision soon proved to be appropriate because even without the presentation of those evaluation questions, the subjects appeared to have become upset by many feedback messages. This was especially true of those wrong links the subjects made by mere slips (e.g., when the subject arrived at the correct link name but drew the link with a wrong one simply because the current link name was a wrong one).

Observations. Both in-session, structured and post-session, unstructured observations were conducted. The former focused on taking notes about certain selected behaviours (e.g., the apparent reaction of the subject to a feedback message), whereas the latter consisted in audio-recording all sessions. Unstructured observation by means of video-recording was also carried out with the first two subjects in the evaluator's own laboratory[2]. In the latter, all user's activities and program's behaviour were recorded, and proved to be a very useful research tool as it allowed subtler information[3] than otherwise. The observer did not normally have to intervene, but sometimes, the subjects were asked to comment and help was provided when they were really stuck.

The observations focused on the subjects' reactions about[4]:

- **Friendliness**. Can the subjects find information easily? Do they understand how to begin, what to do next, and how to proceed? Do they use the short-cuts provided by the mapping tools? Does the subject miss any facility or feature? Do they feel trapped by the program's syntax/semantics? Do they understand the visual clues?
- **Feedback**. Do the subjects receive well-timed feedback? Are the feedback messages appropriate for the subjects' needs? Is the program responsive to the learners' mistakes?

Answers to some of the questions above could also be extracted from the subsequent analysis of the reports[5] generated by MapTutor as well as by using questionnaires.

The Use of Questionnaires. According to Flagg (1990), the great disadvantage of the use of questionnaires is what she calls the response effect, which is the tendency of the subjects to provide inaccurate answers, such as conventional, thoughtless or evasive answers, or even ones intended to please or flatter the

[2] The laboratory in which the other experimental sessions were carried out was also able to provide video-recording facility, but the cost was too high, because by using this facility, the subject would only be able to see the program's output through a television screen with very poor quality of image.

[3] For example, by reviewing the video-tape of a session and observing the mouse movements, one can get some clue as to whether the subject is floundering.

[4] Some of these questions have been suggested by Flagg (1990).

[5] For example, by observing that the learner deleted a link and then drew it again with a new name may indicate that she did not know how to use the Tool Rename. Information like this is readily available in the end-of-session reports.

evaluator[6]. Even so, the use of multiple measures, such as observations and reports generated by MapTutor, helped to alleviate the bias. The questionnaire utilised at the end of each experimental session consisted basically in questions asking the subjects their opinion about their experience after using MapTutor and some more background information (e.g., past experience with computers). All questions, except one, were close-ended ones which use a five-point measurement scale. The questionnaire used in the experimental studies takes as a model the format presented in Shneiderman (1987, Chapter 10), and can be found in **Appendix E**.

Short, Semi-structured Interviews. At the end of each experimental session, and after the subject had answered the questionnaire described above, the evaluator asked further questions in order to clarify some points in the questionnaire and try to elicit more detailed information concerning shortcomings and strengths of the program. The questions depended on both each subject's behaviour during the session (in the case where some interesting events took place) and the program's specific responsiveness. During these short interviews, the participants were asked or prompted to discuss and explain any interesting behaviour they showed during the session.

Think-aloud protocols (Ericsson & Simon, 1984) were not used, because this technique is suspected to increase the cognitive load the learner is faced with (see, e.g., Flagg, 1990; see also Preece, Rogers, Sharp, Benyon, Holland, & Carey, 1994)[7]. Moreover, it was expected that the subjects would be novices to the task and thus would already be burdened by it. Besides, given the background noise in the laboratory where the second run of experiments were carried out, this measure would have been useless anyway. Nevertheless, the subjects were instructed to feel free to verbalise or ask questions when needed. As a result, some did render some spontaneous and useful verbalisation.

6.2.2 Quality of Measures

The external validity of the measures was constrained by time, availability of equipment and volunteers, as well as environmental disturbance. The latter caused trouble not only for the evaluator but also for the subjects. For example, the last session was carried out under intense noise caused by two outsiders chatting and

[6] Given that all volunteers were PhD student colleagues of the evaluator, they would not say that the program was worthless, would they?

[7] Ericsson & Simon (1984) in their influential book on protocol analysis do not believe that reflection and verbalisation interfere very much with one's thinking process, but the comparisons they carried out to support this contention are mostly based on experts' think-aloud, not with learners'.

laughing aloud. Not surprisingly, this subject had the worst performance, and produced the most curious and unexpected mistake.

Regarding the subjects, the sample was a non-random, convenience one. The heterogeneity of the sample was also limited by the volunteers who were readily available to the researcher and willing to take part in the experiments. Hence, it is hard to tell which findings are limited to the particular group of people studied. Nonetheless, despite the apparent non-heterogeneity of the sample, the results presented in **Appendix D** do indicate that both their background knowledge and way of tackling the mapping task were rather varied.

6.2.3 Method

Subjects. The subjects were five PhD students of the School of Cognitive and Computing Sciences (COGS) and two PhD students of the Science Policy Research Unity (SPRU), University of Sussex. They were all volunteers. Most of them were familiar with at least one type of computer, mouse and drawing/painting program, which may have facilitated their handling (e.g., dragging concepts) of MapTutor's objects. Most were also familiar with at least two non-graphical learning strategies[8]. Regarding graphical learning strategies, their prior knowledge is doubtful. For example, some said that they were familiar with networks, but it is very likely that networks (as described in **Chapter 2**) had been mistaken by semantic networks with which many subjects were certainly familiar given that they had major interests in knowledge-based systems. Also, schematisation (as de scribed in **Chapter 2**) might have been mistaken for outlining, since the first is not a very common learning strategy.

The information above about the subjects' background knowledge was gathered by means of analysis of the questionnaires and by observing the subjects' performances. Only one subject actually proved to be familiar with graphical learning strategies. Moreover, it could also be clearly observed that their degrees of expertise with the computer varied a lot, notwithstanding their expertise status stated in the questionnaires.,

Materials. A full version of MapTutor[9], as described in **Chapters 3** to **5**, was used by all the subjects in the second setting of experiments, whereas the two subjects who took part in the first run of experiments used a slightly modified

[8] Almost all subjects used underlining/highlighting plus rereading. Incidentally, they seemed to realise that, as, e.g., Good & Brophy (1990) suggest, underlining/highlighting used alone was useless. That is, these techniques appeared to be useful only as selection tools which helped the learner to focus her attention on the important points of the text when rereading.

[9] The program had a minor bug fixed between the second and third days of experiments.

version (see below). However, some features were either masked out or the subjects were asked not to use them. The masked features included those auxiliary teaching procedures described in **Section 4.9.2**. Also, the subjects were asked not to use either pull-down menus, or the help or tutorial programs. As already mentioned in **Chapter 3**, a biologist evaluated the accuracy of the program's domain.

The experiments were divided (for practical convenience) into two settings. The first setting was the evaluator's own laboratory, and the computer utilised was an Apple Macintosh Performa 630 with a high-resolution 15" colour monitor attached to a video-out recording equipment by FOCUS Enhancements, Inc. Two subjects ($\mathbf{S}_1$ and $\mathbf{S}_2$) took part in this first run of experiments, which was also used for establishing more precisely the procedure to be followed in the second setting, as well as for adjusting some tutoring parameters. Thus, these two initial experimental sessions also served as a preparation for the next series of experiments. Nevertheless, these experiments did provide useful data, so that presenting their analyses is worthwhile. The diagnosis precedence used in these initial experiments was,

$$\textit{MisunderstandsLinkMeaning} \prec \textit{MisunderstandsConcepts} \prec \textit{MisunderstandsRelationship}$$

and the confidence threshold of Procedure *MisunderstandsRelationship* (see **Table 4–2 on page 88**) was set at *0.6*.

The second setting was the multimedia laboratory of COGS, University of Sussex, and the computer utilised was an Apple Macintosh Quadra 950 with a high-resolution 20" colour monitor. Both computers utilised have similar performance, but the first computer runs a bit faster than the second one.

In addition, a micro-cassette audio recorder was used in all sessions as well as a short evaluation questionnaire (see **Appendix E**). Printed forms of end-of-session reports generated by MapTutor were also used.

Procedure. The subjects were given a given brief introduction to both the program and the mapping strategy. Then, they were asked to use the program. Part of the study consisted in observing the subjects interacting with MapTutor and taking notes about some selected events generated by both the program and the subjects, as described above in **Section 6.2.1**. At the end, the subjects were asked to give their final impressions (both verbally and by filling in the evaluation questionnaire) about the program.

The participants were told in general terms what they were expected to do, but no written instruction was provided (i.e., all instructions were verbal). The basic workings of MapTutor, such as how to select a concept or link name, were also sometimes demonstrated upon the subject's request. However, some facilities which were expected to be intuitive were not demonstrated so that the experimenter could evaluate whether those assumptions would hold true.

The contents of the verbal introduction provided by the experimenter was similar to the **Getting Started** section of the tutorial program described in **Section 5.5**. During the introductory instruction, the experimenter also described the purpose of the experiment and stressed that what was going to be tested was the program not the subject. After being briefly told what they were expected to do, they received a brief explanation about how the program works. Then, they were told that the experimenter would not provide any help. However, this rule was to be applied only to either what the participants had already been told or to functions which were expected to be intuitive (i.e., the designer expected that they would appear obvious to the user and would not need further explanation). MapTutor contains many interactive features which are not suitable to be fully described in an brief introduction. Thus, the experimenter did help the participants in cases where a less common feature appeared. For example, they were not told about link ambiguity (see **Section 5.4**) because this anomaly was not expected to happen very frequently. Thus, when MapTutor informed a participant that his or her link was ambiguous, the experimenter explained why MapTutor considered the link as ambiguous and how to answer an ambiguity clarification question (see **Figure 5–5 on page 120** for an example). Another situation where assistance had to be provided was when the participant became rather frustrated, as in the case of the inhabitant mistake made by subject $\mathbf{S}_7$ (see **Appendix D**). In that particular case, clarification was provided so as not to jeopardise the whole experimental session, since the experiment was not making any sense for the subject at all. At any rate, the subjects (except in the latter case) were not given more information than the target learner would have in a real setting[10].

Despite the fact that the participants were told that his or her questions would not be answered during the session, they were encouraged to keep asking whenever they faced any trouble, because by doing so, they would provide useful information about the program. Sometimes, the subjects were asked about the cause of

[10] For example, the information provided about ambiguity could have been obtained in the help program, which was fully operational at the time the experiments were carried out, but the experimenter asked the participants not to use this facility.

a wrong link, but this approach was soon abandoned, because very often, they did not precisely why they committed the error.

The subjects[11] were asked to spend a minimum of 30 minutes in the task, but they were not pressured by the experimenter to give up after this time period.

6.2.4 Results

Analysis of Observations. Given the quality of the measures and the small sample utilised, the observations are better appreciated by means of a qualitative, logical analysis than by means of a statistical (quantitative) one. Nonetheless, to provide some general idea of the performances of both the subjects and MapTutor some quantitative results are also presented. The data gathered during the experiments are presented in **Appendix D**.

It is very hard to compare the subjects' own diagnosis with MapTutor's, because as Feifer (1989) observed, subjects very frequently do not know precisely why they made a wrong link. Many times, however, it is possible to have some clue as to whether a real mistake happened. For example, sometimes it seems clear for a vigilant observer when the mistakes are mere slips or when the subject is puzzled after making a mistake. In the first case, they often became upset with the feedback messages (as if they did not need it at all), and dismissed them without even glancing. On the other hand, in the second case, they looked at the messages very carefully, and in most of the situations they did accept the advice seemingly without disagreeing. Nevertheless, when prompted by the experimenter, about the nature of the error they were not sure about their answers. For example, some could not explain whether their mistake was caused for misunderstanding of the text or the meaning of the appropriate link.

Analysis of Questionnaires. **Table E–1** in **Appendix E** presents a summary of the answers gathered from the post-session questionnaires. The questionnaire model itself can be found in that same appendix. The results presented should be interpreted with caution because, as discussed in **Section 6.2.1**, they are bound to be biased. Even so, the results seem to indicate that MapTutor is a program easy to use and learn to operate. Moreover, it appears that its educational potentiality is very promising.

[11] Except subject $\mathbf{S}_5$ who was very busy and had to leave earlier.

6.2.5 Discussion: Appraising MapTutor

Diagnosis and Feedback. Results from this preliminary evaluation seem to indicate that difficulties faced by a learner in the construction of good maps come from three main sources:

1. Selecting the important concepts in the text before mapping them.
2. Mapping the relationships among those selected concepts onto appropriate links, once those concepts are already selected from the text.
3. Slips caused by mere distraction.

The first of the obstacles mentioned above could not be observed in Feifer's (1989) work because Sherlock's interface was too simplified and idealist so as to take into consideration that learners need to select the concepts in the first place before starting mapping them[12]. That is, it could be observed that some subjects did have trouble in finding important concepts in the text. Thus, although not tested in this preliminary evaluation, the auxiliary teaching procedure *SuggestNext* may prove to be very useful (see **Section 4.9.2**) in future evaluations of the program.

Feifer's (1989) work concentrated on the second cause pointed out above. According to Feifer's theory, learners construct complex plans (i.e., they use procedural knowledge) before mapping a relationship between two concepts onto the appropriate link. Therefore, the most likely cause of such an error would be choosing the wrong plan for doing this mapping. In order to try to discover which plan the learner might have used; Sherlock always asks the learner why she made a given wrong link. These questions are based on the program's library of plans. Once it finds a wrong plan which matches the learner's answer to the *why-did-you-that* question, it will provide the learner with the correct plan. By contrast, MapTutor assumes that (non-slip) errors stem from three sources:

1. Misunderstanding of concepts.
2. Misunderstanding of the relationships in the text.
3. Misunderstanding of the meaning of links.

The results of the experiments seem to support the classification as well as the ordination of these causes. Concept misunderstanding not only was rare but

[12] In other words, Sherlock presented the concepts in the map so that the learner's job was simply looking for the relationships among those concepts in the text and then choosing the appropriate links in order to connect those concepts. Thus, learners would not have any trouble looking for the important concepts in the text, because they were already there, in the map.

also it did not always prevented the subject linking the misunderstood concept correctly. Subject $\mathbf{S}_1$ provides a good example of this: even without being sure about the meaning of concept SNAIL she was able to link correctly this concept to ORGANISMS. On the other hand, the severe misinterpretation of concept HABITAT made by subject $\mathbf{S}_7$ did prevent him constructing a fair map.

In the case of the last two causes of error above, it was difficult even for the subjects to make a finer distinction between errors caused by misunderstanding of the relationship as it appears in the text or in the meaning of the link used to map the relationship. Nonetheless, at least one wrong link can be pointed out as caused by misunderstanding of relationships. As anticipated by the expert who validated MAPTUTOR's knowledge base, the excerpt,

> On the other hand, there are biotic factors, which are determined by the organisms which share the habitat. For example, organisms which eat each other, compete with each other for food or provide shelter.

might have made subject $\mathbf{S}_3$ misunderstand the relationship between ORGANISMS EATING EACH OTHER and ORGANISMS (see the analysis of experimental sessions above). Last, misunderstanding of links was not only indicated explicitly in the questionnaires, but it could also be observed, given the time spent by the subjects looking up the links definitions pane before choosing a less intuitive link.

Currently, neither Sherlock nor MAPTUTOR is able to accurately diagnose the cause of wrong links. MAPTUTOR's basic diagnostic procedures seem to be able to identify causes of errors, but despite this, the decision procedure used to drive the diagnostic process does not appear to possess intelligence enough to choose the correct diagnostic. Nonetheless, the form and contents of the feedback messages provided by MAPTUTOR appear to be much more adequate for teaching mapping than Sherlock's, and all subjects indicated that they learned from the messages presented by the program, notwithstanding its obtrusiveness[13]. For example, the feedback message shown in **Table 6–1** (adapted from Feifer, 1989, p. 12) would be delivered by Sherlock when it suspects the learner is using a wrong plan for using link EQUIV. Note that the variables shown in **Table 6–1** are presented to the learner precisely in that way. Not surprisingly, Feifer (1989) reports that subjects were rarely keen on changing their minds as a result of the feedback provided by Sherlock. In a situation similar to above, MAPTUTOR would present a feedback message according to the schema presented in **Table 4–9 on page 97**, with all slots instantiated to match the particular situation.

[13] Obtrusiveness is an interface trait which will be discussed below.

```
You used the plan:
if
every x is y
it is not as likely that a y is a x
then
make an EQUIV  link from x to y
that is not a good plan
```

TABLE 6–1: PLAN FEEDBACK PRESENTED BY SHERLOCK

The third obstacle mentioned above, which prevents building good maps, seems to be a silly one, but it cannot be overlooked. For example, 10 out of 16 wrong links made by subject $\mathbf{S}_4$ were clearly caused by mere distraction. Other subjects also committed various slips caused by inattention, but the case of subject $\mathbf{S}_4$ provided a most striking case which tutoring system must take into consideration (see more below).

Interface. In general, subjects seemed to have liked to interact with MAPTUTOR's interface. Most found it very friendly and some even said it was very funny. Nevertheless, they also appeared to have found the program obtrusive (notwithstanding that only one subject explicitly stated so). Sometimes, it could be clearly observed that the program was too obtrusive and the subjects seemed to become upset at the tutor when it provided the messages[14], and dismissed the message without even bother having a look at them. This was particularly true of when the feedback was offered as a result of a mere slip (e.g., the subject mapped, in his or her head, the relationship onto a correct link without realising that the current link name was not the desired one). For example, $\mathbf{S}_4$, who provided a very helpful, voluntary think-aloud protocol, clearly proved to know the relationship between HABITAT and MICROHABITAT but used the wrong link simply because she believed that the currently selected name was the one she thought of. This phenomenon occurred with most of the subjects. Thus, the question is, *is there any way of the program knowing when the learner has the right reasoning but draws a wrong link as a slip?* A human tutor can sometimes do this (notably when the learner thinks aloud), but it may be impossible for a computer tutor to uncover those sort of slips given the current state-of-art of human-computer interaction.

[14] Another problem concerning messages was how to dismiss them. The fact that the program hides the cursor in an attempt to prevent sudden moves (see **Implementation Note 3** in **Seção 5.10**) may partially explain the apparent anger of the subjects with the tutorial messages. Nobody likes diving cursors, do they?

Having the concepts selected in the map pane whenever they are selected in the text appears to suggest the learners that they *must* (instead of the intended meaning of they *can*) move the concepts. At any rate, organisation of information is one of the alleged virtues of graphical learning strategies (see **Chapter 2**). Also, selecting a concept in the text before allowing it to be moved (i.e., dragged) around the map pane has not proven to be a good idea, and in some cases was even a nuisance for some subjects. That is, despite the fact that the learner has to return to the text to look for the desired concept before moving it, and consequently being able to acquire some incidental material during the pass (see, e.g., Rothkopf, 1982), this has not been verified in practice. As a matter of fact, the subjects seemed to become a bit frustrated for not being able to move the concepts as soon as they wished to, and no evidence was accumulated that they could learn more by looking for the concept in the text just in order to prepare it to move. A better idea might be to include a *Dragging Tool* in the collection of tools provided by MapTutor. This idea is explored in **Section 7.2**.

Despite the constraints discussed in **Chapter 5**, MapTutor's syntax seems to be very simple and natural, which helps to make the program easy to use. For example, no subject attempted to draw a dangling link first, and then to draw the concepts the link should connect. Thus, the results of this evaluation seem to suggest that those constraints do not appear not so restrictive as to hinder the learner's objectives. Furthermore, the Tools Pane proved to be very useful for most subjects, but some did not make use of the short-cuts provided (e.g., some subjects preferred to delete a link and then draw a new one when they could have simply reverted the link).

Some subjects did not perceive the tutor signalling. Those who perceived the signals never bothered clicking the tutor's window to get more feedback. Also, it was observed that the learner did not seem to pay attention to the link definitions pane at the beginning of a mapping session, but they did when they intended to use those links considered more difficult to employ. Moreover, this simple (but efficient) form of feedback seemed to be sufficient for some subjects.

6.3 Evaluating MapTutor as a Mapping Training Program

The results of the preliminary evaluation indicated several flaws and bugs of the first (operational) version of MapTutor. In particular, it was found that this version of the tutor intervened too much, which could cause confusion for the users. Thus, a new version should be built which incorporates the findings of

the observations. This section indicates how a more extensive evaluation might be carried out on a revised version of MapTutor.

6.3.1 Purpose and Justification of Study

The general purpose of this proposed study is to describe a way in which MapTutor's educational effectiveness could be evaluated, and thus to add to the existing body of knowledge about training graphical learning strategies. The methods of teaching graphical learning strategies (reviewed in **Section 2.6**) do not seem to be entirely satisfactory[15], and although previous work (Feifer, 1989) as well the current research itself argue in favour of a computational approach for teaching a graphical learning strategy, insufficient evidence has been presented to make the argument for or against the individualised teaching of graphical learning strategies by means of the computer. Thus, the data supplied by this proposed study will provide a basis upon which researchers (or practicing teachers) might be able to decide whether to retain their current conventional methods of teaching the mapping strategy or to adopt an innovative computational solution. In other words, the educational implication which follows if MapTutor's method is successful is that researchers will have a more powerful means of teaching the mapping strategy.

The specific purpose of this study is to compare a conventional approach for teaching mapping with the approach followed by MapTutor. The conventional approach chosen is the modelling approach, since this seems to be the most promising one in terms of instructional effectiveness (see **Section 2.6**). Yet more specifically, this study will compare these different methods for teaching mapping with the ultimate objective of finding a relationship between the method used to teach mapping (independent, treatment variable) and mapping performance (dependent, outcome variable).

Research Hypothesis. The particular hypothesis to be tested in the proposed experiment is that the students who use MapTutor should learn how to map more effectively than those who use a conventional modelling approach. That is, the hypothesis embodied in this investigation is that students who learn how to map by using MapTutor will have a better mapping performance than those who learn how to map by means of a modelling approach. The null hypothesis is that there will be no difference in post-test mapping performance between the two groups under scrutiny.

[15] For example, the modelling approaches take long training time, and require expert mappers, who in turn need to be trained.

Definitions. Mapping performance is operationally defined as a score in a post-test consisting of a paper-and-pencil mapping session using a text not related to those used during the training sessions themselves. Informally, a student's mapping performance is a measure to the extent to which she is able to produce good maps. A good map is one which,

- Contains the most important concepts for the understanding of the text.
- Contains the most important relationships found in the text.
- Employs the most appropriate labelled link (extracted from the set of canonical links provided) for each relationship depicted.

The dependent variable will be measured by post-test scores representing the mapping knowledge after the training sessions. Accordingly, high mapping performance means high score in maps according to the criteria above. The scoring procedures to be adopted will be discussed in the instrumentation section.

6.3.2 Background and Review of Literature

Chapter 2, specially **Section 2.6** which deals with graphical strategies training approaches, provides the background and review of literature necessary for carrying out the proposed study.

6.3.3 Methodology and Procedures

Research Design. A *randomised post-test-only control group design* is the one selected to implement the proposed study. **Figure 6–3** illustrates how this design applies to the present study (based on Fraenkel & Wallen, 1993).

Sample. The study will be conducted with undergraduates, who will constitute the population to which the result of the study might (ideally) be generalised. The sample will include at least 60 students[16]. An attempt should be made to keep extraneous variables constant by means of a random sampling. Ideally, the sampling selection would be done randomly, but in face of practical difficulty, it will probably be a convenience one (e.g., by paying subjects willing to take part in the experiments). Nonetheless, the assignments to each experimental groups will be done at random. The treatment group will be assigned to use MapTutor, whereas the comparison group will be taught how to map using the modelling approach. These groups will henceforth be called the **MapTutor Group** and the **Modelling Group**, respectively.

[16] Although the educational research literature does not specify how large groups used in experimental research should be, some researchers believe that a minimum of 30 (see, e.g., Shute & Regian, 1993) to 40 (see, e.g., Fraenkel & Wallen, 1993) subjects in each group should be used.

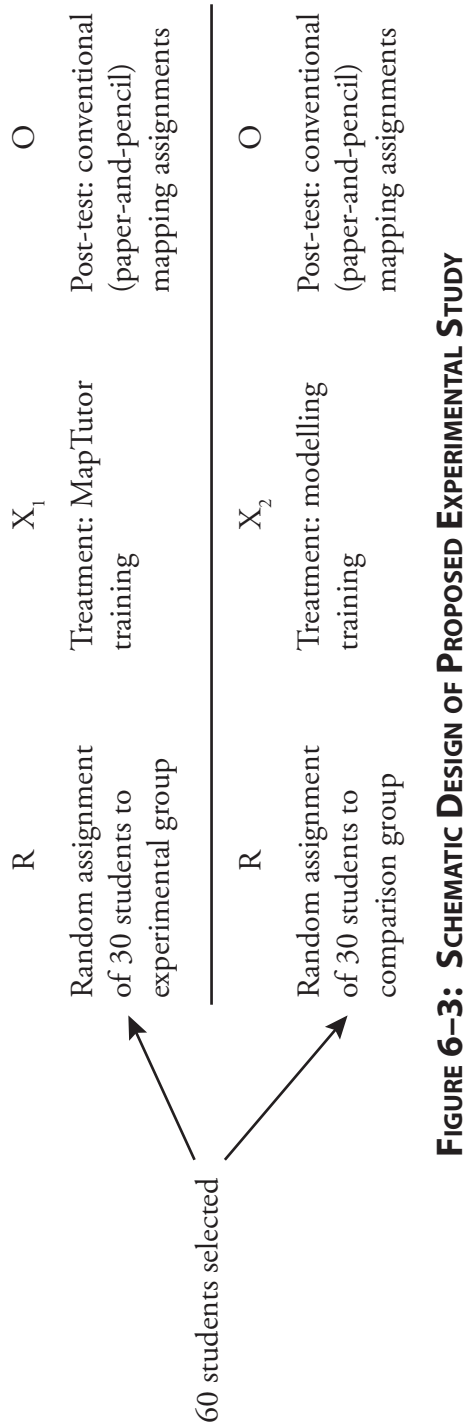

Figure 6–3: Schematic Design of Proposed Experimental Study

Instrumentation. The instruments will consist of paper-and-pencil maps drawn by all subjects. However, the lack of existing instruments for evaluating the quality of maps demands the development of a new instrument of measurement to be used in the post training assessment. Nevertheless, the measures, expressed as scores, pattern themselves on the features which should be present in adequate maps (as described above). The fact that a number of researchers (e.g., Novak &

Gowin, 1984; Holley & Dansereau, 1984a) agree upon those features backs up the validity of such measures.

The measures will result from the judgement of knowledgeable people (i.e., expert mappers who are familiar with the domain from which the to-be-mapped text will be extracted). Also, the use of three independent raters will enhance the validity and reliability of the scores. The measurement procedure to be employed in evaluating maps will be described next.

Each construct in a map consisting of a pair of concepts connected by a labelled link is defined as a proposition. The score obtained by a student's map will be the sum over the ratings obtained by all propositions which constitute the overall map. The procedure for rating each individual proposition is described as follows. Each proposition considered correct and relevant by the panel of experts will be judged according to its import and appropriateness of link type as shown in **Table 6–2**.

Import	Appropriateness of Link Type
High: Add 2	**Most Appropriate**: Add 2
Low: Add 1	**Acceptable**: Add 1

TABLE 6–2: SCORING OF PROPOSITIONS IN EXPERIMENTAL STUDY

It could be helpful to state the scoring criteria presented above in an alternative procedural form as presented in **Table 6–3**.

For each proposition considered *correct* and *relevant* in the map:

1. If it has *high import* and employs the *most appropriate link type*, its score will be 4.
2. If it has *high import* and employs an *acceptable link type*, its score will be 3.
3. If it has *low import* and employs the *most appropriate link type*, its score will be 3.
4. If it has *low import* and employs an *acceptable link type*, its score will be 2.

TABLE 6–3: PROPOSITIONAL SCORING PROCEDURE

Note that a proposition which does not satisfy any of the criteria above (i.e., a wrong/missing proposition) will have a null score. This means that wrong answers will not be penalised[17].

The whole, detailed map scoring procedure will thus consist of the following:

- For each student's map, the experimenter will circle each proposition which makes up the map. He will then number each of these propositions in an ordered way.
- For each map, the experimenter will (partially) fill a grading sheet containing two columns. He will fill the first column in with the numbers corresponding to the propositions in the map. The numbers in the second column will be filled in later with the scores corresponding to the rate of each proposition given by each rater, so that this column will be left empty by then.
- Each map with its propositions highlighted (i.e., circled) and numbered, along with its respective grading sheet, will be photocopied twice, so that there will be three copies of each of these sets — one for each rater.
- Each rater will independently give each proposition in each map a score according to the rules described above[18]. Each score will be written down in the second column of the grading sheet, in the position corresponding to the numbered proposition. The rater will wind up the rating of a map by summing the scores of all propositions, which will result in a total score for the map itself.
- For each map, the experimenter will take the average of the scores given by the three raters as its final raw score. A Kuder-Richardson formula (see, e.g., Ebel, 1972) should also be used to estimate the reliability of the multiple ratings given by the three raters.

Materials. Two pieces of text, extracted from undergraduate text-books, should be used as training materials by both groups. The texts should be chosen based

[17] Penalising wrong answers (or using an appropriate correction formula) is an instrument for correcting scores for guessing. There are good reasons for not taking correction for guessing into consideration here. First, the students will be aware that they will be taking part in an experimental study, so that they will have no apparent reason (e.g., earning more points/credits) for guessing. By the same token, there will be no reason by which students in one group would guess more than students in the other group. Moreover, even in non-experimental situations, correction for guessing does not always result in more reliable scores (see, e.g., Ebel, 1972, pp. 251–8; see also Suen, 1990, p. 73).

[18] Each rater should have a copy of those procedures at hand by then.

on (see, e.g., Heeren & Kommers, 1992): (1) great information density, which is supposed to be adequate for mapping; and (2) fairly little amount of background knowledge needed for understanding based on the subjects' expected prior knowledge. The first text will consist of a short passage (about 300-word long), and the second one would be a medium-sized passage (about 600-word long). Two versions of these texts would be used. The first version of each of these texts should be presented to the students in the comparison group in a sheet of paper, whereas the second version should be entered into MapTutor program as described in **Chapter 3**.

In addition to instruction from a human tutor, a handout should also be used to familiarise the students in the comparison group with mapping and help them to construct maps. This handout should consist of a procedure for applying the mapping technique (see, e.g., Dansereau et al., 1993), and the material included in this handout would be unrelated to the testing material.

A full version of MapTutor, as described in **Chapter 5** and with some of the improvements suggested in **Chapter 7**, should be used by the treatment group. The subjects in the control group should use paper and pencil in their mapping practice. Because the number of subjects in the MapTutor Group will be relatively large, it would help if a number of computers could be used at once.

Procedural Details

The subjects should be divided into two groups: treatment group and comparison group. The treatment group should be assigned to individual training sessions with MapTutor using the computer versions of the texts, and the comparison group should be assigned to (collective) training in a classroom.

First Stage — Introduction and Experimental Sessions. During the introduction for both groups, the subjects should be told that the experimenter is interested in testing the use of a learning strategy that could help them to learn better. This introductory statement is intended to increase the subjects' motivation and interest in the experiments.

Training will be divided into four sessions consisting of 2 hours of instruction/practice each. These training sessions will be distributed over a two-week period, and are summed up in **Table 6–4**. Notice that the subjects in the MapTutor Group will use the program in their practices with the strategy, unless the contrary

is indicated (as in **Session 4**). Subjects in the Modelling Group will only use paper-and-pencil in their practices[19].

The general introduction to the strategy (**Session 1**) will consist of an overview and description of the technique. An instructor will provide this general introduction to both groups. Dansereau et al. (1993) provide an excellent source for the hand-out prescribed in **Session 1** for the Modelling Group. During the overview of the program, the subjects in the MapTutor Group should be presented with MapTutor, the general features of the program (menus, help, panes, etc.) should be explained, and a short (say, 15 minutes) demo session (using a demo-text) should be presented to them. Then, they should be asked to use the tutorial program. To wind up the introductory part of the session, the subjects should be allowed to practice using the program with the demo-text as much as they want, i.e., until they feel comfortable using the program, but the maximum period allowed would not exceed (say) 15 minutes.

Session	Modelling Group	MapTutor Group
1	General introduction to the strategy Subjects receive a training handout Subjects practice mapping on sentences	General introduction to the strategy Overview of the program Subjects work through the tutorial program
2	Subjects practice the strategy on a small passage Subjects receive feedback from the instructor in form of expert-generated maps	Subjects practice the strategy on a small passage Subjects receive feedback from the program as describe in **Chapters 4** and **5**
3	Subjects practice the strategy on a medium-sized passage Subjects receive feedback from the instructor in form of expert-generated maps	Subjects practice the strategy on a medium-sized passage Subjects receive feedback from the program as describe in **Chapters 4** and **5**

[19] The training proposed for the Modelling Group is based on Holley & Dansereau (1984b, pp. 94–5), and is a prototypical sequence of training sessions for the network strategy described in **Section 2.3**. For more. details about how modelling can be implemented, see **Section 2.6.2**.

Session	Modelling Group	MapTutor Group
4	Subjects practice the strategy on a text of their own choice Instructors provide clarification as needed	Subjects practice the strategy on a text of their own choice (using paper-and-pencil) Instructors provide clarification as needed

Table 6–4: Mapping Training in Experimental Studies

Note that, in the treatment group case, the practice task is to use MapTutor and just follow the instructions provided by the program, but the subjects in this group could also receive feedback from the instructor in extreme situations (e.g., when they get stuck due to an unexpected buggy behaviour by the program). Note also that, in **Session 4**, the subjects should not be allowed to practice on material related to the text to be used in the post-training mapping session (i.e., the dependent measures collection).

After their last mapping sessions, the subjects should be told that they would be expected to attend a mapping test a week later.

Second Stage — Dependent Measures. A week after the experimental stage, all the subjects should be asked to draw a map representing a text on some domain extracted from a expository text chosen by the experimenter. Each subject should be asked to reproduce (i.e., map) the text as completely as possible using paper and pencil. They should then be given a fair amount of time (say, 50 minutes) to complete the task. The maps produced by the students would finally be collected and scored by the experts, according to the guidelines described above.

Data Analysis. After the collection of data carried out by the expert-raters selected and trained by the researcher, the means corresponding to the scores of both groups will be calculated. Then, they will be analysed using either t-test or analysis of variance (anova). The ultimate goal is to try to spot any significant differences in mapping performance between the two groups.

6.3.4 Pilot Study

A **pilot study** — consisting of a small-scale trial of the procedures and materials described above — should be included in this proposed investigation. This pilot study aims at detecting (or anticipating) any problems which could occur before the actual large-scale experiment itself begins. In particular, special methodological care should be taken so that treatment and outcome variables are established

in a more careful way as well as to avoid (or at least minimise) confounds (e.g., instructor effects) in the experiments.

Due to the inherent entanglement between mapping and domain knowledge, a potential source of problems might include the text to be used in the dependent measure. For example, difficulty in understanding or lack of interest in the text may cause all subjects perform badly and thus leading to non-significant results. Thus, the texts to be used in the full-scale experiment should be tested with regard to their motivational aspects as well as degree of difficulty with the small sample used in this pilot study.

Special care should also be taken so that the time in all experimental sessions is the same so as to reduce the possibility of confound due to differential training time. Also, in this pilot study, the subjects should be allowed to take the time necessary for training (First Stage) and for accomplishing their post-training maps (Second Stage) so that the amount of time to be allowed for each task can be determined. Moreover, this small-scale trial would serve the purpose of pre-testing the experimental material (text, instruction, etc.) for its usefulness. Last, the expert-raters would have the opportunity for training the scoring procedures described above.

6.3.5 Some Alternatives to the Proposed Study

Some comments about the proposed experimental study described above are called upon. There are two problems which could raise some alternative explanations in case of unexpected results. First, the treatment group will use a computer for their work and then have to use paper and pencil for the post-test. Thus, perhaps the subjects in this group will be at disadvantage (e.g., they will not have a menu from which they can select link types, as in their training sessions). An option which could alleviate this potential problem would be to make the paper-and-pencil post-test similar to the practising tasks carried out by the subjects in the treatment group, but this could introduce bias towards the computer methodology. Thus, it appears there is a trade-off, so that one should not be too optimistic about the result of the experiments.

Note also that the treatment group gets a computer and the teaching methodology described in previous chapters. Thus, if this group performs better than the comparison group, one will not be able to tell whether the cause of any increased score is due to the computer-based teaching methodology implemented by MapTutor or just to the exposure to the computer. The solution could be to evaluate the program's teaching methodology in isolation (as opposed to evaluating

the impact of the computer and the program's method of teaching as a whole). A more thorough design might be to divide the subjects into three groups, with one group being taught by MapTutor, one group receiving human teaching, and the third group having a human teacher but using MapTutor (without the tutoring component) as a tool for producing the maps. This would reveal the possible motivating effect of using the computer.

Finally, a slightly more complex design could also post-test for the amount of domain knowledge retained by the two groups under consideration. Then, it would be expected that the group who performed better at mapping should have learned more of the material in the text at hand.

6.4 Conclusion

One flaw frequently found in research into learning strategies (see, e.g., Goetz, 1984; Anderson & Armbruster, 1982) is that it is not always clear what kind of activity the learner is actually engaged in while executing a given learning strategy. Thus, to be useful as a research tool, a computer tutor for teaching a learning strategy ought to try to uncover, record, and perhaps analyse all the relevant information about what the learners actually do when using the program. MapTutor records and stores information about the learner's performance and makes it available for the researcher in a high-level form by means of end-of-session reports. During the preliminary evaluation described here, this proved to be a very useful research tool, which allowed not only a better understanding of the program but also of the subjects' behaviours.

The preliminary, exploratory experiments described in this chapter served the purpose of gathering information from target-learners in order to (1) assess the accuracy of MapTutor's diagnosis; (2) appraise the effects of the program on real learners; (3) to validate some design decisions described in the previous chapters. Thus, the main purpose of this preliminary study was not to gather quantitative data which could be statistically analysed. Moreover, as Flagg (1990) points out, pretest/post-test studies, typical of the hypothetico-deductive methodology, 'provide little insight as to why the program might be working or might not be working' (p. 147)[20]. Thus, the inductive paradigm seemed more appropriate to this preliminary appraisal of MapTutor. The caveat of this method of inquiry is that there is no well-developed method of data analysis (as in the hypothetico-deductive paradigm). Thus, the conclusions presented in this chapter are best interpreted as tentative. Moreover, some of the results presented here may only

[20] See also Twidale (1993) for a defence of informal, qualitative methods for evaluating intelligent tutoring systems.

apply to the group of people studied. Nevertheless, perhaps more important are the directions that these studies suggest for future research.

Last but by no means the least, despite important from a formative evaluation perspective, the experiments carried out so far do not address the question of educational efficacy of MapTutor properly. Thus, this chapter has also proposed a more formal summative evaluation, which should be carried out by means of experimental research, which could be used to assess whether the program really teaches what it purports to — i.e., how to map. Due to time and resource limitations of this research project, this proposed study has been left out as future work.