

A Survey of Groupware Evaluations in CSCW Proceedings

David Pinelle

Department of Computer Science
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
+1 306 966 6593
mudduck@home.com

ABSTRACT

Groupware software is still not widely used in spite of continued hardware advances in processing power and network speeds. This indicates that developers may not be matching the software they design with the needs of the target user groups. For this reason, a further understanding of groupware evaluation should be sought to help provide a means for identifying the key issues developers must address for successful implementations.

This paper presents the findings of a survey of groupware evaluations in *CSCW Proceedings '90, '92, '94, '96, and '98*. Articles were selected if they included evaluations of groupware applications or included thorough discussions of groupware applications. In order for an application to be classified as groupware, it had to allow multiple users to work collaboratively toward a common goal [23].

A total of 45 articles were included in the study. In order to collect a broad range of information, a set of categories and criteria were developed to serve as a collection framework for the data in the articles. Data in the evaluations found in the articles was then gathered and analyzed.

Of the 32 articles that included evaluations, only 41% evaluated complete implementations of groupware applications in real world settings, and even fewer of these evaluated the social, organizational, or cultural impact the software had on the work settings. Additionally, a high proportion of articles introduced new groupware applications but failed to discuss evaluation or included incomplete information. This paper discusses these findings and makes recommendations for improvements in current research methods and proposes new directions for research on groupware evaluation.

Keywords

groupware, asynchronous groupware, synchronous groupware, CSCW

INTRODUCTION

Developers of collaborative applications must deal with increased levels of complexity during software design. In order to succeed, the developer must address the difficult issues of group dynamics and the idiosyncracies of social and political organization in the workplace.

Whether the developer opts for iterative design or a more traditional software engineering sequence, the issue of evaluation still must be contended with. However, there is no clear consensus on how to carry out an evaluation of groupware software. A range of techniques have been employed by researchers and developers including scientific, engineering, and social science methodologies.

This lack of insight into groupware evaluations provided the impetus for this study. A survey of CSCW proceedings was undertaken in an attempt to enlighten the questions of what, when, where, and how to evaluate.

Grudin [14] has discussed evaluation of groupware at length and has pointed out many of the difficulties inherent in this process. He emphasizes context as a complicating factor of cooperative work and states that differences in personalities and politics in an organization can lead to the rejection of a good piece of software or the acceptance of a bad piece of software. Likewise, Orlikowski [39] conducted a study on groupware implementation and found that the culture and organizational structure in a workplace affects the way the software is utilized by the users. Thus, the group of end users in the context of their workplace must be considered, and laboratory experiments may not necessarily be a good indicator of how a piece of software will be accepted.

Grudin [14] also points out that group interactions which come about with the introduction of groupware unfold over days and weeks, indicating that groupware evaluation takes longer than does the evaluation of a single user application. This brings into question the value of one-shot summative evaluations in evaluating groupware applications.

Developing a system for classifying groupware

In order to carry out this survey, a preliminary classification scheme was devised to organize groupware evaluation data. This proved to be a more difficult task than initially expected. The complexities of groupware and user groups have led to evaluations that cross disciplinary boundaries and that employ methodologies which are not relevant in other areas of software design. Therefore, as the initial analysis of the articles commenced, the preliminary classification scheme was

found to be inadequate at reasonably representing the range of evaluation types, techniques, and findings.

After articles were selected for this study, they were analyzed a total of three times. Much of the work accomplished during the first two analyses involved revising the categories and classifications in order to guarantee that they were truly representational of the information found in the articles. The purpose of the final analysis was to verify the data that had been collected from each article and to ensure that the data had been properly classified.

Since this survey was intended to be exploratory in nature, the categories that were used in data collection were intentionally broad. Each category contained classifications that allowed specification of certain characteristics about the software or the evaluation. If the information was present in the article, it was collected for each of the following categories:

- Classification of the evaluation
- Characteristics of the software
- Characteristics of the evaluation
- Conclusions drawn from the evaluation
- Techniques for data collection
- Placement of evaluation in the software lifecycle

Classification of the evaluation

This study includes articles from *CSCW Proceedings '90, '92, '94, '96*, and '98 with substantial discussion of new groupware applications and/or evaluation of groupware applications. Articles without evaluations were included to help determine the prevalence of formal evaluation in groupware research.

For the purposes of this study, evaluation implies that the researchers attempted to gather information about the software by measuring its use by a group of people, whether they were end users or paid participants in a lab study. The only stipulation on the user group was that they not be directly involved in carrying out the research. Therefore, studies that discussed evaluation based on the experiences of the authors or of members of their research group were not classified as formal evaluations.

McGrath's [29] classification scheme for research strategies was used as a starting point for developing classifications for formal evaluations. However, many of the strategies are inappropriate for groupware evaluations and do not appear in the CSCW proceedings. For this reason, only three of the strategies were adopted: laboratory experiment, field experiment, and field study. As the literature containing groupware evaluations was explored, it became obvious that these were not sufficient for describing all types of evaluations in this domain.

The major differentiating characteristics of the three evaluation categories selected from McGrath were

evaluation setting and formality of manipulation technique (see figure 1). Thus, the lab experiment takes place in a controlled setting and with rigorous manipulation of the evaluation process. The field experiment and field study take place in a naturalistic setting, but the field experiment has rigorous manipulation and the field study has minimal to no manipulation. This left one combination of locale and formality of manipulation unexplored. This combination, controlled setting with minimal to no manipulation, was added to this classification scheme and was called exploratory since it implies that the researcher is gathering general information from the study and is not proceeding with a fixed methodology for controlling the experiment.

Figure 1: Evaluation classifications

		Manipulation	
		Rigorous	Minimal / None
Setting	Naturalistic	Field Experiment	Field Study, Case Study
	Controlled	Laboratory Experiment	Exploratory

Two final classifications were added to accommodate the preliminary findings. First, most field studies entailed involvement by the authors in the development and implementation of the software. However, a portion of these involved an "after the fact" evaluation in which the authors were not involved in development and implementation but instead evaluated a piece of groupware software that had been implemented in a work setting by others. These studies tended to involve a different set of data collection methodologies, and therefore were classified separately from field studies as case studies.

Finally, articles that did not contain formal evaluations were separated into two classifications. Some of these articles drew a set of conclusions based on the usability of the software. These articles were categorized as introspective studies, since presumably they drew on the experiences of the authors. The articles without formal evaluations that did not draw conclusions about the software were placed in the no evaluation category.

Characteristics of the software

A basic classification scheme was adopted for the groupware software itself. First, the collaboration's distribution in time had to be considered, so the software was classified as synchronous or asynchronous. Next, the software was classified based on whether it was a research system or a real world software implementation. Finally, information was collected on whether the authors developed the software, someone else implemented the software, or the software was an off the shelf product. Of

these three dimensions, implementation is the only one that is mutually exclusive.

- Time distribution:
 - Synchronous
 - Asynchronous
- Implementation:
 - Academic/Research
 - Real world setting
- Development:
 - Built by the authors
 - Off the shelf
 - Implemented by others

Characteristics of the evaluation

Twidale, Randall, and Bentley [49] identified four dimensions for classifying evaluations. These are:

- Summative vs. Formative
- Quantitative vs. Qualitative
- Controlled experiment vs. Ethnographic observation
- Formal and rigorous vs. Informal and opportunistic

This classification was initially adopted as is, and an attempt was made to classify evaluations along these lines, but this proved problematic. The controlled experiment vs. ethnographic observations criterion was found to be too restrictive and inappropriate for many studies. In order to make this relevant to all groupware evaluations, this scale was revised to specify experimental setting, with the new criterion being: controlled setting vs. naturalistic setting. This specifies whether the researcher is in control of the experimental setting or whether the evaluation is conducted in a real world setting.

The formal and rigorous vs. informal and opportunistic classification proved to be too vague and was revised as well. McGrath [29] differentiates between techniques for manipulating the experimental procedure and techniques for measuring the experimental results. This distinction was combined with the classification scheme, generating the following classifications:

- Manipulation:
 - Formal / rigorous
 - Minimal manipulation
 - No manipulation
- Measurement:
 - Formal / rigorous
 - Informal

This revised scheme was found to be more precise but still general enough that it could be applied to all articles where adequate details were provided.

Finally, the distribution of users during the evaluation was classified. Evaluations were classified as real distributed when the users were isolated from each other and unable to communicate except through the groupware

application. They were classified as simulated distributed when the researchers attempted to isolate the users but did not do this in an absolute fashion. For example, if a partition was placed between the users, they would still be aware of the others' presence--isolation would not be complete. Finally, evaluations were classified as co-present if the users were present in the same room without any type of isolation.

Conclusions drawn from the evaluation

Initially, one of the goals for this survey was to attempt to discover what the authors were trying to measure in the groupware evaluations. However, it quickly became apparent that this was an unrealistic goal. Many of the studies reported data collection in vague terms and made no mention of, for example, the content of a questionnaire or the purpose of an interview. Therefore, a revised methodology was adopted. Instead, the conclusions drawn from the evaluations were analyzed in an effort to indirectly discover the general classifications of information they were trying to obtain.

The initial set of classifications were motivated by the three categories of benefits realized through using CSCW software cited by Baeza-Yates and Pino [2]. These classifications were then added to and modified to include concepts from a multi-stage evaluation process discussed by Beuscart-Zephir, Molenda, Grave, and Dufresne [4]. Finally, they were modified again during an analysis of the articles included in this survey to guarantee that they were representational of the conclusions drawn from the evaluations.

- Organizational impact / impact on work practices
- End product produced through using the software
- Efficiency of performing a specific task while using the software
- User satisfaction with the software
- Level of support provided by the software for a specific task
- Specific features of the groupware interface
- Patterns of system use
- User interaction while using the software

Techniques for data collection

A classification scheme was developed to record techniques used for data collection during groupware evaluations. The articles included in this survey were analyzed to identify the techniques that had been used, and these were compiled into a final set of classifications, most of which are self-explanatory:

- User Observation
- Interview
- Discussion
- Questionnaire

- Qualitative work measures
- Quantitative work measures
- Collection of archival material (company newsletters, bulletins, email, etc.)

Quantitative work measures were time based and frequently focused on measuring users' efficiency at performing a given task. Thus, an effort was made to record quantitative, objective data about the users' work.

Qualitative work measures attempted to assess some qualitative aspect of the users' work. These usually involved a judge who critiqued the users' task performance or the end product that the users produced.

Placement of the evaluation in the software lifecycle

Grudin [14] stresses the importance of evaluation over a period of time following groupware implementation. He also argues that evaluations of partial prototypes in laboratory settings are not able to address complex social and organizational issues. These arguments motivated the collection of data on the placement of the evaluation in the software's lifecycle.

- Periodic evaluations throughout development process
- Continuous evaluation throughout development process
- Evaluation of a prototype
- Evaluation of a finished piece of software
- Periodic evaluations after software implementation
- Continuous evaluation after software implementation

METHOD

A survey was undertaken of CSCW literature in an effort to identify trends in groupware evaluation. CSCW proceedings from the years '90, '92, '94, '96, and '98 were reviewed and relevant articles were identified. Articles were included in this survey if they met one or more of the following criteria:

- The article introduced a new groupware application.
- The article introduced a groupware toolkit and included substantial discussion of a groupware application created with the toolkit.
- The article contained an evaluation of an existing groupware application that had been implemented in a work setting.
- The article contained a general evaluation of a publicly available groupware application.

The first two criteria allow inclusion of articles that do not include groupware evaluations. The study was designed to allow this in order to discover the prevalence of evaluation in articles that include substantial discussions of new groupware applications.

In order for an article to be included in this survey, the application had to meet the groupware definition

established by Johnson-Lenz and Johnson-Lenz [23]. This definition indicates that the software supports multiple users in working collaboratively toward a common goal.

The proceedings were analyzed to identify articles containing applications which met the groupware definition and which met one of more of the criteria for inclusion above. Initially 55 articles were selected, but this was eventually narrowed to 45 as the selection guidelines were enforced more rigorously.

There were difficulties in deciding which articles met the criteria for acceptance in this study. Many articles clearly qualified, while others only marginally qualified. Some articles were excluded since they did not clearly meet the above definition of groupware. There is room for debate about which articles should have been included or excluded. It is felt that the sample size of articles was large enough to offset any biases that may have been introduced during the selection of borderline articles.

Data collection commenced, and the articles were analyzed to gather and sort relevant information into the initial set of categories and classifications. However, as this process progressed, it became obvious that there was additional relevant information available in the articles. This led to revision of the categories and classifications to better reflect the content. In all, they were revised approximately ten times before it was felt that they could adequately represent the evaluation data found in the articles. This led to three complete analyses of the articles to guarantee that the information included in each article was represented properly.

RESULTS

Many of the classifications used in the survey are not mutually exclusive. For example, a piece of software may have both synchronous and asynchronous components. Additionally, an article may contain formative evaluation information followed by a summative evaluation of the final implementation. Or, an evaluation may focus on collecting both qualitative and quantitative data. For this reason, many of the numbers presented in the categories below represent totals higher than the number of articles in the study.

Three evaluations contained enough information to allow classification in the majority of categories but not in all categories [21, 31, 50]. Information regarding these articles is included in the results below except where noted otherwise.

Characterizing the software

When limiting discussion to the 32 articles where evaluations actually took place, it was found that the majority of these covered synchronous applications. In all, 27 applications allowed synchronous collaboration, and 11 allowed asynchronous collaboration (see table 1).

Most of the systems were academic or research implementation, and a smaller number of implementations were installed with permanence into a real world setting. In the articles including evaluations, 19 of the systems were academic/research and 13 were real world implementations. The percentage of academic/research systems increases when all articles included in the survey (including those without evaluations) are considered. The ratio of academic/research systems to real world implementations becomes 30:15.

Table 1: Characterizing the software

	#	%	References
Distribution in time			
Synchronous	26	70	[3, 5, 7, 8, 10, 11, 15, 17, 18, 19, 21, 22, 25, 28, 30, 31, 35, 37, 38, 42, 43, 44, 47, 48, 49, 50]
Asynchronous	11	30	[1, 7, 10, 33, 38, 39, 41, 42, 44, 45, 51]
Implementation type			
Academic / Research	19	59	[8, 11, 15, 17, 18, 21, 22, 25, 28, 31, 33, 35, 37, 42, 43, 44, 48, 49, 50]
Real world	13	41	[1, 3, 5, 7, 10, 19, 30, 38, 39, 41, 45, 47, 51].

indicates total studies fitting a classification (not necessarily exclusive), % indicates percentage of studies in a given classification. The percentages are calculated separately for each of the two categories and the total studies in each category is used in the calculation.

Evaluation type

Evaluation type, as discussed earlier, is generally a combination of setting and formality of manipulation techniques. Therefore, the field techniques here do not necessarily indicate a “real world” implementation as discussed in the previous section--they indicate an evaluation in a naturalistic setting, which can potentially be carried out with an academic/research piece of software which is not intended to be permanently installed in the setting.

When field based studies (field study, field experiment, and case study) are compared to lab experiments based in controlled settings, the ratio is even (13:13, see table 2).

From the original 45 articles, a total of 13 articles did not contain formal evaluations. Of these, 9 were placed in the no evaluation category, and 4 were categorized as introspective studies since the authors drew conclusions about the software without detailing formal evaluations. Finally, 3 articles did include evaluations, but the sketchiness of the information presented regarding the

evaluation made it impossible to determine how the articles should be classified.

There are a total of 16 articles when combining the “introspection”, “not enough information to classify”, and “no evaluation” categories. All of these articles present no or incomplete information on evaluations. This forms a substantial portion of the articles included in this study (35%).

Finally, a single article [42] defied classification with the current scheme. This is not due to a lack of information provided by the authors but is due to the unique circumstances under which the evaluation was carried out.

Table 2: Studies by evaluation type

Evaluation Type	#	%	References
Laboratory Experiment	13	28	[11, 15, 17, 18, 21, 25, 28, 33, 35, 37, 43, 48,49]
Field Study	8	17	[1, 3, 19, 30, 38, 41, 45, 47]
Case Study	4	9	[5, 7, 39, 51]
Exploratory	3	7	[8, 44, 49]
Field Experiment	1	2	[10]
Does not fit this scheme	1	2	[42]
Not enough information	3	7	[22, 31, 50]
Introspection	4	9	[16, 24, 34, 36]
No Evaluation	9	20	[6, 9, 13, 20, 26, 27, 32, 40, 46]

indicates total studies in a classification, % indicates percentage of studies in a classification. 45 separate articles are included, but one article [49] is included twice. Thus, percentages are calculated out of a total of 46 studies.

Characterizing the evaluation

Formative evaluations were more prevalent than summative evaluations (see table 4). The high number of prototype systems accounts for this in part, as many of the authors were conducting the research to further enlighten system development. In all, 15 evaluations were of prototype systems and 11 were of completed software packages (see table 3).

Only a small number of articles discussed evaluations that were distributed throughout the software lifecycle. There were 2 articles that discussed continuous evaluations throughout the development process, and 3 articles

discussed continuous evaluations after the software had already been implemented. Similar to this, only 3 articles detailed periodic evaluations that were conducted throughout the development process.

Table 3: Placement of evaluation in lifecycle

	#	%	References
Periodic throughout development	3	9	[19, 31, 42]
Continuous throughout development	2	6	[41, 45]
Evaluation of a prototype	15	47	[3, 8, 10, 15, 17, 21, 22, 25, 33, 35, 43, 44, 47, 48, 49]
Evaluation of finished software package	11	34	[1,5,11,18,28,30,31,37,38, 50,51]
Periodic after implementation	0	0	[]
Continuous after implementation	3	9	[7, 39, 45]

indicates total studies in a classification, % indicates percentage of studies in a classification. 32 articles containing evaluations are included, but two articles [31, 45] are classified twice. Thus, classifications are not mutually exclusive and the total of the percentages is greater than 100.

Most of the evaluations were focused on qualitative aspects of the software, with quantitative studies constituting only a small portion of the articles (23 qualitative, 2 quantitative, 7 both).

The level of manipulation used in the evaluations was distributed evenly between formal manipulation and no manipulation. In all, 14 studies employed formal techniques, 2 studies employed minimal manipulation, and 14 studies employed no manipulation. Three studies were not classified due to lack of adequate information [22, 31, 50].

Of the metrics used to collect data in the studies, formal and rigorous measures were the most common (18 formal / rigorous, 12 informal). Two articles did not provide adequate information for classification [22, 50].

Similarly, evaluation locality reflects an uneven distribution with the majority of the evaluations classified as real distributed (19, with 9 co-present and 5 simulated distributed).

Table 4: Characterizing the evaluations

	#	%	References
Formative vs. Summative			
Formative	18	56	[3, 8, 10, 15, 17, 19, 22, 25, 28, 33, 35, 41, 42, 43, 44, 47, 48, 49]

Summative	12	38	[1, 5, 7, 11, 18, 21, 30, 37, 38, 39, 50, 51]
Both	2	6	[31, 45]
Quantitative vs. Qualitative			
Quantitative	2	6	[18, 21]
Qualitative	23	72	[3, 5, 7, 8, 10, 11, 15, 17, 19, 22, 30, 31, 35, 37, 38, 39, 41, 42, 43, 44, 45, 49, 50]
Both	7	22	[1, 25, 28, 33, 47, 48, 51]
Manipulation			
Formal / Rigorous	14	47	[10, 11, 15, 17, 18, 21, 25, 28, 33, 35, 37, 43, 48, 49]
Minimal	2	7	[1, 44]
None	14	47	[3, 5, 7, 8, 19, 30, 38, 39, 41, 42, 45, 47, 49, 51]
Measures			
Formal / Rigorous	18	60	[1, 7, 10, 11, 15, 17, 18, 19, 21, 25, 28, 33, 35, 37, 41, 47, 48, 51]
Informal	12	40	[3, 5, 8, 30, 31, 38, 39, 42, 43, 44, 45, 49]
Distribution in space			
Real Distributed	19	58	[1, 3, 7, 8, 11, 17, 19, 21, 25, 33, 38, 39, 41, 42, 44, 45, 47, 50, 51]
Simulated Distributed	5	15	[15, 22, 35, 43, 48]
Co-present	9	27	[5,10,18,21,28,30,31,37, 49]

indicates total studies fitting a classification (not necessarily exclusive), % indicates percentage of studies in a given classification. The percentages are calculated separately for each of the five categories, and the total number of studies represented in each category is used in the calculation.

Conclusions drawn from the evaluations

Only a small number of studies examined the organizational impact and impact on work practices in a user group when a piece of groupware was introduced. This category obviously pertains to “real world” software implementations. However, with a total of 13 articles with real world implementations, only 8 of these evaluated the impact the software had on the user group itself and on its work patterns (see table 5).

Carrying out an evaluation of this type can be quite time consuming since new work patterns evolve over time. For that reason, the amount of time each researcher was in contact with the user group (regardless of whether this

was continuous or intermittent) and gathering this type of evaluation data was recorded. Although 2 of the 8 studies did not specify time [5, 30], the other 6 did. The following 6 times are given in months: 4.5, 5, 12, 24, 36, 36.

Table 5: Conclusions drawn from evaluations

Conclusion	#	%	References
Patterns of system use	16	50	[1, 11, 15, 19, 25, 28, 31, 37, 41, 42, 43, 45, 47, 48, 50, 51]
Support for specified task	15	47	[1, 3, 8, 11, 15, 17, 19, 28, 31, 33, 41, 43, 44, 45, 49]
User interaction through the system	14	44	[8, 11, 17, 19, 21, 25, 28, 35, 37, 38, 43, 47, 48, 51]
Specific interface features	12	38	[11, 15, 17, 22, 25, 28, 31, 33, 43, 48, 49, 50]
User Satisfaction	12	38	[1, 5, 7, 8, 10, 15, 19, 22, 35, 37, 45, 47]
Organizational / Work impact	8	25	[5, 7, 30, 38, 39, 41, 45, 47]
End product from software use	5	16	[25, 37, 42, 49, 51]
Efficiency of task performance	4	13	[18, 25, 33, 48]

indicates total studies in a classification, % indicates percentage of studies in a classification. The 32 articles with evaluations are included. These classifications are not mutually exclusive. Therefore, percentages total more than 100.

There was generally an even distribution for many of the categories in this section. Conclusions drawn from the evaluations most frequently dealt with: patterns of system use (16 articles), the ability of the system to support a specific task (15 articles), user interaction through the system (14 articles), user satisfaction (12 articles), and specific interface features (12 articles). Less frequently evaluations drew conclusions about the end product produced using the software (5 articles) and the efficiency of task performance using the software (4 articles).

Evaluation Techniques

Observation was by far the most frequently used evaluation technique with 24 studies utilizing this technique, and this was coupled with videotaping of the users in 10 of these cases (see table 6). Interviews were the next most frequently utilized technique, with 12 occurrences noted. This was followed by questionnaire (9 articles) and quantitative work measures (9).

Table 6: Summary of evaluation techniques

Technique	#	%	References
-----------	---	---	------------

Observation	24	83	[1, 3, 7, 8, 10, 11, 15, 17, 19, 21, 25, 28, 30, 35, 37, 39, 41, 42, 43, 44, 45, 47, 48, 49]
Observation with Videotape	10	34	[10, 11, 15, 17, 19, 21, 25, 28, 35, 47]
Interview	12	41	[1, 5, 7, 11, 15, 17, 19, 35, 38, 39, 41, 51]
Questionnaire	9	31	[1, 10, 15, 19, 25, 33, 37, 47, 48]
Quantitative work measures	9	31	[1, 18, 21, 25, 28, 33, 37, 48, 51]
Qualitative work measures	5	17	[18, 25, 33, 37, 48]
Collected Archival Materials	4	14	[5, 7, 39, 47]
Discussion	3	10	[8, 30, 45]

indicates total studies in a classification, % indicates percentage of studies that use a given technique. Of the 32 articles with evaluations, 29 are included ([22, 31, 50] are excluded due to insufficient information). These classifications are not mutually exclusive. Observation with Videotape is a sub-classification of Observation, so any article classified in the former is also classified in the latter.

Experimental methodologies

It was found that experimental studies, including laboratory experiments and field experiments, have a distinctly different distribution of evaluation techniques than do field based studies (see table 7). This reflected a shift toward formalism in manipulation and measurement techniques. Observation was again the most frequently used technique, occurring in 12 articles, but this time it was coupled with the use of videotaping much more commonly with this occurring in 8 articles.

After observation, the quantitative work measure was the next most frequently used technique, and this was utilized in 7 studies. These were time based measurements, and many of these measured the time required for completion of a given task.

In contrast to the overall totals, the questionnaire was used more frequently than the interview in these studies. Questionnaires were used in 6 experimental studies and were administered prior to the experiment in 4 studies and post-experiment in all 6 studies. After the questionnaire, qualitative work measures were the next most common, and these had 5 occurrences. In these studies, the researcher or some other expert graded the quality of the work produced by the users. The interview was utilized less frequently here and was only the fifth most utilized technique with 4 occurrences.

Table 7: Techniques used in Field and Lab Experiments

Technique	#	%	References
Observation	12	86	[10, 11, 15, 17, 21, 25, 28, 35, 37, 43, 48, 49]
-Videotape	8	57	[10, 11, 15, 17, 21, 25, 28, 35]
-Audio tape	2	14	[37, 48]
-Software logged	1	7	[28]
Quantitative work measures	7	50	[18, 21, 25, 28, 33, 37, 48]
Questionnaire	6	43	[10, 15, 25, 33, 37, 48]
-Questionnaire, pre-experiment	4	29	[10, 25, 37, 48]
-Questionnaire, post-experiment	6	43	[10, 15, 25, 33, 37, 48]
Qualitative work measures	5	36	[18, 25, 33, 37, 48]
Interview	4	29	[11, 15, 17, 35]
Discussion	0	0	[]
Collected Archival Material	0	0	[]

indicates total studies in a classification, % indicates percentage (out of 14 studies) of studies using a technique. Classifications are not mutually exclusive. Videotape, Audio tape, Software logged are sub-classifications of Observation so that any article classified in one of these three is also classified in Observation. Similarly, articles classified in Questionnaire pre and post experiment are also classified in Questionnaire.

Field methodologies

Like experimental studies, field based studies, which include field studies and case studies, used a distinct set of evaluative techniques (see table 8). These reflected a more informal and opportunistic style of data collection. Again, observation was the most common technique, with 9 occurrences, but the use of video decreased dramatically to only 2 articles. Interviews occurred nearly as often in 8 articles. Finally, the third most frequent evaluative technique was the collection and analysis of supporting archival materials such as the users' email or a company's newsletters.

Table 8: Techniques used in Case and Field Studies

Technique	#	%	References
Observation	9	75	[1, 3, 7, 19,30, 39,41,45,47]
-Videotape	2	17	[19, 47]
-Audio tape	0	0	[]
-Software logged	2	17	[1, 47]
Interview	8	67	[1, 5, 7, 19, 38, 39, 41, 51]

Collected Archival Material	4	33	[5, 7, 39, 47]
Questionnaire	3	25	[1, 19, 47]
Discussion	2	17	[30, 45]
Quantitative work measures	2	17	[1, 51]
Qualitative work measures	0	0	[]

indicates total studies in a classification, % indicates percentage of studies using a technique. 12 separate articles are included. These classifications are not mutually exclusive. Videotape, Audio tape, Software logged are sub-classifications of Observation so that any article classified in one of these three is also classified in Observation.

DISCUSSION

The role of context in groupware evaluation

Twidale, Randall, and Bentley [49] stress the importance of context in groupware evaluation. However, they point out the usefulness of less authentic evaluation techniques early in development as part of an ongoing formative evaluation. This allows for the elimination of larger, glaring problems early in the development process so that more subtle issues can be dealt with when the software is evaluated with the target user group.

The developer, then, could begin with multiple controlled evaluations to shape the initial prototypes. Many of the problems that would surface early on are obtrusive enough that they would pose difficulties for any user group. As these are overcome, the developer can then move into the naturalistic setting with a piece of software that is refined enough that the target users can begin offering feedback on issues that are unique to them.

From the articles that include enough information to allow classification along these lines, 29 combinations were recorded to allow comparison between placement of evaluation in the software lifecycle and type of evaluation conducted. Of these, 17 were ongoing formative evaluations or evaluations of prototypes [3, 8, 10, 15, 17, 19, 21, 25, 33, 35, 41, 43, 44, 45, 47, 48, 49] and 12 were evaluations of finished products [1, 5, 7, 11, 18, 28, 30, 37, 38, 39, 45, 51]. The formative evaluations were overwhelmingly carried out separately from any kind of work or organizational context. In all, 65% of these were in controlled settings (lab experiment, exploratory) [8, 15, 17, 21, 25, 33, 35, 43, 44, 48, 49], and the remaining 35% were field based (field study, field experiment) [3, 10, 19, 41, 45, 47]. Similarly, evaluations of completed groupware applications were largely carried out in naturalistic settings. In all, 67% of these were carried out using field based evaluations (field study and case study) [1, 5, 7, 30, 38, 39, 45, 51], and the remaining 33% used controlled settings (lab experiments) [11, 18, 28, 37].

These findings are in agreement with Twidale, Randall, and Bentley's hypothesis. This suggests that there may be real value in evaluations conducted out of context early in the development phase. But, as development progresses and the applications become more refined, it becomes increasingly important to move evaluations into the target work setting.

Patterns of evaluation techniques

Distinct patterns of evaluation techniques were used for each of the major evaluation types. These combinations of techniques are not mutually exclusive as presented here (see table 9). Instead, the percentages represent the number of articles in the given category that include the given combination. These are summarized below, and can potentially serve as guides for developers contemplating groupware evaluation:

Table 9: Patterns of evaluation techniques

Evaluation Type	Techniques	%	Citations
Lab Experiment	Qualitative, Quantitative	38	18, 25, 33, 37, 48
	Observation, Quantitative	38	21, 25, 28, 37, 48
	Observation, Interview	31	11, 15, 17, 35
	Questionnaire, Qualitative, Quantitative	31	25, 33, 37, 48
Field Study	Observation, Interview	38	1, 19, 41
	Observation, Questionnaire	38	1, 19, 47
Case Study	Interview, Collection of Archival materials	100	5, 7, 39, 51
Exploratory	Observation	100	8, 44, 49

% indicates percentage of articles using the given evaluation type which contain the specified combination of techniques.

New directions for published evaluations

Many of the complicating factors associated with groupware implementation stem from the difficulties introduced by context. In particular, it is difficult to predict how the software will fit into the organizational and work practices of a group of end users. Therefore, evaluating the impact of a groupware implementation in these areas should be a priority for developers. Additionally, published research literature should help illuminate this process.

However, from the studies included in this survey, 41% of the articles that included evaluations were of actual real world software implementations, but only 25% considered the software's organizational and work impact.

Since 1994, the number of longitudinal, workplace centered evaluations reported in CSCW proceedings has been decreasing (see table 10). *CSCW Proceedings '94* had the highest incidence of field based studies with 5 field studies [1, 3, 19, 45, 47] and 1 case study [7], and 3 of these examined organizational / work impacts [7, 45, 47]. Since then, the '96 proceedings showed a decrease to 1 field study [38] and 2 case studies [5, 51], with only 2 of these examining organizational / work impacts [5, 38]. Finally, the number of naturalistic studies decreased in 1998 to 2 field studies [30, 41], but both of these examined organizational / work impacts.

Table 10: Evaluations by year of Proceedings

	1990	1992	1994	1996	1998
#	4	13	12	9	7
%	9	29	27	20	16
Lab Experiment		5	3	3	2
Field Study			5	1	2
Case Study		1	1	2	
Exploratory			2		1
Field experiment	1				
Cannot classify				1	
Not enough info	2				1
Introspection		3			1
No evaluation	1	4	2	2	

indicates total studies in a classification, % indicates percentage of studies in a classification. 45 separate articles are included.

In order to gain an understanding of how these evaluations can be better carried out, further research needs to be done on longitudinal field based evaluation. The focus must shift toward the users and the organization. Grudin [14] has emphasized the difficulty of learning from previous evaluations due to the broad range of organizations and users that must be dealt with. In spite of this, further work should be done on refining data collection methodologies with a focus on conducting these longitudinal studies in a way that is maximally time and cost efficient.

CONCLUSIONS

The slow uptake of groupware technologies may be an indication that developers are not matching the software they design with the needs of the target user groups. If properly carried out, evaluation can provide a means for bringing the groupware application closer to meeting the users' needs. Little has been written about how such a formative, situated evaluation should be carried out, and

many of the recommendations that have been made [14] are not frequently reflected in CSCW proceedings. When evaluating the final implementation, summative evaluations may not provide adequate information, since work practices evolve over time. In the articles included in this survey, the minimum amount of time spent contemplating software post implementation was 20 weeks, with a maximum time commitment reaching 36 months.

A large number of articles in this study introduced new groupware applications but made no mention of evaluation, and others were vague about the details of the evaluations they performed. If it is possible to learn from experience in this area, evaluation information is vital and should be included.

An effort should be made to identify new evaluative techniques in order to make evaluation more practical in terms of time and cost. One means of accomplishing this may be through adapting single user evaluation techniques to groupware applications. Ereback and Hook [12] attempted this using cognitive walkthrough, but the results were mixed and somewhat inconclusive. It is not altogether clear whether adapting single user techniques will lead to useful groupware evaluation techniques, and it is doubtful that anything short of evaluation in the workplace will yield a complete set of results. However, new evaluation techniques can potentially lead to big payoffs when the system is still at a prototype level.

REFERENCES

1. Ackerman, M.S. Augmenting the Organizational Memory: a Field Study of Answer Garden. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 243 - 252.
2. Baeza-Yates, R. and Pino, J.A. A First Step to Formally Evaluate Collaborative Work. *Proceedings ACM SIGGROUP* (Phoenix, AZ, November 1997), 56 - 60.
3. Bergmann, N.W. and Mudge, J.C. Automated Assistance for the Telemeeting Lifecycle. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 373 - 384.
4. Beuscart-Zephir, M.C., Molenda, S., Grave, C., Dufresne, E. Usability Assessment of Interactive Multimedia Medical Workstation. *Proc. 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Baltimore, November 1994), IEEE Press, p. 1358-9, vol.2.
5. Bikson, T.K. and Eveland, J.D. Groupware Implementation: Reinvention in the Sociotechnical Frame. *Proceedings CSCW '96* (Boston, November 1996), 428 - 437.
6. Boland, R.J., Maheshwari, A.K., Te'eni, D., Schwartz, D.G. and Tenkasi, R.V. Sharing Perspectives in Distributed Decision Making. *Proceedings CSCW '92* (Toronto, November 1992), 306 - 313.
7. Bowers, J. The Work to Make a Network Work: Studying CSCW in Action. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 287 - 298.
8. Brave, S., Ishii, H. and Dahley, A. Tangible Interfaces for Remote Collaboration and Communication. *Proceedings CSCW '98* (Seattle, November 1998), 169 - 178.
9. Brinck, T. and Gomez, L.M. A Collaborative Medium for the Support of Conversational Props. *Proceedings CSCW '92* (Toronto, November 1992), 171 - 178.
10. Brothers, L., Sembugamoorthy, V. and Muller, M. ICICLE: Groupware for Code Inspection. *Proceedings CSCW '90* (Los Angeles, October 1990), 169 - 181.
11. Dourish, P. and Bellotti, V. Awareness and Coordination in Shared Workspaces. *Proceedings CSCW '92* (Toronto, November 1992), 107 - 114.
12. Ereback A.-L. and Hook, K. Using Cognitive Walkthrough for Evaluating a CSCW Application. *Proceedings CHI '94* (Boston, April 1994), 91 - 92.
13. Goldberg, Y., Safran, M. and Shapiro, E. Active Mail--A Framework for Implementing Groupware. *Proceedings CSCW '92* (Toronto, November 1992), 75 - 83.
14. Grudin, J. Groupware and Social Dynamics: Eight Challenges for Developers. *Commun. ACM* 37, 1 (January 1994), 92 - 105.
15. Gutwin, C., Roseman, M. and Greenberg, S. A Usability Study of Awareness Widgets in a Shared Workspace Groupware System. *Proceedings CSCW '96* (Boston, November 1996), 258 - 267.
16. Haake, J.M. and Wilson, B. Supporting Collaborative Writing of Hyperdocuments in SEPIA. *Proceedings CSCW '92* (Toronto, November 1992), 138 - 146.
17. Hindmarsh, J., Fraser, M., Heath, C., Benford, S. and Greenhalgh, C. Fragmented Interaction: Establishing mutual orientation in virtual environments. *Proceedings CSCW '98* (Seattle, November 1998), 217 - 226.
18. Hymes, C.M. and Olson, G.M. Unblocking Brainstorming Through the Use of a Simple Group Editor. *Proceedings CSCW '92* (Toronto, November 1992), 99 - 106.
19. Isaacs, E.A., Morris, T. and Rodriguez, T.K. A Forum for Supporting Interactive Presentations to Distributed Audiences. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 405 - 416.
20. Isaacs, E.A., Tang, J.C. and Morris, T. Piazza: A Desktop Environment Supporting Impromptu and Planned Interactions. *Proceedings CSCW '96* (Boston, November 1996), 315 - 324.
21. Ishii, H.; Kobayashi, M. and Grudin, J. Integration of Inter-Personal Space and Shared Workspace: ClearBoard

- Design and Experiments. *Proceedings CSCW '92* (Toronto, November 1992), 33 - 42.
22. Ishii, H. TeamWorkStation: Towards a Seamless Shared Workspace. *Proceedings CSCW '90* (Los Angeles, October 1990), 13 - 26.
23. Johnson-Lenz, P. and Johnson-Lenz, T. Groupware: The Process and Impacts of Design Choices. In *Computer-Mediated Communication Systems: Status and Evaluation*. Kerr, E. and Hiltz, S. (eds.), Academic Press, 1982, 45-55.
24. Kaplan, S.M., Tolone, W.J., Bogia, D.P. and Bignoli, C. Flexible, Active Support for Collaborative Work with ConversationBuilder. *Proceedings CSCW '92* (Toronto, November 1992), 378 - 385.
25. Kraut, R.E., Miller, M.D. and Siegel, J. Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication. *Proceedings CSCW '96* (Boston, November 1996), 57 - 66.
26. Lee, J. SIBYL: A Tool for Managing Group Decision Rationale. *Proceedings CSCW '90* (Los Angeles, October 1990), 79 - 92.
27. Lee, J.H., Prakash, A., Jaeger, T. and Wu, G. Supporting Multi-User, Multi-Applet Workspaces in CBE. *Proceedings CSCW '96* (Boston, November 1996), 344 - 353.
28. Mark, G., Haake, J.M. and Streitz, N.A. Hypermedia Structures and the Division of Labor in Meeting Room Collaboration. *Proceedings CSCW '96* (Boston, November 1996), 170 - 179.
29. McGrath, J.E. Methodology Matters: Doing Research in the Behavioral and Social Sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000*, 2nd Ed. Baecker, R.M., Grudin, J., Buxton, W.A.S., Greenberg, S. (eds.), Morgan Kaufman Publishers, San Francisco. 1995, 152-169.
30. Moran, T.P., van Melle, W. and Chiu, P. Tailorable Domain Objects as Meeting Tools for an Electronic Whiteboard. *Proceedings CSCW '98* (Seattle, November 1998), 295 - 304.
31. Myers, B.A., Stiel, H. and Gargiulo, R. Collaboration Using Multiple PDAs Connected to a PC. *Proceedings CSCW '98* (Seattle, November 1998), 285 - 294.
32. Neuwirth, C.M., Chandhok, R., Kaufer, D.S., Erion, P., Morris, J. and Miller, D. Flexible Diff-ing In A Collaborative Writing System. *Proceedings CSCW '92* (Toronto, November 1992), 147 - 154.
33. Neuwirth, C.M., Morris, J.H., Regli, S.H., Chandhok, R. and Wenger, G.C. Envisioning Communication: Task-Tailorable Representations of Communication in Asynchronous Work. *Proceedings CSCW '98* (Seattle, November 1998), 265 - 274.
34. Newman-Wolfe, R.E., Webb, M. L. and Montes, M. Implicit Locking in the Ensemble Concurrent Object-Oriented Graphics Editor. *Proceedings CSCW '92* (Toronto, November 1992), 265 - 272.
35. Okada, K., Maeda, F., Ichikawaa, Y. and Matsushita, Y. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 385 - 393.
36. Olsen, D.R., Hudson, S.E., Phelps, M., Heiner, J. and Verratti, T. Ubiquitous Collaboration via Surface Representations. *Proceedings CSCW '98* (Seattle, November 1998), 129 - 138.
37. Olson, J.S., Olson, G.M., Storrøsten, M. and Carter, M. How a Group-Editor Changes the Character of a Design Meeting as well as its Outcome. *Proceedings CSCW '92* (Toronto, November 1992), 91 - 98.
38. Olson, J.S. and Teasley, S. Groupware in the Wild: Lessons Learned from a Year of Virtual Collocation. *Proceedings CSCW '96* (Boston, November 1996), 419 - 427.
39. Orlikowski, W.J. Learning From Notes: Organizational Issues in Groupware Implementation. *Proceedings CSCW '92* (Toronto, November 1992), 362 - 369.
40. Pacull, F., Sandoz, A. and Schiper, A. Duplex: A Distributed Collaborative Editing Environment in Large Scale. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 165 - 173.
41. Prinz, W., Mark, G. and Pankoke-Babatz, U. Designing Groupware for Congruency in Use. *Proceedings CSCW '98* (Seattle, November 1998), 373 - 382.
42. Roseman, M. and Greenberg, S. TeamRooms: Network Places for Collaboration. *Proceedings CSCW '96* (Boston, November 1996), 325 - 333.
43. Shu, L. and Flowers, W. Groupware Experiences in Three-Dimensional Computer-Aided Design. *Proceedings CSCW '92* (Toronto, November 1992), 179 - 186.
44. Sohlenkamp, M. and Chwelos, G. Integrating Communication, Cooperation, and Awareness: The DIVA Virtual Office Environment. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 331 - 343.
45. Star, S.L. and Ruhleder, K. Steps Towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-Scale Collaborative Systems. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 253 - 264.
46. Streitz, N.A., Geißler, J., Haake, J.M. and Hol, J. DOLPHIN: Integrated Meeting Support across Local and Remote Desktop Environments and LiveBoards. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 345 - 358.

47. Tang, J.C., Isaacs, E.A. and Rua, M. Supporting Distributed Groups with a Montage of Lightweight Interactions. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 23 - 34.
48. Toth, J.A. The Effects of Interactive Graphics and Text on Social Influence in Computer-Mediated Small Groups. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 299 - 310.
49. Twidale, M., Randall, D., Bentley, R. Situated Evaluation for Cooperative Systems. *Proceedings CSCW '94* (Chapel Hill, NC, October 1994), 441 - 452.
50. Watabe, K., Sakata, S., Maeno, K., Fukuoka, H. and Ohmori, T. Distributed Multiparty Desktop Conferencing System: MERMAID. *Proceedings CSCW '90* (Los Angeles, October 1990), 27 - 38.
51. Whittaker, S. Talking to Strangers: An Evaluation of the Factors Affecting Electronic Collaboration. *Proceedings CSCW '96* (Boston, November 1996), 409 - 418.