Chapter 6

Evaluation

Designers and programmers of Virtual Environments tend to guess about the best realization and implementation of interaction techniques or even whole applications. Many works have shown that user based assessment is an essential component of developing interactive applications and in this work it is shown that user based assessment is especially important for applications as complex and innovative as CVEs. Already the assessment of parts of the application by different users except the designers can substantiate or refute realizations of a specific CVE.
If those assessments are formalized they are called evaluations. There exist three different evaluation methods which can be applied to Collaborative Virtual Environments. The expert heuristic, the formative and the summative evaluation [53, 54, 74].
The *expert heuristic evaluation* is an analytical method. The evaluator is a field expert who determines problems with usability in the design phase of the CVE. It is important that the evaluating expert is not part of the developer group and not involved in the design of the CVE at all. In addition the expert has to assess the CVE as early as possible in the design phase and also has to determine problems during the entire CVE development cycle. Based on the expert's knowledge, problems concerning usability can be solved following the expert's recommendations. Especially when evaluating Collaborative Virtual Environments this process is of a particular challenging nature, because of the small number of VE experts worldwide. Additionally, there exist only a few VE user interface design guidelines and there is an absence of VE user interface standards.
The *formative evaluation* is an empirical, observational method. Thereby the evaluators assess the CVE throughout the entire development cycle. The output of this evaluation method is a combination of qualitative and quantitative results. The quantitative data evaluates the amount of time, the number of

trials, the number of mistakes etc. while performing a special task. The qualitative data can be obtained by observing so-called *critical incidents* [53]. A critical incident is a problem that occurs while a user is interacting within the CVE. These incidents can be confusion, cancellation, errors, repetition etc. (3.2). Hereby the term critical incident does not necessarily mean that these events have to have negative effects on usability. Positive events can also have an impact on usability and thus user performance and satisfaction. These positive incidents contribute to qualitative evaluation results as well.
The *summative evaluation* is an empirical method. The objective of this evaluation method is to compare between different CVEs designed with the information obtained from the same User Task Analysis (UTA see section 3.2). Hence the output of the summative evaluation method enables the statistical comparison of different realizations of interaction techniques, operations, representation components etc. and the choice of the most appropriate one in terms of usability of the CVE. As this evaluation is performed using nearly final implementations of the CVE the evaluating users are usually those users the CVE has been designed for. However, a more important constraint is that the evaluators have to be non-experts in VEs and have not to be involved in the design process at all.
Best evaluation results can be obtained when combining the three methods described above. With respect to its nature the expert heuristic and the formative evaluation method should be applied in the early phases of the design process. Already short alternating cycles of these two methods can eliminate the biggest problems concerning usability and user satisfaction. For the assessment of more subtle differences in realizations and implementations of CVEs summative evaluation is absolutely essential. However, the most important and often most complex part to manage while planning an evaluation is to determine items to be assessed. This collection of items is necessary to formulate specific questionnaires and hence to find and eliminate disturbance factors within the implementation of the CVE.

6.1 Evaluation of H-C-H Interaction

In order to determine the evaluation items mentioned above the Human-Computer-Human model introduced in chapter 3 is very helpful. As a reminder, in that chapter the three flows within the model have been determined. These are the $H \rightarrow C$, the $C \rightarrow H$ and the $H \leftrightarrow H$ via C flow (see Figure 3.1). The design objective is to enable Human-to-Human interaction in a CVE as if face-to-face where the computer as the mediator for this collaboration becomes omnipresent and transparent. From this point of view it is clear that each user

has to perceive itself and the remote partner being as present as possible in the Virtual Environment. Although the perception of presence is not the only requirement for a good collaboration at least it is the basic constraint for the establishment of the latter. But also appropriate representation of information is important for supporting collaboration. Hence when talking about possible problems and bottlenecks of the Human-Human interaction, the $H \rightarrow C$ and the $C \rightarrow H$ flows are the originators of those problems. For example the $H \rightarrow C$ flow can be disturbed through one or even more of the following disturbance factors:

1. unsuitable graphical and physical user interfaces
2. unsuitable physical input devices and equipment for generating user input
3. unsuitable representations of actions and events

Factors that might disturb the $C \rightarrow H$ flow include:

1. slow data processing and system reaction time
2. low network transfer rate and network drop outs
3. low graphical and acoustical quality

Hence, collaboration can be supported by the designers of the CVEs only if weaknesses such as the ones listed above are eliminated. Back to the assessment of the CVE these considerations imply that the disturbance factors have to be evaluated in order to find the best realization with respect to the User Task Analysis (UTA). However, the disturbance factors listed above are very generic and not linked to CVEs. These factors can be used as macroscopic parameters only. To arrive at a more pragmatic formulation of the disturbance factors, the items above are matched with the taxonomy from section 3.1 and the Awareness-Action-Feedback loops from section 3.6. The resulting evaluation items for the $H \rightarrow C$ flow are respectively:

- unsuitable menu representations (the user does not know where to find the desired function)
- unsuitable tool representations (the user does not know which tool has which functionality)
- unsuitable representation of data and its functionality (the user does not know how to process things and how to fulfill the task)

- unsuitable environmental representations (the user is confused by the surroundings and cannot concentrate on the task or transfer the learned skills to real-world applications)
- unsuitable input devices (the user is unable to work with/handle the input devices and to generate input with them)
- unsuitable physical equipment and annoying cabling (the user is confused by the cabling of input devices and shutter glasses and thus is not acting naturally or is unable to concentrate on the task)

For the $C \rightarrow H$ flow the comparison leads to the following evaluation items:

- real-time system reaction (the user selects items from menus or performs changes on the data set and sees intermediate reaction of the system without perceptible time delay)
- low graphical and acoustical resolution and quality (the user is unable to recognize tools, data set structures, actions of the remote partner and the partner itself, click and warning sounds, or is unable to talk to the remote partner)
- low network transfer rate (the user is unable to recognize the delayed actions of the remote partner, there are interruptions of the audio communication, the user is unable the map actions of the remote partner with occurrences of action feedback such as highlights, click sounds, movements of data sets)

The items mentioned above have great impact on collaboration and can be extremely disturbing to the Human-Human interaction. However, in addition to these items there exist other factors that might also have an impact on collaboration. The character of these factors is based on personal perception of collaboration. In this work evaluation items linked to personal perception are defined that have a quantitative or qualitative nature like :

- perception of the own presence within the CVE
- perception of the partner's co-presence within the CVE
- perception of the collaboration in terms of equality of rights
- perception of the quality of collaboration
- frequency with which the user looked to the partner

- frequency with which the user spoke with the partner

Considering all these evaluation items in one session is almost impossible. The reason is that these items evaluate too many different aspects of the Human-Computer-Human interaction. In order to address this great amount of items special evaluation sessions had to be defined which are able to let assess specific aspects of Human-Human collaboration as it has been defined in the *H-C-H* interaction model.

6.2 Evaluation Sessions

According to chapter 4 three different sessions are implemented that in this chapter are used for extensive evaluation of the interaction taxonomy model and to produce CVE design guidelines. These sessions are:

1. usability session
2. co-presence session
3. co-work session

Before the evaluation an initial session introduction to the system is included. The introduction session itself is not part of an evaluation session. As the evaluators must not be the developers and must not be familiar with Virtual Environments they have to be introduced to VEs. During this introduction they are informed of the display system, the equipment and the environment they are going to work with. The objective and advantage is that this introduction session creates almost same conditions for all evaluators. This is necessary in order to compare numerical results of the summative evaluation. In order to exemplify the evaluation sessions the scenario described in chapter 4 is used in the following subsection.

6.2.1 Usability Session

The usability session is the first evaluation. After the users (evaluators) are introduced to the Virtual Environment they interact autonomously within the VE for about five minutes. During this interaction an observer is taking notes. Beside the overall ability to interact with the system the critical incidents of the formative evaluation are the most interesting to the observer. The application of the usability session offers almost the same interaction techniques, operations, tool representations, menus and feedback components to the user as in the following two sessions. Only the data set is exchanged in order to

ensure that content specific operations have not been learned in this session already. If the observer gets the impression that the user is not yet familiar with the VE, the interaction time is extended. After the usability session is completed a questionnaire is handed out addressing usability assessments. After the questionnaire is completed the user gets another five minutes recovery time before starting the co-presence evaluation session.

6.2.2 Co-presence Session

The idea of this co-presence session is to evaluate the design of the CVE in terms of its support to the evaluator during a certain task using immersive telepresence only. As already mentioned earlier the perception of presence is not the only requirement for good collaboration but it is the basic constraint for the establishment of the latter. It is shown in the following sections that the evaluation results of the co-presence session are of high interest for distance learning applications.
In the co-presence session the user works again in the Virtual Environment but now with another data set. An experienced user who has been involved in the development process is present within the same environment using an remote audio/video connection.
The experienced user explains the task, the data set, the input devices and the tools remotely to the evaluator. Hereby the remote partner who acts like a supervisor does not use any input devices or tools. Only gestures and verbal instructions are used. The task is to position three bones as precise as possible in their correct location on a human female skeleton, as explained in chapter 4. These bones lie in front of the evaluator and look very similar to each other so that it is not obvious where they have to be add to the skeleton. If the evaluator does not know how to achieve the goal the supervisor gives advice about which tool should be used, how to query information about the bones, how to change the viewpoint etc. (see chapter 4). After the co-presence session is completed a questionnaire is handed out addressing co-presence assessments (see section 6.3.3). After the questionnaire is filled out the user gets another five minutes recovery time before starting the co-work evaluation session.

6.2.3 Co-work Session

The idea of the co-work session is to evaluate the design of the CVE in terms of its support for collaborative work and minimum time required to fulfill the task. Now it is important that both partners have equal rights concerning decisions, manipulations and the access to tools. It is interesting to note that in the implemented session both users can complete the task autonomously

as well. So the main question is to evaluate whether the CVE is capable of supporting and encouraging for team work although team work is not really required for this task.
The evaluator works in the VE and the remote partner is present using an audio/video connection again. Although the partner is one of the developers and thus an VE expert both are equal in rights and are using the same tools for interaction. The task is slightly different to the one of the co-presence session. Together the users have to position six bones belonging to three different pairs to complement the human female skeleton. Each bone in a pair belongs to the left or right side of the skeleton (i.e. the femur bone of the right and the left leg). A set of three of these bones lie in front of each user. As the users stand opposite each other on different sides of the skeleton they have to find out which bones belong to their side as the bones are mixed. This can be done by querying the name of the bone or by comparing them directly which is however more complicated. If a user finds out that a bone belongs to the partner this bone can be exchanged by passing it over to the other side. After a bone has been positioned in the skeleton the user can make use of a snap-back tool which lets the touched bone snap into the correct position. Thus it is possible to verify their own or their partner's work. In order to complicate the task the human female skeleton is covered by its own skin. For positioning the bones the particular part of skeleton has to be made visible by cutting away the skin in this region. For doing so one user selects a special cutting tool from the ring menu and apply this to the interesting part of the female's body. It is not possible to cut the skin in this region permanently. This means that the cutting user has to hold the skin cutter while the other user positions the bone. Instead of using a snap back tool to verify the position the user can grow the skin back and verify whether bones stick outside the body or not. After the six bones have been set into the skeleton correctly the co-work questionnaire is handed out (see section 6.3.4).

6.3 Evaluation Questionnaires

The different evaluation sessions are implemented because the evaluation items from section 6.1 assess too many different aspects of the *H-C-H* interaction model. Therefore questionnaires are developed to let the evaluators assess these different aspects.
The items of the questionnaires are enumerated. The usability questionnaire starts with *A*, the co-presence, co-work and observer with *B*. *C* and *D* respectively. All answers are ranged in the interval of 0 - 6. This is done in accordance with other evaluations [91, 92, 97, 103]. In order to support

the evaluator assessing the different aspects of interaction, descriptive text is placed beside the answer possibilities (i.e. 0 corresponds to never/bad/no, 3 to sometimes/acceptable/maybe and 6 corresponds to often/good/yes). The text based support makes it possible to place assessments on a numeric scala more precisely whereas the numeric results are necessary for the statistical analysis.

6.3.1 Introduction Questionnaire

Before the first evaluation session information about the user's profile is queried (appendix C). Interesting in terms of the evaluation is information about the evaluator's profession and experience with computers. Additional information on whether the evaluator is right or left-hander is asked. Early, external observations showed that it is not necessary to make a distinction between different genders. This work shows later again that only age and experience with computers rather than gender has an impact on the ability to get used to interaction in Virtual Environments in contrast to other researchers [92].

6.3.2 Usability Questionnaire

The usability questionnaire is handed out after the first evaluation session (appendix D). This questionnaire addresses evaluation items concerning the *H-C* interaction from section 6.1. Questions are listed assessing the quality and comfort of input devices, tool and device representations, positions of menus and the appearance of text in the VE.

6.3.3 Co-presence Questionnaire

The co-presence questionnaire queries information about evaluation items which have an impact on the perception of co-presence and on the communication between the two partners (appendix E). Social aspects of direct Human-Human communication are addressed here such as the influence of the size and shape of stereo glasses on communication and thus the exchange of information, as well as the position and the size of the partner's representation. In addition to that technical aspects are addressed such as the impact of network delays and drop outs on communication and collaboration.

6.3.4 Co-work Questionnaire

The co-work questionnaire is filled out after the last evaluation session is completed (appendix F). In this questionnaire information concerning the direct

team work is queried such as the frequency with which the evaluator was looking at the partner. Additionally evaluation items concerning the perception of co-knowledge and co-status are of high interest (see section 3.6). Social aspects such as the perception of being equal in rights during the collaboration are queried too. Important is that similar questions are asked in the co-presence and co-work questionnaire. The driving idea behind this is the evaluation of differences related to the social aspects of collaboration. The character of these two questionnaires enables investigations on differences in perception of co-presence. This difference might exist between situations where partners communicate only in comparison to situations where they communicate and work together (see analysis section 6.4).

6.3.5 Observer Questionnaire

This questionnaire is different from the others as it is not corresponding to a particular evaluation session (appendix G). It is filled out by an external observer who is an expert. This VE expert is observing the non-expert evaluator during the usability, the co-presence and the co-work session. Besides querying specific information about the time the user had to think and to debate before performing actions the questionnaire leaves also space for informal observations. These informal observations correspond to the critical incidents of the formative evaluation. These incidents observed from outside give feedback about the abilities of the evaluators. They have great impact on the following statistical analysis of the numeric evaluation results. Especially this questionnaire helps assessing items which are difficult to be assessed by the evaluator itself such as questions D5 or D6:

- Did the user loose concentration during a session ?
- How quickly could the user correct mistakes and continue the work ?

Information whether the evaluator lost concentration during a session has an impact on the analysis and the way the numerical results have to be interpreted. However, this information can also imply the high cognitive load of interaction in the Collaborative Virtual Environment too.

6.4 Evaluation Analysis

The evaluation analysis focusses on the three different evaluation methods introduced earlier in this chapter. These methods are the expert heuristic, the formative and the summative evaluation. In the early CVE design phase alternating cycles of expert heuristic and formative evaluation are performed in

order to eliminate obvious usability problems from the very beginning. For obtaining more subtle results concerning usability and team work summative evaluation is applied.
The CVE evaluated by 60 people has passed through expert heuristic and formative evaluation already. It is improved before given to summative evaluation. Now, the focus of this section is on the analysis of summative evaluation results. The first evaluation results and the design guidelines are given in section 6.4.2.
For analysing the numerical data obtained by the summative method expectancy values $\overline{x_j}$ are computed. In order to handle the uncertainty of the numeric results the standard deviation s is computed from

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where x_i is the value of measurement i, $\overline{x}$ the expectancy value and n the total number of measurements. s^2 is often also denoted as *variance*. Thus follows that the statistical certainty of the average value $\Delta\overline{x}$ is represented by

$$\Delta\overline{x} = \frac{t}{\sqrt{n}}s = \frac{t}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where the factor t is depended from the statistical certainty P and the total number of measurements n. Internationally in industry the statistical certainty is considered to be $P = 95\%$. Hence values for $\frac{t}{\sqrt{n}}$ are [65]:

$P = 95\%$	n	3	4	5	6	8	10	20	30
	t	4.3	3.2	2.8	2.6	2.4	2.3	2.1	2.05
	$\frac{t}{\sqrt{n}}$	2.5	1.6	1.24	1.05	0.84	0.72	0.47	0.37

However, before computing the expectancy values and their statistical certainty from the numerical results of the usability, co-presence and co-work questionnaires an analysis of the evaluators' profile is given.

6.4.1 User Profile

For the assessment of the CVE in expert heuristic and formative evaluation cycles two VE experts have been consulted who are not part of the CVE design team. Throughout the entire summative evaluation process almost 60 non-expert users have been evaluating the CVE. They worked together within the three sessions introduced in section 6.2.

Analysing the introduction questionnaire from appendix C a user profile is created. The age of the 60 evaluators is minimum 17 years and maximum 58 years. The majority are between 22 and 27 years old. Most of these evaluators are university students whereas the diversity of the others' professions reaches from secretaries and journalists over workers, technicians to technical and non-technical university professors and researchers. Although all evaluators are no Virtual Environment experts the knowledge concerning computer hardware and software differs substantially. The group of 22-27 years old uses the computer mostly for web surfing as well as computer games whereas the older evaluators use it for editing with text processing software. This is the reason why this first group is more experienced with hardware devices, such as game joysticks and steering wheels including force feedback. This observation is independent from the subject's profession or field of studies. A contrary result is that the older evaluators use a computer almost twice as long per week as the group of the 22-27 years old. No other significant differences between the evaluators that might have an impact on the analysis of the evaluation results are found.

6.4.2 First Level Analysis

The first level analysis is split into two parts. The first part deals with the results obtained by the expert heuristic and formative evaluation. Although the outcome is taken into account already it is quite instructive to discuss it separately from the results of the summative evaluation.
The usability findings and recommendations of the expert evaluators concern the following items [50]:

1. positioning of the toolbar grouping generic operations
2. handling of the ring menu grouping content specific operations
3. tool representations on menus
4. three button input device and stylus vs. pinch gloves
5. ego-centric vs. exo-centric viewpoint manipulation
6. graphical representation of data set
7. video frame rate

The User+Need Space (UNS) for the considered evaluation scenario of chapter 4 determines different representation forms for generic and content specific

operations. For the generic operations a toolbar is developed whereas the content specific operations are grouped by a special ring menu. In early designs of the CVE the generic toolbar was configurable in position by the user. The idea behind was that a dominant right-handed user might want to position the menu somewhere else in space than a dominant left-hander. Evaluation results showed that configuration of menus has a negative impact on the cognitive load. Additionally it is not really used in limited interaction spaces offered for example by the Responsive Workbench (RWB). Working with both hands at a RWB, the total viewing frustum is accessible in contrast to CAVE-like display systems. Thus during the formative and summative evaluation the toolbar was positioned close to the users body within arm distance corresponding to the vendor's tray metaphor. Working at a RWB this toolbar is fixed whereas it is attached to the user's body position when working in a CAVE or cylindrical and wall display systems.

Similar problems are encountered when using ring menus described in [49]. When a user intersects the data with the menu pick ray in the right hand the ring menu appears attached to the left hand and vice versa. This corresponds to the metaphor of handling a painter's palette with respect to dominant right and left-handers. The advantages were assumed to be the comfortable handling of this ring menu since it does not occlude any object being handled this way. For detaching the ring menu, over the shoulder deletion was integrated (see Figure 2.9). Evaluation results showed that the handling making use of the painter's palette metaphor is not always as comfortable as assumed. The reason is that the user first has to recognize that the status of the hand changed as something is suddenly attached to it. Then the user has to look at the ring menu in order to select a content specific operation using the other hand. This is particularly annoying if the hand is busy with another task already. Additionally this metaphor makes it impossible to concentrate on the data set as the user is forced to turn the head towards the ring menu. In the improved design the ring menu is attached to the calling hand holding the menu pick ray. It follows the translation of the user's hand whereas the rotation of the user's wrist is used to intersect the ring pieces with the pick ray. The advantages are that the menu appears within the user's gaze and disappears as soon as the user releases the stylus button again. The menu is designed to be 70% transparent to avoid occlusion of data (Figure 3.6).

As already mentioned the menus group operations together. In order to apply operations, tools are selected, e.g. the zoom operation requires a special zoom tool. The tools are represented by 3D icons which are attached to the buttons of the toolbar or to the choices of the ring menu. Usability findings showed

that representations for the snap back tool, the information tool and the skin cutting tool were not appropriate in the early CVE design. Now the snap back tool is represented by a three dimensional hook icon, the information by a three dimensional "i" letter icon and the skin cutter tool by a three dimensional knife icon. These virtual tool representations increased the evaluator's tool recognition rate by almost 80%. Evaluation results indicated also that early approaches using two pinch gloves as input devices were not really addressing the user's needs. Reasons are the uncomfortable usage when working stand-alone collaboratively and trying to hand over pinch gloves to another user. Another encountered problem using pinch gloves together with pick rays is that it is almost impossible to keep pointing somewhere and additionally snap with the middle finger and the thumb for selection. Similar problems using pinch gloves have been encountered in [54]. Improvements are made by using a special three button tool in one hand and a stylus in the other. The reason for not using three button tools in both hands refers to the high cognitive load of their usage due to the many buttons. After modification evaluation showed that the stylus is rather used in the dominant and the three button tool in the non-dominant hand.

A *sharing viewpoint* metaphor is implemented for manipulating the users' viewpoint [49] (see chapter 4). Evaluation results showed that an exo-centric viewpoint manipulation is better than an ego-centric when *standing almost beside the partner*. In this context exo-centric manipulation is based on how a user would act in real world by moving laterally. When sharing the same viewpoint (*looking through the partner's eyes*) or sharing the mirrored viewpoint (*looking from opposite the partner*) ego-centric viewpoint manipulation is implemented. This manipulation is realized by pressing and releasing a special button on the three button tool. These observations are valid working at a Responsive Workbench. Because of the limited interaction space it is possible to access the data set visually from all sides by manipulating the viewpoint as described above. However, other own evaluations showed that in the CVE implemented using a CAVE and a cylindrical display no ego-centric viewpoint manipulation is needed. Here users prefer exo-centric viewpoint manipulation due to the larger interaction space and the perception of entire immersion.

In the co-work session the evaluators complement a female skeleton by missing bones (section 6.2.3). There the task is aggravated as the skin of the body is cut in order to make the skeleton visible. Usability findings indicated that users prefer to get a quick overview of the situation. This leads to the implementation of a content specific wireframe operation. The users are able to only render the skin of the body in wireframe and thus have a direct view onto the

underlying skeleton. With this, strategies can be discussed and collaborative tasks can be planned more quickly. This content specific wireframe operation is only usable for getting an overview. For complementing the skeleton the skin has still to be cut.

In addition, observations of critical incidents are made during the co-presence session. These critical incidents occur due to network drop outs, indicating that the perception of co-presence is interrelated with the video frame rate. Further experiments with the video frame rate as parameter showed that the perception of co-presence vanishes completely if the video frame rate sinks below 12 fps.

Statistical Analysis

The next phase of the first level analysis is the statistical analysis of the summative evaluation results. In literature the analysis of numerical data is often restricted to the direct comparison of calculated statistical values [92]. In order to correlate with the work of other investigators direct comparison is performed [103]. Computing the average (expectancy) value and its statistical certainty $\overline{x} \pm \Delta\overline{x}$ for the usability questionnaire from appendix D leads to the following values.

Question	*A1*	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*	*A8*
$\overline{x}$	**5.4**	*4.7*	*4.2*	**3.8**	*4.6*	*4.6*	*4.9*	*4.7*
$\Delta\overline{x}$	**0.54**	*0.68*	*1.05*	**0.99**	*1.05*	*0.77*	*0.76*	*0.79*

A conspicuous feature of the average values is their low statistical certainty represented by the high values for $\Delta\overline{x}$. This is especially the case for low average values like they appear for questions A3, A4 and A5. Obviously one reason is that the statistical certainty of the average value is dependent from the number of evaluators. The larger the number of evaluators the more certain the expectancy value $\overline{x}$ and thus the lower $\Delta\overline{x}$. But there is another reason why the statistical certainty is small for high average values and vice versa. The distribution of answers to questions with small average values is more spread whereas the distribution of answers to questions with high average values is sharper. For example the distribution of answers to a question with an average value of "3" might look like: 2 times "0", 2 times "3", 2 times "6". Here the answers are spread and the statistical certainty is very low. An example for the distribution of answers to a question with an average value of "5" might look like: 2 times "4", 2 times "5", 2 times "6". In this example the answers are distributed much closer around the true expectancy value of "5" and thus the certainty of the average value is higher. In any case, the spread distribution of answers to a question indicates that the evaluators have

different opinions about the particular evaluation item. Although it is much more expressive to have a uniform answer pattern the spread distribution of answers implies problems with usability. If so, more investigations concerning the special evaluation item have to be made.

However, the average values represent important information about usability too. The highest average value with corresponding high certainty is computed for question A1 rating the responsiveness of the CVE (see appendix D). High values are also found for A2 (usability of input devices), A7 (alignment of tools) and A8 (alignment of menus). Lowest average values but with low certainty are computed for A3 (working with stereo glasses) and A4 (working with the cabling). As discussed above the statistical values for A3 and A4 indicate that not all evaluators perceived the work with stereo glasses and cabling as especially uncomfortable. The analysis of these special usability assessments in conjunction with the analysis of the external observations delivered no further results. From this, preliminary conclusions are drawn, indicating that the bad assessment of the stereo glasses and cabling is mainly influenced by personal perception rather than general frustration with the system or even inability to use the CVE.

Proceeding with the analysis of the co-presence questionnaire the following statistical values are computed:

Question	*B1*	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	*B9*	*B10*	*B11*
$\overline{x}$	*5.1*	*4.6*	*4.2*	**5.5**	*5.1*	*4.9*	*4.6*	*3.8*	*3.8*	*4.3*	**3.1**
$\Delta\overline{x}$	*0.83*	*0.77*	*1.22*	**0.52**	*0.77*	*0.88*	*1.05*	*0.92*	*0.71*	*1.07*	**1.35**

The average values representing the answers to questions B1 to B7 are notably high. For example most of the evaluators perceived the partner standing as a real person (B4) on the other side of the table (B1). Also the partner's stereo glasses did not have a very high impact on the perception of co-presence (B2). Remarkably are the average values for questions B5 and B6. It seems that evaluators did not perceive the remote partner to be less present when delays in video and audio transmission occur. Additionally B7 and B8 show that evaluators think the audio and video representation of the partner are necessary to complete the task even though the partner was not looking very often to some of the evaluators. This is indicated by the high uncertainty of the B2 average value. In comparison to this it seems contradictory that B10 shows an average value of about $\overline{x} = 4.3 \pm 1.07$. B10 is assessing a higher transfer rate to be an important factor for increasing co-presence. Very interesting is the high uncertainty $\Delta\overline{x}$ of question B11. B11 is assessing the position and the

size of the partner to be an important factor for increasing co-presence. Observations showed that especially evaluators who were much taller or smaller than the remote partner in reality thought that this parameter increases the co-presence. This explains the spread distribution of answers to B11.

Computing the statistical values for the questions of the co-work questionnaire results in the following table:

Question	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*
$\overline{x}$	*5.7*	**5.8**	*5.0*	*5.2*	**3.1**	*5.1*
$\Delta\overline{x}$	*0.49*	**0.46**	*1.02*	*1.06*	**0.86**	*0.53*
Question	*C7*	*C8*	*C9*	*C10*	*C11*	*C12*
$\overline{x}$	*5.6*	*5.3*	*4.0*	*4.7*	*4.1*	*5.5*
$\Delta\overline{x}$	*0.70*	*0.76*	*0.83*	*1.18*	*1.10*	*0.61*

Average values assessing collaboration are remarkably high in comparison to the sessions A and B. The evaluators assessed the co-work session as a satisfying (C12) event where they really collaborated (C7) with a real person (C1) of equal rights (C8). This assessment is a very interesting result as it seems to disagree with the low value of question C5. Here the evaluators confirm that they were not looking often to the remote partner although they pretended to know always where the partner stood (C4). Which seems a contradiction at first, but becomes explicable when taking the assessment of question C9 into account. Here the evaluators state with an average of $\overline{x} = 4.0 \pm 0.83$ that body and hand gestures did not greatly enhance or support the collaboration. This leads to the preliminary conclusion that the actions and behaviour of the remote partner are adequately rendered by the representations of tools and input devices together with acoustic feedback. This is also underlined by the average values of questions C10 and C11 where the evaluators think they could have completed the task perhaps even without the video representation of the partner.

6.4.3 Simple Guidelines

With the help of the evaluations done so far it is possible to draw and summarize some conclusions.

Highlights of Expert Heuristic and Formative Evaluation

From the analysis of the expert heuristic and formative evaluation cycles the following guidelines can be extracted:

- Static menus should be simple, well positioned and non configurable. They are well suited for grouping generic operations.

- Dynamic menus should appear within the user's gaze (view) and must not occlude data. They must be designed in a way that the user is able to concentrate on the task while using the menu.
- During the task intuitive, recognizable tool representations are necessary. Additionally they are able to render the actions of the remote partner adequately.
- Input devices chosen should be easily passable from one user to another. Input devices should preferably support both, left and right handed users. Additionally they must allow precise application of an operation.
- Possibilities to get quick information or even an overview of the current situation in the CVE must be provided (wireframe technique, World in Miniature technique (WIM).
- Depending on the amount of interaction space and consequently depending on the display system intelligent partitioning of ego-centric and exo-centric viewpoint manipulation is necessary.
- Using immersive telepresence the video frame rate must not sink below 12fps.

Highlights of Statistical Analysis

From the statistical analysis of the evaluation results it is quite difficult to extract guidelines as they reflect the assessment of realizations by the evaluators. However, with the analysis of conspicuous distributions of answers to special questions it is possible to obtain guiding information.

- Cabling of input devices, trackers and stereo glasses are perceived as annoying. Careful handling of loose wires is recommendable.
- In a consultation situation immersive telepresence supports the work flow. In this situation network drop outs do not have negative impact on the perception of co-presence as long as the average frame rate does not go below 12fps.
- In a collaboration situation using immersive telepresence the position of the remote partner representation should be chosen in a way that both partners seem to have same virtual size in the CVE independent from their physical size in real world.
- Appropriate representations of the remote user's tools and input devices support collaboration more than body and hand gestures.

- When using a RWB the perception of co-presence can be increased with a remote partner's video texture representation together with a real background since due to depth perception the user has the impression that the remote partner stands closer to the table.

- When using a CAVE-like display system or cylindrical projections a remote partner's video texture representation without background is recommendable. This is possible, using a uniform background behind the user and a hardware keyer which subtracts this uniform texture pattern from the users outline. Using a real background together with the video texture of the remote partner leads to a mismatch of the non-photo realistic virtual scenery and the real video background (see Figure 5.3).

6.4.4 Group Analysis

Direct comparison of average values and their statistical certainty is a good method for getting initial quick results. In order to obtain more detailed information about usability of the CVE this method is not sufficient. Additionally, for investigating the problem of spread answer distributions another method than the direct comparison needs to be considered.

Human communication serves as an exchange of information. Independent of the appearance of this information the principle basis for communication is alternating cycles of query and answer. On the one hand queries are conceivable which only have one intention or at least expect one precise answer. On the other hand there exist more subtle queries which have more than one intention. According to evaluation it is possible to extract more than one intention from the items of the questionnaires. Within this thesis questionnaires are designed in a way that the same evaluation item could be evaluated by more than one question of different evaluation sessions. With the help of these special questions the following dependencies of evaluation items for the $H \rightarrow C$ and $C \rightarrow H$ flow are decoded:

- Perception of the system's responsiveness and its impact on collaboration (A1, B1, C1, C3)

- Impact of stereo glasses, cabling and input devices on collaboration (A2, A3, A4, B2, B3, D3, D4)

- Perception of tools and tool representations (A6, A7, C3, D1)

- Perception of menus and functionality
  (A5, A8, D1, D5, D6)
- Impact of audio/video transfer rate on co-presence and collaboration
  (B1, B5, B6, B10, C5)
- Perception of telepresence and its impact on collaboration
  (B6, B7, B8, B9, B10, B11, C5, C6, C9, C10, C11, D4, D6)
- Perception of co-knowledge and collaboration
  (B1, B4, C1, C2, C5, C6, C7, C8, C12, D2, D7)

With these decoded dependencies it is now possible to analyse the different evaluation sessions again.

Statistical Analysis

Comparing the average values and their statistical certainty of the questions A1, B1, C1 and C3 leads to the following table.

Question	*A1*	*B1*	*C1*	*C3*
$\overline{x}$	*5.4*	*5.1*	*5.7*	*5.0*
$\Delta\overline{x}$	*0.54*	*0.83*	*0.49*	*1.02*

All four questions assess the perception of system responsiveness, either explicitly or implicitly. In this group comparison question A1 directly queries information about the responsiveness of the system. The questions of the co-presence session (B) and the co-work session (C) assess the system's responsiveness implicitly. These questions query information about the perception of collaboration and co-presence explicitly. Remarkable is that all the average values and their statistical certainty highlight the system's responsiveness equally good and its impact on co-presence and collaboration as positive.

Questions A2, A3, A4, B2 and B3 evaluate the impact of hardware components such as input devices, cabling and stereo glasses on collaboration.

Question	*A2*	*A3*	*A4*	*B2*	*B3*	*D3*	*D4*
$\overline{x}$	*4.7*	*4.2*	*3.8*	*4.6*	*4.2*	*5.4*	*5.2*
$\Delta\overline{x}$	*0.68*	*1.05*	*0.99*	*0.77*	*1.22*	*0.69*	*0.76*

The usability ratings of the evaluators are average. The influence of hardware equipment on collaboration is assessed as average. The diversity of answers is expressed by the high uncertainty $\Delta\overline{x}$ of A3 (*comfort of the stereo glasses*) and B3 (*was your partner looking at you*). In accordance to these assessments the observer questions D3 and D4 confirm that the evaluators had

problems selecting tools and problems with orientation. Here the statistical certainty of observer evaluations are average.
Similar evaluations are encountered for the representations shown in the first part of the following table:

Question	*A6*	*A7*	*C3*	*D1*
$\overline{x}$	*4.6*	*4.9*	*5.0*	*3.2*
$\Delta\overline{x}$	*0.77*	*0.76*	*1.02*	*0.73*

Question	*A5*	*A8*	*D1*	*D5*	*D6*
$\overline{x}$	*4.6*	*4.7*	*3.2*	*4.6*	*4.8*
$\Delta\overline{x}$	*1.05*	*0.79*	*0.73*	*0.59*	*0.85*

The evaluators assessed the tool representations bit better than average suitable (A6) and their alignment as more or less appropriate (A7). With C3 the evaluators confirmed to know about what the partner was doing in the collaboration session which is ascribable to the tool representations. However, the high uncertainty of C3 indicates that some evaluators knew about the partner's actions but others did not. The time which was required to think about an operation before applying is assessed by the observers to be average with a normal certainty value.
The second part of the result table shows the ratings of A5, A8, D1, D5 and D6. Users evaluated the appearance of text (A5) and the alignment of menus (A8) to be a bit better than average but not as good. Informal observations showed that the high uncertainty value for the assessment of text in the CVE is mainly influenced by the physical size of the users. Users which were smaller than the average user had problems to read the text due to the declination of their viewpoint. Obviously text is not readable from any point in the interaction space. The assessment from the observers imply that the evaluators did not have to think too much about performing a certain action implied by problems with the menus (D1). But remarkable is that observers notice a loss of concentration during the sessions (D5). Even more, the highest assessment in this context is encountered for question D6. Here the observers confirm that the evaluators were able quite quickly to correct mistakes and continue the work.

The following table shows statistical values for questions rating the impact of audio/video transfer rate on co-presence and collaboration.

Question	*B1*	*B5*	*B6*	*B10*	*C5*
$\overline{x}$	*5.1*	*5.1*	*4.9*	*4.3*	*3.1*
$\Delta\overline{x}$	*0.83*	*0.77*	*0.88*	*1.07*	*0.86*

Evaluators confirmed that they perceived the remote user as standing on the other side of the table. Even a delay in video and audio transfer rate did not have a negative impact on immersive telepresence and thus on the perception of co-presence (B5, B6, B10). As described in chapter 5 the audio and the video streams are not synchronized. This does not seem to have negative

impact either. Experiments with the network delay showed that this is valid as long as the transfer rate is high enough to guarantee smooth movements. The magic threshold is 12fps. This result is remarkable as most of the evaluators confirm that they did not look often to the partner during collaboration. From this the conclusion is drawn that the video representation of the remote partner serves rather psychological than direct support purposes. This point is discussed in section 6.4.6 again.

The following tables show even more decoded dependencies than the ones above.

Question	*B7*	*B8*	*B9*	*B10*	*B11*	*C5*
$\overline{x}$	*4.6*	*3.8*	*3.8*	*4.3*	*3.1*	*3.1*
$\Delta\overline{x}$	*1.05*	*0.92*	*0.71*	*1.07*	*1.35*	*0.86*
Question	*C6*	*C9*	*C10*	*C11*	*D4*	*D6*
$\overline{x}$	*5.1*	*4.0*	*4.7*	*4.1*	*5.2*	*4.8*
$\Delta\overline{x}$	*0.53*	*0.83*	*1.18*	*1.10*	*0.76*	*0.85*

The perception of telepresence and its impact on collaboration is assessed implicitly by the questions B7, B8, B9, B10, B11, C5, C6, C9, C10, C11, D4 and D6. The table shows a very uniform answer pattern. Most of the average values are distributed into the interval between 3.0 and 5.0. The evaluators have the opinion that the video representation of the remote partner is necessary with an average value of $\overline{x} = 4.6 \pm 1.05$ (B7). Here the statistical uncertainty is quite high. On the other hand the evaluators said that maybe they are able to complete the task even without a video representation of the remote partner (B8). Again this result is accompanied with a high statistical uncertainty.
In accordance to the evaluation results, the perception of co-presence cannot really be improved by better lighting or positioning/scaling the representation of the partner (B9, B11). The audio/video transfer rate is perceived as a factor that might improve the situation (B10). It is conspicuous that the evaluators spoke more often (C6) than they looked at the partner (C5) during the collaboration session. Here the questions C10 (necessity of the partner's video representation) and C11 (ability to complete task without partner's video representation) are asked again (same with questions B7 and B8). The ratings are nearly the same as in the co-presence session which improves the certainty of the result as their statistical certainty of the average values is quite high with 0.92 to 1.18. Additionally C9 assesses the benefit of collaboration from hand and body gestures to be lower than expected. The observers confirm with D4 and D6 that the evaluators were able to correct mistakes and return quickly to work. The users did also not have significant problems with orientation and therefore the following conclusion is drawn.

The video representation of the remote partner has a psychological impact on the perception of co-presence, co-knowledge and collaboration. This means that the partner is contacted visually, (looking at him/her) only in the case of problems. Most of the time the audio connection and appropriate mapping of tool and input device representations seems to be much more important than the video connection. Informal observations emphasize this reflection. Section 6.4.6 deals with this problem again.

The group analysis of the perception of co-knowledge and collaboration is interesting as well as challenging. The following table shows the statistical results:

Question	*B1*	*B4*	*C1*	*C2*	*C3*	*C4*	*C7*	*C8*	*C12*	*D2*	*D7*
$\overline{x}$	*5.1*	*5.5*	*5.7*	*5.8*	*4.6*	*5.2*	*5.6*	*5.3*	*5.5*	*3.1*	*6*
$\Delta\overline{x}$	*0.83*	*0.52*	*0.49*	*0.46*	*0.77*	*1.06*	*0.70*	*0.76*	*0.61*	*0.72*	*6*

The high average values indicate that the collaboration is perceived as quite convincing. The remote partner was perceived as personally (B4,C1) standing at the opposite side of the table (B1). The collaboration was satisfying (C12) since most of the evaluators would like to work with the same remote partner again (C2). During the collaboration they perceived themselves as equal partners (C7) and also treated the partner as a person equal in rights (C8). Although the values are a bit lower than the average the evaluators knew where the partner was standing (C4) and what the partner was doing (C3) during the collaboration session. This values are confirmed by observer assessments of the amount of time the partners had to debate before they came to an agreement about a particular action (D2). The average value of 3.1 indicates that the time needed to do this is average. A clear "Yes" rating is given for the fun the users had while working in the CVE.

6.4.5 Advanced Guidelines

With the help of the group comparisons it is possible to draw some very important conclusions. From this the following guidelines can be extracted:

- Ensure high system responsiveness. It is perceived as having very positive impact on collaboration. Even downsizing the application in order to decrease the CPU load is recommendable. A good system responsiveness is guaranteed if all inputs and outputs are processed and rendered within less than 50ms.

- Although the work with input devices is assessed to have negative influence, this perception seems to be very subjective. However, it is essential

to facilitate the usage of VE input devices as well as shutter glasses and cabling.

- Using descriptive text in a Virtual Environment the developers should ensure that the alignment is realized with respect to the user's physical size. Readability should be provided from any point within the CVE interaction space. This is especially interesting when using a CAVE-like display system or a cylindrical projection. In this case descriptive text can be attached to the user's gaze, body or input devices.

- Appropriate tool and input device representations of the remote partner are adequate means for supporting the perception of co-presence which is the basic requirement for collaboration. With the help of these representations the influence of video is reduced to support collaboration only psychologically.

- When integrating immersive telepresence into a CVE, audio and video streams do not necessarily need to be synchronized. Even the resolution plays a tangential role.

6.4.6 Variation Group Analysis

Comparing statistical values of different evaluation sessions as well as comparing group items suffers from the small number of evaluators. Although it is possible to encounter and interpret trends in the answer behaviour the statistical certainties of the computed average values are very low. In order to overcome this problem special evaluation parameters are determined. These parameters change the initial evaluation conditions for different groups. The presumption is that when changing the initial evaluation conditions by the evaluation parameters special answer patterns are provoked. This produces better results than simply evaluating more users under the same conditions. The following evaluation parameters are defined:

1. Remote user's tool and input device representations.

2. Remote user representation using telepresence.

3. Enhanced collaboration.

In the first case the tool and the input device representations of the remote partner are removed from the co-presence and co-work session. This is done in order to force the partners working together to try and substitute the missing representations by other tools as good as possible. This change is assumed to

have a great impact on the perception of co-presence.
The remote user representation (video texture) is removed in another group and it is interesting to see whether the users are still able to work together. Most of the evaluators said that they would have been able to complete the task even without the remote partner's representation as the statistical results in section 6.4.4 indicate.
The last parameter is the collaboration session itself. An enhanced co-work session is chosen in order to force the users to work together in a more collaborative way. The parameter is supposed to have great impact on the perception of co-presence and co-knowledge and thus collaboration. The co-work session is changed in a way that the users do not have to complement the female skeleton by missing bones anymore while cutting skin. The new task is to assemble a snapshot of a running human out of the lying (standing) skeleton by changing the positions of all bones.
The last change in comparison to former evaluations is that the usability questionnaire is handed out together with the co-work questionnaire. The idea is to investigate the influence of the changed evaluation conditions on usability items too.

Statistical Analysis

The abbreviations for the four different evaluation sessions are *Ref* for the reference group, *NT* (no tools) for the group without tool and input device representations, *NV* (no video) for the group without remote user representation and *EC* for the enhanced collaboration task group. Thereby the reference group (*Ref*) is trying to fulfill the same task as the NT and NV group but working with all representations.
The first part of the statistical analysis deals with the investigation of perception of co-presence and the impact on collaboration. The dependencies which are decoded in the group analysis section above are still valid and are considered here again. Hence the perception of co-presence is intended by the following questions (B4, C1, B1, B6, B10, D4, D6):

Question B4		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.5	0.51
2. Group (NT)	5.8	0.30
3. Group (NV)	5.8	0.30
4. Group (EC)	**5.3**	**0.76**

Question C1		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.7	0.49
2. Group (NT)	5.6	0.55
3. Group (NV)	5.7	0.47
4. Group (EC)	**5.6**	**0.61**

Question B1		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.0	0.66
2. Group (NT)	4.5	0.78
3. Group (NV)	**1.7**	**1.13**
4. Group (EC)	5.2	0.69

Question B6		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	4.9	1.12
2. Group (NT)	5.1	0.82
3. Group (NV)	**4.0**	**0.35**
4. Group (EC)	5.0	0.84

Question B10		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	2.4	0.97
2. Group (NT)	4.6	0.85
3. Group (NV)	**4.7**	**0.90**
4. Group (EC)	3.6	1.45

Question D4		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.6	0.50
2. Group (NT)	5.3	0.59
3. Group (NV)	5.0	0.59
4. Group (EC)	**4.8**	**0.82**

Question D6		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.3	0.90
2. Group (NT)	4.6	0.85
3. Group (NV)	4.7	0.68
4. Group (EC)	**4.5**	**0.97**

Lowest ratings of the evaluators are highlighted using a bold font. The EC group assessed the questions B4 and C1 as lowest although the differences between the average values of the other groups are quite small. For example the difference to the highest ratings (NT and NV) is only about 0.5, which lies within the statistical uncertainty. In general it is conspicuous that the lowest assessments are given by the groups NV and EC, the group without video representation and enhanced collaboration session respectively. So for example did the NV group assess question B1 with a very low value of $\bar{x} = 1.7 \pm 1.13$. This result is not surprising whereas the low NV assessment is astonishing. The NV group says the partner is less present due to delays in audio transmission. It seems that people who suffer from the absence of a visual representation have a higher esteem for an audio connection as it is the only communication link to the partner. From this it is clear that the NV again assessed the transfer rate as the factor which is able to increase the perception of co-presence. This result is closely followed by the NT group which suffers from the absence of tool and input device representations and thus esteem an intact audio connection. Additional observer evaluations confirm that the enhanced collaboration group had less problems with orientation within the CVE (D4) but needed more time than others to correct mistakes and continue the work (D6). The NT and NV groups gave low assessments to question D4 as well. The reference group has the highest evaluations of D4 and D6.

This leads to the following conclusion: the enhanced collaboration and the absence of the remote partner representation show the greatest impact of the

evaluation parameters in contrast to the reference group without any changes. Absence of representations seems to affect the perception of co-presence. These users try to compensate for this absence by complementing representations. Working together more collaboratively on a more challenging task has positive influence on orientation within the CVE but leads to more handling errors. Further observations showed that the evaluators of the enhanced collaboration group (EC) lost concentration quicker than other evaluators.

The second part of the statistical analysis deals with the investigation of perception of co-knowledge and the impact on collaboration. Again the dependencies which are decoded in the group analysis section above are still valid and are considered here again. Hence the perception of co-knowledge and collaboration is intended by the following questions (C2, C3, C4, C6, C7, C8, C12, D2):

Question C2		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.9	0.31
2. Group (NT)	6.0	0.00
3. Group (NV)	6.0	0.00
4. Group (EC)	**5.8**	**0.46**

Question C3		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.0	1.02
2. Group (NT)	4.2	0.66
3. Group (NV)	**3.9**	**0.72**
4. Group (EC)	5.4	0.5

Question C4		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.1	1.06
2. Group (NT)	4.2	0.78
3. Group (NV)	**3.1**	**0.92**
4. Group (EC)	5.2	0.51

Question C6		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	**4.0**	**0.76**
2. Group (NT)	5.8	0.30
3. Group (NV)	5.3	0.59
4. Group (EC)	5.1	0.53

Question C7		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.6	0.69
2. Group (NT)	**5.2**	**1.22**
3. Group (NV)	5.2	0.79
4. Group (EC)	5.6	0.70

Question C8		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.4	0.97
2. Group (NT)	5.6	0.67
3. Group (NV)	**5.2**	**0.79**
4. Group (EC)	5.3	0.85

Question C12		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.5	0.96
2. Group (NT)	4.2	0.73
3. Group (NV)	**4.1**	**0.63**
4. Group (EC)	5.6	0.84

Question D2		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	3.8	0.89
2. Group (NT)	4.9	0.38
3. Group (NV)	4.4	1.12
4. Group (EC)	**5.5**	**1.12**

Again, most of the lowest assessments are given by the NV group, the one without remote partner representation. The enhanced collaboration EC group

assessed question C2 lower than the other groups. However, with a value of $\overline{x} = 5.8 \pm 0.46$ this statistical average still belongs to the generally high ratings of this question. Therefore question C1 will not be considered for further evaluations. Because of the absence of the partner and his/her input device and tool representations it becomes evident that the evaluators did not always know where the partner was standing (C4) and doing (C3). Therefore C3 and C4 show low average values for the NT and NV groups with NV a bit lower than NT. In order to compensate for the missing representations these groups spoke more frequently to the partner than other groups (C6). The lowest average value is encountered for the reference group. Obviously there was no need to talk much with the partner as they were working with all representations on an easy collaboration task.
Although there was more communication within the NT and NV than in other teams the evaluators did not have the impression to be equal in rights during the collaboration session (C7). Also the partner is not accepted to be equal in rights as C8 indicates. The collaboration was especially satisfying for the enhanced collaboration on the first place and the reference group on the second place (C12). NT and NV suffered again from the missing representations. The enhanced collaboration task which forces the partner to work together more collaboratively has positive influence on the overall satisfaction. The work was perceived as a success. The drawback is that these partners had to debate more before they could come to an arrangement about a particular action (D2). The observers gave low assessments for the NT and NV groups too.
After this analysis the following conclusions are drawn: it is confirmed again that users try to compensate for missing representations with other tools or forms of communication. To force users into a situation where they have to search for alternatives has negative impact on the perception of co-knowledge and user satisfaction.

The third part of the statistical analysis deals with the investigation of the impact on usability with respect to the changed initial evaluation conditions. Lowest ratings are highlighted using a bold font.

Question A1		
	$\overline{x}$	$\Delta\overline{x}$
1. Group (Ref)	5.2	0.74
2. Group (NT)	5.7	0.35
3. Group (NV)	5.2	0.57
4. Group (EC)	5.4	0.50

Question A2		
	$\overline{x}$	$\Delta\overline{x}$
1. Group (Ref)	5.5	0.68
2. Group (NT)	4.7	0.38
3. Group (NV)	4.0	0.76
4. Group (EC)	4.5	0.91

Question A3		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.0	0.59
2. Group (NT)	4.1	0.59
3. Group (NV)	**3.9**	**0.77**
4. Group (EC)	4.5	1.02

Question A4		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	4.9	0.47
2. Group (NT)	3.9	0.83
3. Group (NV)	**2.6**	**0.77**
4. Group (EC)	4.2	0.9

Question A6		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.1	0.61
2. Group (NT)	**4.1**	**0.86**
3. Group (NV)	4.3	0.83
4. Group (EC)	4.9	0.76

Question A7		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	4.8	0.57
2. Group (NT)	**4.2**	**0.66**
3. Group (NV)	4.3	0.90
4. Group (EC)	4.7	0.89

Question A8		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	5.2	1.06
2. Group (NT)	**4.3**	**0.59**
3. Group (NV)	4.3	0.77
4. Group (EC)	4.9	0.72

Question D1		
	$\bar{x}$	$\Delta\bar{x}$
1. Group (Ref)	4.3	0.90
2. Group (NT)	5.4	0.30
3. Group (NV)	**5.8**	**0.97**
4. Group (EC)	5.3	0.76

Most of the lowest assessments are given again by the NT and NV groups. Especially the NV assessments of user comfort concerning input devices (A2), stereo shutter glasses (A3) and cabling (A4) are very low. The average values are closely followed by the low assessments of the NT group. Both groups see the user comfort as problematic while working. The low average value of the reference group for question A1 is not fitting into this scheme. A1 is querying information about the responsiveness of the system. Apart from this, the highest assessments are always given by the reference group. Low average values with a quite high statistical certainty are computed for the NT group without remote tool and input device representations for questions A6, A7 and A8. These questions are querying information about tool representations and whether they are intuitive (A6) and about the alignment of tools (A7) and menus (A8). The external observers confirmed that the evaluators on the NV group had to think the longest before they were able to perform a certain action. It is conspicuous that the order of average values from low to high begins with assessments of the NV group, closely followed by the NT and EC groups. Obviously the enhanced collaboration does not affect usability items negatively although they are lower than the assessment of the reference group. These observations enable to draw the following conclusions.

Conclusions and CVE Rating Scheme

With the variation group analysis it is confirmed that the absence of representation forms has a negative impact on usability. It is proved by the statistical

results from the variation group analysis that a missing remote partner representation handicaps the CVE team more than missing remote tool and input device representations. The intensification of a collaborative work session without restrictions in representations shows impact on usability too. Now in conjunction with the conclusions of evaluation analysis sections 6.4.2 and 6.4.4 it is possible to formulate a CVE rating scheme.
This scheme consists of a chain which starts with the most important thing for a CVE, the audio link to the remote partner. Without audio it is impossible to work adequately. The next component is the video representation of the remote partner. Although this representation form is important it is not essential for the completion of the collaborative task. The users are able to compensate for this missing feature with other adequate tools or forms of communication. The third item is the remote tool and input device representation. These representations support completing the collaborative task but they are also not essential. It is proved from the conclusions of the former analysis sections that compensation always performs at the expense of usability or the perception of co-presence and co-knowledge. Users who do not suffer any missing representation features perceive the collaboration in a CVE as most satisfying. If only one feature is missing the users have to compensate for it by adequate other tools and mechanisms. As a consequence the users are unable to concentrate on the task. The compensating tools and mechanisms stress most of the user's senses in a way that these are occupied and overloaded. Therefore the users perceive the usage of equipment, virtual tool and menus as disturbing and confusing. Users who feel supported are rather willing to accept components which are weak in terms of usability.

6.4.7 Advanced Guidelines

Finally it is possible to formulate some further guidelines with the results obtained by the variation group analysis :

- CVE design and realization should consider the CVE rating scheme.
- An audio link to the remote partner(s)/team needs to be more reliable than a video link.
- A synchronization of audio and video streams is not necessary as long as the delay is not bigger than ten frames.
- Appropriate remote tool and input device representations are supportive but with minor importance relative to the video link.

- If appropriate remote tool and input device representations are difficult to realize ensure that equivalent, compensating tools and mechanisms are offered. Action feedback is able to help overcoming this representation drawback.

- Expert heuristic, formative and summative evaluations of the stand-alone Virtual Environment might not be able to identify weaknesses concerning the usability design for a collaborative Virtual Environment. The alignment of virtual tools and menus as well as the usability of input and output device combinations and other equipment should be designed and implemented with respect to CVE evaluation results.

- Work tools and mechanisms should be designed in order to disburden the users senses. High cognitive load, uncomfortable, non-intuitive usability and user fatigue have negative impact on the perception of co-presence and co-knowledge and thus collaboration.

6.4.8 Conclusions

This chapter introduced an evaluation framework for Collaborative Virtual Environments derived from a Human-Computer-Human interaction model. With the help of this evaluation framework 60 non-VE-expert evaluators assessed the CVE in terms of usability and collaborative awareness. For doing so alternating cycles of expert heuristic, formative and summative evaluation are applied. The statistical analysis of the numeric evaluation results are then used for the formulation of CVE implementation guidelines supporting team work. For performing intelligent evaluation, specific questionnaires and evaluation items are designed and a new analysis method (Variation Group Analysis) is developed.