A critical review of argument visualization tools: do users become better reasoners?

Susan W. van den Braak[1] and **Herre van Oostendorp**[1] and **Henry Prakken**[2] and **Gerard A.W. Vreeswijk**[1]

Abstract. This paper provides an assessment of the most recent empirical research into the effectiveness of argument visualization tools. In particular, the methodological quality of the reported experiments and the conclusions drawn from them are critically examined. Their validity is determined and the methodological differences between them are clarified. The discrepancies in intended effects of use especially are investigated. Subsequently, methodological recommendations for future experiments are given.

1 Introduction

Argument diagramming is often claimed to be a powerful method to analyze and evaluate arguments. Since this task is laborious, researchers have turned to the development of software tools that support the construction and visualization of arguments in various representation formats, for instance, graphs or tables. As a result, several argument visualization tools currently exist [3], such as ArguMed [18], Araucaria [5], ATHENA [6], Convince Me [7], Compendium [8], Belvedere [9], ProSupport [4], and Reason!Able [15]. Typically, these tools produce "box and arrow" diagrams in which premises and conclusions are formulated as statements. These are represented by nodes that can be joined by lines to display inferences. Arrows are used to indicate their direction.

Although it is often claimed that structuring and visualizing arguments in graphs is beneficial and provides faster learning, experiments that investigate the effects of these tools on the users' reasoning skills are relatively sparse. Nevertheless, some experiments have been reported and the purpose of this paper is to critically examine their methodological quality and the conclusions drawn from them. Thus we aim to give an assessment of the state-of-the art in empirical research on the use of argument visualization tools, and to make some methodological recommendations for future experiments.

This paper is part of a larger research project on software support for crime investigations. Since reasoning is central to crime investigations and current support tools do not allow their users to make their underlying reasoning explicit, it is important to consider the use of argument visualization during these investigations. In this respect, it is also important to explore the effectiveness of such visualization tools.

The structure of this paper is as follows. Section 2 describes the criteria that will be used to evaluate the methodological quality of the experiments. The methods and results of these experiments are then discussed in Sections 3 and 4. Finally, Section 5 offers methodological recommendations to conduct future research.

[1] Department of Information and Computing Sciences, Utrecht University, the Netherlands
[2] Faculty of Law, University of Groningen, the Netherlands

2 Investigating the effectiveness of argument visualization tools

Among the tools that were experimentally tested for their effectiveness are Belvedere, Convince Me, Questmap, and Reason!Able. These have in common that they are education-oriented and designed to teach critical thinking or discussion skills, and are tested in an educational setting, for instance, on students during a course. Also, important discrepancies exist, for example, Belvedere and Reason!Able are entirely designed to assist argument construction and analysis, while Convince Me produces causal networks. Questmap has different main purposes, namely collaborative decision making, but it supports the construction of argument structures to a certain degree. Furthermore, Belvedere and Questmap are tested during collaborative reasoning, while Reason!Able is used by a single user. Most importantly, differences exist between the intended effects of use. Obviously, the latter affects the measures of effectiveness used and the tasks to be performed. This paper aims to provide an overview of these discrepancies.

In the remainder of this paper, the methods (viz. experimental designs, participants, and procedures) and results of the conducted experiments on argument visualization tools will be described. The aim of this is to find a general pattern or plan that may be followed to conduct research in this area. Moreover, we will determine whether these experiments were able to prove the long existing claim that visualization improves and simplifies reasoning. While describing the experimental methods, two important issues will be addressed, that is, the validity of the experiments and the related problem of finding a measure for the outcome variable, because these may affect their outcomes and the interpretations of their results. For this purpose, a checklist will be presented that allows us to assess their methodological quality. Additionally, this paper provides an overview of the proposed measures and their reliability.

2.1 Validity

If empirical experiments are conducted, it is important to take into account the validity of the experiment. Validity is mainly concerned with the question of whether the experiment really measures what it is supposed to measure. Two important types that we will consider in this paper are internal validity and external validity [2, 19]. Internal validity is the extent to which the differences in values of the dependent variable (the outcome variable) were actually caused by the independent variable (the variable that is manipulated by the experimenter) and not by some other source of variation. The external validity of an experiment is concerned with the following question: how well do the results of the experiment generalize beyond the

sample of subjects in the experiment and the particular experimental manipulations to the rest of the possible situations of the real world?

Besides evaluating the validity of an experiment, it is also important to consider the reliability of the measures used and the experiment conducted. If an experiment or measure is reliable, it means that it yields consistent results. In order for a measure to be reliable (or accurate) the results should be reproducible and as little as possible be influenced by chance.

It should be noted that validity implies reliability but not the other way around. Validity refers to obtaining results that accurately reflect the concept being measured, while reliability refers to the accuracy of the scores of a measure.

Generally, internal validity is assured by assigning subjects to treatment groups and control groups randomly. Experiments that use randomization and that are internally valid are sometimes called "true" experiments. Experiments that approximate these internally valid experiments but do not involve randomization are called quasi-experimental. This means that a valid experiment should at the very least have the participants assigned to conditions randomly, so that the external variables are under control and internal validity is maintained.

However, internal validity is not easy to obtain and is dependent on the chosen design. In a *between-subjects* design the participants are used only once and are part of the treatment group or the control group but differences between participants cannot be completely controlled. To cancel out the influence of relevant pre-existing differences between groups on the results, the treatment and control groups have to be matched or homogenized. For this reason, random assignment of subjects to conditions is crucial. Another solution to avoid effects of external variables is the use of a *within-subjects* design. In such a design all participants are used twice, as they receive both treatments. In order to cancel out any carryover effects, such as learning, practice, or fatigue effects, participants have to be assigned in such a way that different subjects receive both treatments in different orders (i.e. counterbalancing). Basically, these methods of randomization, counterbalancing, matching, and homogenization help to ensure internal validity.

External validity is affected by the design and subjects chosen. In order to assure external validity, the experimenter has to make sure that the experiment is conducted with the right participants as subjects, in the right environment, and with the right timing. Therefore, the experimental environment should be as realistic as possible. Additionally, the subjects should be selected from the population randomly. Finally, to check for external validity, the experiment should be replicated in other settings, with other subject populations, and with other, but related variables.

Table 1. Criteria for experimental validity

	Criteria
Reliability	use consistent measures
Internal validity	use at least one control group assign participants to conditions randomly match or homogenize (between-subjects designs) counterbalance (within-subjects designs)
External validity	draw a random sample from a population use real world settings and stimuli replicate the experiment

Obviously, since experimenters try to prove the effectiveness of their tool by justifying causal relations between the use of the tool and the users' reasoning skills, their research should preferably be done through laboratory experiments that are valid; the criteria are summarized in Table 1. Unfortunately, as we will see below, this is not often the case so that valid conclusions cannot be drawn.

2.2 Measures

The goal of the experiments described in this paper is to measure the effectiveness of a tool. The effectiveness describes the effect on the users' ability to reason (e.g. did these tools make their users better reasoners?). However, defining a measure for this is not straightforward. It is even hard to find an objective, reliable measure, that accurately measures the users' progress in reasoning skills. Moreover, to allow for statistical comparison, a quantitative measure has to be used, but such a generally accepted reliable measure is not available yet, as can be concluded from the large amount of different measures used. Generally, scores on critical thinking tests or assignments assessed by experts are used as measures for learning outcomes. These seem to be the only feasible and most reliable ways to measure reasoning skills in a quantitative way. However, as said, not all tools are designed with the same effects of use in mind. In some cases, the effectiveness of a tool is measured by the quality of the constructed argument. In other cases it is measured by the amount of discussion or the coherence of the arguments. It is important to be aware of these differences and their influence on the experimental tasks and the conclusions drawn from them.

3 Methods and results

In this section a detailed description of the reported methods and results of the experiments on Belvedere, Convince Me, Questmap, and Reason!Able is given. Their validity will be assessed and their conclusions will be critically examined.

3.1 Belvedere

Belvedere [9] is a tool that is designed to support scientific argumentation skills in students and to stimulate discussions on scientific topics among students. With Belvedere students can build and display "inquiry diagrams" to model argumentation (see Figure 1). These diagrams consist of data nodes, hypothesis nodes, and unspecified nodes. Undirected links can be used to connect these nodes by for, against, and unspecified relations.

3.1.1 Method

Belvedere was tested in laboratory sessions and an in-school study [9] that investigated how well Belvedere facilitated the emergence of critical discussion. In the first set of sessions, the participants worked in pairs, using only one computer. The pairs were asked to resolve a conflict that was presented in textual and graphical form. The participants were also allowed to use a database with a small amount of relevant information. The second set was almost identical to the first set except that the participants worked on individual monitors and a shared drawing space. It should be noted that only two pairs of students participated in these sessions.

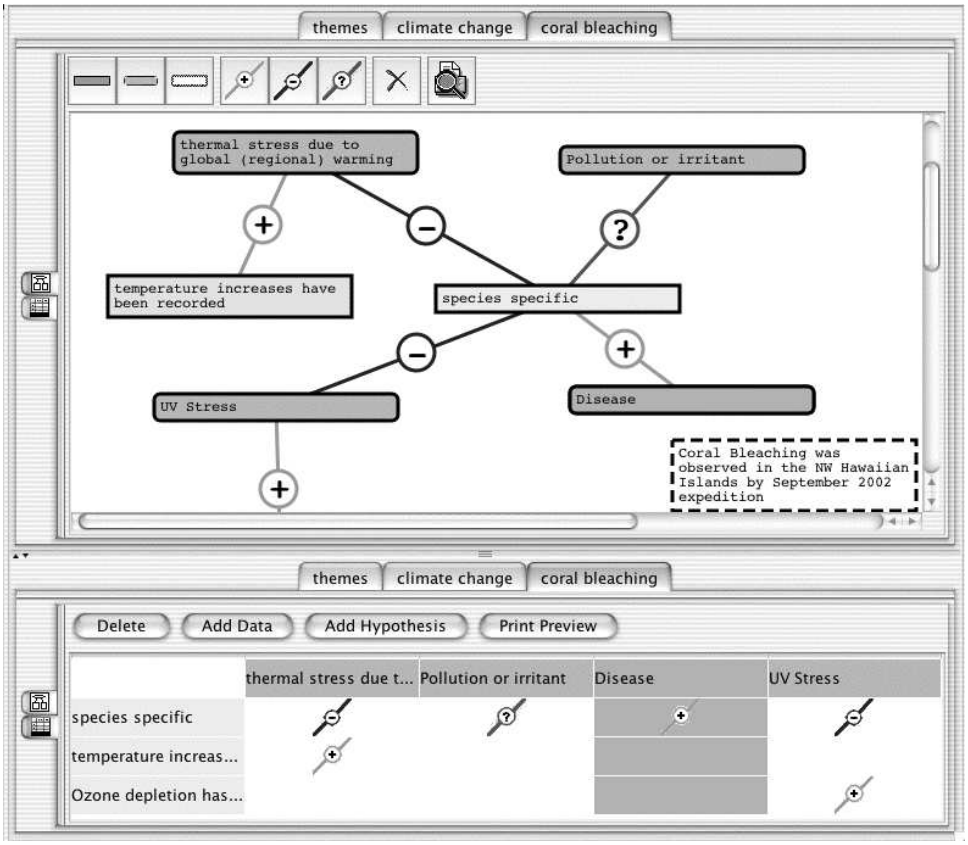

Figure 1. Screenshot of the Belvedere programme

The effect of Belvedere on the participants' critical discussion skills was measured by the amount of discussions that arose. This measure was rather a qualitative than a quantitative one, as the researchers mainly described the students' interactions. This experiment was not valid, because the measure was not valid and no control group was used to compare the experimental group to.

Further, to compare the effect of different representations on the learning outcomes, three different representation formats were tested in [10] and also [11] and [12]. This experiment was internally valid as it was based on a between-subjects design with three groups in which the participants were assigned to groups randomly. Moreover, there were no significant pre-existing differences between the groups' gender balance and mean grade point average due to homogenization. External validity was not guaranteed, because of the artificial nature of the task. It was very limited and was completed in a laboratory setting, while the effect was only measured during the initial use and not over a longer period of time.

The groups, consisting of 20 students each, were defined by the software they used, that is, matrix, graph, or text. All groups had to perform the same task of structuring an unsolved science challenge task into data, hypotheses, and evidential relations. Identical background information was presented to all three groups, one page at a time. The students had to work in pairs and were asked to use the given information in their representation of the problem, before continuing to the next page (the showed information would not remain available for later reference). After finishing their representation of the problem, the students had to complete a post-test containing multiple-choice questions and had to write a collaborative essay.

These essays were scored according to the following measures:

- Evidential strength: the strength of the evidential relationship, on a scale of 0 to 4, with $+$ indicating a supporting relationship, and $-$ indicating a conflicting relationship.
- Inferential difficulty: the number of information pages that must be accessed to infer the relationship, with 0 indicating that the relationship is explicitly stated in the material, and > 1 indicating that the relationship has to be inferred.
- Inferential spread: the difference (in pages) between the first and last page needed to infer the relationship. This is a measure of how well participants integrate information given at different pages.

In order to obtain a measure of the quality of the essay that was produced, an expert completed the task himself and his evidential matrix was used to compare the students' essays to. In this way, the students' ability to list the most important data and relations of the problem was measured. It thus measures the students' collaborative scientific discussion skills.

In sum, Belvedere has two aims: to support the amount of critical discussion and to enhance collaborative learning of reasoning skills. The former was tested in an internally invalid study, while the latter was investigated in an internally valid experiment. The tasks involved constructing arguments based on unstructured information in which the students had to identify data for and against their hypothesis.

3.1.2 Results

For the first set of experiments, the researchers only gave qualitative descriptions of the results. In the first set of sessions, the experimenters found an encouraging amount of discussion. In the second set they found that in one pair the students cooperated to a high degree, but that there was no interaction at all in the other pair.

For the in-school study it was found that sensible diagrams were produced, but that the use of shapes and link types was inconsistent. Moreover, it was found that students incorporated several points of the debate into diagrams.

On the basis of these observations, the authors concluded that Belvedere indeed stimulated critical discussions. However, although a tendency was shown, this experiment did not conclusively prove an effect as it was not internally valid. Conclusions drawn based on these studies are therefore premature. In this respect, the second experiment is more promising, because internal validity was achieved. Moreover, the documentation on the second experiment was considerably more detailed.

None of the test in the second experiment yielded a significant difference between the groups. From these results the researchers concluded that there were no significant differences in performance between the users that used matrix or graph representations and the users that used text only. According to the researchers, the lack of significance of the learning outcomes was disappointing, although the researchers noted that this was not surprising given the fact that the total amount of time spent working with Belvedere was too short for learning outcomes to develop.

It must be said that trends were in the predicted direction but not significant. This means that the students who were allowed to use the Belvedere software that contained matrix representations performed better than the students who used graph representations, who in turn performed better than the students who used text only. Therefore, a tendency is shown that visually structured representations can provide guidance for collaborative learning that is not provided by plain text only, while a significant difference could not be proven. This conclusion is legitimate since the experiment was internally valid.

3.2 Convince Me

Convince Me [7] is a tool for generating and analyzing argumentation and is designed to teach scientific reasoning. In addition, Convince Me provides feedback on the plausibility of the inferences drawn by the users as it predicts the user's evaluations of the hypotheses based on the produced arguments. It is based on Thagard's Theory of Explanatory Coherence [13]. Arguments in Convince Me consist of causal networks of nodes and the users' conclusion drawn from them (see Figure 2). Nodes can display either evidence or hypotheses. Explanatory or contradictory relations are represented as the undirected links between these nodes.

3.2.1 Method

The study described in [7] compared the performance of the participants who used Convince Me to the performance of paper and pencil users. In this study, 20 undergraduate students of Berkeley had to complete a pre-test (in which both groups had no access to the software), three curriculum units on scientific reasoning, integrative exercises (one group is allowed to use Convince Me, the other group is not allowed to do so), a post-test (nobody had access to Convince Me), and a questionnaire (to establish relevant differences between groups).

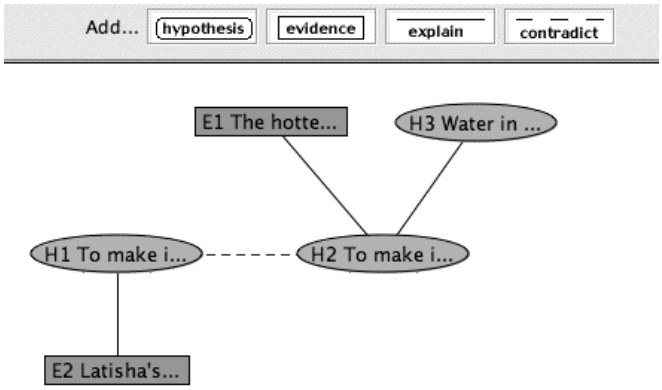

Figure 2. Screenshot of the Convince Me tool

The group that was allowed to use Convince Me consisted of 10 participants, the other 10 participants were part of the group that used paper and pencil only. Both groups received the same instructions and exercises. There were no significant difference between the groups in age, year in school, SAT scores, and total session hours.

This experiment used a between-subjects design. The potential effect of intergroup differences was not an issue here as the experimenters confirmed that the groups were homogeneous with respect to relevant variables. However, they did not mention whether randomization was used while assigning subjects to conditions. Therefore, it will be assumed that this experiment was at least quasi-experimental, but a definitive analysis of the experiments' validity cannot be made.

The following measures were used to measure the utility of the software:

1. How well the participants' beliefs are in accord with their argument structures.
2. The kinds of changes made when arguments are revised.

Only the first measure will be used in the description of the results that will presented below, because this is the most suitable of the two to measure the effectiveness of a tool. The latter only measures the stability of the arguments constructed, not the effect on the users' reasoning skills. The former is a measure of the arguments' coherence, that is, it shows whether people are able to construct arguments that reflect their beliefs properly.

So in short, Convince Me attempts to improve the coherence of its users' arguments so that users become more aware of the believability of their arguments. Note that this differs from the learning effect that was claimed by the developers of Belvedere. Required methodological information is missing so that a genuine assessment of the validity of this experiment cannot be made. Moreover, important details about the nature of the task were not reported.

3.2.2 Results

During the exercises, the participants' beliefs were more in accord with the structures of their arguments if they were using Convince Me, than if they were using paper ($p < 0.05$). Also during the post-test, the belief-argument correlations of Convince Me users were significantly higher ($p < 0.05$) and better than during the pre-test (see Figure 3).

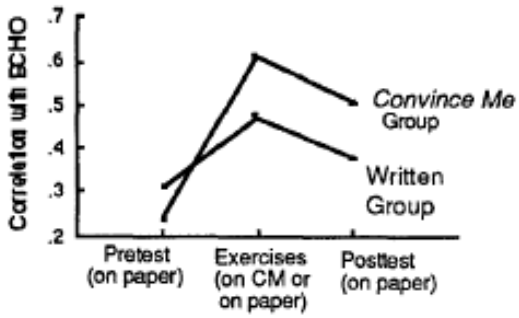

Figure 3. Results of the experimental testing of Convince Me, after [7]

Based on these results the experimenters claimed that the tool improved the users' argumentation skills and made them better reasoners. They also showed that these skills remained when the participants did not have access to the tool and were not supported by it, and that those were still better than the skills of the participants who did not use the tool at all. However, some reservation is appropriate here as the validity of the experiments is unknown.

3.3 Questmap

Questmap is designed to mediate discussions by creating visual information maps (see Figure 4), but is used by [1] to support collaborative argumentation in legal education. It is based on IBIS, an Issue-Based Information System that is designed for collaborative problem identification and solving. IBIS helps multiple users to discuss issues related to a problem and reach a consensus on a solution. Its main procedure involves decomposing the problem into issues. Possible answers to them are recorded as positions. Arguments for and against these positions may be recorded as well. Questmap provides many additional node types, including problems, claims, warrants, backing, and data nodes. By using these nodes, arguments can be constructed.

3.3.1 Method

In [1], the computer-based representational tool Questmap, was tested for its effect on legal argumentation skills.

The most important research question to be answered was: "How does using CSAV, while groups of three or four second-year law students generate arguments throughout the semester, affect the quality and type of arguments generated on a practice final exam (p. 81)". Also, a hypothesis was formulated: "groups using CSAV to construct arguments throughout the study will create higher quality arguments on a practice final exam than those who construct written arguments throughout the study. (p. 81)"

The quality of the produced arguments was measured by:

1. the number of arguments, counterarguments, rebuttals, and evidence present in the practice final exam
2. the scores on the final exams as assessed by the professor
3. the richness of arguments saved in Questmap throughout the semester measured by the number of nodes created (to describe the progress in the treatment group only)

The design was a quasi-experimental between-subjects design. The treatment group consisted of 33 law students who completed the assignments using Questmap in groups of three or four. The control group of 40 students completed the exercises individually using conventional methods. Participants were not randomly assigned to groups, because the participants were allowed to choose the group they wanted to participate in. On the other hand, the pre-test revealed that the groups were in fact homogeneous. This means that at least some internal validity was assured.

The students' argumentation skills were tested and trained throughout the semester. They had to complete five assignments that addressed current legal issues in relation to the admissability of the evidence. Both groups of students were allowed access to the same materials, but only the treatment group was allowed to use Questmap. Two of the assignments of the treatment group were analyzed to measure the progress throughout the semester.

At the end of the semester all participants completed a final exam without the use of Questmap. During this exam the students had to construct all relevant arguments to a given problem individually and without the use of legal resources. These exams were graded by the professor.

To sum up, Questmap claims to improve the quality of the users' arguments so that the users become better reasoners. The assignments involved producing answers to the problem that consisted of arguments, counterarguments, and rebuttals. In the experiment, internal validity was only partially assured.

3.3.2 Results

The found results show that there were no pre-existing differences between the groups ($p > 0.05$), that the arguments did not become more elaborate throughout the semester, and that the treatment group did not have a significantly higher score than the control group ($p > 0.05$). Based on these results, the experimenter claimed that the hypothesis did not hold and that law students who were allowed to use a computer supported argumentation tool did not perform better on the exam than students who only used paper and pencil during the course. On the other hand, it must be said that while the differences between the treatment and control group were not significant, a trend was discovered in the predicted direction (cf. mean $= 5.15$ and mean $= 4.50$ respectively, where $0.05 < p < 0.10$). However, the value of these observations is limited, as complete internal validity was not assured.

3.4 Reason!Able

Reason!Able [15] is educational software that supports argument mapping to teach reasoning skills. It provides support to the users by guiding them step-by-step through the construction process. The argument trees constructed by Reason!Able contain claims, reasons, and objections (see Figure 5). Reasons and objections are complex objects that can be unfolded to show the full set of premises and helping premises that are underlying them.

3.4.1 Method

In [15] and [16], the question of "does it work" was addressed. To answer this question, all students who were part of a one-semester undergraduate Critical Thinking course at the University of Melbourne and used Reason!Able during this project, were asked to complete a pre-test and a post-test that was based on the California Critical

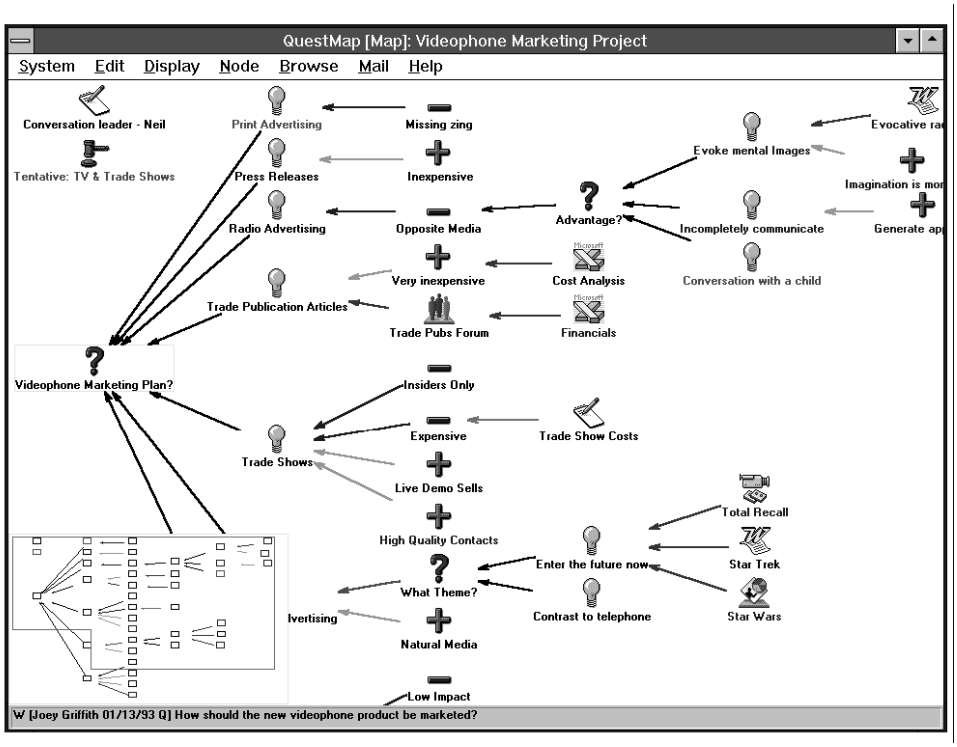

Figure 4. Screenshot of Questmap

Thinking Skills Test. This test consisted of 34 multiple-choice questions. Obviously, this experiment was not internally valid, because no control group was used so that a valid comparison of the results is impossible, although the measure seems to be reliable.

A similar study was reported by [17] in which students were also pre-tested and post-tested using two tests, namely the California Critical Thinking Skills Test and written test in which students had to identify the main conclusions, reformulate the reasoning, and evaluate the reasoning of a short argumentative text. The latter was assessed by two experts. Methodological details were missing so no real assessment of the internal validity can be made. But since no direct control group was available, internal validity will be limited.

Another, more elaborate, study was reported in [14]. Students were learning argumentation skills during a period of 16 weeks; one group of 32 students participated in a traditional course, another group of 53 in a Reason!-based course. The latter was allowed to use Reason! (a predecessor of the Reason!Able programme) to construct argument trees. Both groups were pre-tested and post-tested using the Watson-Glaser Critical Thinking Appraisal; another multiple-choice test. The students in the Reason! group were also asked to complete the written pre-test and post-test. Although two groups were tested, those were not compared to each other. This means that no real between-subjects design was used. Moreover, it was not mentioned whether randomization was used. Therefore, this experiment cannot be considered to be internally valid.

So, similar to Questmap, Reason!Able aims to provide support to make its users better reasoners. Several studies were performed, which were not internally valid. During the course, students had to produce their own arguments but the written pre-test and post-test consisted of the reproduction of an argument from an argumentative text. Similarly, the multiple-choice tests involved identifying proper arguments rather than constructing arguments. This means that the task that measured the students' skills considerably differed from the assignments during the course, although both involved the identification of arguments and counterarguments.

3.4.2 Results

In the first study it was found that the students' scores improved with almost 4 points over the last three years ($SD = 0.8$). Generally, it is assumed that the students' performance in any subject would normally be expected to improve by only 0.5 standard deviation over three years. From this the author concluded that the Reason! approach improved the students' critical thinking skills and was more effective than traditional approaches. Unfortunately, no valid experimental design was used to compare these results statistically.

Similarly in [16] and [17] it was claimed that the approach improved the students' skills more over one semester than traditional approaches that needed the entire undergraduate period to achieve the same result. Reason! was claimed to be three to four times more effective than traditional approaches that do not use the Reason!Able software. However, these claims seem to be premature, as the experiments were not valid.

In the last study, two groups of students were tested but not compared to each other. In [14] significant progress was reported for the Reason! users ($p < 0.05$) on both the multiple-choice and the written test, while the traditional group did not display a significant gain in reasoning skills. But since internal validity was not assured, no safe conclusion can be drawn.

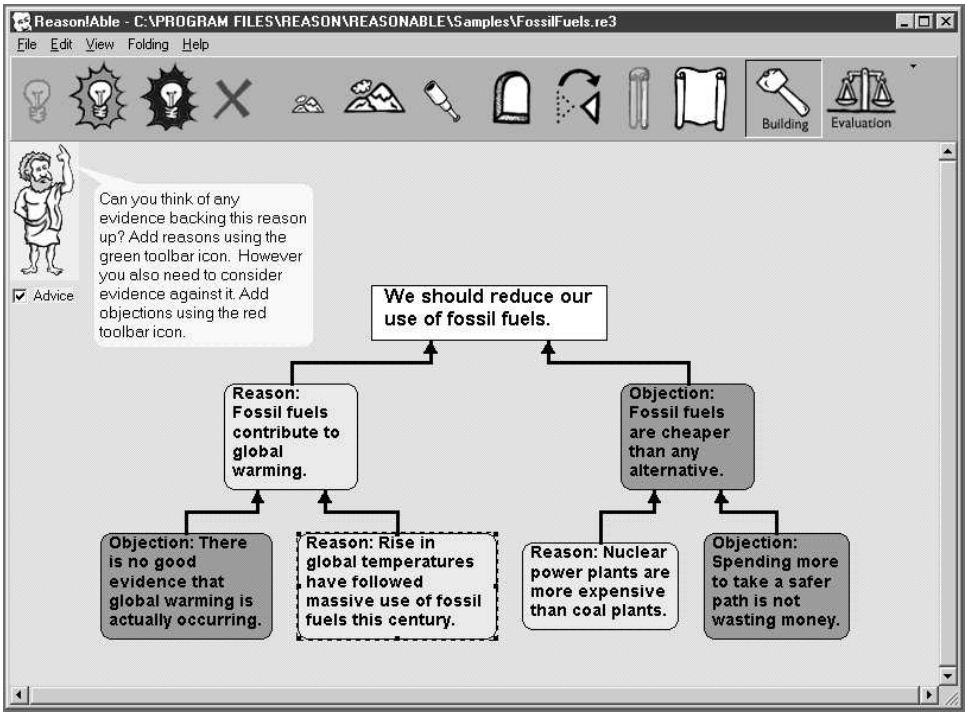

Figure 5. Screenshot of the Reason!Able software tool

4 Discussion

The experiments described above significantly differ. The most important methodological differences are concerned with the nature of the task that had to be performed, the measures used, and the underlying argumentation theory. These differences are summarized in Table 2.

With respect to the task, the main differences had to do with the intended effect of use. Also the nature of the tasks differed, as in some experiments the participants had to produce the arguments themselves, while in other ones reproduction of arguments based on a argumentative text was asked or multiple-choice test had to be completed. Moreover, sometimes collaboration was mandatory, while in other cases users had to work individually. In most experiments subjects had to establish supporting and attacking (or contradicting) relationships.

The measures that were used also differed. Although most of them involved expert assessment, there was a lack of information about the criteria that were used to assess the quality of the users' reasoning. Similarly, little is known about the contents of the multiple-choice tests. As far as the measures of argument quality are concerned, another important distinction has to be made. Two different aspects are measured, firstly, the quality of the arguments' structure. For example, this is measured by the number of nodes used (is there a sufficient amount of detail) or the validity of the structure. Secondly, the quality of the content of the argument is measured, for instance, by expert assessment.

It was found that most results indicated that the tools have a positive effect on argumentation skills and make the users better reasoners. However, most experiments did not yield significant effects. The observation that different underlying argumentation theories were used is relevant for the conclusions drawn. Results that are not significant may be caused by an underlying theory that is not suitable for the task at hand. For example, an IBIS-based system may not be suitable for the task of constructing legal arguments.

The difference in measured effects means that we have to divide our conclusions into three subconclusions on argument quality, argument coherence, and critical discussion skills. Significant effects were only found for argument coherence. For argument quality the effects were not significant, but trends were shown in the positive direction. These trends both concerned argument structure and content. No quantitative results were reported on discussion skills.

5 Conclusion and future work

This paper has provided a critical review of the most recent research into the effectiveness of argument visualization tools. Although it is promising that some researchers at least subjected their tools to testing, most of the experiments described in this paper were not completely valid. Sometimes it was even impossible to determine the validity of the results at all, as many important details were missing in the description of the experiments; in particular methodological and statistical details were not mentioned. As a consequence, due to a lack of internal validity, the differences found may not be completely caused by the use of the visualization tool but may have additional causes and due to a lack of external validity, the results cannot easily be generalized to other populations. Therefore, it is premature to claim that argument visualization tools cause higher quality arguments, critical discussion, or coherent arguments. But given the fact that most results point in the same direction, we think it is reasonable to assume that these tools have a positive effect on the users' argumentation skills.

However, a lot still remains to be done, because until now experiments have failed to provide significant evidence for the benefits of argument visualization tools. After all, significant differences have been found but only in invalid experiments, while in the internally

Table 2. Overview of methodological differences between experiments

	Experimental tasks				Experimental measures	Argumentation theory
	Effect of use	**Production**	**Links**	**Collaboration**		
Belvedere	critical discussion skills and quality of argument structure	production	attack and support	yes	amount of discussion, multiple-choice test, and expert assessment of essay by inferential strength, difficulty, and spread	arguments in terms of inference trees
Convince Me	argument coherence (structure)	*unknown*	*unknown*	*unknown*	correlation with ECHO	Thagard's theory of explanatory coherence
Questmap	quality of both argument structure and content	production	attack and support	yes but not mandatory in control group and not during post-test	the number of argument structures, the richness of arguments, and expert assessment of final exam	IBIS
Reason!Able	quality of argument content	reproduction (pre-test and post-test)	attack and support	no	multiple-choice critical thinking skills tests and expert assessment of written test	arguments in terms of inference trees

valid experiment the results have been not significant. More specifically, based on our assessment of the internal validity, we have to further restate our conclusions and say that with respect to the experiments on Belvedere (the first experiment), Questmap, and Reason!Able, no real conclusions can be drawn. Valid conclusions can be drawn from the second experiment on Belvedere that failed to prove a significant effect on argument quality, although a trend was proven in the positive direction.

Nevertheless, the designs of these experiments and their shortcomings are useful to give recommendations for future research on computer-supported argument visualization. First, the experiment has to be valid, so that the results that are found and the conclusions that are drawn are valid and can be generalized to larger populations. More specifically, at least a between-subjects design should be used with one control group. Second, the chosen measure should be reliable. Therefore, a quantitative, objective measure for the effectiveness of a tool should be developed, but it should be noted that this is not straightforward. The most reliable measure found so far seems to be expert assessment, that is, specialists are asked to assess the quality of the argumentation by criteria such as the completeness and validity of the argument constructed.

Now we have come to the point at which an action plan to conduct research into the effectiveness of argument visualization tools can be given:

1. Formulation of hypotheses.
2. Selection of the variables, especially choosing a dependent variable that is based on a valid measurement.
3. Selection of the subjects, especially choosing a representative sample for the population the results have to be generalized to, other important issues include the sample size.
4. Selection of the design, especially choosing between a within-subjects or between-subjects design, other important issues involve randomization, homogenization (between-subjects design), and balancing (within-subjects design).
5. Selection of the appropriate statistical tests in order to draw valid conclusions.

Preferably, the usability and user-friendliness of the visualization tool is tested first, so that it is easy enough for everybody to understand and use, and its complexity does not limit the constructed arguments. Subsequently, other experiments can be conducted that measure its effectiveness.

In short, this paper has made a contribution to the area of empirical research on argument visualization tools, in that it paves the way for a more scientific approach to this research and provides an action plan to conduct experiments. It is also relevant to our research project on crime investigations, since the effectiveness of the tool we plan to develop will be tested. Unfortunately, to our knowledge no experiments focus on the effects of such tools on police investigations. We are cautious to generalize the results described in this paper to the domain of evidential reasoning in police investigations, as external validity was not assured and the domain differs both in the type and setting of the reasoning (cf. teaching versus crime solving). Most of the described experiments did not concentrate on the effects on evidential reasoning but focus on more general reasoning and conflict resolution skills. Critical discussion and collaborative problem solving are other skills that are of use to police investigators. Taking this into consideration the results on Belvedere and Questmap are most relevant here, though no significant effects were demonstrated. This means that a lot remains to be done in this area and that as far as we know the experiment we plan to conduct on police investigators will be the first of its kind.

ACKNOWLEDGEMENTS

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 634.000.429. Henry Prakken and Gerald Vreeswijk were also partially supported by the EU under IST-FP6-002307 (ASPIC).

REFERENCES

[1] Chad S. Carr, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Using computer supported argument visualization to teach legal argumentation, 75–96, Springer-Verlag, London, UK, 2003.

[2] Thomas D. Cook and Donald T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin Company, 1979.

[3] Paul A. Kirschner, Simon J. Buckingham Shum, and Chad S. Carr, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, Springer-Verlag, London, UK, 2003.

[4] Henry Prakken and Gerard A.W. Vreeswijk, 'Encoding schemes for a discourse support system for legal argument', in *Workshop Notes of the ECAI-02 Workshop on Computational Models of Natural Argument*, pp. 31–39, (2002).

[5] Chris A. Reed and Glenn W.A. Rowe, 'Araucaria: Software for argument analysis, diagramming and representation', *International Journal on Artificial Intelligence Tools*, **14**(3-4), 961–980, (2004).

[6] Bertil Rolf and Charlotte Magnusson, 'Developing the art of argumentation: A software approach', in *Proceedings of the Fifth Conference of the International Society for the Study of Argumentation*, (2002).

[7] Patricia Schank and Michael Ranney, 'Improved reasoning with Convince Me', in *CHI '95: Conference Companion on Human Factors in computing Systems*, pp. 276–277, New York, NY, (1995). ACM Press.

[8] Albert Selvin, Simon Buckingham Shum, Maarten Sierhuis, Jeff Conklin, Beatrix Zimmermann, Charles Palus, Wilfred Drath, David Horth, John Domingue, Enrico Motta, and Gangmin Li, 'Compendium: Making meetings into knowledge events', in *Proceedings Knowledge Technologies 2001*, (2001).

[9] Daniel Suthers, Arlene Weiner, John Connelly, and Massimo Paolucci, 'Belvedere: Engaging students in critical discussion of science and public policy issues', in *AI-Ed 95, the 7th World Conference on Artificial Intelligence in Education*, pp. 266–273, (1995).

[10] Daniel D. Suthers and Christopher D. Hundhausen, 'Learning by constructing collaborative representations: An empirical comparison of three alternatives', in *European Perspectives on Computer-Supported Collaborative Learning, Proceedings of the First European Conference on Computer-Supported Collaborative Learning*, eds., P. Dillenbourg, A. Eurelings, and K. Hakkarainen, pp. 577–584, Maastricht, the Netherlands, (2001).

[11] Daniel D. Suthers and Christopher D. Hundhausen, 'The effects of representation on students' elaborations in collaborative inquiry', in *Proceedings of Computer Support for Collaborative Learning 2002*, pp. 472–480. Hillsdale: Lawrence Erlbaum Associates, (2002).

[12] Daniel D. Suthers and Christopher D. Hundhausen, 'An empirical study of the effects of representational guidance on collaborative learning', *Journal of the Learning Sciences*, **12**(2), 183–219, (2003).

[13] Paul Thagard, 'Probabilistic networks and explanatory coherence', *Cognitive Science Quarterly*, **1**, 91–114, (2000).

[14] Tim J. van Gelder, 'Learning to reason: A Reason!-Able approach', in *Cognitive Science in Australia, 2000: Proceedings of the Fifth Australasian Cognitive Science Society Conference*, eds., C. Davis, T. J. van Gelder, and R. Wales, Adelaide, Australia, (2000).

[15] Tim J. van Gelder, 'Argument mapping with Reason!Able', *The American Philosophical Association Newsletter on Philosophy and Computers*, 85–90, (2002).

[16] Tim J. van Gelder, 'A Reason!Able approach to critical thinking', *Principal Matters: The Journal for Australasian Secondary School Leaders*, 34–36, (2002).

[17] Tim J. van Gelder and Alberto Rizzo, 'Reason!Able across the curriculum', in *Is IT an Odyssey in Learning? Proceedings of the 2001 Conference of ICT in Education*, Victoria, Australia, (2001).

[18] Bart Verheij, 'Artificial argument assistants for defeasible argumentation', *Artificial Intelligence*, **150**(1-2), 291–324, (2003).

[19] Claes Wohlin, Per Runeson, Martin Host, Magnus C. Ohlsson, Bjorn Regnell, and Anders Wesslen, *Experimentation in Software Engineering: An Introduction*, Kluwer Academic Publishers, Boston, MA, 2000.