Methodologies for Evaluating Collaboration Behavior in Co-Located Environments

David Kirsh
kirsh@ucsd.edu

Vision

The challenges facing those who want to understand co-located (as well as distributed) collaboration arise directly from the absence of accepted theory about the nature of collaboration and interaction. Ask anyone what is involved in collaboration, and assuming you can get them to confine their focus to collaboration in a specified task, they will have different ideas about what is involved in the collaborative process. For instance, each researcher will lay less or more stress on the importance of situation awareness, collaborative error, coordination, interruption, attention management, shared representations, group metacognition and the power/social relations holding between collaborative members. There seems to be no agreed upon definition of these theoretical factors and no specification of the way they are expressed in behavior and performance. Even the theories of distributed and situated cognition have no precise statement of the phenomena to be studied or the theoretical constructs that are needed to explain observations. Is there any surprise there are questions about method when so little is known about the subject of inquiry?

My own vision of research draws on what I think most of us share about the near future: context aware everything and data surfaces everywhere. The fact that many of the controllers of these surfaces and devices are handheld or voice activated or will be embedded in our clothing, our glasses, or that the data will be painted on our retinas, is from my point of view a detail. It is environments conceptualized as superpositions of activity spaces that we are designing, and when thinking about such environments we have to focus on the full range of interactivity that human agents engage in.

This is a serious challenge because we do not yet know very much about the nature of this full range. For example, in my own research on interactivity I have been occupied with the many temporal levels and diversity of interaction found in most human environments, and how we humans are good at:

- Creating coordinating representations
- offloading processing or memory needs,
- adapting environments to minimize planning
- stabilizing environments to help us better tolerate interruptions
- stopping errors from becoming real problems
- and basically, adapting our environments to fit our cognitive styles, strengths, while compensating for our cognitive weaknesses

Solitary work on tasks also shows off this interactivity. But collaboration is a particularly rich area for finding and exploring this sort of thing. The point, however, is that we have to have such specific theoretical ideas in mind if we are to discover regularities or patterns that are generalizable or robust.

To be more precise, we cannot marshal evidence that people work better with handheld A than with handheld B when performing some task T_1 unless we also specify the contexts of activity. The notion of a context of activity takes into account the actions they can perform, hence the affordance landscape they are operating in, as well as other constraints on action trajectories and action consequences. Further information has to be supplied about resources or scaffolding as well as their expertise. We need this extra notion of a context of activity because the contribution which a given tool makes to performance has to be evaluated with respect to a large collection of other factors often lumped together under the umbrella term 'context'. Unless we have more structure on that notion of context, however, we will not be able to know enough about a situation to know whether the findings we collected in one situation generalize to the current situation. As always the challenge in design is to know whether a lesson learned in one design context will apply to another design context.

This same worry also applies to tasks. Suppose we have found that use of handheld A leads to better performance in tasks of type 1 than use of handheld B. Can we generalize this finding to other tasks? Under what circumstances? Once again the question can be answered only if we have a method of understanding the way the task shapes activity spaces.

It may seem that this worry is the standard worry of group usability, and in a sense it is, but my own orientation is to focus on how specific theoretical elements of collaboration are affected by changes in activity spaces and environments. The items of greatest interest concern the way different tools and contexts have an impact on vital processes such as

- Being vigilant – that is, tracking progress on certain state variables
- Stabilizing states – that is, immunizing mental or physical states against erosion because of interruption, multi-tasking etc
- Using coordinating representations – that is, creating, following or passing on representations such as lists, tables, and so on, which help a group to manage itself

Work on the mechanics of collaboration by Pinelle et al is helpful in understanding what affects these processes. The question is whether their mechanics are at the right level of analysis to illuminate the processes by which people do, for example, stabilize their activity, or the features of an environment which determine how well a team mate keeps track of the cost structure of putting someone else back on track. To be sure, information gathering is a key factor in my knowing when I can interrupt someone to tell them that I know they are wasting their time. And being able to protect work is a key component in stabilizing a resource or a task state variable. But the level of analysis we need to explore in order to understand whether one team is working better than another, or whether Tool B is more effective than Tool A , is not at the mechanics level, it is at the

level of our theoretical phenomena, and there can be many factors that are invisible from the mechanics standpoint that make the difference.

An example may help. We found in one of our studies that people negotiate rules of use of shared applications such as Chat, or a shared whiteboard, often in response to an error or miscoordination. For instance, in one study we found that team members realized that if they posted lists of to do's in chat then team members would soon lose easy access to those to do's. Chat lacks the right type of persistence to support long term coordinating representations. So the team made a rule not to use Chat for posting significant to do's. Now the value of this observation is only partly about the Chat tool. Focus on that mechanic without having the general concern about rules of use and you miss the real point. To me it is far more interesting that teams negotiate rules of use of tools and applications. That is the kind of question we need to ask in order to understand the dynamics of collaboration and how it is affected by technologies.

Another example we found has to do with vigilance. In my view errors are the norm and the only thing that stops them from becoming too consequential is vigilance. According to roboticists when we walk we are always about to fall and the wonderful thing about our gait is that we invariably catch ourselves before we reach the point of beginning to actually fall. In virtue of our gait and our vigilance we stay up and move forward. And as soon as we feel about to really fall we do something about it. This spotty but near continuous vigilance keeps us on track. IS there something like that in collaboration? Again, on my view, we semi consciously keep track of the cost structure of vigilance, interruption and repair. Naturally having tools and windows into other people's activity helps lower the cost of vigilance, but what I really want is something that supports my vigilance when I want it. This concern with the timing of vigilance and its cost structure, lies substantially above a concern with the mechanics of collaboration because what is important is when to be vigilant, not simply that I can be vigilant. Mechanics help, but we need higher level of analysis to know how to evaluate the mechanics.

The methods I personally use to study interactivity and collaboration run the gamut from fairly narrow experimentation where we can define and operationalize both the outcome space and many of the activity phenomena of interest along the way, to machine observation and low level data mining of 15 or 20 mins activity in games such as StarCraft, to more classical ethnographic study of 1 and 2 hour sessions of people working on a group task, where it is impossible to operationalize all the variables of interest, or even be confident that we have identified the most important determinants of activity and performance. To cope with the huge amount of data and to ensure that we are not losing records of activity that we will eventually realize is vital, I rely on video from camcorders trained on the activity, continuous screen capture of computer desktop activity, audio capture of each participant, and keylogging so we can datamine text input, context switching and the like. We have built a sophisticated digital ethnography environment which I would like to report on, and we have used that to study about 30 hours of distributed and collocated collaboration.

Experiences & Challenges

In our study of collocated collaboration we have slowly developed a methodology for ethnographic coding that is beginning to bear fruit. We ran a few dozen semi-structured experiments of 1.5 to 2 hrs each in which teams of three subjects worked as real time travel agents. In the experiment they received an email every day from their clients who were in Indonesia or Costa Rica or another 3rd world country. Each email contained a list of the day's goals: where should they stay the night, how to get their, what cultural event to see, where to dine for lunch and supper, how to get to the beach etc. A budget was mentioned for each day and the job of the travel agent team was to use the thousand or so web pages we provided as the source material to put together an day plan for the clients. They had thirty minutes to prepare this plan and then email it back to their clients. In a typical experiment they would have to prepare three day plans.

In another study our subjects behaved as interior design advisors. They were shown a collection of pictures of a house, given he desiderata of the mother, father and two children and a budget. With that money they had to furnish and arrange the living room, dining room, parents' room and children's room. They were provided with several hundred pages of furniture inventory as well as MS Office apps (powerpoint, word, excel), chat, ftp, visio, netmeeting whiteboard and a few other apps.

In building our research software our goal was to create an environment which would make it easy to extensively annotate video, often separating the tracks of annotation, between transcription, keyword coding, summarizing activity, and so on so that it would be easy to query a 90 min experiment for say *interruptions* and then review the video, the transcriptions and other annotations of all interruptions, to look for regularities or other noteworthy phenomena.

Our biggest challenge has been to find methods to speed up coding and reviewing activity. To facilitate faster review we have attempted to build filters that allow us to data mine AV and annotation tracks. For instance, we filter audio tracks for when subjects begin and stop talking. This can suggest occasions in the experiment when one or another subject is carrying on a dialogue or one subject interrupts another. We can similarly quantify the amount of time speaking, writing in the chat window, and so on. This can give us a coarse activity profile of users and support views we have about the degree of involvement different participants have. Creating these filters is one of our main goals in designing and creating our analysis software, but it has proven difficult for many reasons. Silence detection, for instance, turns out to be technically harder than one would guess. But it is doable. A bigger challenge is creating a set of primitive predicates and a language for composing them so that we can define arbitrary predicates over activity.

Workshop goals

My main goal is to learn about how other people are creating semi-structured experiments, collecting and then reviewing data. I have long intended to move closer to the CSCW community and this seems an ideal opportunity. I continue to do extensive work on collocated and distributed collaboration.

Bio

I am a professor in the Dept of Cognitive Science at UCSD. I have spent many years trying to understand the nature of interactivity, and for the last few years the nature of collaborative activity. I have been doing semi-experimental work trying to determine the differences between co-located and distributed collaboration, and in a parallel research initiative I have been exploring how experts use gesture, annotation and speech to pass the bubble of situation awareness in the strategy game, StarCraft.