A Job-Shop Scheduling Task for Evaluating Coordination during Computer Supported Collaboration

Desney S. Tan[†], Darren Gergle[‡], Mary Czerwinski[†]

[†] Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{desney, marycz}@microsoft.com

[‡] Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
dgergle@cs.cmu.edu

ABSTRACT
Researchers have begun to explore tools that allow multiple users to collaborate across multiple devices. One class of these tools allows users to simultaneously place and interact with information on shared displays. Unfortunately, there is a lack of good tasks to evaluate the effectiveness of these tools for information coordination in such scenarios. In this paper, we present *collaborative job-shop scheduling*, a task we have designed to evaluate systems and interactions within computer supported collaboration environments. We describe properties that make the task useful, as well as evaluation measures that can be used with this task. We validate the feasibility of the task and demonstrate analysis techniques in an experiment we conducted to compare the differences between presenting information serially versus simultaneously on a large shared display. Results from this experiment show the benefits of shared visual information when performing coordination tasks.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: User interfaces - evaluation, screen design; H.5.3 [Group and Organizational Interfaces]: Computer supported cooperative work.

General Terms
Performance, Experimentation, Human Factors.

Keywords
Job-shop scheduling, task, evaluation, coordination, collaboration, user study, shared visual information.

1. INTRODUCTION
As computing technologies move off the desktop and into the everyday world, methods for examining group interactions in computer supported collaboration environments are becoming ever more essential [16]. Unfortunately, there is currently a shortage of useful experimental paradigms, particularly for evaluating technologies designed to support coordinating information using shared visual displays in collocated environments. In a recent flurry of workshops and papers, researchers have begun work to understand and address this shortage (e.g. [9, 18]).

1.1 Task Framework
In our work, we use McGrath's task taxonomy [15] to provide a conceptual framework that facilitates discussion of existing experimental paradigms. This taxonomy describes eight distinct task types: planning, creativity, contests/battles, performance, mixed-motive, cognitive conflict, intellective, and decision-making tasks. Only some of these types have suitably developed tasks for evaluating collaborative technologies. Working within this framework, we uncover a design space that indicates the need for development of certain classes of tasks that can be used in evaluating collaborative technologies.

In our literature review, we have found that generative tasks aimed at examining group *planning* and *creativity* are fairly well represented with existing experimental paradigms. These are typically tasks that focus on the generation of ideas and plans, as well as the subsequent selection and execution of chosen alternatives. Researchers investigating collaborative systems commonly use tasks like the automated post office design task [17] and furniture layout tasks [22] to assess mediated group performance while generating creative ideas.

Similarly, the number of executable tasks such as *contests/battles* or physical *performance* tasks has seen recent growth. These task areas are traditionally viewed as those that involve physical behavior, as opposed to symbolic, mental, or verbal manipulations. They deal with motor behaviors and expert performance in competitive or non-competitive forms. In a computational world, the pipe construction task [7], the collaborative Lego construction task [3], and collaborative puzzle construction tasks [14] are all representative of this category.

A host of negotiation tasks have also been explored. These are tasks that typically center on a form of conflict and include both mixed-motive and cognitive conflict tasks. Researchers have examined *mixed-motive* tasks, which generate tension between individual and collective rationality, across a range of technologically-mediated environments [19]. These tasks include bargaining, coalition formation, or game theoretic social dilemmas. *Cognitive conflict* tasks are another type of negotiation task in which the conflict resides in participant viewpoint rather than participant interest. An example of such a task is the desert survival task, recently used by Setlock et al. [23].

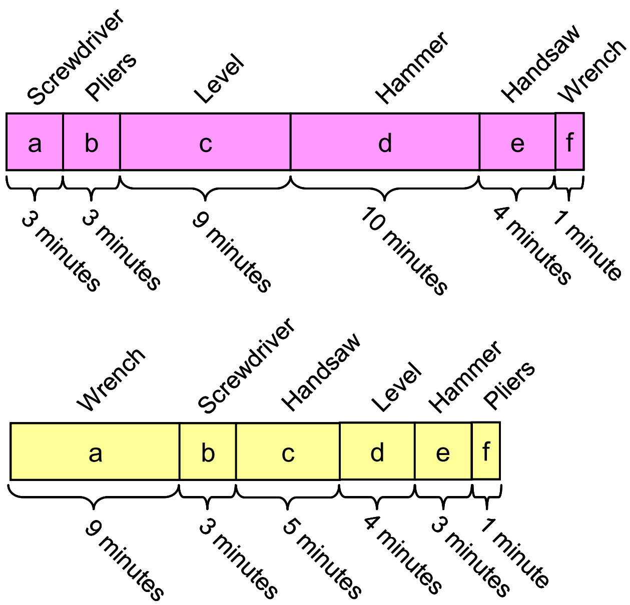

Figure 1. Two example jobs. Each job (color) comprises strictly ordered operations (blocks) requiring specific resources (shop tool) for some time.

The final group of tasks are *intellective* and *decision-making* tasks. These are tasks that involve problem solving with demonstrably correct answers or consensually agreed upon solutions. Although a large number of carefully crafted task paradigms have been designed around written or verbal discussions, real world activities in this space often involve integrating visual information while simultaneously discussing and arranging real world objects. Hence, existing intellective tasks such as logic problems are ill-suited to assessing the value of many modern collaborative systems. In our review of the literature, we found that there exists a major gap in this task space, particularly when considering real-world interactions that often require the use of physical and visual objects. We attempt to fill this gap with our work.

1.2 Possible Approaches to Developing a Task

There are at least two major approaches for generating useful task paradigms. We could either adapt existing group tasks to measure desired effects brought about by technological mediation, or we could scale up tasks designed for individual assessment so that they work for groups. In our work, we take the latter approach. In order to find a group task that tests coordination in intellective and decision-making scenarios, we start with a well-studied task traditionally performed by individuals and modify it so that it works for groups.

1.3 Our Contribution

The primary contribution of this paper is a novel variation of a classic job-shop scheduling task that we assert can be used to evaluate coordination tools in computer supported collaboration environments. This task helps fill the void found in current tasks used for exploring group problem-solving with visual information. We describe properties that make the task useful, as well as evaluation measures that can be used with this task. We then describe an experiment we conducted using this task in order to illustrate possible analysis methods as well as to demonstrate the feasibility of this task for evaluating collaborative software tools. The actual results of the experiment form the secondary contribution of this paper. These results highlight the benefits of concurrent viewing of shared information in coordination tasks.

2. JOB-SHOP SCHEDULING

2.1 Traditional Task

The traditional job-shop scheduling task consists of a finite set of jobs, each composed of a chain of ordered operations. For example, in Figure 1, the jobs are uniquely indicated by color, and the ordered operations are shown as blocks containing sequential letters. Each operation must be performed in strict sequence and cannot begin until the previous one has fully completed. Each operation must be performed in a single uninterrupted period of time using a specific resource (e.g. a shop tool such as a 'hammer'). There exist a finite set of these resources, and each is capable of performing no more than one operation at a time. In other words, operations cannot temporally overlap on a given resource.

To solve the task, the user must schedule all operations while observing the operation ordering and overlapping constraints. An optimal solution is one in which the last operation is completed at the earliest possible time (for an example of an optimal solution to a problem set, see Figure 3). For a more detailed review of scheduling problems as well as computer-based solution techniques, see [11] and [24].

2.2 New Collaborative Task

To extend the individual job-shop task to a collaborative task, we assign each user in the group a subset of the jobs. Users then have to coordinate scheduling operations using the shared resources in order to get all their jobs completed in the shortest amount of time for the group as a whole.

Our collaborative job-shop scheduling task has several nice properties. First, it is simple to explain and easy to learn. Second, it is representative of many real world collaborative tasks in which information is distributed among group members and knowledge must be integrated to formulate a joint solution. Third, interaction with content contains many co-dependencies. In fact, rescheduling one operation typically requires having to move many others around it. This is important because it requires tightly integrated coordination even if users have access to all information. Fourth, the task cannot be solved by a simple algorithmic strategy and iterative improvement is required. Finally, the task has an optimal solution, as well as other metrics that may be useful while evaluating novel collaboration systems.

In real world coordination scenarios, participants typically have a finite amount of time to negotiate the best possible solution. This solution may not be optimal. In fact, it may not even correctly meet all stated constraints. In many cases, iteratively improving the solution until it is optimal may take an arbitrarily large amount of time. This is also true of our task. Hence, rather than trying to measure the time it takes groups to obtain optimal solutions, we have users work for a fixed amount of time and measure the quality of solutions attained. In order to get as complete a description of performance and process as possible, we devised multiple metrics that can be used with this task.

2.3 Evaluation Metrics

There are three interesting classes of evaluation metrics that can be used with the collaborative job-shop scheduling task: outcome measures of performance and communication efficiency, process measures describing social and communication processes, and self-report measures providing subjective ratings of perception of the tools, group processes, and the task.

Outcome measures of performance include the number of times the sequenced letters are placed out of order (ordering error), the number of times any resource is scheduled to simultaneously perform more than one task (overlap error), and the degree to which the groups optimally schedule the group of jobs (measured by solution length).

This quantitative performance data can be complemented with conversational analysis using video and audio transcripts as well as log data representing patterns of tool use. Outcome measures of communication efficiency represent low level communication mechanics that might affect task performance. One such group of measures is a relatively simple count of linguistic components.

Process measures of communication, on the other hand, look at higher level strategies used to solve the task. These measures require some amount of semantic interpretation. Furthermore, analysis can be done to explore social effects such as dominance and leadership, patterns of scheduling and submitting solutions, as well as general use of software tools and interfaces.

Finally, self-report measures using questionnaires, surveys, or interviews are useful for exploring such factors as the level of satisfaction with tools, the perceived distribution of contribution from various users, as well as overall interest in the task.

3. EXPERIMENT

We designed an experiment to assess the feasibility of the collaborative job-shop scheduling task for evaluating collocated small group interactions. In their work, Bly and Rosenberg showed that tiled windows are superior to overlapping ones for individual users performing tasks that require content coordination [2]. They explained that these effects were due mainly to limitations in human short-term memory and attention. Their work suggested that we should design systems that offload the maximum amount of cognitive effort by placing relevant cues and affordances within the interface. We believe that this design requirement is even more important when multiple people are trying to coordinate information that exists on different sources (e.g. each user brings a laptop containing their information into a planning meeting).

Hence, we use the collaborative job-shop scheduling task to examine differences in performance and communication across two methods of sharing display content in a collaborative scenario. In one method, commonly used in conference rooms and meetings today, users take turns projecting information on a large shared display. In the other, a method becoming increasingly available with new software tools, users simultaneously share visual information from multiple sources on the large display.

3.1 Hypotheses

One of the ways large shared displays might benefit group performance is by providing visual information that can be used to support task coordination and the surrounding communication. Research has shown that users utilize visual information to support conversational grounding and task awareness [4, 13]. Hence, we expected that providing a method for groups to share their information in a centralized fashion would facilitate group performance. Specifically,

> *Hypothesis 1: Groups will produce more optimal solutions (fewer errors and shorter solution length) on the collaborative job-shop scheduling task when multiple group members display information simultaneously.*

In addition to the basic outcome measures, we hypothesized that groups would adjust their communication processes to take advantage of the shared visual information. Prior work has demonstrated that conversational efficiencies typically accompany the availability of shared visual information [3, 6, 14]. However, since shared visual information is available in both conditions in this experiment, we expected to see greater communicative efficiency primarily when the shared visual information was more salient as a conversational resource (i.e. when users were able to simultaneously share visual information). Hence,

> *Hypothesis 2: Groups will use more efficient communication techniques when they can simultaneously display shared visual information from all members of the group.*

Finally, we hypothesized that the shared experience would cause users to rate this method more favorably. In fact,

> *Hypothesis 3: Members of the groups will find that simultaneously sharing information is more satisfying and more effective for coordinating information while performing the collaborative job-shop scheduling task.*

3.2 Participants and Setup

Twenty-four (12 female) college students, aged 19 to 31 years old, volunteered for the study. All users spent more than 30 hours a week using a computer, and none had prior experience with the experimental software. Users were divided into eight groups of three people each, with each group consisting either of all male or

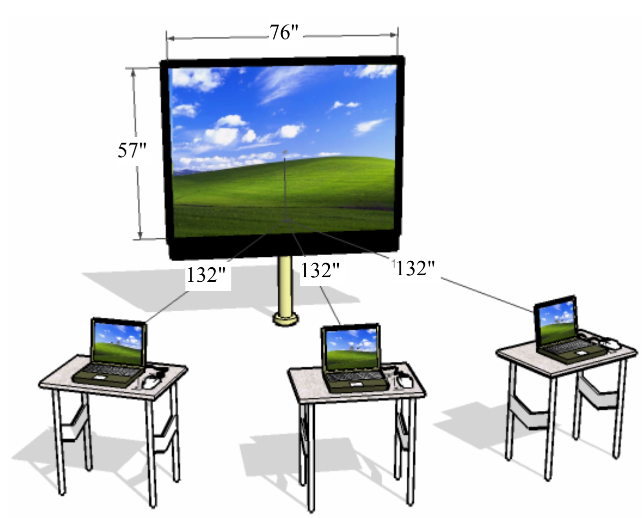

Figure 2. Diagram of the experimental setup.

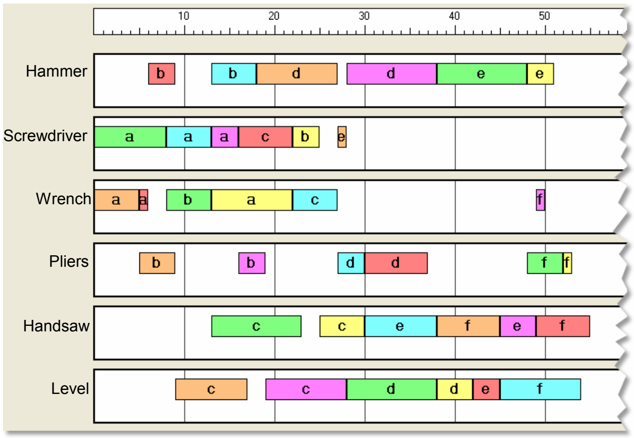

Figure 3. Complete view of optimal solution to the problem. Users never saw this view in the experiment.

all female users. Users within each group did not know each other prior to the study. The study took about an hour and users were paid a small gratuity for participating.

Users from each group sat at three desks each facing a large 95" (~241cm) wall-projected display. Each table was 11' (~3.35m) away from the large display and the two side tables were about 25 degrees off center on either side. See Figure 2 for an illustration of the setup. Each user interacted through an IBM Thinkpad laptop and Microsoft Intellimouse placed on each of their tables. All users had an unobstructed view of other users as well as of the large display, but could not see each other's laptop displays. The laptops were connected over a local network to the desktop computer driving the large display.

3.3 Task

The collaborative job-shop scheduling test we used consisted of six resources and six jobs, each with six operations, as seen in Figure 3. This test was taken from Fisher and Thompson's benchmark tests, which have been used extensively in validating online scheduling algorithms [5]. For similar benchmark tests, see [1].

Each of the three users was responsible for scheduling two of these jobs. Users had to coordinate schedules because they had to share the six resources available to perform the operations. We built a scheduling program that allowed users to adjust their schedules simply by dragging bars representing each operation along a time line (see Figure 4 for an example of what each user saw on their personal laptop display). The interface fixed the specified length of each operation and sorted each by the resource it required. Additionally, jobs were distinctly color coded, and the sequence of operations for each job was labeled 'a' through 'f'.

3.4 Manipulation and Procedure

After balancing for gender, we randomly assigned each group to one of two between-group conditions: Serial Presentation or Parallel Presentation. In both conditions, we used the WinCuts and Visitor systems [26] to replicate content that existed on the local laptop displays onto the large projected display, and to allow each user to use their laptop mouse across the two sets of displays. WinCuts is a system that allows users to replicate arbitrary regions of existing windows into independent windows. Each of these new windows is a live view of a region of the source window with which users can interact. Each window can be shared across multiple devices and hence used to send information to the shared display. Visitor is a system that redirects the input stream over the network so that a user can use a mouse and keyboard connected to one computer to control the input on another computer. We used Visitor to allow users to move their cursor off the top edge of their local laptop screen to control the cursor on the shared display. If multiple users did this simultaneously, they would 'fight' for control. Hence they had to socially mediate the use of the cursor on the large display. We saw no instances in which control collisions were not quickly resolved.

In the Serial Presentation condition, groups could see information from only one of the users on the large shared display at any given time. However, any user could move their cursor to the shared display, click on the taskbar there, and switch to a view of another user's solution space. In the Parallel Presentation condition, groups could simultaneously see information from all three users on the shared display. Using WinCuts, all three users could select relevant regions of their local content to send to the large display for simultaneous viewing. Furthermore, users could rescale and lay out content placed on the shared display. Although updates were seen on the large display in real time, users could only control their own content. It is interesting to note that all four groups in this condition decided to divide the display into thirds and to scale and vertically stack their information. This makes sense, since it allows simultaneous viewing of the information in a way that best aided the task.

Before the test, we gave users written instructions and had them practice on a representative task for 5 minutes. Once they were familiar with the interface and task, they had 20 minutes to work on the actual test. They were warned when they had ten, five, and then one minute remaining. Following the test, users filled out a satisfaction questionnaire.

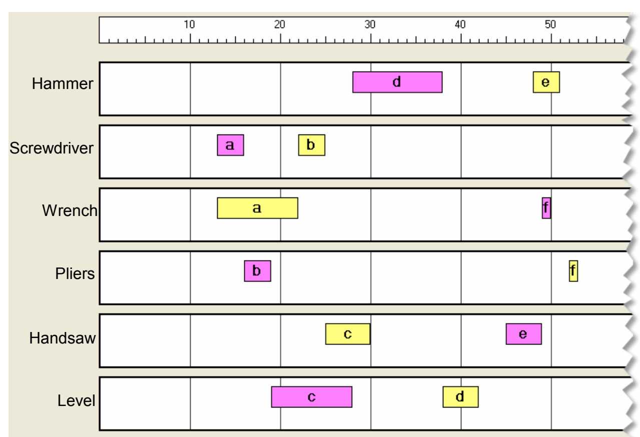

Figure 4. View of one user's schedule. Each user was responsible for two different jobs.

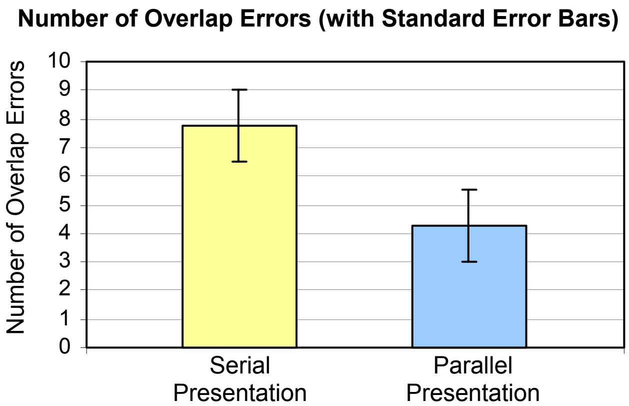

Figure 5. Users made significantly fewer overlap errors in the Parallel Presentation condition.

3.5 Results

3.5.1 Outcome Measures: Task Performance

Using an analysis of variance (ANOVA), we examined two different performance metrics: number of overlap errors and solution length.

To calculate the overlap error, we counted the number of operations scheduled on the same resource at the same time as another. This score could range from 0 (no errors) to 36 (every piece violated the constraint). We observed significantly fewer overlap errors with Parallel Presentation than with Serial Presentation ($F(1,6)=15.47$, $p=.007$, see Figure 5). In other words, groups were more likely to complete the task with fewer errors when they had shared visual information made accessible using WinCuts, rather than having to keep some of this information in their memory or to continually communicate it verbally.

The second dependent variable we examined was possible solution length. It ranged from 10 (the longest single task) to 91 (bounded by space in the interface). The optimal solution for this problem was 55 (see Figure 3). While we found that groups using Parallel Presentation had solutions that were shorter (better) on average (73.5 vs. 78.5 units), the difference was not statistically significant ($F(1,6)=1.65$, $p=.25$).

However, given the small number of groups run in this study and the fact that we were interested in assessing whether or not this measure might be of value in the future, we performed a power analysis to investigate its sensitivity. While caution must be taken when interpreting the findings of power calculations [8], we use it to guide future studies by generating a least significant number (LSN). The LSN indicates the number of observations expected to be needed in order to achieve significance given the existing (or expected) standard deviation, effect size, and alpha-value. The parameters for our analyses were: $\sigma = 5.51$; $\delta = 2.5$; and $\alpha = .05$. This analysis revealed that we would have found a significant difference ($p<.05$) with 21 groups. This suggests that while our measure of solution length is clearly not as sensitive as the overlap measure described above, it is not completely infeasible as a performance metric (11 groups in each condition). Future work needs to be done in order to resolve whether this is a feasible measure to consider for this task.

We should caution that even though it did not happen in our experiment, there could exist an error-optimality trade off. For example, a group could create a really short solution by overlapping all the operations. Thus, even though this might lead to a really short solution length, we would see an increase in overlap error. We believe that the usefulness of either of these metrics must be carefully examined in the context of the specific interface and instructions provided. In our experiment, we explicitly instructed users to aim for the shortest possible valid answer. While we did not see any trade off effects in our experiment, we would advise that other researchers using this task be aware of this possibility.

3.5.2 Outcome Measures: Communication Efficiency

While we found a statistical difference in the quality of the solution (i.e. the groups had less overlapping pieces in the parallel condition), this result tells us little about the communication and strategy used to solve the task. We hypothesized that performance improvements with Parallel Presentation would be partially due to increased communication efficiency brought about by more salient shared visual information. Hence, we expected lower word counts, lower utterance counts, and increased use of conversationally efficient linguistic references such as deictic pronouns [13].

Since groups contained three participants working together and their responses were likely correlated to one another based on the performance of the group as a whole, treating data points from each participant as independent could lead to inflated Type I and II errors. In order to treat data as independent (belonging to the individual) even though they were correlated with actions within the group, we analyzed the data using the mixed model analysis technique described in Kenny et al. [12].

While the means tended to favor shared visual information across our communication efficiency measures (see Table 1), none of the models reached statistical significance. We believe this could be due to the noise inherent in such measures and to the small number of groups we observed.

3.5.3 Process Measures

While the measures of communication efficiency provide aggregated counts of communication features, they provide little detail regarding the actual process engaged in by the groups. The following transcripts present examples of detailed assessments that can be made using this task in order to establish a deeper understanding of the ways in which the technologies affect perform-

Table 1. General benefits of communication efficiency with Parallel Presentation.

	Avg. Utterances per Individual		Avg. Words per Individual		Rate of Deictic Pronoun Use	
	Mean	Std Error	Mean	Std Error	Least Sq. Mean	Std Error
Serial	113.833	18.069791	753.833	93.174195	104.17955	12.037429
Parallel	97.417	18.069791	702.333	93.174195	110.07045	12.037429

ance. This also serves as a preliminary investigation of the content to see where we might focus future efforts on exploring process differences though the development of a behavioral coding system and further data collection.

We expected groups to be more efficient and less error-prone with Parallel Presentation, in which shared visual information was simultaneously available. A detailed exploration of the transcripts and logs seems to confirm this. For example:

Serial [Group 4S – Querying]
3: Ok, um, I guess is anyone's A longer than this?
1: Yea, I have one that's for 8 minutes.
3: Ok uh.
2: I guess I'll be able to move to A.
3: When does your end? When does your A end?
1: Oh mine? I'm sorry, 8.
3: Well my A for red is really small, I'll show you guys. Do you have a bigger A?

Parallel [Group 0P – Demonstrating]
2: Everyone put their A's down. Move everything else away for now.
1: I'll have to start this later than I'd like to, but that's ok.

In the Serial excerpt, it is evident that when the groups are attempting to identify options for their 'a' operation, they use a rather inefficient method of querying one another and then waiting for a verbal response. However, in the Parallel excerpt, one of the users suggests pulling out all 'a' operations for everyone to see. This provides a shared visual resource that can be used for grounding subsequent conversation.

Additionally, the shared visual information provides less ambiguous information than linguistic descriptions. In fact, we observed several instances where errors that would have been caught in the Parallel Presentation condition were missed in the Serial Presentation one. For example:

Serial [Group 4S – Undetected Mistake]
2: We are always using it until 33
1: Yea ok that's fine. Can you put yours after mine? (Error: 3 puts the operation at 36 instead of intended 33)
3: For the D

Parallel [Group 0P – Detected Mistake]
1: Oops, haha, move yours to start at the end of mine, at 35, no 36. (Error: 2 moves the incorrect operation to 36)
1: No no no, not the level, the hammer, yeah that.

3.5.4 Self Report Measures

Finally, we analyzed users' perceptions of satisfaction and their overall level of confidence with their final solution. The analysis was a mixed model analysis (as described above) in order to control for correlated responses within groups. The data showed that the Parallel condition (Mean (SE) = 3.92(.24)) was viewed as significantly more satisfying ($F(1,22)=10.62$, $p=.004$) than the Serial condition (2.83(.24)). The Parallel condition (3.42(.35)) was also considered marginally easier ($F(1,22)=2.88$, $p=.10$) than the Serial condition (2.58(.35)).

4. DISCUSSION

The results demonstrate the feasibility of our collaborative job-shop scheduling task as a candidate task that can be applied to the evaluation of coordination tools in computer supported collaboration environments. In this section we would like to take some time to address the areas where we believe the task worked well and describe some ways in which we believe it may be improved, discuss the overall generalizability of the task and its relation to existing task paradigms, and speculate on the types of coordination and collaborative technologies for which we believe it may be useful.

4.1 Reflection on Findings

In order to evaluate the potential of our task, we decided to use rather stringent testing criteria. First, we used a relatively small number of groups (eight). Since group studies are notoriously difficult to run and require greater resources than studies of individuals, we chose what we felt was a lower bound on an acceptable number of groups. In addition, we used a between-subjects design which has a strong advantage when evaluating a new task in that users are not contaminated by exposure to additional levels of the independent variable. However, doing so makes it rather difficult to find statistical differences between the groups unless the effect sizes are large and the individual differences are minimal. We feel that this is the correct approach to evaluating a new task. It provides information regarding effects that are large enough to be practically relevant (not simply driven by statistical power) and a clean test of experimental factors without having to worry about potential contamination across levels of the independent variables. Keeping this in mind, we now reflect on our findings.

Overall, our results suggest that our task is fairly sensitive at detecting differences on the dependent variable of overlap errors. Using the criteria described above, we were able to obtain mean values that we could claim to be different with over 99% confidence. The second dependent variable that we tested was that of overall solution length. While this measure was not found to be as sensitive as the first at detecting differences in our experiment, we feel that it may still be a practical measure to collect. Our power analysis suggested that we probably needed just more than double our sample size in order to find statistically valid differences. This remains a candidate for future work.

Our raw measures of communication efficiency—as reflected in word counts, utterances, and deictic references—were much less successful. While all of the means were in the hypothesized direction, we did not achieve differences at statistically acceptable levels. We believe this is accounted for, in part, by the fact that measures of communication efficiency are highly variable and group specific. If researchers are particularly interested in using this task to analyze such communication efficiency measures, we would strongly suggest a within-subjects approach in order to control for individual (or in this case group) communication preferences.

Finally, a review of the communication and action transcripts was useful in providing descriptive events that demonstrate how groups adapt their communication to account for the presence or absence of shared visual information. This was demonstrated by the fact that the pairs often times used the visual space to 'demonstrate' their available task objects [3] in the Parallel Presentation condition, while in the Serial Presentation condition they simply used language to describe the potential objects rather than switching views of the workspace. While these findings are not entirely novel, they suggest that our task provides a useful platform for investigating discourse patterns. However, if researchers plan to use this task to perform dialogue or discourse analysis, we would also suggest that they consider within-subjects manipulations in order to help account for the individual differences inherent in communication patterns.

4.2 Relation to Existing Tasks

One of our goals was to design a task that filled the void of collaborative tasks described in the introduction. While a great number of tasks exist in sociology and psychology literatures, we noticed that very few tasks support the integration of visual information or physical objects. For example, most intellective and decision-making tasks have the correct group dynamic structure but reside entirely in the mental domain. These tasks are useful for eliciting group discussion and negotiation, but the information resides solely in an individual's head and only becomes group knowledge when shared through a public spoken discourse. In contrast, most execution tasks focus solely on the physical attributes and their manipulation without the group negotiation and process attributes of a decision-making task. However, many real world tasks share a similar dynamic but have additional physical or visual objects in which the information resides. Even though developers of collocated collaborative systems typically assume that supporting interaction with such objects is useful, there are surprisingly few tasks that can be directly applied to investigate these tools. What are needed are tasks that merge the group dynamics of the former with the tangible and physical attributes of the latter.

Our task fills this void by providing a task environment requiring physical manipulation and shared visual information, elicits discussion and negotiation within a group, and has an ultimate solution that can be regarded as optimal (as opposed to a completely open-ended brainstorming task). One advantage of this is explicit performance metrics that in many other cases are more ambiguous and difficult to form. For example, the furniture layout tasks that are useful for eliciting interaction have less structured solutions that are hard to assess for quality.

4.3 Impact on Applicable Technologies

We believe that our proposed task can be used to evaluate the effectiveness of various emerging technologies for information coordination and collaboration. With further analysis, the task can also be used to provide insight into how users adapt their social behaviors based on the technologies adopted. We hope that this insight will drive not only refinement of existing technologies, but also innovation and design of future ones. Evaluation tasks like this one are particularly important with the recent growth of technologies geared towards distributing computing resources in various form factors throughout the environment.

For example, there has been a recent interest in tabletop displays that allow multiple people, clustered around the horizontal display surface, to view and coordinate a common set of information. These displays are particularly interesting because even though users have a shared view of all information, some of it is usually optimally oriented only for a subset of users. Users on the opposite side of the table have to view this information upside-down or sideways. There has also been much work done on sharing and interacting with information on physically large displays such as wall projections. For reviews of work done in both these areas see [20, 21, 25].

There has also been a large amount of work done in building infrastructures that integrate these technologies into coherent environments, which typically allow multiple users to interact across multiple devices and hence collaborate with each other (e.g. [10]). We believe that this task, and ones like it, will prove useful as these environments mature and researchers move towards evaluating the power of these environments for information coordination and collaboration.

Within each of these technology scenarios, researchers are exploring the effects of particular media on the overall technology-mediated communication experience. For example, they would like to know the differences between text chat and richer visual shared displays, or between virtual and tangible interfaces. Within this domain, the collaborative job-shop scheduling task could be a useful paradigm for eliciting useful performance differences and communications patterns. Finally, even though most of the scenarios we have described have related to collocated settings, we believe that the task is general enough to be useful in evaluating coordination in remote coordination settings as well.

5. CONCLUSION AND FUTURE WORK

We have described the collaborative job-shop scheduling task, which we assert can be applied to evaluate interfaces that support coordination in computer supported collaboration environments. We have discussed evaluation measures and have shown examples, grounded in our experiment, of particular analyses that could prove useful. Results from the experiment demonstrate benefits of shared visual information when performing coordination tasks.

In future work, we would like to explore further variations of this task to test specific properties of group interactions. For example, we would like to explore versions of the task in which users do not have equal access to all information (i.e. a hidden information task). We would also like to explore other analyses, such as correlations between communication and tool usage, more detailed strategic analyses such as the amount of time or effort spent on planning vs. execution, measures of contribution by individual users, and social effects such as leadership and dominance. Furthermore, we would like to explore how tasks like this scale to larger numbers of people, as well as whether or not they allow us to adequately measure the trade offs that exist between the overhead of managing information and the benefits of shared visual spaces.

6. ACKNOWLEDGMENTS

We thank Patrick Baudisch, Andrew Faulring, James Fogarty, Brian Meyers, Jeff Nichols, Judy Olson, Randy Pausch, George Robertson, and Greg Smith for their discussions around the general task space, our specific job-shop task, and the experiment. Kori Inkpen provided valuable feedback on an early version of the paper. Daniel Robbins assisted with producing diagrams.

7. REFERENCES

[1] Beasley, J.E. (1990). OR-Library: Distributing test problems by electronic mail. *Journal of the Operational Research Society, 41(11)*, 1069-1072. For more information, see www.brunel.ac.uk/depts/ma/research/jeb/info.html.

[2] Bly, S., & Rosenberg, J.K. (1986). A comparison of tiled and overlapping windows. *Proceedings of CHI 1986 Conference on Human Factors in Computing Systems*, 101-106.

[3] Clark, H.H., & Krych, M.A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory & Language, 50*, 62-81.

[4] Endsley, M.R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors Special Issue: Situation Awareness, 37*, 32-64.

[5] Fisher, H., & Thompson, G. (1963). Probabilistic learning combinations of local job-shop scheduling rules. In *Industrial Scheduling* (J. Muth and G. Thompson, eds.), Englewood Cliffs, Englewood Cliffs, NJ: Prentice Hall, 225-251.

[6] Gergle, D., Kraut, R.E., & Fussell, S.R. (2004). Action as language in a shared visual space. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work 2004*, 487-496.

[7] Gutwin, C. (2002). Traces: Visualizing the immediate past to support group interaction. *Proceedings of Graphics Interface 2002*.

[8] Hoenig, J.M., & Heisey, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician, 55*, 19-24.

[9] Inkpen, K., Mandryk, R., DiMicco, J.M. & Scott, S. (2004). Methodologies for evaluating collaboration in co-located environments. *Workshop presented at the ACM Conference on Computer-Supported Cooperative Work 2004*.

[10] Johanson, B., Fox, A., & Winograd, T. (2002). The Interactive Workspaces project: Experiences with ubiquitous computing rooms. *IEEE Pervasive Computing Magazine, 1(2)*, 2002.

[11] Karger, D, Stein, C., & Wein, J. (1997). Scheduling algorithms. In *Handbook of Algorithms and Theory of Computation* (M. J. Atallah, ed.), Boca Raton, FL: CRC Press.

[12] Kenny, D.A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D.A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83*, 126-137.

[13] Kraut, R.E., Fussell, S.R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18*, 13-49.

[14] Kraut, R.E., Gergle, D., & Fussell, S.R. (2002). The use of visual information in shared visual spaces: Informing the development of virtual co-presence. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work 2002*, 31-40.

[15] McGrath, J.E. (1984). *Groups: Interaction and performance*. New Jersey: Prentice Hall.

[16] Mennecke, B.E. & Wheeler, B.C. (1993). An essay and resource guide for dyadic and group task selection and usage. *Institute for Research on the Management of Information Systems Working Paper #9309*, Indiana University, Bloomington, Indiana.

[17] Olson, J.S., Olson, G.M., Storrøsten, M., & Carter, M. (1993). Groupwork close up: A comparison of the group design process with and without a simple group editor. *ACM Transactions on Human Computer Interaction, 11*, 321-348.

[18] Pinelle, D., Gutwin, C., & Greenberg, S. (2003). Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Transactions on Human Computer Interaction, 10(4)*, 281-311.

[19] Rocco, E. (1998). Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. *Proceedings of CHI 1998 Conference on Human Factors in Computing Systems*, 496-502.

[20] Ryall, K., Forlines, C., Shen, C., & Ringel Morris, M. (2004). Tabletop design: Exploring the effects of group size and table size on interactions with tabletops shared-display groupware. *Proceedings of the ACM conference on Computer-Supported Cooperative Work 2004*, 284-293.

[21] Scott, S.D. (2005). Territoriality in collaborative tabletop workspaces. *Doctoral Dissertation, Department of Computer Science, University of Calgary*.

[22] Scott, S.D., Carpendale, M.S.T., & Inkpen, K.M. (2004). Territoriality in collaborative tabletop workspaces. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work 2004*, 294-303.

[23] Setlock, L.D., Fussell, S.R., & Neuwirth, C. (2004). Taking it out of context: Collaborating within and across cultures in face-to-face settings and via instant messaging. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work 2004*, 604-613.

[24] Sgall, J. (1998). On-line scheduling. In *Online Algorithms: The State of the Art* (A. Fiat and G.J. Woeginger, Eds.), 196-231. London, UK: Springer-Verlag.

[25] Tan, D.S. (2004). Exploiting the cognitive and social benefits of physically large displays. *Doctoral Dissertation, School of Computer Science, Carnegie Mellon University*.

[26] Tan, D.S., Meyers, B., & Czerwinski, M. (2004). WinCuts: Manipulating arbitrary window regions for more effective use of screen space. *Proceedings of CHI 2004 Conference on Human Factors in Computing Systems*, 1525-1528.