Visualizing Language Use in Team Conversations: Designing through Theory, Experiments, and Iterations

Gilly Leshed
Information Science
Cornell University
301 College Ave.
Ithaca, NY 14850 USA
gl87@cornell.edu

Jeff Hancock
Information Science
Cornell University
301 College Ave.
Ithaca, NY 14850 USA
jth34@cornell.edu

Dan Cosley
Information Science
Cornell University
301 College Ave.
Ithaca, NY 14850 USA
drc44@cornell.edu

Geri Gay
Information Science
Cornell University
301 College Ave.
Ithaca, NY 14850 USA
gkg1@cornell.edu

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Abstract
One way to potentially help people develop effective teamwork skills is to visualize elements of their language use during team conversations. There are several challenges in designing such visualizations, such as how to balance attention between the conversation and the visualization and how much guidance to offer about appropriate behaviors. We discuss the design space around these questions in the context of GroupMeter, a chatroom augmented with visualizations of language use. We generate and critique potential answers to these questions using prior theoretical and empirical work, then describe how the interface evolved and how our answers changed over a series of prototypes we deployed in experimental studies. We conclude with the lessons from our experience that could be used by designers of collaboration-enhancing systems.

Keywords
Teamwork, visualization design, linguistic analysis, behavioral experiment.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces – *CSCW*.

General Terms
Design, Theory, Experimentation.

Introduction

Teams are sometimes ineffective not because they lack the right tools to coordinate, collaborate, or communicate, but because their members don't always have the teamwork skills necessary for effective collaboration. A key aspect of effective teamwork is therefore developing behaviors that support the team interaction process. Recent work shows that analysis of language use can reveal some teamwork-relevant behaviors. For example, using self-references ('I', 'me') is associated with involvement in the interaction [4], frequent agreement ('yes', 'ok') is associated with passivity [25], and the use of justification words (e.g., 'because') indicates task focus [38].

These findings suggest that knowing what language to use during a conversation can be an important skill for effective teamwork. Further, presenting visualizations of linguistic behavior to team members can potentially stimulate reflection and guide team members to adopt effective teamwork behaviors, especially when training teams to collaborate within a mediated communication space. With these goals in mind, we designed GroupMeter, a chat system that visualizes linguistic metrics such as frequency of emotion words, level of agreement, and overall participation level generated from the group conversation.

Systems that present this kind of linguistic information pose a number of design challenges. We address five key questions in this paper: (1) When, during the team interaction, should awareness information be presented? (2) How should attention to feedback and conversation be balanced? (3) Should group or individual level feedback be displayed? (4) What kinds of feedback should be computed and how should they be interpreted? (5) How much normative guidance should be provided in a given context?

In this paper, we explore the design space around these questions in the context of the GroupMeter system. We present our initial answers to these questions, driven by theoretical and empirical work in human-computer interaction, social psychology, and cognitive science. These principles then meet practice, as we describe how GroupMeter's user interface—and our answers—evolved over a series of prototypes we deployed in experiments.

Unlike previous descriptions of GroupMeter [25][26], which looked at specific versions of the system and focused on behavioral experiments, our goal here is to present a case study of how its design changed over time as the research evolved. As HCI scholars and designers, we all face difficult design decisions, trying to settle conflicting goals and balance design tradeoffs. Tools developed in research settings often must deal with extra constraints: they need to account for theory and prior work, and meet research goals while demonstrating ecological validity. We hope that our experience helps other designers building similar systems and facing similar challenges. We also hope that the more general story of how theory and practice shaped our designs will be a useful case study for the HCI design and research communities as a whole.

The GroupMeter System

To ground the discussion of GroupMeter's design goals and evolution, we start with a high-level description of the system's design and architecture. GroupMeter is a web-based system in which groups communicate through chat to perform tasks while receiving dynamic

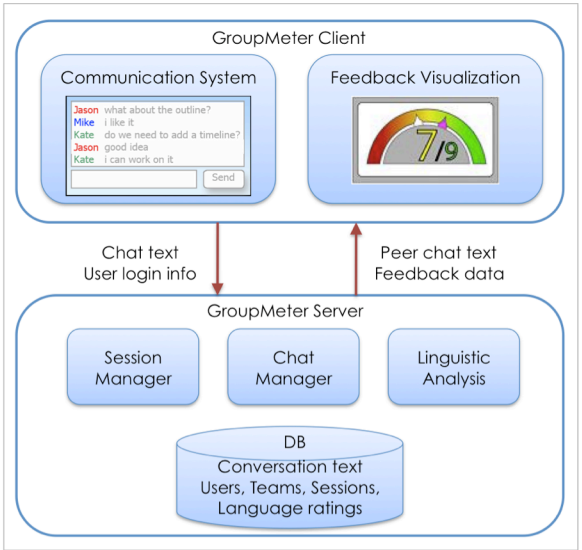

Figure 1. The GroupMeter software architecture.

feedback about features of their language use presented by a visualization that appears near the chat window. The linguistic features are generated using a technique based on Pennebaker's Linguistic Inquiry and Word Count (LIWC) [34], which counts percentages of words in the chat text that fall into categories such as positive and negative emotion, self-references, and justification words. The choice of a chat communication medium over speech enables near real-time computation of these linguistic features; the visualization dynamically updates each participant's scores on the features as they are computed.

GroupMeter is implemented using a web client-server architecture (Figure 1). The server manages sessions that specify the names of team members, linguistic metrics to calculate, and the feedback interface to display. Team members log in to the website and communicate through the chat. As they talk, a chat manager on the server monitors the conversation and sends the chat text to the linguistic analysis module for processing. After analyzing the text, the server sends the computed metrics to the front end, which presents them through the visualization specified for the session. The modular structure allows elements to be replaced easily: changing the linguistic computation leaves the visualization design unaffected, and vice versa.

Theory-Informed Design Decisions

As stated in the introduction, tools designed to raise awareness of social behaviors by visualizing language use in a collaborative activity pose a number of questions. Here, we discuss the questions we faced during the design process and present possible solutions based on theories and prior empirical work from disciplines including human-computer interaction, social psychology, and cognitive science. We use a prototype developed early in the project (Figure 2) to ground the discussion and illustrate some of the choices to be made.

1. When, during the team interaction, should awareness information be presented?

According to Bales' Interaction Process Analysis [1], effective teamwork involves balancing task-related with socio-emotional behaviors. While it is easy to focus on behaviors aimed at completing the task, being aware of socio-emotional behaviors such as language use and word choice helps maintain the team's well-being, and is therefore important for effective teamwork [32].

A key question, then, is when to present awareness information about socio-emotional behaviors during the team interaction process. One option, based on Gersick's punctuated equilibrium model of group development [14], is to provide information about such behaviors at transition points in the team's life span. At such points, team members will be most willing and able to reflect on their behaviors and change them in subsequent team sessions. For instance, visualizing participation patterns after one task was shown to lead to more equal distribution of participation in a subsequent task in face-to-face settings [7][33].

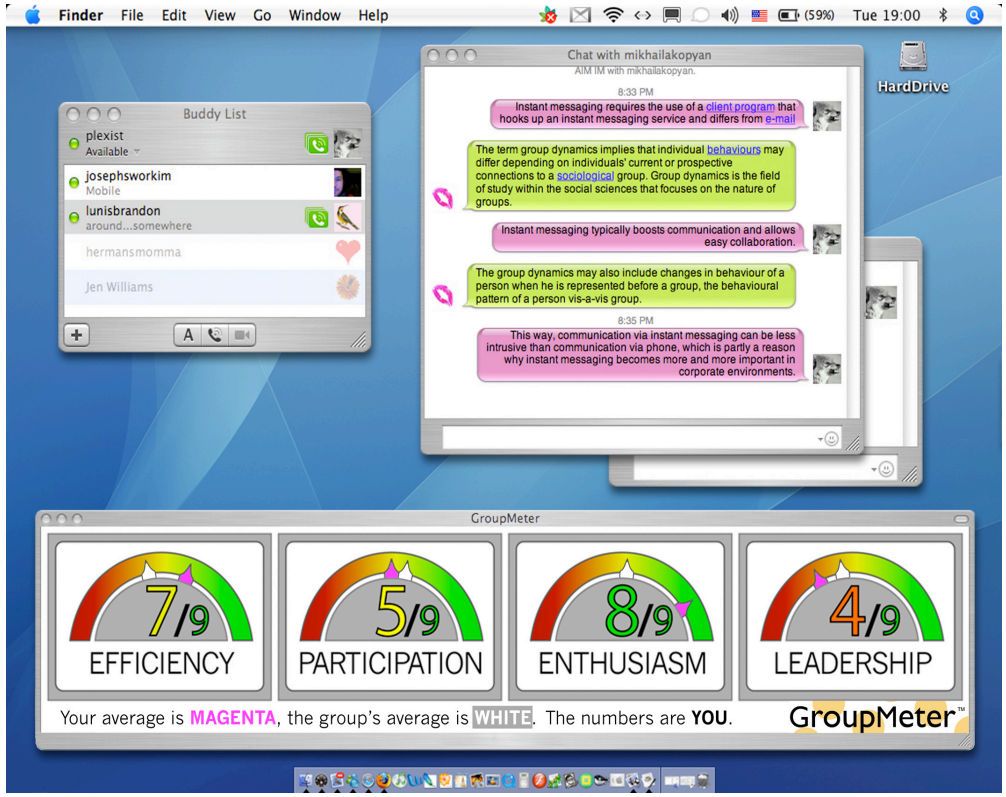

Figure 2. An early prototype of GroupMeter. It interfaces with external chat tools and presents feedback dials in a separate window.

On the other hand, consistent with Kluger and DeNisi's Feedback Intervention Theory [23], an ongoing, dynamic visualization enables individuals to see more clearly how their behaviors are linked to the feedback they receive. This creates a *self-monitoring tool* that allows people to review and modify their behaviors as needed in real-time [12]. Other systems to support ongoing awareness of social behaviors have used a number of dynamic displays, projecting representations of behavior on individual's desktops [8][16], a wall [7], a table [2], and on members' cell phone screens [21].

Because people's production of language in conversation is largely spontaneous and unconscious [27], we initially decided to present the linguistic information dynamically—making people aware of their behavior in the moment—rather than at intermissions between tasks. In Figure 2, as team members modify their behaviors intentionally or unintentionally, the indicators on the dials and the numbers change dynamically.

2. How should attention to feedback and conversation be balanced?

Dynamic visualizations, however, risk distracting team members from the task. Thus, we had to consider ways to stimulate awareness of the team's socio-emotional behaviors without interfering with task performance. We therefore decided to display the information about language use at the periphery of the interface. Such *peripheral displays* promote awareness of background information, are not part of a primary activity, and do not overload users with too much cognitive effort [30]. Peripheral displays have been widely used to present awareness information in collaborative settings (see a review in [19]).

Presenting awareness information in a peripheral display requires individuals to divide their attention and shift smoothly between primary and secondary information sources [42]. Visual design can support or hinder these shifts; for instance, in Figure 2, the feedback meters are physically separate from the chat window. This reflects an initial design goal to make GroupMeter work with many communication tools; however, the

visual distance between the chat window and visualization could make it harder to move back and forth between them.

Both the issues of timing and location call out a key design challenge around managing people's attention: while evaluating the various designs we created for the feedback visualization, we always monitored the tension between awareness of teamwork behaviors and distraction from the team task and conversation.

3. Should group or individual level feedback be displayed?

Another design factor that needs to be considered when presenting awareness information to a group is the level at which this information is aggregated and publicized: should behavioral information be kept private at the individual level, compared to an aggregate, or made publicly available to the group? This may sound like an odd question; after all, behaviors in team settings are inherently public, and team members make judgments based on others' behavior all the time. However, technologies can make aspects of behavior salient that otherwise might go unnoticed or simply require more time to be perceived and made sense of in a mediated environment [40].

The prototype in Figure 2 presents each individual's behavior privately (magenta markers on dials), along with an aggregate average for the group (white markers). Keeping individuals' information private can reduce potential stress caused by exposing behaviors publicly, thus avoiding embarrassment and discomfort [36]. Further, designs that lead to people comparing themselves against each other might encourage competition and negative interpersonal processes such as low trust, low coordination of effort, and attempts to mislead others [20].

An alternative is to make everyone's individual-level information available to all team members. That is, every member sees the feedback information of every other team member, rather than just seeing their own information or seeing an aggregate of the team's behavior. Based on social comparison theory [11], publicly presenting information about individuals allows members to interpret the feedback about their own behaviors in comparison to the behaviors of others. Further, based on functional leadership theory [17], a strong team member or leader can observe others' behavior and intervene actively to motivate and direct others to change their behaviors [22].

Providing public individual-level feedback also supports social translucence [9]. Social translucence emphasizes making social information visible within a system, supporting adherence to social norms through awareness of others and accountability of the individual's own behaviors. Publicly showing each individual's feedback can increase accountability and, as suggested by Festinger [11], affords explicit comparisons.

We decided that the advantages of publicly presenting individual information—social comparison, leadership facilitation, and social translucence—outweigh the concerns users may have with having analyses of their behaviors made public. Still, the design needed to negotiate this tradeoff, and as we will see later, aspects of the design did affect whether people reflected on their language use and whether they changed their language for the good of the team or simply to explore the visualization.

4. What kinds of feedback should be computed and how should they be interpreted?
Another important design question was to select which data to present and how much to process the data. In principle, systems can compute an enormous amount of information about language use—LIWC, for instance, has over 70 linguistic categories [34]. To reduce potential distraction and the complexity of the interface, we chose to focus on linguistic metrics that correlate with the language used by people perceived to be good team members. We found that peer ratings correlated positively with overall contribution (measured by word count) and frequency of achievement-oriented terms, and negatively with frequencies of emotional terms and agreements in a group decision making task [25].

Once we chose these linguistic metrics of word count, self-references, emotional terms, and agreements, the next question was how to process and present them. There is a tradeoff between presenting the behaviors in raw form versus attempting to map them onto higher-level collaborative concepts such as leadership and enthusiasm [13], as shown in Figure 2.

Mapping raw data onto higher-level concepts has some intuitive value; for instance, "high leadership" might be more meaningful to people than "uses many self-references." However, these kinds of mappings put much of the burden of interpretation on the system designers, reducing the flexibility of the system. Less interpretation-laden representations allow users to develop multiple understandings of what the system is for and how to use and experience it in different contexts [15]. For instance, expressing agreements might be valuable for certain kinds of discussions and downright damaging to others; thus, mapping agreement to "enthusiasm" or "friendliness" would make it hard to express disagreement even when it is appropriate. Further, with high-level feedback it might be hard for team members to understand how their language use and choice of words corresponds to the visual display [23].

Thus, rather than mapping raw linguistic features such as self-references onto constructs such as leadership, we chose to simply present linguistic behaviors directly, letting people decide based on the task and the context of the conversation what those behaviors mean[1]. As we will discuss later, users sometimes had trouble understanding how the linguistic metrics might matter to the tasks we posed, leading them to conflicting interpretations or to ignoring the metrics. Based on people's reactions to early versions, over time we presented fewer metrics, focusing on metrics that people found more meaningful and that impacted teamwork more reliably.

5. How much normative guidance should be provided in a given context?
How much to interpret the data is related to another question concerning context: Should the system model contextual factors such as roles and tasks and provide normative goals for language use based on these contexts?

Contextual factors that feed into what normatively counts as good or poor behavior include factors at the individual level, such as members' gender, personality, and skills; at the group level, such as its size, organizational structure, roles, and development phase; and

[1] This may not be the right choice for people participating in groups where they are not speaking their native language; the value of computing and presenting linguistic feedback for people speaking a second language is an open question.

exogenous factors, such the task assigned to the group and its characteristics, the reward structure, and even cultural norms [41]. Based on goal-setting theory [29], introducing such factors and the norms they imply allows team members to easily understand how they are expected to act upon seeing the visualization in any given situation.

However, introducing normative goals may restrict the range of settings where the system can be applied. An alternative is to leave contextual factors out of the system, allowing for greater freedom for the group to appropriate the technology for its own purposes [35]. For instance, if a system visualizes the extent of agreement expression, team members can construct an interpretation of the feedback based on the task and the phase of the group development. They might see high agreement as beneficial when attempting to reach consensus, and as detrimental when critically discussing and negotiating solutions. Thus, we decided that GroupMeter should minimize assumptions about contextual factors and that its design should be mindful about what kinds of norms it suggests.

One important observation is that the graphical representation itself can imply norms. For instance, the green and red colors on the dials shown in Figure 2 imply certain norms to be achieved by team members—toward the green and away from the red. And even without the colors, a meter display might suggest that it should be filled up—that high values on the meters equate to "good" behavior. As we discuss later, despite our goal to keep the graphic representation open to interpretation of behavioral norms, some of our designs suggested normative interpretations of behaviors that were not always effective in the tasks we used.

Principles Meet Practice: Co-evolution Through Use

We now discuss how both the user interface and our answers evolved as we learned from deploying a series of versions of the system. Some changes were motivated by technical issues. Others were driven by results from user studies, controlled lab experiments in which visualizations were compared against each other or against chat use without any visualization. We present the discussion chronologically in an effort to clarify how our thinking evolved with the insights we gained from each study.

Version 1: Unobtrusive bar-charts

Figure 3 presents the first deployed version of GroupMeter. We abandoned the original design of a standalone feedback window, instead choosing to integrate the visualization with a custom-built chatroom. This allowed us to avoid the technical difficulties of interfacing with other systems and to create a more integrated, aesthetically unified experience. We chose to implement the client as a chat window embedded in a web browser. Together with a graphic designer, we created a new design for GroupMeter that included the chat window and a feedback display based on a series of stacked bar meters. In this design, every team member is associated with a color that appears behind their name, as a colored star in front of their chat entries, and in the feedback bars.

The feedback visualization in version 1 consists of horizontal bar charts, one for each linguistic metric. Each bar's length changes based on team members' behavior on the linguistic metric the bar represents. In this version, the interface presents feedback about team members' overall contributions measured by their word

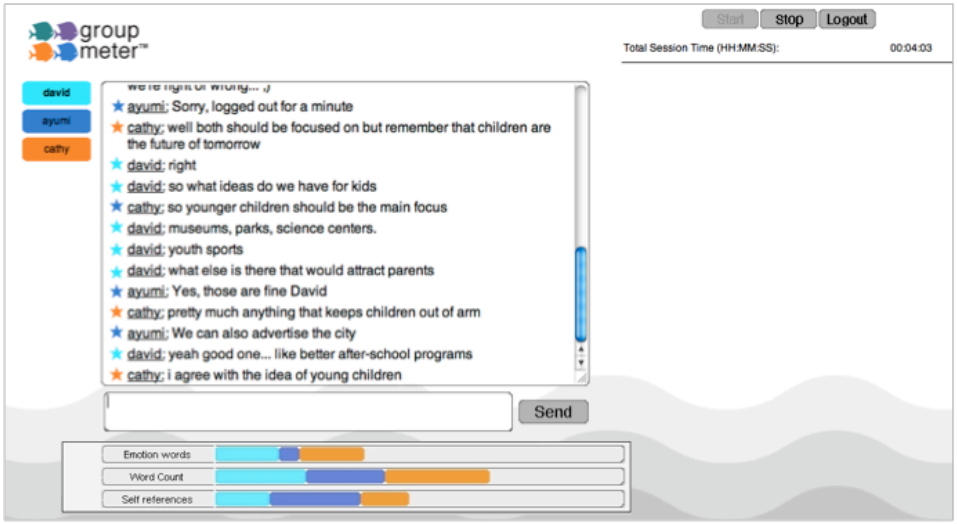

Figure 3. Version 1 of GroupMeter, using a custom-built web-based chat with feedback presented as stacked bar charts at the bottom.

count, their proportion of references to self ('I', 'me'), and their use of emotion-laden words. In a previous study we found people's scores on these linguistic metrics to be associated with peer-ratings on dimensions of teamwork such as participation, friendliness, and task-focus [25]. The linguistic metrics are computed and the bars are updated every minute, using a moving average based on the text entered by each member in the past five minutes. This supported the goal of presenting feedback dynamically, but without constant change or large fluctuations that might be distracting or hard to interpret.

The design of the bar charts and their location below the chat window was intended to make the visualization subtler than the original meter dials. We hoped the proximity would help people make smooth shifts between the chat window and the visualization, reducing distraction.

We chose to present each feedback dimension as an aggregate stacked bar to accomplish our goal of presenting each individual's behavior while reducing the possibility that people would process the visualization in a competitive way. An alternate display using a clustered bar chart could potentially cause people to meticulously compare the length of their bars to others' and attempt to increase their bars' length. Showing how individuals' behaviors accumulate to an aggregate bar also emphasizes the idea of being part of a group or team unit.

We deployed version 1 in a lab experiment to 88 people who worked together in 3- and 4-member teams to complete a decision making task. Half of the teams used GroupMeter with the bar-chart visualization. Because we were interested in seeing what interpretations people came up with for the feedback, we did not tell people how the metrics were computed and we did not present normative instructions or benchmarks. The other half of the teams used a version of GroupMeter with the chatroom alone and no visualization.

We found that the general idea of visualizing language use stimulated reflection on teamwork behaviors, but that compared to the control group, participants did not change their communication patterns in response to the feedback visualization. We interviewed our users, finding that the location and unobtrusive design of the bars might have led them to focus more on the task and not think much about how their word choice would affect the bar lengths.

We also found that the choice of linguistic metrics based on [25] was not always congruent with users' perceptions of how language use corresponds to collaborative behaviors. The emotion words metric was particularly hard to interpret, since it was not obvious which words fall under this category, leading many users to ignore it. This suggests that explaining how a linguistic metric is computed is essential for making a link between behavior and feedback. Finally, not providing norms led to conflicting interpretations of what counts as good or poor behavior. For instance, one participant understood the *self-references* bar such that high levels of it are undesirable:

> *"I was looking at the meter that was talking about how much you talk about yourself, and I was hoping it would be lower, because I didn't want to be that person that's just talking about themselves all the time."*

However, another participant said:

> *"I don't think it's bad to say 'I', cause sometimes it's better to convince people by saying, well this is how I feel, but I might be wrong. You know you've said 'I' twice there, but you're just trying to be nice by saying, you don't have to think what I think."*

In a second experiment with 25 participants, teams completed a task in which they brainstormed for solutions for a problem and then discussed and decided on the top three alternatives. While communicating in the chat they saw the bar chart visualization presenting information about proportion of agreement words and word count (see [26]). We chose the agreement words feature to see if there would be differences in agreement expression between the brainstorming and decision making sub-tasks. In this experiment we told participants how the language metrics were computed.

As in the former study, the bars made participants aware of their use of language compared to not seeing any visualization, and were considered unobtrusive: "*the bars were just there*" and "*could be ignored if wanted*." Unlike the former study, however, users changed their language use in response to the bars, expressing more agreement toward the end of the task when they had to reach consensus. However, we also found that across the two sub-tasks, users who saw the bars visualization expressed more agreements compared to those who did not. Despite the divergent thinking required by the brainstorming task and our explicit avoidance of normative guidelines, our decision to represent agreement as a bar might have implicitly guided this behavior: an embodied view of linguistic representations [24] assumes that people interpret "long" as better than "short", and agreeing more makes the bar longer.

Version 2: Playfulness and the fish metaphor

Our next design was aimed at both trying to reduce the implicit norms of bar charts and to see how a more visible and aesthetically pleasing visualization would affect people's use of GroupMeter. We therefore chose a more playful, abstract design using the metaphor of a school of fish. We found the school of fish inspiring because it symbolizes "togetherness", breaks from conventional forms of data presentation, and has a natural and serene connotation that "tells a story about the data" [37]. Other researchers previously used visualizations of fish to persuade people to engage in physical activity [28] and to represent workplace activity in a shared display [10].

In this visualization (Figure 4), colored fish represent individual team members, matching members' colors in the chat window. The fish start in a circular formation, all at the same size and equidistant from the center. We placed the fish visualization to the right of the chat window instead of below it so that, like the bars in version 1, it would be visible without scrolling the page.

The visualization is animated based on the moving average, dynamically changing the size of the fish to represent conversational activity (measured by word count) and their distance from the center to represent agreement with the group (measured by proportion of agreement words). We chose the circular form to better convey a sense of unity and community, as well as providing a natural mapping for "teamness" as measured by agreement: the more team members agree, the closer to each other their fish appear in the visualization. Again, despite our intention not to design for a specific context or to encourage certain behaviors, this decision implied the norm that more agreements are desirable—since the unity of fish closer together could easily be interpreted as preferable to fish scattered all over.

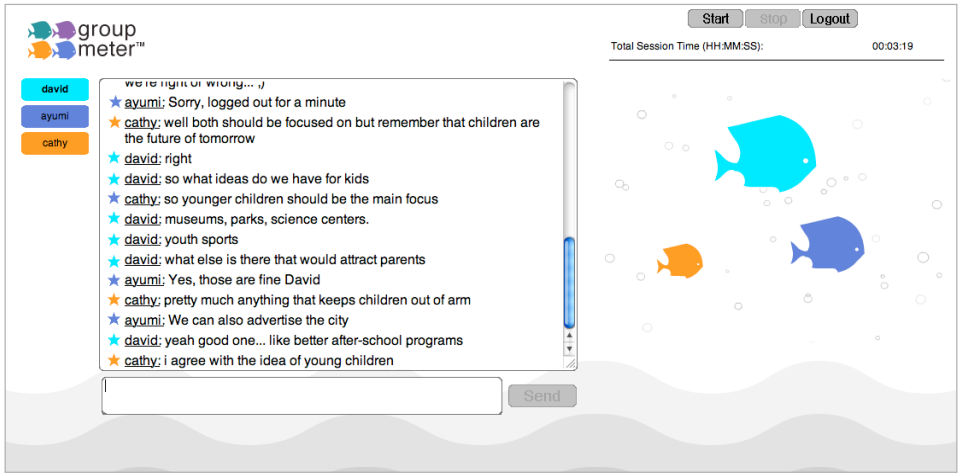

Figure 4. Version 2 of GroupMeter: feedback is visualized using fish in a circular formation that change their size and distance from the center.

This visualization was studied in comparison to the bar chart visualization and to a control group that saw no visualization [26]. Like the bars, the fish seemed to encourage users to reflect on their communication behaviors, but unlike the bars the fish were referred to as "cute" and "fun to watch" by participants. However, to notice changes in the fish visualization, users felt they had to constantly monitor it, distracting them from the conversation. Self-reports of distraction and chat about the fish during the task itself also implied that we might have gone too far toward engagement, disturbing people's ability to balance task and process.

Communication patterns also changed compared to not seeing any visualization: teams seeing the fish expressed more agreement with each other across the sub-tasks of brainstorming and decision making, at the cost of conducting less discussion. This suggests that our design decisions might have caused people to respond to the visualization in ways that potentially sabotage effective teamwork.

Version 3: Glanceability, history, and guidance

The goal of version 3 was to refine the feedback visualization to address the problems we found in version 2. One primary interface change was to improve the visualization's glanceability [31] by enriching the display with a history view [18]. Historical information can be important for understanding changes in social be-

haviors within a collaborative space [39]. This enables users to consider their behavior not in isolation, but in relation to trends they see in the past [5]. Similar to [3], if a user had been occupied by the primary team task and did not look at the display for a while, he or she could glance at the display and catch up with not only what is going on right now, but also with the trend of the feedback over the course of the conversation. We expected that this would help people balance attention between the primary conversation about the task and the peripheral feedback display.

We represented the history as trails of bubbles, as shown in Figure 5. To simplify the history view and the number of feedback dimensions represented by the visualization, here the visualization presents only one linguistic metric: fish move higher on the vertical axis based on the proportion of agreement words people used. With this design we also wanted to avoid imply-

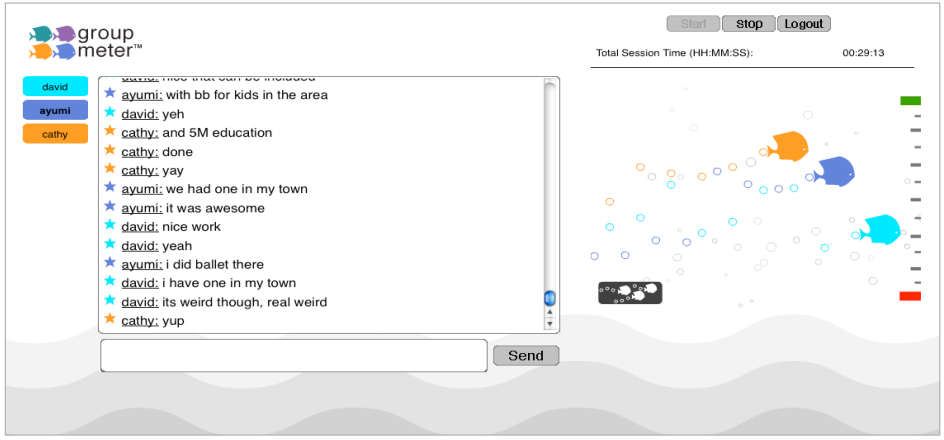

Figure 5. Version 3 of GroupMeter: fish move up and down in response to one feedback dimension and leave bubble trails behind them as they update their position. A ruler on the right with green and red marks serves as a normative cue.

ing that a team that agrees with one another, represented by closer fish, is more desirable. Every minute, the system calculates the linguistic metric, moves the fish to its new position on the vertical axis, and leaves a bubble behind it in its previous location. As a result, the bubble trails appear on the horizontal axis, giving an impression of the fish swimming from left to right.

Because the display can only show ten minutes of feedback, a button at the bottom left of the visualization opens a window with the full history view. This allowed us to experiment with another design consideration: the timing of the feedback. In our next deployment of the GroupMeter interface, we assigned two tasks with a pause between them (similar to [7][33]), during which participants opened the full history view and were encouraged to reflect on their teamwork and linguistic behaviors. This procedure provided both real-time display of feedback and explicit periods for reflection.

In this version we also made a major change in our thinking about providing normative goals. Initially, we attempted to avoid designing for a single interpretation of behavior, with the goal of allowing teams to develop their own meanings for the feedback. However, without guidance, people using version 1 developed mixed interpretations of what level of self-references was appropriate. Further, in version 2 users tended to agree more with implicit guidance (e.g., bring the fish closer together)—going against our views of appropriate behavior of divergent thinking in a brainstorming task.

To explore whether providing normative guidance would drive behavior changes in a certain direction, we added a ruler at the right hand side of the visualization. The ends of the ruler were green and red; the green

marker could either be placed at the top and the red at the bottom of the ruler, as in Figure 5, or their order could be reversed. We hoped this would encourage users to choose words that move their fish toward the green and away from the red.

This last version of GroupMeter was deployed in an experiment with 123 users. In a 2-by-2 design, 3-member teams were first given explicit instructions that either encouraged them to agree more or less with each other. They then completed two tasks with brainstorming and decision making segments similar to the tasks used in the former experiment; half the teams saw the visualization, while the other half did not. Teams receiving feedback saw the green and red ends of the ruler in congruence with the instructions they received—teams instructed to agree more saw the green at the top, and teams instructed to agree less saw the green at the bottom. Between the two tasks, teams receiving feedback reviewed the full history visualization and were prompted to reflect on it, while the other teams completed a filler survey.

Our results show that providing behavioral guidance stimulated change in language use in response to the visualization, although the changes were asymmetric. In particular, seeing the visualization induced users to be more agreeable when instructed to do so. However, people using the visualization did not agree less when instructed to do so. This may again be because of the notion of embodied representations of language [24]: people tend to perceive up as more—and more as better. Because the fish move up in response to more agreements, guiding people to agree less by moving their fish down works against natural perceptual interpretations. Perhaps reversing the entire display such that the fish move up with *less* agreement would stimulate a more critical and less agreeable conversation pattern, instead of only reversing the green-red ruler.

Further, our concerns about presenting individuals' behavior leading to "gaming the system" came true. Analyzing the team conversations, we found that people sometimes talked not to further the team goals but simply to make their fish move up in the display:

> *A: yes yes yes yes! hahaha sorry.*
> *A: let's make sure we all use the word "yes" at least once in every comment*
> *B: yes, let's do that*

This is a difficult problem worthy of further research. Researchers need to examine how to design feedback that explicitly links behavior with onscreen representations, but that resists gaming behaviors and guides people toward the adoption of behaviors that benefit the team.

In all of our experiments we found no differences between visualization conditions on performance measures such as number of brainstorming ideas and decision quality. We focused on metrics such as agreement expression that correspond to *social* team behaviors and that are relatively easy to compute. This might have been at the expense of choosing metrics of *task*-related behaviors that could directly support higher performance, for example, number and quality of brainstormed ideas. The right move might be to decouple learning social behaviors from task behaviors. By training both independently, people might come to learn how to effectively balance their use of behaviors that support the social environment and behaviors that support task performance [32].

Summary

To summarize our experience, and to call out aspects we hope will be useful for other designers, we present the questions we posed earlier and how our answers changed over the evolution of GroupMeter's design.

1. When, during the team interaction, should awareness information be presented? We started by visualizing feedback dynamically so team members could continually monitor it and connect changes in their behavior to changes in the display. Later, we added a history view that allowed for a fuller depiction of how behavior during the conversation unfolds, and introduced pauses between tasks in which teams had a chance to reflect more deeply on their teamwork behaviors. We sense that the combination of dynamic and punctuated feedback was especially useful in raising awareness of unconscious behaviors such as word choice.

2. How should attention to feedback and conversation be balanced? We used peripheral, glanceable displays to support quick transitions between the conversation and the visualization. The design of the visualizations affected how people managed their attention: bar charts were unobtrusive but easy to ignore, whereas fish were fun and playful but distracting. Somewhere in between, we hope, other designers will find visualizations that are both stimulating and effective.

3. Should group or individual level feedback be displayed? Our versions of GroupMeter visualized all individuals' behavior publicly, facilitating social comparison and leveraging the idea of social translucence. Users reported using the feedback to think about their language use and how it affected the group. However, they also sometimes competed against or played with their team members, changing their language use not for the goal of adopting beneficial communication skills but rather to manipulate the display.

4. What kinds of feedback should be computed and how should they be interpreted? We chose to compute metrics of language use that correlate with peer ratings of effective teamwork behaviors. Keeping these metrics at the raw data level instead of mapping them onto high-level concepts enabled users to see a link between their language use and the visualization and to make their own interpretations. Over time we had to adjust our choice of metrics and provide explicit explanations of the computation behind the visualization to help create the behavior-feedback link.

5. How much normative guidance should be provided in a given context? Our initial answer was to give only enough guidance for people to understand the visualization, allowing for flexible appropriation in many contexts. However, we found that this could lead to competing interpretations of the appropriateness of behaviors and that our graphical designs sometimes encouraged behaviors ineffective in the context in which they were applied (e.g., seeking consensus in a brainstorming task). We later explored how to provide guidance that would drive behavior in a certain direction using visual cues and explicit instructions. Our limited success cautions designers to be attentive to how design details, especially those not thought of, might influence behavior.

One topic for further discussion is our observations of gaming behaviors. One of our concerns, justified by our participants' behavior, was that team members would play with GroupMeter's feedback. For example, partici-

pants typed agreement words to make their fish move up or down regardless of their relevance to the conversation. We have several responses. First, play is natural and often a sign of engagement, and it might actually be positive to encourage exploration and discussion of the feedback as a way to encourage people to reflect on their behaviors. Second, gaming the system was another indicator that feedback can distract from the primary task. As such, this kind of system may be best applied in a training context, rather than in the field. This would allow participants to engage with the feedback without concerns of interfering with the team's central purpose.

Conclusion

In this paper, we make three main contributions. First, we lay out important questions in the design space for systems that visualize behavior to help people collaborate better: managing the task-social process balance through effective timing and positioning of the feedback, deciding whether to display information publicly or privately, choosing which data to display and how much to interpret it, and weighing how much the system should try to account for the context of use and provide normative guidelines. Although every system will face unique challenges in its particular context, the issues we call out are likely to apply across a broad range of systems for supporting teams.

Second, we call attention to studies and theories of cognitive and social behavior that seem relevant for designing collaboration-enhancing systems. We discuss how we used these theories, prior studies, and our intuition to select reasonable candidate answers to the questions we faced in building GroupMeter. Our design solutions are not the only possibilities. For instance, some groups or cultures might be sensitive to individual criticism and thus prefer group-level feedback to presenting information about each person's behavior. Also, designers of tools that are specific to a given context may want to leverage that context. Research on cross-cultural differences in responses to linguistic feedback visualizations in other contexts such as creative design teams is currently underway [6]. In general, however, presenting theoretical perspectives on potential solutions opens up the design space for educated exploration and consideration.

Third, we explore how our initial theory-driven answers fared in practice through a series of deployments and experiments using versions of the GroupMeter system. GroupMeter was designed and deployed as a research platform, but it also provides practical lessons for designers. By calling out the design questions, presenting theoretical perspectives on potential solutions and tradeoffs, and reviewing the lessons we learned, we hope other designers can make use of our analysis and experience to make good choices and systems.

References

[1] Bales, R.F. (1950). *Interaction process analysis: A method for the study of small groups*. MA: Cambridge.

[2] Bergstrom, T., & Karahalios, K. (2007). Conversation clock: Visualizing audio patterns in co-located groups. *HICSS'07.* 78-86.

[3] Bezerianos, A., Dragicevic, P., & Balakrishnan, R. (2006). Mnemonic rendering: An image-based approach for exposing hidden changes in dynamic displays. *Proc. UIST '06,* Montreux, Switzerland. 159-168.

[4] Cegala, D.J. (1989). A study of selected linguistic components of involvement in interaction. *Western Journal of Speech Communication, 53*, 311-326.

[5] Consolvo, S., McDonald, D.W., & Landay, J.A. (2009). Theory-driven design strategies for technologies that support behavior change in everyday life. *Proc. CHI '09,* Boston, MA. 405-414.

[6] Diamant, E.I., Echenique, A., Lim, B.Y., Leshed, G., & Fussell, S.R. (2009). Supporting intercultural collaboration with dynamic feedback systems: preliminary evidence from a creative design task. *CHI '09 Extended Abstracts*, Boston, MA.

[7] DiMicco, J.M., Hollenbach, K.J., Pandolfo, A., & Bender, W. (2007). The impact of increased awareness while face-to-face. *HCI, 22*(1), 47-96.

[8] Dourish, P., & Bly, S. (1992). Portholes: Supporting awareness in a distributed work group. *Proc. CHI '92,* Monterey, CA. 541-547.

[9] Erickson, T., & Kellogg, W.A. (2000). Social translucence: An approach to designing systems that support social processes. *ACM ToCHI, 7*(1), 59-83.

[10] Farrell, S. (2001). Social and informational proxies in a fishtank. *CHI '01 Extended Abstracts.* 365-366.

[11] Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117-140.

[12] Fogg, B.J. (2003). *Persuasive technology: Using computers to change what we think and do*. San Francisco, CA: Morgan Kaufmann.

[13] Freeman, M., & McKenzie, J. (2002). SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects. *British Journal of Educational Technology, 33*(5), 551-569.

[14] Gersick, C.J.G. (1988). Time and transition in work teams: toward a new model of group development. *Academy of Management Journal*, 31(1), 9-41.

[15] Gaver, W.W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. *Proc. CHI '03,* Ft. Lauderdale, FL. 233-240.

[16] Greenberg, S. (1996). Peepholes: Low cost awareness of one's community. *Proc. CHI '96,* Vancouver, BC, Canada. 206-207.

[17] Hackman, J.R., & Wageman, R. (2005). A theory of team coaching. *Academy of Management Review, 30*(2), 269-287.

[18] Hill, W.C., & Hollan, J.D. (1994). History-enriched digital objects: Prototypes and policy issues. *The Information Society, 10*(2), 139-145.

[19] Huang, E.M., & Mynatt, E.D. (2003). Semi-public displays for small, co-located groups. *Proc. CHI '03*, Ft. Lauderdale, FL, 49-56.

[20] Johnson, D.W., & Johnson, R.T. (1975). *Learning together and alone: cooperation, competition, and individualization*. Englewood Cliffs, NJ: Prentice-Hall.

[21] Kim, T., Chang, A., Holland, L., Phillips, J.M., & Pentland, A. (2008). Meeting mediator: Enhancing group collaboration using sociometric feedback. *Proc. CSCW '08,* San Diego, CA. 457-466.

[22] Klein, K.J., Ziegert, J. C., Knight, A.P., & Xiao, Y. (2006). Dynamic delegation: Shared, hierarchical, and deindividualized leadership in extreme action teams. *Administrative Science Quarterly, 51*(4), 590-621.

[23] Kluger, A.N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

[24] Lakoff, G., & Johnson, M. (1981). *Metaphors we live by*. Chicago: University of Chicago Press.

[25] Leshed, G., Hancock, J.T., Cosley, D., McLeod, P.L., & Gay, G. (2007). Feedback for guiding reflection on teamwork practices. *Proc. GROUP '07*, Sanibel Island, FL. 217-220.

[26] Leshed, G., Perez, D., Hancock, J.T., Cosley, D., Birnholtz, J., Lee, S., McLeod, P.L., & Gay, G. (2009). Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. *Proc. CHI '09*, Boston, MA. 537-546.

[27] Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

[28] Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G., & Strub, H.B. (2006). Fish'n'Steps: Encouraging physical activity with an interactive computer game. *Proc. UbiComp '06,* Orange County, CA. 261-278.

[29] Locke, E.A., Shaw, K.N., Saari, L.M., & Latham, G.P. (1981). Goal setting and task performance: 1969-1980. *Psychological Bulletin, 90*(1), 125-152.

[30] Matthews, T., Dey, A.K., Mankoff, J., Carter, S., & Rattenbury, T. (2004). A toolkit for managing user attention in peripheral displays. *Proc. UIST '04*, Santa Fe, NM. 247-256.

[31] Matthews, T. (2007). *Designing and evaluating glanceable peripheral visualizations.* Unpublished PhD, EECS Department, University of California, Berkeley,

[32] McLeod, P.L., & Kettner-Polley, R.B. (2004). Contributions of psychodynamic theories to understanding small groups. *Small Group Research, 35*(3), 333-361.

[33] McLeod, P.L., Liker, J.K., & Lobel, S.A. (1992). Process feedback in task groups: An application of goal setting. *Journal of Applied Behavioral Science, 28*(1), 15-41.

[34] Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). *Linguistic inquiry and word count: LIWC2001*. Mahwah, NJ: Erlbaum.

[35] Poole, M.S., & DeSanctis, G. (1989). Use of group decision support systems as an appropriation process. *HICSS'89,* 149-157.

[36] Rofé, Y. (1984). Stress and affiliation: a utility theory. *Psychological Review, 91*(2), 235-250.

[37] Tufte, E.R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

[38] Turner, G.A., & Schober, M.F. (2007). Feedback on collaborative skills in remote studio design. *HICSS'07*, 44-53.

[39] Viégas, F.B., Golder, S.D., & Donath, J. (2006). Visualizing email content: Portraying relationships from conversational histories. *Proc. CHI '06,* Montréal, Québec, Canada. 979-988.

[40] Walther, J.B., & Parks, M. (2002). Cues filtered out, cues filtered in. *Handbook of interpersonal communication*, 529-563. Thousand Oaks, CA: Sage Publications.

[41] Wang, H.C., Fussell, S.R., & Setlock, L.D. (2009). Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. *Proc. CHI '09*, Boston, MA. 669-678.

[42] Wickens, C.D., & Hollands, J.G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.