Towards supporting social deliberative skills in online classroom dialogues and beyond

EXTENDED DRAFT of ITS-2012 Short Paper

Tom Murray, Beverly Park Woolf, Xiaoxi Xu,
Stefanie Shipe, Scott Howard and Leah Wing

Contact: tmurray@cs.umass.edu
University of Massachusetts, Amherst, MA

Abstract. 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.

Keywords. social deliberative skills, dialogue and deliberation, automated text analysis, reflective engagement, online skill scaffolding

1   Introduction

The capacity to deliberate with others about complex issues where interlocutors have differing viewpoints is paramount for so many life contexts, including citizen engagement, collaborative problem solving, knowledge building, and negotiating needs in personal relationships (Spragens 1990; Kögler 1992; Toulmin 1958). We have coined the term "social deliberative skills" to point to a set of skills that are important to success in such deliberative contexts. Social deliberative skills include the skills of perspective-taking, social inquiry (perspective-seeking), meta-dialog, and reflecting on how one's biases and emotions are impacting a deliberative process. Our research is looking into how to support higher quality deliberations in online contexts by supporting such skills. We are investigating a number of deliberative contexts, including online dispute resolution (for e-commerce, divorce settlements, and workplace disputes), online civic engagement, and online discussion forums on topics of importance to participants (including college students). Our approach is not to teach or tutor such skills, but rather to provide scaffolding that helps interlocutors bring existing skills to bear in contexts where they otherwise may not. I.E. we assume that participants have some generic

capacity for perspective taking, self-reflection, etc., but don't often enough bring such skills to bear "in the heat of the moment" in dialogue with others who have different goals, values, or belief systems. We are thus scaffolding within and along participants' "zone of proximal development" (Vygotsky, 1978; Murray & Arroyo 2002).

We are interested in supporting higher quality deliberations in both facilitated (with mediators, arbitrators, moderators, etc.) and non-facilitated dialogues. For facilitated dialogues we are designing a Facilitator's Dashboard that will allow a facilitator to get a birds-eye-view of one or more dialogues, and monitor key indicators to help decide when and where to make useful interventions. For both facilitated and non-facilitated dialogues we are interested in how a variety of scaffolding features (some adaptive) in the dialogue software can support increased use of social deliberative skills.

A key technology in our research is automated text analysis to characterize participant posts along a number of relevant dimensions, such as emotional tone, self-reflection, topic abstraction, etc.   We are investigating whether text analysis methods developed by Pennabaker et al. (2007) and Graesser et al. (2011) can measure characteristics relevant to supporting quality deliberation. If we identify valid classification models, we plan to use text classification for three purposes: (1) for post-hoc analysis to supplement the hand-coding of dialogue text in evaluations of experimental trials of interventions; (2) for near-real-time analysis to feed information into the facilitator's dashboard; and (3) for real-time analysis to support intelligent adaptation of scaffolding features of the dialogue software (e.g. to enable or disable prompting features).

We eventually intend to apply our interventions to situations involving authentic contexts such as online civic deliberation and online dispute mediation. We have obtained text from prior online deliberations in these domains and are analyzing this text in various ways—this work will be reported on elsewhere. To date the experimental contexts we have worked with are classroom online dialogues.   In this paper we report on one set of classroom studies. We report on our first experimental study, an online dialogue among college students on three "hot topics" in which three sets of software features were compared. We hypothesize that software features meant to support higher quality dialogue and the use of social deliberative skills will have these intended results, as compared with a control group not using these features.

2   Social Deliberative Skills

Rosenberg (2004) claims that "A good deal of research on small group behavior and communications provides evidence of people's evident inability to understand and fairly consider other people's perspectives, to think critically about their own position or the social conventions to which they adhere, or think about problems

creatively and generate novel alternatives" (p. 4). He advocates for public processes that develop the "cognitive capacities, emotional orientation and social context" for democratic deliberation. According to Inglis and Steele (2005) "if attempts to bring citizens together to grapple with complex social issues are not designed to consider their diverse worldviews, capacities and complexities, the best in people will not be brought forward, and participatory projects will flounder, leaving people frustrated and eventually apathetic" (and see Ross 2005).

In framing the work in terms of "social deliberative skills" we empathize the following:

1. *Skills*: Our approach to supporting dialogue and deliberation is *skills*-based. I.E. we are interested in supporting underlying skills rather than, or in addition to, more surface features of dialogue.
2. *Social*: We use the term "social" rather then "collaborative" — the skills of navigating diverse goals and viewpoints come into play in both cooperative and non-cooperative contexts; and in various social contexts just as diplomacy, not only group dialogue and problem solving.
3. *Deliberative*: "Dialogue" often involves the sharing of opinions where there is no particular shared goal and/or nothing at stake for the participants. In such situations interlocutors often focus on shared understandings and avoid differences or areas of conflict. The term deliberative is used to explicitly include dialogue where there is enough at stake that interlocutors are challenged to deal with their differences.

Social deliberative skills are best defined in relationship to other reasoning skills. A significant volume of prior work has focused on supporting the overlapping reasoning skills of argumentation, critical thinking, inquiry, and meta-cognition, which play important roles in successful social/collaborative deliberation. This skill set includes: being able to coordinate between factual (or more concrete) and theoretical (or more abstract) forms of knowledge; monitoring and self-correcting one's thought processes; backing up claims or counter-claims with evidence and valid reasons; arguing logically; self-directed learning and information gathering; and evaluating the quality of information or arguments (Kuhn, 1999; (Kuhn [1999; 2000; 2008] differentiates metacognitive, metastrategic, and metatask knowledge, and her work illustrates the thick interrelationship between all of these intellectual skills, which, though related constructs, are often studied separately from each other.

This research focuses on what we call "social deliberative skills" which relate to but contrast the more cognitively oriented skills just described. At the center of social deliberative skills is a deep consideration and response to the opinions of others, which involves the skills of reciprocal role taking and cognitive empathy (Kögler, 1999; Habermas 1971, 1999; Goleman 1995). However, this central capacity directly implies many others, expanding the list of important skills. First, entering into another's perspective, if it differs from ones own, requires taking a more reflective and distanced stance form ones own beliefs (see Kögler on "self-distanciation"). Second reciprocal engagement implies a type of curiosity about other's opinions, which implies question asking (or "perspective seeking" as well as perspective taking) skills. Third, participants will want to check whether one has

correctly understood another and that another has correctly understood one (a specific type of question asking).

High quality deliberation, then, is more than merely expressing ones' opinion, and even more than arguing rationally for one's opinion in the face of competing opinions. It involves, at its best, (1) a sincere desire to know the truth about the *objective world*, through inquiry that is as unbiased and comprehensive/systematic as can be managed. And since deliberation happens in social contexts usually with diverse perspectives and goals, it involves (2) an inquiry into other's *inner worlds* (opinions, feelings, goals, etc.). (The moral/ethical implications are unavoidable here—see Habermas, 1999.) Finally, such opening to increasing circles of fact and opinion means one must *also* bring to bear a number of meta-skills in order to cope with both the increasing amount of information and the fact that much of it is uncertain or conflicting. Meta-cognition, meta-dialogue, and tolerance for uncertainty are among these skills.

In sum, Social deliberative skills are those skills which, in conjunction with the more cognitively oriented skills such as problem solving and critical thinking mentioned above, are needed when collaborators or interlocutors must navigate the terrain of differing goals, values, or assumptions. Social deliberative skills include:

- Perspective taking & cognitive empathy
- Perspective seeking (curiosity/inquiry and question-asking)
- Self-reflection: on one's biases, intentions, emotional state
- Meta-dialog: Reflect on the quality of the dialog
- Epistemic skill: e.g. treating facts/data differently from opinions/hypotheses
- Tolerance for uncertainty, ambiguity, disagreement, paradox

Though this skill set has significant overlap with what has been called emotional and social intelligence (Goleman, 1995; Mathews et al, 2002), rather than frame them in terms of the cognitive vs. emotional-social dichotomy, we see them as depicted in Figure 1. Social deliberative skills involve the application of cognitive-like higher order skills, which are usually described as focusing on objective problems "out there," *to* subjective and intersubjective domains.
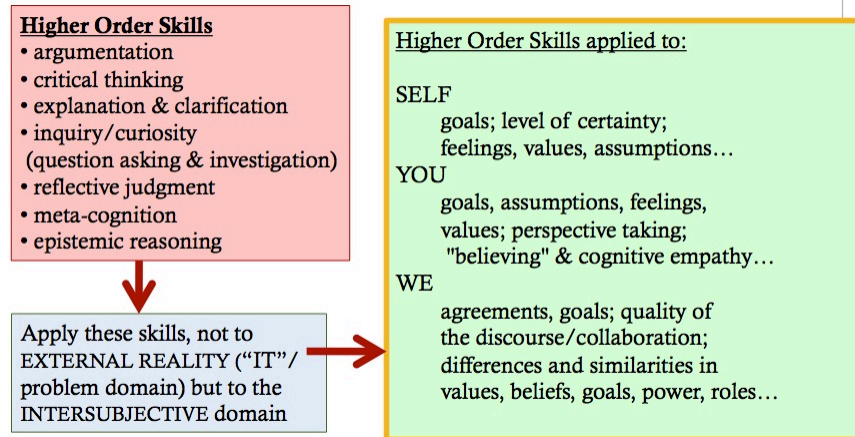
Fig. 1. Social Deliberative Skills as higher order skills applied to intersubjective domains.

We have suggested a number of software features that might support the use of many of the social deliberative skills we have described (Murray, 2007). However, in this study we used an existing commercial software platform that already contained features intended to support several of the key skills (more on the Mediem platform later).

3   Related Research

The support and measurement of quality in dialog and deliberation.   This work extends prior research into assessing (and eventually improving) communication and skillfulness in dialogue and deliberation. We will with Stromer-Galley's definition of *deliberation*: "a process whereby groups of people…engage in reasoned opinion expression on a social or political issue in an attempt to identify solutions to a common problem and to evaluate those solutions" (Stromer-Galley, 2007, p. 3). ["carefully examine a problem and arrive a t a well-reasoned solution after a period of inclusive, respectful consideration of diverse points of view" (Gastil & Black, 2008, p. 1) Stromer-Galley notes that "this definition aligns most closely with that of Schudson (1997), Habermas (1984), as well as Gastil (2000)." She notes that "a number of…experiments have found that deliberation affects political knowledge…considerateness of opinion…and attitude and opinion change…to name a few effects" (p. 1). For us the scope of deliberative dialogue includes various forms of dialogue-based conflict resolution and also includes two individuals working things out, as well as group and inter-group dialogue. Social deliberative skills are important in all of these contexts, though how they manifest or are measured may differ in each context.

Stromer-Galley (2007) defines six "elements of deliberation," which I paraphrase as follows: supporting arguments with reasons and evidence; referencing sources of information (a specific form of evidence); a diversity (heterogeneity) of perspectives; equality among participants (no one dominates or controls others); having a specific focal topic; and reciprocal engagement (considering and

responding to each other's opinions). In a study of European Union online civic deliberation (28 forums, 28 countries discussing "The social and economic future of Europe") Karlsson (2010) found that diversity of opinion and higher levels of engagement both correlated with more deliberation. Social deliberative skills focus primarily on reciprocal engagement, but we are also interested in reasoned arguments, referencing sources. (For this study we will treat diversity of opinion, having a focal topic, and equality as aspects of the dialogue that are pre-given and maximized as best we could prior to the online dialogues.)

Stromer-Galley (ibid) operationalizes her elements of deliberation and other dialogue properties in a coding scheme that was used to asses a civic engagement process in which 568 Pittsburgh residents deliberated for three weeks on school policy using three modalities: face-to-face, moderated online discussion, and "individual contemplation." Our coding scheme is very similar to hers but (for this study) does not code for the topic of conversation; and includes some additional codes focusing on aspects of reciprocal engagement.

Tools for online deliberation.

Social deliberative skills have some overlap with other higher order skills such as meta-cognition and epistemic skill, and prior research in computer-based support of such skills has provided us with initial models for cognitive capacities, scaffolding methods, representational parameters, and measurement metrics (Alven et al. 2006; Azevedo et al 2004; Bromme et al 2003; Winne 2001; Conati & Vanlehn 2000).   Much of the research into technology-supported dialogue and deliberation (situated within the fields of computer-supported collaborative learning and collaborative work) has focused on tools for *knowledge building, argumentation, or debate*. The main methods involve helping participants isolate and classify individual concerns, issues, or opinions, and supporting explicit relationships between contributions, such as pro, con, explanation/elaboration, questioning, meta-commenting, etc.   Some of these adapt textual interactions, essentially adding structure to discussion forum or chat-like interfaces (Klein, 2010), and others use on graphical interfaces that represent typed contributions as nodes and typed relationships as links in a concept or knowledge map (Scardemalia & Bereiter, 1994; Jonassen et al 1997; Nussbaum et al. 2007; Conklin 2005; Scheuer et. al. 2010). In focusing on social deliberative skills we are more interested in perspective taking and reciprocal engagement, an area where less work has been done.   Also, as we are interested in supporting participants without the need for significant training on unfamiliar tools, we are focusing on text-based interfaces that minimally constrain the flow of ideas (while not capturing as much of the structure of ideas as in graphical systems).

Suthers et al. have investigated how students engaged in computer-based collaborative problem solving engage in the "uptake" of each other's ideas (Suthers 2005; Suthers et al. in press). The focus on intersubjective processes is common with our work, but he is more interesting in supporting knowledge building and

meaning-negotiation in relatively well-defined problem solving tasks, where we focus on supporting social deliberative skills in more open-ended contexts. In resonance with our work, he is interested in exploring the unique affordances of technology to explicitly support collaborative processes through their ability to reify and visualize (externalize), and thus draw attention to, important aspects of a deliberation that would otherwise be more intangible and less salient.

Automatic Text Analysis Methods. Though we are hand-coding many of the deliberative dialogues used in our project, we are also interested in computer-automated analysis of discourse text. Automated analysis of important features of the text will allow us to (1) evaluate larger sets of data than would be manageable by hand coding, and (2) provide real-time analysis for adaptive (perhaps "intelligent") deliberation tools for participants and facilitators. Our interest is not in natural language processing that tries to understand the meaning of the text, but in classifying text according to relevant features. There have been a number of research projects processing large corpuses of data to determine general patterns of categorization. Tausczik and Pennebaker claim "we are in the midst of a technological revolution whereby, for the first time, researchers can link daily word use to a broad array of real-world behaviors" (2010, p. 24). Graesser et al. note that there "has been a dramatic increase in computer analyses of large text corpora during the last decade. This can partly be explained by revolutionary advances in computational linguistics…discourse processes…the representation of world knowledge…and corpus analyses…" (Graesser et al. 2007, p. 199).

Our present work aims to use build upon the results of two large relatively mature corpus analysis projects to help us evaluate deliberative texts and social deliberative skills, the LIWC project (Pennebaker et. al 2007) and the Coh-Metrix project (Graesser & McNamara, in press), explained in more detail below. At this time we are post-processing dialogue text through each of these software applications. (Real-time analysis will require additional software development.) We will be using the indices (measures, categories, metrics, output variables) output by these programs for several reasons: (1) the indices may be directly relevant to dialogue assessment; (2) the indices may have statistically significant relationships to our hand-coded categories, allowing us to build computational models to automate some of the hand-coded measures; (3) the indices may be useful as features input into machine learning algorithms that would produce models of the hand-coded categories. In this paper we focus on the first of these.

LIWC, Linguistic Inquiry Word Count. LIWC 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. For example, the word cried is part of five word categories: sadness, negative emotion, affect, verb, and past tense verb. Hence, if it is found in the target text, each of these five scores will be incremented. The classification scheme is hierarchical. For example, sadness is a subset of negative emotion, which is a subset of affect. The input for a LIWC run is a

sample of text (large or small) and the output is the number of words counted in each of its 80 categories. A word is scored as being in a category if it (or its stem) is a member of the dictionary words for that category.

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).   An overview of the categories output by LIWC is given in the following quote.

> [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 (Graesser et al., 2007, 2011) is a more sophisticated text analysis system than LIWC. It 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 categorized individual words (as does LIWC) and also uses "deeper or more processing-intensive algorithms that analyze 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. It also outputs a range of traditional textual measures such as average word length, average sentence length, and the readability formulas of Flesch Reading Ease and Flesch-Kincaid Grade Level (Klare 1974-1975).

Coh-Metrix uses a "multilevel discourse framework where the levels include the surface code (wording and syntax), the explicit textbase, the referential situation model, genre and rhetorical structure, and pragmatic communication" (Graesser et. al. 2010, p. 2). 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; and studies of gender differences across texts.

Of the approximately 100 measurements output by Coh-Metrix we focused on four composite measurements (or major factors) called Narrativty, Referential Cohesion, Syntactic Simplicity, and Word Concreteness. Graesser et al. (2011) conducted a principal component analysis over the set of Coh-Metrix indices and found eight major factors that accounted for most of the variance in texts across many grade levels and a variety of text genres. The study used the TASA corpus of 37,520 texts of approximately one paragraph in length. The corpus was representative of texts that students would encounter in grades K to 12, and each text is assigned an estimated K-12 grade level based on readability measures that provided an objective gold standard for the corpus analysis. The TASA text samples are categorized by genre including language arts, science, and social-studies/history, which gives some indication of placement on a spectrum of more narrative to more information-dense styles. Principal component analysis was performed to reduce the full set of Coh-Metrix measures to 8 functional dimensions or principal components, which together accounted for 67% of the variability among sample texts. The current version of Coh-Metrix outputs indices for the 8 principal components using Z-scores normed to the TASA corpus. In a sense each of these principal components reflects a composite or summary over (five to over 20) related primary language/discourse measures.

Though the principal components were derived from non-dialogical texts (books, articles, etc.) we expect that they will form a useful basis for studying some aspects deliberative dialogue. We are interested in them as possibly revealing measures of dialogue quality, and are also interested in whether these Coh-Metrix measures are predictive of social deliberative skills, as determined by our hand coding. Following is a brief description of the four principal components we used in this study: **Narrativity**: captures the extent to which the text conveys a story, a procedure, or a sequence of episodes of actions and events with animate beings. It measures a spectrum from more narrative to more information-dense texts. **Syntactic Simplicity**: scores are higher when sentences have fewer words and simpler, more familiar syntactic structures. At the opposite end of the continuum are structurally embedded sentences that require the reader to hold many words and ideas in working memory. **Word Concreteness**: Scores are higher when a higher percentage of content words are concrete, are meaningful, and evoke mental images—as opposed to being abstract. **Referential Cohesion**: the extent to which words and ideas in the text are connected with each other as the text unfolds. (Several of the

individual indices that loaded onto this factor came from latent semantic analysis.) "Cohesion helps most readers comprehend more texts, but under some conditions, texts with lower cohesion stimulate more knowledgeable readers to generate more inferences and meaningful explanations." (Greaser et al. 2011, p. 224).

4   Method

Forty college students in students in an Alternative Dispute Mediation courses were assigned a series of discussions to be had online. 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). The class had a face-to-face brainstorming discussion to decide three topics (from about 20 suggested during brainstorming) that would be most interesting to talk about. As our goal was to encourage spirited conversation with differing opinions—the topics did not have to be about topics covered in the class. The research team modified the chosen topics to turn them into focal questions. Students engaged in a sequence of three online dialogues, one per week, over three weeks. The three questions were:

- Week 1: *Discuss the pros and cons of legalizing marijuana.* (To focus the conversation, we invite you to assume you are on an advisory panel for the state legislature, having some preliminary conversations online, and you will eventually be drafting a group recommendation. Consider not only your own preferences but what is best for the state (or society).)
- Week 2: *Sex – what's the big deal? What values are most important in making sexual choices? How do you explain the diversity of opinions in society (or on campus) on this topic?* (We would like you to share your views and also express yourself mindfully to provide a dialogue space that feels safe for everyone to participate in.)
- Week 3: *Discus the pros and cons of the death penalty (capital punishment).* (To focus the conversation, we invite you to assume you are on an advisory panel for the state legislature, having some preliminary conversations online, and you will eventually be drafting a group recommendation. Consider not only your own preferences but what is best for the state (or society).)

Software used. For the online discussions we used the Mediem software created by Idealogue Inc. 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.
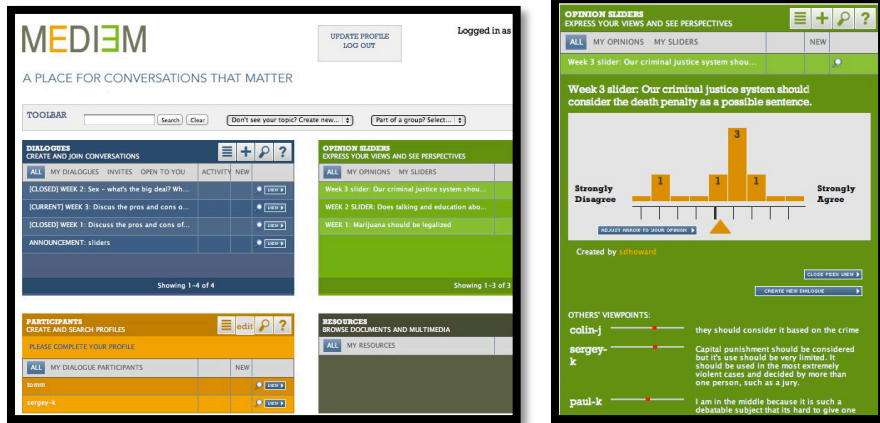
Fig. 2. A, B: Mediem Home Screen and Sliders

Mediem has been used in a number of dialogue contexts, including interfaith discussions among college students. Figure 1A 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 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. In our study we had three pre-determined dialogue topics, as described above. 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 1B shows the detailed view of the Opinion Slider feature. (As with Conversation topics, participants usually set up their own Opinion Slider questions, but ours were pre-defined for the classroom dialogues.) The slider gives a summary view of where participants stand on an issue. They can move the slider to change their opinion at any time and the current opinion state of all participants is shown graphically. Participants can optionally type in an explanation for their viewpoint, which is shown below the slider. One or more opinion questions can be linked to a discussion topic. In our trials we created one slider for each topic (as noted below, this was done only for certain groups).
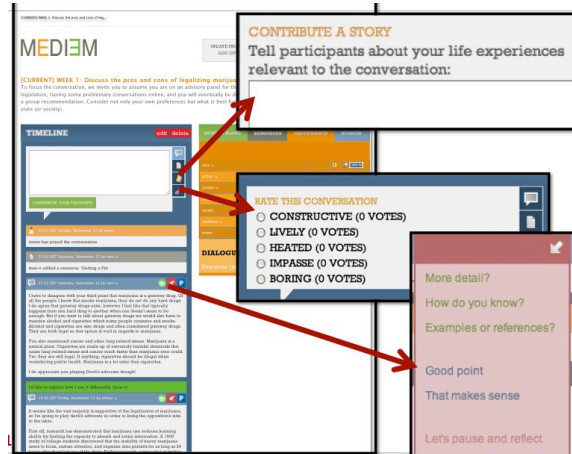
Fig. 3. Mediem Reflective Tools

The Mediem software was chosen for our study because it has a number of features designed to support deeper reflection and engagement. Figure 2 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 2 shows three reflective tools (the sliders also support more reflective dialogue). 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; Goodman et al. 2005). The fixed vocabulary, which can be customized the administrator, can be tailored to support or remind participants of the types of skills or attitudes that constitute a quality conversation.

Experimental Groups. The class of 40 students was partitioned semi-randomly into six groups of 6 or 7 each, in alphabetical order by last name. (Group sizes assigned: 1A=7, 1B=7, 2A=7, 2B=7, 3A=6, 3B=6. Group sizes actually participating: 1A=7, 1B=5, 2A=6, 2B=6, 3A=6, 3B=6.) We used the semi-randomized method of alphabetical grouping rather than fully randomized assignment because from past experience we have found that the easier and shorter the instructions are, the less likely that students will make mistakes in following them. We created two

discussion groups (A and B) per experimental condition (1, 2, 3) in order to limit the size of the groups to allow for more intimate and/or manageable engagements.

The instructions included the guideline: "It is expected that you will take the time to read other posts in your group and offer thoughtful responses."   The rest of the instructions were logistical and not related to encouraging the use of social deliberative skills.

Students in each group were given instructions to use software features as follows:

Condition:	V (control/vanilla) GROUPS 1A, 1B	S (sliders) GROUPS 2A, 2B	R (reflective features) GROUPS 3A, 3B
Use these Mediem tools only:	- Contribute & Reply	- Contribute & Reply - Opinion Sliders	- Contribute & Reply - Share a story - Thermometer - Contribution tags
Do NOT use these Mediem tools:	- Opinion Sliders - Share a story - Thermometer - Contribution tags	- Share a story - Thermometer - Contribution tags	- Opinion Sliders

When we want to speak of the sliders and reflective features together, we will call them "reflective engagement" features or tools.

Data collected. Post text and "reply" connections between posts were collected. Data was collected on Slider, Story, Conversation Tag and Thermometer use. Subjects were given a post-survey including the 18 questions using a 5-point Likert "agree...disagree" scale.

5   Experimental Hypotheses

This, our first experiment in this area, is an exploratory study. Below we describe our hypotheses for how the reflective engagement tools would affect various aspects of the dialogue. However, these should be taken as "soft hypotheses" in that, exploring some uncharted territory in an area with many variables, we are taking a broad brush approach and trying to identify parameters that will be promising to focus in on in future studies. Our hypotheses are:

H-1 (a,b,c,d) **basic quantitative statistics** : The Slider and Reflective features will result in more substantial dialogue as measured by these quantitative statistics a) number of posts, b) size of posts, c) connectivity of replies *by* participants and d) *to* participants.

H-2 (a,b,c,d) **opinion survey**: Subjects in the Sliders and Reflective conditions will indicate more positive attitudes on opinion survey questions related to: a) enjoyment, b) engagement, c) learning from others, and d) feeling understood (vs. the control group).

H-3 (a-l): **LIWC measures vs experimental condition**: We hypothesized that participants in the reflective tools groups would show the following patterns.

- An increase in: use of (a) "we" and (b) "I".
  - ◦ Reasons: tools that encourage participants to reflect on each others' viewpoints and reflect on the dialogue as a whole might cause more references to "us;" and more disclosure about one's subjective state (vs. keeping things completely objective and at a psychological arms length).
- An increased use of (c) affective (emotional) words, (d) agreement (assent) and disagreement (dissent) words
  - ◦ Reasons: again, we expect that self/you/we reflective tools will generate a more personal (thus emotional) engagement with the topics; and that more engagement would allow for greater critical thinking and opportunities to both agree and disagree (compare and contrast).
- An increase in words about (e) causation (e.g. because, effect), (f) thinking/insight (know, consider).
  - ◦ Reason: deeper engagement and reflection will support more arguments supported by reasons, relationships between events and ideas; and will cause more explicit reference to one's thought processes.
- An increase in (g) tentative words (e.g. maybe, guess), and (h) a decrease in certainty words (e.g. always, never).
  - ◦ Reason: deeper reflection and support of reciprocal perspective taking may lead to less certain, rigid, or dogmatic statements, and a greater sense of safety to express uncertainty.
- An increase in (i) number of words, words per sentence, (j) large (>=six letter) words, (k) use of quotation and (l) question marks.
  - ◦ Deeper engagement and reflection might lead to more volume of dialogue and more sophisticated language; tools supporting perspective taking may lead to more references to other's thoughts and thus more quoting and more asking each other questions.

We were also interested in any relationship between experimental group and positive vs. negative emotion, discrepancy words (e.g. should, could); cognitive processes (e.g. know, think, ought).

H-5 (a,b): **Coh-Metrix measures vs. experimental condition**: We hypothesized that participants in the reflective tools groups (using the sliders or the other reflective tools) would show (a) increased Referential Cohesion and (b) less Narrativity. Reason: Tools supporting deeper engagement may lead to more coherent and organized thoughts, and thus more referential cohesion; and may also lead to more information-rich dialogue that is less chatty or linear (though an encouragement to tell one's personal story may work in the other direction).

We were also curious to see whether there was a relationship between use of the reflective tools and the other Coh-Metrix measures of Narrativity, Syntactic Simplicity, and Word Concreteness.

6   Results

Data analysis is still in progress, and we give preliminary results here. We will start with basic descriptive measurements of the dialogue activity: the number of posts, the length of posts, and the response-connectivity between posts.

Number and length of posts. There were a total of 318 posts in the data set (covering 36 participants discussing 3 topics in 6 groups). 36 of the 40 students participated. However, participation varied greatly, as measured by total number of posts or average characters per post. Figure 4 shows the distribution: Ave. Characters per post: Mean 580, SD 215; Ave number of posts: Mean 8.9, SD 3.4. (There was no significant relationship between the number of posts and the length of posts among participants. We will use the 0.05 level for significance in this paper, unless otherwise noted.) Given that there were three topics over the three weeks, participants posted an averaged of 2.9 times in each discussion topic (or two in additional to an initial post). This is considered low, especially given that credit was given based upon participation.

There was no significant relationship between the number or length of posts and the experimental group (or the dialogue group of which there were two per experimental group). Given the instructor's impression that some topics generated more interest than others, it is somewhat interesting that there was no correlation between number or size of posts and the topic (i.e. participants engaged at about the same volume each week). Analysis of variance showed a significant (p=.007) and moderate difference in post length due to gender, with males averaging 521 and females averaging 648. Hypotheses H1-a ad H1-b are not supported in this experiment.
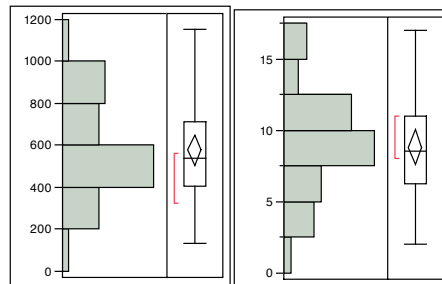

Fig. 4. Ave. characters per post and total number of posts for participants

Social network connectivity. On average students replied to others 4.3 times (of the 8.9 posts on average; with SD 2.9); i.e. about half of the posts were explicit replies to others. On students were replied to 4.0 times (SD 2.2). Figure 4 shows the distribution of the Replies_by_me and Replies_to_me for all participants, showing a significant range of behavior. There was no significant relationship between the Replies_by_me or Replies_to_me and the experimental condition. There was a significant (p=.0007) positive linear relationship between Replies_by_me and Replies_to_me (Replies_by_me = 1.1 + 0.69 * Replies_to_me) showing that overall the more often someone replied, the more often they were replied to. (There was no significant relationship between Replies_by_me or Replies_to_me and gender.) Hypotheses H1-c and H1-d are not supported by this experiment.
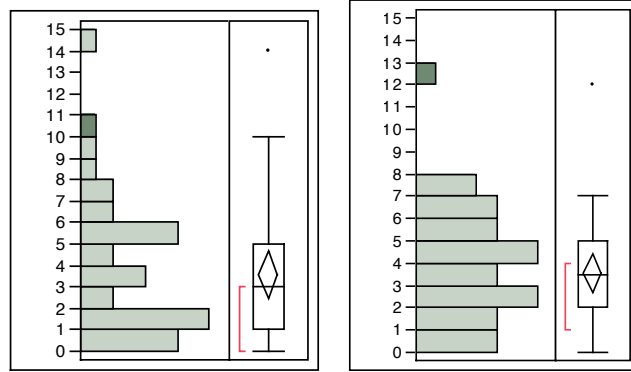
Fig. 5. Ave. Replies_by_me, Ave. Replies_to_me

Special tool use. Next we report on how often participants used the special features of the software, see Table 1. Groups 3A and 3B were assigned to use a variety of Reflective features, including the personal Story, the Contribution Tags, and the conversation Thermometer. (They also made use of the feature to post a resource file or link for group consideration.) We can see that Group 3B was about twice as active in using these reflective features as group 3A.[1] This may have been because of the group effect that once someone starts using a feature others are more encouraged to. Given that each group had three discussions and about six participants, or 18 total participant-discussions, these numbers reflect a rather sparse overall use of the features.

	Group 3A	Group 3B
Story	0	4 (4)
Tags	2 (2)	6 (2)
Thermometer	8 (3)	15 (6)
Resource	4 (3)	4 (3)
	Group 2A	Group 2B
Slider use	17 (6)	14 (5)

Table 1. Tool Use: Total uses (participants using)

Table 1 shows that Groups 2A and 2B, which were assigned to use the opinion sliders, had about the same amount of slider use. Over the three discussions, almost all participants used a slider at least once, with an average of about 3 times overall, or about one use per question. We consider this amount of use to be fairly low.

The low use of all of the reflective engagement tools may explain the lack of significant findings in the measurements described up to this point.

Survey Questions. Subjects were asked to take a post-survey including the 18 questions using a 5-point Likert "agree...disagree" scale. 29 of the 36 participants

---

[1] It is also notable that in 13 of the 15 slider uses from Group 3A an Explanation text was added (see Figure 1B); while for Group 3B there was no use of the Explanation text feature.

took the survey.   Below is a summary of the average responses (which, as discussed below, tend not to differ among treatment groups). (Lower scores in the 1-5 Likert scale indicate more agreement and a more positive response, since all questions were phrased positively.)

Question	Ave.	
I **enjoyed** the online dialog process.	2.72	neutral
The dialog software was **easy** to use.	2.21	agree
I felt **engaged** by the topics and the dialog.	2.64	neutral
The dialog felt free and **safe** enough for me to honestly voice my opinions.	2.03	agree
I **learned** something by reading others' postings.	2.28	agree
My opinion on the main topic **changed** over the course of the dialog.	3.55	disagree
Others understood and responded to my postings.	2.28	agree
Overall, factual statements were accurate and arguments were **well supported**.	2.41	agree

Table 2. Opinion survey overall results

Subjects in general thought the software was easy to use, thought it was a safe environment to share, felt understood and responded to, learned something in the process, and thought conversations were accurate and well supported. However, we can see that, even though subjects were allowed to generate topics of greatest interest, that they did not feel particularly engaged in the process nor was it enjoyable (neutral responses). Subjects did not feel that their opinion changed during the dialogue. Comments from the students indicated that, even though the topics chosen seemed interesting (or controversial and important), there was not much diversity in opinions among participants. In future trials we will take more care to select topics that are not only controversial *in general,* but controversial *among* the participants.

Hypothesis #2 is that subjects in the Sliders and Reflective conditions will indicate more positive attitudes on survey questions related to: a) enjoyment, b) engagement, c) learning from others, and d) feeling understood (vs. the control group).   None of these questions showed a significant relationship to experimental treatment at the .05 alpha level, but the enjoyment and engagement questions showed significance at the 0.1 alpha level, indicating suggestive trends. Post-hoc or pair-wise comparisons based on the Tukey-Kramer method were performed to detect significant differences. The analysis showed no difference between condition R and the control, but a notable difference between condition S and the control, in the direction of more positive results for the slider condition, shown in Table 3.

	Overall p	V mean (control)	S mean, p value vs. V	R mean, p value vs. V
enjoyment	0.10	3.1	2.3 (0.085)	2.7 (not sig)
engaged	0.092	3.0	2.1 (0.075)	2.7 (not sig)

Table 3: Survey questions showing near-significant trends

Thus some elements of hypothesis H-3 are supported for condition S, and none for condition R. But in light of the low use of the slider features, we consider this anomalous.

Analysis of LIWC and Coh-Metrix output.   Based on ANOVA, several of the 12 hypotheses related to LIWC outputs showed significant differences between the group (R) using reflective tools and the control group (V). Group R using reflective tools showed a significant (p=.00004) increase in the use of first person singular (I, me mine); an increase (p=.005) assent (agree, OK, yes, etc.) words; an increase (p=.0008) in cognitive process words (cause, know, ought); an almost significant increase in words per sentence (p=.06); and in increase (p=.015) in the number of large (>=six characters) words.   The group using the sliders (S) did not show any significant differences from groups V or R, which is compatible with the low use of this feature. Thus, some elements of hypothesis H-3 were supported for group R and none for group S.

We do not have the space here to speculate on the implications or reasons for each of these findings (nor the individual non-supported hypotheses), but suffice it to say here that with five of the 12 investigated indices showing a significant difference in the hypothesized direction (counting the p=.06 item as supportive), that LIWC-based automated text analysis shows promise for differentiating the use of reflective tools and also for differentiating some aspects of higher quality deliberative dialogue.

Of the two Coh-Metrix measurements hypothesized to increase, neither showed any significant differences between experimental groups. The one near-significant finding (p=.055) was an increase in Syntactic Simplicity for the group using the opinion sliders (S). Until we can imagine some explanation for this, we will consider it an anomaly.

7   Discussion

Our exploration of features of dialogue software that support social deliberative skills is just beginning, as is our exploration of the use of automated text analysis to provide adaptive support for deliberative dialogues. In our first classroom-based study we tested features that gave passive prompts or scaffolding for framing productive comments on others' ideas; reflecting on the quality of the dialogue as a whole; and relating the topic to what is personally at stake. We also tested an opinion polling and visualizing tool. We found very modest effects of using these features. There are two factors that might explain the near lack of significant findings. One was the fact that participation was on the low side (an average of 3 posts per discussion) and, despite several urging reminders, the student participants made very little use of the special features of the software. In future trials we will build more hearty forms of encouragement into the design. The other

was that there was not enough diversity of strong opinions to generate much interest in the topics.

Despite participating in a brainstorming activity to select interesting topics for discussion, students did not feel engaged. Classroom and survey comments indicated that, while the topics may have been interesting and controversial *in general*, that among the students there was not sufficient diversity or personal buy-in to generate enthusiastic dialogue. In future trials we will aim to pick topics that are controversial among the participants, rather than in general.

In the automated text analysis 5 of the 12 LIWC measures showed significant differences between the control and reflective tools groups (none of the four Coh-Metrix indices did). This is encouraging evidence that it is worth further pursuing automated text analysis with the eventual goal of building adaptive scaffolding features and guidance messages in a Facilitators Dashboard we are prototyping. We also plan to use the text analysis output as input features for machine learning algorithms to build predictive models of social deliberative skills, using our hand-coding of the dialogue text as the comparison standard.

8   References

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.

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.

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.

Bromme, R., & Stahl, E. (2003). The impact of epistemological beliefs on e-learning: the case of help-seeking. In F. W. Hesse & Y. Tamura (Eds.), Presented at the joint workshop of cognition and learning through media-communication for advanced e-learning (pp. 29-35), Berlin.

Charniak, E. (2000). A maximum-entropy-inspired parser. Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics (pp. 132-139). San Francisco: Morgan Kaufmann Publishers

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.

Conklin, J. (2005). Dialogue Mapping: Building Shared Understanding of Wicked Problems. John Wiley & Sons.

Fishkin, J. S. (1991). Democracy and deliberation: New directions for democratic reform. New Haven, CT: Yale University Press.

Fishkin, J.S. (1995). The voice of the people: Public opinion and democracy. New Haven, CT: Yale University Press.

Gastil, J. (2000). By popular demand. Berkeley: University of California Press.

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

Gastil, J., & Dillard, J. P. (1999). Increasing political sophistication through public deliberation. Political Communication, 16, 3-23.

Gastil, J., & Sawyer, K. (2004, February). Why process matters: An exploration of different operational definitions of public deliberation. Paper presented at the Annual Convention of the Western States Communication Association, Albuquerque, NM.

Goleman, D. (1995). *Emotional intelligence*. New York, NY: Bantam Books.

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., Cai, Z., Wood, J., Hatfield, D., Bagley, E., Nash, P., et al. (2010). Comments of journalism mentors on news stories: Classifications and epistemic status of mentor contributions. In C. Lynch, K. Ashley, T. Mitrovic, V. Dimitrova, N. Pinkwart & V. Aleven (Eds.), *Proceedings of the 4th International Workshop on Intelligent*

Tutoring Systems and Ill-Defined domains held at the 10th International Conference on Intelligent Tutoring Systems, ITS 2010 (pp. 21-28).

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.C., McNamara, D.S. (in press). Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*.

Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), pp. 223-234.

Habermas, J. (1962/1989). The structural transformation of the public sphere: An inquiry into a category of bourgeois society (T Burger, Trans.). Cambridge, MA: MIT Press.

Habermas, J. (1971). *Knowledge and Human Interests*. Boston: Beacon Press.

Habermas, J. (1984). The theory of communicative action: Reason and rationalization of society (T. McCarthy, Trans. Vol. 1). Boston, MA: Beacon Press.

Habermas, J. (1999). *Moral consciousness and communicative Action* (C. Lenhardt & S. W. Nicholsen, Trans.). Cambridge, MA: MIT Press.

Inglis, J. (2007). Matching Public Interaction Skills with Desired Outcomes by International Journal of Public Participation. Volume 1, Issue 2, 2007.

Jonassen, D, Reeves, T. Hong, H., Harvey, D., & Peters, K. (1997). Concept Mapping as Cognitive Learning and Assessment Tools. *J. of Interactive Learning Research* Vol 8 No. 3/4, pp. 289-308.

Karlsson, M (2010). "What does it take to make online deliberation happen? A comparative analysis of 28 online discussion forums." Presented at the IV International Conference on Online Deliberation in Leeds, June 2010.

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

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.

Kögler, H. H. (1999). *The power of dialog: Critical hermeneutics after Gadamer and Foucault*. Cambridge, MA: MIT Press.

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.

Kuhn, D. & Pease, M. (2008). What needs to develop in the development of inquiry skills? Cognition and Instruction, 26:512-559.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Matthews, G, Zeidner, M, and Roberts, R (2002). *Emotional intelligence: Science & myth*. Cambridge, MA: Bradford Book/MIT Press.

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.).

Murray, T. (2007). Toward Collaborative Technologies Supporting Cognitive Skills for Mutual Regard. In Proc. of Computer Supported Collaborative Learning, Rutgers: ISLS, July 2007.

Nussbaum, E. M., Winsor, D. L., Aqui, Y. M., & Poliquin, A. M. (2007). Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. International Journal of Computer-Supported Collaborative Learning, 2, 479 – 500.

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.

Rosenberg, M. (2003). Non-violent communication: A language of compassion. Encinitas, CA: Puddledancer Press.

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.

Ross, S. (2005). Toward An integral process theory of human dynamics: Dancing the universal tango. Integral Review, 1(1) 64-84.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge building communities. *Journal of the Learning Sciences*, *3*(3), 265-283.

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

Schudson, M. (1997). Why conversation is not the soul of democracy. Critical Studies in Mass Communication, 14, 297-309.

Soller, A.L. (2001) Supporting Social Interaction in an Intelligent Collaborative Learning System. International Journal of Artificial Intelligence in Education, 12(1), 40-62.

Spragens, T.A (1990). *Reason and Democracy*. Durham, NH: Duke University Press.

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

Suthers, D. (2005). Technology affordances for intersubjective learning, and how they may be exploited. In P. Dillenbourg (Series Ed.) & R. Bromme, F. W. Hesse, & H. Spada (Vol. Eds.), *Biases and barriers in computer-*

mediated knowledge communication: And how they may be overcome (pp. 295e319). Boston, MA: Kluwer Academic Publishers.

Suthers, D., Vatrapu, R., Median, R. Josphe, S. & Dwyer, N. (in press). Beyond Threaded Discussions: Representations Guidance in Asynchronous Collaborative Learning Environments. To appear in Computers & Education.

Tausczik, Y. R. & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*.

Toulmin, S.E. (1958). *The uses of argument.* Cambridge: University Press.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, (Eds.). Cambridge, MA: Harvard University Press.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman and D. H. Schunk (Eds.), Self-regulated learning and academic achievement (2nd ed.). Mahwah, NJ: Erlbaum Asso.