A Comparison of Usage Evaluation and Inspection Methods for Assessing Groupware Usability

Michelle Potts Steves and Emile Morse
National Institute of Standards and Technology
Gaithersburg, Maryland, 20899 USA
+1 301 975-{3537, 8239}
{msteves, emile.morse}@nist.gov

Carl Gutwin[a] and Saul Greenberg[b]
Department of Computer Science
[a]University of Saskatchewan, Canada
[b]University of Calgary, Canada
gutwin@cs.usask.ca; saul@cpsc.ucalgary.ca

ABSTRACT

Many researchers believe that groupware can only be evaluated by studying real collaborators in their real contexts, a process that tends to be expensive and time-consuming. Others believe that it is more practical to evaluate groupware through usability inspection methods. Deciding between these two approaches is difficult, because it is unclear how they compare in a real evaluation situation. To address this problem, we carried out a dual evaluation of a groupware system, with one evaluation applying user-based techniques, and the other using inspection methods. We compared the results from the two evaluations and concluded that, while the two methods have their own strengths, weaknesses, and trade-offs, they are complementary. Because the two methods found overlapping problems, we expect that they can be used in tandem to good effect, e.g., applying the discount method prior to a field study, with the expectation that the system deployed in the more expensive field study has a better chance of doing well because some pertinent usability problems will have already been addressed.

Keywords

Evaluation, groupware usability, inspection evaluation techniques, usage evaluation techniques.

INTRODUCTION

Evaluation of groupware has received attention from researchers in the Computer-Supported Cooperative Work (CSCW) and groupware communities, e.g., [1,14,16]. However, evaluation is still considered a difficult problem [8], and many researchers feel that the only way to get a true picture of a groupware system is to study it in an actual context with real users. Although field methods are able to contextualize the evaluation, they can be both time-consuming and expensive; in addition, they can be difficult or impossible to perform if a system is not fully developed.

Recently, different types of groupware evaluation methods based on usability inspection techniques that do not utilize a real work situation have been proposed, e.g., [1,6,9]. These techniques are much less costly than field methods, and they can often be used earlier and more frequently in the development cycle. However, since these techniques are not used in the actual context of work, it is unclear whether the usability information they provide is valid for real users.

Other studies have compared various evaluation methods, e.g., [4,10]. We wish to build on this work by determining how, and if, inspection methods complement field methods for evaluating group software. In particular, we want to determine what kind of usability problems the techniques find, and whether the inspection method can provide an overall assessment of a system.

To explore these issues, we carried out two separate evaluations of the Teamwave Workplace (TW) groupware tool[1]. The first evaluation was a user-based study of how collaborators used the tool for real work performed over several months at the National Institute of Standards and Technology (NIST). We collected log data, had users self-report through diaries, and conducted a survey questionnaire and interviews with them. The second evaluation was an inspection of the tool in the Human-Computer Interaction laboratory at the University of Saskatchewan. In this much shorter study, independent evaluators assessed TW by trying the system over several different use scenarios, and examined how well it fit several inspection criteria. Evaluators then jointly synthesized the results into a problem report.

In the remainder of this paper, we report the results of these evaluations, and then use these results to compare how usage evaluation methods and inspection methods assess groupware usability. First, we describe the collaborative scenario where the tool was used. Second, we outline a set of usability principles called the *mechanics of collaboration* [9] that we used to orient both studies. Third, we report on the methodology and the main results of both the usage study and the inspection study. We then compare the

[1] Any commercial product identified in this document is for the purpose of describing a collaborative software environment. This identification does not imply any recommendation or endorsement by NIST.

techniques and their results. Our conclusions from this comparison are:

- The two different techniques provide a core of similar results, although each has particular strengths.
- The details of actual work practices and organizational context are needed to determine whether a system will be successful in a particular group-work situation.
- Inspection techniques are able to find a wide variety of usability problems specific to group activity, problems that overlap with those found by contextual methods.
- The low cost and versatility of inspection methods allow for the earlier and more frequent evaluation of groupware, even during development stages.
- Contextualizing inspection techniques through methods such as task-centered walkthrough can improve the focus and specificity of these low-cost evaluations.

GROUP-WORK SCENARIO: WELDING EXPERIMENTS

An ongoing research project at NIST investigates interface standards for automated, robotic-welding components. In this project, several welding researchers form a geographically-dispersed team working to define and test interface standards between robotic, arc-welding, work-cell components, controllers, and power supplies. The research is carried out using a welding testbed at the NIST facility in Gaithersburg, Maryland. New or modified equipment and interfaces can be plugged in and then tested during welding experiments. Analysis of completed welds is performed to verify effective operation of interfaces, equipment, and controllers [15]. Figure 1 shows the welding testbed, featuring the robotic arm with a welding torch and the fixturing table.

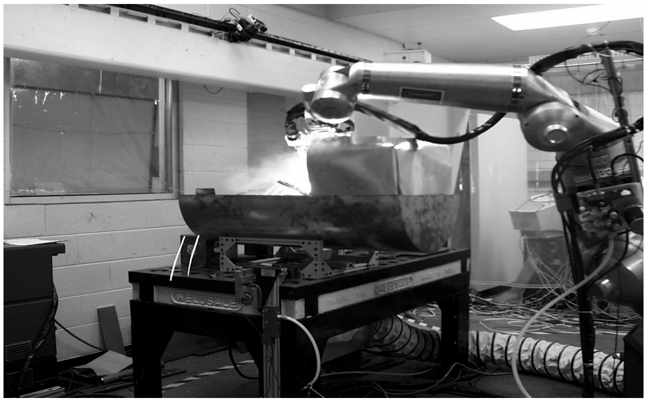


Figure 1. The remote welding testbed used by the group.

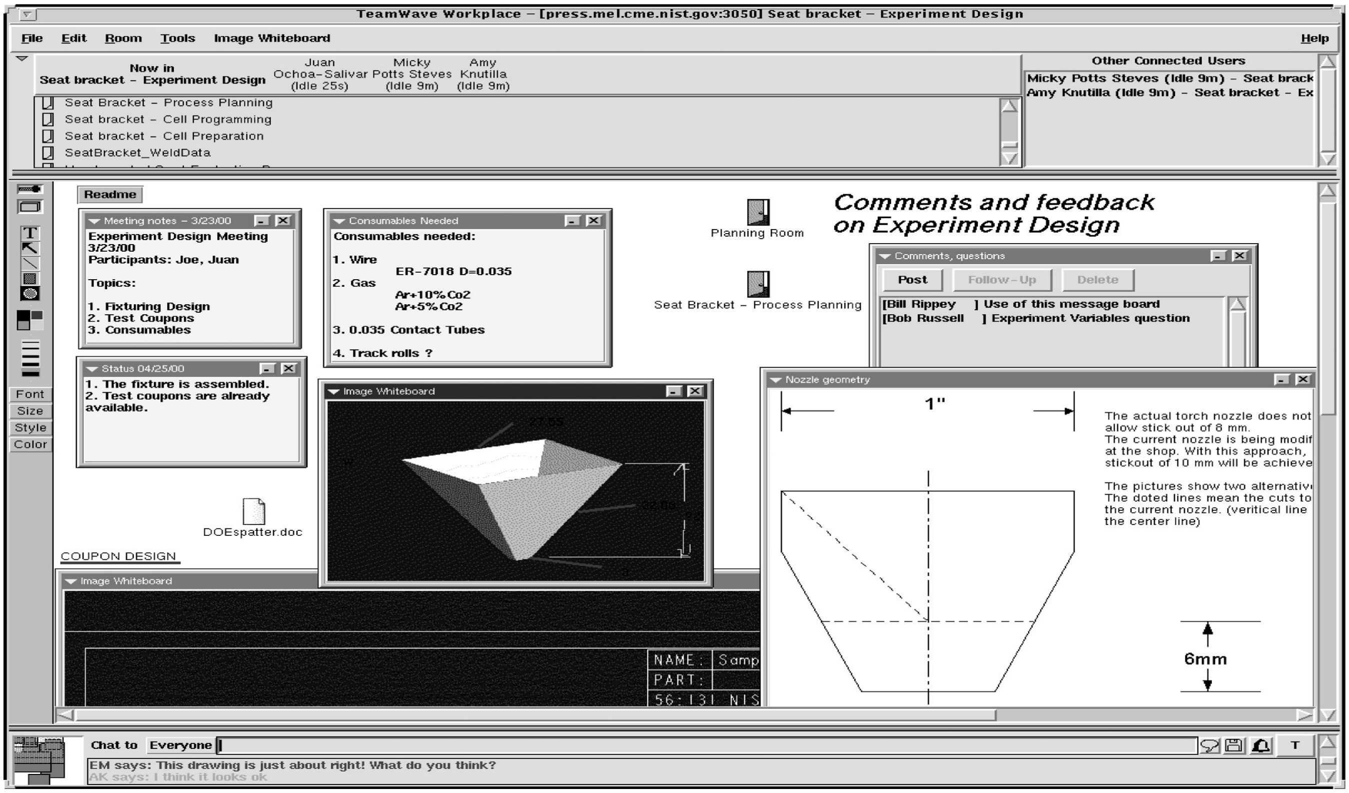
The research group consists of five to seven welding engineers and computer scientists. Members of the core group are located in various buildings at the NIST site, and guest researchers are generally located off-site.

The work in the welding experiment has both synchronous and asynchronous aspects. The team holds several full group meetings for planning and coordination, interspersed with periods of asynchronous individual activity, and smaller coordination meetings of two or three team members. A more detailed description of the welding scenario is given in [18].

Prior to using the TW groupware system, collaboration in this group was carried out using ad hoc methods. Meetings were often conducted face-to-face, with support for remote team members through telephone, email, and occasionally video or audio conferencing tools. Some asynchronous coordination efforts used email. Document transfer and distribution was accomplished using File Transport Protocol (FTP) and email.

GROUPWARE SYSTEM

Teamwave Workplace (www.teamwave.com) was selected as a groupware tool to support the distributed welding experiments (see Figure 2). TW is a room-based collaborative system with a relaxed-WYSIWIS (What You See Is What I See) whiteboard backdrop. “Rooms” in TW provide boundaries for data groupings and user interactions, and provide a metaphor for easing the transitions between synchronous and asynchronous work [7]. Occupants organize data spatially within rooms by placing various tools, documents, and graphics on the whiteboard backdrop. Objects and data within the virtual space are persistent between sessions. The TW system provides for synchronous and asynchronous user interactions, but, importantly, these interactions are in the context of relevant data. Figure 2 shows the Experiment Design room in TW that was used by the welding group. The pull-down menus allow users to add new tools to the room and to find out about other users. The main panel is the room’s whiteboard, on which several tools and datasets have been placed. Participants can draw on the backdrop using drawing tools shown on the left. At the bottom left of the screen is a radar overview, e.g., [17], that shows the entire room and each person’s current viewport into the room. To the right of the radar screen is a chat window where messages can be typed to some or all of the occupants of the room. To the right of the chat area are several controls for the chat tool; in particular, the “bell” button allows participants to send an alert beep to the other people in the room.



TeamWave Workplace – [press.mel.cme.nist.gov:3050] Seat bracket – Experiment Design

Now in Seat bracket – Experiment Design

Juan Ochoa–Salivar (Idle 25s)

Micky Potts Steves (Idle 9m)

Amy Knutilla (Idle 9m)

Other Connected Users

Micky Potts Steves (Idle 9m) – Seat brack

Amy Knutilla (Idle 9m) – Seat bracket – Ex

Seat Bracket – Process Planning

Seat bracket – Cell Programming

Seat bracket – Cell Preparation

SeatBracket_WeldData

Meeting notes – 3/23/00

Experiment Design Meeting
3/23/00
Participants: Joe, Juan

Topics:

1. Fixturing Design
2. Test Coupons
3. Consumables

Consumables Needed

Consumables needed:

1. Wire
ER-7018 D=0.035
2. Gas
Ar+10%Co2
Ar+5%Co2

3. 0.035 Contact Tubes

4. Track rolls ?

Status 04/25/00

1. The fixture is assembled.
2. Test coupons are already available.

Comments and feedback on Experiment Design

Comments, questions

[Bill Rippey] Use of this message board

[Bob Russell] Experiment Variables question

Nozzle geometry

The actual torch nozzle does not allow stick out of 8 mm.
The current nozzle is being modif at the shop. With this approach, stickout of 10 mm will be achieve

The pictures show two alternativ The doted lines mean the cuts to the current nozzle. (veritical line the center line)

Chat to Everyone

EM says: This drawing is just about right! What do you think?

AK says: I think it looks ok

Figure 2. The Experiment Design room in TW used by the welding engineers.

ORGANIZATION OF THE DUAL EVALUATION

Since the two evaluation techniques used in the comparison are very different and can yield very different types of results, we looked for ways to organize our approach so that the techniques would be comparable. In this section we describe a set of usability principles called the mechanics of collaboration that were used in both studies, and a high-level rating system that allowed us to compare study results.

The Mechanics of Collaboration

Groupware usability problems can be roughly divided into two parts: contextual issues resulting from social and organizational factors, and problems resulting from insufficient support for the *mechanics of collaboration*—the basic collaborative acts that are common across many work contexts.

The mechanics represent several basic activities of collaboration—the small-scale actions and interactions that group members must carry out in order to get a shared task done [9]. These actions are part of the teamwork (the work of working together) rather than part of the taskwork (the work that carries out the task). There are seven activities that are covered by the mechanics of collaboration.

Explicit communication. Group members must be able to provide each other with information. Verbal, written, and gestural communication are cornerstones of collaboration.

Implicit communication. People also pick up information that is produced implicitly by others going about their activities—information from artifacts being manipulated, or information from others' movements and actions.

Coordination of action. People organize their actions in a shared workspace so that they do not conflict with others. Shared resources and tools require that turns be taken, and some tasks require that actions happen in a particular order.

Planning. Some types of planning activities are carried out in a shared workspace, such as dividing up the task, reserving areas of the workspace for future use, or plotting courses of action by simulating them in the workspace.

Monitoring. People generally need to keep track of who else is in the workspace, where they are working, and what they are doing. In addition, situations such as expert-novice collaboration require more explicit monitoring.

Assistance. Group members provide help to one another when it is needed. Assistance may be opportunistic and informal, where the situation makes it easy for one person to help another, or it may be explicitly requested.

Protection. One danger in group-work is that someone may alter others' work inappropriately. People therefore keep an eye on their artifacts and take action to protect their work.

A Scale-Based Rating System

In addition to the individual usability problems we identified using each method, we wanted another way to compare the techniques at a higher level. Since usability studies are used generally to identify problem areas rather than to give summative assessments, the comparisons that can be made between two different techniques are limited.

To address this, we assigned a rating for each mechanic of collaboration as part of both evaluations. Although we realized that a single score for each mechanic can not represent the range of individual problems and strengths, we used the scores as a way to look for large differences in the results of the two methods. A description for the ratings is given in Table 1.

Rating	Description
+2	Very successful, few reports of problems
+1	Often successful, but some awkwardness
0	Adequate: no major problems or major benefits
-1	Useful in some situations but many drawbacks
-2	Rarely successful with many failures
N/A	Not enough information to rate

Table 1: Ratings and associated descriptions.

STUDY 1: THE USAGE EVALUATION

We carried out a usage evaluation with a group of welding engineers and researchers working on a welding experiment using Teamwave Workplace. The goals of the evaluation were to determine the strengths and weaknesses of the groupware system for the welding scenario, and to assess the impact of this technology on the group's collaborative processes.

The welding research team was composed of five people with six roles divided among them. The participants were all experts in the domain of automated, robotic welding and robotic control. Four of the five participants had worked together previously. The person new to the group, a guest researcher, held two roles and was involved in all aspects of the experiment.

Four of the five team members had offices in a building that is located 0.4 km from the welding testbed, the guest researcher was more remotely located. Each of the participants had access to the collaborative, virtual space via TW from their desktop computers, as well as from computers in the welding testbed. Although TW can work with audio and video conferencing tools, they were not used in this deployment.

The version of TW used in the study produced a server-based log file that contained information about the identity of users entering and leaving the distributed application, the identity of the rooms through which users navigate, room creation and deletion, file uploads, messages sent between users, and tool invocation, creation, modification, and deletion.

Usage Study Methodology

The study involved five steps: training, logging, questionnaires, interviews, and analysis.

Training. A three-hour training session and several days of individual experimentation preceded the actual work. The training session introduced the TW interface to the participants, including hands-on use of the main features. At this time, participants were asked to maintain a notebook to record any group activities in the experiment that were conducted outside of the TW environment.

Logging and monitoring. The welding experiment ran for 2.5 months. During this time, the evaluators periodically monitored the collaboration as observers within TW. The system logged 7620 events during individual and shared use of the tool.

Questionnaire. After the experiment, the welding researchers completed a questionnaire asking them to rate the TW interface (using a zero-to-ten scale) on several criteria including the mechanics of collaboration. Free-form comments were also gathered.

Interviews. Each of the participants was interviewed to follow up survey answers and to confirm initial interpretations of the evolving data analysis. Participants also ranked the importance of individual mechanics of collaboration for their welding work. We used this ranking to order the identified usability issues with respect to the welding researchers' priorities.

Analysis. Survey data were examined using statistical analysis to determine trends in satisfaction across the group. The system data was explored using a log visualization tool called the CollabLogger [11] and statistical analysis. Additionally, the chat logs recorded by the system were analyzed to determine major topics of conversation and general communication pathways.

Main Findings from the User-based Study

The user-based study was time-consuming and effort-intensive to set up and analyze. Many preparatory activities were carried out, including customizing TW to augment its logging capabilities, developing the log visualization tool, and preparing training materials, questionnaires, and interviews. The study produced a large amount of information, and two analysts spent more than five months compiling and analyzing the data.

These experiences are similar to other field-based methods, which are often costly to implement and analyze, e.g., [12,14]. However, these approaches provide an understanding of users in their own environments and provide clues about critical aspects of the real work situation that can be used to increase the realism and relevance of a usability test [19]. Our user-based study revealed how the welding team members worked together, their patterns of interaction, how meetings were conducted, and how they used the groupware tool to accomplish their work goals. This information provided a context for the usability analysis and a way of assessing the severity of identified usability issues.

Below, we first review some of these situation-specific findings, and then outline the main usability problems.

Observations of the collaboration

The study revealed several aspects of the group's collaboration, both in their work practices and in the ways that they used the groupware system. These observations are summarized below.

Use of the tool. The participants used TW for the full course of the study and seemed generally satisfied with the system. The system appeared to support certain aspects of the group's collaboration better than past practices. Participants liked the ability to leave notes and artifacts in one place and have that relevant data at hand later when conducting a meeting. Participants also liked that all team members could add artifacts and modify the space, as opposed to their previous situation where all files went through the one person in charge of the FTP site. The group also used the tool to augment their practices. In particular, they often saved the text of their chat sessions for others to review later. In fact, participants saw this capability as one advantage to chat-based communication that would be lost if audio tools had been used.

Organization of the work. There was considerably more asynchronous use of the system than synchronous use. TW space was only occupied by two or more persons approximately twenty-five percent of the total time it was occupied at all. Most of this time was spent in meetings. The welding researchers generally did not carry out lengthy individual work on the TW artifacts inside the system, so there were few chance encounters. Notification of changes to artifacts in the space was carried out by an email-notification scheme that the participants devised early in the study. In general, these results corresponded closely to the way that the group organized their work before the study. It points out, however, that the group was able to work in the same manner with TW that they had before the tool was used.

Communication. Conversations in the chat tool tended to be efficient, with more than eighty percent of messages pertaining directly to the subject matter and organizing the meeting. Of the remaining messages, four percent were social in nature, leaving about 12 percent that related to TW itself (questions, problems, or requests to another person in the space). Relative to the team's face-to-face meetings, the use of the chat tool did not change the way this group conducted their meetings. Larger coordination meetings were led by the project manager with input provided by other team members. Smaller meetings, generally one-on-one, had roughly equal amounts of give and take between peers.

Usability Assessment

The usability problems were primarily in the areas of explicit communication and monitoring, although problems were also found in the areas of implicit communication, planning, and protection. In order to compare our results with those from the inspection study, we compiled a list of the ten most obvious problems found, in order of severity. The associated mechanic of collaboration is identified with each problem.

1. (Explicit communication) Chat was awkward for meetings. We observed participants having difficulty following threads in large meetings (four or more participants); this observation was supported strongly by questionnaire and interview responses. Participants also reported difficulty ignoring non-relevant threads during meetings and difficulty following relevant side-conversations.
2. (Explicit communication) People found it difficult to communicate using the whiteboard, and found the whiteboard tools awkward and rigid. The single-user usability problems with these tools (e.g., people could not easily determine how to set properties such as color and font size) led to a group usability problem. That is, people would attempt to use the whiteboard tools to communicate, but would be unable, or slow, to do so.
3. (Explicit communication) Room boundaries, and the fact that chat in TW is visible only within the room, caused a variety of communication problems. Some participants had difficulty establishing a dialogue with a colleague who was in another room. In other cases, a participant would need to leave a room during a meeting, in order to fetch or modify an artifact located elsewhere—missing some of the meeting to accomplish the task.
4. (Explicit communication) Participants felt that they were unable to adequately communicate using visual artifacts, due to the limited types of artifacts possible within TW. Many of the group's conversations involved either pictures of previous welds or schematic diagrams of unit assemblies. The picture formats available in TW gave neither the resolution nor the expressiveness needed for proper communication. In particular, participants wanted higher-resolution image formats (e.g., Tagged Image File Format (TIFF) rather than Graphics Interchange Format (GIF)) and manipulable representations of objects and models (such as Virtual Reality Mark-up Language[2]) to support design review.
5. (Explicit communication) Participants felt that the more task-oriented communication tools in TW were too rigid for their work. Using the log data, we found that the two most popular tools (Post-it™ and Message Board) were also the most general-purpose.
6. (Implicit communication) Participants reported that they were unable to effectively and efficiently detect modifications to artifacts made by others in the rooms they were using without the email notification scheme they devised.

[2] http://web3d.org/technicalinfo/specifications/vrml97/

7. (Monitoring) From the log files, we observed participants occasionally taking a long time to notice chat messages or sometimes missing a chat message altogether. Participants also reported this in the questionnaire and interview responses.
8. (Monitoring) Even through tools exist to monitor others' activities, our participants reported that generally they did not use them, or felt they were ineffective. They felt that once two or more team members were present in the space, the color-coded radar boxes and pointers were not useful. For the one color-blind participant in our study, this was always true, even for small meetings.
9. (Planning) Participants reported that the functionality for planning group-work was inadequate. For example, items on a ToDo List could not be reordered and items on an Agenda could not be edited. We observed some initial use and then no use of these tools, as participants started using more general tools to accomplish their planning work.
10. (Protection) The participants noted that the lack of an 'undo' function for modifying or deleting tools and objects would make it difficult to protect their work. Perhaps as a result, more than half of the artifacts in the space were created by one person, the project manager. The manager would also have liked to be able to make certain artifacts non-editable by others.

To further summarize the findings and relate them to the tool's suitability for the welding scenario, we assigned a rating for each mechanic of collaboration (see Table 2).

Mechanic	**Rating**
Explicit communication • Many successes, but some awkwardness	+1
Implicit communication • Change detection emphatically needed, but manual work-around succeeded	-1
Coordination of action • No reported problems and we did see some instances of coordination, but not enough evidence to evaluate confidently.	N/A
Planning • Many complaints about some tools, however, the team found the needed functionality to be effective.	-1
Monitoring • Even through the tools exist to monitor, our group didn't always use them effectively or use some of them at all.	-1
Assistance • We saw two instances of assistance being requested and given, but not enough evidence to evaluate confidently.	N/A
Protection • No specific protection capabilities, but the group was able to use social protocols and not delete or modify other users' artifacts. Lack of 'undo' function was considered a notable protection problem.	0

Table 2. Ratings and rationales from user-based study.

Overall assessment of the tool

Overall, Teamwave Workplace was a suitable groupware tool for the welding work scenario despite its identified usability issues. The welding researchers did accomplish their work goals using the tool. Following the study, participants expressed their overall satisfaction with the groupware system in helping them collaborate with remote partners. Participants felt that they collaborated more effectively and efficiently than with their previous ad hoc collaboration tools.

STUDY 2: THE INSPECTION EVALUATION

In addition to the user-based evaluation, we carried out a separate inspection study of the groupware system. The evaluators in the inspection were different from those in the user-based study, and were unaware of the results from that study. The goal of the inspection was to look for usability strengths and problems, and to determine whether Teamwave Workplace would be a good tool for supporting collaboration in the welding scenario.

Inspection Study Methodology

The inspection was a structured assessment of the collaborative tool by a set of evaluators, using the mechanics of collaboration as evaluation criteria. Inspection is a widely-accepted, discount evaluation method used for diagnosing usability problems in user interfaces, e.g., [5]. In this technique, several evaluators examine an interface and judge its compliance with recognized usability principles. Problematic aspects of the interface are identified, as well as, their potential severity.

In our inspection, four evaluators familiar with the welding scenario examined TW. They were not experts in the domain and were not intimately familiar with the ways that the welding group carried out its work. Our intent was to do a broad inspection that would look at all areas of the system. The study contained the following activities:

- an orientation to the system, the welding scenario, and the data-collection tools
- a one hour exploration of the tool, mixing synchronous and asynchronous interaction, and using both voice and chat-based communication
- a one hour synchronous task where the evaluators were asked to make a group decision based on several types of data present in two rooms of the system
- the initial inspection itself
- an asynchronous task carried out for three days after the initial inspection, where evaluators were asked to collaboratively write a document
- a final revisitation and revision of the inspection

Data was collected through an inspection booklet that listed each mechanic as an evaluation heuristic and provided the evaluators with space to list problems and comments. After all of the inspections were completed, the evaluators worked together to compare their assessments and determine the main strengths and weaknesses of the system.

Main Findings from the Inspection

The inspection identified several potential usability problems in the system. These problems were primarily in the areas of explicit communication, coordination, and planning. In order to compare our results with those of the user-based study, we organized the inspection results into a list of the ten most obvious issues found.

1. (Explicit communication) It is difficult to notice new messages in the chat tool. Users' attention was rarely on the chat tool unless they were already engaged in a conversation. The problems were most severe when people wanted to start a conversation.
2. (Explicit communication) It is difficult to determine whether the intended audience is "listening." It was hard to tell whether anyone had read a chat message, and the 'bell' feature was considered too heavyweight for attracting people's attention.
3. (Explicit communication) There is no way to communicate with people outside your current room. It is impossible to say something to everyone in the system. There is a separate chat session for every room, so it is difficult to continue conversations across rooms.
4. (Planning) Group organization and project planning is clumsy for asynchronous work. There were few tools available to help assist the organization of project activities (e.g., what tasks were assigned to whom, or what progress was being made on assigned tasks).
5. (Coordination) It is difficult to determine what others are doing. In several cases, multiple people would initiate an activity without knowing that another person was doing the same thing. For example, several people would respond to a chat message, or would attempt to change an option in a tool.
6. (Coordination, Monitoring) It is difficult to determine who is using a tool. This led to confusion and coordination errors. For example, people would type over top one another in a Post-it note or overwrite fields in the ToDo list.
7. (Monitoring) It is hard to notice when someone enters or leaves a room; in several cases, people would continue a conversation even though the other person had left the room.
8. (Monitoring) It is difficult to determine what has happened previously in a room. Unless explicit, written notes were left, it was impossible to ascertain who had visited a room, what the current state of an activity was, and what had changed since the last visit.
9. (Monitoring) It is difficult to determine the identity of participants. Participants are represented by color, but with several people in a room, it was difficult to remember how colors mapped to people. There is a participant list in the system, but it was considered too effortful to use.
10. (Protection) There is no way to claim ownership of a tool and prevent others from changing the data. Tools can be moved while a person is typing, and it is possible for several people to type into a tool at the same time.

To summarize the inspection study's findings and relate them to the tool's suitability for the welding scenario, we assigned a rating for each mechanic of collaboration. The ratings and rationales are shown below in Table 3.

Mechanic	Rating
Explicit communication • Communication through the chat tool was very difficult and caused several problems • Reasonable support allowing integration of workspace objects with verbal conversations	-1
Implicit communication • Basic representations of activity exist, but there are several tools and areas that show no activity information	-1
Coordination of action • Some tools show feedback that assists coordination, but several do not	0
Planning • Support for project planning is rudimentary	-1
Monitoring • Reasonable tools for tracking others within a room, but poor representation of people outside the current room	0
Assistance • Difficult to determine when people needed assistance	-1
Protection • No facilities for locking objects or tools; no means for stating intentions to work in a particular area	-1

Table 3. Ratings and rationales from the inspection study.

Overall assessment of the tool

Although summative assessments of entire systems are not part of the traditional inspection methodology, we asked the evaluators to state whether they felt that TW would be a successful tool in the welding scenario. Based on the results of the inspections, all of the evaluators felt that TW had too many usability problems to support the welding group's collaboration adequately. Their main reasons were the difficulties in explicit communication and the lack of support for longer-term collaborations. This conclusion is different from that found in the user-based study, and the reasons for these differences are discussed below.

COMPARISON OF TECHNIQUES

It is clear that user-based methods and inspection methods are very different, and the comparison that we undertook did not attempt to treat the two techniques as somehow interchangeable. Rather, they are complementary techniques with some areas of overlap, and can both be used to identify usability problems. The sections below attempt to determine differences and commonalities that will help practitioners decide when and where to use them.

As expected, the user-based study required far more time and effort, and the data gathered was considerably more difficult to analyze because of its volume and detail. The inspection study was relatively simple to design and conduct, and the synthesis session, with all four evaluators, was completed in an afternoon. The main methodological difference, however, between the user-based study and the inspection study was that the former gathered data from real users working in a realistic situation, and the latter used evaluators who were experts in usability and groupware, but not in the application domain. Given this difference, we expected to see at least some variation in the results obtained by the two evaluations.

We used three means of comparing the results of the two evaluations. First, and most importantly, we compared the "top ten" lists of usability issues. Second, we looked for large discrepancies in the ratings for the mechanics of collaboration. Third, we considered the overall assessment for the tool in the welding work scenario.

Differences in Usability Problems

In comparing the "top ten" usability lists produced by each evaluation, we find that just over half of the issues in one list have an equivalent issue in the other. In particular, both evaluations found that (1) chat-based communication was difficult to initiate and notice, (2) that it was difficult to keep track of who was currently in a room, (3) that the lack of project planning tools made asynchronous work more difficult, (4) that it was difficult to determine what had previously happened in a room, and (5) that there was some missing protection functionality. Conversely, both studies found usability issues that the other study did not, or at least that did not make their "top ten" lists.

From a practitioner's perspective, this is an exciting result. While inspection methods are not a substitute for field studies, it means that many pertinent usability problems can be found without having to resort to the difficulty, time, and cost of a field study. The cost-saving makes it feasible for earlier and more frequent evaluation of groupware systems. This research is a validation of the idea that discount methods can be used effectively to find user-relevant, usability problems in groupware.

The key to comparing the usability issues found by these two evaluations was in how we categorized them. The mechanics of collaboration were effective for this purpose; however, our results suggest that other criteria could be added. In particular, the mechanics are oriented more towards synchronous work than to asynchronous collaboration, and this led to some of the differences between the two studies. We are currently determining additional mechanics that are specifically oriented to asynchronous work: for example, two activities seen in the welding scenario are *document transfer* and *change notification*. Additionally, while we recognize that usability evaluation typically considers a set of issue areas larger than that contained in the mechanics of collaboration, e.g., social and organizational aspects, having a set of organizing principles that span contexts is useful.

Differences in Ratings

Since the ratings were intended as a general summary in each of the seven categories, we were only looking for large discrepancies between the two techniques. The assigned ratings from the two techniques we studied were relatively close, often with only a one-point difference for the each categories, and so we draw only a few conclusions from this comparison.

First, the similar ratings stand as another indication that the methods can find similar problems. It is interesting to note that where the two methods differ, the user-based study generally rates TW higher than the inspection, a trend that was more fully realized in the overall assessments of the system (see below). Second, it is notable that the user-based study was unable to comment on two of the mechanics because of insufficient data. This highlights the fact that a user-based study can only evaluate areas that intersect with the users' work, whereas an inspection is able to cover any areas for which it has assessment criteria. Therefore, when choosing between the two methods, the goal of the evaluation must be well understood, e.g., whether a broad array of identified usability issues is desired or a more situated study.

Differences in Final Assessment of TW

We considered two competing hypotheses when comparing the overall assessments produced by the different techniques. On one hand, we thought that if the groupware system were a good (or a bad) tool, its qualities should be apparent no matter what evaluation technique was used. Therefore, the two evaluations should come to roughly the same overall conclusion. On the other hand, we expected that the realistic setting of the user-based study would reveal things specific to the culture and context of work that could affect the summative evaluation.

In the end, the latter was true—it appears that a list of usability problems is not a good predictor of a system's success in a real work setting. This, of course, is what some CSCW researchers have been saying all along, e.g., [13]. Nevertheless, it is useful to take a closer look at why TW was acceptable overall, even though it had several usability problems. There are two main issues related to the work context. First, work practices of the group allowed them to avoid the problematic areas of the tool. Secondly, the tool was an improvement over the system that the group had used previously.

Structure of work practices. Several of the problems found by the inspection method did not affect the welding group greatly because of the two primary ways that they carried out their collaboration: leaving documents and holding meetings. The basic collaboration requirements of these two activities were in fact supported adequately by TW. Meetings were relatively formal, were held in a pre-appointed room of the system, with certain people taking responsibilities for actions such as starting room tools, editing text, and so on. With this more formal approach and structured division of responsibility, there was much less chance of encountering the major problems identified by the inspection. It is unlikely the group intentionally structured their collaboration in this way to avoid problems in the tool, but it is clear that they found a way to get their work done successfully.

The tool was an improvement. The welding group had been sharing documents before the introduction of Teamwave Workplace, primarily by leaving documents in a central FTP site and by sending email messages. This set-up caused the collaborators two problems: The FTP site required an administrator to move files from incoming areas to outgoing directories, and it was difficult to refer to images and documents in email messages. The Teamwave Workplace system solved these two problems, and so the welding group may have been happy with the tool simply because it was an improvement.

The fact that the two techniques came to different conclusions is perhaps not surprising, but highlights the fundamental difference between user-based methods and inspection. The divergence was clearly caused by the user-based study's ability to see the context as well as the tool, and by the inspection's underlying orientation towards finding problems over a wide range of use.

CONCLUSIONS

There are several conclusions that we can draw from the comparison of usage methods and inspection techniques.

1. The details of the work situation are needed in order to adequately assess a groupware tool for a particular scenario. Naturalistic and user-based techniques are currently the only methods that can provide that information, e.g., [3]. However, as discussed below, it may be possible to contextualize inspection methods and make them more sensitive to various work scenarios.
2. Discount usability techniques, such as inspections, do have a valid role in groupware evaluation. Our dual study shows that inspections can find many of the same problems that are found by real users in real work situations, even when the inspection takes place far from the actual work setting. In addition, these problems can be found for a fraction of the time and effort required by a user-based study.
3. The two types of techniques will work well in combination: inspections to find early and major usability problems, and user-based techniques to find more subtle, contextual issues. The low cost of inspections means that several evaluations can be carried out before undertaking an expensive user-based study.
4. The generality of inspection techniques can be an advantage when developing certain types of groupware. Groupware developers face a difficult problem when they have a general-purpose group support tool (such as TW) that can be used in a wide variety of actual work situations, for a wide variety of work practices. With this type of system, it could take several situated evaluations before the developers would notice a majority of the main usability problems. This happens because there are contextual and cultural factors that may hide the presence of problems in different areas of the system.
5. It may be possible to contextualize inspection techniques in order to make them more sensitive to work and organizational situations. If our inspectors had better understood the kinds of collaborative activities that were common in the welding group, the inspection could have produced an evaluation that was better tailored to the welding scenario. Certain techniques from singleware usability, such as task-centered walkthroughs, manage to add context to discount methods, and these may be applicable to groupware evaluation as well. Additionally, this contextualization would help prioritize the many issues returned by discount methods.
6. Research is needed to find additional assessment criteria for discount groupware evaluation. One direction is in extending the mechanics of collaboration to capture aspects of asynchronous work. However, other principles that are common across contexts may also be found or distilled from existing field studies.

These findings sound much like what is already common practice in singleware evaluation. However, it is novel that the discount approach can be applied to groupware situations, which have often been considered too complex for low-cost techniques. We do not suggest that inspections are all that is required for adequate evaluation of multi-user systems; only that there are advantages to both types of techniques, and both should be considered as part of the practitioner's toolbox. User-based methods find situated usability problems, help prioritize issues, and can accurately assess system success in particular work contexts. Inspection techniques are far cheaper and less time consuming, are able to find a wide variety of usability problems, but do not traditionally prioritize the issues. With more inspection criteria like the mechanics of collaboration, and with ways to contextualize inspection studies, a variety of new groupware evaluation techniques seem possible.

REFERENCES

1. Baker, K., Greenberg, S. and Gutwin, C. (2001) Heuristic Evaluation of Groupware Based on the Mechanics of Collaboration. Proceedings of the 8th IFIP Working Conference on Engineering for Human-Computer Interaction (EHCI'01), May 11-13, Toronto, Canada.
2. Cugini, J., Damianos, L., Hirschman, L., Kozierok, R., Kurtz, J., Laskowski, S., and Scholtz, J. Methodology for Evaluation of Collaboration Systems. Available at http://zing.ncsl.nist.gov/nist-icv/documents/method.pdf.
3. Dix, A., Finlay, J., Abowd, G., and Beale, R. *Human Computer Interaction, 2nd Ed.*, Prentice Hall Europe, 1998, Chapter 11.
4. Doubleday, A., Ryan, M., Springett, M., and Sutcliffe, A. A Comparison of Usability Techniques for Evaluating Design. *Proceedings of Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (The Netherlands, August 1997), ACM Press, 101 – 110.
5. Dumas, J.S. and Redish, J.C. *A Practical Guide to Usability Testing*. Ablex, 1993.
6. Greenberg, S., Fitzpatrick, G., Gutwin, C. and Kaplan, S. (2000). Adapting the Locales Framework for Heuristic Evaluation of Groupware. *Australian Journal of Information Systems* (AJIS) 7(2), 102-108, May.
7. Greenberg S. and Roseman, M. (Forthcoming). Using a Room Metaphor to Ease Transitions in Groupware. In M. Ackerman, V. Pipek, V. Wulf (Eds) Beyond Knowledge Management: Sharing Expertise, MIT Press.
8. Grudin, J. Groupware and Cooperative Work: Problems and Prospects. *The Art of Human-Computer Interface Design*, B. Laurel ed., 1990, Addison-Wesley, 171-185.
9. Gutwin, C. and Greenberg, S. The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces. *Proceedings of WETICE 2000, Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (Gaithersburg, MD, June 2000), IEEE Computer Society, 98-103.
10. Jeffries, R., Miller, J., Wharton, C., and Uyeda, K. User Interface Evaluation in the Real World: A Comparison of Four Techniques. *Proceedings of Human Factors and Computing Systems* (New Orleans, LA, April 1991), ACM Press, 119-124.
11. Morse, E. and Steves, M. CollabLogger: A Tool for Visualizing Groups at Work. *Proceedings of WETICE 2000, Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (Gaithersburg, MD, June 2000), IEEE Computer Society, 104-109.
12. Nielsen, J. *Usability Engineering*, Academic Press, Boston, MA, 1993.
13. Nielsen, J. and Phillips, V. Estimating the Relative Usability of Two Interfaces: Heuristic, Formal, and Empirical Methods Compared. *Proceedings of Human Factors in Computing Systems* (Amsterdam The Netherlands, April 1993), ACM Press, 214-221.
14. Pinelle, D. and Gutwin, C. A Review of Groupware Evaluations. *Proceedings of WETICE 2000, Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (Gaithersburg, MD, June 2000), IEEE Computer Society, 86-91.
15. Rippey, W. and Falco, J. The NIST Automated Arc Welding Testbed. *Proceedings of the Seventh International Conference on Computer Technology in Welding* (San Francisco, CA, July 1997), 203-210.
16. Ross, S., Ramage, M., and Rogers, Y. PETRA: Participatory Evaluation Through Redesign and Analysis. *Interacting with Computers 7,* 4 (1995) 335-360.
17. Smith, R., Hixon, R., and Horan, B. Supporting Flexible Roles in a Shared Space. *Proceedings of CSCW'98* (Seattle, WA, November 1998), ACM Press, 197-206.
18. Steves, M. and Knutilla, A. Collaboration Technologies for Global Manufacturing. *Proceedings of the ASME International Mechanical Engineering Congress and Exposition (IMECE): Symposium on Manufacturing Logistics in a Global Economy* (Nashville, TN, November 1999), 541-555.
19. Wixon, D. and Wilson, C. The Usability Engineering Framework for Product Design and Evaluation. *Handbook of Human-Computer Interaction, 2nd Ed.*, Elsevier Science B. V., Amsterdam, The Netherlands, 1997, 653-688.