A Review of Groupware Evaluations

David Pinelle and Carl Gutwin
Department of Computer Science, University of Saskatchewan
[dlp849, gutwin]@cs.usask.ca

Abstract

A better understanding of how groupware systems have been evaluated in the past can help to frame the discussion of what methods and techniques should be considered for future evaluations. We reviewed all papers from the ACM CSCW conference (1990-1998) that introduced or evaluated a groupware system. Forty-five papers were included in the review. The main findings are that almost one-third of the groupware systems were not evaluated in any formal way, that only about one-quarter of the articles included evaluations in a real-world setting, and that a wide variety of evaluation techniques are in use. Our main conclusions from the review are that more attention must be paid to evaluating groupware systems and that there is room for additional evaluation techniques that are simple and low in cost.

1. Introduction

As groupware systems become more common, more attention is being paid to the question of how to evaluate multi-user systems. Researchers and developers have employed a range of techniques including scientific, engineering, and social science methodologies, but there is no clear consensus on which methods are appropriate in which circumstances. In addition, groupware is traditionally considered to be difficult to evaluate because of the effects of multiple people and the social and organizational context [14, 39].

In order to gain a better understanding of how different evaluation techniques have been used in the past, and of the strengths and weaknesses of the different methods, we reviewed a sample of CSCW research that discussed groupware systems and evaluations of groupware. This review provides a foundation from which we can consider questions of when and how to evaluate future systems.

This paper presents our review of 45 papers from the ACM CSCW conference 1990 – 1998. Articles were included in this survey if they introduced a new groupware application or contained an evaluation of an existing groupware application. We only considered applications if they were intended to support specific task goals; basic tools such as email systems were not included.

We begin by outlining a set of categories and criteria that we used to analyze the papers, and then detail the results of the review.

2. Classifying groupware evaluations

We analyzed each article in five areas: type of evaluation, characteristics of the evaluation, data collection techniques, placement of the evaluation in the software development cycle, and type of conclusions drawn from the evaluation. Each of these areas is described below.

2.1. Type of evaluation

McGrath's [29] classification scheme for research strategies was used as a starting point for assessing the type of evaluation carried out. The major differentiating characteristics between the strategies are level of experimental manipulation and evaluation setting (Figure 1).

Two additional categories were added to McGrath's scheme. 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 were classified as case studies.

Figure 1. Evaluation types

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

Second, articles that did not contain formal evaluations were separated into two groups. Articles that drew conclusions based only on the researcher's experience were classified as introspective studies. Articles without formal evaluations that did not draw conclusions about the software were placed in the no evaluation category.

2.2. Characteristics of the evaluation

Evaluations were further classified according to the rigor of the experimental manipulation and the type and rigor of measurements. Twidale, Randall, and Bentley's [49] dimensions for classifying evaluations were combined with McGrath's [29] techniques for experimental manipulation and techniques for measuring experimental results to produce the classification scheme below.

- Formative vs. Summative

- Quantitative vs. Qualitative

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

- Measurement:
- Formal / rigorous
- Informal

2.3. Techniques for data collection

We also assessed each evaluation's data collection techniques. There were seven main techniques:

- User Observation
- Interview
- Discussion
- Questionnaire
- Qualitative work measures
- Quantitative work measures
- Collection of archival material

2.4. Placement of evaluation in 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. We considered six potential placements of the evaluation:

- Periodic evaluations throughout development process
- Continuous evaluation throughout development
- Evaluation of a prototype

- Evaluation of a finished piece of software
- Periodic evaluations after software implementation
- Continuous evaluation after software implementation

2.5. Focus of the evaluation

A set of categories was developed to identify the focus of the evaluation. An initial set of categories was found in the three categories of benefits cited by Baeza-Yates and Pino [2]. These were extended to include concepts from a multi-stage evaluation process discussed by Beuscart-Zephir et al [4]. Types of evaluation focus include:

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

3. Results

We include results from our analyses in each of the areas outlined above.

3.1. Characterizing the software

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]

	#	%	References
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].

The majority of the papers introduced synchronous applications. Most of the systems were academic or research implementation, and a smaller number of

implementations were installed with in a real world setting. Results are summarized in Table 1[1].

3.2. Evaluation type

The two main types of evaluations seen in the review are laboratory experiments and field studies. However, several other types of evaluation were seen that are related to field methods. When these groups are taken together, the ratio of field to laboratory studies is even (see Table 2).

Almost forty percent of the articles contained no formal evaluation at all. Of these 17 papers, four contained introspective evaluations, and nine did not report on any study. Four more did not include enough information to determine if an evaluation had been 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]
Introspection	4	9	[16, 24, 34, 36]
No Evaluation	9	20	[6, 9, 13, 20, 26, 27, 32, 40, 46]
Not enough information	4	9	[22, 31, 50, 42]

3.3. Characterizing the evaluation

Formative evaluations were more prevalent than summative evaluations (see Table 3). 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 4). Only a small number of articles contained evaluations that were distributed throughout the software lifecycle.

---

[1] Tables show the number of studies (#) and percentage of the total (%). Classifications are not necessarily mutually exclusive, and percentages may not always add to 100%.

Table 3: Characterizing the evaluations

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]

Table 4: 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]

3.4. 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 5). 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 5: 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]

3.5. Focus of the evaluation

Only a small number of studies examined the organizational impact or 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 6).

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, and ranged from 4.5 to 36 months.

Table 6: Focus of the evaluation

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	5	16	[25, 37, 42, 49, 51]
Efficiency of task performance	4	13	[18, 25, 33, 48]

4. Discussion

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.

4.1. Evaluation in the work place

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. 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.

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 time and cost efficient.

4.2. New techniques for groupware evaluation

An effort should be made to develop new evaluation techniques in order to make evaluation more practical in terms of time and cost. By improving techniques used prior to workplace evaluations, many problems can be eliminated early on, thus improving the efficiency of evaluation as it progresses into the workplace.

One means of developing intermediate level evaluation techniques may be through adapting techniques used to evaluate single user applications. Ereback and Hook [12] attempted this using cognitive walkthrough, but the results were mixed and somewhat inconclusive.

5. Conclusions

This paper explores groupware evaluation trends and techniques by surveying a wide variety of articles from the ACM CSCW conference. We closely analyzed all articles that introduced or evaluated groupware applications. After analyzing forty-five papers, we discovered that almost one-third of the groupware systems were not evaluated in any formal way and that only about one-quarter of the evaluations involve a real-world setting. We suggested future directions for groupware evaluation. In particular, evaluation techniques should be developed to help eliminate problems with the application before it progresses into situated workplace-centered evaluations.

6. References

[1] Ackerman, M.S. Augmenting the Organizational Memory: a Field Study of Answer Garden. *CSCW '94*, 243-252.

[2] Baeza-Yates, R. and Pino, J.A. A First Step to Formally Evaluate Collaborative Work. *SIGGROUP '97*, 56-60.

[3] Bergmann, N.W. and Mudge, J.C. Automated Assistance for the Telemeeting Lifecycle. *CSCW '94*, 373-384.

[4] Beuscart-Zephir, M.C., Molenda, S., Grave, C., Dufresne, E. Usability Assessment of Interactive Multimedia Medical Workstation. *Proc. IEEE Engineering in Medicine and Biology Society*, p. 1358-9, vol.2.

[5] Bikson, T.K. and Eveland, J.D. Groupware Implementation: Reinvention in the Sociotechnical Frame. *CSCW '96*, 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. CSCW '92*, 306-313.

[7] Bowers, J. The Work to Make a Network Work: Studying CSCW in Action. *CSCW '94*, 287-298.

[8] Brave, S., Ishii, H. and Dahley, A. Tangible Interfaces for Remote Collaboration and Communication. *CSCW '98*, 169-178.

[9] Brinck, T. and Gomez, L.M. A Collaborative Medium for the Support of Conversational Props. *CSCW '92*, 171-178.

[10] Brothers, L., Sembugamoorthy, V. and Muller, M. ICICLE: Groupware for Code Inspection. *CSCW '90*, 169-181.

[11] Dourish, P. and Bellotti, V. Awareness and Coordination in Shared Workspaces. *CSCW '92*, 107-114.

[12] Ereback A.-L. and Hook, K. Using Cognitive Walkthrough for Evaluating a CSCW Application. *CHI '94*, 91-92.

[13] Goldberg, Y., Safran, M. and Shapiro, E. Active Mail--A Framework for Implementing Groupware. *CSCW '92*, 75-83.

[14] Grudin, J. Groupware and Social Dynamics: Eight Challenges for Developers. *Commun. ACM* 37, 1, 92-105.

[15] Gutwin, C., Roseman, M. and Greenberg, S. A Usability Study of Awareness Widgets in a Shared Workspace Groupware System. *CSCW '96*, 258-267.

[16] Haake, J.M. and Wilson, B. Supporting Collaborative Writing of Hyperdocuments in SEPIA. *CSCW '92*, 138-146.

[17] Hindmarsh, J., Fraser, M., Heath, C., Benford, S. and Greenhalgh, C. Fragmented Interaction: Establishing mutual orientation in virtual environments. *CSCW '98*, 217-226.

[18] Hymes, C.M. and Olson, G.M. Unblocking Brainstorming Through the Use of a Simple Group Editor. *CSCW '92*, 99-106.

[19] Isaacs, E.A., Morris, T. and Rodriguez, T.K. A Forum for Supporting Interactive Presentations to Distributed Audiences. *CSCW '94*, 405-416.

[20] Isaacs, E.A., Tang, J.C. and Morris, T. Piazza: A Desktop Environment Supporting Impromptu and Planned Interactions. *CSCW '96*, 315-324.

[21] Ishii, H.; Kobayashi, M. and Grudin, J. Integration of Inter-Personal Space and Shared Workspace: ClearBoard Design and Experiments. *CSCW '92*, 33- 42.

[22] Ishii, H. TeamWorkStation: Towards a Seamless Shared Workspace. *CSCW '90*, 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, 42-55.

[24] Kaplan, S.M., Tolone, W.J., Bogia, D.P. and Bignoli, C. Flexible, Active Support for Collaborative Work with ConversationBuilder. *CSCW '92*, 378-385.

[25] Kraut, R.E., Miller, M.D. and Siegel, J. Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication. *CSCW '96*, 57-66.

[26] Lee, J. SIBYL: A Tool for Managing Group Decision Rationale. *CSCW '90*, 79-92.

[27] Lee, J.H., Prakash, A., Jaeger, T. and Wu, G. Supporting Multi-User, Multi-Applet Workspaces in CBE. *CSCW '96*, 344-353.

[28] Mark, G., Haake, J.M. and Streitz, N.A. Hypermedia Structures and the Division of Labor in Meeting Room Collaboration. *CSCW '96*, 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. *CSCW '98*, 295-304.

[31] Myers, B.A., Stiel, H. and Gargiulo, R. Collaboration Using Multiple PDAs Connected to a PC. *CSCW '98*, 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. *CSCW '92*, 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. *CSCW '98*, 265-274.

[34] Newman-Wolfe, R.E., Webb, M. L. and Montes, M. Implicit Locking in the Ensemble Concurrent Object-Oriented Graphics Editor. *CSCW '92*, 265-272.

[35] Okada, K., Maeda, F., Ichikawaa, Y. and Matsushita, Y. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. *CSCW '94*, 385-393.

[36] Olsen, D.R., Hudson, S.E., Phelps, M., Heiner, J. and Verratti, T. Ubiquitous Collaboration via Surface Representations. *CSCW '98*, 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. *CSCW '92*, 91-98.

[38] Olson, J.S. and Teasley, S. Groupware in the Wild: Lessons Learned from a Year of Virtual Collocation. *CSCW '96*, 419-427.

[39] Orlikowski, W.J. Learning From Notes: Organizational Issues in Groupware Implementation. CSCW '92, 362-369.

[40] Pacull, F., Sandoz, A. and Schiper, A. Duplex: A Distributed Collaborative Editing Environment in Large Scale. *CSCW '94*, 165-173.

[41] Prinz, W., Mark, G. and Pankoke-Babatz, U. Designing Groupware for Congruency in Use. CSCW '98, 373-382.

[42] Roseman, M. and Greenberg, S. TeamRooms: Network Places for Collaboration. *CSCW '96*, 325-333.

[43] Shu, L. and Flowers, W. Groupware Experiences in Three-Dimensional Computer-Aided Design. *CSCW '92*, 179-186.

[44] Sohlenkamp, M. and Chwelos, G. Integrating Communication, Cooperation, and Awareness: The DIVA Virtual Office Environment. *CSCW '94*, 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. *CSCW '94*, 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. *CSCW '94*, 345-358.

[47] Tang, J.C., Isaacs, E.A. and Rua, M. Supporting Distributed Groups with a Montage of Lightweight Interactions. *CSCW '94*, 23-34.

[48] Toth, J.A. The Effects of Interactive Graphics and Text on Social Influence in Computer-Mediated Small Groups. *CSCW '94*, 299-310.

[49] Twidale, M., Randall, D., Bentley, R. Situated Evaluation for Cooperative Systems. CSCW '94, 441-452.

[50] Watabe, K., Sakata, S., Maeno, K., Fukuoka, H. and Ohmori, T. Distributed Multiparty Desktop Conferencing System: MERMAID. *CSCW '90*, 27-38.

[51] Whittaker, S. Talking to Strangers: An Evaluation of the Factors Affecting Electronic Collaboration. *CSCW '96*, 409-418.