

Workshop: Methodologies for Evaluating Collaboration in Co-located Environments
CSCW 2004
November 6, 2004

Disclaimer: as they stand, these notes reflect the workshop through the eyes of a single person and may or may not have accurately captured the thoughts and intent of the group as a whole or any individual contributor.

Introductions

- Kori Inkpen
 - How to measure effective collaboration
- Regan Mandryk
 - Using physiological indicators as metrics of collaboration
- Stacey Scott
 - Interested in how to integrate interaction with digital media into a richer environment – added benefit from paper environments.
- Joan Morris DiMicco
 - Interested in displays that are augmenting but not mediating communication in information sharing/decision making tasks.
- Mark Hancock
 - Awareness/speed/accuracy tradeoffs in group interactions.
- Darren Gergle
 - Interested in collaborative physical tasks (co-located and distributed).
Using discourse analysis and sequential analysis.
- Sukeshini A. Grandhi
 - Role of pace in design of interactive location-based systems.
Understanding communication needs in different place types.
- J. Karen Parker
 - Evaluating Tractor Beam interaction technique in a group work scenario
- Lisa Kleinman
 - Uneven technology use in a group collaborations. Examining the positive and negative aspects and how to evaluate it.
- Markus Klann
 - Evaluation methodology for mobile collaboration with wearables – want to be able to compare across applications domains. Co-presence of people that are communicating across rich channels (not necessarily co-location)
- Meredith Ringel Morris
 - Supporting collaborative work around tables including the ways to switch between personal and group work. Ways of supporting individual and group needs including interaction techniques and privacy issues.
- Nikhil Sharma
 - Working on the Collaborative Sensemaking project – looking for information on-line and making sense of it together. Effective collaboration is hard to evaluate – interested in getting a taxonomy of tasks for collaboration.

- Sheelagh Carpendale
 - People visualize things to share with others. Observational studies have been informed by complexity theories. Will the metrics that come from an understanding of complex systems aid with evaluating collaboration (another complex system).
- David Kirsh
 - Foundational questions required for articulating a design science. Understanding how people can collaborate effectively (co-located and remote) if share a digital environment. Context of a mediated coupling are so rich, entities coupled are deep and profound – current approaches are simple and may not illuminate the key issues.

Definitions of Collaboration

- People come together: task and a goal.
- Benefits to collaborating – often a form that is not easily measurable.
 - Emotional support
 - Affirmation of ideas
 - Often intangible
 - § Remote collaboration seems easier – face to face is given as the ideal, anything is better than nothing. But with co-located we are already starting from the ‘ideal’ – harder to define/measure the increased
- Cross fertilization
 - § Everyone coming to the table has their own knowledge
 - § skill sets (structure makers)
 - § Increasing the number of perspectives. Redefining the search space.
 - § Window of vitality – need a certain number of ingredients to reach that collaborative vitality, but also a threshold - too much can make worse
 - Essential elements:
 - Diversity
 - Communication
 - § Full spectrum, both explicit and implicit. Serendipitous exchange of knowledge as well as formal learning.
 - § Being able to communicate with each other. Can share knowledge by doing – tacit knowledge, learning by demonstration. Tacit knowledge can be hard to quantify, can happen more readily with proximity. Feedback is available. Certain situations can’t be created unless in the co-located collaborative environment – are exclusive benefits of collaboration.
- Unexpected results

- More interesting to collaborate than work alone – gain energy and motivation.
- Increases the discourse domain – social activity. Introduces contrary-wise thinking that's essential to growth or can reinforce beliefs.
- Essential elements
 - Coordination
 - § Mechanisms (often very similar to benefits if rephrased).
 - Levels of mechanics
 - Atomic elements
 - Gutwin's mechanics of collaboration
 - § Hard to cut across and say 'in this case, need to know when people will interrupt in a positive way' – need interruption, how to facilitate good interruptions
 - Sharing of information
 - Integration of information
 - Coordination of (physical or cognitive) actions
 - Division of work – some tasks encourage parallel work
 - Different roles
 - Different granularities
 - Use of coordinating representations
 - Linguistic development – convergence on common definitions, developing a domain of discourse
 - Gestures
 - Interaction with artifacts/knowledge
 - Negotiation
 - Cooperation
 - Task dependent – goals may not be common, but still collaborating.
 - Levels of cooperation. 2 lawyers, single goal of reaching an agreement sets the framework. Within that, goals may be very different and may change over time.
 - Benefits can be at the group level, at the individual level...
 - Cooperation and collaboration and not the same thing.
 - § CSCW – cooperative work
 - § Collaborator – cooperating with the enemy – negative connotation particularly in Europe.
- Domains:
 - § Game playing, problem solving, etc.

- Face-to-face technology (collaboration and evaluation)
 - § Asynchronous co-located collaboration (shift changes, moving between collaboration and individual work, attention changing)
 - § Tabletop displays
 - § Multiple displays (handhelds)
 - § Wall displays
 - § Wearable computing (head-mounted, hands-free)
 - § Tactile interfaces
 - § Tangible user interfaces (digitally augmented)
 - § RFid
 - § Non-digital physical artifacts (traditional)
 - Laser pointers – augment coordination
 - § Inputs
 - Eye tracking (additional input on what somebody is doing)
 - Biometrics

Measuring Successful Collaboration

- What is effective/successful collaboration?
 - § Did people accomplish their explicit goals?
 - § Did people have serendipitous benefits – extras?
 - § Did people have appropriate awareness of each other (which then leads to a benefit of collaboration)?
 - Actions
 - Intentions
 - Feedback
 - § Were people satisfied with the process? Did they enjoy it?
 - § Perceived effectiveness/benefit which can lead to willingness to continue collaboration over time?
 - § Need to differentiate between the other collaborators and the technology.
 - § Are people interacting with each other (and having fun)?
 - § Has existing collaboration been enhanced by the technology?
 - § What is the impact of the collaboration? (Personal, group, result)
 - § Was the collaboration efficient?
 - § How easy is it to interact (with each other, with artifacts)?
 - § Is the group aware of their own dynamics? Reflection on group performance?
 - § Seamlessness / fewer ‘breakdowns’
 - § Sense of community?
 - § Skill transfer?
 - § Does technology take you away from the collaboration?
 - § Scales well with different levels of coordination?
 - § Does the technology frame the interaction appropriately for the group dynamics. Are the people supported in the roles that they wish to play.

- § Assessing something that exists or evaluating the things that have been discovered as people collaborate (serendipitous from the researcher's point of view)
- § People are on-track, able to manage the processes.
- § We get the most from each participant. Competency and knowledge of participants. Look at propagation of ideas rather than frequency.
- § Does the collaboration make best use of existing resources? What resources and tools are you using, which technologies best support those?
- § Does the collaboration adapt well to environmental changes? How robust is the collaboration? Is it flexible, does it adapt when it needs to?
- § Collaboration A is better than B? Or is it a different style of collaboration? If you have collaboration A as the baseline, can then look at collaboration B.
- Information gathering session – talk and report on the types of methodologies you know in particular fields – benefits for studying collaboration
 - § Qualitative (observational)
 - Stacey
 - Talked about different methods of observation – involving videos, ability to break down interactions later
 - Pro/con: a huge amount of data. Rich, but time consuming. Filter it by what you focus the camera on and then when analyzing it – what is transcribed/coded.
 - Hard to figure out what to focus on when in the field environment.
 - Are you capturing all data or sampling some data? What do you choose to sample.
 - Coding scheme motivated by an ontology. Could be developed collaboratively.
 - Retrospective – video tape somebody, review it with them later, what were you thinking, what motivated you? Get the subject's insight instead of the experimenter's judgment. Issues of sense making of actions.
 - Contextual interviews. In a natural environment, probing them about what they are doing.
 - David's methodology. Disaster response teams. Had teams run through the script of what they expected to happen. Props in the structure of the space. Simulate everything by hand. Took videos. Gave them a new board, asked to recreate the simulation. Interesting things about what they did/didn't remember.
 - How involved should participants be in your research? How many participants can you ask to give you that time. In early stages of design, may be necessary.

- Traditional subject interviews – telling you what they think you want to hear or what they think is appropriate to say.
 - Participant diaries, beeper studies (experience sampling methods)
- § Quantifying the qualitative processes – qualitative measures (eye contact, deictic references) – measuring behaviour rather than performance – things that generally aren't logged. Can be very time consuming.
- Regan, Darren (has background in linguistics), Kirstie, Kori, Suke
 - Coarse representation: overall amount of communication info (word counts), rates (speech per unit of time), as action goes up, speech rates can go down, deictic use (help to create commonality)
 - How to automate the process
 - Can work better in the distributed community – references can be very subtle and may not be easily observed – can be implicit coordination.
 - How to quantify the fluidity.
 - Lexical entrainment – agreeing on a common term to use – shows a shared understanding - fluidity in conversation, grounded in that conversation, less error and ambiguity. Happens over time. May be able to measure the compression of terms (Darren's new cool idea)
 - Collaboration changes in fluidity as well as speech.
 - Big drawback is the time consuming nature. Machine learning doesn't work well unless it's a very constrained task.
 - Higher levels – brainstorming, refinement, resolution. Ann Anderson (?) has developed some ways of coding)
 - Are there set ways to evaluate?
 - § Linguistic analysis
 - Roger Baakemn (?)
 - Darren is looking at sequential analysis, looking at speech stream, action stream, etc. – is the action associated with verbal action happening (efficiency/fluidity measure)
 - General process that the group is doing in a mediated and non-mediated solution – process changes are interesting even when some things like time and solution stay the same. Not just an outcome. Olsens have published in this area.
 - You can know how the process differs, but don't know if it's better or worse? Does it matter? Could maybe regress with group satisfaction, etc. Correlations between process or performance. (ie. Groups that do

more brainstorming end up with higher value about group).

- Gestural stuff: eye contact, gaze, pointing, body language
- Hard to replicate the codes across people, hard to quantify. If all 'one-offs', then no external validity. As a community can we start developing our own. When drawing from other fields, metrics may not be specifically for collaboration.
- Time consuming to set up and run – could there be a shared corpus? A repository of coding schemes (here's the task, here's the coding scheme we used...) (This was also discussed in the qualitative section)
- Analyzing a rich data set and being able to report on it in a way that others can understand. Would be easier with similar coding schemes.
- Difficult to do comparisons. Coding schemes may not apply to all scenarios (control and condition). Need to develop methods that don't require controls. Could be creative about it. Might be able to define theoretical background conditions that constitute controls.
- Need to discuss more after the workshop. What we've done, Language, gesture, face,
- Constrain for individual differences (realism vs. precision)

§ Quantitative metrics (speed, timing)

- Joannie, Nahil, Karen, Mark, Meredith
 - Quantitative outcomes
 - Lots of challenges
 - Related to effectiveness of collaboration
 - § Explicit goals accomplished
 - § How efficient were they
 - § Perceived effectiveness
 - § Learning from the experience (did become better)
 - Easy to measure things, defining as being 'good' or 'bad' is subjective. Ie. Is efficiency necessarily a good thing?
 - Outcome of the task can be a metric of success. Is a task a recommended real use scenario or just something to evaluate technology with.
 - Individual difference, need large populations, hard to get, but important for quantitative results.
 - Do you do the analysis at the individual level or the group level? Individual behaviour is correlated with group membership. David Kenny (U of Connecticut) has a good paper about dealing with this issue statistically.
 - Using standard deviation as a metric.

- Use Myer's Briggs to help separate out individual differences. Bate's sim log will help give metrics for the group and their roles. Can create hand-tailored metrics.
- Quantifying questionnaire data – perception vs. metrics.

Designing Collaborative Tasks

- Task framework – categorized them on various dimensions. Haven't been able to synthesize yet.
- Brainstorm:
 - Tasks that have used:
 - § Groups of users select songs for a soundtrack – visible representations of scenes from a movie and song. Objective to evaluate a shared system with audio.
 - Engaging and fun, have to manipulate lots of objects
 - Not a real-life task, no right or wrong answer (group consensus)
 - § Biologists used a system to categorize things
 - Realistic system
 - Hard to find a group of users that are experts
 - § Memory game
 - Relies on memory – lots of individual differences – too cognitive
 - § Magnetic poetry – selecting words to copy a poem. Search task.
 - Performance metric was hard. 4 faster than 3 faster than 2.
 - § Slow tetris over PDAs – collaborative (one could rotate one way, one the other way. Goal: see if 2 distributed colleagues
 - Spatial reasoning skills play a big problem. Was screening for a certain level.
 - § Puzzle solving
 - Goal to look at collaborative process
 - Were able to look at collaborative process – able to manipulate orientation
 - Realistic task? Reviewers resistant to games.
 - § Referential communication
 - Field work added too much noise
 - Moving it into a lab w/ lego like structures
 - § Climbing – math collaboration game. Goal: how does display impact collaboration
 - Led to shared understanding
 - § Furniture layout task.
 - Looking for inspiration for design of tabletop displays
 - In-depth spatial analysis of what they did – interested building widgets, not supporting collaboration
 - § Laptop use in class.
 - Note taking – ½ screen, IM windows, etc.

- Ok to tune out in classroom
- Trying to understand the methodology
- § Real-time travel agents
 - Co-located vs distributed
 - How to get enough interactivity with people
 - 30-90 minutes collaboration
 - Outcomes: task completeness, task goodness, robustness, satisfaction, beliefs about performance, vs. actual performance
 - Process constraints: requires interactivity and sharing, need a set of resources that can be used, needs structure with subgoals so can look at intervals
 - \$ figure, travel constraints, come up with an answer w/in 30 minutes.
 - But expertise required.
 - Watch over time to see if microgenetic change over time.
 - Not just looking at errors – looking at stabilization (immunization from interruptions) – theory of stabilization – virtue for inclusion in model. Can now measure info encoded in the state. Go back to the data and re-do it looking at items.
- § Interior design problem – each fake family measure has wants, budgets, pictures, etc. re-decorate.
 - Lots of interaction, highly collaborative tasks
- Pros / cons with tasks:
 - Realist tasks require multiple experts to evaluate
 - § Not product development, but has to work for users' purpose. Research stage prototypes aren't sufficient
 - Motivation issues with mandated tasks
 - Realism vs. precision
 - Generalization
 - Individual differences as a result of poor task selection can impact ability to see effects from variable manipulation
- What are you trying to evaluate and what metrics are you using? Having to conjecture what might be a task that can show something. Generalization issues – what conditions make it generalizable or not.
- Integrate theory and re-visit the data. Difficulty if you haven't designed the study to collect that kind of data. Violation of statistical testing theory.
- Do observation, develop theory, re-assess data in fine-grain and see if theory is supported in data.
- Repeat experiment slightly tweaked to see if theory is supported by other data.
- Validation of ad-hoc observations? Or fishing?
- Task standardization (Darren)

- Characteristics of the task are re-used in a number of studies
- Driven by a real task in the domain (remote instruction/guidance)
- Based on task from the 60's – how people describe things and objects in an environment to one-another
- Found referential communication tasks in psychology. Not looked at in modified settings. Added that to the task. Started high level. Too noisy to learn what's valuable and what's not. Controlled field of view. Refined to puzzle task. Realistic task down towards something very constrained – retains one component. Can it be taken back up to the realistic task? Can a task be decomposed so far that it is no longer useful?
- Task dimensions that might be important
 - Generalizability
 - # people
 - Simple vs complex
 - Can it be done by one person
 - Still needs to be framed with respect to task qualities important to collaboration(?)
- Where does collaboration naturally occur? Need to ground in real world, but may miss opportunities that technological innovations that may provide.
- Are there genres of problems for which different styles of tasks might be suitable?
 - – interested in coordination. One aspect is style of coordination – divide and conquer, mutual coordination. Photo layout task (Friends, Lord of the Rings) – create a collage of 4 themed pages for fan layout. More pages than people. Could split it up if wanted to.
- Is there an obligation to study the task without the technology. Perform the treatment. Now have something with which to contrast. Do you then have a background against which to place useful observations.
- Piggy-back onto another field and what they do without technology. Can use as a base. Many previously defined social/psychology theories can help explain what's going on. “How to do things with things” Streak
- How people collaborate co-located now and why do they do it (like use email for file storage instead of ftp). What works/doesn't. Try to develop things that build on it.
- Input literature has a number of different tasks: funneling task, 5 kinds of pointing tasks. Collaboration is so heavily influenced by tasks at use. Collaborative tasks are much more complex though. Need to take it down to something lower. Mechanics.
- Task vs scenario vs. task elements
- Standardized tasks
- Standardized task characteristics
- Standardized metrics
- McGrath's taxonomy gives some guidelines: degree of interdependence, etc. Can be somewhat useful. But can be very complex in the end.

- Task dimensions (cultural biases have an impact)
 - Fun factor
 - Amount of interaction
 - § Between people
 - Verbal
 - Touch gestures
 - Eye contact
 - Gaze
 - § With resources
 - Passing between people
 - Added onto by people
 - Manipulation
 - Add/delete/modify
 - Nature of coordinating mechanisms
 - Clock
 - Other objects (have an effect on collaboration)
 - § Format (keeps on task)
 - hierarchical task / non-hierarchical
 - roles – social hierarchy
 - Number of people, performance (relationship)
 - § Assign each task a performance contour (brainstorming – would larger groups subdivide)
 - Homogeneity of the population

Metrics: measurable aspects – develop a code. What kind of metrics have we looked at to see if collaboration is happening.

What is effective collaboration –benefits of collaboration – measuring collaboration

- Did people accomplish their task goal?
 - Pre-defined task: yes / no, % completed
 - Open-ended task (no right answer):
 - § Individual answers? Group consensus?
 - How fast was the goal achieved?
- Appropriate level of awareness of others' contribution
 - Eye-contact: more or less could be good
 - Self-reporting
 - Response time
 - Time dimension – overall awareness
 - how to measure awareness?
 - § Can be if somebody is doing something else and you've reacted to it
 - § Mistakes made because of non-awareness
 - Collisions
 - § Watching video and see if any verbal indications of awareness

- § Priming task after the fact (Tan & Czerwinski)
 - § Interrupt and ask what the current state is
 - § Awareness plotted to time
 - § cost-benefit of being aware at times
 - cost of maintaining awareness, interrupting self, benefit of knowing, constantly fluctuating
 - function of media to know when to interrupt based on cost-benefit
 - § long-term awareness vs. short-term (second by second)
- User satisfaction
 - Self-reporting
 - Will they actively try to use it again – adopt the technology
 - Cardinal (actual value (user survey)) or ordinal (lots of things to choose), how much would you pay for this, how much effort would you make to use it,
- Perceived effectiveness
 - Self-reporting
- Interaction with each other
 - Mirroring
 - cultural differences, gender differences
 - Towards the group goal
 - Task dependent
 - Amount and balance of interaction – floor control – depends on goal, roles, etc. – equal may not always be optimal
 - Discourse analysis – how long a topic stays alive even if not constantly interacting.
 - How much did you learn about the other person
 - Gestures – richness of dynamics – spectator vs. active participant
 - Physiological measures - many factors contribute
 - Can simulate over time? Hard to visualize.
 - Get a measure of satisfaction about the interaction level (rather than the whole collaboration)
- Outcome/impact
 - Have had aspects of impact in previous levels
 - Immediate impact or impact on subsequent collaborations
 - Impact on what? Knowledge, perceptions of others...
 - Pre and post. Get a prediction from them and ask them after the fact. Can see how the technology impacted their perceived notion of the task.
 - Speed/accuracy/efficiency.. (not sure if this was supposed to be under goal)
- Communicative efficiency
 - Time takes to complete & efficiency of communicating it
 - Amount of discourse it takes
 - Speed may not be an indicator of good collaboration – task dependent

- Units: linguistic tokens – durations, # words used, speech rate - picture may be very useful, but won't show up the same as with tokens. Semantic units.
- Seamlessness / lack of breakdowns
 - Characterize a breakdown or interaction depending on the task and code (taking somebody's object without permission, interfere with somebody else)
 - § Depends on what group is trying to accomplish – often inferred by researcher (implicit permission given?)
 - § People are doing social things the best they can – not judging the people, but how the technology surprises them and does something intrusive – but we do design technology to intentionally interrupt
- Sense of community
 - Did become a member of something
 - Was a shared understanding reached
 - Did they become 'collaborators'
 - § How much interaction after than before
 - May need a 'real' group to be able to know
 - How much do individual team members perceive how much a part of the community they really are? Self-reporting may not be enough
 - Entitativity (sp?) – how tight do people feel – measure in psychology literature
- Not detract from face-to-face interaction
 -
- Process management
- Many different view points, ways of looking at this

Open Issues, Publishing

- Individual differences – how to design groups
- Impact on collaboration and task
- Knowing each other
- Introversion/extroversion
- Everything is interdependent
 - Hypothesis testing – need independence
 - Hierarchical individual modeling
 - § Assessing amount across each dimension
 - § Judith Singer - Harvard
 - Inter-class correlation coefficient – calculate difference – can decide whether to treat as a group or as individuals
 - David Kinney – statistical methods
 - May get richness from the qualitative analysis
 - Normalize data – if function is too complicated outside boundaries, can perhaps know things about parts of activity that are well-behaved.

Partition what can be checked over everybody and what is reduced to individual differences. Vision problems use this type of approach.

- Post-normal approach

Challenges of publishing with non-standard approaches:

- ‘difficulties’ with getting published
- Some people’s publication rates may be high, but they’ve been through several submission rates. Taking a 2-3 year stand on getting a paper out. Conferences used to be about fast turnaround – now turning into journal quality.
- Difficulty in getting co-located collaboration stuff published – it’s messy. If tightly controlled, get problem with no generalizability – ‘toy problems’. If not controlled, it’s very hard to get something solid.
- Sometimes it’s methodological issues – different expectations within the community. If we have a common ground, it might be easier to meet the expectations of reviewers in the community.
- Address the reviewers comments you know that you’re going to get about the non-standard approaches. May need to convince them that there is rigor in qualitative research.
- Eco-system studies also don’t fit into the ‘scientific approach’ – destroys the eco-system to reduce the complexity of the problem. Still no solution in that community. Group collaboration has similar problems. If you make it a lab experiment with metrics to collect, does it still have realism? Will the things learn in the lab still be true in the real world. Need to devise a new formulation that we can accept in the community.

Next Steps

- Special issue on co-located collaboration
 - Introduction to the special issue could address the methodological issues
- Include reference samples of what a good qualitative paper would look like.
- Need to get more keywords added so that we get more appropriate reviewers
- Consciously try to re-use tasks and methodologies. If reference others’ works, it may help the perception that this is a published standard. Also gives something to compare to with results. Build on each other’s works.
- Repository
 - WIKI to help find the links to social theory that may help us deal with our data?
 - Converge on task elements? Find more consistent task elements.
 - Repository of experimental software. Description of task, pictures, code if possible.
 - Put the workshop account on the WIKI

Deliverables (specified early in the day):

- Specific things we might be able to get out of today – something tangible to show at the end of the day
 - § List of benefits of collaboration and at least one metric for quantifying each of the benefits

- § Taxonomy of tasks which can be employed to evaluate benefits/problems
- § Types of metrics for measuring collaboration
- § Integration between quantitative methods and qualitative evaluation
- § Coordinate