Supporting Common Ground and Awareness in Emergency Management Planning: A Design Research Project

GREGORIO CONVERTINO, HELENA M. MENTIS, ALEKSANDRA SLAVKOVIC, MARY BETH ROSSON, and JOHN M. CARROLL, The Pennsylvania State University

We present a design research project on knowledge sharing and activity awareness in distributed emergency management planning. In three experiments we studied groups using three different prototypes, respectively: a paper-prototype in a collocated work setting, a first software prototype in a distributed setting, and a second, enhanced software prototype in a distributed setting. In this series of studies we tried to better understand the processes of knowledge sharing and activity awareness in complex cooperative work by developing and investigating new tools that can support these processes. We explicate the design rationale behind each prototype and report the results of each experiment investigating it. We discuss how the results from each prototyping phase brought us closer to defining properties of a system that facilitate the sharing and awareness of both content and process knowledge. Our designs enhanced aspects of distributed group performance, in some respects beyond that of comparable face-to-face groups.

Categories and Subject Descriptors: H.5.m. [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous

General Terms: Design, Experimentation, Measurement

Additional Key Words and Phrases: Activity awareness, common ground, CSCW, design, colocated and distributed geo-collaboration, emergency management planning

ACM Reference Format:
Convertino, G., Mentis, H. M., Slavkovic, A., Rosson, M. B., and Carroll, J. M. 2011. Supporting common ground and awareness in emergency management planning: A design research project. ACM Trans. Comput.-Hum. Interact. 18, 4, Article 22 (December 2011), 34 pages.
DOI = 10.1145/2063231.2063236 http://doi.acm.org/10.1145/2063231.2063236

1. NTRODUCTION

Group psychologists have repeatedly found evidence of lower productivity of distributed groups, in relation to face-to-face groups. They have attributed these performance issues to the extra communication costs for distributed groups, due to reduction of useful cues in the distributed setting. McGrath and collaborators reported "a reduction in cues such as eye contact [and] head nods [. . .] creates disruptions in the flow of communication" and these discrepancies due to the setting are greater "for tasks requiring higher

This research was supported by the Collaboration and Knowledge Interoperability Program of the Office of Naval Research under award N000140510549. Data capture and analysis tools for the lab experiments were provided by Information Technology (Lucas Noldus, Andrew Spink, Tobias Heffelaar) [Noldus].
Authors' addresses: G. Convertino, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 9304; email: gconvertino@gmail.com; H. M. Mentis, Microsoft Research Cambridge, 7 JJ Thomson Avenue, Cambridge CB30FB, United Kingdom; email: hmentis@microsoft.com; A. Slavkovic, Department of Statistics, The Pennsylvania State University, University Park, PA 16801; email: sesa@stat.psu.edu; M. B. Rosson and J. M. Carroll, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16801; email: {mrosson, jcarroll}@ist.psu.edu.

DOI 10.1145/2063231.2063236 http://doi.acm.org/10.1145/2063231.2063236

levels of coordination" [Straus and McGrath 1994, p. 87-88]. Furthermore, research on common ground has found that as communicators move from face-to-face to distributed settings they lose the useful constraint of copresence and thus need to make more effort to reduce ambiguity [Clark and Brennan 1991].

These challenges for distributed groups pose a problem for emergency management groups who often work together in a distributed fashion—at times not ever having had the chance to meet face to face beforehand. A lack of any previous joint actions presents a large hurdle for newly formed groups to quickly and effectively share the relevant content and coordinate the work process; not least in part due to the group members' different areas of expertise or job roles that imply different languages, responsibilities, and priorities [McCarthy et al. 1991]. This increases the need for group members to quickly build enough shared knowledge as well as to continually maintain awareness of one another's actions and intentions in order to communicate, coordinate, and perform well. Our research responds to this growing need.

Another challenge is posed by the complexity of the *collaborative planning* activities of emergency management groups. Planning activities allow such groups to envision, discuss and practice emergency management strategies. In a planning session, such as a tabletop exercise [Schafer et al. 2007], a group of collaborators holding diverse job roles must analyze, filter, share, and manage large amounts of content, while understanding the emergency scenarios, evaluating possible responses, and planning the best response procedure. This is all in order to build a resource of knowledge about collaborators' roles and expertise, and prepare the decision-making processes that may be called upon during the event of an actual emergency response incident. Currently, little research and very little technology support are available to improve this knowledge building process for distributed groups.

In this article, we use a design research approach to address the problem of supporting knowledge sharing and activity awareness in groups performing emergency management planning (EMP) tasks. In previous papers [Carroll et al. 2007; Convertino et al. 2008a, 2008b, 2009], we reported results from two of a series of three lab studies; we demonstrated that a paper-prototype and first software prototype provided an incremental gain in shared knowledge for EMP groups over multiple runs of a collaborative task. This article describes the full design research program that produced these two prototypes and their evaluations, as well as a second iteration of the software prototype (not yet reported). We explicate the rationale behind the design of the prototypes and the results of lab studies that evaluated the development of shared knowledge and activity awareness in each condition. We discuss how the results from each prototyping phase brought us closer to defining properties of a system that better supports knowledge sharing and activity awareness, and thus performance. Finally, we characterize our approach and draw implications for theory, design, and methods.

In summary, the contributions are the full report on a design research program, which includes both reported and unreported findings and prototypes from three experiments, the characterization of the design research approach used that has been unreported, and the implications derived from a full analysis and comparison of the findings for theory, design, and methods, all of which are new contributions.

2. BACKGROUND

2.1. Activity Awareness and Common Ground

Early research on group decision-making conceptualized shared knowledge as shared mental models (knowledge structures held in common by members), and focused on how shared mental models predicted group performance (e.g. Cannon-Bowers et al. [1993]). More recently, researchers studying group decision-making and collaborative technology have shifted attention from team mental models to transactive memory-knowledge of who knows what within a group. Transactive models are particularly appropriate for group tasks like emergency planning that involve interdependency and role specialization; several researchers have argued that the work of groups making complex decisions is more accurately represented by these models (e.g., Cooke et al. [2000]; Mohammed and Dumville [2001]). This reasoning suggests that the ability of multi-role groups to share large amounts of content could be productively augmented through tools that afford selective sharing of information when it is needed during the collaboration process, as opposed to sharing all information in advance regardless of need.

A related construct is common ground, which recognizes that communicators have a mutual understanding of the content and process of their communication and further that they all know that they have this mutual understanding. This shared understanding enables greater efficiency, or minimum effort for communication—the functional role of common ground. Clark and his collaborators proposed that specific communication contexts could be described in terms of specific sets of grounding constraints [Clark and Brennan 1991; Monk 2003]. The more constraints on ambiguity in communication a media can provide, the better the media is for facilitating common ground and thus efficient communication. This framework has enabled HCI researchers to make systematic comparisons, generalizations, and predictions among media. The establishment and maintenance of common ground can be seen as an ongoing dynamic process; thus tools that can reinforce shared understandings of who knows what may augment performance.

Finally, researchers studying groups in dynamic situations have suggested that to perform effectively, the group members need to share preexisting strategic knowledge, which allows them to develop expectations for cue/action sequences (which cue should trigger which action), cue patterns and their significance (cue patterns associated with task strategies), group resources and capabilities (roles, resources, and expertise), and strategies most appropriate for the current task (see review in Convertino et al. [2005]). For example, consider a well-trained basketball team, where a successful blind pass from one player to another requires both to simultaneously assess a pattern of cues in the shared dynamic environment of the game and promptly respond to the current situation using an agreed tactic with no need for explicit communication. If such strategies can be visualized or reinforced in a computer-mediated group activity, they might provide important input to the activity awareness process, as they allow group members to set up and monitor a range of useful expectations.

In a prior theoretical contribution, we described the process of shared knowledge development such as common ground creation and maintenance as a subprocess that should be supported for activity awareness to adequately develop, and consequently, for group performance to improve [Carroll et al. 2006]. Activity awareness is an awareness of project work that enables effective group performance in complex tasks [Carroll et al. 2003]. The term "activity," which is adopted from activity theory, refers to long-term endeavors directed at major goals [Bodker 1996; Bardram 1998]. We argue that through the support of knowledge-sharing and maintenance of knowledge in the minds of collaborators, we can effectively support activity awareness and achieve better group performance.

In this article we extend our earlier discussion, articulating specific hypotheses tested through a series of experiments where we measured the effects of technology interventions on groups' sharing process and knowledge development, as well as their performance. Thus, when we refer to knowledge-sharing we are specifically referring to the development of shared knowledge as integral to transactive memory, common ground, and strategic knowledge. We are not discussing knowledge management tools but rather tools for collaboration that provide assistance to groups who need to develop shared knowledge that can be called upon in future collaborative efforts. In addition, when we refer to collaborative awareness we are focusing specifically on a particular type of awareness—that of activity awareness. We investigate both knowledge-sharing and activity-awareness with groups performing emergency management planning tasks.

2.2. Emergency Management Planning

The design research program documented in this article originated in the findings of a field study of emergency management groups in two local communities in Central Pennsylvania, USA [Schafer et al. 2007, 2008]. One result of the field study was a focus on the planning phase of the emergency management work conducted by the groups in these communities.

Prior research has postulated four phases of emergency management: planning, response, recovery, and mitigation [Haddow and Bullock 2003]. This article focuses on activity-awareness and knowledge-sharing in the planning phase. In contrast, much prior research in emergency management has focused on the response phase (e.g., as when fire engines and personnel are called out to a site and must coordinate their actions in response to changing and evolving circumstances). Importantly, while it may be less salient, the planning phase, occurs much more widely than response. The planning phase lays the foundation for the response and recovery phase in that it provides an opportunity for emergency management coordinators to work with local officials, first response agencies, and the vast array of stakeholders and volunteers, to develop emergency plans so that they can become, and remain, prepared for any emergency event that may happen. In terms of day-to-day activities, planning is a major responsibility of emergency management groups, simply because real emergencies are rare and care must be taken to prepare for as many exigencies as possible. Thus, the planning phase consists of the cooperative work that precedes any actual emergency disaster and allows stakeholders to discuss and practice emergency management strategies. The response groups rehearse interdependencies and build a shared planning experience. This experience supports improvisation during actual crises, because the groups can draw upon more shared knowledge and awareness and this facilitates more effective and efficient work.

A common activity for emergency management planning (EMP) professionals is the tabletop exercise. In such an exercise, group members walk through a scenario, identifying but also reconsidering existing response procedures on a shared paper map positioned on a central meeting table. After observing such sessions, we decided to begin work on a tool for distributed emergency planning exercises. Our motivations in this were three-fold: first, the group members' abilities to build these shared experiences together was constrained by their availability for colocated meetings; there was little to no existing support for distributed work and the meetings that did occur were few and far between. This provided a large hurdle for continuous shared knowledge development and maintenance. Second, when meeting in person, the group's ability to share and manage large amounts of content was limited by a lack of interactivity in the physical tools they used. Maps and paper-based annotations do not support exploration and retrieval of information, distributed annotations, reeditable shared annotations, or reuse of prior annotations or their categorizations. A third key issue was the absence of explicit acknowledgements or reinforcements of the roles that are enacted by members representing different agencies (e.g., police, civil government) and expertise (e.g., healthcare, engineering). These roles, which consist of sets of agency-specific competences and responsibilities, are a central organizing rubric for EMP. The absence of role-oriented tools limited the group's ability to select the right content to share (e.g., what should a civil engineer contribute at any given point); at the same time they limited their ability to implicitly share the act of sharing itself. By this we mean a shared awareness of how the activity's data, judgments, or actions are related to roles (see agency-specific data, judgments, or actions), as well as the strategy for tackling the overall task (e.g., first share key facts, then discuss these by area).

Building from the fieldwork [Schafer et al. 2007, 2008], we began to investigate regional EMP more systematically and to design a system that supported more flexible and distributed forms of these planning activities. Our primary design focus was on features that might increase the level of shared knowledge that the groups developed during each iteration of their tabletop exercises. Our design goal was that the resulting software system should meet the following objectives.

(1) It should provide at least the same level of activity awareness as a paper-based system used in a face-to-face setting (the planning staff's current system and thus our baseline).
(2) It should provide, through tools and functionality not available in a paper-based system, a greater level of activity awareness.

To pursue these design objectives, we began an investigation of specific properties of this emergency planning activity that might be used to facilitate activity-awareness in distributed working groups. We first studied a paper prototype in a face-to-face environment to establish a baseline. In close comparison with this physical prototype used for collocated group work, we studied two software prototypes. The first mirrored many of the features seen with the paper prototype but also supported distributed work. The second added a new feature aimed at facilitating activity awareness in a way that deliberately capitalizes on the shift to distributed group activity. The comparison between the collocated and remote settings as well as the three media (paper prototype and the two software systems) provides a basis for inferences about what qualities of the medium facilitate activity awareness as well as a more general understanding of group coordination and activity awareness in distributed work.

This work we report differs from previous work in this area and contributes to the research community in two distinct ways. First, we focus on the knowledge-sharing processes in support of activity-awareness that occur in groups who perform complex cooperative work. Our research originates from fieldwork on real groups and complex tasks that motivate the work conditions that we reproduce and study in the lab. Second, our program is design research; it is not a series of isolated psychology experiments on specific research questions. The prototypes are not mere experimental instruments to answer research questions for basic research. Instead, these prototypes are concrete research questions in themselves [Zimmermann et al. 2007] and specifically apply to the context of cooperation in emergency management planning.

3. DESIGN RATIONALE: FUNCTIONALITIES TO SUPPORT KNOWELDGE-SHARING AND ACTIVITY AWARENESS

The design rationale for the software tools was grounded in both our fieldwork [Schafer et al. 2007, 2008] and prior research on group cognition [Convertino et al. 2005]. The following are three explicit design choices we made based on our general intention to enhance knowledge-sharing and activity-awareness and our specific objectives for supporting distributed EMP.

3.1. Dual-View Map for Strategic Knowledge Sharing

During our field study, we observed that the emergency planning professionals used one large map in the middle of the table to share and manipulate information. Thus, the map served as the location for the aggregation and assimilation of knowledge for each individual collaborator as well as the group as a whole. On the one hand, this led to a large amount of information being shared on the map but on the other, it led to the need to continually negotiate coordination and sharing strategies. We wanted to reduce the need to explicitly share information by allowing the groups to share strategic knowledge more effectively.

Based on this premise, and in order to provide support for knowledge-sharing and activity awareness, we introduced the following features.

(1) A prominent distinction between role (private) and group (public) spaces. The primary function of the system is to coordinate multiple unshared role-specific views of the map with one shared group view of the map. Kraut and collaborators (e.g., Kraut et al. [2002]) found that a shared visual space improves not only communication efficiency (i.e., content management), but also the knowledge of the task structure and situation awareness (useful to managing the process), especially in complex problem solving tasks. However, there was far too much information to be understood by each member of the group without some form of organization and continual manipulation external to the shared space of the group. The dual-view map provides a working space for individuals to then choose what information to share at the optimal point in the discussion process. In addition, by constantly presenting to group members the contrast of private, role-specific versus shared, general information views, we hoped that the software would reinforce the importance of role-specific information, instilling in members a more proactive approach to their responsibilities in the task (which helps to counterbalance the systematic bias of group members for knowledge in common; see Stewart and Stasser [1995]). This dual-view exists in the paper prototype as well as the two software prototypes.

(2) We ensured that there was a low-cost but explicit sharing mechanism for transferring selected information from the personal, role-specific map to the public, group map through the use of a "Copy-To" button. This feature provides a lightweight punctuation to the act of sharing—it gives each member explicit control over their acts of sharing and it emphasizes the distinct goals and responsibilities among the members while their information is deliberately added to the shared map. This mechanism for sharing exists only in the two software prototypes.

3.2. Clear Role Indicators

We introduced the feature of *role indicators* to communicate individual group member's actions on the shared map. By identifying all shared information in two ways (during the act of sharing and afterwards), we hoped that members would be continually reminded that different group members (and thus different roles) were contributing different types of information in addition to any specific information that had been contributed by each of their partners during discussion. This should help the group to generate a sense of who knows what that is useful in organizing and managing their shared information space. These mechanisms for role-specific sharing exist in all three prototypes; however, these mechanisms are supported more in the two software prototypes than in the paper prototype. In the software prototypes, the map actions associated with particular group members are reinforced through visual cues that indicate their roles (e.g., color and label indicate the role while each role selects shared objects, move the telepointer, etc.). In addition, the second software prototype displays the roles that are associated with the annotations in the Annotation Browser.

3.3. Reviewable History of Activity

As Clark and Brennan [1991] suggested, forms of communication media that enable reviewability and revisability of shared content (e.g., text or drawings) should better support the formation of common ground, an integral part of knowledge-sharing and activity-awareness. This enables the group members to explicate and review traditionally tacit aspects of individual and group activity and make such information visible and permanent throughout the collaborative activities. The history and record of the shared work provides selected aspects of past activities as a visible and tangible resource to group members, relieving some of the weight placed on members' short term memories [Clark 1996; Convertino et al. 2008a, 2009]. Although all three prototypes provided some support for reviewability and revisability, the annotated notes on the geographic maps of the paper and first software prototypes did not provide any context beyond their location and role-specific color. For the second software prototype, we added the Annotation Browser (AB) to augment the earlier prototype and better support reviewability and revisability.

4. PROTOTYPES

4.1. Paper Prototype

In devising the paper prototype we sought to balance the simulation of the tabletop exercises observed in the field with constructing a paper version of the new design we were considering for the collaborative software tool. So, for example, because we planned to build a software system with multiple, coordinated map-based views, we gave group members playing each role their own individual role-specific maps. The role-specific maps presented features that were specific to a particular role player's understanding of the area (e.g., the environmental expert had weather information and the public works expert had utility information). For collaboration and coordination, a group map with only the basic, shared map features was provided.

The map data layers used in role-specific and team maps were based on real GIS map data. The map data layers were adapted from official GIS map data of Centre County region (Pennsylvania, USA).[1] We rotated and relabeled towns and other features of the maps to control for prior experience (no participant reported recognizing locations in the maps). In addition to the maps, each group member (one per expert role) was provided a different colored pad of Post-its and a different colored pen as tools for annotation and information sharing.

For the study, as shown in Figure 1, our laboratory was configured with three tables at right angles to one another. This provided an individual working area for each participant (for role-specific map and written materials) as well as a common tabletop working area at the intersection of the three tables (group map).

4.2. First Software Prototype

Our first software prototype featured a group map and a role-specific map placed side-by-side (Figure 2). Each map displayed multiple layers of authentic geographic data: the same maps were used in the paper prototype study. For example, the group map was a shared object that was used collaboratively by all the group members. The role-specific maps contained unshared data layers that were used privately by each user. In addition, the role-specific maps shared common data layers with the group map (shared data). A toolbar had tools for navigation (e.g., zoom-in and zoom-out) and tools for annotation (e.g., add a text note, scribble tool, and various conventional symbols).

Four new features existed in the first software prototype, which were impossible in the paper-based system (Figure 3 and Figure 2).

(1) *Copy-to Button*. Allowed for sharing annotations quickly and seamlessly from the individual view to the group view;

[1] http://www.co.centre.pa.us/gis/.

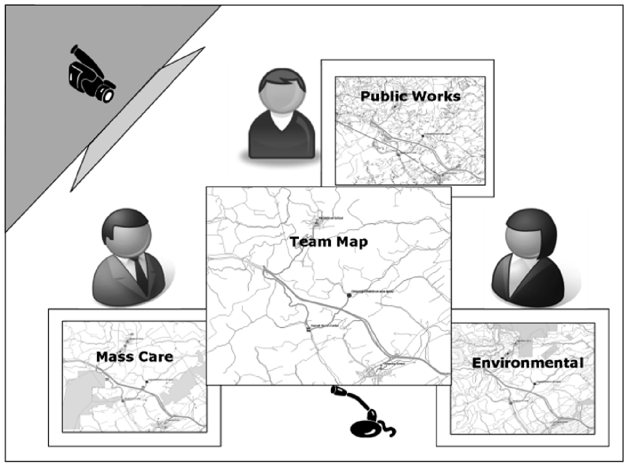

Fig. 1. Paper-prototype experimental setup.

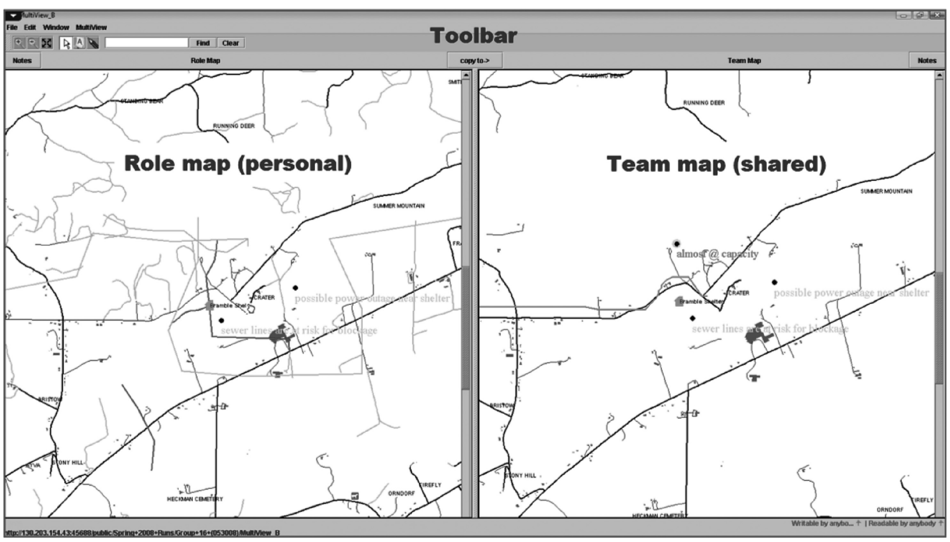

Fig. 2. First software prototype: User interface. It includes the toolbar at the top with tools for navigation and annotation, the role map on the left showing a portion of the Power Works role map, with role-specific information (e.g., damaged power lines indicated in red) and personal annotations, and the team map on the right with shared (color-coded) annotations from two roles.

(2) *Dual-pointer*. Provided ease of coordination between the individual view and the group view for each group member;
(3) *Telepointer*. Provided a constant indicator of role.
(4) *Role-affirming Annotations*. Since the annotations were the same color as the role related telepointer, this was a constant reminder of role-related knowledge.

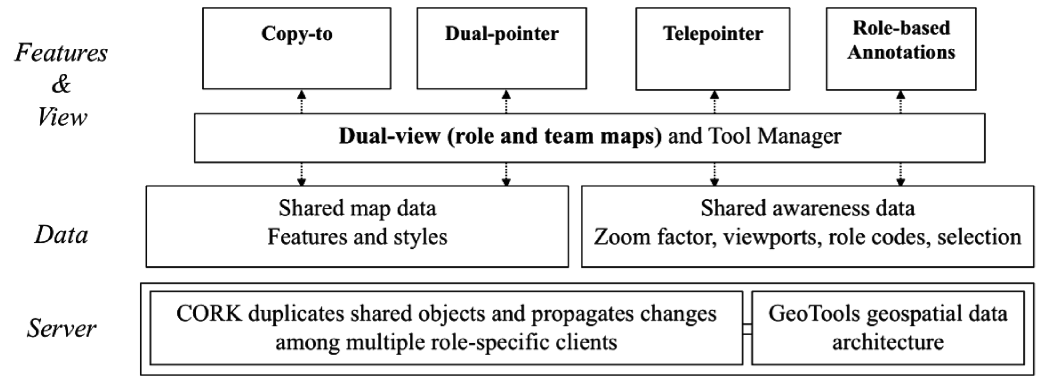

Fig. 3. First software prototype: Architecture (view, data, and server layers).

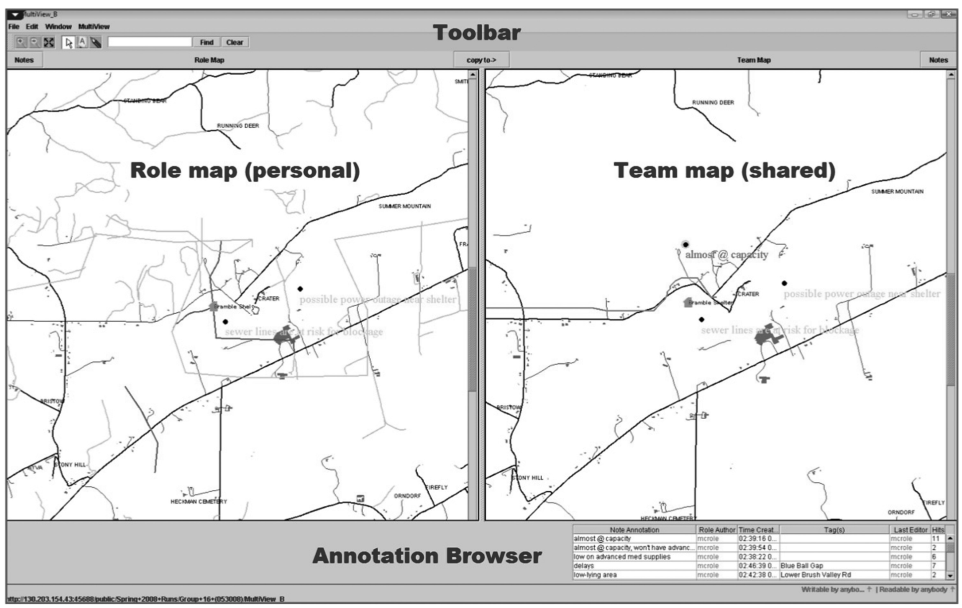

Fig. 4. Second software prototype: User interface with the Annotation Browser. As in Figure 2, it includes the toolbar at the top, the role map on the left, the team map on the right, but also the Annotation Browser at the bottom, which shows the content and various metadata of each annotation in a sortable table.

4.3. Second Software Prototype

Our second software prototype leveraged the lessons learned from the paper and first software prototypes. In both of the previous prototypes, the notes (e.g., text annotations or marks) added to the geographic map did not provide any context beyond their location and role-specific color. For this study, while using the same maps, we added the Annotation Browser (AB) to augment the earlier prototype and better support the development of common ground.

The AB presents the annotations in a tabular format (Figure 4, lower right hand corner). Each row in the table represents a note annotation (as opposed to line annotation or a scribble). The columns indicate the text (Note Annotation), note creator (Role Author), timestamp, tags, last editor of the note, and frequency of note selections on the map (Hits). The note author and last editor columns hold the role of the group member who created or edited the note; we chose to emphasize roles rather than person names because it is the emergency planning roles that are associated with specialized knowledge or decision strategies. The colors assigned to roles in the AB are identical to those used on the map for telepointers and map annotations. The reviewable and revisable columns that display each note's text and tags give the collaborators immediate access to the content of the note annotations and are editable within the table; this offers an additional mechanism for updating shared information. Also, to support reviewability, as the user clicks on an annotation in the AB the system scrolls the team map to display the portion of the map that contains that annotation. Semantically, the creation time and hits contextualize the notes with respect to when they were introduced into the collaboration and when they became a focus for more work. Moreover, in terms of procedural knowledge, the reviewable columns of creator, last editor, creation time, and hits explicate aspects of the interactions with the data. This supports the mutual awareness of who has edited and contributed what note and what information has been added or changed over time.

About the simple style of the maps used, it should be noted they were maps that were obtained from municipal governments and are maps that EMP groups really use, as opposed to the very best designs cartographers could provide. Our main focus was not on improving geo-collaborative tools or cartography. Instead, we were interested in giving groups realistic maps of a region such as the Centre County region so that we could study knowledge-sharing and activity-awareness during emergency management planning. Thus the maps used may not be the best possible map designs, but they are representative of real world resources and practices.

5. RESEARCH QUESTIONS AND HYPOTHESES

For each of the three conditions (paper prototype, first, and second software systems; we refer to this variable as medium), a different group of participants was recruited. For convenience, in the following sections we refer to the groups working with the three mediums as the PAPER, SW1, and SW2 groups.

In contrast to the PAPER groups, both the SW1 and SW2 groups were faced with the challenge of learning a new piece of software (the collaborative tool) and working remotely. On the other hand, the tools used by the software groups had extra computer-based features that we hypothesized could better support knowledge-sharing and activity-awareness over that of the PAPER groups. Moreover, the addition of the annotation browser in SW2 was intended to further help the distributed group members with reviewing, synthesizing, and revising the shared information. The annotation browser should also facilitate knowledge-sharing and activity-awareness in comparison to the PAPER and SW1 conditions, thus helping the SW2 groups to overcome some of the costs of being distributed that were faced by the SW1 groups. In summary, two general research questions guided our formulation of more specific hypotheses.

RQ1. What setting-dependent decrements occur in the quality of communication of SW1 and SW2 groups due to the extra process costs imposed on their members by the distributed setting, compared to the PAPER groups, who worked in a collocated setting?
RQ2. What medium-dependent increments occur in the quality of knowledge-sharing and performance of SW1 and SW2 groups (software medium) over PAPER groups (paper medium) and of SW2 groups (enhanced software medium) over SW1 groups?

Based on prior evidence [Clark and Brennan 1991], we hypothesize that extra communication costs will be observed when groups are in a distributed rather than a face-to-face setting (an effect of setting). At the same time we hypothesize that the features introduced in the first software prototype will offset some of the extra costs due to setting; specifically, those that address knowledge-sharing and activity-awareness.

Furthermore, with the additional features introduced in the second software prototype, we expect any effect of setting on communication to disappear due to our greater support for knowledge-sharing and activity-awareness. Thus, we hypothesize the following improvements from condition to condition.

H1. An increment in costs is associated with the communication process between the PAPER groups and the SW1 groups.
H2. An increment in knowledge-sharing and activity-awareness occurs over the three studies.
  H2a. between the PAPER groups and the SW1 groups; and
  H2b. between the SW1 groups and the SW2 groups.

Finally, we expect that the improvements in knowledge-sharing and activity-awareness building as stated in H2 will lead to correlated changes in the average performance of the groups across the conditions, suggesting that increased knowledge-sharing and activity-awareness may help to compensate for the communication costs that come from the distributed setting.

H3. An increment in performance occurs over the three studies:
  H3a. between the PAPER groups and the SW1 groups; and
  H3b. between the SW1 groups and the SW2 groups.

It is worth noting that our study, while it was conducted in a lab, was a valid model of the real situation in the field. Great effort was put into ensuring that the research questions were relevant to problems experienced by real EMP groups and that the experimental conditions were representative of the real work situation. The field study by Schafer and collaborators [Schafer et al. 2007, 2008] had identified plans and information issues that were quite analogous to those we studied in our experiments. This prior field study had pointed to the need for better tools that could improve the quality of knowledge-sharing and activity-awareness, and as a result, the coordination and performance of the EMP groups. This led us to focus on the listed hypotheses. Moreover, the lab study adopted task and materials that were representative of those observed in the field (see the following for more details).

6. EXPERIMENTAL DESIGN

To evaluate our design rationale, we first designed a reference task that would enable us to study complex emergency planning processes in a controlled laboratory setting. We developed a simulated version of the tabletop exercise as our reference task to be used in our lab studies of emergency management planning.

A key issue in the design of the study was balancing between ecological validity with respect to field observations of emergency planners (roles they play, information they manage, how they use maps, etc.) and the systematic manipulation or control of factors like access to information in a laboratory setting. To make the reference task realistic, we developed the three roles from the FEMA emergency planning roles (DHS-FEMA 1997) and the Red Cross Rescue Scenario, a notional scenario on a multi-expert evacuation operation by the US Office of Naval Research [Carroll et al. 2007]. In addition, the maps used in the task were created adapting real GIS maps of the Centre County region. Considerable research effort was devoted to the design and specification of the task structure and content through an extensive process of pilot testing (6 pilot groups) and revision.

6.1. Emergency Management Roles

We designed the experiment to study small groups of three members, each playing distinct and stable roles (Public Works, Environmental, and Mass Care expert). Each role contributed unique expertise and information—all of which was required to perform the collective task successfully.

Because the participants were not actual emergency management planning experts, we developed procedures designed to impose the beliefs and knowledge that an expert would have. Each participant was given a detailed description about his/her role and role-specific background information. The role description included three parts.

(1) A list of items the expert knew about and that were displayed on the role-specific maps (e.g., the Public Works expert knew about roads and bridges and utilities);
(2) Problems that typically arise about each of the items s/he was responsible for (e.g., downed power lines: sparks from wires can trigger fires);
(3) An example illustrating how, given a problem, the expert would use his/her knowledge to evaluate tradeoffs and find a solution (e.g., Choose the route through the north valley because it would be safer in the storm).

It is important to note that our empirical studies were designed to systematically investigate whether the software design features would increase knowledge in common and improve communication processes over time. Thus, the use of novice emergency planners not only reduced costs of experimentation, but also controlled for individual differences that would be introduced by differing levels and sources of expertise. We wanted to avoid the problem of prior experience changing the weight of the cons or introducing more variance and noise into the experimental design.

6.2. Emergency Management Task

The experimental task required each group to create three different plans over three different emergency planning problems; in our experimental design, we view these as repeated trials and refer to them as runs. Members played the same roles across the three runs. The task scenarios for each run were similar in that there was a family in need of rescue in a floodplain, there were four possible shelters, only one route to each, and the same amount of information was provided.

The experimental task we used adapted the “hidden profile” paradigm of Stasser and Titus [2003] developed to study information-sharing within group decision-making. A hidden profile is embedded into the data provided to each participant; that is, an overall superior decision alternative exists but is hidden from individual members, each of whom is biased toward suboptimal solutions. The superior alternative can be discovered only if the members effectively pool relevant individual knowledge and reason as a group. This creates a logical dependence of the quality of the group decision on the quality of group sharing. It also allows measuring if (and how often) group members disanchor from their initial preferences to choose the group-level optimal solution. However, the participants were not made aware of the existence of an optimal solution, and thus they perceived the problem before them as sufficiently wicked.

In our task, the key information items are risks or constraints (*cons*) pertaining to each of four possible solutions—four shelters to which the family could be evacuated. Example cons are “Garbrick St. goes through a floodplain” and “The Brown Clinic Shelter holds 30 people and currently has 25 people”. Each member of the group is provided with 9 cons; each individual’s set of cons are biased toward choosing one shelter—having only one con—and against another shelter—having four cons—with the remaining two shelters having an intermediate number of cons—two cons each—(Table I). At the individual level, each group member would thus advocate for one shelter and against another. As shown in Table I, we allocated cons for three of four shelters (rows A, B, and C) according to a 3×3 Latin Square.

The fourth shelter in our task design has the hidden profile. Each group member is given one distinct con for that shelter, and one shared con for that shelter. From each

Table I. Cons Distribution Matrix

		Public Works	Environment	Mass Care	Total Cons
Shelter A	unshared	1	2	4	7
Shelter B	unshared	4	1	2	7
Shelter C	unshared	2	4	1	7
Shelter D	unshared	1	1	1	4
	shared	1	1	1	

individual's perspective, the fourth shelter has two cons, neither the worst nor the best alternative (see three intermediate columns and row D in Table I). But if the group effectively pools relevant information across roles, then that shelter is clearly the best alternative: it has only four total cons versus seven each for the other three shelters (see rightmost column and row D in Table I).

It is important to note that from the perspective of the study participants, as for real EMP group members, the group task did not have a single correct solution but multiple solutions had to be compared and prioritized. The group had to choose the best shelter and then rank two alternative solutions in order of preference. The task is not designed to measure the success of the team in finding the single correct solution (as for an intellectual task such as the Tower of Hanoi). It was designed to measure the efficiency of the group in sharing knowledge and gaining awareness. It also allowed measuring how this efficiency impacted the performance (the optimality of the ranked set of solutions chosen), since the experimenters had designed the alternative solutions with this property. The task was challenging, so that the expected performance would typically be far from optimal. The group members had to overcome the hidden profile effect, but also successfully integrate quite a bit of information, just to attain the rather so-so performance levels they had.

6.3. Participants

In total, across the paper prototype study and the two software prototype studies, 156 participants were recruited from a university located in the northeast part of the United States. In each study the participants were grouped into three-person groups: 12 groups, or 36 participants, in the PAPER study and 20 groups, or 60 participants, in each of the two software studies.

To encourage equal participation and avoid male dominance [Malz and Borker 2007], we created same-gender groups (except for one of the 52 groups). In the PAPER study, there were six male and five female groups, and one mixed-gender group with one female and two males. In each of the two software studies, there were ten male and ten female groups.

The participants' ages ranged from 20 to 45. All tasks were performed in English and all participants were fluent English speakers. The non-native speakers were less than 30% in each of the studies. About 68–70% had undergraduate college degrees and the remaining had graduate degrees. About 60–65% reported to be students, 20–25% were workers, and the others were unspecified. The participants had little prior experience with EMP or related operations.

In order to be able to consider additional control variables, at the beginning of the study each participant filled out a background questionnaires rating on Likert scales for the following variables: level of comfort with using maps on paper and maps on a computer; level of expertise with relevant tools (video conferencing, audio conferencing, interactive maps, GPS/GIS/satellite imagery tools); level of expertise for the emergency management domain and each of the three roles of the experimental task (Public Works, Environmental, Mass Care); her/his preferences for working in a group alone, via a computer, and on paper. These measures were collected to control for possible unusual individuals (outliers) across and within experimental conditions. Participants also completed the Santa Barbara Sense-of-Direction (SBSOD) scale [Hegarty et al. 2002], a validated self-report measure of spatial skills; the Metacognition Awareness Inventory (MAI), a measure of individuals' knowledge and regulation of their own cognition; and the Adjective Check List (ACL), a scale assessing personality factors like agreeableness and extroversion (see also Carroll et al. [2008]; Convertino and Carroll [2010]). No significant differences were found across the experimental conditions. For example, for the background questionnaire, the differences between the average ratings across the samples in the three conditions were even smaller than within-sample standard deviations. The average measures of spatial skills obtained from the SBSOD scale were not different between the three samples (this was confirmed by a t-test between pairs of samples). Since none of these control variables differed across the samples, they were not included in the final analysis.

6.4. Procedure

For the PAPER groups, each participant was given an assigned place at the joined table. For SW1 and SW2, participants were seated at one of three different workstations in three adjoining rooms. At each workstation, a participant had a Dell Optiplex with a 19" widescreen LCD monitor, a microphone, and set of speakers for verbal communication among group members and with the experimenters. For all three studies, the participants read descriptions of their individual roles and the task scenario and read the role-specific information sheet relating each piece of information to their role-specific map.

At this point, the participants began to collaborate on the planning task. The information given to the participants was presented in three different forms: individual role-specific maps, information sheets with role-specific information, and a shared scenario with background information. Participants were instructed to share information as needed with the group, by copying information onto the shared map. When they reached a decision, they were to write down the final plan along with three alternatives in order of preference in a final plan document. Groups were given about 20 minutes to complete this task. After the task, participants completed, (1) a questionnaire that asked them to rate the quality of the groups' process and performance; and (2) a set of open-ended questions that assessed their recall of both the solutions generated, and the information considered for each solution (i.e., cons). This process was repeated three times (we refer to this as three repeated runs), with new scenarios and information presented each time. The order of the scenarios used was counterbalanced across the groups.

7. DATA TYPES COLLECTED AND ANALYZED

7.1. Recall of Task Information

A first set of measures assessed group members' retention of task-relevant information. At the beginning of each task session (i.e., each of the three runs), the group members worked individually. They considered how each con they had been given was related to their role-specific map and judged its relevance on a 5-point scale. The individual ratings of each member were analyzed to keep track of what cons had been acknowledged individually and judged relevant. Then, in the second part of the task session or run, the members shared any con they considered relevant and co-constructed the rescue plan on the shared map. We transcribed the group's verbal interactions, and for each turn of transcribed conversation, we recorded whether a con was mentioned, by noting the unique identifier of that con from the list of cons in the reference task materials.

The results of this analysis were used to assess which cons had surfaced to the level of group discussion, and later if these cons were recalled after the task session.

At the end of each task session or run, the members filled out a post-task survey. The survey was used to assess the recall of the three best solutions chosen along with the recall of the cons discussed in relation to each of these solutions. The expectation for the recall measures is that, as knowledge-sharing and activity-awareness improve, the members will recall more about the solutions discussed, including the cons shared by the group (e.g., Monk et al. [1996] and McCarthy et al. [1991]). Therefore we measured the recall of the shelters chosen (the first, second, and third best shelters) and how many available cons the group members remembered after each task session (for more details on the recall measures see [Convertino et al. 2007]).

7.2. Communication Structure

A second set of measures assessed the turn-taking structure of communication. As knowledge-sharing and activity-awareness improve, communication becomes more efficient, because the shared understanding means that conversation topics need less introduction or clarification. Conversational turns occur more rapidly and utterances are more compact (e.g., more turns, fewer words, more synchronicity). Researchers have relied on these measures to assess the effects of different communication settings on communication efficiency and amount of common ground (e.g., Sellen [1995]; Sanford et al. [2003]). We expected to observe similar trends, with our groups becoming more efficient in the turn-taking structure of communication as knowledge-sharing and activity-awareness improved.

We transcribed the communication records of the groups during Runs 1 and 3. We adapted the analysis scheme used by Sellen [1995], which breaks a dialogue into turns and pauses. It also codes simultaneous speech, including speech that causes an interruption (SSI, taking the floor), and noninterruptive simultaneous speech (SSNI). We compared counts and durations for runs 1 and 3, normalizing for the length of the run (see Convertino et al. [2008a, 2009]).

7.3. Communication Content: Dialogue Acts

A third set of measures assessed changes in the types of dialog acts used, as knowledge-sharing and activity-awareness improved. Prior research has contrasted specific categories of dialog acts (e.g., changes in the proportions) to measure changes in the common ground building process (e.g., Sanford et al. [2003]).

We categorized the content of the transcripts, using an adaptation [Convertino et al. 2008a, 2009] of the Conversation Game Analysis method [Anderson et al. 1991; Sanford et al. 2003]. This scheme classifies the communicative functions of dialogue acts (what the speaker is trying to achieve) rather than their linguistic form or meaning.

We chose this coding method for two reasons. First, the method has previously been applied and validated specifically for dialogs of groups or dyads performing map-based collaborative tasks [Anderson et al. 1991; Carletta et al. 1997; Sanford et al. 2003]. Second, previous studies have shown that changes in the proportion of some categories of dialog acts—categories such as Check, Align, and Query—are associated with changes in the common ground building process. For example, Doherty-Sneddon et al. [1997] observed a significant increment in the proportion of Align acts with participants using the audio-conferencing medium compared to those using the video-conferencing medium. Analysis of the conversations suggested that the Align dialog acts were being used as substitutes for some nonverbal forms of alignment that are missing in audio conferencing. Similarly, Sanford et al. [2003] found significant increments in the proportion of Align and Query acts in video-mediated communications on maps, with respect to comparable face-to-face communications. This was seen as evidence of the

Table II. Dialogue Act Codes and their Descriptions

Class	Dialogue Act	Description
Transfer Info (to share)	Add Info (AI)	Provides new information, not elicited.
	Query (Q)	Question used to elicit new information.
	Reply (R)	Reply to query to provide new information.
Check Understanding (to coordinate sharing)	Check (CH)	Verify *own understanding* of information *previously presented* by others.
	Align (AL)	Verify *partner's understanding* of information *previously presented* to others.
	Clarify (CL)	Clarifies or restates information already presented.
	Acknowledge (AC)	Signals receipt of information, understanding.
Manage Process & Decision (to perform task)	Manage (MN)	Instruction, command, direct, or indirect request for action; orchestrating strategy, how to do the work.
	Summarize (SA)	Summarizes information previously presented.
	Judge (J)	Individual judgment, opinion, or preference.
	Confirm (CO)	Requests partners' agreement on a proposed decision.
	Agree (AG)	Indicates approval for a prior judgment or decision.

extra effort imposed by video as a medium for building common ground. Compared to the face-to-face participants, the video-mediated participants "spent more time and effort establishing common ground with their partners" [Sanford et al. 2003, p. 1084].

Working from findings such as these, we adapted the Conversation Game Analysis coding method for our task domain. This was necessary because the prior studies had analyzed communication tasks on maps [Anderson et al. 1991; Sanford et al. 2003], while our tasks required both communication and decision-making [Convertino et al. 2008a]. Table II summarizes the codes and descriptions. We distinguish three subsets of dialog acts, based on their contribution to the collaborative work: the acts aimed at transferring information (AI, Q, R); those aimed at building and verifying common understanding or common ground (CH, AL, CL, AC—see the 'action ladder' in Monk [2003]); and those aimed at managing the process and content (MN, SA), and making the decision (J, CO, AG).

Two coders were trained in parallel to use the coding scheme. For better reliability they coded the dialogues separately while reviewing the video of the interaction together. For every few minutes of a session, they viewed the video and coded the transcript, then compared their codes. At each review step they negotiated and agreed about any conflicting codes, referring to the coding scheme. Samples of coded transcripts were compared to check the intercoder agreement. On average the intercoder conferral was about 80% based on Cohen's Kappa (about 82% for the paper prototype study, about 75–80% for the first and second software prototype studies). The coders were coached directly by the researcher and one of them served in this role across all three conditions. This setup ensured consistency in the coders' application of the coding scheme across the three sets of data.

7.4. Post-Task Questionnaire

Finally, we used a questionnaire that produced seven self-reported indices of group process (gain of shared knowledge; quality of communication; communication means; understanding and expression; ease of referencing and planning; interpersonal awareness; and awareness over time) and two indices for perceived group product (group performance and satisfaction). We administered the questionnaire after each run, resulting in three sets of nine scores per participant for each study (see Convertino et al. [2008a, 2009]).

Table III. Measures of Structure of Communication for the Three Study Conditions. In the Rightmost Column are Indicated the Significant ($p < .05$) Effects of Study (**), Run (*), Study X Run (***), or *P, *S2,*S2 for Effects of Run in the PAPER, SW1, or SW2 Condition Only

	PAPER (N = 12)		SW1 (N = 16)		SW2 (N = 20)		
Measures	R1	R3	R1	R3	R1	R3	Effect
#Turns	381	312	155	118	242	171	**, *
#Words/turn	**8.6**	**7.0**	**10.9**	**9.6**	**8.6**	**8.2**	***,**, *
#Turns/minute	**17.3**	**20.7**	**8.0**	**10.4**	**15.1**	**20.5**	**,*
#SIT/minute	4.4	6.1	2.5	3.5	5.1	6.6	**, *
#LIT/minute	6.7	7.1	4.3	5.3	6.2	7.5	
#IT/Turns	63%	64%	87%	85%	76%	73%	**,*S2
#LIT/Turns	40%	35%	56%	51%	42%	41%	**,*
#SIT/Turns	26%	30%	31%	34%	33%	32%	*P
#SS/Turns	**34%**	**32%**	**13%**	**14%**	**20%**	**22.5%**	***,**,*S2
#SSI/Turns	11%	12%	4.5%	6.6%	8.5%	9.9%	**,*S1*S2
#SSNI/Turns	23%	20%	8.5%	7.8%	11.2%	12.6%	***,*P

7.5. Capture Software

The interaction of the distributed groups via the software prototype was recorded through tools provided by Noldus Information Technology:[2] uLog 2.0 (with Camtasia by TechSmith) to log keystroke-level events and The Observer XT 7.0 to integrate the multiple data types (keystroke-level logs and videos) for each group and run.

8. RESULTS

In this section we present the findings in relation to each of the three hypotheses that we presented in Section 5. We report findings from various comparisons: within each study condition, between PAPER and SW1, or between SW1 and SW2, and among all three conditions.

8.1. Extra Communication Costs due to the Distributed Setting

For H1, we hypothesized that communication costs will inevitably increase as we move from a collocated setting to a distributed one. In support of H1, we found various consistent indicators of extra communication costs in the SW1 groups compared to the PAPER groups.

First, with respect to differences in the structure of communication, we found a lower frequency of speaker turns in the SW1 groups (9.2 turns per minute, on average across the three runs) than in the PAPER groups (19 turns per minutes, on average across the three runs), and more words per turn in the SW1 groups (10.3 words per turn, on average across the three runs) than in the PAPER groups (7.8 words per turn, on average across the three runs); results of MANOVA confirmed statistical significance of these findings, $F[2,25] = 32.9$, $p < .001$. Moreover, we observed that the proportion of simultaneous speech (SS) turns was significantly greater in the PAPER (33% of all the turns were SS turns) than in the SW1 groups (14%). Note that prior studies have already used these measures as measures of efficiency in communication [Sellen 1995, Sanford et al. 2003]. Table III summarizes the results about structure of communication for run 1 and run 3 in the three conditions.

Second, with respect to differences in the proportions of dialog acts, we found more acts for explicitly managing the work process in the SW1 groups than in the PAPER groups (ANOVA $F[1, 26] = 7.0$, $p < .01$; see manage acts (MN) in Table II and the results in Table IV). Slightly more acts for summarizing the content were also visible in the SW1 groups (especially in the first run) than in the PAPER groups (see Table IV).

[2]www.noldus.com/human-behavior-research.

Table IV. Effects on Dialog Acts (Average % per Run) for the Three Study Conditions. In the Rightmost Column are Indicated the Significant ($p < .05$) Effects of Study (**), Run (*), and Study X Run (***). The Number of Groups Considered (N Values) is the Subset of Groups that had their Transcripts Fully Transcribed and Coded

	PAPER (N = 12)		SW1 (N = 16)		SW2 (N = 20)		
Dialog acts	R1	R3	R1	R3	R1	R3	Effect
Add Info (AI)	**7.8%**	**9.2%**	**9.5%**	**12.2%**	**11.3%**	**13.2%**	*, **
Queries (Q)	5.6%	4.9%	4.9%	3.6%	2.0%	1.0%	*, **
Replies (R)	**4.8%**	**4.0%**	**3.7%**	**2.3%**	**2.2%**	**1.0%**	*, **
Checking Understan. (CH, AL, CL, AC)	**36.2%**	**41.3%**	**33.7%**	**34.4%**	**50.1%**	**48.5%**	**
Judge (J)	20.2%	19.5%	21.8%	23.8%	16.3%	20.0%	*, ***
Manage (MN)	**7.9%**	**5.5%**	**10.6%**	**9.4%**	**9.8%**	**7.6%**	**
Summarize (SA)	**4.8%**	**4.7%**	**6.9%**	**3.2%**	**2.1%**	**1.2%**	**
Confirm (CO)	2.2%	2.1%	2.0%	2.0%	1.2%	1.5%	
Agree (AG)	**10.6%**	**8.8%**	**7.0%**	**9.1%**	**5.1%**	**5.9%**	**, ***

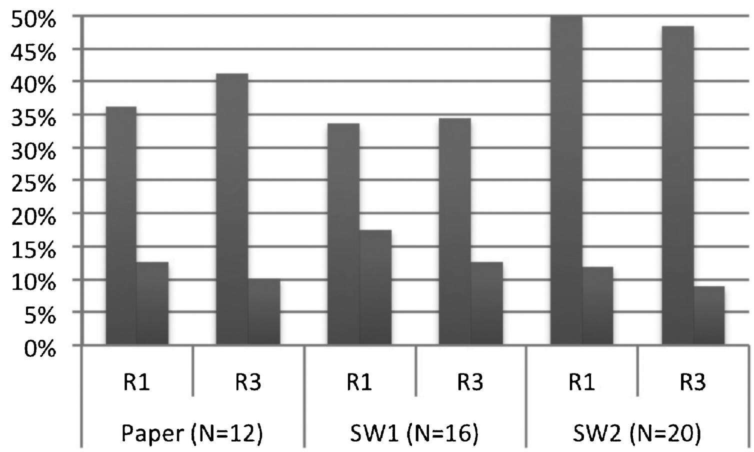

Fig. 5. Differences in dialog acts: Checking acts (blue bar, sum of CH, AL, CL, AC) and management and summarize acts (red bar, sum of MN and SA). Average percentage of acts per run (runs 1 and 3) in the three study conditions.

Thus, extra process costs were required for managing the process and content (MN, SA) in the distributed setting. This finding extends prior work from computer-mediated communication research (e.g., Clark and Brennan [1991], Sanford et al. [2003]) to the case of groups that perform complex decision-making on maps. While more acts are used to manage the process or summarize the content, less time is left for acts aimed at building and verifying mutual understanding in the SW1 groups than in the PAPER groups. In fact, 34% (SW1) versus 39% (PAPER) of these acts were exhibited on average per run if we consider the check (CH), align (AL), clarify (CL), and acknowledge (AC) acts together (see Table IV and Figure 5). These acts indicate the throughput in the process of establishing and maintaining shared understanding.

Third, as a converging measure related to these slower turn-taking rates and smaller throughput in checking understanding of the SW1 groups, the analysis of the information discussed by the group (analysis of verbal transcripts) revealed that the total number of cons discussed by the SW1 groups was smaller than in the PAPER groups (SW1: 52%, sd = 23%; PAPER: 72%, sd = 16%; ANOVA $F[1,26] = 15.8$, $p < .001$) (see Figure 6).

Finally, the subjective ratings from the post-task questionnaire about quality of communication were lower in the SW1 groups than in the PAPER groups (SW1: 5.56, sd = .08; PAPER: 5.83, sd = .11; ANOVA $F[1, 90] = 5.6$, $p < .05$)) (see Figure 7).

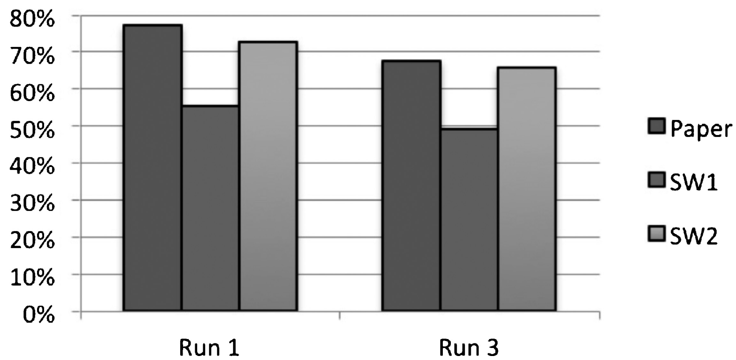

Fig. 6. % Cons discussed in the three study conditions. Average percentage per run (runs 1 and 3).

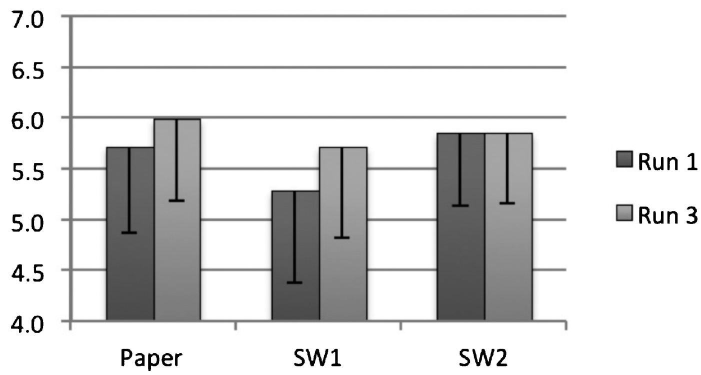

Fig. 7. Differences in the subjective measures of the quality of communication. Average rating per run (runs 1 and 3) in the three study conditions. $N = 156 = 3 \times 12$ (PAPER) $+ 3 \times 20$ (SW1) $+ 3 \times 20$ (SW2).

8.2. Improved Knowledge Sharing and Activity Awareness in the SW groups

8.2.1. Offsetting of Costs. In H2 we hypothesized that the costs predicted by H1 will have minimal bearing on the overall functioning of the SW1 groups (H2a) and in fact may be fully overcome by the improvements in knowledge-sharing and activity-awareness by SW2 groups (H2b). This hypothesis was motivated by the features of the new software medium, which we designed to support knowledge-sharing and activity-awareness. In support of H2, in particular H2b, the results show that the extra communication costs that had been observed in the SW1 groups were no longer visible in the SW2 groups.

In several of the measures assessing communication costs, the SW2 groups had values similar to the PAPER groups and better than the SW1 groups. This pattern was observed across several measures: communication structure (or turns), such as the frequency of turns and the number of words per turn (17.8 turns per minutes, 8.4 words per turn, see Table III); the number of cons discussed (SW2: 66% of cons, see Figure 6); the percentage of dialog acts for managing process and summarizing content (see Table II and the red bars in Figure 5) and the subjective ratings such as the quality of communication (SW1: 5.83, sd = .08, see Figure 7). Post hoc tests in the ANOVA analyses confirmed the statistical significance of the differences for each of these measures ($p < .05$).

However, some communication process effects of the distributed setting were still visible for the SW2 groups. These groups had values that were poorer than the PAPER groups but slightly better than the SW1 groups for the proportion of simultaneous

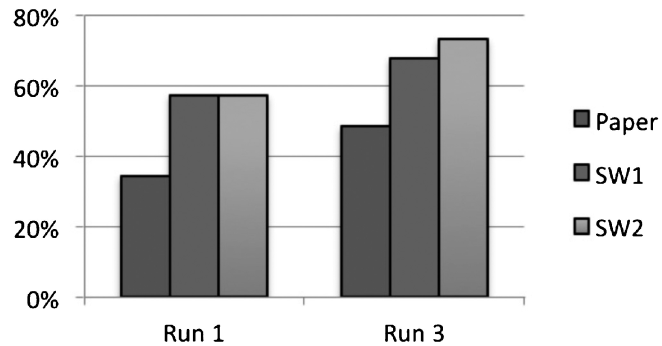

Fig. 8. % Cons recalled (runs 1, 3) in the three study conditions.

Table V. Recall of Final Shelter Choice (2nd and 3rd shelter) for the Three Study Conditions. Significant Increases Within Each Study Condition are Marked in Bold

	PAPER Prototype		First Software Prototype		Second Software Prototype	
	Run 1	Run 3	Run 1	Run 3	Run 1	Run 3
2nd Shelter						
Full	**75%**	**94%**	89%	89%	**88%**	**98%**
Partial	8%	3%	0%	3%	5%	0%
Wrong/No	17%	3%	11%	8%	7%	2%
3rd Shelter						
Full	64%	75%	79%	89%	76%	93%
Partial	3%	0%	0%	0%	6%	2%
Wrong/No	25%	8%	16%	11%	13%	5%

speech (SS) turns (communication structure, 21% of all the turns were SS turns) (see Table III).

Overall, these findings suggest that the features introduced by the SW2 prototype were able to offset most of the extra costs imposed by the distributed setting visible in the SW1 groups. Also, these features seemed to facilitate improvements in activity awareness, which we discuss in more depth in the next subsection.

8.2.2. Gains in Knowledge-Sharing and Activity-Awareness. As a counter-tendency to the communication trends across settings, we found evidence of similar or increased knowledge-sharing and activity-awareness in groups supported by software. On many indicators of knowledge-sharing and activity-awareness the SW1 groups did as well, or better than, the PAPER groups (in support of H2a). More importantly, the SW2 groups evidenced a clearer trend towards improved knowledge-sharing and activity-awareness over the course of the study (in support of H2b).

One indicator of knowledge-sharing and activity-awareness that we considered in this research was the amount of post-task recall of information. We collected multiple measures of recall. First, we measured the recall of the cons that each individual had acknowledged before starting the group discussion. For PAPER, SW1, and SW2 groups, we found that the percentage of cons recalled increased from 34%, 57%, and 57% in run 1 to 48%, 68%, and 73% in run 3, respectively. As shown in Figure 8, the recall of cons in the SW1 and SW2 groups tended to be higher than in the PAPER groups. Although the difference across the three study conditions is not significant (mostly due to large variability in recall ability among individuals), the trend in favor of the last study is consistent with the other measures of recall reported in the following (see Table V and Figure 8) (partially supporting H2a and H2b).

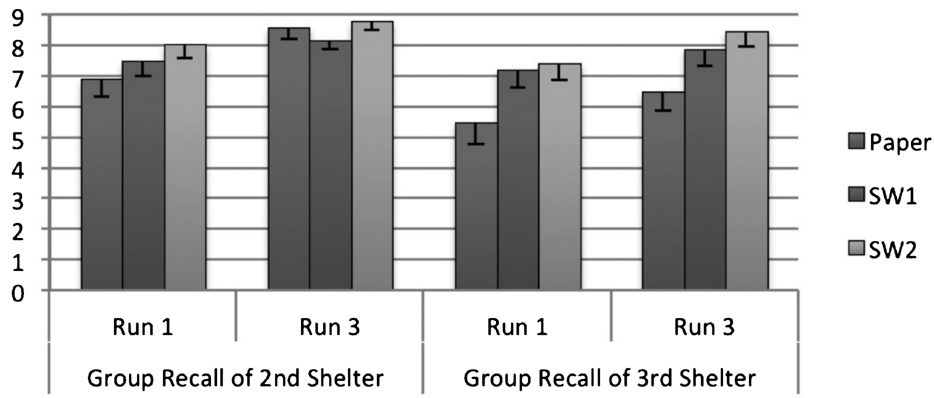

Fig. 9. Group Recall of the 2nd and 3rd shelters chosen by the group: average scores (maximum average score: 9) and standard deviation (vertical line in each bar).

We also measured the recall of the cons that were relevant to the most preferred shelter. This measure increased from run 1 to run 3 across the three study conditions. The level of recall improved at about 11% (from 32% to 43%), 10% (from 44% to 54%), and 17% (from 30% to 47%) for PAPER, SW1, and SW2 groups, respectively. These increments were significant when running a t-test for each study condition (PAPER: $t[11] = -2.3$, $p < .05$; SW1: $t[11] = -2.2$, $p < .05$; SW2: $t[11] = -2.7$, $p < .01$) and also for a repeated-measure ANOVA ($F[1, 45] = 11.33$, $p < .005$) (partially supporting H2a and H2b).

Another recall measure collected was the group members' recall of the shelters that their group had listed in their final recommendations; in other words what they proposed as the first, second, and third best shelters. Unsurprisingly, almost everyone recalled the shelter chosen as most preferred, so this measure could not reveal any differences across the study conditions. However, with respect to recall of the second and third choices we found a trend of growing recall (see Table V and Figure 9) from the PAPER, to the SW1, to the SW2 study, which is consistent with the trend in the recall of cons reported in the preceding (partially supporting H2a and H2b).

Through more robust nonparametric tests, we confirmed that the increment in the recall for second and third choices increased over the three runs; the differences were significant for both PAPER and SW2 conditions. In the PAPER groups, the recall for the second shelter is significantly higher on run 3 than on run 1 (from 75% to 94%, Wilcoxon Signed-Rank test: $p < .01$, $N = 36$) (see Table V). The recall of the third shelter has the same growing trend, although it is not statistically significant. In the SW1 groups, there was a greater percentage of recall in the first run and less evident increments from run 1 to run 3 than in the PAPER groups (see Wrong/No values in Table V). In the SW2 groups, the increments of the recall over time are again evident. They are significant for both the second and the third shelters (second shelter: 88% to 99%, third: 76 to 93%, Wilcoxon Signed-Rank test: $p < .05$, $N = 60$) (partially supporting H2a and H2b).

The percentage of cases of wrong or no recall for the second and third shelters in the two software groups was smaller than in the PAPER groups, as shown in Table V. Specifically, the likelihood ratio chi-square test indicated that the recall of the second and third shelters, during the first run, was different among PAPER, SW1, and SW2 groups: L.R. Chi-Square = 9.1, df = 4, $p < .06$, for recall of the second shelter, and L.R. Chi-Square = 8.3, df = 4, p = < .08, for recall of the third shelter (partially supporting H2a and H2b).

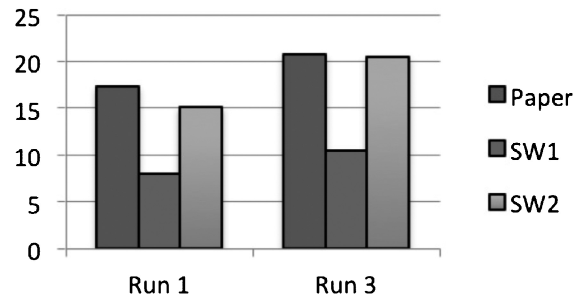

Fig. 10. Turns per minute (averages).

Finally, we computed a group-level measure of recall for the first, second, and third shelters chosen in each run. For each run, this recall measure was computed by assigning to each member a score of 3, 2, 1, and −1 for full, partial, no, and wrong recall, respectively (we coded as partial recall, for instance, when the member recalled the map area but not the shelter name). The group-level recall measure was the sum of the scores of the individual members (this measure has not been reported in previous papers, e.g., Convertino et al [2007]). We ran a MANOVA test with the group recall measure for the second and third shelters chosen and for the cons (those that had been acknowledged by the members before the task) and found that both run and medium had significant effects (Run: $F[3,43] = 6.2$, $p < 0.001$, Study: $F[6,88] = 2.3$, $p < 0.05$) (partially supporting H2a and H2b).

A second indicator for knowledge sharing and activity-awareness was measures of communication efficiency calculated from turns. The results for these measures are summarized in Table III. In the statistical analysis, we focused on a core subset of these measures: number of words per turn (#Words/turn), number of short turns per minute (#SIT/minute, 1.5 seconds or shorter), number of long turns per minute (#LIT/minute, longer than 1.5 seconds), and global rate of simultaneous-speech turns (#SS/turns, simultaneous turns among all turns). As shown by prior studies, fewer words per turns, shorter turns, and a good level of simultaneous speech, typically denote greater communication efficiency—an indicator of knowledge-sharing and activity-awareness (see measures used in Sellen [1995]; Sanford et al. [2003]). We ran a MANOVA test with these measures as response variables and repeated runs and study conditions as independent variables. We found an effect of run (with efficiency increasing over time) and study condition, and also a modest but significant interaction effect between run and study condition. Thus, the efficiency increases over time and varies considerably between the two software conditions (greater in SW2 groups) and between PAPER and SW1 (Run: $F[4,42] = 8.36$, $p < 0.001$, Study: $F[8,86] = 5.2$, $p < 0.001$, Run x Study: $F[8, 86] = 8.36$, $p < 0.05$) (partially supporting H2a and H2b). In the following, we report in detail the results and the effects observed for each study condition and for specific measures.

For the PAPER groups, the frequency of speakers' turns increased (see Figure 10) and the frequency of words-per-turn decreased (see Figure 11) from run one to run three. More specifically, the frequency of turns increased from 17.3 to 20.7 turns per minute (repeated measures ANOVA, $F[1,11] = 8.2$, $p < 0.05$, see turns #Turns/minute in Table III). The number of words per turn decreased from 8.6 to 7.0 (repeated measures ANOVA, $F[1,11] = 22.9$, $p < 0.001$, see #Words/turn in Table III). Distinguishing between short and long speaker turns, we found that the percentage of long turns (longer than 1.5 seconds) diminished (from 40% to 35%; $p < .05$, paired t-test, see

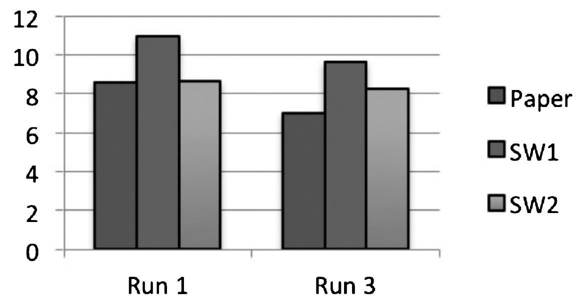

Fig. 11. Words per turn (averages).

#SIT/Turns in Table III) and the percentage of shorter turns increased (from 26% to 30%; $p < .05$, paired t-test, see #LIT/Turns in Table III).

In the SW1 and SW2 groups, similar increments occurred over the three runs to the frequency of speakers' turns and the frequency of words per turn (see Table III). The frequency of turns increased from 8 to 10.4 turns per minute in the SW1 groups (repeated measures ANOVA, $F[1,15] = 11.981$, $p < 0.005$) and from 15.1 to 20.5 turns per minute in the SW2 groups (note that in these groups, the increment does not reach statistical significance due to larger variability across the groups). The number of words per turn decreased from 10.9 to 9.6 in the SW1 groups (repeated measures ANOVA, $F[1,15] = 7.1$, $p < 0.01$) and decreased only minimally from 8.6 to 8.2 in the SW2 groups. Also, when distinguishing among different types of speaker turns for the software groups, we found that in the SW1 groups, the long speaker turns (longer than 1.5 seconds) diminished from 56% to 51% ($p < .05$, paired t-test) and the shorter speaker turns slightly increased (from 31% to 34%, with an increment that approached significance). In the SW2 groups, the longer turns remained stable around 42% (run 1) and 41% (run 3), but were overall fewer in relative frequencies when compared to the PAPER groups. The shorter speaker turns also remained relatively stable around 32-33% (partially supporting H2a and H2b).

A third indicator of knowledge-sharing and activity-awareness pertained to changes across runs and study conditions in the categories of dialog acts. The trends observed indicated steady increments in knowledge-sharing and activity-awareness development from the PAPER groups to the SW1 groups and even further for the SW2 groups (partially supporting H2a and H2b). So for example, one trend was an increment in the proportion of dialog acts for pushing information (Add Info) directly into the discussion, and a related decrement in the proportion of the dialog acts for providing information in response to pulling acts by other members (Reply). The increment was significant not only between the first and third runs of each study but also across the three study conditions. The changes became increasingly evident from PAPER to SW1, to SW2 groups (see visible trends in Figure 12 and the top three data rows of Table IV). Like other researchers who observed changes in dialog acts across different communication media (e.g., Sanford et al. [2003] found significant effects of media on the proportion of Query acts) we interpreted these patterns as evidence of accommodation to the needs of the task and the media being used to support the task. Over time, group members' activity awareness of when their information was needed improved and they began to push relevant information directly into the discussion rather than waiting for a request (an indicator of activity awareness). This makes the sharing process more efficient: a push act (Add Info) replaces two or more request and response acts (Query and Reply). The finding that these differences increased across the three media conditions suggests

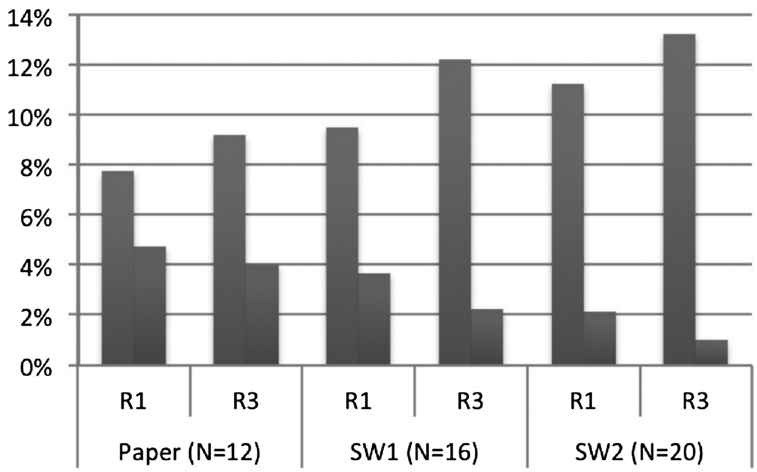

Fig. 12. Differences in dialog acts: Add Info (blue bars) vs. Reply (red bars). Average percentage of acts per run (runs 1 and 3) in the three study conditions.

that the computer-mediated tasks may have promoted more attention to efficiency in one's dialog acts.

We also found other indicative improvements in the SW2 condition compared to the two prior conditions. Despite the distributed setting, the proportion of the checking acts in the SW2 groups is similar to the PAPER groups. More importantly, the SW2 groups used significantly fewer explicit acts aimed at summarizing the previously shared information and fewer acts aimed at explicitly agreeing on judgments previously expressed (see Table IV). Post hoc tests in the ANOVA analyses confirmed that these differences were significant ($p < .05$). This suggests that the introduction of the Annotation Browser, which makes the cons shared visible in a table, may have reduced the need for explicit summarization of information and made the judgment process faster (in support of H2b). This change is consistent with the better recall of cons (previously mentioned) and the improvement in performance (see next section).

8.3. Increases in Performance Measures

Our final hypothesis, H3, addressed the effect of the development of knowledge-sharing and activity-awareness on the enhanced performance of the groups. Thus, in parallel with the hypothesized improvement in knowledge-sharing and activity-awareness development, for each new prototype, we also expect an increase in the group performance. We collected three performance-related measures—a set of subjective ratings from the questionnaire and the two objective measures of task completion time and optimality of the final decision.

We found that individuals' subjective ratings of performance and satisfaction increased significantly over the three runs across the PAPER, SW1, and SW2 conditions (Performance: $F[1,151] = 23.1$, $p < .001$; Satisfaction: $F[1,151] = 10.2$, $p < .005$). However, these ratings showed no significant increase (nor decrease) from the PAPER to the SW1 or from the SW1 to the SW2 groups. With respect to the performance ratings, we observed that SW2 groups' ratings were higher in the first run, but diminished from the first to the third runs, when compared to the ratings of both the PAPER and SW1 groups (considering the three study conditions, we find a Run X Study effect on Performance: $F[2,151] = 3.4$, $p < .05$). (Thus, while showing improvements over the three runs, the subjective ratings did not support H3a and H3b.)

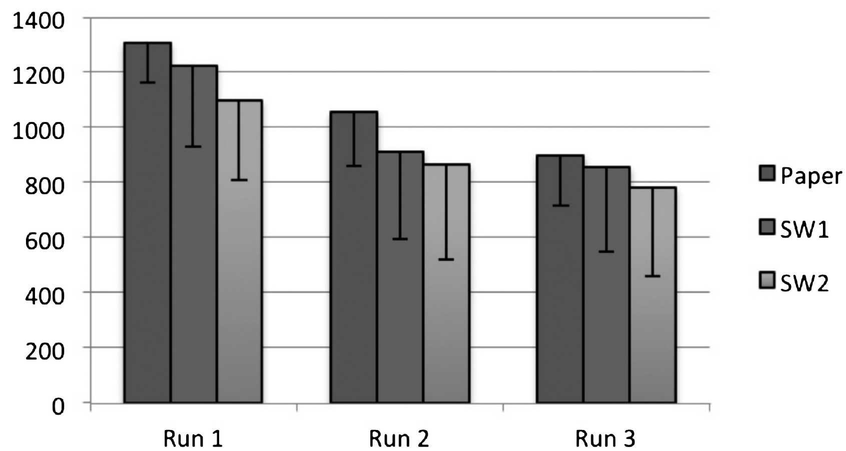

Fig. 13. Average task completion time (in seconds) for the conditions. The vertical lines in each bar indicate the per-run variability (standard deviation) across the groups.

Table VI. Effects on Performance: Completion Time (Average Seconds Per Run)

	Run 1	Run 2	Run 3	*Average per cond.*
PAPER	1306	1054	898	*1086*
SW1	1225	913	856	*998*
SW2	1098	864	780	*914*
Average per run	*1194*	*927*	*836*	

A clear pattern of differences emerged from the objective measures of performance. First, by measuring the task completion time we found that on average, the SW1 groups completed the task in less time than the PAPER groups (supporting H3a) and the SW2 groups completed the tasks in less time than the SW1 groups (supporting H3b). The improvement from the PAPER to the SW2 groups is statistically significant ($F[1,30] = 4.4$, $p < .05$). The SW1 groups exhibit a performance that is intermediate between the PAPER and the SW2 groups. The trends are clearly visible in Figure 13 and Table VI. Not surprisingly, the time to complete the task decreased significantly from run 1 to run 3 in all three conditions (PAPER: $F[1,11] = 31.3$, $p < .001$; SW1: $F[1,16] = 57.9$, $p < .001$; SW2: $F[1,19] = 30.1$, $p < .001$). A Linear Mixed Models analysis with Run (3 runs) and Study (three study conditions) as independent factors and task completion time as response variable, confirmed that Run (as repeated effect, $F[2, 50] = 37.5$, $p < .001$) and Study ($F[2, 50] = 5.2$, $p < .01$) have significant effects on task completion time.

Finally, when measuring the optimality of the final decisions of the groups (see Figure 14), we found that the PAPER groups were more likely to make optimal decisions in run 3 than run 1 (this improvement was not statistically significant). In the SW1 groups, the number of groups making an optimal decision was also greater in run 3 than run 1. More importantly, the SW1 groups in both run 1 and run 3 did markedly better than the PAPER groups. Oddly, the SW2 groups did slightly more poorly in both run 1 and run 3 than the SW1 groups although they still did better than the PAPER groups. Thus, results concerning optimality of final answers supports H3a but does not support H3b.

In summary, there is an increase in speed to completion (performance efficiency) facilitated by the additional features of SW2. However, this increase in speed does not lead to an increase in the probability of arriving at the optimal answer. It is also

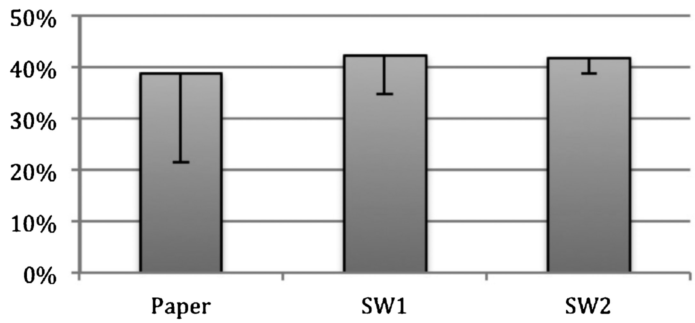

Fig. 14. Average percentage of optimal answers per run for the three conditions. The vertical lines indicate the per-condition variability (standard deviation) across the 3 runs.

worth noting that the overall trend of improvement from PAPER to SW1 to SW2 in performance efficiency (Figure 13) is consistent with the other improvements observed: the quality of knowledge-sharing and activity-awareness processes, presented earlier (e.g., Figure 8, Figure 9, and Figure 12). The implications of these trends and their relations are discussed in the following.

9. DISCUSSION

In the prior sections we reported in detail the rationale and evaluation of our collaborative system design explorations that were aimed at supporting knowledge-sharing and activity-awareness in complex collaborative planning tasks. In this section, we point to differentiating aspects in our approach compared to prior work, summarize the main findings, and draw implications from what we have learned through our design research program.

9.1. The Design Research Program

Our overarching research objective is to investigate software designs that can better respond to the need that distributed EMP groups have for quickly building shared knowledge and activity awareness so as to coordinate task content and process during complex response collaborations.

In emergency management, the distributed setting of many of the group's activities, the complexity of their tasks, the members' multiple roles, and the need to analyze a large amount of content, make knowledge-sharing and activity-awareness development challenging and critical to good performance. Current emergency management groups have little support for distributed EMP; their planning meetings are constrained by physical collocation and as a result are few and far between. By providing effective technology for distributed collaboration, we can increase the opportunities for EMP groups to interact and facilitate their shared work. At the same time, prior research on distributed groups demonstrates that remote interaction will always pose problems for communication and coordination (e.g., Straus and McGrath [1994, p. 87–88]). Therefore, in this research we gauged the extra costs that the distributed setting can impose on the processes of knowledge-sharing and activity-awareness, while at the same time experimenting with new functionalities in collaborative tools that might compensate for these process losses. In particular we sought to capitalize on opportunities not available in face-to-face collaborative settings.

Two main aspects distinguish our study of knowledge-sharing and activity-awareness: first, we adopted a design research approach; second, the task context for our investigations was complex, as we replicated in the laboratory, real collaborative conditions that had been previously characterized through fieldwork of actual EMP groups (see tabletop exercises in Schafer et al. [2007]).

The emergence of design research in HCI has been motivated by the need to formally address the increasing complexity of situations that designers are expected to support through software, for example the control of airplanes and battleships [Zimmermann et al. 2007]. We argue that the same motivations inhere in complex socio-technical systems such as collaborative EMP activities. Design research differs from the professional practice of design because the goal is not only to design more effective artifacts for consumption but also to develop new knowledge that can inform future design activities. It differs from research in psychology in that the software tools are not viewed simply as experimental instruments used in basic research. Instead, they embody a nexus of research questions in themselves [Zimmermann et al. 2007]. Carroll and Campbell [1989] offer an early conceptualization of artifacts as psychological theories, and how design rationale is one way to integrate the many levels of theory that bear on designs.

Building on fieldwork with practicing EMP groups [Schafer et al. 2007, 2008] and prior research on team cognition (see review in Convertino et al. [2005]), our design research program comprised an iterative design of a map-based collaborative prototype and a refinement of its design rationale as we empirically evaluated subsequent versions of the prototype.

First, we built and evaluated a paper prototype that resembled the tools observed in the field: the groups used paper maps (3 personal and 1 share map), Post-it notes, and markers. We developed a reference task for studying EMP in a laboratory setting and with our paper prototype. We conducted preliminary sessions with the paper-based prototype and the reference tasks to refine our procedures and measures. We then conducted a lab experiment with face-to-face groups using a paper prototype. This paper system was the first embodiment of a role-based multiple view system for geo-collaborative planning. At the end of this experiment we elicited requirements about functionalities that would be desirable in the subsequent software version of the prototype to support distributed collaboration [Carroll et al. 2007; Convertino et al. 2008a].

Second, we built and evaluated the first software prototype. We incrementally tested the features through individual and collaborative pilot sessions. The early tests with individuals led to modest changes in the user interface; the sessions with groups allowed us to refine the collaborative features (e.g., telepointer with color-coded role labels). After these refinements, we conducted an experiment with distributed groups using the first software prototype. We used the same reference tasks and map data as in the paper prototype study. Thus the first software prototype was the second embodiment of our design concept. At the end of the study we collected feedback on the current features of the prototype and suggestions about possible extensions [Convertino et al. 2007, 2009].

Third, informed by our study of the first software prototype, we built and evaluated an enhanced software prototype, which extended the prior software prototype with an Annotation Browser to better support knowledge-sharing and activity-awareness. The Annotation Browser provided an alternative view of cons shared on the group map. It supported reviewability and revisability of these pieces of shared content, and awareness of who contributed or viewed which piece of content. This prototype was the third embodiment of our design concept. The corresponding experiment evaluated the consequences of the refined design. Thus, it is the entire progression from EMP fieldwork, to the simulated tabletop exercise, and through the three different prototypes, that we hold up as our design research program.

Continuing in the same direction, we recently implemented a version of our SW2 prototype for the Web [Convertino et al. 2008c; Wu et al. 2009]. When we demonstrated the

Web-based tool informally to regional emergency managers in Centre County, PA, these professionals were quite enthusiastic about the multiple map views. They suggested that multiple-view tools like this could support aggregation of individual information into a composition situation map, and also serve as a training mechanism, allowing trainees to organize situation information on a map, and then to receive feedback from a more expert emergency professional. Although this feedback channel has been relatively informal thus far, the interactions have helped us to identify a wish list of additional features and refinements that we are implementing and will experiment with further.

9.2. The Findings from the Three Experiments

In the three lab experiments, we measured changes in process and performance variables while the groups performed repeated runs of a planning task in three different conditions: collocated and with a paper-prototype (PAPER); distributed and with a first software prototype (SW1); distributed and with a second software prototype (SW2). Through specific comparisons we tested our hypotheses about the extra process costs expected for the distributed groups (H1: Paper vs. SW1, Paper vs. SW2), the expected improvements in knowledge-sharing and activity-awareness from Paper to SW1 to SW2 (H2), and the associated improvement expected in performance (H3).

We found that the distributed setting does indeed impose extra process costs on collaboration. We documented some specific communication costs borne by the groups who worked in a distributed setting (in support of the first hypothesis). Especially when considering the groups using the first prototype (SW1), we found that compared to face-to-face groups the SW1 groups took slower and more extended turns (i.e., less frequent communication turns and more words per turn, on average) and used more of certain types of dialog acts to coordinate their collaborative work (e.g., the acts for explicitly managing the process or summarizing the shared content). The relative slowdown in turn-taking was also consistent with a relatively smaller volume of cons (pieces of information about risks) discussed.

We also showed that the communication costs may be offset via a tool that enhances specific aspects of knowledge-sharing and activity-awareness, such as the reviewability and revisability of the content shared by the group (see Clark and Brennan [1991]). The groups using the second software prototype (SW2) exhibited more efficient communication than the SW1 groups. For SW2 groups, several measures that could have replicated the process costs seen in the SW1 condition (turn frequency, words per turn, volume of cons discussed) appeared instead very similar to those of the PAPER groups. This offset effect can be explained by the extra support for knowledge-sharing and activity-awareness that is enabled by the changes to the medium: the introduction of the annotation browser enabled the reviewability and revisability of the shared content.

With respect to knowledge-sharing and activity-awareness, we found an increasing trend across a set of measures, with the highest values in the SW2 groups (in support of the second hypothesis, especially H2b). This trend is visible in the measures of post-task recall and related changes in the categories of dialog acts used; an increase in push acts and corresponding decrement of pull acts across PAPER to SW1 and to SW2; a reduction of summarization dialog acts in SW2 compared to the other two conditions; an increase in dialog acts for establishing and checking understanding in SW2 compared to the other two conditions; and a decrease in dialog acts for agreeing on or confirming judgments from PAPER to SW1 and to SW2. This set of changes suggests that when fewer explicit acts are needed to request new content, summarize, and confirm or agree with prior judgments, the group may devote more effort to directly pushing relevant content and ensuring comprehension, thus making communication more efficient. Note that these results not only document gains in content knowledge, but also point to improvements in sharing about the collaborative process and in the meta-communication about content and judgments (see related theory and experimental findings in Clark and Brennan [1991], Monk [2003], Sanford et al. [2003], and Doherty-Sneddon et al. [1997]).

With respect to group performance, we found a growing trend from PAPER to SW1, to SW2 in the speed of task completion (in support of the third hypothesis). This increase in efficiency was not associated with increases in the quality of the group solutions, which were similar across the conditions. However, as for the gains in knowledge-sharing and activity-awareness, the improvement in performance efficiency can be explained by the additional support from the features of the two software prototypes with respect to the paper prototype. For example, the revisability and reviewability afforded by the Annotation Browser, in the second software prototype, provided group members with a mechanism to implicitly share information.

Group members did not exhibit differences across the study conditions in subjective assessments of their experiences, even though they implicitly exhibited behaviors denoting improvements in the sharing process and in the performance efficiency. It may be that these subjective assessments were reflecting lack of certainty about their decision; as we have noted none of the conditions had high levels of correctness. Put another way, these ratings may simply reflect participants' perceptions that the task was indeed difficult. We turn now to a discussion of implications for theories of knowledge-sharing and activity-awareness in complex collaboration, the design of systems that can support these processes, and the methods for studying these systems.

9.3. Implications for Theory and Design: A Broader View on Knowledge-Sharing and Activity-Awareness

9.3.1. Beyond Knoweldge-Sharing and Acitvity-Awareness of Content: Focusing on Process. One implication from the findings we have presented is based on the visible gains in the amount of shared knowledge and activity-awareness built during the collaborative EMP activities. To support complex cooperative tasks, a system needs to attend to more than content-sharing. In prior literature this sharing activity is commonly discussed as the formation and maintenance of common ground (e.g., the drugstore scenario [Monk 2003]). Common ground results from exchange of content and mutual checking and signaling understanding: "I know that you know that I know what." This view of common ground emphasizes the shared understanding about the subject and focus of interaction.

In a complex decision-making domain like EMP, sharing how to approach a problem and bring it to a solution is at least as important as sharing content about the task itself. Indeed, it was our empirical findings concerning process common ground that led to new design ideas for enhancing the implicit sharing of procedural and strategic knowledge among group members. In the software prototype we implemented features (role indications associated with actions, and shared annotations within the team map) and provided empirical evidence of their benefits in collaboration (e.g., reduction in explicit agreement or confirmation acts). In research on common ground, process can be summarized as "I know that you know that I know how." This view refers to a shared understanding of the rules, procedures, timing, and manner by which the interaction will be conducted. This sort of sharing is like a blind pass in basketball. Not only do group members share an understanding of how to do the work but they also learn the signals of when to apply different shared tactics. In a blind pass, I see my teammate looking towards the right but moving the ball to her left. Thus, I prepare myself to catch the ball as opposed to moving down the court.

9.3.2. Beyond Knowledge in Common: Toward Transactive Knowledge-Sharing and Activity-Awareness. A second implication concerns the design of systems for groups that include members with complementary job roles (i.e., different areas of expertise that imply different languages, responsibilities, and priorities). For these situations, system designers must move beyond mere knowledge in common. That is, the conceptualization of content to be shared in group work must be broader than that implied by traditional models of individual cognition (e.g., Cannon-Bowers et al. [1993]). In heterogeneous work groups, diverse experts come together so as to know more and make more accurate judgments than any of the collaborators can achieve individually. In this task context, transactive models of sharing are more suitable than shared team mental models [Cooke et al. 2000; Mohammed and Dumville 2001]. Collaborative tools should encourage experts to bring to their collaborators' attention the unique knowledge they possess that is relevant at a particular point in time. If low-level details are managed individually by each expert and shared only when needed, then the group can focus on the collaborative activity of aggregating and synthesizing data, judgments, and strategy.

9.3.3. Beyond Being Aware: Capitalizing on the Benefits of Distributed Work. Third, and more generally, designers of systems for distributed collaboration must move beyond the idea of reproducing desirable properties of face-to-face collaboration, to stop assuming this as the universal reference model for collaboration [Hollan and Storenetta 1992]. We maintain that well-designed collaborative systems may allow distributed groups to perform better (and accomplish more) than face-to-face groups. Collaborative technologies can provide benefits that are not possible in unmediated conditions [Carroll et al. 2009]. For instance, prior research has shown that radar views can suggest a partner's current information needs, and over time reveal a partner's priorities and plans. More generally, we have argued that members of mediated groups may develop enhanced mutual awareness of each other due to the additional cognitive work they do to establish and maintain awareness. Thus, instead of considering the extra effort directed at coordination and awareness as a simple cost, we propose that such effort be regarded as enabling greater awareness and more effective coordination.

To move beyond being aware, the design must be grounded in task-specific empirical research, and the affordances introduced should seek to do more than approximate face-to-face collaboration. Thus an overarching design goal has been to articulate, facilitate, and enhance what is already good about distributed and computer-mediated environments, with the corresponding design objective to create a tool that increases distributed group performance beyond that of a face-to-face group. We investigated specific tool functions that were able to not only offset the costs associated with the distributed environment, but also capitalize on potential benefits that can inhere in a distributed, computer-mediated environment.

We made various design decisions aimed at capitalizing on these new benefits: separating the spaces for private and public information and making the sharing act a visible and explicit act; including explicit role indicators of user actions and traces; providing a reviewable history of the information that has been shared (Annotation Browser). One example is the function that allows members to share their private annotations onto the public map, generating wholly new and editable objects (annotations) for use by the group. The annotations record the role of the creator and the use made of them (e.g., number of views). This meta-information can be then analyzed and re-presented for other purposes (e.g., to summarize or infer relevance).

9.4. Implications for Design Research Methods

9.4.1. The Need for Reference Tasks and Validated Measures. Beyond the empirical findings and prototypes central to our design research, a secondary contribution is the experimental methods we have developed to study knowledge-sharing and activity-awareness in EMP groups. Elsewhere [Convertino and Carroll 2010] we have argued that the field of research on collaborative computing has a clear need for more studies that develop and validate measures, procedures, and reference tasks (see the call for reference tasks in Whittaker et al. [2000]). Given the variety of definitions of awareness and models of knowledge-sharing in groups this need for standard tasks and measures is particularly evident for research on knowledge-sharing and activity-awareness in the context of complex collaboration [Convertino 2008].

A first step that we made in this direction was to develop a reference task for geo-collaborative planning [Carroll et al. 2007], adapting the hidden profile paradigm from social psychology [Stasser and Titus 2003] as a model of obstacles to effective coordination in complex planning tasks. In the task, we sought to balance ecological validity of the results and the ability to manipulate or control factors. On one hand, we used multiple research tactics to investigate EMP work in realistic conditions: our reference task modeled in the lab the tabletop exercise, which is part of professional training, as documented by prior fieldwork with practicing EMP groups [Schafer et al. 2007, 2008]; we created roles and a task scenario that were based on real-world emergency planning roles and a notional scenario of a multi-expert evacuation operation as defined by FEMA [Carroll et al. 2007]; finally, we created the maps based on real GIS maps of the Centre County region. On the other hand, our version of the reference task allowed us to track the pieces of information analyzed and shared (e.g., recall of cons) and to measure the quality of the sharing process and the group outcome. In fact our adaptation of the hidden profile paradigm ensured that the best decision alternative could only be discovered if all members shared their content efficiently and filtered out irrelevant information. It is worth noticing that the built-in differences in optimality among the four alternative solutions (i.e., shelters) were known to the experimenter only, as a research tactic to connect the quality of knowledge-sharing and activity-awareness to the quality of the final group decision. The participants however, were not asked to merely discover the best solution; they had to compare the alternative solutions and pick the top three, in order of preference; which is more representative of real EMP tasks.

In the context of this reference task, we leveraged prior experimental procedures [Sellen 1995; Sanford et al. 2003; Anderson et al. 1991] to operationalize multiple measures that allowed us to gauge the changes in knowledge-sharing and activity-awareness in tandem with those in performance, which was afforded by the reference task. The convergence between different measurements of the same constructs, such as measures from post-task recall (stored knowledge), turns analysis (communication structure), and dialog acts analysis (communication content), increased the reliability of the observed changes induced by the study conditions and the repeated runs.

9.4.2. Reapplying Sensitive Measures in Field Studies and Analytics Tools. The measures of communication structure such as turn frequency and words per turn appeared to be sensitive measures of these changes in communication efficiency. A novel implication is that some of these identified markers of efficient knowledge-sharing and activity-awareness or trends-over-time of these markers could be reapplied in field studies (e.g., for communication structure: higher turn frequency, fewer words per turn, greater rate of simultaneous speech; for communication content: fewer dialog acts focusing on process coordination and explicit agreement on judgments).

Recent years have seen the emergence of various new technologies that are enabling the online storage, indexing, and reuse of various traces of collaborators' actions. A system recently developed at IBM Research allows distributed group meetings to be recorded, with speech and visuals automatically transcribed and indexed using speech-to-text technology, and then make the traces shareable and searchable. Similar commercial products are also being offered by other companies, such as Nuance.[3] While these technologies currently are focusing on facilitating discovery and rapid access to information contained in recordings, in the future they could also allow feeding back to the work groups (synchronously or asynchronously) aspects of their own process in relation to suitable baseline parameters (e.g., showing the level of communication efficiency in comparison with prior meetings by the same group or by other groups on similar tasks). Some early studies of the impact of process visualizations show that groups can self-regulate their work process if adequate information is fed back [DiMicco et al. 2007]. Managers of groups and divisions could also make use of metrics of efficiency of knowledge-sharing and activity-awareness to gauge how healthy the work processes are. In summary, this implication points to the possibility of transferring some of our lab measures into field investigations of the same phenomena, and possibly, into the design of new applications for group-work visualizations and organization-wide business analytics.

10. CONCLUSIONS AND FUTURE WORK

The research presented here raises many further questions. For example, we described aspects of shared group knowledge, and drew distinctions between the sharing of content and process knowledge, but we did not investigate how knowledge-sharing and activity-awareness practices develop as groups collaborate over long periods of time. A developmental analysis of this sort could further clarify the process/content distinction and other aspects of group knowledge.

Our study focused on geospatial information sharing in the context of relatively complex EMP activities. It provided a rich context for investigating multiple-view information management and display techniques. For example, we are experimenting with alternative visualizations that could help with sharing of process knowledge [Convertino et al. 2008c; Wu et al. 2009]. We are also continuing this design research program by investigating even more complex decision tasks that entail a wider range of supporting information tools. Currently we are investigating an information analysis task in which groups analyze a large collection of facts, 222 propositions, as compared with the 25 propositions that constituted rescue scenario our (see the CACHE study for a similar reference task for groups of intelligent analysts in Convertino et al. [2008d]). Rather than emphasizing information only within a geospatial context, we have included information about meetings and social ties, personal schedules, financial transactions, and so forth. This growing space of systems for complex cooperative tasks requiring high levels of coordination calls for more systematic research on a wider variety of visualizations, such as social networks and timelines, and thus presents many new possibilities for developing and investigating multiple-view problem-solving tools.

ACKNOWLEDGMENTS

Several close collaborators contributed to this multi-year design research project: Dejin Zhao and Blaine Hoffman developed the software prototypes; Craig Ganoe and Wendy Schafer engineered the underlying Java-based software architecture; Alex Ting, Caitlin Ferro, Prajakta Bhambare, and Su Chuen Foo worked scrupulously with us on running the experiments and performing the rigorous data analyses.

REFERENCES

ANDERSON, A. H., BADER, M., BARD, E., BOYLE, E., DOHERTY, G., GARROD, S., ISARD, S., KOWTKO, J., MCALLISTER, J., SOTILLO, C., AND THOMPSON, H. 1991. The HCRC map task corpus. *Lang. Speech 34,* 4, 351–360.

BARDRAM, J. 1998. Designing for the dynamics of cooperative work activities. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW).*

[3] http://www.nuance.com.

BØDKER, S. 1996. Applying activity theory to video analysis: How to make sense of video data in human-computer interaction. In *Context and Consciousness: Activity Theory and Human-Computer Interaction,* Nardi, B. A., Ed., MIT Press, Cambridge, MA, 147–174.

CANNON-BOWERS, J. A., SALAS, E., AND CONVERSE, S. A. 1993. Shared mental models in expert team decision- making. In *Individual and Group Decision Making: Current Issues,* N. J. Castellan Ed., 221–246.

CARLETTA, J., ISARD, A., ISARD, S., KOWTKO, J., NEWLANDS, A., DOHERTY-SNEDDON, G., AND ANDERSON, A. H. 1997. The reliability of a dialogue structure coding scheme. *Computat. Lingui. 23*, 1, 13–32.

CARROLL, J. AND CAMPBELL, R. 1989. Artifacts as psychological theories: The case of HCI. *Behav. Inform. Technol. 8*, 4, 247–256.

CARROLL, J. M., ROSSON, M. B., CONVERTINO, G., AND GANOE, C. H. 2006. Awareness and teamwork in computer-supported collaborations. *Interact. Comput. 18*, 1.

CARROLL, J. M., CONVERTINO, G., GANOE, C., AND ROSSON, M. B. 2008. Toward a conceptual model of common ground in teamwork. In *Macrocognition in Teams: Understanding the Mental Processes that Underlie Collaborative Team Activity*, M. Letsky, N. Warner, S. Fiore, C. Smith, Eds., Elsevier, New York.

CARROLL, J. M., MENTIS, H., CONVERTINO, G., ROSSON, M. B., GANOE, C. H., SINHA, H., AND ZHAO, D. 2007. Prototyping collaborative geospatial emergency planning. In *Proceedings of the 4th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

CARROLL, J. M., NEALE, D. C., ISENHOUR, P. L., ROSSON, M. B., AND MCCRICKARD, D. S. 2003. Notification and awareness: Synchronizing task-oriented collaborative activity. *Int. J. Hum.-Comput. Studies 58*, 605–632.

CARROLL, J. M., ROSSON, M. B., FAROOQ, U., AND XIAO, L. 2009. Beyond being aware. *Inform. Org. 19*, 3, 162–185.

CLARK, H. H. AND KRYCH, M. A. 2004. Speaking while monitoring addressees for understanding. *J. Mem. Lang. 50*, 1, 62–81.

CLARK, H. H. 1996. *Using language*. Cambridge University Press, Cambridge, UK.

CLARK, H. H. AND BRENNAN, S. E. 1991. Grounding in communication. In, *Perspectives on socially shared cognition*., L. B. Resnick, J. Levine, and S. D. Teasley, Eds., APA, Washington, DC.

CONVERTINO, G., GANOE, C. H., SCHAFER, W. A., YOST, B., AND CARROLL, J. M. 2005. A multiple view approach to support common ground in distributed and synchronous geo-collaboration. In *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV)*.

CONVERTINO, G., MENTIS, H. M., TING, A. Y. W., ROSSON, M. B., AND CARROLL, J. M. 2007. How does common ground increase? *In Proceedings of the International Conference on Supporting Group Work (Group)*.

CONVERTINO, G., MENTIS, H., ROSSON, M. B., CARROLL, J. M., SLAVKOVIC, A., AND GANOE, C. 2008a. Articulating common ground in cooperative work: Content and process. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

CONVERTINO, G., MENTIS, H. M., BHAMBARE, P., FERRO, C., CARROLL, J. M., AND ROSSON, M. B. 2008b. Comparing media in emergency planning. In *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

CONVERTINO, G., ZHANG, X. L., WU, A., GANOE, C. H., HOFFMAN, B., AND CARROLL J. M. 2008c. Designing group annotations and process visualizations for role-based collaboration. In *Proceedings of the International Workshop on Social Computing, Behavioral Modeling, and Prediction*.

CONVERTINO, G., BILLMAN, D., SHRAGER, J., PIROLLI, P., AND MASSAR, J. P. 2008d. The CACHE study: Group effects in computer-supported collaborative analysis. *J. Comput. Support. Coop. Work 17*, 353–393.

CONVERTINO, G. 2008. Awareness and knowledge sharing in collaborative computing experimental methods. Ph.D. Thesis, College of Information Sciences and Technology, Pennsylvania State University.

CONVERTINO, G., MENTIS, H. M., ROSSON, M. B., SLAVKOVIC, A., AND CARROLL, J. M. 2009. Supporting content and process common ground in computer-supported teamwork. In *Proceedings of the Conference on Human Factors in Computing (CHI)*.

CONVERTINO, G. AND CARROLL, J. M. 2010. Toward experimental methods in collaborative computing research. In *The Handbook of Emergent Technologies in Social Research*, Sharlene Hesse-Biber, Ed., Oxford University Press, Oxford, UK.

COOKE, N. J., SALAS, E., CANNON-BOWERS, J. A., AND STOUT, R. 2000. Measuring team knowledge. *Hum. Factors, 42*, 151–173.

DIMICCO, J. M., HOLLENBACK, K. J., PANDOLFO, A., AND BENDER, W. 2007. The impact of increased awareness while face-to-face. *Hum.-Comput. Interact., 22*, 47–96.

DHS-FEMA (US DEPARTMENT OF HOMELAND SECURITY, FEDERAL EMERGENCY MANAGEMENT AGENCY). 1997. Federal Response Plan (for Federal Law 93-288, first published April 1992). http://library.findlaw.com/1992/Apr/1/127810.html.

Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., and Bruce, V. 1997. Face-to-face and video mediated communication: A comparison of dialogue structure and task performance. *J. Exper. Psychol. Appl. 3,* 2, 1–21.

Engeström, Y. 1990. *Learning, Working and Imagining. Twelve Studies In Activity Theory.* Orienta-Konsultit, Helsinki.

Fussell, S. R., Kraut, R. E., and Siegel, J. 2000. Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*.

Haddow, G. D. and Bullock, J. A. 2003. *Introduction to Emergency Management*. Butterworth-Heinemann, Amsterdam.

Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., and Subbiah, I. 2002. Development of a self-report measure of environmental spatial ability. *Intell. 30*, 425–447.

Hollan, J. D. and Stornetta, S. 1992. Beyond being there. *In Proceedings of the Conference on Human Factors in Computing (CHI)*.

Kraut, R. E., Gergle, D., and Fussell, S. R. 2002. The use of visual information in shared visual spaces: Informing the development of virtual co-presence. *In Proceedings of the Conference on Computer Supported Cooperative Work (CSCW).*

Malz, D. N. and Borker, R. A. 2007. A cultural approach to male-female miscommunication. In *A Cultural Approach to Interpersonal Communication*, L. Monagham and J. Goodman, Eds., Blackwell Publishing, Malden, MA.

McCarthy, J. C., Miles, V. C., and Monk, A. F. 1991. An experimental study of common ground in text-based communication, In *Proceedings of the Conference on Human Factors in Computing (CHI)*.

McGrath, J. E. 1984. *Groups: Interaction and Performance*. Prentice-Hall, NJ.

Mohammed, S. and Dumville, B. C. 2001. Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *J. Organiz. Behavior. 22*.

Monk, A. 2003. Common ground in electronically mediated communication: Clark's theory of language use, in *Toward a Multidisciplinary Science of HCI,* J. M. Carroll Ed., MIT Press.

Monk, A., McCarthy, J., Watts, L., and Daly-Jones, O. 1996. Measures of process. In *CSCW Requirements and Evaluation*, P. J. Thomas, Ed., Springer-Verlag, Berlin.

Nardi, B. A., Ed. 1996. *Context and Consciousness: Activity Theory and Human-Computer Interaction.* The MIT Press, Cambridge, MA.

Sanford, A., Anderson, A. H., and Mullin, J. 2003. Audio channel constraints in video-mediated communication. *Interact. Comput. 16*, 6.

Schafer, W., Carroll, J. M., Haynes, S., and Abrams, S. 2008. Emergency management planning as collaborative community work. *J. Homeland Security Emergency Manage. 5,* 1, 1–17.

Schafer, W. A., Ganoe, C., and Carroll, J. M. 2007. Supporting community emergency management planning through a geocollaboration software architecture. *J. Comput. Support. Coop. Work, 16*.

Sellen, A. J. 1995. Remote conversations: The effects of mediating talk with technology. *Hum.-Comput. Interact. 10*, 401–444.

Stasser, G. and Titus, W. 2003. Hidden profiles: A brief history. *Psychol. Inq. 3–4*, 302–311.

Stewart, D. D. and Stasser, G. 1995. Expert role assignment and information sampling during collective recall and decision making. *J. Personal. Social Psychol. 69*, 619–628.

Straus, S. G. and McGrath, J. E. 1994. Does the medium matter? The interaction of task type and technology on group performance and member reactions. *J. Appl. Psychol. 79*, 1, 87–97.

Topkara, M., Pan, S., Lai, J., Wood, S., and Boston, J. 2010. Tag me while you can: Making online recorded meetings shareable and searchable. IBM Research rep. RC25038 (W1008-057).

Whittaker, S., Terveen, L., and Nardi, B. A. 2000. Let's stop pushing the envelope and start addressing it: A reference task agenda for HCI. *Hum.-Comput. Interact. 15*, 75–106.

Wu, A., Zhang, X., Convertino, G., and Carroll, J. M. 2009. CIVIL: Support geocollaboration with information visualization. In *Proceedings of the International Conference on Supporting Group Work (Group)*.

Zimmerman, J., Forlizzi, J., and Evenson, S. 2007. Research through design as a method for interaction design research in HCI. *In Proceedings of the Conference on Human Factors in Design (CHI)*.

Received March 2011; revised July 2011; accepted September 2011