NSF # 0968536 Final Report

SoCS: The Fourth Party: Improving Computer-Mediated Deliberation through Cognitive, Social and Emotional Support

December 2014

Tom Murray (PI)
Co-PIs: Beverly Woolf, Ethan Katsh,
Leon Osterweil, Lori Clarke, Leah Wing

Research Assistants:
Xiaoxi Xu, Lynn Stephens, Natasha Shrikant, Stefanie Shipe,
Scott Howard, Stephanie Kent, Roger Ledgister

Project website: www.socialdeliberativeskills.com

Results Overview

Executive Summary of Outcomes: People are increasingly engaged in online dialogue, deliberation, and collaboration. The Internet provides opportunities for increased exchange of ideas, particularly with others who we may not have a chance to engage face-to-face. There is under-explored opportunity for online systems and tools to directly support participants in having higher quality and more skillful engagements. The overall goal of the project is to support higher quality online deliberation, especially by supporting number of "***social deliberative skills***" such as *perspective taking, empathy, self-reflection, tolerance for uncertainty, listening and question-asking skills, and meta-dialogue*—in online contexts.

We attempt to do this through software tools and features, some of which directly support participants, and others which support a facilitator or mediator as they engage

with participants. **To support participants** we implemented unobtrusive **scaffolding features**, and **to support facilitators** we implemented a "**dashboard" visualization tool**. We also investigated using state-of-the-art **text analysis and machine learning** to measure important properties of deliberative dialogue. In addition to the development and formative evaluation of these tools, we conducted experimental trials that showed, for a population of college students engaged in online discussion of controversial issues, that our "**reflective tools**" did indeed lead to deeper, more skilled, and reflective dialogue. Our work with text analysis had given early indications that automated methods for assessing the quality of online dialogue can be used productively to support higher quality communication (for instance through visualizing this information in the Facilitator's Dashboard).

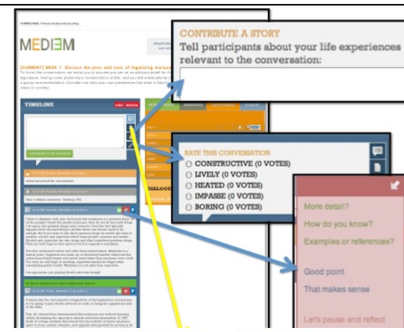
Wider potential impacts: Our unobtrusive scaffolding methods and facilitator dashboard concepts will transfer to use in collaborative work, civic engagement, and online dispute resolution. Discussion forums and commenting features are widely used in educational contexts and social media, but little exists to support higher quality deliberation in these environments. Our methods should also be applicable to "flipped" classrooms and MOOC (Massive Online Open Course) environments, which are in need of better tools for interaction, dialogue, and management.

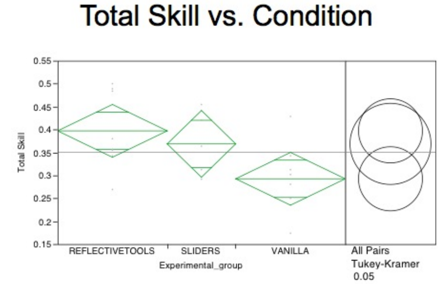
Work and Results Summary. "The Fourth Party: Improving Computer-Mediated Deliberation through Cognitive, Social and Emotional Support" was funded for three years and was extending into a fourth year (no-cost extension). The project was interdisciplinary and exploratory, meant to integrate research from several areas and map out future directions in an emerging sub-field. The primary goal of the project was to study methods to support skills that help individuals successfully negotiate deliberative dialogues in which they are challenged by differences in perspectives, goals, assumptions, etc.—in online contexts specifically. The project produced 17 peer-reviewed research papers and an additional 10 articles, workshop presentations, and radio show interviews. Significant progress was made over the grant period in several areas that can be summarized as follows (thumbnails are show of figures that appear later):

1. Experimental trials of scaffolding features

Experimental trials in college classrooms were used to study the effectiveness of passive scaffolding features hypothesized to support social deliberative behavior.

Results summary: A combination of the following Reflective Tools were found to have a significant effect (large effect size) on total deliberative skillfulness: meta-dialogue support; personal-stake-and-story support; and productive-reply-reminders. Intersubjective speech acts (such as questioning, reflecting back, and perspective taking) were particularly affected by these tools.

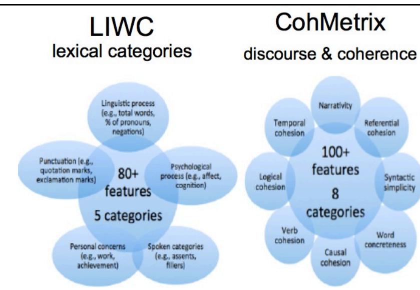


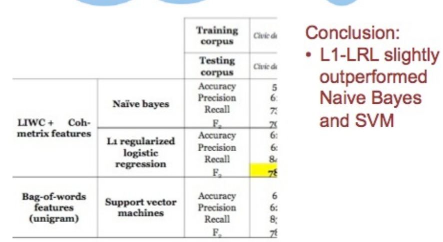


2. Text Analysis for dialogue quality

Research on the use of state-of-the-art text analysis methods to identify indicators of deliberative dialogue quality, including participant skillfulness.

Results summary: the combination of L1 Regularized Logistic Regression with psycholinguistic features (LIWC) showed the best performance on predicting overall deliberative skill and intersubjective speech.

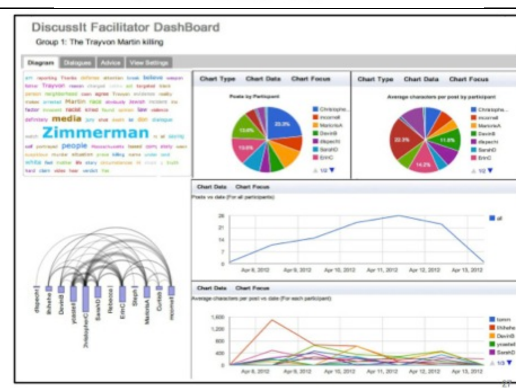




3. Facilitator's Dashboard

Prototyping of a Facilitator's Dashboard for assessing and visualizing dialogue quality in online deliberation.

Results summary: The Dashboard was designed over several iterations with input from teachers and professional mediators and online facilitators. Visualization tools include participation charts and trends, demographic differences, word cloud, and

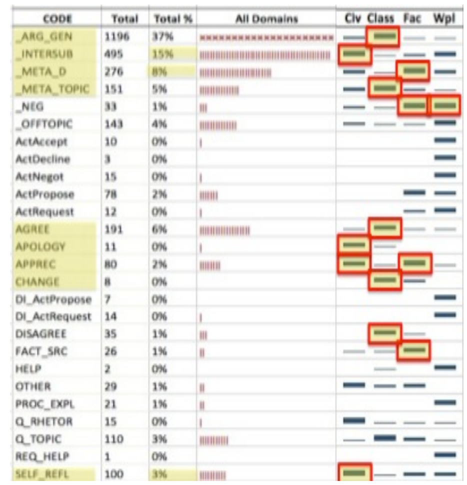


social network diagram. "Intelligent" indicators using text analysis for several analysis types were implemented. Evaluation for strictly formative. Students were interviewed about the possibility of peer-based visualization tools as well.

4. Comparative analysis of diverse dialogue domains

We began a comparative analysis of a number of dialogue domains assessing dialogue properties and predictive potential of text analysis.

Results summary: Initial work was done to compare dialogue properties across diverse domains and populations, with pre-existing online discussion data, for these domains: college classes, deliberation within a professional association, a teenage social networking forum, online dispute resolution services, and e-commerce dispute resolution. Notable patterns were observed regarding these speech act types: intersubjective speech, meta-dialogue, agreement, appreciation, referencing sources, and self-reflection.



CODE	Total	Total %	All Domains	Civ	Class	Fac	Wpl
_ARG_GEN	1196	37%					
_INTERSUB	495	15%					
_META_D	276	8%					
_META_TOPIC	151	5%					
_NEG	33	1%					
_OFFTOPIC	143	4%					
ActAccept	10	0%					
ActDecline	3	0%					
ActNegot	15	0%					
ActPropose	78	2%					
ActRequest	12	0%					
AGREE	191	6%					
APOLOGY	11	0%					
APPREC	80	2%					
CHANGE	8	0%					
DI_ActPropose	7	0%					
DI_ActRequest	14	0%					
DISAGREE	35	1%					
FACT_SRC	26	1%					
HELP	2	0%					
OTHER	29	1%					
PROC_EXPL	21	1%					
Q_RHETOR	15	0%					
Q_TOPIC	110	3%					
REQ_HELP	1	0%					
SELF_REFL	100	3%					

5. Theoretical framework

Developing a theoretical framework for the analysis and support of social deliberative skills.

Results summary: Related literature from cognitive science, communications theory, deliberative theory, conflict and peace studies, and educational pedagogy were synthesized and interrelated. A novel conceptual framework was developed that sees deliberative skills as the application of cognitively oriented skills to the emotional/social/intersubjective realms.

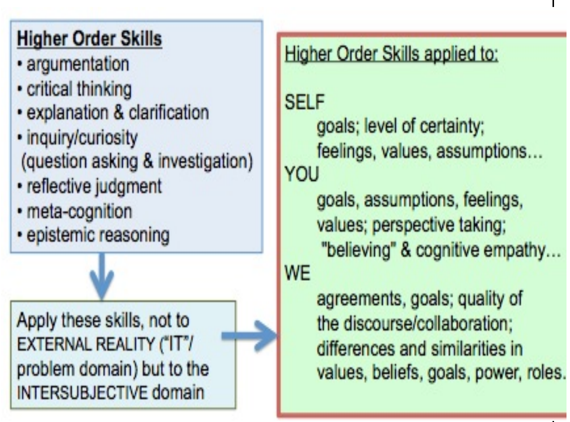


Table of Contents

Project History

This work constitutes a new sub-field of research that combined interdisciplinary elements from collaborative online systems, dialogue and deliberation theory, machine learning, and a novel approach to online dialogue quality that focuses on skill-support (skills that we call social deliberative skills). When we started this grant project some of the domains of study and the research methods were new to the research team. It took a full year to "get our feet under us" and build the necessary tools (see Codoole application below), relationships, expertise, and data sets to ground the research work. In year 2 we began our deep dive into research and development work, and in years 3 and 4 we were able to produce a satisfying number of publications and report our work in about a dozen conference presentations (see Publications section). As exploratory research, this project has just cracked the surface in terms of research in the field of supporting higher quality social deliberation in online contexts, and we have submitted grant proposals to continue the work.

The project began with a focus on online dispute resolution in areas such as e-commerce, workplace disputes, and divorce settlements. We had industry connections and collaborators in these areas (including a company working on eBay and PayPal online dispute resolution, a lead person in the U.S. National Mediation Board, and a company offering online dispute resolution services in areas such as divorce and workplace disputes). Through the generosity of our contacts we were able to obtain (redacted) prior data for analysis. However, it became clear in Year 1 that it would be very difficult to experiment with interventions, i.e. new software systems or features, with the participants in these populations. Participants were engaged in solving important real disputes using known methods, expecting strict confidentiality, and there was insufficient incentive for the managing professionals in these organizations to experiment with novel approaches and new software (and go through the laborious steps to obtain consent). These connections provided important data for our comparative analysis of domains, and allowed us to get up to speed with text analysis methods. Also our contacts provided important consultation about real-world contingencies and needs throughout the project.

However, for testing novel interventions and tools we turned to college classrooms. These had the benefit of readily available populations who could be engaged in online dialogue about controversial topics, were tolerant to imperfections in the methods and prototype software, and were available for follow-up focus group interviews. The downside of this population was that the conversations were more contrived, i.e. less was at stake than in real-world disputes. Nevertheless, participants engaged fairly energetically.

Initially our goal was to incorporate two state-of-the-art computational methods: Prcess Modeling and Text Analysis. In moving from our industry partners, many of whom used specific conflict resolution processes, to open classroom (and civic engagement) contexts, we left behind the Process Modeling theme and focused on Text Analysis (and machine learning). Process Modeling research was continued under the umbrella of other grants lead by Osterweil and Clarke.

At the completion of the project we have two journal papers in process. As is expected from exploratory studies, we have many research and data analysis questions and investigations that remain on the "to do" list, and many aspects of the collected data that remain to be analyzed.

Introduction and Motivations

Deliberation in society. Consider the variety of types of challenging deliberative dialogues that people engage in. Students in a science course and citizens discussing current news or social trends might be talking about birth control, internet privacy, gene therapy, climate change, or gun control. They are not tasked with arriving at any collective conclusion, but through interaction become informed about the important issues of the day. These dialogues require considering divergent perspectives, complex relationships of facts, and the reliability of those facts. At work, or in neighborhood civic meetings, participants having diverse goals or values *are* tasked with finding consensus on complex issues so that projects can move forward in ways that meet as many stakeholder needs as possible. The same is usually true for formal and informal dispute resolution and for facilitated group problem solving processes. Finally on the home and relationship front, as in all of the above scenarios, we have difficult conversations with others and are challenged to make our points or get our needs met while yet maintaining (or even deepening) the relationships involved. In so-called "inter-group" dialogues, achieving mutual understanding and mutual regard are top priorities, often above achieving consensus or action plans.

Much has been said and written about productive or effective dialogue methods and skills in a variety of academic disciplines. However, where these concerns overlap with R&D in the Learning Sciences, Educational Technology, and Computer Supported Collaborative Work (taken together) the focus is largely lopsided in a particular dimension. R&D in these areas focuses on helping participants build cognitive and analytical skills so that they can achieve optimal solutions, but does not often focus on building skills for mutual understanding and mutual regard *for their own sake*. Emotional/social skills are certainly considered, including collaboration, group creativity, and peer teaching/learning skills, but predominantly they are in service of learning content or solving problems. Mutual understanding, including "uptake" and grounding, is also studied, but again as an element of problem solving, knowledge building, or collaborative learning. Yet if one imagines oneself in each of the deliberative contexts mentioned above, the desired outcomes are not merely impersonal and logical, as if participants were expendable or interchangeable means to an end. It is not only the analytic mind that must open and reach for new and higher quality knowledge and solutions, but we are also challenged to reach across intersubjective space and connect with other participants with care and/or respect. Care and respect are ends in themselves, even though creating effectual relationships is also an important means toward other ends.

In this research we join a small and growing number of other scholars who focus on the intersubjective skills that participants must bring to bear to build mutual understanding and mutual regard in deliberative dialogues about complex or controversial topics (we call these skills social deliberative skills). We believe (as is supported later) that successful dialogue and deliberation in all of the domains mentioned above share a basic set of social deliberative skills that can be supported, and in particular supported in the context of online interactions.

Deliberation Online. People are increasingly engaged in online dialogue, deliberation, and collaboration. The Internet provides opportunities for increased exchange of ideas, particularly with others who we may not have a chance to engage face-to-face (F2F). There is under-explored opportunity for online systems and tools to directly support participants in having higher quality and more skillful engagements. We have prototyped both deliberation tools that support *participants* directly and Dashboard tools ("4th Party" tools) that support *third parties* (mediators, facilitators, teachers, moderators, etc.) in supporting higher quality deliberation among participants. A key emerging element of this work is the application of state-of-the-art computational techniques to analyze the textual content and social network relationships of an exchange. Recent advances in computational psycholinguistics allow for a more systematic and deeper analysis of dialogues, which is necessary to uncover subtle cues that might be diagnostic of critical deliberation characteristics.

The overall goal of the project is to support higher quality online deliberation, especially by supporting number of "social deliberative skills" such as perspective taking, self-reflection, social inquiry, and meta-dialog. We attempt to do this through software tools and features, some of which directly support participants, and others which support a facilitator or mediator as they engage with participants. The software features and interventions range from simple prompts and reminders to suggestions based on automated content analysis of the dialog text. Our goal is to evaluate these software tools in a diverse set of line deliberation contexts, as mentioned below.

Information for such analysis can be considered for a variety of uses, such: (1) feeding into visualization and "awareness" tools that help participants reflect on process and outcomes; (2) informing third parties of dialogue trends to enable better facilitation; and (3) adaptive user interfaces that intelligently turn on or off features meant to support quality dialogue. The UMass team has already tested preliminary versions of #1 and #2 above, and plans to explore #3 in the future. This project focuses on deepening the theory and methodology behind the analysis of deliberative dialogue (taken from online interactions). Unlike some social computing projects that focus on big data taken from 1000s of participants, our focus is on supporting smaller group engagements—anywhere from two-parties engaged in a challenging online dialogue such as dispute mediation, to dozens or hundreds of individuals engaged in discussion or decision making process, such as an online civic engagement or community decision-making process.

Deliberative Dialogue Text Analysis

Dialogue Data Sources

One of our goals was to include a variety of types of deliberative dialogue in our analysis so that we could build domain-generic models, and learn something about how deliberative domains differ in terms of SD-skill behavior. Our analysis to date has focused on (1) text from pre-existing online dialogues; and (2) text from experimental trials of online dialogues among college students. The pre-existing dialogues come from sources including: online civic engagement, a group of faculty using an email listserve to deliberate about a conference venue (a discussion that became contentious); online dispute resolution with mediators and two parties; and e-commerce dispute resolution involving buyers, sellers, and mediators. (See the Appendix "Samples from online dialogs from our data sets".) Table 1 summarizes the data that was analyzed over the course of the project. Only a subset of these data sets underwent full analysis, as described later.

- College classroom dialogs — live participant data
  - UMass graduate classes in: (E)ducation, (L)egal Studies, and (C)ommunication Studies
  - *Experimental & Control groups in each study*
  - Spring2011-L, Fall2011-E, Fall2011-L, Spring2012-E, Fall2012L, fall2012-E, Spring-2013 -L
- Online Dispute Resolution — pre-existing data
  - E-Commerce (~3000 e-bay auto disputes)
  - Juripax – divorce settlement & workplace dispute (2 cases)
- Civic Deliberation — pre-existing data
  - E-Democracy.com (Minnesota neighborhood) (3 discussions)
  - Mass Dept of Dispute Resolution —Forest Futures process delib.
- Misc — pre-existing data
  - GovTeen.com (Philosophy & Ethics forums)
  - Bi-community faculty deliberation on conference venue (10 participants)

Table 1: Text Analysis Domains

A subset of these data sets was hand-coded for deliberative skill behavior and other indicators of dialogue quality. Table 2 summarizes the data sets that were coded using the coding manual that we developed through a synthesis of schemes found in the literature (see below). The table shows the percent agreement and Cohen's Kappa value for each data set. The Workplace Grievance data set is considered an atypical outlier here, and is not included in the Total/Average row in the Table.[1] Posts (from discussion forums, emails, etc., depending on the domain) are first divided into segments manually according to speech act types (typically of 1-4 sentences). Coding is done per segment, usually by two coders. The table shows the total number of posts and segments, and the number of participants in each discussion.

Finally, as seen in Table 2, there is much more data from the classroom discussion domain, so we must keep in mind that any analysis results from combining all domains will be highly influenced by this domain.

Domain Statistics and InterRater Agreement. Two independent trained human judges annotated the data sets based on the coding scheme described above. As shown in Table 2 our interrater reliability shows a Cohen's Kappa of 71% and a 76% agreement. The reliability scores for the first three domains are considered "good" (Altman, 1991; Carletta,1996), and we consider them quite acceptable given the complexity of our scoring system). The low inter-rater score for the Workplace domain is another reason we excluded it from some of the analysis reported.

Domains	No. Segments	No. Posts	No. Particip.	Cohen's Kappa	% Agreement
ODR Workplace Grievance	637	56	3	47%	54%
E-democracy Forum	396	151	31	78%	81%
Faculty Listsrv Negotiation	438	72	16	70%	74%
UMass Grad F'11 Forum-E	324	238	13	71%	78%
UMass Grad S'11 Forum-L	748	231	41	65%	70%
UMass Grad F'11 Forum-L	781	294	36	70%	76%
UMass Grad S'12 Forum-E	954	369	28	72%	77%
TOTAL/AVE (w/o WG)	**3641**	**1355**	**165**	**71%**	**76%**

Table 2: Hand Coded domains

Below is a description of the four corpora of online conversations that our team has coded:

Civic Deliberation: postings from a neighborhood civic engagement online discussion forum at e-democracy.org. 31 participants were discussing racial issues and tensions about their multi-racial community.

Faculty listserve: email exchanges from 16 participants on a faculty listserv with geographically dispersed participants. Two research communities were engaged in a discussion about how to organize a conference addressing overlapping interests. The discussion became contentious. Though the discussion was cordial and professional, some participants expressed increasing levels of disappointment at what they thought was a non-cooperative attitude by some in group leadership positions. The level of intelligence and sophistication of the participants is assumed to be be quite high because they presumably have doctoral degrees.

Classroom discussions: postings from online discussions on controversial issues from four college classrooms, assigned as homework. All students in each class discussed the same questions, and students were assigned to small groups (4-6), and each group discussed at least two topics (3 for some classes) in online homework assignments lasting 1-3 weeks. Topics were

[1] All the other coded sets involved group discussion, whereas this one involved two parties and a mediator; and all the other data sets had no or minimal mediation or facilitation. Also this domain involved much more action-negotiation speech acts that other data sets, for which there was less interrater agreement, and our coding category definitions were not as well developed.

determined in brainstorming sessions with the students. The topics include “should the legal drinking age be lowered in Massachusetts?,” The Trayvon Martin shootings in Florida, and “pros and cons of using FaceBook or other social networking software as part of high school curriculum”. Discussions were cordial and sometimes engaged, but not particularly so, and post-activity surveys showed that participants level of interest was on average "neutral." (These were part of experimental trials with online discussion from features meant to support SD-skills—see below.)

Workplace dispute mediation: exchanges from an online dispute mediation session about a workplace dispute, involving one employee, one supervisor, and one mediator. This data was not used for some of our analysis because it has unique characteristics compared to the others. The fact that it was a three-party exchange rather than a group dialogue is one difference, as is the fact that it was the only facilitated dialogue (some codes are specific to facilitators). Also, it was not really a dialogue as the parties communicated mostly with the mediator, and occasionally passed messages to each other through the mediator. Unlike the other domains, mutually acceptable outcome was essential to the parties (we have additional data from an online e-commerce service that we will later compare with this data set).

Participation Statistics. Table 3 shows additional details on the four domains used for the cross-domain comparisons, including the average words per post and posts per participant. The college discussions constituted a large proportion of the total data points (and participants).

Domain	Posts	Segments	Participants	SD-Skill seg codes	% Codes SD-skills	Words / Post	Posts / Partic	Seg. / post
Civic deliberation	51	396	31	225	57%	352	1.6	7.8
Faculty negotiation	72	438	16	231	53%	195	4.5	6.1
College discussions	768	1783	9	565	32%	88	8.5	2.3
Workplace	56	637	3	251	39%	200	18.7	11.4
All domains	947	3254	140	1272	39%	117	6.8	3.4

Table 3: Participant Descriptive Statistics

In online dialogues one typically observes something like the "80/20 principle" in which a minority of the participants produce the majority of the posts. The **posts per participant** observed say something about the group dynamics of the domains. The college domain had twice as many posts per participant as the faculty domain, which in turn had about twice as many as the civic domain. This corresponds with expectations. The college students were required to participate. The faculty discussion was a semi-closed group of participants who probably knew most of the participants in their cohort, participants had moderate incentive to participate. The civic dialogue was in a more open-enrollment context, with perhaps thousands of people living in a city neighborhood in the pool of appropriate participants. The commonly observed internet dynamics of a few participants doing the majority of the participation was even more exacerbated here, as there was little consequence for non-participation or low participation.[2] The

[2] At a further extreme would be an entirely open internet forum, such as the Comments section at the bottom of online news articles, where one can often see inappropriate posting and disrespectful dialogue. For our current work we are not interested in these types of domains because there is little "deliberation" in the open-web forums, and less force of accountability for participants who have no past or planned future relationship with each other. In such contexts the typical aim of intervention

workplace domain had the highest posts per participant, but it involved a very different process which was not a real dialogue.

The **words per post** is in the reverse order as the posts per participant, with the college domain having about half as many as the faculty domain, which again had about half as many as the civic domain. Students were posting more but shorter posts, in part because their interchange was daily and they were asked to respond to several others. The group of professionals might be less likely to "carry on" with long posts, while some community members in the civic domain can be seen to expound at length on their opinions and philosophy, and some also provided concrete 'stories' as evidence (also, the topic of discussion was relatively complex and multi-faceted). The segments per post, unsurprisingly, follows the words per post, with longer posts containing more speech act-type segments.

Table 4 shows the **gender distributions** for the group discussions. Gender was balanced in the Classroom domain, moderately similar in the Civic domain, and highly skewed toward male participants in the Faculty domain. (We have not yet done a gender-based analysis of dialogue across domains.)

Domains	Female participant counts	Male participant counts	Segment counts (female)	Segment counts (male)
Civic and ethical deliberation	19 (59%)	13 (41%)	250 (63%)	146 (27%)
Professional community negotiation	3 (19%)	13 (81%)	113 (26%)	325 (74%)
College classrooms	46(51%)	44(49%)	1022 (57%)	761 (43%)

Table 4: Gender Distribution

Coding Scheme

SD-skills overlaps with but are distinct from other cognitive constructs that have been studied in depth, including collaboration skills, metacognition, reflective reasoning, social intelligence, argumentation skills, and critical thinking (Lin & Sullivan, 2008; King & Kitchener, 1994; Kuhn, 1999, 2000; Block-Lerner et al., 2007; Graesser et al., 2005, 2008; Dawson & Stein, in press). We differentiate our research from others that focus on *argumentation*, which aims to help learners generate logical, well-formed, well-supported explanations and justifications (Andriessen et al., 2003; Baker et al 2007; Scheuer et al., 2010), usually framed in objective rather than intersubjective terms. That is, they are about finding the right answer or the most efficient and effective solution to a technical or scientific question—but don't address, as we do, the skills need in those moments during deliberation or collaboration containing opportunities for mutual understanding and mutual recognition.

We are using both manual coding and automated text analysis. We have developed and refined a (approximately) 30-category hierarchical coding scheme for human raters to code segments of the text. This scheme was derived from several prominent schemes found in the literature (Black et al., 2011; Klein, 2010; Stromer-Galley, 2007; and Stolcke et al. 2000) and adds codes specific to social deliberative skills (Murray et al. 2012). The scheme is unique in its inclusion of social deliberative skills with indicators of dialogue quality. Our achievement of adequate interrater reliability was noted above. The coding scheme includes 10 codes for social deliberation behaviors (including perspective taking, asking clarifying questions, mediation actions, and meaning generation and repair actions), 7 codes for additional deliberation quality indicators (including weighing alternatives, citing sources, changing ones mind, and apologizing), and other categories for action negotiation and argumentation.

---

is more to minimize flaming and off topic posts. The contexts we are interested in involve quasi-boundaried groups with some degree of mutual accountability, so we set the bar higher in our aim of supporting SD-skills.

A description of the text coding scheme used for this research, and its development, appears in The Appendix. We include a figure of the categories in Figure 1 for reference.

Soc. DELIBERATION Skill Evidence	MISC CODES	ACTION NEGOTIATION	ARGUMENT CODES
SELF_REFLection	Q_TOPIC	(External actions)	_ARGument_GENeric
_INTERSUBictive	CHANGE_mind	ActRequest	FACT_NOSRC
Q_INTERLocutor	UNCERtainty	ActPropose	GENERAL_SOLUTN
REF_INTERLocutor	OTHERS_THNK	ActAccept	EXPER_OBSERV
PERSPECTIVE_taki ng	APOLOGY	ActDecline	ARG_OPINION
	HELP	ActNegot	OPINION_ONLY
_META_Dialog	REQ_HELP	(Dialogue_Actions)	OVER_GEN
MEDIATE	PROCESS	DI_ActRequest	SUPPORT
META_CONS	AGREE	DI_ActPropose	SUM_MY-argumt
META_CONFL	DISAGREE	DI_ActAccept	EXAMPLE
META_SUM	_NEGative-emotion	DI_ActDecline	ELAB
META_CHECK	NEGEMO_INTer locutor	DI_ActNegot	
_META_TOPIC		(Facilitators only)	
WEIGH	NEGEMO_Topic	WELCOMING	
SYSTEMs_thinking	_OFFTOPIC	PROC_EXPL	
FACT_cite_SouRCe	TECHnical	MOTIVATE	
SOURCE_REFerence	SOCIAL		
APPRECiation			

Figure 1: Text Coding Scheme

The "Core Set" those we consider the most central skills/behaviors for social deliberative capacity. The "Additional Deliberation Quality Indicators" are also considered good things to have. Our research focus differs from related studies in the Learning Sciences in that focus on cognitive/analytical speech acts that we list under Argument Codes, while we were more interested in intersubjective deliberative skills. In the domains we have codes thus far, all Argument Codes were coded using the meta-code ArgGen, and coders were not trained to differentiate the various types of Argument Codes. (We could of course go back and recode these segments at the more specific level of detail.) The Action Negotiation category was only used for two domains: Workplace Dispute and Faculty Negotiation. These were the only domains in which participants were discussing real proposed extra-dialogue actions or decisions—all other domains were discussions with no off-line decisions or actions involved. Unlike the other codes, the Action Negotiation codes were developed ad-hoc and do not reflect a careful analysis of the literature.

Other than Argument Codes, which were all coded using the meta-code ArgGen, all other codes were coded at the base code level. However, all later analysis, including the interrater reliability numbers shown above, are done at the meta-code level. For example, the Question_Interlocutor, Reference_Interlocutor, and Perspective_Taking codes were abstracted to the _Intersubjective code. Misc Codes cover other categories of interest.[3] The full coding instructions ("manual") is in the Appendix. Table 5 shows a sample four of the codes with example segments from our database.

Deliberative Skill	**Description**	**Examples**
Perspective taking	Social Perspective taking -- putting yourself in another's shoes (of an interlocutor OR a	(1) From both of you I have now a little insight into how you view the problem and what the possible solutions

[3] Note that Negative_Emotion is coded, but we found that Postive_Emotion was usually too ambiguous to code so we do not include it—it may be coded as Appreciation, Agree, Help, etc.

	group you are not a member of). The tone is more one of empathy vs. critique.	could be. (2) I cant help but imagine what that is like for her and for her family. (3) I know firsthand how scary that can be. (4) This might be their only sense of closure after losing someone very dear to them. (5) The reality is that many people of color in this community feel like they are not seen or heard. (6) But I do see how people can link his last name to a specific race.
Systems Thinking	Introducing a larger set of concerns in: time; geography; causality; level; part-to-whole systems. Moving the conversation from individual examples and factors to more inclusive, abstract, or big picture systems of things or factors. (About the *topic*, not the dialogue).	(1) When considering society as a whole I think there is a larger issue at hand. (2) Competition could just put us into a petty fight about territory...in the end, we will all lose and not deliver the potential technology that can improve the [rest] of our world (3) This law allows people to kill based on assumption, which is dangerous in a society where racism, discrimination, and racial profiling is so ingrained in our culture. (4) Considering our country is in so much debt already I'm not sure if we can afford to add to the debt we have already created.
Meta-dialogue	Birds eye view of the discussion, including participant mediation moves, summarizing the conversation, and "how are we doing?" inquiries.	(1) I will collect all the suggestions and put them together. If there is any suggestions feedback, please try your best to provide it in three days.. (2) It would be great if we could summarize some of our previous comments to the group (3) However, it sounds like there are some remaining points of confusion. (4) I am also impressed by the way that the conversations have now started to coalesce. (5) But this conversation is much more than male female. (6) What do you guys think? (6) The majority opinion is clearly against mandatory drug testing...
Uncertainty	Explicitly express uncertainty, ambiguity, confusion, or ignorance (about the topic, or uncertainty about another interlocutor's thoughts).	(1) I'm a little confused. (2) That's the limits of my insight -- I still have no idea how to begin this process of building a truly inclusive community. (3) I have to say I'm confused as to why an all white group on election night wouldn't vote for you, a white woman to join the board. (4) I'm not sure I understand the benefit of limiting the amount of alcohol people under 18 can purchase. (5) I don't know how much I can speak to my own white privilege, (6) But I'm still debating my own answer to the question.

Table 5: Sample From Coding Manual

In addition to the Base-code and Meta-Code levels, we calculated a higher level construct Total-SD-Skill, a Boolean indicator of whether any of the Core or Additional dialogue quality indicators described the segment. This allowed us to test for general "skillfulness" improvements in our experimental trials, and also gave us a Boolean measurement to use in text classification of deliberative skill, making a difficult task more tractable. (Total-skillfulness at the participant level was the percentage of segments classified as Total-skill=1.)

Coding Tool

As part of this project we have developed a database tool for the manual coding and processing of text, because we had requirements not met by any single available application, including:

- encoding the **reply structure** of dialogue text (i.e. social network structure),
- having **second-choice codes** for tie breaking (i.e. coders could specify a second segment code and specify weather it was "-also-," meaning both codes were applicable, or "-else-" meaning the coder was not quite sure which of two codes applied),

- advanced **searching and sorting** features,
- allowing base level codes and "**meta-codes**" that span several base codes, with analysis supported for both levels;
- allowance for more than one coding scheme to be used,
- password-based login accounts for each coder, with the ability to limit read/write access on a fine grained level for each user type,
- allowance for more than two coders on a project,
- **cloud-based** application allowing multiple coders to add to a common data base,
- storing of **participant types** (e.g facilitator), **demographic** data (gender, age, etc.), and experimental groupings (e.g. control, intervention-A, intervention-B),
- encoding of **time-based "phases**" of each post (e.g. introductions, brainstorm, critique, agreements),
- additional fields **characterizing discussions** or domains, e.g. synchronous/asynchronous; facilitated/non-facilitated.

A screen shot from our coding tool, called Codoole, is shown in Figure 2. The tool is a cloud-based application built by our team using FileMaker Pro. (The system is an 'alpha' version prototype, not yet in a stage where it can be used by other research groups without 'handholding' by us.) The menu at the top shows how the structured database can be viewed from the perspective of Forums/Sites/Domains—each of which has several Topics/Discussions, each of which has many Posts, each of which has one or more Segments. The Codes menu includes reference material for codes, and the ability to define new coding schemes and meta-coding schemes. Within the screen for each data level are tabs for various functions. Tabs showing Summary/Overview lists and descriptive statistics are available at each level of data (Forum, Topic, Post, Participant, and Segment).

Figure 2 shows the Posts data level screen, within which there are several Tabs, including one used during coding and one used to compare codes across raters (which coders are forbidden to look at while they are coding).

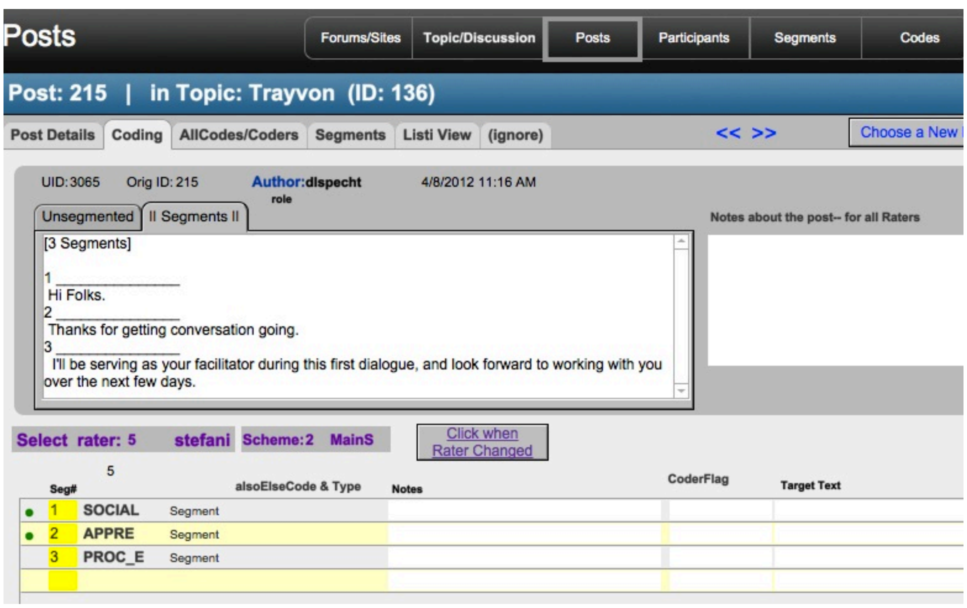


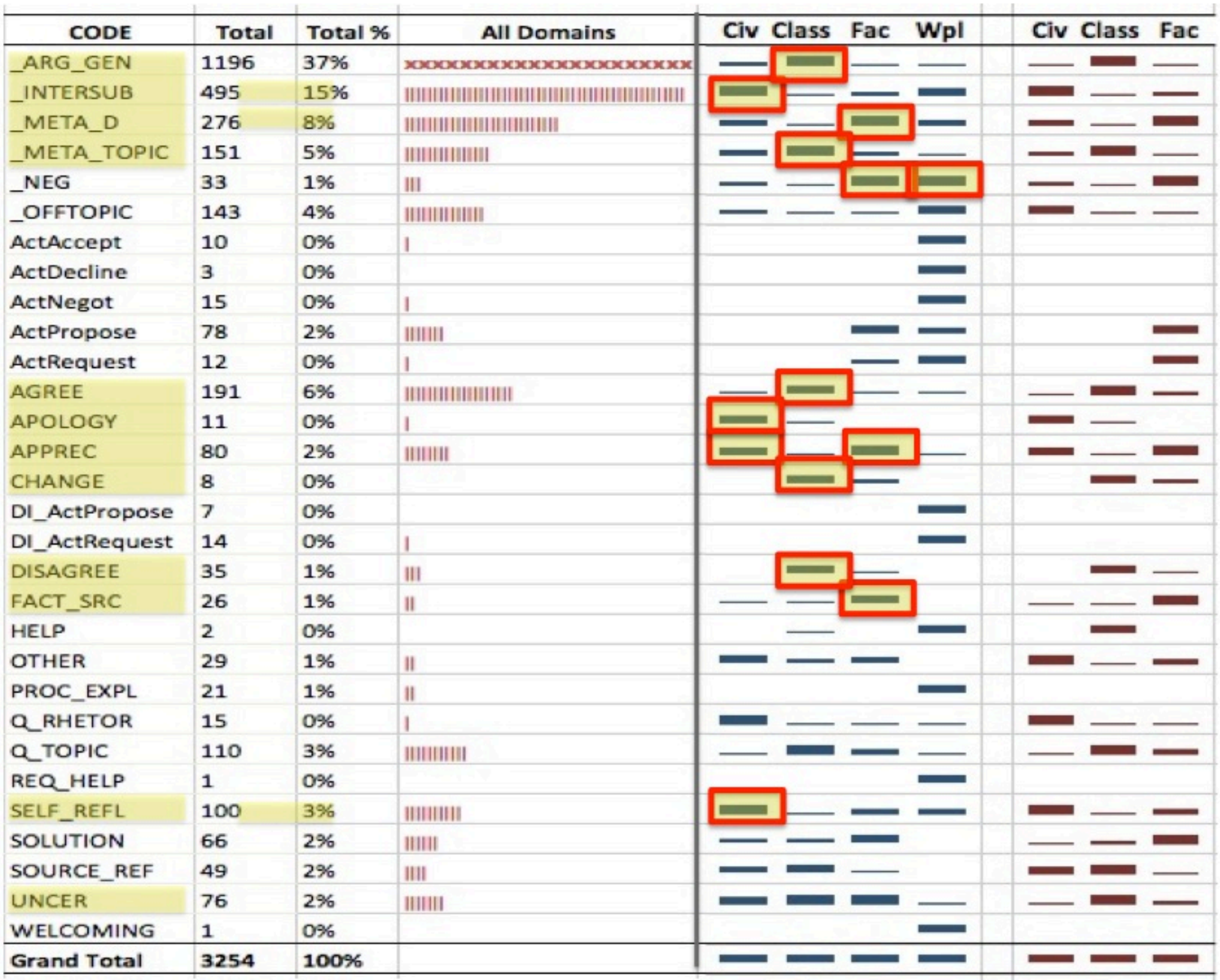
Figure 2: Codoole--Coding tool

Data analysis overview

We report on several streams of analysis below (in addition to the descriptive statistics noted above): (1) a comparison of characteristics of the domains, to (a) help interpret later results that are domain-specific, and (b) take initial steps toward a large goal of designing a domain-characteristics scheme for use in deliberative skill research; (2) an analysis of code frequencies in each domain to investigate the relationship between domain characteristics and deliberative skillfulness; (3) investigations of the correlations between automated linguistic measurements and hand-coded skill scoring; (4) efforts to build a classifier that will predict overall SD-skill, (5) efforts to build classifiers to predict individual SD-skills; (5) efforts to improve total-SD-skill classification by finding domains that serve as the best training sets; and (6) efforts to determine which set of linguistic measures allow the best total-SD-skill classification.

Multi-domain Descriptive Statistics

We did some preliminary analysis of how dialogue characteristics differ for different types of domains. This allows for cross-domain comparisons and to situate the measurements of each domain in a larger context. Figure 3 illustrates the frequencies of code (meta-code) categories found in the four domain data sets of 3254 coded text segments (codes not shown had zero occurrences; in the figure Total % is rounded to the integer, so "0%" can be misleading): Civic deliberation, College class discussions, Faculty negotiation, and Workplace Dispute.



CODE	Total	Total %
_ARG_GEN	1196	37%
_INTERSUB	495	15%
_META_D	276	8%
_META_TOPIC	151	5%
_NEG	33	1%
_OFFTOPIC	143	4%
ActAccept	10	0%
ActDecline	3	0%
ActNegot	15	0%
ActPropose	78	2%
ActRequest	12	0%
AGREE	191	6%
APOLOGY	11	0%
APPREC	80	2%
CHANGE	8	0%
DI_ActPropose	7	0%
DI_ActRequest	14	0%
DISAGREE	35	1%
FACT_SRC	26	1%
HELP	2	0%
OTHER	29	1%
PROC_EXPL	21	1%
Q_RHETOR	15	0%
Q_TOPIC	110	3%
REQ_HELP	1	0%
SELF_REFL	100	3%
SOLUTION	66	2%
SOURCE_REF	49	2%
UNCER	76	2%
WELCOMING	1	0%
Grand Total	3254	100%

Figure 3: Code Frequencies for all domains

The red horizontal bars illustrate percent of the total across all domains (with Arg_Gen removed to better discriminate the lower frequency codes). The first set of (blue) vertical bars illustrate the percentage of each code within each domain. how these totals are distributed over the domains (for example, for INTERSUB the large box under Civic, next to the smaller boxes for other domains, indicates that the Civic domain had a larger percentage of its posts as INTERSUB). The second set of (red) vertical bars reiterates the first set, but comparing only three domains rather than all four, because the fourth domain is in many ways an outlier.

Our analysis of the frequencies of codes in each of the four domains (and total for all domains) shows several things:

- The **Arg_Gen**, or generic argumentation codes predominate, as might be expected. 37% of all segments were Argumentation moves, which include explanations, elaborations, self-supports, opinions, and fact-giving (without citing a source). (The horizontal bar for ARG_GEN is not shown, and the horizontal bars are scaled to give INTERSUB, at 15%, the full width.)
- **Intersubjective** speech acts were the second highest frequency overall, and were also second in the classroom (14%) and workplace (20%) domains. This is an indication that all conversations were at least moderately reflective and interactive.

- **Meta-dialogue** had the third highest frequency overall (8%), and meta-topic was also high (5%). It makes sense that the highly educated Faculty group would have a high level of meta-dialogue. Also, because one of the issues for these participants was group dynamics or decision making power, the quality of the deliberative dialogue was more often an explicit topic of conversation. At first glance it seems unusual that the College students had higher meta-topic percentage than Faculty, but this is probably because the topics of conversation for the students were general, while the Faculty conversation was oriented to concrete logistical matters.
- **Agreement** constituted a relatively large percentage of segments (5.9%), while disagreement was much lower (1.1%)—and this trend held within each domain also. But the vast majority of these posts were in the classroom domain (see Classroom below)—88% of the Agree and 94% of the Disagree codes were in the Classroom domain.
- There were a significant number of **Uncertainty** codes (the Classroom, Faculty, and Civic domains all had similar frequency percentages, spanning from 2.5% to 2.9%). However there were very few Changes of mind coded (7 in Classroom and 1 in Faculty). The literature on deliberative dialogue confirms that deliberation is likely to lead to better mutual understanding and higher uncertainty, but relatively little changes of mind. As noted below, SELF_REFLection was high in the Civic and Faculty domains, and notably low in the Classroom domain. This may indicate the difference in average age and maturity, but it also may be because the Civic and Faculty discussions involved read decisions in real life.
- For the **Faculty** domain **Meta_dialogue** had the second highest frequency (19%) and INTERSUB was third (16%). The Faculty domain also had the highest number of sources cited, which is congruent with the academic background of participants.
- The **Civic** domain was especially high on **SELF_REFLection** (8.8% of its posts, vs 0.8% for Classroom, 5.2% for Faculty, and 4.4% for Workplace). Self reflection often implies a move away from certainty into more nuance and contingency of opinions. In this domain participants were more part of an established community with longer term relationships, and presumably the trust level was higher so that there was more self-reflection because increased trust can lead to increased allowance for uncertainty. The Civic domain contained most of the APOLOGYs also, possibly related to trust, but also because much of the conversation reflected on past actions.
- The **Classroom** domain had particularly high percentages for ARG-GEN (48%), AGREEment (9.5%), DISAGREEment (1.85%), and asking questions about the topic (5%) (in each of these the codes were at least twice as high a percentage of all codes compared with the other domains). These findings are consistent with the fact that in this domain participants were specifically tasked with discussing a single topic, and unlike the other domains, the topic of discussion was not directly related to their personal relationships or external activities.
- The **Workplace** dialogue, which was already noted to have some characteristics distinct from the other three domains, is seen to have distinct characteristics in Tables 1, 2. Because the dialogue was a negotiation about specific actions, it contains the vast majority of the action-based codes (ActXXX, DI_ACTXXX) (there were a small number in the Faculty dialogue and none in the other two domains). It also had the

only codes for PROCedural_EXPLanation, REQest_HELP, and WELCOMING, because it included heavy facilitation.

Total SD-skill. As described above, we also calculated a Total-SD-skill values for each domain, the percent of segments coded as core skills or additional dialogue quality indicators: 57% for Civic, 32% for Classroom, 53% for Faculty, and 39% for Workplace; with an overall value of 39%. Consistent with the above analysis, the Civic and Faculty domains showed a higher degree of overall deliberative skill. The Workplace domain is not very comparable, as many of the posts were by a professional facilitator.

Our above analysis of code frequency data serves several purposes. The alignment of much of the above analysis with what would be expected given the characteristics of a domain gives credence to the face validity of the coding scheme. The above analysis is one step toward developing a framework to characterize the differences between deliberative domains. This can be used (1) in understand the reasons that various findings differ by domain, and (2) in the future in attempts to create a general scheme for comparing the deliberation-relative properties of deliberative domains. Finally, it informs the question of which codes are most revealing, and thus also may influence which codes should have more priority or weight as we attempt to build machine classifiers.

Skew in Code Frequency. As mentioned, data skew or imbalance is one of the challenges of working in this domain, and Table 6 shows a comparison of domains that speaks to this issue. It shows the percent of the segments coded as one of the Total-SD-skill set, the percent coded as ARG_GEN, which tended to dominate the other codes, and all others. From this table we can see that the classroom domain has the highest skew, with the largest percent of ARG_GEN and the smallest percent of SD-skill. The civic domain had the least skew, with the largest percent of SD-skill, and a relatively low percent of ARG_GEN. Later we report that the domain with the lest skew was the optimal one to use in machine learning for text classification.

	Civic	***Class***	***Faculty***	***Workplace***	***Total***
Total_SDskill%	55%	29%	53%	39%	38%
ARG_GEN%	27%	48%	24%	22%	37%
OTHERS	18%	23%	24%	39%	26%

Table 6: Analyses of Data Frequency Skew by Domain

Classroom Experiments

One of the goals of education is to produce competent national and global citizens capable of participating in democratic self-governance and capable of wrestling with the difficult questions and dizzying array of information and opinion they face in our technologically advanced society. Engaging with others on complex topics requires not only learning the relevant facts and concepts and making logical inferences, but also engaging with the perspectives and opinions of others who may not share one's views or goals. Doing so requires skills that can be systematically supported (King & Kitchener, 1994; Rosenberg, 2004; Herzig & Chasin, 2006; Holman et al., 2007).

We differentiate our research from others that focus on *argumentation*, which aims to help learners generate logical, well-formed, well-supported explanations and justifications (Andriessen et al., 2003, Baker et al. 2007), usually framed in objective rather than intersubjective terms. That is, they are about finding the right answer or the most efficient and effective solution to a technical or scientific question—but don't adequately address the specific moments of deliberation or collaboration where opportunities for mutual understanding and mutual recognition arise.

We use the term "social deliberative skills" (SD-skills) to indicate the capacity to deal productively with heterogeneous goals, values, or perspectives, especially those that differ from ones own. SD-skill includes social perspective taking, meta-dialogue, social inquiry, systems-thinking (complexity thinking), and self-reflection. Though the teaching/learning/support (including computer-based support) of these related skills have been researched intensively, the prior research does not adequately address some key challenges in building mutual understanding and mutual regard when interlocutors encounter the disequilibrium of diverse perspectives. This research makes an incremental contribution in this area.

An important research question for this grant is whether social deliberative skills can be supported in online dialogue environments through passive scaffolding features that bring attention to concepts or behaviors related to higher quality dialogue and deeper engagement. An experiment with college students engaged in online dialogue about controversial topics indicated that indeed such software features can result in improved dialogue, and with a large effect size. In the 2013 papers "Supporting Social Deliberative Skills Online: the Effects of Reflective Scaffolding Tools" and "Toward Defining, Justifying, Measuring, and Supporting Social Deliberative Skills" we report on an evaluation of an experimental study summarized below.

We are in the process of analyzing data from several other classroom-based studies, including one that attempts to replicate the results of the study below in a new domain, and another series of experiments in which have a phase-in design, in which reflective tools are phased in one at a time, and we look for subsequent changes in dialogue quality. In addition, in the new data sets we have gathered pretest data on student personality type and reflective skill level, which we will correlate with social deliberative skill.

Method

For the online discussions we used the Mediem deep dialogue discussion forum software created by Idealogue Inc. (see Figures 4, 5, 6).[4] In addition to standard (semi-threaded) discussion forum features, Mediem has a number of features intended to support deeper engagement and reflection (based in part on the designers' many years experience with members of the National Coalition on Dialogue and Deliberation).

Hypothesis. Participants were put in three experimental groups: 1) the "Vanilla" (control) group using only plain discussion forum features; 2) the "Sliders" group using a slider tool to rate opinions; and 3) the "Reflective tools" group using tools designed to support meta-dialogue, good question asking, and self-reflection (described below). The primary research hypothesis was that the features intended to support SD-skills, i.e. in groups 2 and 3, would be shown to do so based on hand coding of participant posts. We are also interested in relationships among skill use, posting activity, response relationships, and survey results.

Mediem features. Mediem has been used in a number of dialogue contexts including interfaith discussions among college students. Figure 4 shows the Mediem home screen, with sections listing Dialogues ("Conversations"), Opinion Sliders, Participants, and Resources. Each section lists items that can be expanded for full view. Dialogues are semi-threaded discussion

[4] We worked with Idealogue to create an API for exporting the data from the dialogue (posts and other user actions) for our monitoring and data analysis. We also worked with them to build additional customization features supporting experimental trials.

forums with additional features mentioned below. Normally participants in open-ended discussion will propose their own dialogue topics and "set the table" for a conversation by specifying certain parameters (number of participants, demographic information, etc.) and inviting others to join; however, in our study we used pre-determined dialogue topics entered by the facilitator. The Participants section shows participant profiles, and the listing can show graphical indications of demographic and other participant information. The Resources section allows participants to upload documents and links related to the conversation. We did not use the Participants or Resources features for this study.

Figure 4: Mediem Home Screen

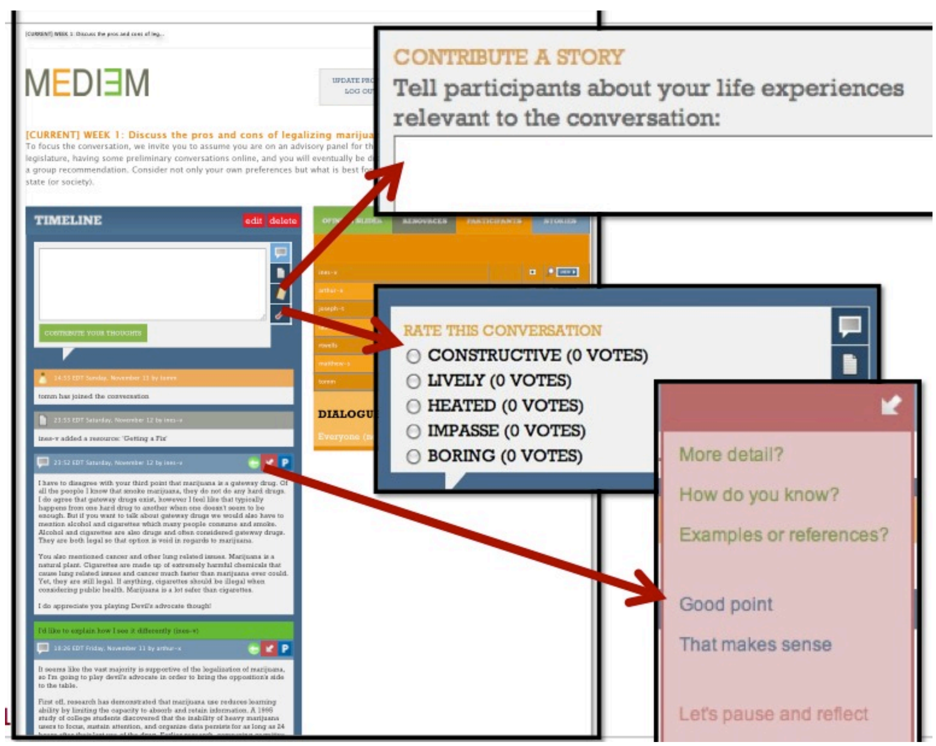


Figure 5. Mediem Dialogue Forum and Reflective Tools

The Mediem software was chosen for our study because it has a number of features designed to support deeper reflection and engagement. Figure 5 illustrates the expanded view of a Conversation (Dialogue), showing three such features illustrated separately. The discussion is viewed in the "Timeline" with most recent activity on top. Participants type their thoughts in the empty box at the top and submit. The Timeline shows posts and also other events (resources posted, conversation ratings, etc.) in temporal order. Posts are replied to using the arrow-shaped button above a post. To the left on the screen are tools for viewing participants, sliders, stories, and resources associated with the particular Conversation.

Figure 5 shows the three reflective tools used in the "Reflective tools" group. First is the Story feature, which gives participants a special place to say how the issue at hand relates to them personally, including relevant background information about themselves and "what is at stake" for them in the issue. Second is the Conversation Thermometer, a meta-dialogue tool that allows participants to rate (vote on) the quality of the conversation at any time. The choices can be customized by the administrator. Third is the Contribution Tag feature, which allows participants to give brief comments on other's contributions. It provides a fixed vocabulary similar to the sentence starters (or locution openers) used in other dialogue software, but the tags remain attached to the target post rather than starting a new post (see Soller, 2001).

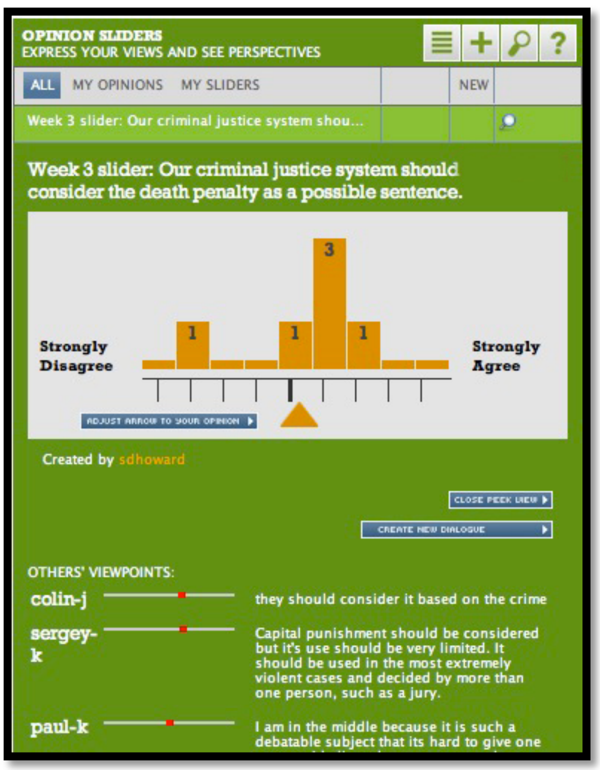


Figure 6: Mediem Opinion Slider

The software includes an Opinion Slider, a polling feature used in the "Sliders" group, shown in Figure 6. (As with Conversation topics, participants usually set up their own Opinion Slider questions, but ours were pre-defined for the classroom dialogues.) Sliders are thought to provide a motivational, brainstorming, and group-awareness function similar to Student Response Systems ("clickers"), which draw attention to differences, similarities, and diversity of opinion within the group as a whole. The slider gives a summary view of where participants stand on an issue.

Participants and discussion questions. Twenty six students in a college Alternative Dispute Mediation class discussed two topics (one each week over two weeks) in moderated discussions using Mediem. Students were randomly broken into three discussion groups of 8-9 members each, with all groups given identical questions. The activity was a required assignment that was part of the course and students were given class credit based on participation alone (not the content of participation). They were required to post at least once every day. In a prior class session students had brainstormed interesting and controversial topics for this activity. The discussion topics chosen were 1) Trayvon Martin killing in Florida, and 2) Gun Control.

Facilitation. We employed the service of experienced facilitators. To keep the control and experimental groups comparable, the facilitators were asked to keep their interventions to a minimum, and if they made an intervention in one group to do something similar, or at least something of similar length, in the other groups. Facilitator #1 facilitated all three groups during the Trayvon Martin discussion (week #1) and facilitator #2 facilitated all three groups during the Gun Control topic (week #2).

Data collected and analyzed. The three groups had similar numbers of students participating in the discussions (Vanilla 9, Sliders 8, Reflective Tools 9). There were 8 males and 14 females ranging in undergraduate grade level from sophomores to seniors, with one non-degree student. All text from student posts was collected; in addition, "reply" connections between posts were collected. Data were collected on Slider, Story, Conversation Tag and Thermometer use in the groups where these features were offered. All subjects were given a post-survey including 18 questions using a 5-point Likert "agree...disagree" scale.

Coding. Text of student posts was divided into segments and coded by two independent coders using a coding scheme developed by our group that focuses on social deliberative skills and other indicators of dialogue quality. Our coding scheme has 42 categories, 17 of which indicate deliberative skill. This scheme synthesizes prominent frameworks found in the literature (Black et al., 2011; Stromer-Galley, 2007; Stolcke et al., 2000) and adds codes for dialogue quality specific to SD-skills (Murray et al., in preparation). Cohen's Kappa Interrater reliability measure for this coding scheme is 71%, (76% agreement) averaged over five dialogue domains we have used it in(this level is considered "good" (ref) and is particularly good given the complexity of our coding scheme). For this classroom data that is the subject of this paper the interrater agreement is 77% and the Cohen's Kappa is 72%.

For this experiment, 7 codes were singled out for data analysis: *Intersubjectivity*: perspective awareness, perspective taking or question asking; *MetaTopic*: Birds eye or systemic view of the topic (related to complexity or systems thinking); *MetaD*: Meta-dialogue, discussing the quality of the dialogue and proposing changes to its structure; *Appreciation:* Gratitude, affirmation of another's idea or situation); *Apology:* noting and/or taking responsibility for one's errors; and *Source Referencing*: mentioning a source for a fact or idea. A Total-SD-Skill score was computed for each segment by adding the scores of the seven skill measures for that segment. An average Total-SD-Skill score per segment was then computed for each student in each discussion.

Students who posted fewer than 5 times for both topics combined are excluded from statistical comparisons. Also, preliminary analysis revealed several issues with the Sliders group sufficient to lead us not to include this group; we compare only the Vanilla and Reflective tools groups.5 Students in these two groups who met the criteria for inclusion happen to be balanced in total number and in gender, though not in grade level. Although the individual codes included in the study had been determined to show no effect due to grade (within-group ANOVAs ranging from p = 0.25 to p = 0.78), due to the difference in distribution of juniors and seniors, we continue to include grade as a potential factor in correlations.

Classroom Study Results

In this section we will report on: (1) the main question of whether the group using reflective tools showed higher (total and subskill) skill levels than the control (Vanilla) group; (2) look for possible relationships between SD skill scores and gender, post size and frequency, post reply statistics, tool use statistics, and survey results. Participation and basic statistics for the Vanilla and Reflective Tools groups:

- The data set over the two groups contains 241 posts and 516 segments; for an average of 15.06 posts for each student over both topics (*SD* = 7.45).
- The mean words per post was 53.60 (*SD* = 42.12) and the mean characters per post was 299.40 (*SD* = 241.95).
- We found no significant relationship between the number of posts and the length of posts among participants.

The average student skill scores as percentage of each student's segments were:

Intersub	Meta_D	Meta_Topic	Apology	Apprec.	Fact_Src	Src_Ref
25.08%	0.88%	5.51%	0.22%	1.30%	0.28%	1.20%

The main results of the study include (see Stephens et al. 2013 for more detail):

- A main effect between Total-SD-Score and grouping, $F(1, 14) = 6.89$, $p = 0.02$*, $d = 1.46$ (a very large effect) in favor of the Reflective Tools group. Thus our main hypothesis was confirmed.
- A significant relationship between Intersub and grouping, $F(1, 14) = 4.81$, $p = 0.05$*, $d = 1.05$ (a large effect) in favor of the Reflective Tools group. Intersub was strongly correlated with Total-SD-Skill, indicating that most of the effect of Total-SD-skill comes from the Intersub subskill. There was no significant relationship between any of the other subskills and group.
- ANOVAs revealed no difference due to gender on the Total-SD-Skill score or on any of the subskills except for Appreciation, where females scored higher, F (1, 14) = 5.59, p = 0.03. Six females had at least one segment coded Appreciation; none of the males did.
- From the survey, there was some positive correlation between Total-SD-Skill and self-reported Engagement (r = 0.44) and Learning (r = 0.21). These results conform to our intuitions that those exhibiting more skill would find the experience more positive, though we cannot infer causation in either direction.

[5] In the Sliders group one student failed to follow instructions (did not use the sliders). This student dominated the discussion, contributing over a third of the total posts. Two other students in this group did not post enough to be included in the analysis. One student wrote a note to the facilitator claiming that one student in this group seemed overly critical and not respectful, which affected her feeling of safety.

Next we look more closely at two phenomena: the use of the reflective tools, and the reply structure among participants.

Student replies to each other. The number of contributions that reply to (or refer to) other contributions is one indicator of the robustness of deliberation (Stromer-Galley, 2007; Suthers 2008). We analyzed several quantitative metrics related to this phenomenon. Our hypotheses were: 1) that students with higher skill (especially the intersubjective code), and 2) students showing positive survey opinions—would post more replies; and 3) that the reflective tools would support more replies. Results are:

- The average number of posts per student that were explicit replies to posts of another student (Replies_by_me) were 10.59 (*SD* = 3.41), about 71%. This is relatively high, suggesting engagement. The average number of replies each student received (Replies_to_me) was 10.35 (*SD* = 6.86)—a high reciprocity.
- There was a correlation between Replies_by_me and Replies_to_me: R = 0.8284. In other words, students who replied more to posts of their fellow students received more replies in return.
- There was no main effect on Replies-by-me or Replies_to_me due to experimental group and no significant relationship between Replies_by_me or Replies_to_me and grade level within either group.
- There was no significant difference between genders in the numbers of Replies_by_me or Replies_to_me within the Reflective Tools group. However, within the Vanilla group, females replied to others significantly more often than males, t = 2.68, p = 0.04*; females replied more than twice as often as males.

In summary, our hypothesis that reflective tools would support more replies was not supported. A majority of the posts were replies to other posts and were replied to in turn; students who replied more to posts of their fellow students received more replies in return; and, interestingly females in the Vanilla group replied to the posts of others more than twice as often as did the males within that group.

Use of Reflective Tools. The reflective tools group had at their disposal a set of three tools that constitute innovations over what is offered by most discussion forum software: the scaffolded post comment tool, a discussion temperature rating, and a story tool where participants could write personal stories about the topic. We hypothesized that there would be a positive relationship between the amount of tool use and evidence of social deliberative skill (presumably because making use of the scaffolding supports bringing skills to bear in the dialogue, but causation can not be inferred from the data). This hypothesis was confirmed in finding a positive correlation between intersubjective speech acts and total tool use (R=.54) and dialogue temperature tool (R=.85) (this was for the Trayvon topic, as discussed below).

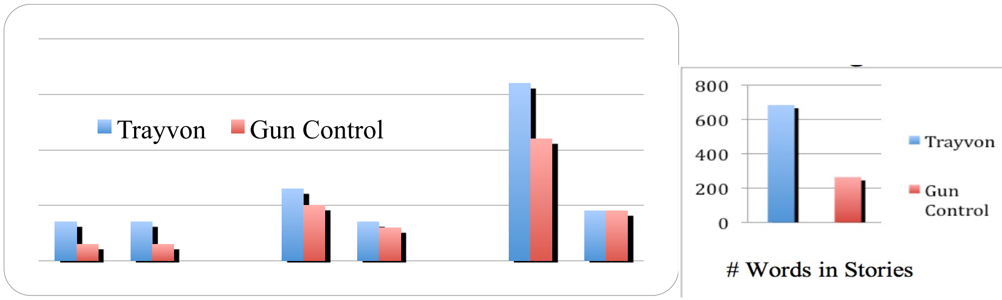


Figure 7 A: Reflective tool use vs. topic; B: story words vs. topic.

The amount of tool use is shown in Figures 7a and b. For this analysis we separated the data by discussion topic because we noted more tool use in the first topic, Trayvon Martin. As can be seen, students posted less in their stories, used the discussion temperature tool less, and posted fewer comments for the second topic (Gun Control). We believe this could have been due to several factors: the novelty and motivation to do this homework task could have worn off after the first topic; the Trayvon topic was more specific (involved specific people) and could have been related to more easily than the Gun control topic; Trayvon was salient from recent news reports; the second topic was facilitated by the less experienced facilitator. In general, the participation levels in the reflective tools group were acceptable but not particularly high. This concurs with the average survey Engagement rating, 4.0 in a 1-5 scale. This analysis also highlights the potential large effects of choice of topic and other context variables on measures of student deliberation and problem solving.

Facilitators Dashboard

Dashboard Development

Our goals for research into supporting deliberative skills include supporting participants directly (as in the reflective tools described above) and supporting *facilitators*—who can be teachers, mediators, peers, moderators, etc.—as they in turn attempt to help participants engage in skillful dialogue and deliberation. In our interviews with facilitators and teachers we have identified the following types of phenomena that are indicators of **potential problems in dialogue**, and which are particularly important in online facilitation:

- Low or no participation of individuals or groups, or silences or lulls on the part of individuals, the entire group, or sub-groups
- Conversation domination by an individual or group
- Overconfidence or unwillingness to consider other perspectives
- Individuals (or a group) who only communicate with one other or a subgroup, or ignores a particular other or subgroup
- Inappropriate or disrespectful behavior
- Off-topic conversation
- Negativity or negative emotion; demotivating, discouraging, or impeding language
- Tension-filled disagreements, or highly emotional content
- Too much agreement or politeness
- Misunderstanding due to missing communication skills normally available in face-to-face communication
- Violation of rules (e.g. confidentiality, no advertising, etc.)

Note that these are described in terms of negative behaviors, yet for many of them we could define positive behaviors corresponding to their opposites, for example identifying off-task behavior is analogous to identifying on-task behavior. In addition there are a number of **other positive behaviors** that facilitators indicated are important to monitor, many of which correspond to our deliberative skills (this is in part because in interviewing facilitators we showed them our deliberative skills list):

- Question asking and curiosity
- Empathy and perspective taking
- Citing reliable sources in arguments
- Appreciations and apologies
- Considering and weighing alternatives beyond initial beliefs
- Helping others or the group to come to agreement, harmony, or understanding
- Using "I language" and avoiding judgmentalness and sermonizing others

We have prototyped a "Facilitator Dashboard" that provides third parties a "birds-eye view" of the state and flow of online engagements—see Figure 7. We have also piloted its use as a feedback and "awareness tool" for participants. As we describe below, we have built visualization tools to monitor many of the phenomena listed above and we have plans to build tools to monitor all of the listed phenomena eventually. Some of this analysis will require emerging text analysis methods described in the Text Classification section of this report.

We have built an API that allows the Dashboard to receive real-time updates on the dialogue state and text from the Mediem discussion forum software. See Figure 7. Pie and bar charts show participation levels (number and size of participants) for individuals and subgroups (based on any demographic categorization, such as religion, gender, grade, etc.). Trendlines show trends in these same metrics. A social network diagram shows who is replying to whom. A Word Cloud graphically shows word frequencies. As described in the Year 2 Report, we have prototyped the Dashboard with two skilled facilitators, and have received usability feedback from a panel of 10 skilled facilitators and mediators. We have had extremely positive feedback about these tools and their potential to help online facilitation.

As an example, Figure 8 shows pie charts for men vs. women for two measures of participation: number of posts and average size of posts. As you can see the women had more posts but the men had longer posts. This was of interest to the facilitator, first because seeing only one of these indicators might be misleading in terms of evaluating participation, and second because these metrics seemed to correspond with some research on gender differences in dialogue.

Figure 9 shows the Dashboard's Dialogue view, which shows the text in a format similar to what participants see in the Mediem discussion tool (and is accessed through the Dialogue tab). The facilitator can use the Dialogue view to focus on a particular subset of posts, for example those by an individual or demographic group, those within a particular time period, or those containing certain words.

Our Dashboard contains analyzers that watch for patterns in the dialogue and flag occurrences that reach a specified threshold. The current version of the Dashboard uses a dictionary-based keyword matching method to identify characteristics of the dialogue and potential hot spots. The dictionary was derived from the Linguistic Inquiry Word Count application, which contains sets of words that have been heavily researched for their psychological and social significance in many textual contexts (see Tausczik & Pennebaker, 2010; Pennabaker et al. 2007). The LIWC system has 84 output measures for each text sample, most of them frequencies of words occurring in dictionary categories. For the Dashboard analyzers we identified 14 LIWC dictionary categories that related to what experts said was of interest, and added three additional categories (always-never, question words, and should; for a total of 17 Analyzers). The these text analysis categories are listed in Table 7, along with the number of word stems in each dictionary entry. For example, the category You has 20 entries: thee, thine, thou, thoust, thy, y'all, ya, yall, ye, you, you'd, you'll, you're, you've, youd, youll, your, youre, yours, and youve. These categories have been developed over several iterations with experts and experimental data (ibid). Note that some categories match a large number of word stems, for example positive emotion contains 504 stems.

Category	Number of word stems
always never	2
anger	185
anxiety	91
assent	31
certainty	85
negation	57
negative emotion	500
positive emotion	404
question words	8
sad	101
sexual	97
should	1
swears	54
tentative	54
they	10
we	12
you	20

Table 7: Dictionary Categories for Dashboard Analyzers, with number of word stems in each

In Figure 9 one sees how the results of the Analysis processes are displayed in the Dialogue view, with yellow-colored annotations below each post, and the corresponding matched words highlighted in red in the original text. Note that the concept of Analyzers does not need to be constrained to keyword (or dictionary word stem) matching, but can be extended to other types of analysis (which we plan to do, for example classifying text as formal vs. informal or abstract vs. concrete). Our formative trials have shown that the representation of Analyzer results shown in the Dialogue view is too busy and complex for most facilitators. Figures 10 and 11 shows a more graphical view, which is accessed through the "Advice" tab. For each of the 17 Analyzers two pie charts are shown. One illustrates the percentage of all of the posts that are flagged by that Analyzer. The other illustrates the breakdown by participant among the posts that were flagged positively.

As we add more computationally intensive forms of analysis for the facilitator to visualize, we acknowledge that the metrics or classifications produced will be machine-generated heuristic guesses. We rely on the intelligence of facilitators to determine which results make sense, and which warrant a dialogue intervention. Figure 12 shows the Settings screen which illustrates that facilitators can choose to ignore (turn off) any alert. The human intelligence of the facilitator works in coordination with the "intelligent" analysis of the system.

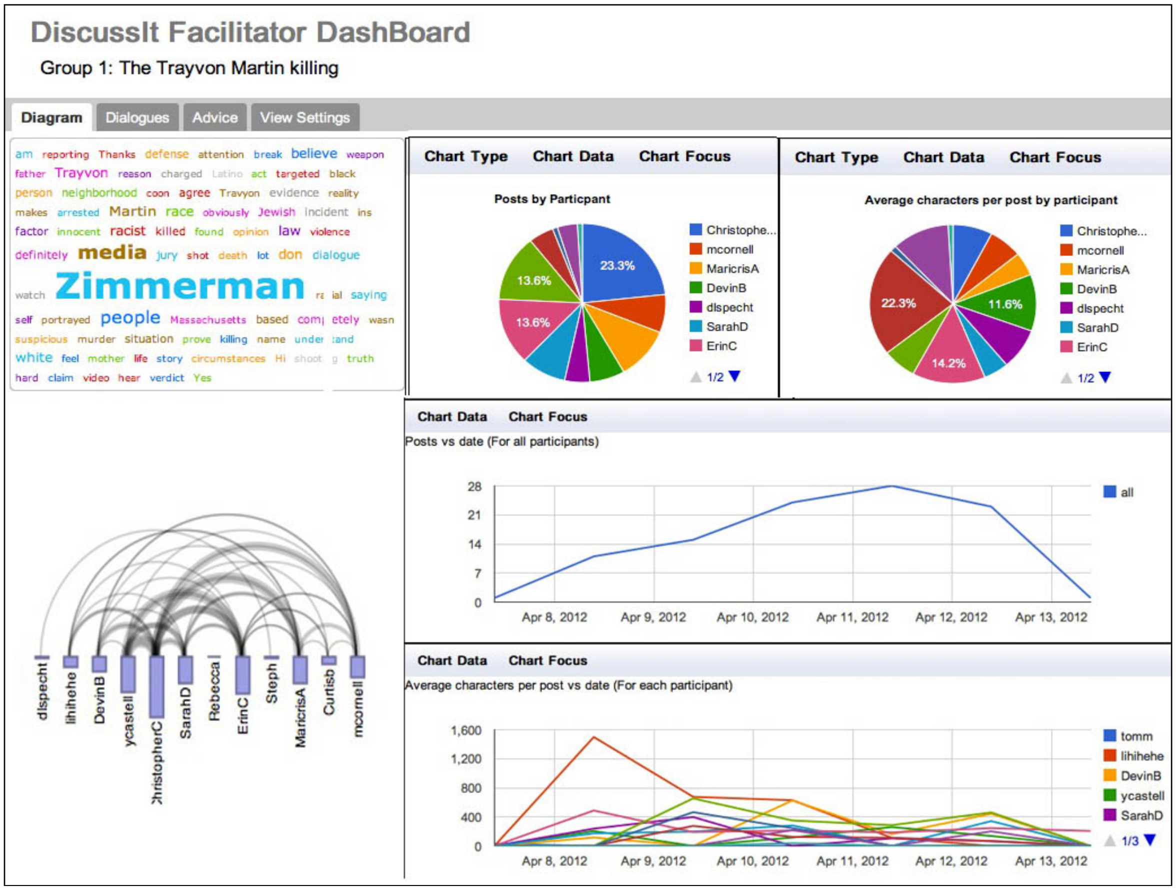


Figure 7: Facilitator Dashboard

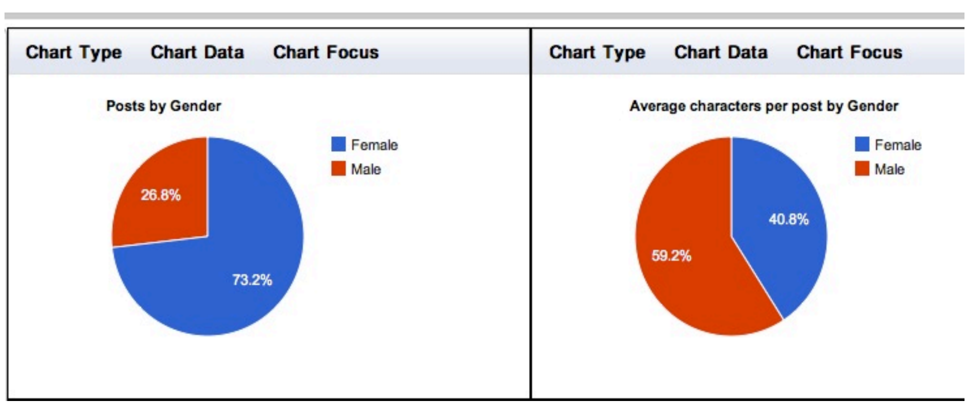


Figure 8: Posts vs. Post size for Males vs. Females

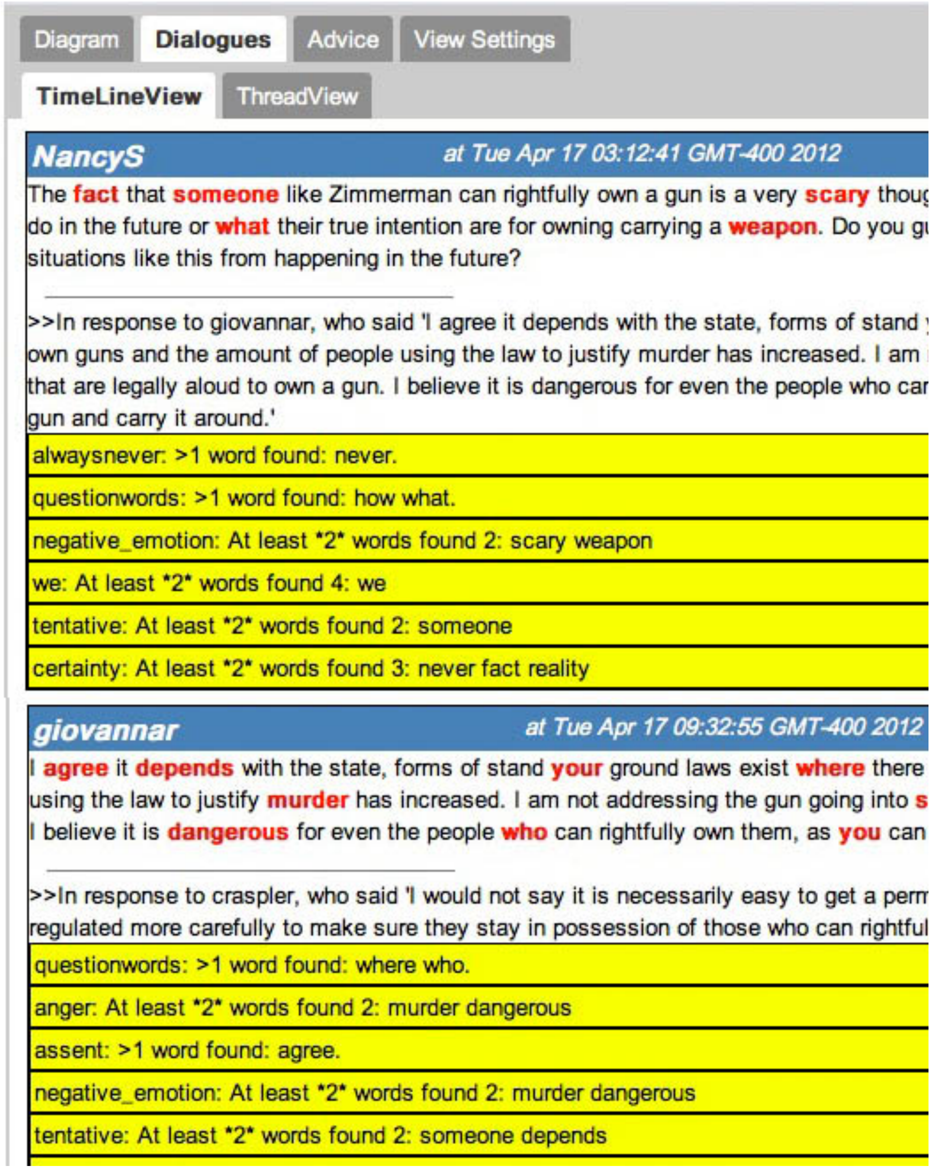


Figure 9: Dashboard -- Text View

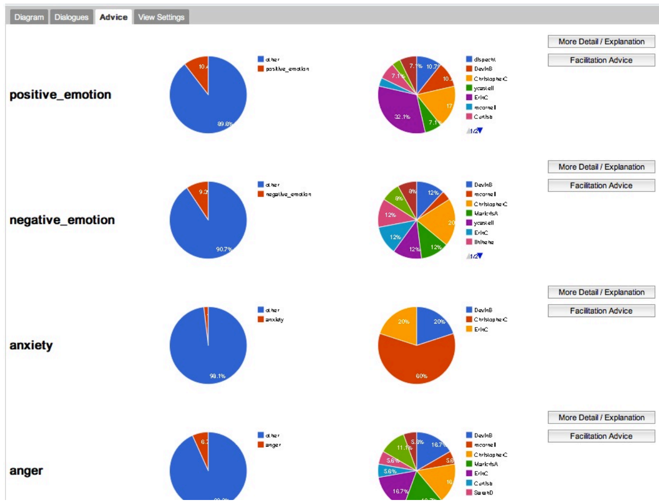


Figure 10: Dashboard -- Advice panel

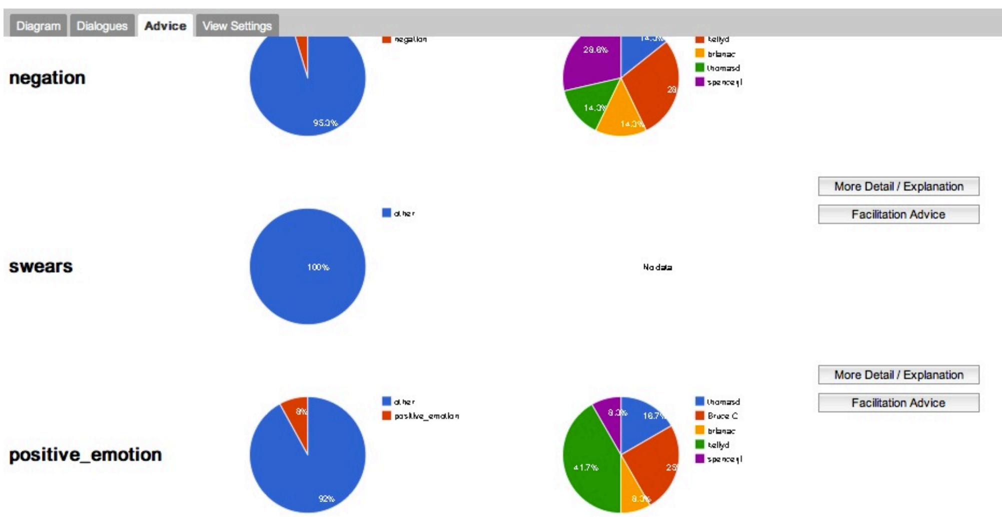


Figure 11: Dashboard -- Advice panel Expanded

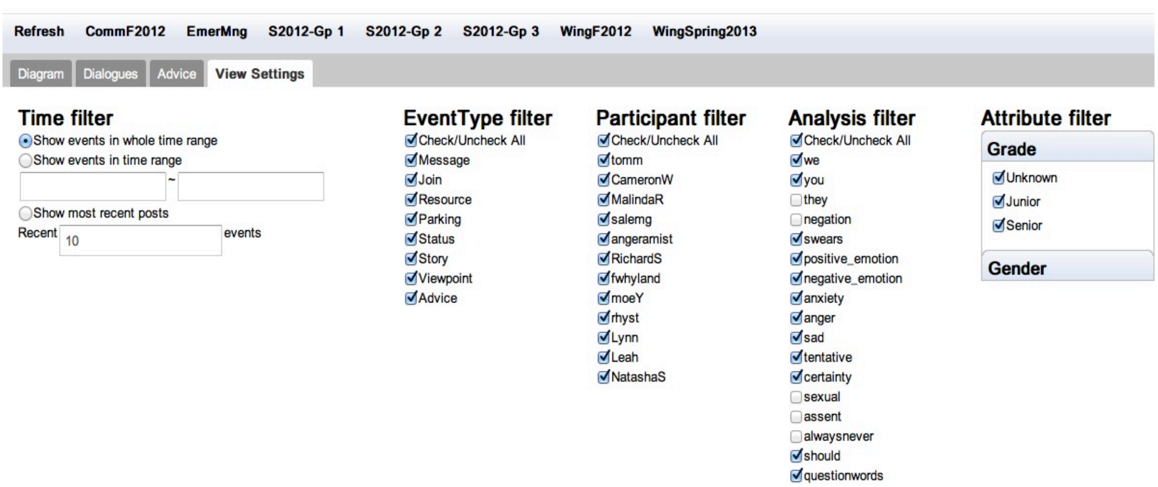


Figure 12: Dashboard -- Settings panel

Development Plans

We will briefly describe our plans for extending the dashboard. First, we plan to build analyzers using more sophisticated computational methods (such as machine learning and latent semantic analysis) to identify more of the important dialogue properties and behaviors listed above. We also plan to improve the linked representations in the interface, as illustrated in Figure 13 illustrates how clicking on elements of the visualizations in the Diagram or Advice panel will bring up text in the Dialogue pane that is filtered to show a specific person, group, time period, conceptual focus, or analysis result (e.g. all posts showing negative emotion words).

We plan to subject the Facilitators Dashboard to more stringent, if still formative, evaluations under more diverse conditions, and to use survey and interview methods to evaluate its usefulness and impact. (It will be difficult to find a statistically significant number of facilitators and to set up appropriate control conditions to be able to experimentally compare facilitation with and without the Dashboard, so more qualitative methods will be used.) We also plan to further evaluate how the Dashboard, or some of its components, would be useful to participants to visualize important dialogue attributes.

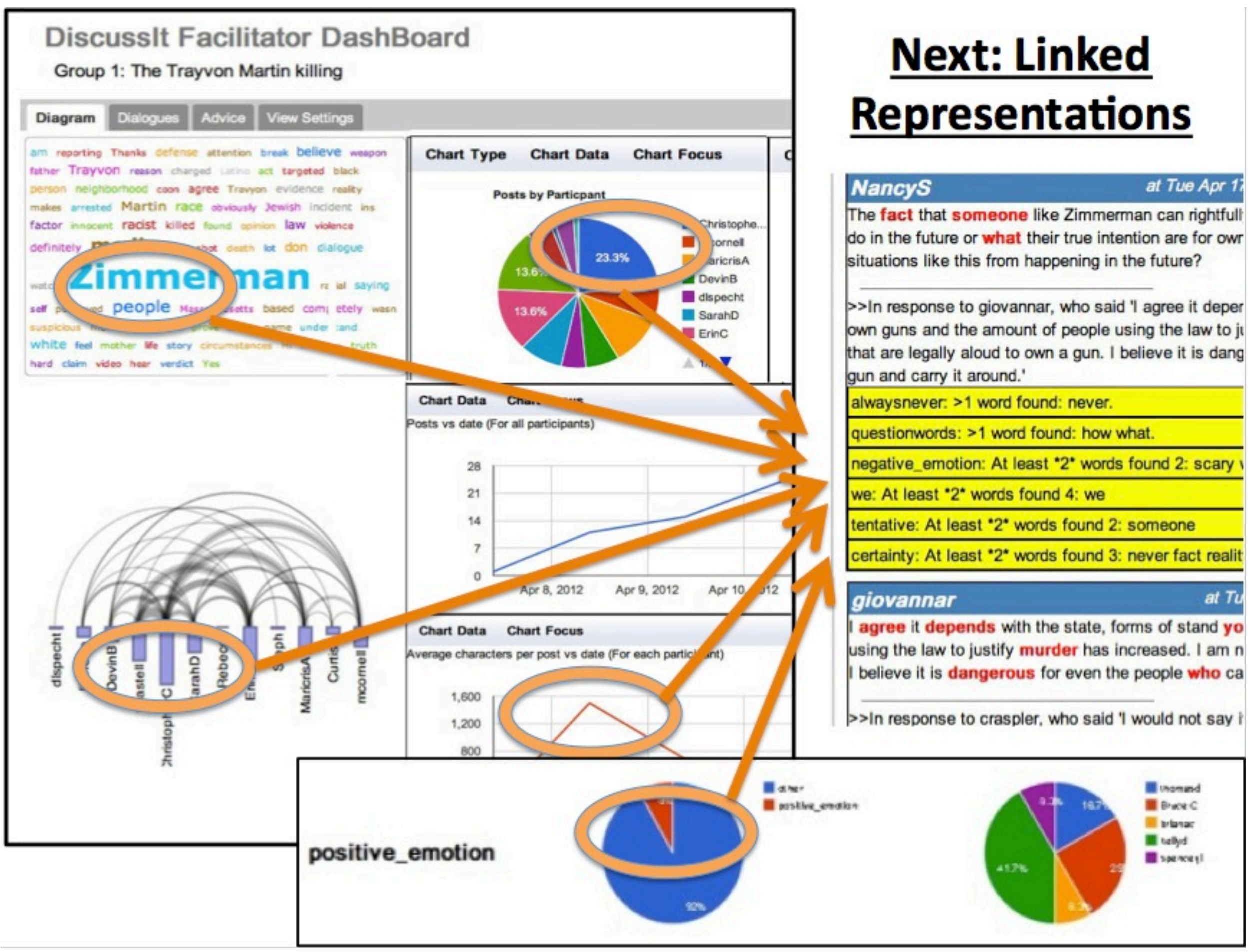


Figure 13: Dashboard -- Linked Representations (planned)

Dialogue Text Classification Research

Our overall research goals are to better understand, assess, and support SD-skills in online contexts. A prerequisite to researching how to support SD-skills is being able to measure, identify or assess them. Most of the work described above was done with hand-coding of dialogue text. Below we describe our ongoing attempts to assess SD-skills using computational linguistic models. As part of our work investigating online support of SD-skills we have developed a hand-coding scheme for categorizing segments of online text, described above, which has been used to evaluate software features in college classes, with encouraging results. In parallel we are using text classification tools and machine learning to develop automated methods to categorize text to ascertain SD-skills and related indicators of deliberative dialogue quality, which we report on here. We are using automated assessment of SD-skills for two purposes: (1) to assess skill differences and correlations in our evaluative research, and (2) to display facets of social deliberative skill in a Facilitators Dashboard (see above) that gives facilitators and teachers a birds-eye view of important deliberative properties of an online conversation.

Text analysis: Background and Related Work

Automatic text analysis (including a wide variety of computational methods: supervised learning, latent semantic analysis, topic modeling, etc.) has been used successfully for a wide variety of purposes in educational contexts, including to: grade essays], analyze content for conceptual understanding, discover topics or themes, score text sophistication, writing quality, and reading grade level, detect off-topic behavior, assess learning styles , and score argumentative and question-answering quality. As far as we know, we are the only ones researching text analysis to assess social deliberative skills such as perspective taking and meta-dialogue in educational contexts or in human dialogues of any sort. There has been related work in non-educational and non-dialogical contexts to identify psycho-linguistic and socio-linguistic phenomena such as emotional states and sentiment, personality traits; and even to predict health improvement based on essay writing. Text analysis methods have been used to classify speech acts (including dialogue moves, tutorial acts, argument moves, etc.).

Text analysis has been used successfully for a wide variety of purposes, including to: grade essays (Shermis & Burstein 2003; Dikli 2006), analyze content for conceptual understanding (Lintean et al., 2011; Azadevo see rose;), discover topics or themes (ref), score text sophistication, writing quality, and reading grade level (McNamara et al., 2010;), detect off-topic behavior (Ceninas…), and score deliberative, argumentative, and question-answering quality (Rose et al. 2008; Kim et al., 2006; Ravi & Kim 2007; Yoo et al., 2012; Rus et al., 2007). It has been used to identify: dialogic moves and patterns (ref), tutorial behaviors (D'Mello et al., 2007; Graesser et al. 2007); agreement/disagreement (ref), deception (Warkentin et al 2011;), trustworthiness (Feng Kim, 2006;), participant roles; influence (including control and involvement, Strazalikowski et al 2010), text style or genre (Graesser.. ref). It has been used to identify psycho-linguistic and socio-linguistic phenomena such as emotional states and sentiment (D'Mello & Graesser, 2011; CRTD 2009;), personality traits (ref); learning styles (Ozpolat in rose..), and even to predict health improvement based on essay writing (Pennebaker xxx). Finally, and most closely related to our work, text analysis methods have been used to classify speech acts (including dialogue moves, tutorial acts, argument moves, etc.).

Past research exploring linguistic and discourse features in dialogues has proven moderately successful in predicting complex phenomena such as personality type, status, deception behavior, metacognition, speech acts, intention, and affect states (Hancock et al. 2010; Keshtkar et al. 2012; Mairesse, et al. 2007; Campbell & Pennebaker, 2003; D'Mello, Dowell, & Graesser, 2008). Therefore, it is plausible to expect that a linguistic and discourse analysis of deliberation dialogues would provide valuable insights into predictors that are diagnostic of deliberation dynamics and skills. Our research question is whether such methods can be used to predict SD-skills.

Our work uses the output of sophisticated text analysis systems (LIWC and Coh-Metrix) as feature inputs for machine learning algorithms—described in Appendix A2. LIWC (Linguistic Inquiry Word Count) is a well researched but "shallow" dictionary-matching text categorization system yielding about 80 linguistic categories (e.g. positive emotion words, pronouns, and causation words). Coh-Metrix performs a series of deep-processing analysis (including semantic cohesion, latent semantic analysis, and reading complexity level) yielding about 100 metrics. The Appendix contains a short description of each system. See descriptions of these systems in Appendix A2.

A simplistic comparison of these systems is that LIWC categorizes speech acts based on *what* participants are saying, and Coh-Metrix produces measurements related to *how* participants are speaking. LIWC features are derived across topic domains and from people from all walks of life; Coh-Metrix features are generated across text genres from a wide spectrum of disciplines. Though LIWC's dictionary-matching method is simple (like keyword-matching), hundreds of studies have been done using it (and contributed to its development) so many of the categories it uses are well researched in terms of how use of these linguistic categories correlate with important psychological or social phenomena. LIWC and Coh-Metrix measurements are ideal for this study, where the discourse data comes from participants across a variety of topic domains and online contexts. Both LIWC and Coh-Metrix features have been shown to be valid and reliable markers of a variety of psycholinguistic phenomena.

Early Research

In early studies we used text analysis in conjunction with multi-class machine learning methods to build models for individual deliberative skills. This proved to be challenging for the methods available to us at the time, and we shifted to the more tractable task of building models for a total or composite deliberative skill measure that was the aggregate of the individual sub-skills (later to return to individual skill modeling). A series of experiments, reported in several papers, refined our ability to automatically assess deliberative skill across multiple domains of online engagement. These experiments were conducted with a data corpus consisting of online interactions from three domains. Participant posts were first partitioned into segments if the type of speech act changed within a post (usually there were 1-4 segments per post). The domains were: an online civic engagement dialog (32 participants with 396 segments of text), two faculty communities engaged in logistical decision making (16 participants and 438 text segments), and, the largest set, college classroom online discussions of controversial topics (90 participants and 1783 text segments). Training was done based on human-rated assessment of deliberative skill, using a coding scheme that had shown inter-rater Cohen's Kappa statistics of 71% on average across the domains (average percent agreement of 76%), which is quite good for a scheme of its complexity.[6] Ten-fold cross validation over the data set was used in all cases.

Early work compared various machine learning methods including Naïve Bayes, Support Vector Machine, Topic Modeling, and Regularize Logistic Regression methods (experimenting

[6] Our coding scheme has 42 categories, 17 of which indicate deliberative skills.

with a number of parameters within each). We found L1 Regularized Logistic Regression to be the preferred model (though we continued to include comparison with other models though subsequent experiments to validate this finding). Next we compared the success of various feature sets including bag-of-words, LIWC, Coh-Metrix, and combinations of these. We found that using text analysis (LIWC or Coh-Metrix) outperformed bag-of-words methods, that LIWC features usually outperformed the Coh-Metrix features, and that combining these feature sets lead to worse performance that using them individually. Finally, we did cross-domain studies showing that superior models resulted from using certain domains as the training set. Specifically, the model developed using the faculty community showed better performance on all three domains than either drawing training data from the entire corpus or drawing the training sample from the domain to be tested. It appears that this is because the faculty domain had the most balanced (least skewed) data, i.e. there was a sufficiently large percentage of text segments tagged as deliberative skills vs. others (about half).

We continue our research in the study reported here by: (1) applying methods developed previously to a new set of classroom online dialogue data and (2) adding demographic information, gender and grade level, to the models feature set. In this study we extend out prior research on building machine learning models to predict an aggregate (total) social deliberative skill measure.

SD-skill correlations with LIWC/CohMetrix measures

A summary of the correlation analysis of CohMetrix and LIWC measurements vs. coding scores is given in Table 7. This table includes only those items for which the p value is < .00005 and the absolute value of the correlation coefficient (r) > 0.1. Because we are looking for significant effects within a space of about 200 features (LIWC and CohMetrix) the usual p value criteria of p <.01 needs to be modified to (.01/200=) p < .00005. (Note: all four domains were used for this analysis.)

Code		R2	Explanation
LIWC measures			
_INTERSUB	Qmark	1.2%	
AGREE	Assent	8.4%	
AGREE	I	1.0%	
APOLOGY	Negemo	1.7%	negative emotion
APPREC	Affect	2.3%	
APPREC	Posemo	2.9%	positive emotion
DISAGREE	Negate	1.4%	
DISAGREE	Negemo	1.2%	
Q_RHETOR	Conj	1.2%	Conjunctions
Q_TOPIC	QMark	2.3%	
UNCER	Anx	2.9%	Anxious words
UNCER	Tentat	1.4%	Tentative words
DI_ActPropose	Numerals	1.4%	
ActDecline	Cause	1.2%	casual words
SOURCE_REF	AllPct	5.3%	all punctuation
CohMetrix measures			
_INTERSUB	DENPRPi	1.0%	Personal pronouns
_INTERSUB	DENSPR2	1.2%	Ratio of pronouns to noun phrases
_INTERSUB	PRO2i	4.4%	Second person pronoun

AGREE	INTEi	1.2%	Incidence of intentional actions, events, and particles.
APPREC	CONTEMPEXi	1.4%	Temporal Connectives
DI_ActPropose	DATTIMi	1.0%	Date/Time
SELF_REFL	PRO1i	1.2%	1st person pronoun
SOURCE_REF	ZRef	1.0%	??
SOURCE_REF	READFKGL	2.6%	Flesch-Kincaid Grade level
SOURCE_REF	READASW	4.8%	Average syllables per word

Table 7: Correlations between Codes and LIWC/CohMetrix measures

There were a number of LIWC and CohMetrix measures that showed slight but highly significant correlations to SD-skill codes (R2 2-10%, $p < .00005$). The following correlations conform to expectations based on the definition of the codes: Qmark and pronoun occurrence (DENPRPi, DENSPR2, PRO2i) with INTERSUB; Assent with AGREE; Posemo with APPRECiation; Negate and Negemo with DISAGREE; Qmark with QTOPIC; Anxious and Tentative with UNCERtaint; SELF-REFLection with 1st person pronouns. Other correlations in the table are more difficult to imagine explanations for. The modest size of the R squared values indicates that some LIWC and CohMetrix values should be useful features in models that predict the corresponding SD-skills, but that by themselves they have little predictive value.

As part of the L1 machine learning analysis described later, we produced values indicative of the correlations between the *Total*-SD-skill composite metric and the LIWC and CohMetrix measures. The results are summarized below in terms of trends:

Total-SD-Skill Positive Correlations	**Total-SD-Skill Negative Correlations**
More negative addictive connectives	Less narrativity
	Less lexical diversity
More negations	Lower reading ease
More pronouns	Less connectives
More second person pronouns	Less concrete words
More meaningful words	Lower number of words
More punctuation	Less spatial location & motion words (i.e., upon, into)
More verbs	
Older age of acquisition for words	Less causal verb and particles
	Less nouns

Table 8. Correlations between Total-SD-skill and LIWC/CohMetrix measures

We would expect that SD-Skill, by the definition of the component scores, would correlate with pronoun use (i.e. intersubjectivity). Several of the correlations align with intuitions that those with higher SD-skills would in general be more sophisticated or mature, including: word acquisition age (i.e. grade level), more meaningful words (i.e. more concrete and less vague words), lower reading ease (i.e. more complex sentence structure), and more punctuation. Some correlations are counter to our current intuitions, including: less lexical diversity, less connectives, and lower number of words.

Automated Text Analysis: Method

Data set. Twenty six students in a college Alternative Dispute Mediation class discussed two topics (the Trayvon Martin killing in Florida and Gun Control, one each week over two weeks) in using the Mediem deep dialogue discussion software. Students were randomly broken into three discussion groups of 8-9 members each, with all groups discussing these topics. There were 8 males and 14 females ranging in undergraduate grade level from sophomores to seniors, with one non-degree student. Each of the three groups used a different set of software features based on our protocol for an experimental study of the effects of tools to support social deliberative skills. In Murray et. al. [6] we discuss our findings that "reflective tools" showed a significant effect size in deliberative skills as measured by human coding, but for this paper we ignore the grouping of students as we are only interested in trying to model the human rating of total deliberative skill using computational methods. The data set consisting of 829 text segments from 369 posts. 43% of the segments were coded under the "deliberate skill" meta-category (vs. 57% "other").

Machine learning method. In this study, we used our highest performing machine learning method, L_1 regularized logistic regression (L_1RLR) to model social deliberative behavior and predict its occurrences. L_1RLR is also preferred in this research because it not only works well with high dimension feature space and small data sets, but also is able to automatically select features and learn an easy-to-interpret (transparent) model. Being able to automatically select features mitigates the problem that little precedent research exists in this new area that is suggestive of features predictive of social deliberative behavior. Being able to yield an interpretable model presents fewer challenges for researchers in social science and communication science to understand the efficacy of a computational model for social deliberative behavior.

Before we describe L_1RLR, let us recall that the logistic loss function is defined as:

$$p(y|x; \mathrm{W}) = \frac{1}{1+\exp(-\mathbf{W}^T x)}$$

where x is the training data, y is the response variable, and W is the model we learn.

In regularized logistic regression, we solve the following optimization problem:

$$\underset{\mathbf{W}}{\operatorname{argmax}} \sum_i \log(p(y_i|x_i; \mathbf{W}) - \lambda * \Omega(\mathbf{W})$$

where $\Omega(\mathrm{W})$ is a regularization term used to penalize large weights.

In the case of L_1 regularized logistic regression, L_1 norm [27], or least absolute shrinkage and selection operator (Lasso) is used to induce the penalty. Previous research [28] has shown that L_1 regularization logistic regression requires the number of training examples that grows logarithmically with the number of features to learn well, which favors this study.

In our experiments, we used the l_1 regularized dual averaging algorithm [29] for solving l_1RLR. We trained l_1RLR (i.e., λ=1, γ=2) with various feature sets and carried out 10-fold stratified cross-validation.

Results and Discussion

We performed a set of experiments by exploring the effectiveness of different types of features on predictive accuracy, precision, recall, and F_2 measure (the harmonic mean of precision and recall that weights recall twice as high as precision). In Table 8, we report the average performance across cross-validation runs.

	LIWC features	CohMetrix features	LIWC+gender+ gradLevel
Accuracy	61.41	60.68	60.81

Precision	54.30	54.31	53.78
Recall	68.52	57.94	67.41
F_2 measure	65.11	57.18	64.16

Table 8: Predictive performance (in %) of L1 regularized logistic regression built using different type of features

Predictive performance and feature comparisons. As can be seen in Table 8, with computational models, we are able to predict social deliberative behavior with up to 61% accuracy, 54% precision, 68% recall, and 65% F_2 measure. LIWC features outperformed Coh-Metrix features by a slight margin overall, which confirms earlier findings (we did not model using combined LIWC and Coh-Metrix features as prior work suggested this would not help). Surprisingly, adding the demographic information of gender and grade level as machine learning inputs did not improve performance (it degraded it slightly).[7] This suggests that variations due to grade and gender are already encoded in the text analysis features (of both LIWC and Coh-Metrix)—a hypothesis we will pursue in further research.

The performance of the L1-RLR on this data set outperformed the models reported en earlier studies of classroom data. In general, prior studies of multi-domains showed that prediction in the classroom domain was worse than in the other domains (civic engagement and faculty logistical decision-making). More specifically, the results reported here improved over previous results of classroom domains by 8% on precision and 64% on recall. We believe that this is mostly due to the newer data set having less data skew (43% deliberative skill on this set vs. 32% on the prior classroom data set). We are looking into methods to compensate for data skew, including training our models on the most robust data sets as opposed to the testing data sets [26].

In a larger sense, the results suggest that it may be feasible to train machine learning models to automatically analyze conversations in online communication to identify high-order communication skills such as social deliberative behavior.

Parameters in the learned model. As mentioned, one of the benefits of using L1-RLR is that the relative importance or weights of each feature can be inspected (they are related to the coefficients of the regression model). The L_1 regularized logistic regression learned a model with 27 features in this domain. In other words, 55 out of the 82 LIWC features were shrunk by L_1RLR (which automatically prunes features, another advantage vs. other machine learning methods). In Table 9, we show the top 10 most salient features of the learned model. The rest of the 17 features have absolute feature weights less than 0.01 and are commented below.[8]

LIWC feature	Interpretation	Weight
assent	assent	0.335
WC	word counts	0.223
social	social processes	-0.051
Dic	dictionary words	-0.045
i	1^{st} pers singular	0.028

[7] Indeed, when examining the learned feature space, we found that both gender and grade level features were shrunk by the L_1RLR model. In other words, both features were assigned zero weights in the final model.

[8] Note, the absolute value of the weights is meaningless and dependent on tuning parameters of the algorithm, and in general are not comparable from one model to the next. Only the relative sizes of the weights within a model are meaningful.

funt	total function words	-0.024
posemo	positive emotion	0.023
AllPct	total punctuations	0.023
affect	affective processes	0.023
period	punctuation	0.022

Table 9: Top 10 LIWC features learned by L1 regularized logistic regression

Next we summarize the characteristics of social deliberative behavior in the language of LIWC features. LIWC was not designed to measure deliberative skill or any sort of dialogue-quality related speech act categories, and predictive relationships between its categories and deliberative skill are expected to be secondary (i.e. resulting from more clearly relevant intermediate factors). Compared to "other speech acts", social deliberative behavior has: more assent words, longer messages, more 1^{st} person pronouns, more positive emotions, more total punctuations, more affective processes, more certain words, more pronouns (i.e., personal pronouns and impersonal pronouns), more cognitive process, more auxiliary verbs, fewer social processes, fewer dictionary words, fewer functional words, fewer relative words, fewer words per sentence, fewer prepositions, fewer big words, fewer dashes, fewer words about time, fewer commas, fewer space words, fewer present tense, and fewer articles.

Assent-words (31 word stems including absolutely, agree, alright*, haha*, ok, yes, yup…) and the segment word count (WC) were by far the largest factors in this model. Pennebaker & King say the following about assent and word count. Higher word count is related to better group performance. Lots of assents and questions stimulate better team performance. "Later in a group task, assents may signal consensus, early assents may indicate blind agreement by unmotivated group members" and "in a cooperative coordination context, higher total word count may signal better communication and agreement, whereas in a negotiation context it may signal a breakdown in agreement." (p. 35). Our related analysis of the faculty dialog also showed that word count was highly related to human assessment of deliberative skill, but, curiously assent was not so related. Further work is ongoing to determining the domain-dependent aspects of deliberative behaviors.

Collaborators

Below we list the organizations who helped us in this research.

Idealogue Inc.

Idealogue offers an online dialogue environment called Mediem that is uniquely tailored to support deeper dialogues. They have allowed us to use the software for free for a number of research studies, and have worked with our staff in brainstorming ideas to support quality dialogue. We also paid them a small fee to customize the software to make it "research-ready"--so that it can output the dialogue information to our text analysis tools in real time. Our team and research has also helped them think about product improvements and new project opportunities. We have created a Facilitator Dashboard software that draws data from Mediem in real time through an API developed jointly.

Mass Office of Public Collaboration

Mass Office of Public Collaboration, MOPC, has been working with us to identify civic engagement projects in the state that could serve as test beds for our research. They would benefit from having a research component to one of their projects. They are particularly interested in the Dashboard tool See http://www.umb.edu/modr/.

DeMars & Associates, Ltd.
DeMars & Associates, Ltd. works with eBay and PayPal to offer online dispute resolution services. They are interested in our research and whether it can improve online mediation. They have brainstormed with us on software features that might improve the quality of online dispute mediation. They have also given us sample data from online mediations (cleaned up to be anonymous before research use). One of our publications is based on their data.

Juripax BV
Juripax BV is a small business in The Netherlands that does online dispute resolution in areas such as workplace disputes and divorce mediation. They have been helping us brainstorm the design our dialogue support tools, have given us sample data to analyze, and are looking for application/test contexts for our systems. We have coded their data and it is still being analyzed.

Franklin Pierce Collage's New England Center for Civic Life
We are working with Zan Goncalves, a faculty member at Franklin Pierce Collage and an associate at the college's New England Center for Civic Life (www.franklinpierce.edu/institutes/neccl). Goncalves teaches writing classes specifically aimed to improve deliberative skills. She consulted with us on theory and practice and we prototyped some of our methods in her summer class in the first year of the grant.

Interactivity Foundation.
The Interactivity Foundation works to engage citizens in the exploration and development of possibilities for public policy through small group discussions.(www.interactivityfoundation.org). We are in communication with them to determine how their F2F engagement and deliberation processes could be translated to online tools.

References

For publications by our group members (Murray, Woolf, Xu, etc.), see Appendix A5.

Aleven, V., McLaren, B., Roll, I., Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. Int. J. of Artificial Intelligence in Education, 16 (2006) 101-128.

Altman DG (1991) Practical statistics for medical research. London: Chapman and Hall.

Andriessen, J., Baker, M., & Suthers, D. D. (Eds.). (2003). Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments. Dordrecht: Kluwer.

Azevedo, R., Guthrie, J.T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. Journal of Educational Computing Research, 30 (1), 87-111.

Baker, M. J., & Lund, K. (1997). Promoting reflective interactions in a computer-supported collaborative learning environment. Journal of Computer Assisted Learning, 13, 175–193.

Baker, M., Andriessen, J., Lund, K., Amelsvoort, M. v., & Quignard, M. (2007). Rainbow: A framework for anaysing computer-mediated pedagogical debates. International Journal of Computer Supported Collaborative Learning, 2(2-3), 315-357.

Batson, C. D. (2009). Two Forms of Perspective Taking: Imagining How Another Feels and Imagining How You Would Feel. In K. Markman, W. Klein & J. Suhr (Eds.), Handbook of imagination and mental simulation (pp. 267). New York: Psychology Press.

Black, L., Welser, H., Cosley, D., and DeGroot, J., Self (2011). Governance Through Group Discussion in Wikipedia Measuring Deliberation in Online Groups. Small Group Research 42(5) pp. 595-634.

Block-Lerner, J., Adair, C., Plumb, J. C., Rhatigan, D. L., & Orsillo, S. M. (2007). The case for mindfulness-based approaches in the cultivation of empathy: Does nonjudgmental, present-moment awareness increase capacity for perspective-taking and empathic concern? Journal of Marital & Family Therapy, 33, 501-516.

Boland, R. J., & Tenkasi, R. V. (1995). Perspective making and perspective taking in communities of knowing. Organization science, 6(4), 350-372

Bromme, R., Kienhues, D., & Stahl, E. (2008). Knowledge and epistemological beliefs: An intimate but complicate relationship. In Knowing, knowledge and beliefs (pp. 423-441). Springer Netherlands.

Campbell, R. S., and Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. Psychological Science, 14, 60-65.

Carletta, Jean (1996). Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2):249–254.

Collins, H. (2010). Tacit and explicit knowledge. University of Chicago Press.

Conati, C. & Vanlehn (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. Int. J. of Artificial Intelligence in Education, 11.

Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)

Cross, D. R. & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. Journal of Educational Psychology, 80(2),131-142.

D'Mello, S. K., Dowell, N., Graesser, A. C. (2008). Cohesion Relations in Tutorial Dialogue as Predictors of Affective States. Paper Presentation at the 14th International Conference on Artificial Intelligence in Education. Brighton.

Dawson, T. L., & Stein, Z. (in press). We are all learning here: Cycles of research and application in adult development. In C. Hoare (Ed.), The oxford handbook of reciprocal adult development and learning. New York: Oxford.

Desiato, C. (unpublished, 2012). Online Political Deliberation and Diversity: Users' Communicational Strategies to Facilitate Consensus Maintaining Heterogeneity. Dissertation proposal, Communication and Information Sciences University of Hawai'i at Manoa.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*(3), 231-264.

Duell, O. K., & Schommer-Aikins, M. (2001). Measures of people's beliefs about knowledge and learning. Educational psychology review, 13(4), 419-449.

E. Michael Nussbaum & Denise L. Winsor & Yvette M. Aqui & Anne M. Poliquin (2007). Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. Computer-Supported Collaborative Learning (2007) 2:479–500.

Edelson, D.C., D.N. Gordin, and P.D. Pea (1999). "Addressing the Challenges of Inquiry Based Learning Through Technology and Curriculum Design." The Journal of Learning Sciences. 8(3&4): 391-450. 1999.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. Motivation and Emotion, 14(2), 131-149.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. The California Academic Press.

Ford, C. L., & Yore, L. D. (2012). Toward convergence of critical thinking, metacognition, and reflection: Illustrations from natural and social sciences, teacher education, and classroom practice. In Metacognition in science education (pp. 251-271). Springer Netherlands.

Freund, Y., Schapire, R. (1996). Experiments with a new boosting algorithm. In: Saitta, L. (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, San Francisco, 30 CA.

Fuhs, C. (2010). An Integral Map of Perspective-Taking. Integral Theory in Action: Applied, Theoretical, and Constructive Perspectives on the AQAL Model, 273.

Fuhs, C. (2013). Researching Perspectival Growth in Organizational Leaders. Integral Theory and Practice, 8(3&4), 1-18.

Gastil, J., & Black, L. W. (2008). Public deliberation as the organizing principle in political communication research. Journal of Public Deliberation, Vol.

Goodman, B., Linton, F., Gaimari, R., Hitzeman, J., Ross, H., & Zarrella, J. (2005). Using dialogue features to predict trouble during collaborative learning. Journal of User Modeling and User Adapted Interaction, 15(102), 85–134.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science, 3, 371–398.

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. Information Design Journal, 15, 199–213.

Graesser, A., & McNamara, D. (2010). Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science 3(2), 371–398. 2010.

Graesser, A., & Person, N. (1994). Question asking during tutoring. American Educational Research Journal, 31, 104-137.

Graesser, A., McNamara, D. & VanLehn, K. (2005). Scaffolding Deep Comprehension Strategies Through Point&Query, AutoTutor, and iSTART. Educational Psychologist, 40(4), 225–234.

Graesser, A., Vasile, R., & Zhiqiang, C. (2008). "Question classification schemes." Proc. of the Workshop on Question Generation.

Graesser, A.C., Dowell, N., & Moldovan, C. (2011). A computer's understanding of literature. Scientific Study of Literature 1:2 (2011).

Graesser, Art, Vasile Rus, and Zhiqiang Cai (2008). "Question classification schemes." Proc. of the Workshop on Question Generation.

Haller, E. P., Child, D. A., & Walberg, H. J. (1988). Can comprehension be taught? A quantitative synthesis of metacognitive studies. Educational Researcher, 17(9), 5-8.

Halpern, D.F., Millis, K., Graesser, A.C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. Thinking Skills and Creativity, 7, 93-100.

Hancock, J., Beaver, D., Chung, C., Frazee, J., Pennebaker, J., Graesser, A., Cai, Z.: Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. Behavioral Sciences of Terrorism and Political Aggression 2(2), 108–132 (2010)

Hennessey, M. G. (1999). Probing the dimensions of metacognition: Implications for conceptual change teaching-learning. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston, MA.

Herzig, M. & Chasin, L. (2006). Fostering dialogue across divides: A nuts and bolts guide from the Public Conversations Project. Watertown, MA: The Public Conversations Project. Available at www.publicconversations.org

Holman, P., Devane, T. Cady, S. (2007). The Change Handbook: The Definitive Resource on Today's Best Methods for Engaging Whole Systems. San Francisco: Berrett-Koehler Publ.

Jarvela, S., & Hakkinen, P. (2000). Levels of Web-Based Discussion: Theory of Perspective-Taking as a Tool for Analyzing Interaction. In B. Fishman & S. O'Connor-Divelbiss (Eds.), Fourth International Conference of the Learning Sciences (pp. 22-26). Mahwah, NJ: Erlbaum.

Johnson, D. W., & Johnson, R. T. (1999). Learning together and alone: cooperative, competitive, and individualistic learning (5th ed.). Boston: Allyn & Bacon.

Kaplan, J. M. (2005). Rational Decision Making: Descriptive, Prescriptive, or Explanatory?. A Companion to Rationalism, 425-449.

Keshtkar, F., Burkett, C., Li, H., & Graesser, A.C. (2012). Detecting players personality behavior with any effort of concealment. In A. Gelbukh (Ed.), Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (pp. 503-515). Heidelberg, Germany: Springer.

King, P.M. and Kitchener, K.S. (1994). Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults. San Francisco: Jossey-Bass.

Klein, M. (2010). Using Metrics to Enable Large-Scale Deliberation. Collective Intelligence In Organizations: A Workshop of the ACM Group 2010 Conference. Sanibel Island, Florida, USA.

Knight, S., Buckingham Shum, S., & Littleton, K. (2013, April). Epistemology, pedagogy, assessment and learning analytics. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (pp. 75-84). ACM.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. Computers in human behavior, 19(3), 335-353.

Kuhn, D. (1999). A Developmental Model of Critical Thinking. Educational Researcher, March, 1999, pp. 16-25.

Kuhn, D. (2000). Metacognitive Development. Current Directions in Psychological Science, 9(5), pp. 178-181.

Lin, X. & Sullivan, F. (2008). Computer contexts for supporting metacognitive learning. In J. Voogt, G. Knezek (eds.) International Handbook of Information Technology in Primary and Secondary Education, 281–298. Springer Science+Business Media, LLC.

Linn, M. C. (2000). Designing the knowledge integration environment. International Journal of Science Education, 22(8), 781-796.

Lintean, M., Rus, V., & Azevedo, R. (2011). Automatic Detection of Student Mental Models Based on Natural Language Student Input During Metacognitive Skill Training. International Journal of Artificial Intelligence in Education, 21(3), 169-190.

Lipman, M. (1991) *Thinking in Education*, Cambridge: Cambridge University Press.

Louca, L., Elby, A., Hammer, D., & Kagey, T. (2004). Epistemological resources: Applying a new epistemological framework to science instruction. Educational Psychologist, 39(1), 57-68.

Luxton-Reilly, A., Bertinshaw, D., Denny, P., Plimmer, B., and Sheehan, R. (2012) The impact of question generation activities on performance. In Proceedings of the 43rd ACM Technical Symposium on Computer Science Education (Raleigh, North Carolina, USA, February 29 - March 03, 2012). SIGCSE '12. ACM, New York, NY, 391-396.

Lyon, T. D., & Flavell, J. H. (1994). Young Children's Understanding of “Remember” and “Forget”. *Child Development*, *65*(5), 1357-1371.

Mairesse, F., Walker, M., Mehl, M., Moore, R.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30(1), 457–500 (2007)

Manevitz, L.M & , Yousef, M. (2002). One-class svms for document classification. Journal of Machine Learning Research archive 2(3 ,139-154.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written Communication, 27(1), 57-86.

Melville, Keith, Willingham, Taylor L., Dedrick, John R. (2005). National Issues Forums. A network of communities promoting public deliberation. In John Gastil and Peter Levine (Eds.), The deliberative democracy handbook. Strategies for effective civic engagement in the 21th century (pp. 37-58). San Francisco: Jossey-Bass.

Mischel, Walter. "Metacognition at the hyphen of social-cognitive psychology (1998)." Personality and Social Psychology Review 2.2 : 84-86.

Neuberg, S. L., Kenrick, D. T., & Schaller, M. (2010). Evolutionary social psychology. Handbook of social psychology.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. L., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: www.LIWC.net.

Puntambekar, S. & du Boulay, B. (1997). Design and development of MIST—A system t help students develop metacognition. J. Educational Computing Research, 16(1), 1-35.

Ravi, S., Kim, J. (2007). *Profiling student interactions in threaded discussions with speech act classifiers*. IOS Press.

Reiser, B.J, Tabak, I. & Sandoval, W.A. (2001). BGuILE: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. In S.M. Carver & D. Klahr (Eds.) Cognition and Instruction: Twenty-five years of progress. Mahway, NJ: Erlbaum.

Rish, I.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol. 3, pp. 41–46 (2001)

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning. Volume 3, Issue 3, September 2008.

Rosenberg, S. (2004). Reconstructing the concept of deliberative democracy. University of California, Irvine: Centre for the Study of Democracy. Retrieved December 22, 2008, from http://repositories.cdlib.org/cgi/view-content.cgi?article=1031&context=csd.

Rus, V., Cai, Z., & Graesser, A. C. (2007, April). Evaluation in natural language generation: The question generation task. In Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation (pp. 20-21).

Salonen, P., Vauras, M., & Efklides, A. (2005). Social Interaction--What Can It Tell Us about Metacognition and Coregulation in Learning?. European Psychologist, 10(3), 199.

Scardamalia, M. (2003). Knowledge-building environments: Extending the limits of the possible in education and knowledge work. In A. DiStefano, K. E. Rudestam, & R. Silverman (Eds.), Encyclopedia of distributed learning (pp. 269–272). Thousand Oaks, CA: Sage Publications.

Scheuer, D., Loll, F., Pinkwart, N. and McLaren, B. M. (2010). Computer-supported argumentation: A review of the state-of-the-art. International Journal on Computer Supported Collaborative Learning, 5(1), 43-102. Springer.

Schommer-Aikins, M., & Hutter, R. (2002). Epistemological beliefs and thinking about everyday controversial issues. The Journal of Psychology, 136(1), 5–20.

Schommer-Atkins, M. (2002). An evolving theoretical framework for an epistemological belief system. Chapter 6 in B.K. Hofer & P.R. Pinrich (Eds.) Personal Epistemology: The psychology of beliefs about Knowledge and Knowing. Mahwah, NJ: Lawrence Erlbaum Assoc.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*, 111-139.

Seih, Y. T., Chung, C. K., & Pennebaker, J. W. (2011). Experimental manipulations of perspective taking and perspective switching in expressive writing. Cognition and Emotion, 25, 926-937.

Selman, R. L. (1980). The growth of interpersonal understanding: Developmental and clinical analyses. New York: Academic Press.

Shermis, M. & J. Burstein (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum.

Smith, Adam (2006). Cognitive empathy and emotional empathy in human behavior and evolution. The Psychological Record. 56.1 (Winter 2006): p3-18.

Smith, K. & Berg, D. (1997). Paradoxes of Group Life: Understanding conflict, paralysis, and movement in group dynamics. Jossey-Bass: San Francisco.

Smith, K. A., Sheppard, S. D., Johnson, D. W., & Johnson, R. T. (2005). Pedagogies of engagement: Classroom-based practices. Journal of Engineering Education, 94(1), 87-101.

Soller, A.L. (2001). Supporting social interaction in an intelligent collaborative learning system. Int. J. of Artificial Intelligence in Education 12(1) 40-62

Stephens. L., Murray, T., Shrikant, N., Wing, L., Xu, X. & Woolf, B.P (2013 in submission). An Evaluation of Online Forum Features Supporting Social Deliberative Skills: Preliminary Results.

Stillings, N. A., Ramirez, M. A., & Wenk, L. (1999, March). Assessing critical thinking in a student-active science curriculum. In meeting of the National Association of Research on Science Teaching, Boston, MA.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, J., Bates, R., Jurafsku, D., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3), 39– 373.

Stromer-Galley, J. (2007). Measuring Deliberation's Content: A Coding Scheme. Journal of Public Deliberation, 3(1).

Suthers, D. D. (2008). Empirical studies of the value of conceptually explicit notations in collaborative learning. In A. Okada, S. Buckingham Shum & T. Sherborne (Eds.), Knowledge Cartography (pp. 1-23). Cambridge, MA: MIT Press.

Suthers, D. D., Connelly, J., Lesgold, A. M., Paolucci, M., Toth, E. E., Toth, J., Weiner, A. (2001). Representational and advisory guidance for students learning scientific inquiry. In K. D. Forbus & P. J. Feltovich (Eds.), Smart Machines in Education: the Coming Revolution in Educational Technology (pp. 7-35). Cambridge, Massachusetts: AAAI Press, The MIT Press.

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. Science Education, 86(2), 264-286.

Veenman, M. V. J. (2012). Metacognition in Science Education: Definitions, Constituents, and Their Intricate Relation with Cognition. Contemporary Trends and Issues in Science Education Volume 40, 2012, pp 21-36.

West, R.F., Toplak, M.E., Stanovich, K.E. (2008). Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions. Journal of Educational Psychology, 2008, Vol. 100, No. 4, 930–941.

White, B., Shimoda, T., Frederiksen, J. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: computer support for metacognitive development. International J. of Artificial Intelligence in Education Vol. 10, 151-1182.

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning*, *4*(1), 63-85.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. American Educational Research Journal, 46, 1060-1106.

Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. Handbook of self-regulation of learning and performance, 15-32.

Winne, P. H., Nesbit, J. C., Kumar, V., & Hadwin, A. F., Lajoie, S. P., Azevedo, R. A., & Perry, N. E. (2006). Supporting self-regulated learning with gStudy software: The Learning Kit Project. Technology, Instruction, Cognition and Learning, 3(1), 105-113.

Appendix

A1. Coding Scheme

Code with _underdash before them are meta-codes that are categories of other codes. They are also designated with "**" in the Descriptions table.

Codes Overview

Our coding scheme was developed after synthesizing several existing schemes—as described above. The "Core Set" those we consider the most central skills/behaviors for social deliberative capacity. The "Additional Deliberation Quality Indicators" are also considered good things to have. As is noted abve, our research focus differs from many related studies in the Learning Sciences in that most related studies focus on the more cognitive speech acts that we list under Argument Codes, while we were more interested in intersubjective deliberative skills. Thus, in the domains we have codes thus far, all Argument Codes were coded using the meta-code ArgGen, and coders were not trained to differentiate the various types of Argument Codes. (We could of course always go back and recode these segments at the more specific level of detail.) The Action Negotiation category was only used for two domains: Workplace Dispute and Faculty Negotiation. These were the only domains in which participants were discussing real proposed extra-dialogue actions or decisions—all other domains were discussions with no off-line decisions or actions involved. Unlike the other codes, the Action Negotiation codes were developed ad-hoc and do not reflect a careful analysis of the literature.

Other than Argument Codes, which were all coded using the meta-code ArgGen, all other codes were coded at the base code level. However, all later analysis, including the interrater reliability numbers shown above, are done at the meta-code level. For example, the Question_Interlocutor, Reference_Interlocutor, and Perspective_Taking codes were abstracted to the _Intersubjective code. Misc Codes cover other categories of interest.[9]

SD-skill -- CORE Set	**Additional Delib. Quality Indicators**	**MISC CODES**	**ACTION NEGOTIATION**	**ARGUMENT CODES**
SELF_REFLection	_META_TOPIC	Q_TOPIC	(External actions)	_ARGument_GENeric
	WEIGH	OTHERS_THNK	ActRequest	GENERAL_SOLUTN
_INTERSUBjective	SYSTEMs_thinking		ActPropose	EXPER_OBSERV
Q_INTERLocutor		HELP	ActAccept	ARG_OPINION
REF_INTERLocutor	FACT_cite_SouRCe	REQ_HELP	ActDecline	SUPPORT
PERSPECTIVE_taking	SOURCE_REFerence	PROCESS	ActNegot	SUM_MY-argumt
			(Dialogue_Actions)	EXAMPLE
_META_Dialog	CHANGE_mind	AGREE	DI_ActRequest	ELAB
MEDIATE	UNCERtainty	DISAGREE	DI_ActPropose	
META_CONS	APOLOGY		DI_ActAccept	*low-skill:*
META_CONFL		_NEGative-emotion	DI_ActDecline	OPINION_ONLY
META_SUM		NEGEMO_INTerloc	DI_ActNegot	OVER_GEN
META_CHECK		NEGEMO_Topic	(Facilitators only)	FACT_NOSRC
		_OFFTOPIC	WELCOMING	
APPRECiation		TECHnical	PROC_EXPL	
		SOCIAL	MOTIVATE	

Codes Description

Descriptions of each code were kept relatively brief, see the table below (an *asterisk in the Description column notes a Core skill or Dialogue quality code from the table above). ID's are unique Ids for each code. If a code is subsumed under a meta-code, that is listed in the Metacode column. We found it very useful to specify other codes that could possibly be confused with a particular code—in the "1st make sure its not" column. This allowed a prioritizing of codes. In defining categories describing what it is *not* can be extremely useful. For

[9] Note that Negative_Emotion is coded, but we found that Postive_Emotion was usually too ambiguous to code so we do not include it—it may be coded as Appreciation, Agree, Help, etc.

example, Self_Reflection may sometimes include Uncertainty, ChangeMind, or PerspectiveTaking. The later three are more specific, and should be used instead of Self_Reflection. The Order column indicates the placement of the code in the pop-up menu in the Codoole coding tool. The table below is ordered by this number, but for some uses it was ordered by another column. Common or prototypical "Trigger words and phrases" were also included to help define and identify codes.

ID	**CODE name**	**Metacode**	**1st make sure its not**	**Description** (*sdskillis , and other dialogue properties)	**Order**	**Trigger words & phrases**
48	SELF_REFL		Uncer Change Perspective	* Self-reflection: reflecting on (or commenting on) one's own assumptions, values, biases, or emotions (other than UNCER, CHANGE). (If its clearly empathy and perspective taking use Perspective.)	1	When I... I am probably... I did not realize... It really makes me... My assumption was... I admit...
33	APPREC			* Appreciation, gratitude, affirmation of another's idea or situation. - Agreement is not Appreciation.	2	Thank you... I appreciate...
34	APOLOGY			Apology.	3	Sorry... Oops...
35	HELP		Meta_Clarify	Helping, explaining something to another about the Topic (not about one's own ideas --see ELAB) or helping with the process or technology.	4	
99	REQ_HELP		Q_Interl Q_Topic	ask for help from interlocutor or mediator/facilitator.	4.5	
71	AGREE			Express agreement with interlocutor or more than one specific interlocutor.	5	
72	DISAGREE			Express Disagreement with interlocutor. (Does not include disagreement with a general idea by people outside the discussion.)	6	
60	FACT_SRC			* Stating a fact and noting the source in the same post (a “source” is not a personal observation, see below). Facts are specific, not general.	8	
69	SOURCE_REF		Fact_Src	* Mentioning a source, with a pointer or quote (where the source IS publicly available text). It can mention a resource that does not support any specific fact.	9	
46	UNCER		Weigh	* Explicitly express uncertainty, ambiguity, confusion, or ignorance (about the topic, or uncertainty about another interlocutor’s thoughts) (e.g. “I’m not sure what he means” or “I’m not sure if it would cost that	10	I’m not sure I don’t know... I am torn I’m not sure what he means.

				much”). If clearly weighing alternatives use Weigh.		I’m not sure if it would cost that much...
47	CHANGE			* Acknowledges a change of opinion/perspective about topic OR an interlocutor. Including: admission of error "I was wrong"; Yielding: "I am starting to change my view…"	11	I was wrong... I am starting to change my view... Maybe my idea was... you're right... After reading more, I am convinced...
45	Q_TOPIC		Q_Interl Q_Rhet	* Content questions: Wh questions, and other Qs that are about the topic in general rather than the thoughts or feelings of participants. Not rhetorical Qs. (If its trying to see what another participant thinks about a topic, use Q_Interl.)	12	Is it just so … ? Is there … ? Wouldn’t there be … ? How many … ? Is the .. ? What makes … ? If it was … ? Is it … ? How would .. ? Isn’t there … ? Does anyone [in the world]...? What about …? What would happen …? What would be … ? What if …? If the … ?
86	Q_RHETOR			Can be rhetorical question about topic or interlocutor. "if you are old enough to kill for your country why can't you…" "what makes you think that…"	12.5	
81	_NEG	>>		General negative emotion; Includes the following codes:	13	“I was blazing mad”... She is just a mean person... This is what makes me so tired...
31	NEGEMO_INT	_Neg		Negative Emotion (about interlocutors or dialog process) Including: disrespect/insult, anger/rage, threat/hostility (even implicit), dismissiveness/contempt, distrust/suspicion (e.g. doubting that interlocutor has good intentions). - Should not be vague or general dissatisfaction or critique, but more direct expression of negative emotion. - Must include words that are usually interpreted negatively (i.e. code the words not the tone exclusively). -Note: disagreement and aggressively promoting an idea are not necessarily	15	

				negative emotion.		
32	NEGEMO_TP	_Neg		Negative emotion or distrust, about a topic, i.e. not about interlocutors.	16	
82	_INTERSUB	>>		* Intersubjectivity; connecting with ideas of specific (usu. one) interlocutor.	17	
42	Q_INTERL	_Intersub	Meta_check	* Asking Questions to discover more about a single interlocutor's thoughts or feelings (curiosity). Can be an assertive: "give me an example." Must have actual or implied “you”--second person singular. Includes: - Explicitly reflecting back words or ideas to check for understanding. - - If asked to the whole group, see Meta-Check. -If the question is asking to *do* something, see the Action codes . - If the question is clearly rhetorical, code as Q_Rhetorical. - Don’t use this code for the mediator/facilitator, see Mediate codes]	18	Are you worried … ? You’re saying that …? To what extent do you believe …? Don’t you consider …? What makes you think … ? How do you ...? Do you mean …? If I heard correctly you said... Let me know if that's not what you meant...
44	REF_INTRL	_Intersub	Q_Interl tech, any _meta_d	* Referencing what another said, including quoting, summarizing, mentioning it. -If it references what *everyone* is saying, use Meta_Check.	20	As has been stated by X… As X suggested … A a previous post mentioned … X’s point... To respond to X... To expand on [X’s idea]… Continuing the idea…. What X said about ….
58	PERSPECTIVE	_Intersub	any _meta_d	* Social Perspective taking -- putting yourself in another's shoes (of an interlocutor OR a group you are not a member of). The tone is more one of empathy, not critique (though critique may occur in another post/segment).	21	I understand how that might feel... I can see why they might think/do that... If I were in this situation... I hear you... I understand where you’re coming from... You experience this as...
49	OTHERS_THNK	_Intersub	Perspective	* Assessing or reflecting on the ideas, assumptions, values, biases of others outside the discussion - or if its about someone in the discussion as a representative of a group (e.g. you Catholics always..."	22	They may still … The vast majority of people … Conservatives think... They deny...

				- If its clearly empathy and perspective taking use Perspective.		
83	_META_D	>>		* Birds eye view of the discussion (see sub-categories)	23	
50	MEDIATE	_Meta_D	Q_Interloc Meta_Cons, Meta_Conf, Meta_Sum Meta_Check Welcome DL_Act*	* -Make a *suggestion* about how interlocutors should communicate or how the conversation should proceed; redirect conversation toward *higher quality*; e.g. - Attempting to *clarify* what another party has said, for a *third* party (not just for oneself). - Suggesting that interlocutors take each others *perspective*. - Does not have to be "skilled" mediation; i.e. can be a poor attempt.	24	Maybe if we took small steps... We should really be looking into... Cathy is indicating that... She needs... Why don't we focus on…
51	META_CONS(ensus)	_Meta_D		* Highlight agreement (consensus) on some point - for entire group or part of group (not just self and self and one other);	25	I think we are all in agreement...
52	META_CONF(lict)	_Meta_D		* Highlight disagreement on some point. for entire group or part of group (not just self and self and one other)	26	I don't think the two of them agree... We don't have consensus here...
53	META_SUM	_Meta_D	meta_cons, meta_sum	* Summarizing the conversation -may include both consensus and conflict	27	There are many of us who think... Many people have suggested... The discussion focused on... Our common themes are...
56	META_CHECK	_Meta_D	Elab, tech	* "How are we doing" questions or reflections about the whole group or context. Usually has “we” stated or implied (we as this group, not we as in “my people”). - Includes: to note or ask about the purpose or context of the discussion; - (Re)framing the issue. (not Framing one's own ideas - see ELAB) - If used for a mediator/facilitator, the reflection may be on you-all participants, rather than "all of us."	28	Does anyone else agree …? What do people think …? Who said anything regarding …? Does anyone … ? No one has given any reasons for... The root of this question... We need to remember...
38	PROCESS	_Meta_D	Use Proc_Expl for facilitators	* Process talk, questions or discussion about the dialogue group process or rules, that are not categorized by other META_xxxx	29	3/27 (SS): do not worry about spelling errors; you have posted; I want to ask you a favor

84	_META_TOPIC	>>		* Birds eye or systemic view of the topic	30	
59	WEIGH	_Meta_Topc	Solution Mediate	* Weighs alternatives; Identifying trade-offs. -If possible separate out the weighing from a solution or mediation in different segments - The trigger words are not exact indicators but possibilities, e.g. "however" may not always indicate weighing alternatives.	31	However... On the other hand... Realistically... While...could... But then again... At the same time... If, then... But... Still... The flip side... Although... Either...or... [cause/effect systems] [pros and cons]
57	SYSTEM	_Meta_Topic	_ArgGen, Meta-Check, Solution	* About the *topic* (not the dialogue). Introducing (for the **first time** in a dialog) a larger set of concerns in: time; geography; causality; level; part-to-whole systems. Moving the conversation from individual examples and factors to more inclusive, abstract, or big picture systems of things or factors. - If its not newly introduced but referring to something already brought up, it is not necessarily showing systems level thinking.	32	If the same logic were applied to another context... There is a larger issue at hand, When considering society as a whole... The wider economical implications... Our society has caused...
61	_ARG_GEN	>>	Give ALL other codes priority	Generic or Unspecified Argument element (usu. one of below; quick way to tag as an argument element without being more specific)	33	(Trigger notes are sparse for this category because we did not focus on it)
68	FACT_NOSRC	_ArgGen	Exper_Observ Fact_Src	Stating a fact without noting the source in this post. - If personal experience see Exper_Obsrv. - If source given use Fact-Src. - Facts are specific, not general.	34	
??	GENERAL_SOLUTION	_ArgGen	ActPropose; Di_actPropose;	Proposing or hypothesizing a solution to a problem being discussed. It is a general solution rather than something the participants could actually bring about.		
78	EXPER_OBSERV	_ArgGen		Experiential/observational fact; use this rather than Fact or Example, when the fact or example is from personal experience; includes "anecdotal" argument support if from one's experience. - Note: "in my experience…" and "in my observation…" can be used to falsely introduce opinions to sound like facts (or	35	

				observations). If its not an actual observation or experience, code it as an Opinion (or other code).		
62	ARG_OPINION	_ArgGen	_General_Solution	Giving an opinion with a reason or argument later in same post (another segment)	36	
63	OPINION_ONLY	_ArgGen		Opinion without support or elaboration (in this post)	37	
79	OVER_GEN	_ArgGen		Over-generalization; statements using ("universal") qualifiers like all, none, every, never, always, everyone, no-one... that are almost surely hyperbole or not accurate. They are opinions presented as fact. (Use this code, not Opinion or Fact.) When moderated with words like "almost", or “usually” then DON'T code them as Over_gen.	38	Use of: all, none, every, never, always, everyone, no-one
64	SUPPORT	_ArgGen	Fact Example	A later support, argument, or evidence for my opinion; that is not in another category (FACT, EXAMPLE,...)	39	
65	SUM_MY	_ArgGen		Summarize my own ideas (and see META-SUM)	40	
66	EXAMPLE	_ArgGen	Fact	Giving Examples or analogies to explain an idea or give evidence for opinion (not as a fact to support an argument, see FACT); grounding in real situations. Should be specific, not general (“young adults drink and drive” is general (usu. Support);	41	
67	ELAB	_ArgGen	Fact Example Support	A later elaboration of one's ideas : - giving background info - definitions, hypotheticals, consequence, - further/deeper explanation - framing or setting context for ones ideas (This is a general tag, if you can't code it more specifically as SUPPORT, EXAMPLE, FACT, etc.)	42	
85	_OFFTOPIC	>>		(See Base codes below)	43	
39	TECH	_OffTopic		Off-topic talk about the software being used.	44	
40	SOCIAL	_OffTopic		Social Talk, chit-chat, off topic and not covered by above; includes **greetings** and	45	Hello.. Good-bye...

				closings.		
41	OTHER			Not categorizable or difficult to code/uncertain.	46	
	[Action Codes]			**-** Note: **action codes** below are for discussing actions outside the dialogue. I.E. not for coordinating “speech acts” (for those see Meta-dialogue codes).		
88	ActRequest			Requesting action from others; assigning tasks; directives;	47	Would/could you...
89	ActPropose			Proposing or suggesting action/action or solution planning. Must be about actions of the participants (or any group of “we” that can actually take action; i.e. “our club” is OK, but “our city” is not directly actionable).	48	I think we should... What if we...
90	ActAccept			Committing/Agreeing to or Accepting Request	49	OK, I'll do it...
91	ActDecline			Declining/Rejecting Actions	50	I can't/won't do that...
92	ActNegot			Concessions/Negotiations/Proposal Amendments ; Maybe/Accept Part;	51	I will if you... I might...but...
	[Dialogue Action Codes]			*Dialogue* action codes; see Action codes above, but these apply to speech-acts in the dialogue, not to external contexts outside the online dialogue.		
93	DI_ActRequest			(see similar Action code above)	52	
94	DI_ActPropose			(see similar Action code above)	53	
95	DI_ActAccept			(see similar Action code above)	54	
96	DI_ActDecline			(see similar Action code above)	55	
	DI_ActNegot			(see similar Action code above)		
97	WELCOMING		Proc_Expl	-For mediators/facilitators ONLY. - Welcoming, introduce context, invite participants to introduce themselves.	46.2	
98	PROC_EXPL			-For mediators/facilitators ONLY. - Explaining the dialogue/deliberation process/plans/agenda.	46.3	
100	MOTIVATE			-For mediators/facilitators ONLY. - Speech acts to encourage participation	46.4	

				(that are not in another category)		
80	~~SOLUTION~~ (now-replaced with actPropose and ActREquest, General_So ution)		Mediate	(stopped using 4.18.12) A proposed solution to a problem being deliberated, or suggests what participants should do to solve the problem or take action--*outside* of the online dialog (if its about what to do in this dialog, see MEDIATE). (see Strzalkowsik et al.)		

A2. Descriptions of LIWC and CohMetrics text analysis systems

LIWC

LIWC (Linguistic Inquiry Word Count) uses a straightforward dictionary-based method of classifying texts. The LIWC dictionary contains about 4,500 words and word stems, and each dictionary word can be in one or more of the 80 word categories that the LIWC2007 software assesses. LIWC has been evolved and validated in a series of studies using independent judges (details are beyond the scope of this paper, see Tausczik & Pennebaker, 2010). "By drawing on massive amounts of text, researchers can begin to link everyday language use with behavioral and self-reported measures of personality, social behavior, and cognitive styles. Beginning in the early 1990s, we stumbled on the remarkable potential of computerized text analysis through the development of [LIWC]." (ibid, p. 25). In LIWC there are 4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (periods, commas, etc). (Pennabaker et al. 2007, p 4.) Of the 84 output measures given by LIWC we focused on 19 that seemed relevant to deliberative skills and the quality of dialogue, including: pronoun use (first and second person singular and plural), indicators of affect including positive and negative emotion, assent, certainty, and number of big words (>6 letters).

Coh-Metrix

Coh-Metrix combines a wide variety of text analysis methods and indices that have been developed and validated in the text processing and language analysis fields into one analysis system. It assesses syntax, referential cohesion, semantic cohesion, dimensions of the situation model, and rhetorical composition" (Graesser & McNamara, in Press). Coh-Metrix processes texts for 89 indices of cohesion, language, and readability. It contains modules including syntactic parsers (Charniak 2000); latent semantic analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2006), and other computational linguistics features. The Coh-Metrix team has collected and evaluated hundreds of measures of text over the last decade in the process of developing the approximately 100 measurements the system outputs (Graesser et al. 2011). Several studies have validated the Coh-Metrix indices (Graesser & McNamara, in press). Coh-Metrix has been used to help establish a wealth of evidence on a variety of text analysis topics, including detecting authorship through writing style, assessing temporal and structural cohesion in narrative, historical, and science genres; estimating human assigned grade levels of text books; assessments of formal/informal and spoken/written distinctions across genres; investigations of political leaders discourse strategies; and studies of gender differences across texts. We have investigated the language and discourse patterns of Arabic, Chinese, and Spanish leaders to examine how political discourse varies over time, across different cultures and languages, and for different parameters of history (economy, war, domestic uprising). In line with this, we have also examined how political leaders might strategically use discourse for manipulative and persuasive purposes.

A3. Early Text Classification studies, Year 2

In preliminary work we have shown that text analysis can provide useful information about extended deliberative engagements. Figure A1-A shows a comparison of some LIWC-based metrics vs. participant role (buyer, seller, and mediator) in e-commerce dispute resolutions. It graphically shows clear differences between these roles. For example, the mediator (Neutral) has less self-reference, less negative emotion, and uses longer words than the participants. Sellers use less social and less positive words than buyers. We are combining this analysis with data about successful mediations with the goal of giving mediators or participants advice on how to have successful engagements. Figure A1- B below shows some CohMetrix metrics vs. the temporal phase of a listserve-based dialogue among professionals. It shows clear trends from the beginning to the end of an online deliberation (the end involved a kind of impasse): syntax becomes more complex and cumbersome; there less lexical co-reference and more negative language as the engagement proceeds over several weeks. This type of analysis may help provide feedback to avoid or repair impasses.

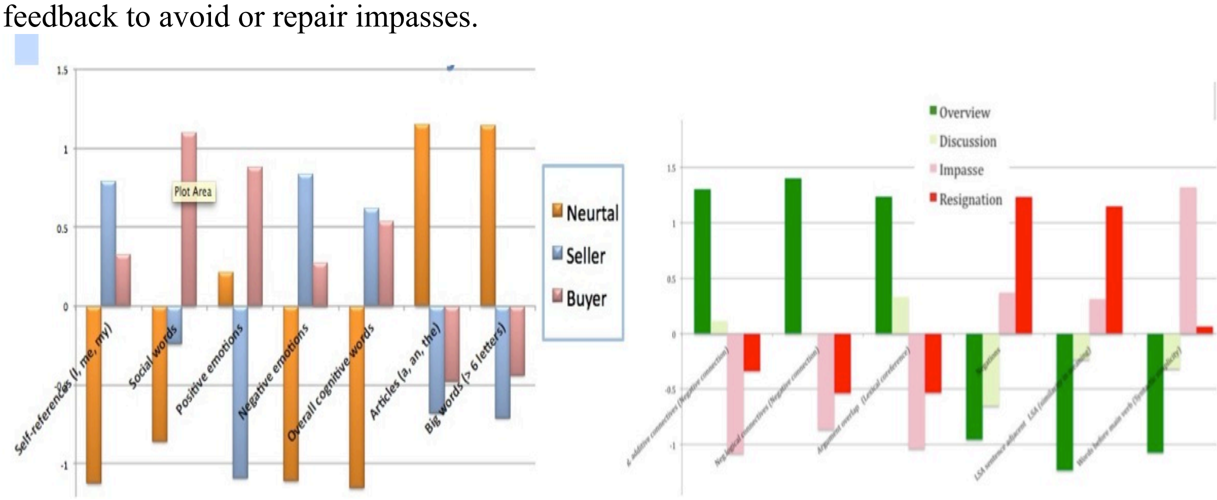


Figure A1 A, B: Ebay and Professional Dispute analysisWe are developing machine-learning algorithms that combine multiple factors, as explained below.

In one study (Xu et al. 2012a) we explored the possibility of predicting settlements in online disputes by performing text-analysis on conflict narratives from disputant parties. The domain was eBay Motor vehicles, in which disputants try to resolve complaints, possibly working with online human mediators. The conflict discourse was analyzed based on the divergence of topic distributions in a generative model extending Latent Dirichlet Allocation (LDA) to include role (buyer, seller, third party) information (Blei, 2011). We analyzed the quality of discovered topics in terms of topic coherence and evaluate settlement classification and prediction power. Experimental results showed that the unsupervised model outperforms a state-of-the-art supervised learner on precision, recall, and F-measure. The performance of a supervised learner with derived features from this model outperformed the unsupervised version using bag-of-words in terms of precision and efficiency. This work shows promise for being able to inform participants of when their interaction styles are likely to lead to unsatisfactory results.

SVM (10-Fold CV)	
True Positive = 175	False Positive = 65
False Negative = 36	True Negative = 51
Accuracy = 69.11% Precision = 72.92% Recall = 82.94% F-measure= 77.61%	

DNM+SVM (10-Fold CV)	
True Positive = 159	False Positive = 52
False Negative = 52	True Negative = 64
Accuracy = 67.89% ***Precision = 75.36%*** ***Recall = 75.36%*** ***F-measure = 75.36%***	

Performance of SVM (left panel) and DNM + SVM (SVM using derived features from DNM, right panel) for settlement prediction

Figure A2: Ebay analysis

In another set of studies we tested the hypothesis that a statistically significant correlation exists between discourse features (recognized automatically) and dialogue skills observed by human annotators (Xu et al. 2012b; Woolf et al. 2012). See Tables A1, A2. We found significant correlation between computational deliberation measures and human annotations for the same online dialogue. We combined over 3000 coded text segments from six domains. First, using multi-level analysis, we identified LIWC and CohMetrix features that had significant correlations with key constructs coded by humans (we are currently focusing on five composite codes: meta-dialogue, perspective-taking, big-picture-thinking, question-asking, and self-reflection. Next we used these metrics as features in machine learning algorithms. Preliminary results show that the best feature set and classification method is SVM (vs. Random Forest and Naïve Bayes classifiers), which is showing 50-70% class accuracy. This shows great promise for developing domain-independent classifiers for important speech acts, deliberative skills, and dialogue quality indicators. But our analysis is preliminary and much remains to be done.

Codes	significant measures	correlation coefficient (r)	pvalues (at the 0.01 level)	
_INTERSUB	DENPRPi	0.10	0.00	Personal pronouns
_INTERSUB	DENSPR2	0.11	0.00	Ratio of pronouns to noun phrases
_INTERSUB	PRO2i	0.21	0.00	Second person pronoun
AGREE	INTEi	0.11	0.00	Incidence of intentional actions, events, and particles.
APPREC	CONTEMPEXi	0.12	0.00	Temporal Connectives
DI_ActPropose	DATTIMi	0.10	0.00	DateTime
SELF_REFL	PRO1i	0.11	0.00	1st person pronoun
SOURCE_REF	ZRef	-0.10	0.00	?
SOURCE_REF	READFKGL	0.16	0.00	Flesch-Kincaid Grade level
SOURCE_REF	READASW	0.22	0.00	Average syllables per word

Table A1: Cohmetrix vs. Hand coding correlations

Codes	significant measures	correlation coefficient (r)	pvalues (at the 0.01 level)
_INTERSUB	**QMark**	0.1048512	2.03E-09
AGREE	**assent**	0.2893145	0
AGREE	i	0.1025945	4.47E-09
APOLOGY	negemo	0.1298163	1.05E-13
APPREC	affect	0.1461053	0
APPREC	posemo	0.1735854	0
DISAGREE	negate	0.1193964	8.34E-12
DISAGREE	negemo	0.1075049	7.85E-10
Q_RHETOR	Conj.untion	0.1144694	5.80E-11
Q_TOPIC	QMark	0.145354	0
UNCER	**Anx.ious**	0.1648797	0
UNCER	**Tentat.ative**	0.1161037	3.08E-11
DI_ActPropose	Numerals	0.1161791	2.99E-11
ActDecline	cause	0.1067837	1.02E-09

Table A2: LIWC vs. hand coding correlations

We are using the highly correlated LIWC and CohMetrix measurements as features in machine learning algorithms to build predictive models of social deliberative skills. Methods include:

- Mixed Method and Hierarchical/Multilevel linear and logistic modeling (Raudenbush, S. & Bryk, A. (2002).
- Latent Dirichlet Allocation (LDA) and Topic Modeling (David Blei, Andrew Ng, & Michael Jordan 2002)
- SVM (support vector machine) classifiers (Vapnik and Corinna Cortes in 1995).
- Generative and discriminative Linear Discriminant Analysis (supervised and unsupervised) Abdi, H. (2007)
- Breiman's random forest algorithm (a multi-decision tree ensemble classifier) (Prinzie et al. 2007)
- Sentiment and Emotion analysis (Pang, Bo; Lee, Lillian 2008)

A4. Samples from online dialogs from our data sets:

EBay (e-commerce):
A Buyer: “This seller is fraudulent and should be removed from eBay. Why should a eBay buyer have to be put through this.”

A Seller “…my good feedback be tarnished by these bottom feeders. That lay and cheat honest people out for there hard earned money.”

Forum: TeenGov, Topic: So Defensive
Post #1: Why are we so defensive when someone says that they don't like homosexuality. … I wouldn't get bent out of shape about it, I'd just calmly ask them why, and be done with it. Don't make it a biased and utterly offensive to the defending party…

Post #2: Because the spirit of the time demands it- we rarely determine our responses to particular social problems based merely on an abstract notion of rights…

…Post #20: Prejudice cant and never will be eliminated, but we can make it simmer down a little. Specific prejudices can be largely eliminated; very few Americans still hold a strong Irish prejudice for example

e-democracy: Minneapolis Powderhorn Neighbors Forum
[51 posts — by 31 authors, Dec. 2010]
Post #1: …while I still love my neighborhood for all its arty, community garden, Fair Trade goodness, I am disappointed -- and yes, angry…. these past few weeks [by what] feels disturbingly like [racial] targeting. This, coupled with the [documented] surveillance of parents of color…

Post #2: I'm so sorry that you are having this experience, especially in a neighborhood that prides itself on diversity. Thank you for sharing here so people can be more aware that this is still happening.

…Post #6: …The whites in Powderhorn pride themselves on diversity, but few actually mingle with their neighbors of color. They tend to reach out to the other liberal artsy gardening whites…

Workplace dispute: Intake summary
Boss (Grieta) to the Moderator: “Ryker has created a situation in which a continuation of the work relationship is no longer possible. What I am concerned, we are talking about terminating the work relationship. I will of course cooperate fully with a constructive mediation and hope for the best. It is unlikely that I myself can come to a solution with Ryke.r”

Employee (Ryker) to the Moderator: “Since late last year it has been a mess in the company. Management is unclear and inconsistent. The work relationship is disrupted. They want to get rid of me. I am literally "sick" of it. My confidence in the company has been shaken to such a point that I am not sure if I want to stay.”

Two professional communities deliberating on Conference Planning
"... there is a significant contingent of us who do not want to be a part of that type of community. Instead of welcoming our input, the leadership has consistently tried to stop the conversation."

"The problem with the current process is that members not on the executive committee do not feel they are a part of the process. And apparently, somewhere from 1/3 to 1/2 of the advisory board feels the same."

"If the leadership, …, does not feel this is important. . . then I will elect to spend my energies elsewhere and resign."

A5. Publications and Presentations

Peer Reviewed Publications

Xu, X., Murray, T., Woolf, B.P. & Smith, D. (2014). Identifying Social Deliberative Behavior from Online Communication -- A Cross-domain Study. Proceedings of FLAIRS: Florida Artificial Intelligence Research Society Conference, May, 2014. Pensacola Beach, Florida.

Xu, X., Murray, T., Woolf, B.P. & Smith, D. (2014). Social Network Signatures of Effective Online Communication. Proceedings of 12th International Conference on Intelligent Tutoring Systems (ITS-2014). June, 2014, Honolulu, HI.

Murray, T., Wing, L., Woolf, B., Wise, A., Wu, S., Clarke, L. Osterweil, L., Xu, X. (2013). A Prototype Facilitators Dashboard: Assessing and visualizing dialogue quality in online deliberations for education and work. Proceedings of The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE-2013). Las Vegas, July 2013.

Xu, X., Murray, T. , & Woolf, B. (2013). Text Analysis of Deliberative Skills in Undergraduate Online Dialogue: Using L1 Regularized Logistic Regression to Model Psycholinguistic Features. Proceedings of The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE-2013). Las Vegas, July 2013.

Murray, T., Stephens, A.L., Woolf, B.P., Wing, L., Xu, X., & Shrikant, N. (2013). Supporting Social Deliberative Skills Online: the Effects of Reflective Scaffolding Tools. Proceedings of 15th International Conference on Human-Computer Interaction (HCII-2013). Las Vegas, July 2013.

Murray, T., Xu, X. & Woolf, P.B. (2013). An Exploration of Text Analysis Methods to Identify Social Deliberative Skills. In Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED-2013). Memphis, TN, July 2013. K. Yacef et al. (Eds.): LNAI 7926, pp. 811–814.

Xu, X., Murray, T., Smith, D. & Woolf, B.P. (2013). Mining Social Deliberation in Online Communication: If You Were Me and I Were You. Proceedings of Educational Data Mining (EDM-2013). Memphis, TN, July, 2013.

Murray, T., Wing, L., Woolf, B., (2013). A Dashboard for Visualizing Deliberative Dialogue in Online Learning. Proceedings of 2nd Workshop on Intelligent Support for Learning in Groups—in association with AIED 2013 (Kim & Kumar Eds.). July, 2013, Memphis, TN, USA.

Murray, T. (2013). Toward Defining, Justifying, Measuring, and Supporting Social Deliberative Skills. Proceedings of Workshop on Self Regulated Learning — in association with AIED 2013 (Weerasinghe, du Boulay, & Biswas Eds.). July, 2013, Memphis, TN, USA

Shrikant, N. (2013). The trajectory of resistance to authority in online academic institutional talk. Proceedings of International Communication Association. London, June 17-21, 2013.

Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S. & Wing, L. (2012). "Supporting social deliberative skills in online classroom dialogues: Preliminary results using automated text analysis." Proceedings of 11th

International Conference on Intelligent Tutoring Systems (ITS-2012). S.A. Cerri and B. Clancey (Eds.). LNCS 7315, pp. 669–671, 2012. June 2012, Chania, Greece.
Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S. & Wing, L. (2012). "Towards Supporting Social Deliberative Skills in Online Group Dialogues." Presented at The 7th Annual Interdisciplinary Network for Group Research Conference (**InGroup**). Chicago July 12-14, 2012.
Xu, X., Smith, D., Murray, T. & Woolf, B. (2012). "Analyzing Conflict Narratives to Predict Settlements in EBay Feedback Dispute Resolution." Proceedings of the 2012 International Conference on Data Mining (DMIN12), July, 2012, Las Vegas.
Woolf, B.P., Murray, T., Xu, X., Osterweil, L., Clarke, L., Wing, L., Katsh, E. (2012). "Annotations and Computational Predictors in Online Social Deliberation." Presented at the 6th International **AAAI** Conference On Weblogs And Social Media (ICWSM). June 4, Dublin, Ireland.
Murray, T. (2009). Online Curriculum and Dialog Design for Ethics Skills for Science and Engineering Students. In World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (Vol. 2009, No. 1, pp. 555-564).
Murray, T. (2007). Toward collaborative technologies supporting cognitive skills for mutual regard. In Proceedings of the 8th international conference on Computer supported collaborative learning (pp. 542-544). International Society of the Learning Sciences.
Murray, T. (2003). Toward supporting information quality in rhetorical, dialogic, and collective on-line communication. In Proceedings of Workshop on "Metacognition and Self-regulation in Learning with Metacognitive Tools" R. Azevedo (Ed.).

Other Publications and Presentations

Murray, T. (2014). Supporting Deeper Deliberative Dialogue Through Awareness Tools. Presented at Build Peace, MIT, May, 2014.
Stephens. L., Murray, T., Shrikant, N., Wing, L., Xu, X. & Woolf, B.P. (2013, in submission). An Evaluation of Online Forum Features Supporting Social Deliberative Skills: Preliminary Results.
Wing, L. & Murray, T. (2014). "Supporting Conflict Resolution Skills in Social Media and Online Forums." An online radio show interview with Patricia Porter, host of the *Conflict Connection* radio show, Nov., 7 2013. http://www.adrhub.com/profiles/blogs/supporting-conflict-resolution-skills-in-social-media-and-online.
Wing, L., Murray, T., Woolf, B., & Katsh, E. (2013). "An Online Deliberation Facilitators' Dashboard: Visualizations and Text analysis to support quality dialogues." Presented at the 12th International Online Dispute Resolution conference (ODR Forum), July 17, 2013, Montreal.
Jeghelian, S., Palihapitiya, M.P., Murray, T. & Shore, N. (2013). "The Future of Collaborative Governance: Integrating F2F with online public engagement." Presented at UNCG 2013, Annual meeting of the University Network for Collaborative Governance, June 2-4, 2013; Malibu, CA.
Murray, T. (2012). Supporting Social Deliberative Skills in Online Contexts. Poster presentation at National Conference on Dialogue and Deliberation, Seattle, WA, October 2012.
Xu, X., Murray, T. & Woolf, B. (2012). "Analyzing Social Discourses to Predict Social Deliberation and Understand Social Intent." Poster presentation at the NSF Social Computing Doctoral Consortium.
Wing, L., Katsh, E., Murray, T., and Woolf, B. (2011). Supporting Social Deliberative Skills in Online Dialog, Deliberation, and Dispute Resolution. Presentation at The Tenth International Online Dispute Resolution Forum (ODR). Chennai, India, February, 2012.
Murray, T. (2003). A framework for developing cognitive tools that support critical, reflective, and multi-perspectival thinking. Poster presentation at the AACU Technology, Learning, and Intellectual Development Conference, October 2003, Cambridge, Massachusetts.

Selected Paper Abstracts

Murray, T., Xu, X. & Woolf, P.B. (2013). An Exploration of Text Analysis Methods to Identify Social Deliberative Skills. In Proceedings of 16th International Conference on Artificial Intelligence in Education (**AIED-2013**). Memphis, TN, July 2013. K. Yacef et al. (Eds.): Springer LNAI 7926, pp. 811–814.

We report on text processing and machine learning methods with the goal of building classifiers for social deliberative skill, i.e. the capacity to deal productively with heterogeneous goals, values, or perspectives. Our corpus includes online deliberative dialogue from three diverse domain contexts. We use the LIWC and CohMetrix linguistic analysis tools to generate feature sets for machine learning. We report on our evaluation of various machine learning algorithms, feature selection methods, and cross-domain training methods.

Xu, X., Murray, T., Smith, D. & Woolf, B.P. (2013) . Mining Social Deliberation in Online Communication: If You Were Me and I Were You. Proceedings of Educational Data Mining (EDM-2013). Memphis, TN, July, 2013.

Social deliberative skills are collaborative life-skills. These skills are crucial for communicating in any collaborative processes where participants have heterogeneous opinions and perspectives driven by different assumptions, beliefs, and goals. In this paper, we describe models using lexical, discourse, and gender demographic features to identify whether or not participants demonstrate social deliberative skills from various online dialogues. We evaluate our models using three different corpora with participants of different educational and motivational levels. We propose a protocol about how to use these features to build models that achieve the best in-domain performance and identify the most useful features for building robust models in cross-domain applications. We also reveal lexical and discourse characteristics of social deliberative skills.

Murray, T., Stephens, A.L., Woolf, B.P., Wing, L., Xu, X., & Shrikant, N. (2013). Supporting Social Deliberative Skills Online: the Effects of Reflective Scaffolding Tools. Proceedings of 15th International Conference on Human-Computer Interaction (HCII-2013). Las Vegas, July 2013.

We investigate supporting higher quality deliberations in online contexts by supporting what we call "social deliberative skills," including perspective-taking, meta-dialog, and reflecting on one's biases. We report on an experiment with college students engaged in online dialogues about controversial topics, using discussion forum software with "reflective tools" designed to support social deliberative skills. We find that these have a significant effect as measured by rubrics designed to asses dialogue quality and social deliberative behaviors.

Murray, T., Wing, L., Woolf, B., Wise, A., Wu, S., Clarke, L. Osterweil, L., Xu, X. (2013). A Prototype Facilitators Dashboard: Assessing and visualizing dialogue quality in online deliberations for education and work. Proceedings of The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE-2013). Las Vegas, July 2013.

The emerging next generation ("Web 3.0") of socio-technological tool development is adding additional support for reflecting on and improving the quality of online information, communication, and action coordination. An important opportunity is that online systems can include tools that directly support participants in having higher quality and more skillful engagements. We are evaluating dialogue software features that support participants directly and "dashboard" tools that support third parties (mediators, teachers, facilitators, moderators, etc.) in supporting higher quality deliberation. In this paper we will focus on our work in educational settings (college classes) and on our development of a Facilitators Dashboard that visualizes dialogue quality indicators for use as facilitation tools or participant social awareness tools. We are particularly interested in supporting the "social deliberative skills" that interlocutors need to build mutual understanding and mutual regard in complex or contentious situations.

Xu, X., Murray, T. , & Woolf, B. (2013). Text Analysis of Deliberative Skills in Undergraduate Online Dialogue: Using L1 Regularized Logistic Regression to Model Psycholinguistic Features. Proceedings of The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE-2013). Las Vegas, July 2013.

We report on a text analysis and machine learning study of social deliberative skill using online dialogues on controversial topics from a college class. We report on our comparison between using the LIWC and Cohmetrix text analysis feature sets, as well as demographic feature information in an L1 Regularized Logistic Regression machine learning algorithm.

Murray, T. (2013). Toward Defining, Justifying, Measuring, and Supporting Social Deliberative Skills. Proceedings of Workshop on Self Regulated Learning — in association with AIED 2013 (Weerasinghe, du Boulay, & Biswas Eds.). July, 2013, Memphis, TN, USA

Social deliberative skill is the capacity to deal productively with heterogeneous goals, values, or perspectives, especially those that differ from ones own, in deliberative situations. In other papers we describe our team's initial results in exploring this domain, which includes evaluating software features hypothesized to support SD-skills in participants, using machine learning and text analysis methods to recognize SD-skills and other indicators of deliberative quality, and prototyping a Facilitators Dashboard to help third parties get a birds-eye-view of important aspects of an online deliberation so that they can better help participants bring SD-skills to bear within dialogues on controversial topics. In this paper we take the opportunity to expand upon the nature and importance of SD-skills as we currently understand them at a more theoretical level.

Murray, T., Wing, L., Woolf, B., (2013). A Dashboard for Visualizing Deliberative Dialogue in Online Learning. Proceedings of 2nd Workshop on Intelligent Support for Learning in Groups— in association with AIED 2013 (Kim & Kumar Eds.). July, 2013, Memphis, TN, USA.

New and emerging online trends in group education, work and communication have lead to a dramatic increases in the quantity of information and connectivity without always supporting—and sometimes sacrificing—their quality. An important opportunity is that

online systems can include tools that directly support participants in having higher quality and more skillful engage- ments. We are evaluating dialogue software features that support participants directly and "dashboard" tools that support third parties (mediators, teachers, facilitators, moderators, etc.) in supporting higher quality deliberation. We will focus on our work in educational settings (college classes) and on our develop- ment of a Facilitators Dashboard that visualizes dialogue quality indicators for use as facilitation tools or participant social awareness tools. The Dashboard makes use of text analysis methods to highlight indicators of dialogue quality. We are particularly interested in supporting the "social deliberative skills" that interlocutors need to build mutual understanding and mutual regard in complex or contentious situations.

Shrikant, N. (2013). The trajectory of resistance to authority in online academic institutional talk. Proceedings of International Communication Association. London, June 17-21, 2013.

The ambiguous hierarchy existing in academia is a source of tension in academic discussions, where deliberation is encouraged, but those who rank highly are more likely to control the decision making process. This paper takes a conversation analysis (CA) approach to analyze online academic interactions among an Advisory Committee formed, in part, to solve a conference scheduling issue. This analysis will examine how participants invoke and negate hierarchy during these interactions. Robert G, the appointed leader of the listserv discussion group, OrgE, consistently tries to control the conversation and make decisions without the input of OrgE members. OrgE members resist Robert's control by constructing strong disagreements, negative assessments, and performing intersubjectivity work. This paper presents Robert's first email to the group and then follows by presenting three of the many resistance episodes to Robert. These emails illustrate the increasing hostility participants express, which leads to their resignation, causing a breakdown in communication on the listserv.

Xu, X., Smith, D., Murray, T. & Woolf, B. (2012). "Analyzing Conflict Narratives to Predict Settlements in EBay Feedback Dispute Resolution." Proceedings of the 2012 International Conference on Data Mining (DMIN12), July, 2012, Las Vegas.

We explore the possibility of predicting settlements in online disputes by performing text-analysis on conflict narratives from disputant parties. The experiment domain is eBay Motor vehicles, in which disputants try to resolve complaints, possibly working with online human mediators. The conflict discourse is analyzed based on the divergence of topic distributions in a generative model extending Latent Dirichlet Allocation (LDA) to include role information. A set of distance schemes and a heuristic are designed for various negotiation scenarios to predict settlements. We analyze the quality of discovered topics in terms of topic coherence and evaluate settlement classification and prediction power for settlements on unseen data. Experimental results show that this unsupervised model outperforms a state-of-the-art supervised learner on precision, recall, and F-measure. The performance of a supervised learner with derived features from this model outperforms that using bag-of-features in terms of precision and efficiency.

Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S. & Wing, L. (2012). "Supporting social deliberative skills in online classroom dialogues: Preliminary results using automated text analysis." Proceedings of 11th International Conference on Intelligent Tutoring Systems (ITS-2012). S.A. Cerri and B. Clancey (Eds.). LNCS 7315, pp. 669–671, 2012. June 2012, Chania, Greece.

We present results from a study in which we tested features of online dialogue software meant to scaffold "social deliberative skills," which include social perspective-taking, question-asking, meta-dialog, and reflecting on how one's biases and emotions are impacting a dialogue. Social deliberative skills are important capacities in a wide array of social contexts in which people with differing goals, values, or perspectives deliberate toward some end, including civic engagement and dispute resolution. In this study we look at online dialogue on controversial topics in a college classroom. In addition to hand coding of the dialogue text we are exploring the use of automated text analysis tools (LIWC and Coh-Metrix) to identify relevant features. Automated analysis might allow for adaptive or intelligent scaffolding of dialogue software features, and could also be used in a Facilitator Dashboard, which we are now prototyping, to bring a facilitator's attention to critical junctures in deliberative dialogues. In our preliminary analysis we found suggestive evidence that LIWC-based automated text analysis can differentiate the use of reflective tools and also differentiate some aspects of higher quality deliberative dialogue. In addition to the empirical results, this study contributes to a theoretical framework for the study and support of social deliberative skills.