Assessment and support of the idea co-construction process that influences collaboration

Gahgene Gweon

CMU-HCII-12-101
April 2012

Human Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis committee:
Carolyn P. Rosé, Chair
Sara Kiesler
Dan Siewiorek
Bhiksha Raj
Aimee Kane (Duquesne University)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

This work was supported by the National Science Foundation under grants EEC0648487, EEC0935127, and SBE0836012. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect those of the funding agencies.

Keywords: Knowledge co-construction, idea co-construction, transactivity, group assessment, automatic assessment

Abstract

Research in team science suggests strategies for addressing difficulties that groups face when working together. This dissertation examines how student teams work in project based learning (PBL) environments, with the goal of creating strategies and technology to improve collaboration. The challenge of working in such a group is that the members frequently come from different backgrounds and thus have different ideas on how to accomplish a project. In these groups, teamwork and production of successful solutions depends on whether members consider each other's dissimilar perspectives. However, the lack of a shared history means that members may have difficulty in taking the time to share and build knowledge collectively. The ultimate goal of my research is to design strategies and technology to improve the inner workings of PBL groups so that they will learn from each other and produce successful outcomes in collaborative settings.

The field of computer supported collaborative learning has made much progress on designing, implementing, and evaluating environments that support project based learning. However, most existing research concerns students rather than instructors. Therefore, in my initial research, I explore the needs of the instructors in conducting student assessments (studies one, two). These studies identify five different group processes that are of importance from the instructors' perspective. My subsequent research focuses on one of them, namely the process of knowledge co-construction, which is a process that instructors have significant difficulty in assessing. In order to support the assessment of the knowledge co-construction process, my research has progressed along two axes: (a) identifying conditions that support the knowledge co-construction process and its relationship to learning and knowledge transfer (studies three, four, and five), and (b) automatically monitoring the knowledge co-construction process using natural language processing and machine learning (studies six ~ nine). Studies five and eight look at a specific type of knowledge co-construction process called the idea co-construction process (ICC). ICC is the process of taking up, transforming, or otherwise building on an idea expressed earlier in a conversation. I argue that ICC is essential for groups to function well in terms of knowledge sharing and perspective taking.

Acknowledgements

I have spent much of my time studying knowledge sharing between people and its benefits, so I fully realize that my dissertation would not have been possible without knowledge sharing from so many smart and kind people. My advisor, Carolyn P. Rosé, is one of the brightest and most knowledgeable people I have ever known, and I thank her for introducing me to the world of research and for her never-ending, passionate support of my work. My thesis committee members, Sara Kiesler, Dan Siewiorek, Bhiksha Raj, and Aimee Kane, each helped me overcome the "bumps" along the way. Each of them gave me ideas, as well as political advice, that proved valuable in becoming a skilled and professional researcher.

Over the years, I have also worked with Pulkit Agrawal, Emil Albright, Jaime Arguello, Lawrence Bergman, Rachel Bellamy, Regan Carey, Vittorio Castelli, Yue Cui, Susan Finger, Iris Howley, Mahaveer Jain, Soojin Jun, Rohit Kumar, Joonhwan Lee, John Levin, John McDonough, Richard Moreland, Tim Nokes, Margaret Schervish, Asim Simailagic, Marietta Sionti, Mikesh Udani, Laura Willson, and Sam Zaiss. I am lucky to have worked with so many talented people, and I thank each of them for helping me design and carry out countless experiments, and for taking part in many, sometimes wearying, coding sessions.

My journey would also not have been possible without the support of friends. I would like to thank the KCCP Praise Team, all of the Teledia members, and fellow HCII students and staff for countless, pleasant memories of Pittsburgh. Many thanks in particular to Turadg Aleahmad, Moira Burke, Justine Cassell, Marian D'Amico, Laura Dabbish, Scott Davidoff, Tawanna Dillahunt, Gregory Dyke, Matthew Easterday, Philip Gianfortoni, Jinhyuk Hong, Mahesh Joshi, Matthew Kam, Seungjun Kim, Sunyoung Kim, Ken Koedinger, Bob Kraut, Queenie Kravitz, Matthew Lee, Minkyung Lee, Bilge Mutlu, Ian Li, Elijah Mayfield, Sharad Oberoi, Amy Ogan, Choonsung Shin, Peter Simon, Eliane Stampfer, Leonghwee Teo, Erin Walker, Jason Wiese, and Ruth Wilye. Thank you.

Thanks, too, to Donghyun Joshua Ku for staying with me through the hardest moments as well as the happiest ones. His prayers helped me push through this long journey. I also want to thank the countless, unnamed people who prayed for me through personal hardships. Our Lord has always provided for us, and I have witnessed many miracles come about because of our relationships with each other and our Lord.

Finally, I thank my family. Even though we were thousands of miles apart, I knew in my heart that I could depend on them. I could always turn to my family for support no matter what situation I was in. Every time I visit my grandparents, I could see it in their eyes: they truly would do anything for the sake of my happiness and success. I thank my grandparents for their support, as well as my parents, Jeongjoo Gweon and Youngsook S. Gweon. They are the best parents I could ask for and they, along with my brother Dahgene Gweon, ceaselessly cheered me along the way.

For my parents,

Jeongjoo and Youngsook

Table of Contents

CHAPTER 1

Introduction

Research in team science suggests strategies for addressing difficulties that groups face when working together (Cummings & Kiesler, 2005). This dissertation examines how student teams work in project based learning (PBL) environments, with the goal of creating strategies and technology to improve knowledge collaboration between members. In the particular setting of the current thesis, student teams work to generate solutions to engineering design problems. Like other groups, they experience process losses (Steiner, 1972) and a lack of perspective taking (Krauss & Fussell, 1991; Schober, 1993; Schober & Brennan, 2003). Therefore, one challenge of working in project based learning groups is overcoming such problems and integrating different ideas on how to accomplish a project given that the members frequently come from different backgrounds. In these groups, teamwork and production of successful solutions depends on whether members consider each other's various perspectives, but the lack of a shared history means that members may have difficulty in taking the time to share and build knowledge collectively. The ultimate goal of my research is to design strategies and technology to improve the inner workings of project based learning groups so that they will learn from each other and produce successful outcomes in a collaborative setting.

Project-based learning (PBL), especially in courses where students work in groups on real world problems for and in collaboration with industry sponsors, is commonly believed by educators and administrators alike to have great value for engineering students (Dutson et al. 1997; Adams 2003). These courses are often situated in engineering curricula as capstone design courses that offer students the opportunity to integrate and apply the knowledge they have acquired in their more theoretical courses. Although there is much benefit in participating in such project courses, student groups do not always function in an ideal manner. In an attempt to address this problem, students learn the social skills required for working together in a group under the supervision of an instructor who acts as a group facilitator.

Although such guidance helps students in overcoming some of the problems that occur during group work (Hmelo-Silver, 2004; Meloth & Deering, 1999; McGrath, 1984), instructors may have difficulty discerning when support is needed because much of that work is done when instructors are not present. In fact, instructors have raised concern. They observed that problems are frequently intentionally hidden behind the well-functioning facade of groups until the final weeks of the semester, or until conflict becomes irreconcilable partly because students worry about breaking trust within the group (Gweon, et. al 2011). Recently, in order to support instructors in managing group work, researchers have developed various automatic assessment and reporting tools (e.g., Soller & Lesgold, 2003; Kay et al., 2006; Pianesi et al., 2008).

The field of computer supported collaborative learning has made much progress on designing, implementing, and evaluating collaborative learning environments that provide automatic assessments in project based learning (PBL). However, most existing research concerns students rather than instructors who provide guidance to those students. Open questions remain about what such technology can and should track, as well as how and when such information should be communicated. In particular, the questions on what types of information would be the most beneficial for the facilitator are as important as the technical challenges. For instance, if a system that supports a PBL environment captures information that is already apparent to the instructor, then such a tool would not be very useful. Therefore, in this dissertation, I address research questions that provide design insights for focusing research effort where it is needed. In order to examine what type of information should be captured for systems to be "useful" to expert facilitators/ instructors, I explore their needs during the assessment of student project groups (studies one and two).

In study one, I conduct an interview study with the goal of identifying and categorizing group processes that instructors believe are important in accomplishing successful group work. Thus, this study contains interviews with experts, i.e. instructors, who provide guidance to students who participate in project based learning groups. Study two is a classroom study conducted in an engineering design class over the course of a semester. The goal of study two is investigating the instructors' current assessment practices and observing instances where the instructors are having trouble in making accurate assessments. These two studies have identified five different group processes of importance from the instructors' perspective. The processes are goal setting, progress, knowledge co-construction, participation, and teamwork. My subsequent research focuses on one of them, namely the process of knowledge co-construction, which is a process that instructors were observed to have the most difficulty with in their assessment practices.

In particular, at the heart of my research is the idea co-construction (ICC) process, which is a specific type of knowledge co-construction process. ICC is the process of taking up, transforming, or otherwise building on an idea expressed earlier in a conversation. The ICC process is more commonly referred to as transactivity, which was originally defined by Berkowitz & Gibbs (1983). The idea is also related to various constructs from the computer supported collaborative learning community such as "intersubjective meaning making" (Suthers, 2006) and "productive agency" (Schwartz, 1998). The idea first comes from the neo-Piagetian perspective on learning where we understand that optimal learning between students occurs when students respect both their own ideas and those of the others that they are interacting with (de Lisi & Golbeck, 1999). Transactivity is theorized to reflect this balance within a collaborative learning setting. Since peer-to-peer knowledge sharing and integration, as part of collaborative learning, is also essential in a group-work setting, I argue that the ICC process is valuable for insuring that groups function well.

In order to support the assessment of the knowledge co-construction process in general and the ICC process in specific, my research has progressed along two axes: (a) identifying conditions that support knowledge co-construction and its relationship to learning and knowledge transfer (studies three, four, and five), and (b) automatically monitoring group processes using natural language processing and machine learning (studies six, seven, eight, and nine). The research questions addressed in this dissertation contribute to the fields of behavioral science, computer supported cooperative work, computer supported collaborative learning, and human computer interaction. Figure 1.1 presents a diagrammatic summary of this thesis.

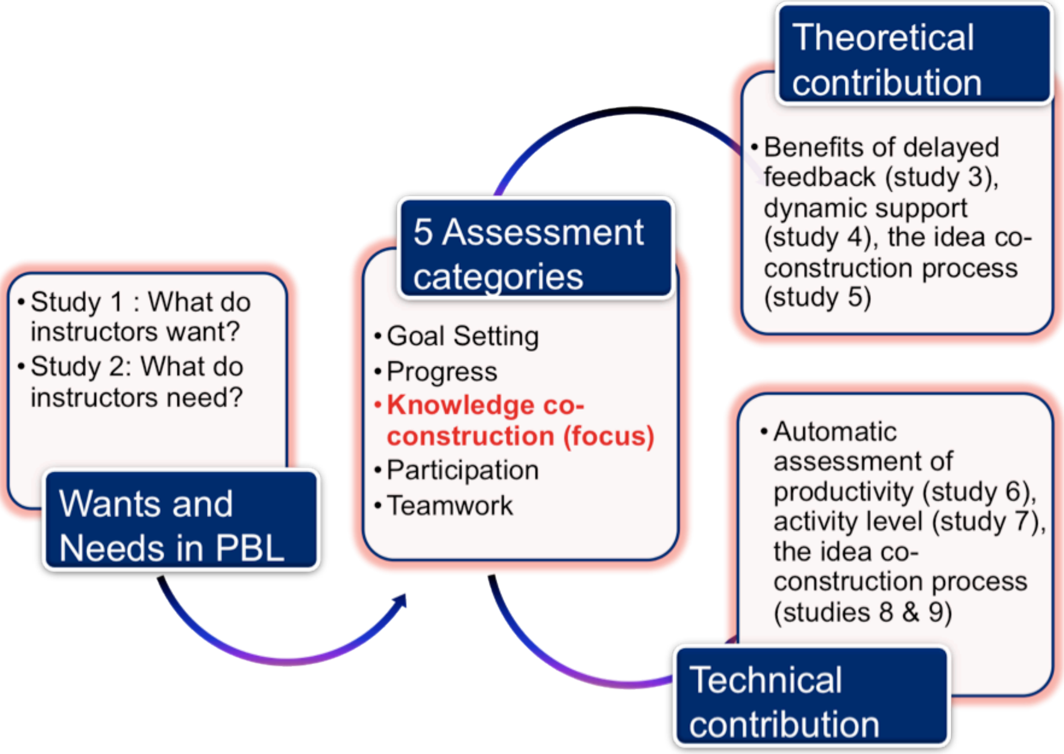

Theoretical contribution
• Benefits of delayed feedback (study 3), dynamic support (study 4), the idea co-construction process (study 5)
5 Assessment categories
• Study 1 : What do instructors want?
• Study 2: What do instructors need?
• Goal Setting
• Progress
• Knowledge co-construction (focus)
• Participation
• Teamwork
Wants and Needs in PBL
• Automatic assessment of productivity (study 6), activity level (study 7), the idea co-construction process (studies 8 & 9)
Technical contribution

Figure 1.1. Thesis overview

The research questions addressed in each of the nine studies are summarized in table 1.1. Although my broad research interest is in supporting collaborative work in general, I started investigating the factors that impact collaborative work in an educational setting. The studies were conducted in two types of collaborative learning environments; project oriented classes (studies one, two, five, six, seven, eight and nine) and intelligent tutoring systems used in a collaborative setting (studies three and four).

I start by addressing what the instructors value and need in project based learning environments (studies one and two). Among the five assessment categories identified from these studies, I focus on one type of knowledge co-construction process, namely idea co-construction (ICC). Study five examines the value of ICC in collaborative discussions. Based on the demonstrated value of ICC from this study, I examine methods for automating the ICC process in later studies (studies seven, eight, and nine). Studies three and four look at conditions for supporting the knowledge co-construction process in an intelligent tutoring environment used collaboratively. Because the automatic assessment of speech conversations that occur in project oriented courses are where the bulk of my technical contribution lies, the body of this dissertation presents only those related studies that present a coherent story leading up to this technical contribution. Studies three, four, and six are presented in the appendix (E, F, and G respectively) for interested readers.

The studies presented in the body of this dissertation are face-to-face interactions where the medium of communication was speech rather than text. In contrast, the studies presented in the appendix (studies three, four, and six) are text based interactions. Table 11.1 provides a road map of the research questions addressed in each study, as well as the location of each study.

The next chapter (chapter two) describes broadly related background work. After the presentation of main studies, chapter nine provides a general discussion of the results, limitations of the work and a vision for future work.

Table 1.1. Overview of studies 1 ~ 9

Goal	Study	Question	Context	Loc
Instructor needs in problem based learning environments	1	What do instructors want students to do and learn in project based learning environments?	Project class / speech	Ch 3
	2	How much do instructors really know about what is going on during group projects?	Project class / speech	Ch 4
Identifying conditions that support the knowledge co-construction process	3	Is providing delayed feedback more effective than immediate feedback in collaborative learning environments, given that delayed feedback provides students more chances to engage in the knowledge co-construction process?	Intelligent tutoring system / text	App E
	4	Would we still see learning benefit with dynamic support rather than the state-of-the art static support, given that dynamic support provides less help but more chances for students to engage in the knowledge co-construction process?	Intelligent tutoring system / text	App F
	5	Are students more likely to transfer knowledge when they engage in the idea co-construction processes?	Project class / speech	Ch 5
Automatically monitoring the group processes	6	To what extent can the rudimentary features extracted from message board discussions, be used to predict the amount of knowledge co-construction?	Project class / text	App G
	7	To what extent can the rudimentary features extracted from speech recordings of collaborative discussions, be used to predict various group processes, including the process of knowledge co-construction?	Project class / speech	Ch 6
	8	To what extent can the rudimentary features extracted from speech recordings of collaborative discussions, be used to predict where idea co-construction is occurring in those discussions?	Project class / speech	Ch 7
	9	Can the insights from sociolinguistics be used to improve the prediction of where idea co-construction is occurring in collaborative discussions?	Project class / speech	Ch 8

CHAPTER 2

Background

This chapter surveys three existing areas of research; existing frameworks for group learning support, the role of the knowledge co-construction process in group work, and existing technology for automatic assessment of conversational characteristics. Each area provides background in addressing the research goal of designing strategies and technologies to improve the inner workings of project based learning groups.

Section 2.1, "Frameworks for group learning support", provides a review of the current state of assessment tools in collaborative learning environments. These tools provide support in assessing various types of group processes such as the processes of division of labor and knowledge co-construction. Since the latter part of this dissertation is focused on supporting the knowledge co-construction process, section 2.2 reviews the importance of the knowledge co-construction process as well as one specific example process, the idea co-construction process. Verifying the important role of the knowledge co-construction process during group work would make a valuable theoretical contribution. In addition, I apply these findings to make a practical contribution as well. I accomplish this by addressing the technical challenge of automatically monitoring group processes in conversations that occur during group work. As a comparison to my approach, section 2.3 reviews existing technologies in automatic monitoring of social processes in text and speech.

2.1. Framework for group learning support

In the past decade, there has been an increasing interest in developing automatic assessment technologies to aid facilitators of group work who oversee numerous groups (Jochems & Kreijns, 2006; Wong et al., 2007; Ren et al., 2008; Phielix, Prins, & Kirschner, 2010; Phielix, Prins, & Kirschner, 2011). Among other potential uses, such technologies can assist expert facilitators by capturing and summarizing the group work conducted in their absence. With the help of such assessment technologies, the hope is that expert facilitators can detect potential problems and intervene before conflicts escalate and become unmanageable. Given this need, researchers have developed various automatic assessment and reporting tools that assist facilitators in providing feedback to groups (e.g., Kay et al., 2006; Kochakornjarupong & Brna 2010; Ma, et al., 2010; Pianesi et al., 2008; Soller & Lesgold, 2003).

More recent advances have demonstrated the potential for automatic collaborative process analyses (Donmez et al., 2005; Joshi & Rosé, 2007; Rosé et al., 2008; Gweon et al., 2011b) using machine learning. The need for such tools is motivated by the typically low facilitator-to-group ratios. Facilitators often do not have enough time and resources to provide adequate attention to all the groups that they supervise, let alone individual group members. This problem is exacerbated by the increasing amounts of data that instructors need to manage (e.g., messages from online discussion boards), owing to the rising levels of technology assimilation in instruction. Many of the existing assessment tools suggest technological solutions in distilling and presenting summaries of the data, to provide facilitators with the extra information that might otherwise have been overlooked.

A conceptual framework referred to as the Collaboration Management Cycle is shown in figure 2.1 (Soller, et al., 2005). This foundational work was influential in forming a vision for work on dynamic support for collaborative learning. In this work, Soller and colleagues provided an ontology for types of support for collaborative learning, which is still a useful lens through which to examine state-of-the-art approaches in light of prior work. They illustrated the central role of the assessment of group processes underlying the gamut of support approaches, including (a) mirroring tools that reflect the state of the collaboration directly to groups, (b) meta-cognitive tools that engage groups in the process of comparing the state of their collaboration to an idealized state in order to trigger reflection and planning for improvement of group processes, and finally, (c) guiding systems that offer advice and guidance to groups. At the time, guiding systems were in their infancy and all of the systems reviewed were research prototypes, mostly not evaluated in realistic learning environments.

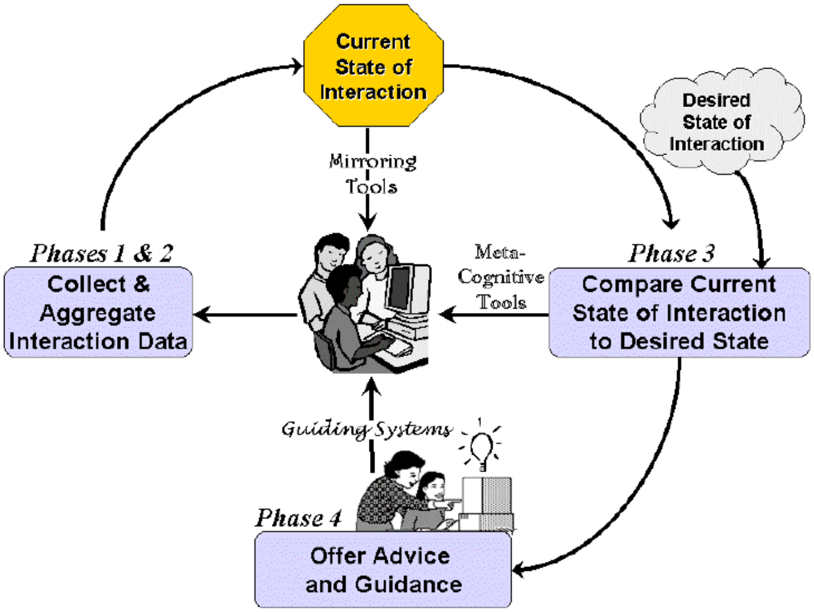

Current State of Interaction
Desired State of Interaction
Mirroring Tools
Phases 1 & 2
Collect & Aggregate Interaction Data
Meta-Cognitive Tools
Phase 3
Compare Current State of Interaction to Desired State
Guiding Systems
Phase 4
Offer Advice and Guidance

Figure 2.1 The Collaboration management cycle, reproduced in exact form from (Soller, Mones, Jermann & Muehlenbrock, 2005).

Walker (2010) and Kumar & Rosé (2011) were the first to develop full-fledged "guiding systems" that have been evaluated in large scale studies in real classrooms. My early work, discussed in studies 3 & 4, spans the gap between the vision from Soller et al. (2005) and its recent realization by Walker (2010) and Kumar & Rosé (2011). The vision of dynamic support for collaborative learning can be seen as emanating from the visionary work of Soller and colleagues. However, at the time of writing that article, the state-of-the-art in support for collaborative learning was a static form of support known as script based support (Kollar, Fischer, & Hesse 2003). Static support is offered through structured interfaces that scaffold the interactions by making it clear what types of contributions are preferred. It was believed at the time that feedback offered to groups based on their perceived needs are more effective, especially in the long term where constant support can be stifling. Most of the insights into offering feedback to students in the process of problem solving have come from the intelligent tutoring research community. My early work (study three) contributes to this effort by providing insights into the ways in which feedback offered to groups might function differently from the feedback offered to individual problem-solvers.

A major question related to dynamic support for collaborative learning is whether the benefits afforded by tailored feedback outweigh the effects of constant scaffolding offered by interfaces that continuously offer support regardless of detected need. My early studies of supported calculus problem solving address this question, offering evidence that dynamic support that targets specific needs as they surface within collaboration, while infrequent, is sufficient to have a significant impact on learning (study four).

While there are many technical differences between the Walker (2010) and Kumar et al. (2011) architectures, what they have in common is the application of machine learning to the assessment of group processes. Both of these approaches employ the TagHelper tool kit for analyzing conversational data (Rosé et al., 2008). Early work with the development of TagHelper (Donmez et al., 2005; Wang et al., 2007; Rosé et al., 2008) focused on the automatic assessment of knowledge co-construction from text based interactions. They achieved reliabilities at the level of .69 Kappa, in comparison with human coding. Other recent work demonstrates the same level of performance on transcriptions of face-to-face large group interactions in a classroom discussion context (Ai et al., 2010). However, none of this work addresses the new challenge of detecting social processes using speech. I address this challenge in the latter part of my dissertation (studies seven, eight, and nine).

2.2. Role of the knowledge co-construction process in group work

The great value of collaboration, especially collaboration between multiple experts, is that when it is managed well it creates an environment in which creative solutions to difficult problems can be generated. Generating solutions to engineering design problems is no exception; ideally, members with different expertise work together to produce successful products by sharing their expertise and learning from each other. Nevertheless, this ideal is rarely realized because of various difficulties that occur frequently during group work. In my dissertation work I address one difficulty that is critical in accomplishing a successful project, namely integrating differing perspectives via engaging in the process of knowledge co-construction.

The knowledge co-construction process can be observed from an individual who takes the initiative to use his unique knowledge and skills to contribute to the group work. In addition, when knowledge co-construction occurs within a group, members mentor others on skills, share ideas, and engage in meaningful conversations that may lead to learning and project advancement. Note that not all conversational contributions would be considered instances of a knowledge contribution in the truest sense. Evidence of reasoning, either of an individual's own or other's ideas/ knowledge, is necessary. In this sense, a more specific type of knowledge co-construction, called idea co-construction (ICC) can be defined. ICC is the process of taking up, transforming, or otherwise building on an idea expressed earlier in a conversation. It comes from the neo-Piagetian perspective on learning. Piaget suggested that through the balance of assimilation and accommodation process, children advance through the stages of cognitive development (Piaget 1985). Likewise, learning science researchers have adopted Piaget's theory and demonstrated that optimal learning between students occurs when students both respect their own ideas and those of others' (de Lisi & Golbeck, 1999). The ICC process is also referred to as transactivity, which was originally defined by Berkowitz & Gibbs (1986). Because of their similar form, "transactivity" is frequently confused with "transactive memory" (Moreland 1999), which is another popular term in group work. However, they are almost opposing concepts. Therefore, in order to reduce confusion in this dissertation, I use a new term ICC, which reflects the essence of the process and distinguishes it from "transactive memory". I argue that ICC, which is at the heart of collaborative learning, is also essential in a group work setting.

In investigating the process of idea co-construction and its role in collaborative learning environments, I ran studies three, four, and five. Studies three and four look at the role of idea co-construction more broadly from the level of knowledge co-construction. In study three, I raise questions related to design principles for offering interactive support within a collaborative setting, offering evidence that wholesale adoption of insights from the field of intelligent tutoring based on individual learning with technology is not warranted. In study four, I offer evidence that supporting group processes through dynamic feedback has a positive effect on learning. Both types of support in these two studies are ones that would allow for more frequent knowledge co-construction in collaborative learning settings. These two studies have helped spawn a new area of research that includes not just my own dissertation research, but that of others as well (Kumar, 2011; Walker, 2010).

In study five, I expand on previous research that examined constructs similar to the idea co-construction process, such as transactivity, in two ways, (a) investigating the role of the idea co-construction process on knowledge transfer between team members, (b) looking at engineering problems whose solutions require deeper knowledge integration across members of the group than the problems investigated in previous work.   Other researchers have verified the positive impact of transactivity in the context of well-defined problem solving tasks, or within a single domain (Azmitia and Montgomery, 1993; Tudge, 1992).   I argue that their findings could generalize to ad hoc teams because idea co-construction exchanges can mark places in a conversation where the integration of alternative perspectives is taking place. In addition, idea co-construction exchanges may point to places where members are acknowledging each other and new knowledge is constructed that did not originate within the individual minds of any of the team members prior to the interaction.

Establishing the relationship between the idea co-construction process and knowledge transfer requires the measurement of the knowledge integration process itself. Although using surveys for analyzing group process can yield a good measurement of perspectives, impressions, or reactions to group processes, these methods do not measure the prevalence of the process itself. Therefore, I apply a method that enables an in depth analysis of the process itself, namely conversational analysis.

Beyond this practical contribution, this dissertation research connects theories from collaboration to a work context as well. The concept of idea co-construction is one that has been the focus of much work in the computer supported collaborative learning community, and is new as a focus in a work context. My application of it to the Kane dataset later in this dissertation (study five) can be viewed as forging a new bridge between theories of collaborative learning and theories of group work, and there I provide quantitative evidence that this process that has long been valued within the collaborative learning community has a positive association with knowledge transfer and productivity in working groups.

2.3. Automatic assessment of conversation characteristics using natural language processing and machine learning

Once I verify the important role of knowledge co-construction contributions in engineering design discussions, instructors can look for them when they monitor conversational interactions. Because such monitoring is difficult given the instructor's limited vantage point with respect to group work, my dissertation work includes automatically analyzing the speech recorded from team discussions to detect various types of group processes, mainly focusing on the process of idea co-construction.

I build on recent efforts to support instructors in managing group work by offering them forms of automatic assessment and reporting (e.g., Soller & Lesgold, 2003; Kay et al., 2006; Pianesi et al., 2008). In prior work, researchers have looked at automatically detecting various aspects of student activities during group work (e.g., Kay et al., 2006; Pianesi et al., 2008). Multiple types of data have been used including message board postings (Kim et al., 2007), chat data (Soller & Lesgold, 2003), video (Chen, 2003), and audio (DiMicco et al., 2004). As a starting place, I look at message board data to detect the levels of student knowledge contribution. Although this preliminary work shows promise for increasing the baseline value of a correlation between measured level of student contribution and instructor assigned course grade from 0.16 to 0.63 (study six), this approach is not novel in that it only uses prior existing technology. The unique contribution of my research work is that I utilize recorded speech instead of text to predict group processes (studies seven, eight, and nine). I have chosen speech as the input modality because it is a more natural form of interaction than text.

An obvious first step in automatic assessment from speech, because of its relative ease of implementation, is detecting the total amount of speech[1] in the recorded utterances of individual speakers. Amount of speech has been used as an indicator of group meeting status. For example, DiMicco, Pandolfo, & Bender (2004) have looked at the amount of participation in meetings by displaying the ratio of speech contributed by each group member to the total amount contributed by the group in real time for participants in four-person group meetings. In response to the visualizations, DiMicco and colleagues observed a change in behavior of those who over-participated, or under participated in the meetings, by attending to an interface that displayed an indicator for the amount of speech relative to other group members. Similarly, volume of speech has been used as a quality indicator in a student project meeting setting as well. For instance, Chen (2003) displayed an estimate of each student's participation levels by automatically detecting the volume (loudness) of student speech. He suggested an intervention for those students who are less active in the class by observing the display speech activities. However, the effect of this intervention was not evaluated.

In addition to the amount[1] and volume of speech, which is a reasonable starting point for detecting levels of student activity, researchers have looked at other aspects of speech, such as pitch and energy levels, which can provide information about the nature of an interaction. Dabbs and Ruback (1987) show the usefulness of the style of speech in gaining insights into group work processes. They have analyzed the number of turns, pauses, and interruptions and showed that individuals who talk more are rated more favorably, as are individuals who pause more during their speech. Recent research has demonstrated the success of using sophisticated language technologies in addition to basic speech features for automatically tracking group activities effectively in the face of practical challenges. For instance, the CHIL (computer in the human interaction loop) project (Waibel, Steusloff & Stiefelhagen 2004) investigates the problem of automatically tracking the current speaker, detecting emotion, and transcribing speech. In the AMI (Augmented multiparty interaction) project (Renals, Hain, & Bourlard 2007), researchers

[1] Amount of speech is looking at the amount of contentful speech, excluding silence

have automatically identified speakers and segmented speech files collected during the meetings. The CALO (Cognitive Assistant that Learns and Organizes) project (Kaiser, et al 2004) also focuses on problems such as automatic speech transcription and topic identification.

Although using speech recognition technologies to automatically transcribe speech recordings and then applying models for prediction has been shown to be a promising approach (Ai et al, 2010), the state-of-the-art in speech recognition is still too poor to make this a viable option with raw speech as input. Current automatic speech recognizers are not robust enough (typical error rate 25~30%) to generate accurate speech transcriptions of group meetings (Stolcke, Friedland, & Imseng 2010). Some tutorial dialogue systems such as Scot (Pon-Barry, et al 2006) and ITSPOKE (Forbed-Riley & Litman 2009) have used speech recognition technology to detect uncertainty in student responses. However, neither system requires high accuracy of the content detection to perform this task. For instance, both systems use speech recognition to detect hedges (e.g. I think, I thought, maybe) or pauses to detect uncertainty in student responses.

Therefore, in this dissertation, I detect idea co-construction by using features related to the style of speech rather than using the content from speech recognizer output. Although the speech, used as the data source has been transcribed prior to the annotation process, the automatic analysis technique I describe does not use the transcriptions as input. Rather, the speech signal is first processed to extract features from segments of speech using basic audio processing techniques, which then become the representation used for classification by a machine learning model. For example, Ranganath et al. (2009) have used acoustic and prosodic features extracted from speech data to predict whether a speaker came across as flirting or not in a speed dating encounter. Ang et al. (2002) and Kumar et al. (2006) have applied a similar technique to the problem of detecting emotions such as boredom, confusion, or surprise. In addition, Liscombe et al. (2005) have applied the technique to the problem of detecting student uncertainty, which indicates whether students are unsure about their problem solving states. All of this work makes use of signal processing techniques that are able to extract basic acoustic and prosodic features such as variation and average levels of pitch, intensity of speech, amount of silence and duration of speech.

As used in this work, acoustic and prosodic features are frequently associated with intuitive interpretations. Therefore, they are an attractive choice for use in baseline techniques for stylistic classification tasks. For example, increased variation in pitch might indicate that the speaker wants to deliver his ideas more clearly. Likewise, volume and duration of speech may signal that the speaker is explaining his ideas in detail, and is presenting his point of view about the subject matter. Such interpretations are grounded in sociolinguistic work. Sociolinguists study how speech styles (Coupland, 2007; Eckert & Rickford, 2001) reflect intentional and subconscious aspects of the way that a speaker positions him or herself within an interaction at multiple levels. This work builds on decades of foundational work beginning with Labov's work on speech characteristics that signal social stratification (Labov 1966) and Giles' work developing Social Accommodation Theory (Giles 1984), which describes how speech characteristics shift within an interaction, and how these shifts are interpreted. A simplistic interpretation of this work would lead us to believe that hidden within the speech signal are features that enable prediction of social meaning. The work related to detection of flirting by Ranganath et al. (2009) supports this view. It can be argued that while the essence of the idea co-construction (ICC) process is related to content level distinctions, it also has a social interpretation, and therefore can be detected from speech as well. For example, consider that externalizations position students as intellectual leaders within a conversation. True leadership requires that the leader be received as such by the other group members. Because ICC contributions indicate the expected reception, then the occurrence of ICC contributions indicate that there is respect between speakers. It can then be expected that stylistic features that predict positive reception between conversational participants may also predict the ICC process.

In this dissertation, studies seven, eight, and nine address the technical challenges of using speech to detect the group processes reviewed in this section. In study seven (chapter six), I use activity levels in speech to predict the five assessment categories identified through the interview study presented in chapter three. Study eight (chapter seven) focuses on predicting the idea co-construction process using speech. In this study, I advance the area of automatic analysis of idea co-construction by identifying important aspects of speech dynamics that have predictive value with respect to idea co-construction (ICC). Finally study nine (chapter eight) incorporates ideas from sociolinguistics to gain more understanding about how a measure of the social nature of the conversation provides a predictor of the ICC process.

CHAPTER 3

Identifying valuable assessment categories[2]: Study 1

This chapter establishes a framework for assessing group work from an instructors' point of view. In order to determine what instructors value and look for when they conduct project group assessments, I conducted an interview study with instructors about their experiences conducting group work assessments. This chapter describes the findings of my interview study, as well as providing evidence of the value of the knowledge co-construction process in group work assessment. The key takeaways of this chapter and study are: (1) The knowledge co-construction process is valuable, including the idea co-construction process, and (2) Instructors find it difficult to assess the knowledge co-construction process from their vantage point. The difficulty instructors face in assessing knowledge co-construction is subsequently explicated with quantitative evidence in chapter four.

Borrowing from a basic interview study methodology, I asked the following two questions: (1) What types of group work processes are desired in a successful group project?; and (2) How can the group work processes be identified? I interviewed nine instructors who oversee group work, each of whom is an expert in evaluating group work processes. These instructors teach students how to collaborate in groups, skills that are desirable for successful group work at the

[2] This work is published in Gweon, G., Jun, S., Lee J., Finger, S., Rosé, C. (2011). A Framework for Assessment of Student Project Groups Online and Offline. Puntambekar, S. & Hmelo-Silver, C. (Eds.) Analyzing Interactions in CSCL: Methodology, approaches, and issues, Springer.

undergraduate and graduate levels. Transcripts of instructor interviews were then used to develop a framework for assessment of group work processes, which resulted in five pairs of assessment categories.

3.1. Motivation and Design

Identifying the specific interaction processes that are important to successful group work is necessary in order to achieve the goal of this dissertation, which is to aid facilitators in fostering collaborative environments. The goal of this instructor interview study is to develop an ontology of group processes that are believed to be causally related to group work outcomes. If instructors can provide support that fosters those group processes, students are more likely to learn from each other and produce successful solutions to difficult engineering design problems. The literature investigating the connection between group processes and outcomes of group work is both vast and diverse in focus, investigating factors such as communication, coordination, conflict, and conformity (Faidley et al, 2000; Fussell et al., 1998). However, those group processes that are most important to "trace" for the purpose of assisting instructors in creating ideal environments for teamwork have not been directly addressed in the literature.

By and large, previous research on these issues has been motivated by scientific questions, wherein only one or a small number of processes have been examined, usually to evaluate their effect on group work outcomes, such as the relationship between information exchange or learning and team innovation (Drach-Zahavy & Somech 2001). These kinds of scientific questions are usually related to social engineering questions, and these kinds of studies have not focused on a set of criteria complete enough to be of much use for instructors who are motivating the design of a technical solution for supporting facilitators. Other studies have even presented an ontology or categorization of group processes, such as Hackman's (1987) categorization of processes into task functions and maintenance functions, and although these kinds of studies are useful for revealing the relationship between certain group processes and outcomes, the selection of processes that have been investigated is not ideal for the purposes of this research. To address the more specific questions relevant to my research, an interview study with experts on group work was necessary in order to enable me to formally specify the type of group processes that are considered vital for group success.

I chose project course instructors as the target of my interview study because they are experts who regulate group processes as part of their professional practice. Insights about group work from experts can be used not only to better understand the important processes that impact collaborative work, but also to better observe how these processes are being nurtured. Experts from both educational contexts (instructors) and working contexts (managers) evaluate group work, which involves monitoring, assessing, and mediating group processes. However, there is a notable distinction between instructors and managers: the primary role of an instructor is to help group members develop optimal group-work practices through group tasks, whereas in a work context both managers and their employees usually think of evaluation as a necessary evil. Consequently, I selected the educational context as my research environment because evaluation and intervention are considered the norm and are received more positively than in working environments. The interviews were conducted with instructors and the goal was to identify those group processes that they believe have causal connections with group-work outcomes, identifying how such processes are operationalized by instructors. I interviewed both undergraduate and graduate-level instructors who teach students how to accomplish successful group projects. This interview study resulted in a framework that consists of five types of processes, which are detailed in the following sections.

3.2. Procedure

This section presents the interview method I employed for the study. I began each instructor interview with two focus questions. First, I asked instructors about the difficulties they face when they attempt to diagnose problems during group work. Secondly, I asked how instructors categorized the observations they make, such that they can make a general assessment of a group project, in such a way as to reduce the range of reported issues, difficulties, and reported practices into a manageable list.

The interview data was collected by three interviewers, who ran nine focused interviews with instructors. All nine instructors had taught at least three university-level group-project courses. These instructors where all from the same university and included two from design, two from social sciences, and five from engineering. Each interview lasted between 30 minutes and an hour. Due to technical difficulties, only six of the nine interviews were recorded, and all of the six recorded interviews were transcribed for further analysis. Background questions for instructors included the types of projects their students worked on, as well as the general characteristics of the students who participated in their courses. The purpose of background questions for instructors was to get a sense of what kinds of group work their students were doing. Next, instructors were asked to describe the syllabus for their course, including their course requirements and how they assigned grades. This information revealed what instructors regarded as important, since this is what they assessed, and what they wanted students to learn from the course. Questions about syllabi led to answers that oftentimes made specific mention of the procedures they used to assign grades, and their methods for assessing peer evaluations. I also asked questions about instances when problems arose in their student teams. Instructors responses included details about the causes, detection, and solution of problems in their student groups. Lastly, interviewers explained to the instructors that these interviews were meant to inform a reporting system, or summary tool, that was meant to help instructors get more insight into group-work processes. After explaining the objective of the summary report, the instructors were also asked what they would want from such a tool, and how they would imagine using it.

Each instructor interview was observed by at least two researchers, and after each interview the researchers typed notes and discussed their observations. After every few interviews, all three researchers consolidated their data and identified themes based on their respective notes. Multiple meetings allowed the researchers to balance their desire to discuss interview content while it was fresh in their minds with the competing desire to base conclusions on deep reflection and comparison of data across interviews, which often leads to a revision of earlier interpretations or the emergence of new themes. I employed this sort of iterative process in order to obtain meta categories of group processes that instructors look for as they evaluate student groups. From this iterative process, three meta assessment categories emerged: *learning goals*, *process*, and *product*.

To verify that these three meta categories sufficiently covered all interview data, the six recorded interviews were transcribed and segmented into sentences. For the six interviews, this yielded a total of 2320 sentences. The segmented sentences were then categorized for further analysis, which I refer to as "assessment categories." In order to generate these assessment categories, two of the researchers coded sentences related to what instructors wanted to know about the student groups. I excluded sentences on background information, rephrasing of interviewer questions, elaborations meant for clarification, and greetings. Next, to differentiate group-process types within the three themes, the researchers coded short, descriptive labels consisting of 3-5 words, wherein to the sentences identified as belonging to the 3 meta categories. The short labels were grouped to form 15 assessment categories. The resulting hierarchy of the 3 meta categories as well as the 15 finer-grained assessment categories that emerged is displayed as Figure 1 in the following section (3.3. Findings). Once the meta categories and the finer-grained assessment categories were determined, I conducted two types of coding to test their reliability. These coding rounds are referred to as the "assessment coding" and "evidence coding."

Assessment coding. Three rounds of coding occurred during the assessment category coding stage. In the first round, the sentences were each assigned to one of the three meta assessment categories to see if they could be reliably differentiated, and also to see how much data each of the meta categories covered. In the second round of assessment coding, two coders annotated the fifteen assessment categories to verify that they had been reliably coded under the three meta categories. Finally, in the third round of coding, two coders coded the five pairs of detailed assessment categories, under the meta category of *process* because that meta category is of interest for the purpose of gaining insight into the group processes.

Evidence coding. After "assessment coding," another round of coding was conducted, which I refer to as "evidence coding." For each of the five pairs of process assessment categories, the coders went back to the transcripts and identified indicators that instructors mentioned using in order to assess group work. For example, the extent to which students were each able to articulate what aspects of group work they were taking ownership of was used as evidence that students were equally dividing up their work. What is important here is the extent to which instructors rely on their intuition for group processes they are not able to observe directly.

3.3. Findings

Using a grounded theory-based approach to my analysis, I conducted an iterative coding process that resulted in five pairs of assessment categories and a list of indicators that instructors mentioned relying on to make those group work assessments. In addition to the processes, I also found that instructors value the learning goals they set for the course as well as the final product of group work. Figure 3.1 shows the detailed categories for *learning goals*, *process*, and *product.*

Between learning goals, process, and product, process is of greatest interest. Group difficulties can be revealed through interaction processes that display such things as amount of effort offered by group members or the characteristics of group dynamics. Learning goals, on the other hand, are set by instructors, and products do not show where and when the students are having difficulty in doing group work. Although in real work settings, it is mainly the success of the product that matters, instructors regard the process to be important for the purpose of giving students the opportunity to learn. By influencing the process, instructors have the opportunity to enhance both the learning experience as well as to facilitate the accomplishment of a higher quality product. By the time the product has been produced, and the learning objectives of the course have been accomplished, it is too late for the instructors to intervene. Finally, processes are also a more appropriate focus for a tool like this one, which is meant to be general across multiple disciplines. The same group processes are relevant in teamwork within the domains of design, behavioral science, or engineering. However, the learning objectives, as well as the group products, differ across disciplines and even within the same discipline.

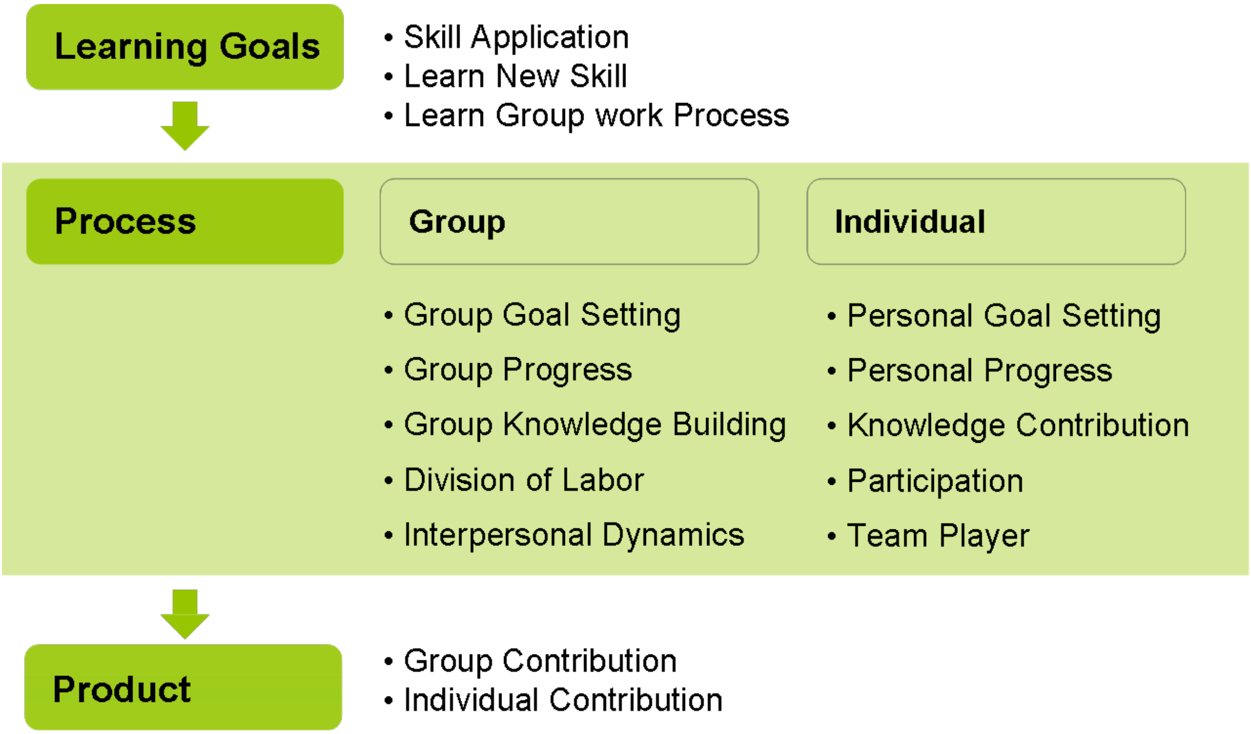


Figure 3.1. Assessment Categories

The importance of looking at process was evident in the data as well. Instructors mentioned assessment categories under process more often (70% of the instances) than under learning goals (15%) or product (15%). In addition, the number of more detailed assessment categories under the general heading of process (10) was higher than those mentioned under learning goals (3) or product (2) as seen in Figure 3.1. Given the importance and interest, I focus on the meta category of process rather than learning goals or product in the rest of this study. Note that the ten assessment categories under the general heading of process can be paired into corresponding individual and group level assessments. Individual assessments relate to an individual student in isolation, whereas group-level assessments relate to students in connection with their group or in comparison to other team members.

Table 3.1 shows the definitions of the five pairs of process assessment categories. Note that the coders verified that the three instructors whose data was not transcribed also mentioned the same types of categories, but were not included in the count due to unavailability of transcripts.

For each of the three rounds of assessment category coding, coders achieved the following kappa values. In the first round of coding, where two coders coded the three meta categories, a kappa agreement of 0.88 was achieved between two coders over 20% of the data. This kappa value is an acceptable rate of agreement. For the second round of coding, where two coders coded the 15 categories, a kappa value of 0.72 was achieved for the 20% of data. For the third round of coding, where two coders looked at the five pairs of assessment categories under process, the coders coded 20% of the data on the 5 assessment categories and achieved a kappa of 0.90. After calculation of the kappa in each round, disagreements were settled by discussion among coders.

Table 3.1. Five pairs of process assessment categories

Individual	Definition (# of instructors mentioned)	Group	Definition (# of instructors mentioned)
Personal goal setting	Making individual plans for next steps (4/6)	Group goal setting	Making team plans for next steps (4/6)
Personal progress	Fulfilling personally stated goals through producing work (5/6)	Group progress	Fulfilling group goals through producing work as a group (4/6)
Knowledge Contribution	Taking initiative to use knowledge or skill (1/6)	Knowledge co-construction	Exchanging skills, ideas, or conversations that lead to learning & project advancement (3/6)
Participation	Being involved in work (6/6)	Division of labor	Contributing work for the group relative to other group members (5/6)
Team player	Attitude toward interaction with teams (5/6)	Interpersonal dynamics	Interaction with the team due to personality & relationship (5/6)

After categorizing the list of processes that are used for evaluating group work, I conducted an “evidence coding” round to see what pieces of evidence are currently used by instructors in order to track the five pairs of categories under the meta category of “process.” The list of types of evidence mentioned by the instructors contained both directly observable and inferable evidence. Directly observable evidence is most visible, and therefore more straightforward to track. This issue will resurface when I discuss automatic assessment in later studies (studies six, seven, eight, and nine). For instance, the number of postings on a message board is directly observable, but inferences from conversations that take place in group meetings are not. Although inferable evidence is harder to track, it is as frequently mentioned as the directly observable evidence. Therefore, methods of detecting inferable evidence should be investigated in future studies.

Next, a table that summarizes each of the five pairs of *process* assessment categories is presented (Table 3.2). What follows from this are more detailed observations of each category, along with the pieces of evidence that the instructors mentioned in connection with each of the categories.

Table 3.2. Instructor identified needs and problems

Assessment categories	Example processes wanted by instructors	Example problems observed
Personal & group goal setting	Selecting own research methods and putting together own research plans	Spending too much time meeting without productive results
Personal & group progress	Steering and controlling process, checking in on accomplishments, keeping track of where they are in the project	Not meeting production goals; bottlenecks occur when part of the team is not delivering
Knowledge Contribution/ Co-construction	Sitting all together, physically close and being in constant communication with a tight feedback loop	Producing reports that are not united and have clearly separate sections
Participation/ division of labor	Contributing and presenting their work to teams	Supporting team members but complaining about the work
Team player/ interpersonal dynamics	Working together, collaborating with each other	A dysfunctional team, wherein members are not even talking to each other

Personal goal setting and group goal setting

The first pair of assessment categories is personal goal setting and team goal setting. Goal setting means making concrete plans for the project's next steps. For instance, instructors assessing personal goal setting might look for students selecting methods and putting together a project plan with explicit milestones. For group goal setting, instructors might examine whether the whole team is setting an appropriate goal. Also, having all of the team members buy into the same vision is important for group goal setting.

To assess student's personal goal setting, some instructors observed whether a student produced a list of tasks to accomplish using schedules or activity charts. Other instructors looked at publicly stated goals that each student made during weekly meetings. Some instructors were more explicit and required students to submit lists of tasks as well as report the time spent on each task. The frequency of the submission of such lists varied from weekly to monthly, depending on instructors' preferences. In addition to personal goals, instructors looked for team goals. One instructor mentioned that in order to see whether a team had a goal, she observed group meetings. If the meeting lasted too long or did not have any explicit agenda, that indicated the team should have made more specific plans. Instructors also looked at schedules produced by the group that show dependencies between their tasks to see whether groups are doing an effective job of coordinating across activities.

Personal progress and group progress

In addition to suggesting and providing goal setting help for students, instructors followed up to see if the students were fulfilling their stated goals. Instructors mentioned that they observed whether students fulfilled promises they made, whether students steered and controlled the process of their work, and whether they checked in about their accomplishments along the way. For group progress, instructors checked whether groups explicitly tracked progress towards the milestones they had agreed upon as a group. To assess personal progress, instructors looked at schedules to see whether planned items were finished on time and what action items were accomplished. For group progress, instructors observed scheduled team meetings, which varied in frequency from once a week to three times a semester depending on the course. In these meetings, instructors looked at students' presentations about what they had done so far as a group, and at the team's progress by examining the list or resolutions they had made. In addition to the meetings, instructors also looked at midterm and final presentations with similar assessments.

Knowledge contribution and group knowledge building

Knowledge contribution and group knowledge building is the next pair of assessment categories. At the individual student level, instructors observed students taking initiative by using use their unique knowledge and skills in group work. In addition, at the group level, instructors looked for evidence of group members mentoring other students, passing along skills, sharing ideas, and engaging in meaningful conversations that may lead to learning and project advancement. Note that not *all* conversational contributions would be considered instances of a knowledge contribution in the deepest sense. What instructors want to see is perspective sharing and perspective taking, which requires group members to display their reasoning to one another, and ideally for them to search for connections that allow them to build on one another's reasoning.

Overall, the pieces of evidence instructors used to assess knowledge contribution and group knowledge building were not very concrete or direct. In general, instructors looked more for evidence of breakdowns of these processes, rather than positive evidence of the occurrence of these processes. For example, instructors mentioned that one sign of trouble is when students come to talk to them about the absence of communication with other students in their group. Note that the same piece of evidence can be used in connection with other categories as well. For instance, absence of communication can be used as an indicator in the participation category. However, evidence such as attendance at group meetings can only be used for participation and *not* knowledge sharing. Knowledge sharing involves exchange of ideas and skills, and thus likely requires traces of communication as pieces of evidence. In addition to communication, another interesting indicator that instructors used for knowledge contribution is when the overall productivity of a group was low. An unintegrated work product, such as the patched report described in the introduction of this chapter, indicated group trouble for the instructor. To assess knowledge contribution and group knowledge building, it might be possible to infer knowledge building breakdowns from the available traces of group work. For instance, a low number of discussion threads initiated by students or a low number of replies by students on a group message board may indicate that communication is not active. Other sources of information, such as exchange of emails or number of group meetings, could also be used as additional indicators.

Participation and division of labor

Another pair of assessment categories is participation and division of labor. To assess participation, instructors observed whether each student contributed to the group effort, or whether some students were not working. If a student was not working in an obvious way, such as not attending group meetings or classes, instructors could easily detect such instances. However, instructors were also concerned with not knowing about students who seemed diligent, but who in reality were "slackers." Instructors also wanted to know what each student's contribution was in a given project in order to see whether the work was done by only a subset of the group members.

The importance of participation was articulated by all six of the instructors. To observe the assessment category of participation, instructors examined a variety of sources including self-reported work logs, peer-evaluation forms, and group message boards. From these various sources, they looked at student attendance in classes or group meetings, number of hours worked, number of action items accomplished, and the number of messages posted in group message boards. Indicators used for the assessment category of division of labor also came from the same sources as the individual assessment category of participation. Instructors used the same indicators, such as number of hours worked, and compared them to those of other group members to see whether the distribution of work is equivalent among members, across groups.

Team players and interpersonal dynamics

The last pair of assessment categories is engagement and interpersonal dynamics. Engagement is attitude towards participation in group work. Instructors looked for students that were dedicated, emotionally invested, and who loved their project. One instructor noted that when she listened to students presenting, she looked for engagement: "I think with experience I see when somebody is really honestly engaged with the work or whether they are faking it. So that's the first thing I look for.” The group level assessment category that relates best to attitude is interpersonal dynamics, which is interaction within the group resulting from the personalities and relationships within a group. Instructors observed that group chemistry was also an important factor in assesment; for example, were students were having difficulty getting along?

To assess students’ level of engagement, instructors observed their behavior. For instance, if a student created posts and avidly replied to other students’ posts, instructors inferred that the student was engaged. Another instructor noted that when students were enthusiastic about their project, they came to the instructor or other team members willingly for more work. When instructors saw students out in the field building things rather than just browsing the web for “research,” they inferred those students were actively involved in the project. Instructors also stated that although group dynamics is an assessment category they would like to gain more insight into, currently no good indicators exist other than spending time and being involved with the team.

However, language researchers provide an alternative perspective. Some researchers have inferred team dynamics by looking at the type of language used in the team. For instance, examining the usage of positive words versus negative words used by students in their correspondence (Pennebaker, 2008) may signal a particular pattern of group dynamics. Another way to infer team dynamics is to observe the cohesiveness of conversation by looking at the words used in message board or documentation produced by the group, where a low degree of cohesiveness may signal conflict between team members (Dong, Hill, & Agogino, 2002).

3.4. Discussion

During the interviews, I found a great deal of overlap in the types of problems instructors mentioned, which demonstrates a certain consistency in the gap between the instructors’ perceptions of the student groups and reality. Two sample stories, offered below, illustrate some of the problems that instructors reported experiencing in trying to assess group work.

One type of problem that was reported multiple times during the interviews was that students' contributions vary greatly within a team. This is a typical problem in groups and is oftentimes referred to as "social loafing" or "free riding" (Karau & Williams, 1993). The first story is from an instructor, here referred to as K, who taught a graduate level capstone design project class. To gain insight into how the groups were functioning, K asked students to perform a peer evaluation in the middle and at the end of the term. To K's surprise, group C rated one member of the group, here referred to as Brian, much lower than the rest of the team. K had not anticipated Brian's low score because Brian attended the weekly group meetings, wherein the group reported their progress to K. Members of group C had covered for Brian during the face to face group meetings with the instructor. K expressed dismay at not having had enough insight into the group work processes that occurred outside the weekly meetings. His limited view of what was happening in the group prevented him from detecting when a student was not pulling his own weight.

Another type of problem reported by a number of instructors was when groups completed their project through a divide-and-conquer approach; i.e., by splitting the whole project into parts to be integrated later, but then not communicating with each other while they were working on their respective parts. In one such situation, an instructor for an undergraduate level course on web design called S had a pair of students who were working on a project related to building and critiquing websites. The pair submitted all the required materials at each of the milestone points. Therefore, S did not see any problems with the group until they turned in the final report and product. The report was a conglomeration of two obviously separate parts, each with its own distinct writing style. As in the case of K, this problem occurred for S because the majority of group work was done outside of the instructor's view, and the instructor did not have enough insight into the group processes to detect the coordination problems so that he could have intervened.

As seen in the stories from instructors, one can see that instructors value the process of knowledge co-construction, where students share knowledge and learn from each other. In addition, this is an area where they feel they lack the ability to provide adequate assessment. In relation to the knowledge contribution assessment, most of the evidence currently used by instructors is indirect and somewhat nebulous. This provides a strong indication that instructors may value and benefit from additional insight into what happens during group meetings related to knowledge co-construction. Chapter four provides quantitative evidence that shows how the indirect evidence of knowledge co-construction creates a mismatch between the instructor's view of the inner-workings of their student groups and the view that can be obtained from within the group meetings themselves.

CHAPTER 4

Investigating current assessment practices: Study 2

In order to design and build a tool that is "useful" to expert facilitators (or instructors for the purposes of this dissertation), one should survey what type of information is important and valuable to expert facilitators and instructors. Therefore, after identifying the qualities of group work that are important for instructors (study 1), I conducted a field study in the classroom to better detect when instructors have difficulty making accurate assessments due to limited exposure to actual group work. This chapter describes a classroom study wherein I collected data from student groups working on an engineering project at the graduate level.

The classroom study addresses two primary research questions: (1) Are the group processes identified in the interview study observable, and can they be reliably tracked by human annotators in real time?; and (2) What evidence do we have of a need for support on the part of the instructors? As in the previous chapter, my focus here is broader than just knowledge co-construction. I demonstrate the importance of the knowledge co-construction process as one of the key areas that instructors can benefit from by having more insights into group processes. This both motivates the specific focus of the technical work I present later in this dissertation (studies 6 ~9) and paves the way for a broader research program that can be addressed in future research that's presently beyond the scope of this dissertation.

4.1. Motivation and Design

Important questions about how to provide instructors with information that would assist them in their assessment practices have not yet been adequately explored (Dimitracopoulou et al. 2004). Yet, such questions are as important as technical challenges. For instance, if an automatic assessment system captures information that is already apparent to the instructor, then such a tool would not be very useful. Therefore, in this study, I address research questions that provide design insights for focusing research effort where it is needed. More specifically, I am interested in what type of information should be captured in order for the tools to be "useful" to instructors.

The core of this research is an exploration into the problem of cognitive biases that influence assessment under conditions of incomplete and imperfect information. The literature on cognitive biases and how they influence social perceptions is vast, and I will not attempt an exhaustive treatment here. Instead, I have chosen two specific cognitive biases that are related specifically to social perception in conditions of incomplete information. I apply these in situations where the perceptions are specifically measured by people in higher authority positions on people in lower authority positions; for example, when instructors or facilitators are assessing group work. In particular, I focus on the "halo effect" (Thorndike 1920) and the fundamental attribution error (Ross 1977). I hypothesize that such errors would interfere with assessment by instructors and would result in a difference in perspectives among instructors and those with a more complete view of the group work environment. A demonstration of support for my hypothesis would be valuable in the design of assessment tools. First, it would provide motivation for researchers to then seek to build tools that address the root cause of difficulty, in this case, to help instructors to be less susceptible to these types of biases. Furthermore, as the exploration of the impact of cognitive biases on assessment practices is a relatively unexplored area, this foundational work would motivate a broader investigation into the impact of a wider range of cognitive biases on group assessment.

In order to address when instructors have difficulty in assessing student groups, I wanted to see whether instructor difficulties are large enough to warrant further investigation. I saw the need for such investigation in two types of observations, which are presented in section 4.1.1.

Following these observations, I identified principles from the social psychology of group work that explain these observed problems, which are presented in section 4.1.2. This attempt to apply concepts from social psychology to illuminate struggles faced by engineering project course instructors is a novel one. This aspect of novelty as well as the advantages of measuring such struggles in a real classroom environments are both presented in section 4.1.3.

4.1.1. Investigating the gap between instructor and student perspectives

My goal was first to investigate current instructor assessment practices and to detect when instructors are having difficulty. I measured the alignment between instructor and student perspectives in order to determine where there are gaps in perspectives. To this end, I conducted a classroom study over the course of a single semester. I collected assessment scores for students in a graduate level engineering course that was offered at a private university in the eastern United States. In this study, the clients were instructors of a capstone project course where students were working on a real engineering problem for an industry sponsor.

Assessment scores were collected from three different sources, from observers, from instructors, and from the students themselves. I refer to these assessment scores as observer assessment scores, instructor assessment scores, and peer evaluation scores, respectively. Two researchers observed weekly student group meetings related to the course project. The decision to observe group meetings was based on reported experiences of instructors who felt that much of group work occurs during meetings outside of class rather than in class (Gweon et al., 2011a). Their assessment scores allow us to estimate how much is missed by instructors from not seeing what happens during these group meetings.

I hypothesized that the correlation between observer assessment and peer evaluation scores would be higher than the correlation between instructor assessment and peer evaluation, because observers would have first hand exposure to group meeting observations and would therefore have a viewpoint that would be more in line with that of the students. For the observer and instructor scores, I collected ratings along five different assessment dimensions identified from the interview study (study one) and computed an average score for each student. I then compared these average scores to the peer evaluation score by conducting a correlational analysis. As I hypothesized, the data from the classroom study showed that the correlation between the average observer assessment scores and peer evaluation scores (0.79) was higher than the correlation between the average instructor assessment scores and peer evaluation scores (0.38).

This data from the classroom study is also consistent with some of the concerns expressed by instructors that I documented in my interview study (study one). One type of problem that was reported multiple times during the interviews was that students' contributions vary greatly within a team, but instructors often do not know about this unequal contribution until very late in the semester. For instance, remember instructor K, who taught a capstone design project class in chapter three (section 3.4).

Although the existence of a gap between the students' and the instructor's perspective might not be a surprising finding by itself, I am not aware of many investigations of this phenomenon supported by quantitative analysis of observational data, although Heller and her colleagues compared student and faculty perceptions of engagement in engineering, where she showed the difference of perspective from both parties using surveys (Heller, et al. 2010). Building on their research, wherein they looked at one type of process, namely engagement, I explore other types of processes. In addition, the measurements are collected as repeated measures throughout the semester in the form of weekly assessments. Although the data collected during this study is a contribution by itself, the more interesting research question is to ask why such a gap in perspectives might occur.

4.1.2. Two hypothesis; the halo effect and fundamental attribution error

Given that my data shows a greater gap between instructor and student perspectives than between that of observers and students, my research goal is to investigate whether principles from social psychology can be used to explain this discrepancy. If an explanation for the phenomenon is found, one can use this understanding to motivate the design of tools for instructors that might help them be less prone to blind spots in the assessment of group work.

The two well known principles of social psychology that I hypothesize may be applicable are the halo effect and the fundamental attribution error.

The halo effect. The halo effect is a cognitive bias leading people to perceive a person's traits in a way that is consistent with previous impressions of that person's other traits (Thorndike, 1920). The halo effect was first documented by Thorndike when he observed that supervisors seemed unable to rate their subordinates independently of the different aspects of their character. Other situations where it has been documented are when commanding officers are rating their soldiers, where a boss is evaluating employees (Beehr et al., 2001), where a customer is evaluating sales people (Lambart et al., 1997), or where a student is evaluating an instructor (Becker & Cardy, 1986; Feeley, 2002).

Based on this prior work, one would expect to find similar patterns in an assessment context. However, I am not aware of any prior research where the halo effect has been researched from the perspective of instructors evaluating students relative to their specific assessment goals. As in previously documented research, instructors are in a situation where they have to evaluate students based on their limited interactions with them. Therefore, I believe that, similar to other situations wherein evaluations occurs, instructors also may be strongly biased to rate students similarly on a variety of characteristics, including both those directly observed and those predicted, even when their stated desire is to differentiate between these characteristics (Gweon et al., 2011a). This raises important questions as to what types of assessments of student characteristics are affected most by this bias. Without these types of insights, it is not possible to predict exactly what types of errors in judgment instructors will make, or in what circumstances.

Previous studies that demonstrate the halo effect have reported different conceptual and operational definitions for measurement of this cognitive bias (Balzer & Sulsky 1992; Becker & Cardy 1986; Cooper 1981). This confusion between conceptual and operational definitions is partly due to the two different types of measurements that Thorndike used in his original study (Thorndike 1920). Therefore, in this classroom study, I measure and report both types of correlations that were used in the original study; namely, (1) intercorrelations among specific performance dimensions, and (2) correlations between overall ratings and specific performance dimensions. The specific performance dimensions used were based on what other instructors agreed were valuable in evaluating individual and group project performance (e.g., the framework from study one). Certain categories of these dimensions, for example the dimension of knowledge co-construction, which measures whether students are taking initiative to use knowledge or skills, might be difficult to assess without direct observation of group work. Therefore, I hypothesize that observers with direct experience observing group work may be less prone to the halo effect in making assessments compared to instructors, (i.e., group work overseers). If this hypothesis is correct, one would expect to see a higher correlation with the instructor's scores compared to those of a direct observer of group work.

Hypothesis one: Group work overseers who are not direct observers of group work make more errors of judgment consistent with what is predicted from findings related to the halo effect than direct observers do.

If the results from this study support this hypothesis, we can gain deeper insight into how this effect plays out in assessment of group work and better explore whether having more of a direct view into group work can lessen the halo effect.

Fundamental attribution error. The fundamental attribution error is the frequent tendency for people to underestimate the effect of context, and to overestimate personality based influences when explaining behaviors observed in others (Ross, 1977). Although limitations and difficulty in determining the accuracy of the fundamental attribution error have been found (Harvey, Town & Yarkim, 1981), it has been documented across multiple environments; generally, the fundamental attribution error is considered to be a robust phenomena. For example, Cook and Klumper documented the effect of the fundamental attribution error regarding perception of leadership. They argue that although individuals can command authority and demonstrate leadership skills in certain circumstances, the skills do not necessarily transfer to other situations (Cook & Klumper, 1999). Similarly, others have shown the effect of the fundamental attribution error where the Board of Probation and Parole evaluated parole cases for offenders (Carroll, 1978), and where students evaluated teachers (Kelsey, et al., 2004; McPherson & Young, 2004).

However, like the halo effect, I am not aware of prior research on the effect of the fundamental attribution error from the perspective of instructors evaluating students in a classroom setting. As in previously documented research, I expect instructors to underestimate the impact that their presence has on the behavior of their students in the context of their direct experience with them, and thus imagine instructors would make strong assumptions about how students behave in group work settings that don't necessarily carry over into that context. More specifically, instructors are vulnerable to the fundamental attribution error because they have a limited vantage point, wherein they only see the students in the context of class sessions or check-up meetings where students display their best behavior. Certain students may "put on a show" at these meetings in order to appear to be taking on leadership in their group.

An observer who attends group work sessions would not be subject to errors in judgments about group work because they see the students performing in the actual work context. Thus, I can test the extent to which instructors are subject to the fundamental attribution error by comparing their perceptions of group work to that of the students, then seeing whether the difference in perception is greater than the comparison between the observer and student perspectives. The fundamental attribution error predicts that the further removed someone is from directly observing behavior, the greater their errors of judgment should be. This being the case, one would expect to see more similarity between the perspective of observers, and that of the students themselves, than in comparing the perspective of the instructor and that of the students. Thus, the second hypothesis:

Hypothesis two: Group work overseers who are not direct observers of group work make errors of judgment consistent with what is predicted from findings related to the fundamental attribution error.

Beyond seeking evidence to support or refute these hypotheses, the concrete and multi-dimensional operationalization of perspective on group work developed and employed in this work is valuable. For example, it allows the measurement of the extent of errors in judgment relative to different types of assessment goals. Thus, this investigation has the potential to yield rich insights into the inner workings of group work overseers' assessment practices.

4.1.3. Importance of investigating the problem in an actual classroom

Investigations of instructor's cognitive biases in real classroom settings are sparse. Identifying such errors from instructors in a real assessment setting is essential in order to observe how those errors actually interfere with assessment. We need that information if we want to design assessment tools that address the root cause of group assessment difficulties. In this case, that would mean helping facilitators to be aware of such types of errors with the hope of helping them overcome them as much as possible.

The cognitive biases that I am investigating are applicable to instructors and facilitators in various environments, not just in educational settings. For example, two common environments where a facilitator oversees a group is in a work place or in a classroom setting. In this chapter, I begin the investigation of cognitive biases in a setting that bridges the workplace and the classroom, namely with a capstone project course at a university where the client is not hypothetical, but real. In such courses, groups of students work under the supervision of an instructor who acts as a group facilitator, but the project is real. As introduced in the introduction chapter (chapter one), problem (and project) based learning (PBL) courses are popular in engineering courses due to their educational merits. Students get a chance to work on real world problems where they can integrate and apply knowledge acquired in more theoretical courses (Dutson et al. 1997; Rohde 2007). However, despite much benefit in participating in capstone project courses, student groups do not always function in an ideal way and need guidance. As a result, numerous efforts to support PBL resulted in innovative software tools for both individual and collaborative creative design (Guzdial *et al.* 2001; Kolodner *et al.* 1998). For example, the Progress Portfolio (Loh *et al.* 1998) is a software application designed to support student's long-term PBL activities. It assists students in documenting their work and supports reflection on the learning process with tools for screen capture, annotation, organization and presentation. Scardamalia's CSILE/Knowledge Forum approach also supports PBL, knowledge-building, and the aggregation of student progress in problem solving around a given learning domain (Scardamalia 2004).

Although such tools and methods have advanced PBL practices, most of this work targets students as users rather than teachers. In my interview study (chapter three), instructors expressed a need for assessment tools that would help them capture parts of students' group work that they do not have access to. For example, they raised concerns about group problems being intentionally hidden behind the well-functioning part of groups until the final weeks of the semester, or until conflict became irreconcilable, partly because students worry about breaking trust within the group (Gweon, et. al 2011a). Others have expressed similar need for teacher support (Dimitracopoulou 2004; Soller 2005).

Consequently, I conducted a field study in a classroom in an attempt to better detect when instructors have difficulty making accurate assessments because of their limited exposure to actual group work. Data from field study contexts tend to be messy compared to data from laboratory studies, and can yield only correlative results. However, a field study is more suitable than a laboratory study for my purposes because assessing a simulated group is certainly different from assessing an actual student group, wherein relationships between the instructor and group develops over time. In addition, different types of group problems occur over time in real groups, and simulating multiple problems in short laboratory sessions would result in unrealistic groups and group behavior. Moreover, simulated groups may not experience common problems that "real" groups experience, such as interpersonal issues or cliques. Additionally, conducting a laboratory study with actual instructors, rather than participants who sign up for an experimental session, is difficult due to a lack of availability (e.g., most instructors are too busy teaching to spend time simulating teaching). Because assessing student groups requires experience and expertise, conducting the study with non-experts as assessors would also not address how to find what expert instructors are "missing." Most importantly, a laboratory study would not answer the question regarding the specific context I am investigating; for example, although participants in a laboratory study may detect group problems in a simulated group, this does not mean that instructors would also detect the same type of problems in a classroom setting. A classroom environment is noisier in that many more variables are in play. For example, in a simulated group, members do not have a vested interest in maintaining their long term relationship. Therefore, detecting the problem that students may hide in order to avoid breaking trust with other group members is probably less likely than in "real" groups. Thus conducting a classroom study would give us knowledge of the types of problems that instructors have difficulty assessing "on the scene," given the noisy setting. In addition, this is exactly the type of data needed for deriving design principles for tools that support assessment, which was identified as an important research goal at a recent computer supported collaborative learning symposium (Dimitracopoulou 2004).

4.2. Procedure

The course that provided the context for my data collection effort was a graduate level engineering course offered spring of 2008. The whole group of students worked together on a common project sponsored by a corporate client. Four subgroups of students were formed to carry out the project. Because the class is a project-oriented class, a major component of the grade assigned by the instructor is based entirely on their productivity, and this portion of the grade is explicitly indicated by the instructors, separate from the part of the grade related to the quality of their result. There were two instructors and 22 students in the class. Data collected from this class included messages on discussion boards, reports, and weekly work logs from each student. However, this information alone was not enough to address the gap between what instructors would like to know about the groups they are overseeing, and what they actually see. In order to get a clearer picture of what information instructors were missing, I instrumented the course in order to collect extensive observational data from the four groups of students. Specifically, I collected audio recordings of group meetings as well as video tapes of their classroom activities. The semester long course was divided into three phases. The grouping of students into subgroups changed for each of the phases. Each phase lasted five instructional weeks. Although data was collected throughout all three phases, the most usable data was collected during phase two. Data collection during phase one was used for calibrating observer and instructor instruments. Data collection for observers during phase three didn't occur frequently because the students didn't hold many structured meetings at the end of the semester. Therefore, the instructor and observer assessment data used in this chapter is based on data collected during phase two.

As discussed in the introduction of this section (4.1.1), three types of data were collected: *observer assessment scores*, *instructor assessment scores*, and *peer evaluation scores*. In order to collect data for this analysis, I asked the observers and instructors to make weekly evaluations of students in the five areas of the assessment framework that was established through the interview study described in chapter three.

Again, my observers were researchers who attended group meetings, but did not participate in the group work. Two researchers attended weekly group meetings and evaluated the five pairs of assessment categories identified in the first stage of this study. Group meetings were chosen as the target of the observation since the bulk of the project work was accomplished during group meetings, although instructors are not able to attend group meetings due to time constraints. The two sets of assessment categories scored by the researchers were used to calculate a reliability measure, which would show whether the assessments can be made reliably from these observations. In observing the group meetings, the researchers remained uninvolved in the group meetings, as "flies on the wall." The main goal of this method is to observe the environment and the social interactions as they occur without influencing the participants. In order to achieve this goal, the students were assured that their grade would not be affected in any way due to the presence of the researchers and that there would not be any communication regarding the group meetings between the researchers and the instructors. Analysis of our data included group members' negative discussions of instructors and, as promised, both the observers' assessment scores and the recordings of the group meetings were kept confidential and never given to the instructors.

Unlike the observer and instructor assessments, the peer evaluation was conducted once, two weeks after the class ended. Although conducting the survey multiple times would result in richer data, one motivation for waiting till the end of the semester was to avoid having students report dishonest evaluations or "gang up" against one unpopular student, which would result in biased data (Byard 1989). But more importantly, multiple evaluations may be socially awkward and take time away from regular group work activities (Falchikov 1995). I used this peer evaluation score as the standard and compared assessments made by instructors and observers to those of the students. I acknowledge that, as with any type of assessment, student assessments can also be subject to potential biases. However, the decision to use peer evaluation as our standard was based on an assumption that, compared to instructors or observers who conduct evaluations based on their observations, the students who are doing the actual work would know more about the details of how much each student really contributed to the project.

4.2.1. Instrument for measuring observer perspective

The first instrument developed was for use with direct observers of project group meetings. For each of the five group process dimensions displayed in Table 4.1, I went back to transcripts to study the data. From that data I identified different questions that instructors used in association with each dimension. Using those questions as a foundation, I constructed statements that described the positive and negative student behaviors. Table 4.1 shows a sample question for each category.

Table 4.1. Sample questions used for the 5 assessment dimensions

Dimension	Sample question
Goal Setting	Is the student suggesting next steps (plans, high level steps) for himself or for the team?
Progress	Were there items that he finished during the past week?
Knowledge Co-construction	Did the student present new ideas or solutions for problems being discussed during the meeting?
Participation	Does the student seem engaged in the meeting by giving full attention to the meeting itself? Ex. Not playing games, not checking email constantly.
Teamwork	Did the student respect others' opinions by allowing them to speak/ respond?

These questions were used by the observers attending group meetings to make assessments by responding to them with yes or no answers. Based on the pattern of yes and no answers related to a category, coders then assigned an overall score for each dimension with a number between "-2" and "2". The range has both negative and positive numbers so that the scorers can easily map negative behaviors to negative scores and positive behaviors to positive scores. The five point range is also convenient in terms of its correspondence with typical grading scales (e.g., A~F).

The observers evaluated each student on the five assessment dimensions each week. For each meeting attended, observers evaluated each student on all five dimensions of group process identified in the framework. In addition, an average of these five scores was computed. Because students did not hold formal meetings during phase three, observer data during this phase could not be collected. Therefore, I have phase one and phase two scores for observer assessment scores.

The reliability of the coding scheme for the ten categories was verified by calculating the correlation between the scores assigned by the two researchers, which was 0.81 for the group level categories and 0.64 for individual level categories. I refined the assessment instrument during the first third of the course. In the second third of the course, I used the instrument to record assessments for scheduled group meetings. In the final third of the course, the observers continued to attend scheduled group meetings, however the frequency of such meetings dropped considerably during the final third of the course, until they eventually ceased altogether in favor of impromptu small group meetings that occurred on as-needed basis.

4.2.2. Instrument for measuring instructor perspective

Once the instrument for quantifying the observations of direct observers based on the five assessment categories was developed, I constructed an isomorphic instrument to be used by instructors to track their assessments of students throughout the semester. The questions were completely isomorphic to those on the observer instrument (Table 4.1). For example, the knowledge contribution question for the student is isomorphic to that of the instructors; i.e., “Did the student present new ideas or solutions for problems?” Note, too, that the only difference with this sentence is the removal of the phrase “being discussed during the meeting.”

The instructor evaluated each student on the five assessment dimensions each week. Because the instructor evaluation started in phase 2, I have phase 2 scores and phase 3 scores for each of the five dimensions on a weekly basis. These weekly scores were then averaged for each phase. Instructors continued to assess students during phase 3 because unlike direct observers' assessments, the instructor assessments were not based on any actual observations of group meetings. Again, the averages for each phase were also computed for instructor scores.

4.2.3. Instrument for measuring student perspective

Access to students was even more limited than the instructors in terms of validating an instrument for measuring perspective. I developed a questionnaire that was patterned after typical peer evaluations used in project courses, which was separate from the instruments that the observers and the instructor used. I used a simple questionnaire that consisted of one question that the students were required to answer separately for every student in the class including themselves. The question required them to rate every one of their classmates from -2 to 2 in terms of their level of contribution to project work, regardless of whether those students were in their team or not. 15 out of 22 students responded to the survey. The average of the scores that a student received from other students was calculated so that each student would have a single peer evaluation score.

4.3. Findings

The instruments described in the previous sections allow us to compute through a correlational analysis the degree of match or mismatch between instructor and observer perspectives on relevant judgments about the quality of student participation in their project groups. The pattern of match and mismatch is consistent with both of the hypotheses, as described in this section.

4.3.1. Hypothesis one

Group work overseers who are not direct observers of group work make errors of judgment consistent with what is predicted by the halo effect.

One can evaluate the extent to which hypothesis one is consistent with the data by comparing correlations across dimensions of the observer scores and the instructor scores. I ran two types of analysis that Thorndike used in his original study (Thorndike 1920): 1) intercorrelations among specific performance dimensions; and 2) correlations between overall ratings with specific performance dimensions.

The intercorrelation score is computed using correlations among differing dimensions, in this case; between the five assessment categories identified from the interview study in section three. Firstly, because the halo effect is the tendency to rate different performance dimensions similarly, the intercorrelation score from a person that is more affected by this cognitive error would be higher. In the classroom study, because direct observers do not have an instructor-student relationship with the students in the groups they observe, they may be less likely to strongly associate the separate characteristics together. Secondly, direct observers have first hand experience observing the students in their groups. Taken together, I hypothesize that the direct observers, who are themselves not part of the working group, may be less susceptible to the halo effect, which means that they should be able to differentiate across the various assessment categories even better than instructors can. If so, the correlation between those categories of observer scores should be lower than those computed from instructor scores.

Figure 4.1 shows three types of correlation scores between four of the assessment categories; namely, goal setting, progress, knowledge co-construction, and participation. The assessment category of "teamwork" was excluded from the analysis due to insufficient variance among the scores to compute a correlation. Each set of three bars correspond to a pair of assessment categories and represent the following: the bottom bar is the correlation of observer scores between the two dimensions during phase two, the center bar is the correlation of the instructor scores between the two dimensions during phase two, and the top bar is the correlation of the instructor scores between the two dimensions during phase three. Figure 4.1 shows that the center and the top bars in each set, which represent the correlations of instructor's scores, are highly correlated. In contrast, the bottom bars, which represent the observers' scores, do not demonstrate such a high correlation. This is consistent with hypothesis one as evidenced by the higher correlations of the instructor scores. The data shows that the instructor has more difficulty in differentiating among the various assessment categories, which may be explained by the presence of the halo effect. I cannot draw strong conclusions about the trend between phase two and three because of the limited data during phase three. However, by comparing the phase two instructor score correlations (center bar) to those of phase three scores (top bar), I suspect that the halo effect may actually worsen as the semester progresses, as evidenced by higher correlation scores of the phase three scores.

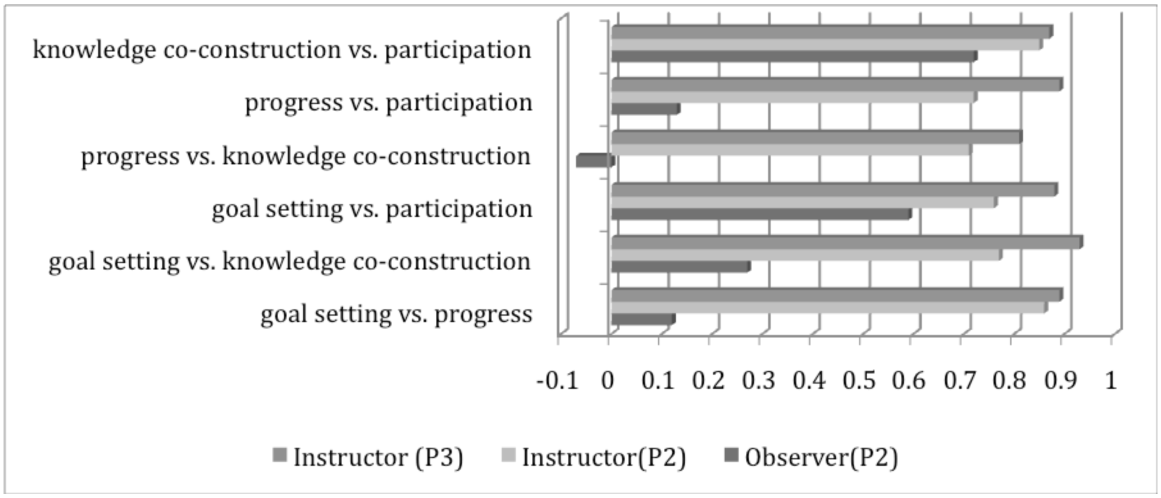


Figure 4.1. Correlation among various assessment categories

The second type of measurement that is indicative of the halo effect is the correlation between overall ratings with specific performance dimensions. Only the instructor overall ratings are available, which are the actual grades that the instructor assigned to students for the class. The correlation scores between overall ratings and the five dimensions for phase two and three are r = 0.56 and r = 0.48 respectively. This is comparable to the correlations mentioned in Thorndike's research, which were in the range of r = 0.51~ 0.64 (Thorndike 1920). Thus this second correlational measurement also supports hypothesis one.

4.3.2. Hypothesis two

Group work overseers who are not direct observers of group work make errors of judgment consistent with what is predicted from findings related to the fundamental attribution error.

The fundamental attribution error predicts that group work overseers may make errors in judgment by attributing a student's behavior or performance to the group environment rather than the student's own individual capabilities or motivations. Thus, a high correlation between the group's score and individual's score could suggest that the overseer is influenced by the group's performance when assessing the individual student. In order to test the second hypothesis, I first compared the correlation between the group score and the individual student scores for each group of students across the five assessment categories. The group score was computed by calculating the average of all members in a given group. Figure 4.2 shows the resulting correlations. As predicted by the hypothesis, the correlation associated with the instructor's score was higher (r = 0.40) as compared to the one based on direct observer's score (r = 0.23).

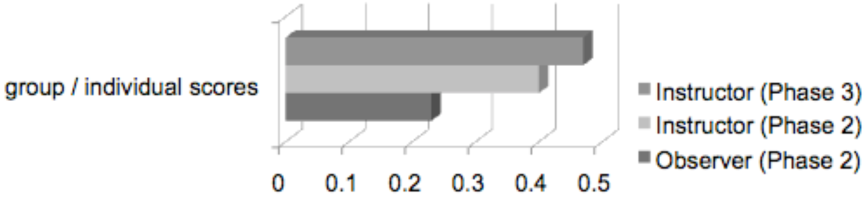


Figure 4.2. Correlation between group and individual scores

In addition, the correlation of the instructor's score, which was based on a later phase of the class, was even higher (r = 0.47). As with the hypothesis, I cannot draw strong conclusions about the trend between phase two and three because of the limited observer's data during phase three. However, from the comparison of correlations of the instructor score for phase two and three, I suspect that the degree of fundamental attribution error may worsen as class progresses.

Another statistic that could signal a fundamental attribution error during assessment is the variance of the individual's scores. If the overseer attributes much of the individual's performance to the group performance, the scores of the individuals in the same group would be similar, resulting in low variance. Indeed, the variance among the students in a given group was 0.43 for the observer's assessment scores, and 0.30 for the instructor's scores. Thus, this second analysis also tells a similar story, which is consistent with hypothesis two.

4.3.3. Additional Observations

After collecting observer and instructor assessment scores along the five dimensions, I computed a correlation of these scores with the peer evaluation scores. The five group processes are ones that instructors identified as valuable assessment criteria in the interview study; namely, goal setting, progress, knowledge co-construction, participation, and teamwork. The top row of Table 4.2 displays the correlations between the observer assessment scores and peer evaluation scores. The average was computed by taking an average value of the five dimensions for each student before calculating the correlation. The bottom row of Table 4.2 displays the correlations between instructor assessment scores and peer evaluation scores. Note that the correlation between peer evaluation scores and observer scores along the dimension of teamwork could not be computed due to insufficient variance.

Table 4.2. Correlation between student's survey score and observer/instructor across the five different dimensions of framework.

	Goal setting	Progress	Knowledge co-construction	Participation	Teamwork	Average
Observer & student	0.40	0.01	0.77	0.85	.	0.79
Instructor & student	0.36	0.42	0.08	0.32	0.36	0.38

As hypothesized, the data from the classroom study showed that the correlations between observer assessment scores and peer evaluation scores were higher than the correlations between instructor assessment scores and peer evaluation scores. The high correlation between observer scores and peer evaluation scores gives us some confidence that both the peer evaluation scores and the observer ratings are capturing something veridical. Therefore, one can interpret the low correlations between instructor scores and peer evaluation scores to be an indication that there's something important that instructors are missing by not attending to the different dimensions of assessment categories because they do not have an opportunity to observe group work directly.

The many low correlations in Table 4.2 showed that the perspective of instructors, observers, and students are indeed different. In addition to this main finding, the occurrences of some key high correlations tell other interesting stories as well.

Additional observations about the data come from the different perspectives of instructors and students. When students rate group work, they tend to focus on phenomena that are immediate, such as amount of knowledge shared, participation, or teamwork. Dimensions that require evaluation of a bigger picture are harder to evaluate from a novice's point of view. However, instructors attend to the bigger picture and assess students at a high level. Unfortunately, because of this, and the fact that they have limited opportunities to directly view group work, they are more prone to the fundamental attribution error and tend to make assessments based on the high level phenomena that they have access to. However, direct observers are trained to focus on making assessments of students on each individual dimension, without looking at the bigger picture, and are thus less prone to the fundamental attribution error. The ideal perspective would combine both.

Correlations between instructor ratings and the student ratings are uniformly low, and only in the case of progress is the correlation between student and instructor scores higher than that of the correlation between student and observer scores. This exception could be predicted because the progress dimension is heavily focused on tangible work products, which instructors have firsthand exposure to, as well as reports, which instructors use for progress updates.

The high correlation between student scores and observer scores gives us some confidence that both the student surveys and the observer ratings are capturing something veridical. Therefore, one can interpret the low correlations between instructor scores and student scores to be an indication that there is something important that instructors are missing by not attending to different dimensions of assessment categories because they have not had the opportunity to observe group work directly.

This analysis was also useful for identifying the dimensions where instructors are more likely to make errors of judgment. In Table 4.2, one can see that the largest instructor error occurs in the knowledge co-construction dimension, where the correlation for observer/student is 0.77, and that of instructor/student is only 0.08. This points to an opportunity for supporting assessment with technology that can better monitor group processes related to this dimension.

4.4. Discussion

By presenting an assessment model and using it as a tool to assess group work in a classroom setting, I showed the usefulness of the model by demonstrating how it can be used to explain assessment problems that observed in a classrooms study. The findings from the classroom study can be used for effective group assessments in two ways. In this section, I discuss: 1) Future use of the instrument used for measuring the observer/ instructor perspectives,; and2) Design implications for an automatic assessment tool.

4.4.1. Future use of the instrument for measuring group work

The instruments that were used for measuring group work processes in the classroom study could be used for assessing group work beyond the scope of our research. The instrument used by observers is targeted for assessors who have direct access to student group meetings, whereas the instrument used by instructors can be used more broadly by group work facilitators in various settings. However, one could potentially also use the same instrument in both kinds of contexts because the five assessment dimensions are broadly defined. On the other hand, further research in a broader range of settings would be required to verify whether the instruments are general enough to be used across different cultures, or across various types of project-oriented working groups. The full set of questions for each category used in this study is presented in appendix one.

Several observations were made during the group coding process that might be reflected in the next version of the assessment instrument for observers and instructors. First, assessing all ten categories was too much of a burden for so few coders. One possible direction for the future is to limit the focus to individual level assessments. I expect that even if insights are offered at the individual-level categories, by seeing which individuals are in trouble, teachers may be in a better position to allocate their time effectively, and to do more detective work to determine whether the individual performance would affect the group in each problem case.

A second observation made was that although the coders felt that they had a harder time assigning the goal setting and group progress assessment categories than the knowledge building, division of labor, and interpersonal dynamics categories, the correlations of scores between the coders were actually higher for goal setting and progress ($r = 0.92$), as opposed to the latter three ($r = 0.75$). This indicates that the confidence that instructors may feel about their judgments related to the different assessment categories does not predict the reliability of such judgments.

A third, and possibly the most important, observation was that depending on the type of the meeting, the type of assessment categories that were observable differed. The two main types of meetings that the students held were administrative meetings wherein most of the discussions were related to handling administrative matters such as scheduling and work assignments, and work meetings when the bulk of the meeting time was spent doing actual work such as building a conceptual framework for the project. For the administrative meetings, goal setting and progress assessment categories were easier to observe, whereas almost no knowledge sharing occurred, making it hard to assess the knowledge building. However, for the work meetings, knowledge building was easier to assess whereas the goal setting and progress behaviors were rarely discussed. For the purposes of automatic assessment, one would have to explicitly take the type of meeting into account when using data collected from the meetings to make assessments, taking into consideration what assessment categories we can get a reliable assessment for depending upon the type of meeting.

4.4.2. Design implications for an automatic assessment tool

In this section I discuss how the classroom study findings could be used for designing tools that provide effective group work assessment. Deriving design principles based on empirical study has previously been identified as an important research direction, most recently at a computer supported collaborative learning symposium (Dimitracopoulou 2004). This classroom study presents data that pinpoints where discrepancies lie, as well as the magnitude of those discrepancies. More specifically, this study has shown that such instructor discrepancies may be explained in part by two psychological phenomena that contribute to errors in judgment, namely the halo effect (Thorndike 1920) and the fundamental attribution error (Cook & Klumper 1999). These findings indicate that we should draw attention to differences between distinct assessment dimensions.

Ultimately, we ought to automatically trace the five pairs of assessment categories based on the findings from this study. In the remainder of this section, I present a summary report that the instructors can use to categorize and diagnose complex problems in assessing group work. The research questions that should be addressed in building such a summary report for instructors are related to issues addressed in the conversation visualization community (Shneiderman, 1992; Smith & Fiore, 2001). Similar to the research goals addressed in this community, my vision on building a summary for instructors is about displaying information inferred largely from conversation data. In addition to the display of representation of the conversations, what is learned about the groups using that data is important as well. Existing work such as the Wattle tree (Kay, Maisonneuve, Yacef, & Reimann, 2006) and Group Awareness Widgets (Kreijns et al., 2002) have attempted to addressed this issue. For example, using the Wattle tree, one can infer some group processes, for example, the relative amount of each student’s contribution, by examining the number of commitments to work activities, number of files committed to a code management system, or amount of time spent on some aspects of group work. What is different from the vision of my work is that these tools present the evidence that instructors use to make an assessment, rather than making and displaying the assessment itself at the level of granularity that instructors mentioned in the interviews.

I have developed an experimental testbed and reporting infrastructure, the Group Learning Assessment Platform (hereafter GRASP) to better illustrate the kind of reporting that my work is meant to facilitate. GRASP is a technology designed to enable unobtrusive, real time assessment of group dynamics from digital recordings of student speech. Figure 4.3 illustrates the GRASP framework. In stage one, students carry headphones and mini digital recorders with them so that whenever their project teams meet in any location, they can record their meetings, with the speech from each student recorded in a separate file. In the next stage, speech is preprocessed to create a representation that consists of feature-value pairs. Next, I compute indicators from speech recordings, related both to an individual's participation in a meeting and to a group's overall successs in meeting together, computations that are based on comparisons of the speech for the individuals present in the group meetings. Using regression models that are trained over speech data, paired with human assessment ratings, I am able to make automatic predictions about how humans would rate group meetings along selected assessment dimensions. Finally, these automatic predictions can be displayed for an instructor to observe.

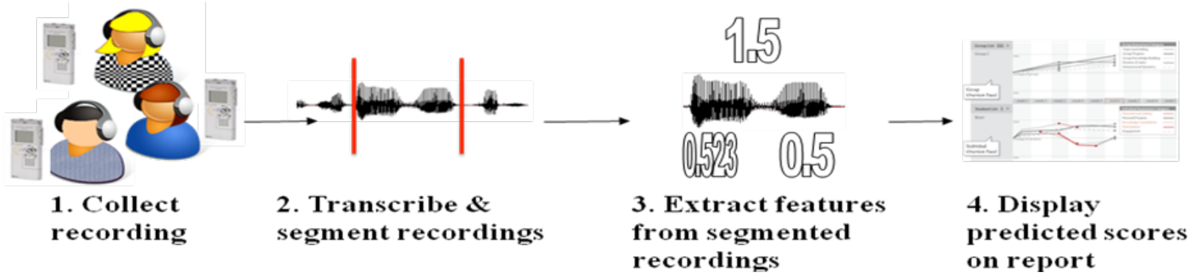

1.5
0.523
0.5
1. Collect recording
2. Transcribe & segment recordings
3. Extract features from segmented recordings
4. Display predicted scores on report

Figure 4.3. Overview of the four stage automatic assessment process

The display of GRASP (shown in figure 4.3, step four), is what I refer to as a "browsable summary." A browsable conversation summary presents a digested view of an on-going, potentially multi-party conversation; for example, it can present the high points of an extended conversational interaction. Such a conversation summary might make it possible for one to get the gist of a conversation at a glance without having to spend the time to actively participate in the conversation as it was occurring or listen to the entire conversation after the fact. In addition, these browsable summaries would present the user not only with a summary at the most abstract level, but also with links that allow the user to drill down and explore the data on multiple levels of abstraction based on their interests and needs.

This system makes an assessment about the level of student productivity over time based on a model trained to features extracted from student conversation behavior within a groupware system. Such a summary enables instructors to get a sense of how an individual student or groups of students are doing without having to read *all* the messages posted in the groupware system. In this sense, the summary I am presenting is not an "informative summary," which would be merely a summarization of the content. Rather, it is an "indicative summary," which presents meta information or indicators of characteristics of the interaction, such as the style or the manner that various types of conversational processes, such as grounding, planning, negotiating, brainstorming, etc., play out.

Figure 4.4 shows the display of GRASP in more detail. An important feature of my prototype is that its interface consists of two main panels; the group overview panel (upper half) and the individual overview panel (lower half) in Figure 4.4. Displaying both views simultaneously helps overseers compare the degree to which group's performance might affect the assessment of an individual, and responds to the observation that group work overseers are susceptible to the fundamental attribution error.

The first hypothesis from the classroom study showed that instructors who did not have direct access to group work were more prone to the halo effect, resulting in similar scores across differing assessment categories. Therefore, in order to dampen the halo effect, the assessment tool should make an explicit distinction between different assessment categories, providing information related to each category separately, so that the overseer would be better prepared to differentiate among these various dimensions. As seen in Figure 4.4, this design recommendation is implemented by showing the progress of each assessment category using a separate line. The y-axis of the graphs are goal setting, progress, knowledge co-construction, participation, and teamwork respectively. The x-axis of the graphs are the number of weeks in a semester. The categories showing a negative progress are highlighted with different colors to catch the attention of the instructor. When the instructor detects trends that are different from his own observation, he can be more alert and check and see whether there are indeed differences between various assessment categories that he missed.

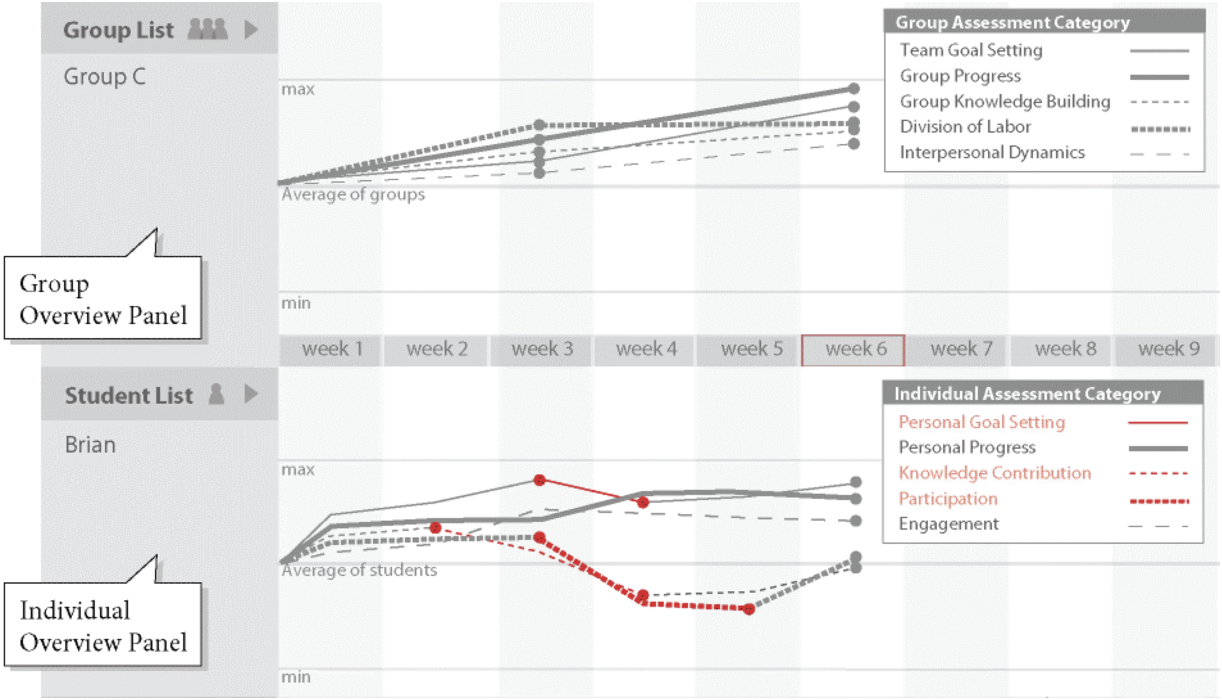


Figure 4.4. Graphs displaying group and individual assessment categories

Recall that in chapter three (Section 3.4), I introduced two stories that instructors reported as problematic during the interview study. With the help of GRASP, instructors could have diagnosed such problems earlier in the semester. The two vignettes presented below illustrate how the problems faced by instructors can be solved using the browsable summary in GRASP if it was implemented fully. Unlike the visualization presented in Figure 4.4, which is implemented, the figures presented for the two vignettes (Figure 4.5 ~ 4.8) have not been implemented. They are design prototypes used for illustrative purposes.

Vignette One

One problem reported multiple times during instructor interviews was students' unequal contributions. This first vignette was from an instructor K, who had a limited view of group work. K discovered that a student was not pulling his weight too late in the semester. Using the browsable summary tool, K potentially could have gained insight into the group work processes earlier and might have been able to detect the students under performance. The following paragraph illustrates how K could have used the browsable summary tool to detect Brian's low contribution in group C at week 4.

Imagine that at the start of week four, K loads the browsable summary tool that he uses approximately once a week. He first selects a group of students using the drop down menu in the group overview panel in the upper left of figure 4.4. The top half shows progress for the group as a whole, while the bottom half presents an individual student's progress. Figure 4.4 shows the graph through week six as a way to show the results of K's intervention, but for now imagine that K's current view is only up to week four. At week four, K examines group C, which is Brian's group. The group overview panel displays five graphs with assessment categories at the group level. The y-axis of the graphs are: team goal setting, group progress, group knowledge building, division of labor, and interpersonal dynamics respectively. The x-axis is the number of weeks in a semester. K observes that, compared to all groups, group C is doing well. K notes that the assessment categories of division of labor and group knowledge building are constant from week three to week four, but are still above average compared to other groups. Moreover, the other assessment categories are increasing compared to the other groups.

Even without the browsable summary tool, K made similar observations through his weekly meetings with group C and concluded the group was doing well. However, with the browsable summary tool, K can also observe how individual students are doing compared to other group members by selecting a member of the group, using the drop down menu in the individual overview panel shown on the bottom left of figure two. While examining each student in group C as part of his weekly routine, he sees that Brian has not done well. Figure 4.4 shows five graphs in the individual overview panel. The y-axes of the graphs are the five individual assessment categories: personal goal setting, personal progress, knowledge contribution, participation, and engagement. The x-axis of the graphs is the number of weeks in the semester. K sees that Brian's personal goal setting, knowledge contribution, and participation have all declined from week three to week four. K suspects that Brian is having trouble and investigates further.

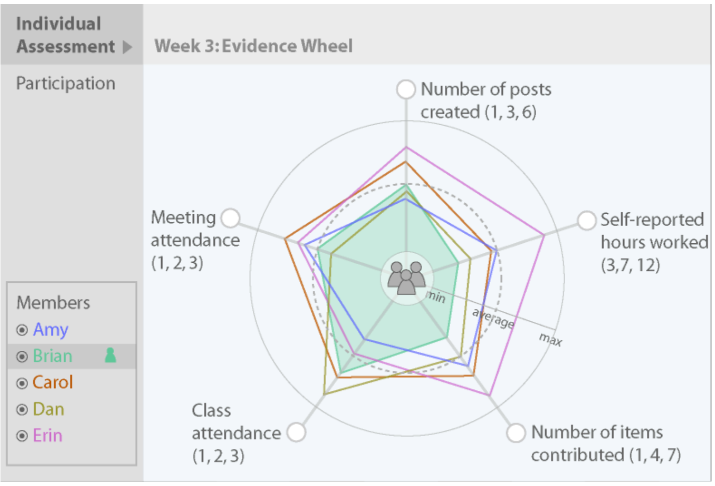

Individual Assessment ▶
Week 3: Evidence Wheel
Participation
Number of posts created (1, 3, 6)
Self-reported hours worked (3,7, 12)
Number of items contributed (1, 4, 7)
Class attendance (1, 2, 3)
Meeting attendance (1, 2, 3)
min
average
max
Members
Amy
Brian
Carol
Dan
Erin

Figure 4.5. Evidence wheel for Vignette One

K can select any of the individual assessment categories to look further at the indicators used to measure the categories. K decides to examine the participation assessment category first because he observes that, of the three declining categories, Brian’s participation dropped the most. Because the evidence wheel shows the quantitative values of evidence used to compute the selected assessment category, K selects the previous week to examine how Brian was doing. After K selects participation from the drop down list of individual assessments on the left side of figure 4.5, the evidence wheel for the selected week is updated on the right.

The evidence wheel contains three concentric circles that show the minimum (innermost), average, and maximum (outermost) values computed from the values of all group members. Because Brian’s self-reported hours worked and number of items contributed are close to the minimum, K sees this as indicative of a problem. The evidence wheel shows how an indicator, such as “number of posts created,” compares across selected students by the relative position of students on the radius of the circle. The points for an individual student are connected to form a pentagon as shown in figure 4.5. Brian’s pentagon is shaded, and by looking at each of the vertices of the shaded pentagon, K can see the relative quantitative values of each indicator for Brian. Each indicator also lists the minimum, average, and maximum values below its label.

Because some of the indicators for participation suggest that Brian is not doing his share of the work, K talks to Brian in person. Unlike the first vignette, K could detect that Brian was having a problem and therefore could intervene to give advice. In this case, Brian listened to the instructor's advice and by week six, when the midterm evaluation occurs, his assessment categories at the individual level were at the average level as seen in figure 4.4.

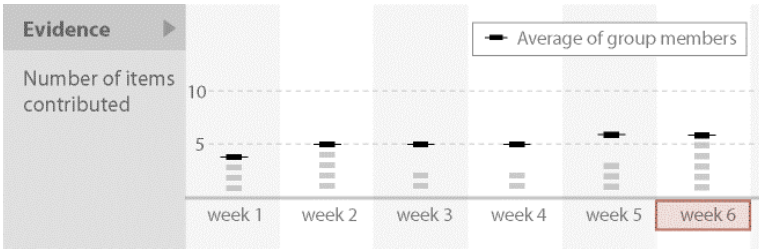


Figure 4.6. Detail of Number of Items Contributed

At week six, in order to verify that Brian has indeed taken his advice, K looks at the value of each piece of evidence over time as shown in figure 4.6. The y-axis of the graph is the number of items contributed, and the x-axis of the graph is the week in the semester. The darker orange bar is the average of the group members for that week. Indeed, K can now see that, although the number of items Brian contributed was low from week three to week five, in week six Brian is contributing as much as other members in his group. K mouses over the square bar in the graph to see details on the items Brian has contributed. As a result of gaining more insight into the work process of group C, K is able to provide advice to a student in a more targeted and timely manner.

Vignette Two

Another type of problem that was identified as problematic by instructors was when groups used a divide-and-conquer approach, without communicating with each other, in order to complete group projects. In such situation, S, an instructor for an undergraduate-level course on web design, had expressed concern about a pair of students who did not integrate their work at all. As in the earlier vignette, this problem occurred because the instructor did not have enough insight into the group processes to detect the problem. If instructor S had the browsable summary tool, he could have detected the lack of communication between group members as illustrated in the following scenario.

Imagine that at week three, S examines the evidence wheel shown in figure 4.7 and sees that the number of attachments on the message board for this group is about the same for each of the two students, but that there is a large difference between the number of posts from the two students, on the one hand, and the number of replies between the two students, on the other. Because the number of attachments is similar, the two students may be contributing similar amounts of knowledge (or doing similar amounts of work), but the other evidence signals inadequate communication between the two members about their work. In addition to the evidence wheel, S looks at the word cloud that displays the words used by the group members during a particular week (Figure 4.8). Similar to the concept of a tag cloud, a word cloud shows more frequently used words in larger size (Rivadeneira et al., 2007). Seeing that the word cloud is populated and contains meaningful words relevant to the project, S has another piece of evidence that this group is doing well in terms of contributing knowledge.

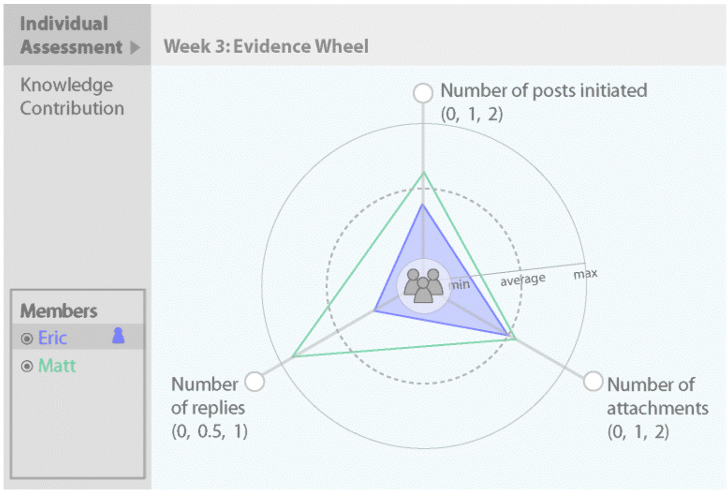


Figure 4.7. Evidence wheel for Vignette Two

Based on the evidence wheel and word cloud data, S concludes that the group members are contributing knowledge to the project, but are not communicating enough with each other.

Therefore S intervenes and advises the students to communicate more often, thereby preventing the group from producing a patched together report.

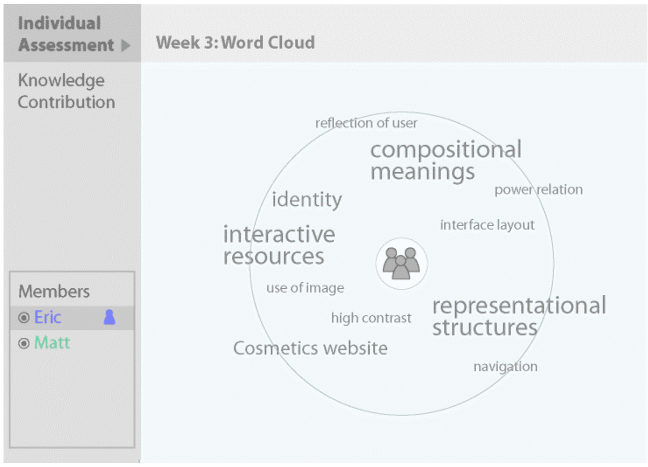


Figure 4.8. Word Cloud

4.4.3. General discussion

This chapter addresses important questions related to the extent and nature of mismatches in perspectives among interested parties in group work, whose understanding of the group varies in their degree of exposure to the context wherein that group work occurs. The formal classroom study reported in this chapter was based on data collected from a project class. The chapter describes insights gained from this investigation as well as how two principles from social psychology (the halo effect and fundamental attribution error) can be used as a lens for exploring important design questions in engineering education. However, important differences may exist across settings that would threaten the applicability of these findings. Therefore, the proof of concept offered may mainly be applicable only in the context of engineering project courses similar to the one presented in this study.

I have shown that the halo effect and fundamental attribution error are principles that can be used to explain data that showed a gap between instructor and students' perspectives on group and individual performance during group work. I have also presented how the findings from the classroom study can be applied in research that supports assessment practices of project course instructors by providing guidelines that would enable instructors to be less prone to cognitive errors and biases. Previous research has shown that, although eliminating the effect of cognitive biases is difficult, there are certain techniques that can reduce the effect (Tetlock 1985; Burger 1991). Likewise, the design guidelines should be tested in a formal study to see whether they can lessen the effect of cognitive biases because they were suggestions based on interpretation of the data used for this study.

I have only scratched the surface of what could be investigated. This investigation was based on a single class. It would be interesting to see if the findings replicate in other types of project courses, and to test what results generalize beyond the single context reported in this study. Furthermore, although I have only investigated two types of cognitive biases, researchers could adapt this methodology and test other types of psychological principles that would provide an evidence base for designing assessment tools for group work. For example, in order to test the existence of a different type of cognitive bias, one would first identify the root cause of the bias and compare the current practice with an alternative practice that is hypothesized to lessen that bias. In this study, I identified the root cause of these two biases, by demonstrating that instructors who missed the "action" that occurs during group meetings were more prone to errors in group work assessment than direct observers of the work.

Another avenue for future work could be to test how the information provided on these five processes could influence instructors' behavior and assessments. In this regard, one idea that is not addressed fully in this chapter is to examine methods for communicating information about group processes to instructors. Section 4.4.2 presents one design that could be used for such communication. However, the usefulness and effectiveness of the design should be tested.

Even though this research was conducted in an educational context, rather than in an industrial one, I expect to see similar needs between class and work environments. While the choice to conduct the work in an educational setting was made partly for practical reasons, it has advantages from the standpoint of the research focus on assessment, especially at an early stage. Consider that in a work context the idea of evaluation carries more serious consequences for the people involved than in the classroom. Thus, conducting research that focuses on and involves evaluation could be viewed as a threat, and therefore may possibly interfere with obtaining participants. Furthermore, such a study could potentially have a negative impact on work groups if they suspect that the purpose of evaluation is measuring worker performance rather than to provide formative feedback meant only for their benefit. Assessments in project course like this are more likely to be seen positively in an academic context where formative feedback is considered normal. The assumption within an educational context is that student teams participate in group work as a means for improving their skills, which prepares students to be more receptive to feedback in work environments. At the same time, to the extent that issues that trouble groups within both contexts are general, insights gained from observing groups in this context may offer some insight into other contexts as well. Therefore, this work could serve as a basis to conduct future research in a workplace setting. For example, in a work environment it might not be feasible to assess all five assessment categories given its limitations and resources. Because I know from this research that knowledge co-construction is an area that is likely to be most affected by cognitive biases when compared to other categories, one could choose to focus on knowledge co-construction only.

Given this finding on the importance and the need for supporting the process of knowledge co-construction, my subsequent research focuses on this assessment category specifically. The next few chapters detail my research on: (a) identifying conditions that support the knowledge co-construction and its relationship to learning and knowledge transfer (studies three, four, and five); and (b) automatically monitoring the group processes using natural language processing and machine learning (studies six, seven, eight, and nine).

CHAPTER 5

The role of the ICC process in group discussions and its measurement: Study 5

In this chapter, I examine the role of the idea co-construction process (hereafter ICC) during group work, particularly in knowledge transfer between work groups at key transition times, such as when group membership changes. I argue that the idea co-construction process is the essence of what it means for groups to function well in terms of knowledge sharing and perspective taking. Using methods from discourse analysis, I measure ICC statements, and test whether they have a predictive value with respect successful knowledge transfer. I hypothesize that the concentration of idea co-construction during discussion of group work is positively correlated with occurrence of knowledge transfer because idea co-construction involves reconciling different perspectives, which can be seen as offering potential opportunities for adopting others' knowledge.

In addition to verifying the relationship between the ICC process and knowledge transfer, the other key contribution of this chapter is to establish a coding manual for quantifying the concentration of ICC within a group discussion. I developed a coding manual using three different data sets called "eggdrop," "origami," and "ottoman." In this chapter, I present the coding manual using the "origami" data set, a dataset used for testing the relationship between the ICC process and knowledge transfer.

5.1. Motivation and Design

The goal of study five is to verify the role of idea co-construction (ICC) contributions during group work discussions. ICC (first mentioned in chapter one, and again in chapter 2.2), is the process of taking up, transforming, or otherwise building on an idea expressed earlier in a conversation. I argue that this ICC process, which is at the heart of collaborative learning, is also essential in group work settings, especially during key transition times. For example, the entry of a new member into a group creates an opportunity for knowledge transfer, as a group can implement new ideas, routines, and practices that its new member brings (Argote and Ingram 2000).

The theoretical link between idea co-construction and group outcomes, such as design effectiveness, has already been documented in previous research. Idea co-construction is more than a simple rote sharing of knowledge; i.e., information sharing. Instead, idea co-construction reflects on the process, drawing an inference or conclusion through the use of reason. Such statements are transactive in nature; i.e., team members use their knowledge to operate on the reasoning of their partner, or clarify their own ideas (Teasly 1997). Prior work on the relationship between the amount of transactive statements and group outcomes has shown a positive relationship between the two. This prior work has focused on topics of discussion in domains such as moral reasoning, as well as in conceptual, mathematical and scientific reasoning (Kruger & Tomasello 1986, Damon & Phelps 1989, Tudge 1992). In those studies, participants worked on solving problems in domains such as logic, science, and mathematics.

However, the relationship between idea co-construction and knowledge transfer has not yet been investigated. It is plausible to expect that the level of respect for one another's ideas that is presupposed in the ICC process creates conditions conducive to knowledge transfer. Thus, I hypothesize that when group members engage in the ICC process, knowledge transfer will occur more often, which could then lead to positive group outcomes; e.g., increased productivity in workers. I also examine whether ICC is a process distinct from other related facilitators of knowledge transfer, such as superordinate social identity (hereafter SSI) and knowledge consideration (Kane, 2010). A SSI is a person's identification with a superordinate group, such as a department or an organization. Kane showed that work groups were more likely to transfer knowledge when they shared a superordinate identity than when they did not share such an identity with other group members. This identity effect was moderated by how apparent the merit of their knowledge is (knowledge demonstrability). In addition, her study demonstrated that superordinate identity induced knowledge consideration, or the focusing of group attention on determining the value of another's knowledge (Kane, 2010).

The relationship between the variables involved in the main hypothesis of verifying the predictive value of idea co-construction on knowledge transfer is presented in the figure 5.1.

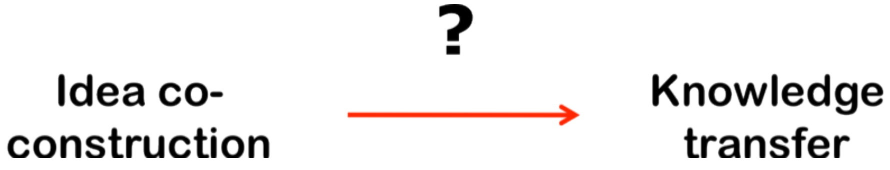


Figure 5.1. Hypothesis

5.2. Procedure

In order to test the hypothesis, I used data that was collected in an experimental study wherein a new member rotated into a group of three others. The group itself was in the process of making origami sailboats in an assembly line (for details, see Kane, 2010). This data was collected by Kane (2010) in an experimental study. Kane conducted a 2 (superordinate social identity: yes, no) x 2 (knowledge demonstrability: less, more) x 2 (production trial: trial 1, trial 2) mixed model design study. Here, superordinate identity and knowledge demonstrability were *between* subject variables, whereas the trial was a *within* subjects variable. Volunteers for the study participated in a one-hour session and were assigned into one of two different work groups; a recipient and a source group. The study consisted of the following three phases.

- Phase one: Introduction and superordinate social identity manipulation. In this phase, a total of 6 people entered the room and were assigned to either one of two teams: the recipient or source team. In the superordinate identity condition, participants were introduced as a single six-person group comprised of two subgroups. In the no superordinate social identity condition, participants were introduced as two separate three-person groups.

- Phase two: Training and knowledge demonstrability manipulation. In this phase, the recipient team learned a less efficient 12-step origami routine for making a paper sailboat. In another room, the source team learned a more efficient 7-step origami routine, which was the superior routine. The participants in both teams were not aware of the fact that they learned different origami routines, nor that one method was technically more efficient than the other.
- Phase three: Personnel movement and group production. The second person in the assembly line of each group switches groups. At this point, I am interested in the performance of the recipient team who now has the new member with knowledge of the superior 7-step routine. Will the recipient group adopt the new knowledge or retain their old, less efficient methodology for making paper sailboats?

Here, knowledge transfer is measured during phase three. When groups use the newcomer's superior routine he learned in phase two, knowledge transfer occurs. I focus the current analysis on the half of the sample where the newcomer's knowledge is low in demonstrability, usually because the merits of a less demonstrably superior production routine are concealed for the sake of social variables. Under such conditions, Kane (2010) found a significant effect of superordinate social identity, namely a psychological sense of belonging to the same group. Here, knowledge transfer was mediated by knowledge consideration, or the amount of collective attention paid to the newcomer's routine during a key part of the group's conversation. In the current study, I examine groups engaged in the ICC processes during these key moments in conversations, which always temporally precede any incidence of knowledge transfer. In order to examine the uniqueness of ICC compared to superordinate identity and knowledge consideration, I will also compare ICC with two dichotomous variables - the dummy variable, which indicates whether or not groups were induced to share a superordinate identity, and the coded variable, which indicates whether or not groups engaged in high or low levels of knowledge consideration, or the focusing of attention on the new member's new routine (for details, see Kane et al 2010). Also consistent with Kane (2010), knowledge transfer is treated as a repeated, dichotomous measure that reflects whether a group has adopted the newcomer's superior routine.

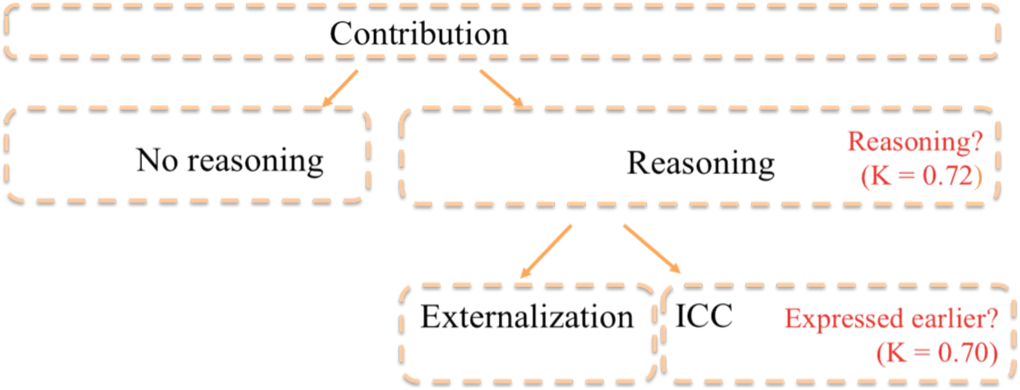


Figure 5.2. Two step coding process for ICC

My measure of ICC was a dichotomous measure in that it reflects deep, collaborative reflection wherein participants take the time to display their reasoning, *and* others build on that reasoning. I looked for evidence of ICC across the units of speech that group members expressed during the key conversation that transpired after the newcomer shared a better routine, but before the group resumed production. In order to be coded as an ICC speech unit, a statement should first contain a display of reasoning. Also that display of reasoning should be related to a previous statement. Therefore, I begin by differentiating non-reasoning and reasoning statements, and then focused on differentiating between reasoning statements that represent new directions, from those statements that build on prior contributions (i.e., ICC contributions). This two-step coding process is shown in figure 5.2, and detailed in section 5.2.1.

The two-step coding process described in this section is also used to measure ICC in later studies (studies eight and nine). The basic elements and the structure of the coding process stay consistent across all three studies. However, the concepts, which are used in the first step of detecting the reasoning process (operationalization step 1), are redefined depending on the context in which the study was conducted. The coding manuals for the studies can be found in appendix B, D and F.

5.2.1. Operationalization of the idea co-construction process

Usually, when students are working on a given task or a project in a team, they receive a certain amount of information that will help them solve the problem, usually in the form of a task statement and training materials. In order to solve the given problem, students discuss the materials and try to apply them to a potential solution to their problem. I am interested in capturing instances when students display reasoning during group discussions that goes beyond what is given and displays some understanding of a causal mechanism, since typically some causal mechanism would be referenced in a discussion of how something works, or why something is the way it is. One purpose in segmenting student talk by identifying those segments that display reasoning is so that the amount of reasoning displayed can be quantified. What I am coding for is attempts at displayed reasoning. Thus, I need to allow for displays of incorrect, incomplete, and incoherent reasoning to count as reasoning, as long as in our judgment we can believe an attempt at reasoning was going on. Such judgements will necessarily be quite subjective – especially in the case of incoherent explanations.

At the heart of this operationalization is the idea of causality. Causality may be indicated with a "because" type clause, or be implicit, where the meaning of the verb communicates the mechanism, (e.g., "Folding the paper like so reduces the number of necessary folds to achieve a boat"). I refer to this distinction between non-reasoning and reasoning statement as "operationalization step one." But in "operationalization step two," I focus on the distinction between reasoning statements that represent new directions within a conversation (i.e., externalizations) from those that build and connect on prior contributions (i.e., ICC contributions). The notion of what counts as a connection between a reasoning statement and a prior statement is a simpler distinction to make. Any overlap in content, such as references to the same objects, goals, or actions, counts as a connection for the purpose of distinguishing ICC from new directions. The measure of ICC used in this research is computed by summing up the ICC statements for each group, carried out by a coder with extensive experience coding for ICC, and using a corpus from a different domain. This count measure was dichotomized at its median (2) to facilitate the analysis of its role in knowledge transfer across time.

Operationalization step 1: Reasoning process

My formulation of what counts as a "reasoning display" borrows from the Weinberger & Fischer's (2006) notion of what counts as an "epistemic unit," where they look for a connection between some detail from the given task and a theoretical concept. For example, when students have seen enough text such that they can see in it mention of a case study detail, a theoretical concept, and a connection between the two, Weinberer and Fischer place a segment boundary around it. Occasionally, a detail from a case study is described, but not in connection with a theoretical concept. Or occasionally a theoretical concept may be mentioned, but not tied to a case study detail. In these cases, the units of text are considered degenerate, they're not quite epistemic units.

I have adapted this notion of an epistemic unit from Weinberger & Fischer, but rather than using it the same way they did, I have adapted it because the topic of conversations are very different in nature. The goal of detecting the ICC process is to distinguish instances when students are making statements based on "reasoning" contributions, not just parroting what they have heard. In my current formulation, I am considering the task and training materials that the experimenter has provided as *given*, and mark a distinction between what is given to the students up front and what the students contribute themselves beyond that.

The conversations that I analyzed come from three participants who were asked to build as many origami paper boats as they possiblely could. As in Weinberger & Fisher's (2006) notion of "epistemic unit," I look for a connection between two or more concepts. Unlike Weinberger & Fisher's operationalization of reasoning, where one of the concept is details from the task and the other is a theoretical concept, both concepts can be of either type. Details of what a "concept" is are listed in the sub section below. Insofar as my operationalization of reasoning is concerned, here are a few examples.

First, examine a segment of this conversation where I have highlighted instances of displayed reasoning with italics:

s1: *I think her way was faster [comparison]*
s2: yeah.
s3: You think?
s1: *Yeah, because there's a lot of bending with ours [causal relationship]*
s3: *Okay, then why don't we learn her way so that we can be more efficient? [causal & comparison]*

The simple way of thinking about what constitutes a reasoning display is that it has to communicate an expression of comparison and contrast, or some causal mechanism. Oftentimes that will come in the form of an explanation, such as "X is faster than Y," or "X because Y." However, it can be more subtle than that, for example in the case of "Push it further up [,] just enough for it to have a base." The basic premise was that a reasoning statement should reflect the process of drawing an inference or conclusion *through* the use of reason. Note that in the more subtle example, although there is no "because" clause, one could rephrase it in the following way, which does contain a "because" clause: "Push it further up just enough [*because that will cause the paper boat*] to have a base." Reasoning statements stand in contrast to information sharing statements, which can be thought of as sharing of rote knowledge.

Concepts

The basic building block of a reasoning statement is a concept. I identified five types of concepts relevant for this domain, namely theoretical concepts, prior knowledge, physical system properties, emergent system properties, and goals. For each concept, the definition and an example are given in table 5.1. The examples in the table are from the dataset described in section 5.2, where students are discussing the best way to build origami paper boats. Note that the "system" in this case is the origami paper boat and the pieces of papers used for making them.

Table 5.1. Definition and examples for the 5 concepts.

Type	Definition	Example
Theoretical concept	principles (i.e. physics principle) and theories that can be applied in decision making	knowledge about origami steps
Prior Knowledge	information that helps decision making based on common sense	folding several papers together is faster than folding each paper separately
Physical system properties	elements and characteristics of elements that are available for the system	paper is difficult to fold
Emergent system properties	characteristics of elements that appear in the process of making origami	properties of the origami boat
Goal	general beliefs and perspectives, or anything associated with strong expectations related to points of view. Label incorporates unstated opinions or perceptions.	efficiency, or making as many boats as possible

Relationship

The presence of multiple concepts in a statement by itself does not determine whether a statement contains reasoning. Rather, the expressed relationship between multiple concepts is the determining factor. For example, a simple list of concepts (e.g., this cup is round, and it is also white) is information sharing, not reasoning. I identified two types of relationships that signal a reasoning process: 1) Compare and contrast; and 2)Cause and effect.

1. Compare and contrast: When the speaker compares two concepts, the speaker is making a judgment, which necessarily involves thinking about how two concepts are related to another.

    - The speaker says that the other method might be better compared to his own method because he has learned this new method. Speaker compares the two methods: *"It might be better, now that I've figured out what I am supposed to do."*

2. Cause and effect: When the speaker uses a cause-and-effect relationship, this process involves establishing the relationship between two concepts through a reasoning process. The general relation in this category is "doing x helps you achieve y." Examples are illustrated below:

    - Let's do A because of B: *"Let's use our method because we are already used to it."*
    - Let's do A in order to achieve B: *"Let's fold multiple papers at once for efficiency."*

Operationalization step two: Idea co-construction vs. Externalization

Statements that display reasoning can be either externalizations, which represent a new direction in the conversation, and not building on prior contributions, or ICC contributions, which operate on or build on prior contributions. In the distinction between externalizations and ICC contributions, I have attempted to take an intuitive approach wherein I determine whether a contribution refers linguistically in some way to a prior statement, such as through the use of a pronoun or deictic expression.

Consider the sample conversation I used earlier to illustrate reasoning contribution. The lines marked (E) are contributions that are categorized as externalization, whereas the ones (I) are ICC contributions. The first statement by s1 is an externalization, since s1 starts a new topic. Thus this contribution is not building on a prior contribution. Subsequent reasoning contributions in this discussion are coded as ICC because they each build on statements that directly precede them.

s1: *I think her way was faster (E)*
s2: yeah.
s3: You think?
s1: *Yeah, because there's a lot of bending with ours (I)*
s3: *Okay, then why don't we learn her way so that we can be more efficient? (I)*

5.2.2. Reliability of Annotation

Two coders were trained using a manual that describes operationalization of reasoning displays and idea co-construction. They were also given an extensive set of examples. After each coding session, researchers discussed disagreements and refined their manual as needed. Most of the disagreements were due to the interpretation of what the students meant, rather than with this definition of reasoning itself. Therefore, later efforts focused more on defining how much the context of a statement could be brought to bear on the interpretation. In a final evaluation of reliability for reasoning process, I calculated a kappa agreement of 0.72 between both coders,whose disagreements were settled by discussion between the coders. For detecting instances of idea co-construction and externalization, the kappa value was 0.70.

5.3. Findings

Before examining the main hypothesis, I investigate whether ICC is a process distinct from the facilitators of knowledge transfer found in Kane (2010). I conducted an analysis of bivariate correlations as well as a factor analysis. I found neither a significant bivariate correlation between the ICC process and superordinate identity (r = -0.25, *ns*), nor a significant correlation between the ICC process and knowledge consideration (r = 0.25, *ns*). Consistent with these results, a principal components analysis with varimax rotation revealed two separate factors (eigenvalues ≥ 1.00), and together explained 85% of the variance. SSI and knowledge consideration loaded on one factor (item loadings ≥ .80), while the ICC process loaded on its own factor (item loading ≥ .90). These results suggest that ICC is distinct from previously discovered facilitators of knowledge transfer between groups via membership change.

I now ask whether the ICC process is associated with the occurrence of knowledge transfer between the groups. Because I measure knowledge transfer across two time periods, I used a weighted least squares regression technique for analyzing repeated-measures of categorical data (for details see, Stokes et al., 2000). I modeled the correlated marginal proportions of knowledge transfer with an underlying contingency table, and then with 2 (ICC) x 2 (SSI) classifications and four response profiles for knowledge transfer in trial one and trial two: YY, YN, NY, NN. This analysis revealed the following two significant main effects. Consistent with Kane (2010), there was a significant effect for SSI, X^2 (1, N = 24) = 12.15, p < .001, wherein the marginal probability of knowledge transfer was 21 percent (*SE*=.08) higher for groups with a SSI compared to those without such an identity. As hypothesized, there was also a significant main effect for ICC process, X^2 (1, N = 24) = 7.18, p < .01. The marginal probability of knowledge transfer was 18 percent (*SE* = .07) higher in groups whose members took up, transformed, or otherwise built on an idea expressed earlier in a conversation.

5.4. Discussion

Preliminary results suggest that ICC plays a unique role in knowledge transfer, which is a complex, multiply determined social phenomenon. Whereas Kane (2010) revealed that knowledge transfer between groups via membership change is determined in part by the *type* of social psychological relationships between groups (superordinate identity), itself mediated through the *depth* of attention paid to the knowledge (knowledge consideration), this current research reveals that knowledge transfer is also determined by the *way* that group members engage in dialogue (ICC).

Although this study shows the potential for the ICC process to be a unique contributor to knowledge transfer in group work, these analyses do not allow us to say anything about causation—that is, whether elevating the extent to which groups engage in the ICC process would causally lead them to perform better in terms of knowledge transfer. However, my results are consistent with research that shows relationships between constructs similar to ICC and related group outcomes such as learning (Joshi & Rosé, 2007). A future research direction would be to investigate the causal relationship between ICC and knowledge transfer. For example, an experimental study where the amount of idea co-construction is manipulated would allow for such an investigation. Other research might include the use of different tasks to investigate the generalizability of these findings independent of the task type (in cases other than folding paper boats).

CHAPTER 6

Automatic assessment of group processes using conversational speech data[3]: Study 7

While verifying the important role that different types of group processes play in engineering design discussions would make a valuable scientific contribution, I must apply these findings in order to make a practical contribution as well. Studies one and two, introduced earlier in chapters three and four, demonstrate the importance and need for supporting assessment of group work processes. In this latter half of the dissertation (Chapters six, seven, and eight), I present my research machine learning technology that automatically provides instructors with insights about group work processes as they unfold during group work. If such automatic assessment can be computed from collected classroom data, a report wherein individual students' weekly statistics are displayed could be given to the instructor as a means of better assessing the students and the group.

I focus specifically on technology that detects processes related to knowledge co-construction from speech. I argue that this focus serves as an interesting first attempt to computationalize theories from sociolinguistics and pragmatics for automatic assessment. Before embarking on that discussion in chapters seven and eight, I begin by describing my earlier technical work (study 8) Then I present my preliminary work towards assessment of the idea co-construction

[3] This work is published in Gweon, G., Kumar, R. Jun, S. Rosé, C. (2009). Towards Automatic Assessment for Project Based Learning Groups, In Proc. Artificial Intelligence in Education.

process from speech. Lastly, in chapter eight I describe my final approach (study 9), which is the computationalization of theories from pragmatics and sociolinguistics.

I now present a specific technique for predicting activity levels and the amount of overlapping speech in recordings of actual student group meetings. The data was collected from a graduate engineering design project course over the course of a semester. This preliminary work shows promise in automatically detecting the level of knowledge contribution and simple turn taking dynamics.

6.1. Motivation & design

Study two illustrated the difference in perspectives between observers, students, and instructors, and it suggested that some types of reporting might improve an instructors' ability to make more accurate assessments of group work. To this end, researchers have looked at automatically detecting various aspects of student activities during group work (DiMicco, Hoolenbach, & Bender 2006; Kay et al., 2006; Pianesi et al., 2008). For example, various forms of data including text data (Soller & Lesgold 2003), video (Chen 2003), audio (DiMicco, Pandolfo, & Bender 2004) and even galvanic skin response (Madan, Caneel, & Pentland 2004) have been measured by previous researchers (review presented in section 2.3). Within the scope of this work, several researchers have used the amount of speech as an indicator of participation during group meetings (DiMicco, Pandolfo, & Bender 2004, Chen 2004). Although the amount of speech is a reasonable starting point that can be used to detect a level of student activity, I hypothesize that more useful information can be extracted from speech. To that end, I tried to automatically predict student activity levels during group meetings and to investigate the relationship between the activity level and various types of group processes identified from study two.

More specifically, I use features extracted from speech, such as pitch and energy level, to monitor student activity levels. Using features that reflect the style of speech, rather than those that are directly from speech content, has its advantages. For example, although, interpreting the content of speech might be useful, current technology is not at a state where one can obtain accurate transcripts using automatic speech recognition from real meetings audio data (Hain et al., 2005). In addition, not looking at the content of speech can preserve the privacy of the speakers. The choice to examine the style of speech as a way to predict student activity level is based on existing literature that examines the role of speech styles in conversations. For instance, Dabbs and Ruback shows the usefulness of style of speech in gaining insight into group processes that occur among individuals who participated in group work (Dabbs & Ruback 1987). Dabbs and Ruback manually analyzed the number of turns, pauses, and interruptions in a group conversation, and they showed that individuals who talked more were rated more favorably, as well as individuals who paused more during their speech. Although Dabbs & Ruback's work is influential, their analysis was conducted manually. In this study, I use speech processing and machine learning to examine whether the machine generated indicators can be used to distinguish students who are actively participating in group work.

6.2. Procedure

A graduate level engineering course offered in spring of 2008 provides the context for my data collection effort. Students enrolled in the course worked on a single project sponsored by a client, and four subgroups of students were formed to carry out this project. Because the class is a project oriented, a major component of the grade assigned by the instructor is based entirely on their productivity, and this portion of the grade is explicitly indicated by the instructors and treated separately from that part of the grade related to the quality of the result. There were two instructors and 22 students in the class. Various types of data were collected in this class, including instructor assigned grades, instructor assigned scores and observer assigned scores. The "instructor assigned grades" are formal grades assigned to the students for the class for the duration of the observation period. The "instructor assigned scores," and "observer assigned scores," are weekly evaluations of students in the five areas of the assessment framework established through my interview study (study one).

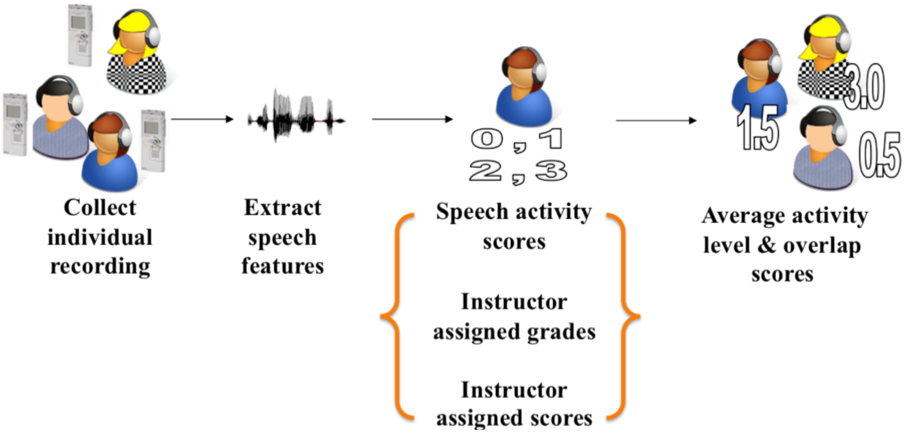

Collect individual recording
Extract speech features
0, 1
2, 3
Speech activity scores
Instructor assigned grades
Instructor assigned scores
3.0
1.5
0.5
Average activity level & overlap scores

Figure 6.1. Process overview for study seven

However, this data was not enough to address the gap between what instructors would like to know about the groups they are overseeing and what they actually see. In order to get a more specific picture of what information instructors are missing, I instrumented the course in order to collect extensive observational data from the subgroups. Specifically, I collected audio recordings of group meetings as well as video tapes of classroom activities. In this study, I compared the correlation between the observer assigned scores and three types of scores: 1) instructor assigned grades; 2) instructor assigned scores; and 3) Speech activity scores predicted by machine learning. The overview shows the relationship between these three scores and the machine learning process in figure 6.1. The details of instructor and observer assigned scores are detailed in study two (chapter four, section 4.2). Using speech activity scores, two types of measurements were additionally computed: 1) The averaged percentage a student talked during group meetings (average activity level); and 2) The averaged percentage a student talked over another group mate (amount of overlap). The average activity level is an approximate measure of the amount of talk that a student contributed during group meetings. The average amount of overlap says something about how seriously a student's group mates take his or her contributions. If overlap is high, then it may be the case that the student's group mates don't find it valuable to stop and listen when he or she speaks. Therefore in the remainder of this section, I present the process used for obtaining the speech activity scores, average activity level, and overlap scores.

As a first step towards automatic assessment from speech, I started with the relatively simple task of assessing the level of contribution. Using recordings collected from students during project group meetings, I computed the amount of activity for each student using machine learning technology. Before the activity level can be computed from speech, it must be segmented, and each segmented must be coded for the amount of speech by the associated student. I chose to segment the speech into 10 second intervals so that it would be reasonable to assume that, for most segments, there would be at most a single dominant speaker. That allowed us to utilize a relatively simplistic approach to coding activity level for individual segments. I adopted the following 4-point scale for activity level: 0 - no speech from primary speaker; 1 - primary speaker back-channels, where back-channeling is a way of showing a speaker that you follow and understand their contributions, often through interjections such as , "I see," "yes," "OK," "uh-huh"; 2 - primary speaker speaks but holds the floor for less than half of the 10 seconds; 3 - primary speaker speaks but holds the floor for more than half of the 10 seconds.

I first verified that human annotators could make this judgment reliably from the audio recordings of individual segments. Using this coding scheme, the inter rater reliability evaluated for two coders over 144 segments was 0.78 Kappa. With the reliable coding scheme, a single coder then coded 1132 segments (distributed evenly across students from a project course). The largest proportion of segments was coded as 0, which amounted to 47.5% of the segments. 8.5% were coded as 1, 30.5% as 2, and 13.5% as 3.

Next, in order to apply machine learning to speech, each segment of speech must first be transformed into a set of feature-value pairs. The activity level that I am trying to predict from speech is related to *how* the words are spoken rather than the content of those words. Such structural aspects of speech are captured by speech prosody. Similarly, other speech applications such as emotion detection are also more concerned with speech prosody rather than content (Ang et al., 2002), and thus use features similar to ours. In contrast, speech applications such as dictation software use content related features such as spectral features processed through a speech recognition system. Therefore, the features extracted from speech for the experiment are variations of prosodic features such as pitch, power and amount of silence. A total of 39 features were extracted for each of the 10 second segments using Wavesurfer (Beskow & Sjlander, 2000).

I extracted pitch and power contours for each of the 10 second segments. The variations of prosodic features that I extracted using these contours are detailed in the next paragraph. The numbers in brackets indicate the number of features. All features are computed automatically.

The feature set includes structural aspects of speech, features related to F0 and power. For the pitch related features, I calculated average, maximum, minimum, and range of F0 (4) and delta F0 (4). All F0 parameters are computed over voiced frames in the utterances. The voiced frames are identified while computing F0 in Wavesurfer. Power features were computed similarly to pitch features. They include average, maximum, minimum, and range of power (4) and delta power (4). Unlike F0 related features, these features are computed considering both the voiced and unvoiced frames. I added total power and power in voiced frames (2) as well. Also, a ten point power contour was calculated. The ten points are computed by dividing the segment into uniformly sized one second sub-segments, then computing the average power of each sub-segment (10). I also computed average, range, and standard deviation over the ten point power contour (3). Average, maximum, minimum, range, and standard deviation of the deltas of the ten point power contour (5) were included. A feature that counted the number of points on the ten point power contour that were above the average delta was also used (1). Finally, duration related features included duration of voicing and duration of initial silence (2). Duration of initial silence was computed automatically using heuristics computed over power and F0 contour.

After this coded speech data had been transformed into a vector representation, I evaluated whether it was possible to use machine learning to automatically assign segments of speech to one of these four categories with a high enough accuracy to be helpful. Because the feature space was small and because the possibility of interactions between features within the vector representation existed, I used Weka's SMO (Weka's implementation of a support vector machine) learning algorithm (Witten & Frank, 2005). In order to avoid the evaluation results from being inflated due to overlap in speakers between train and test sets, I adopted a cross-validation methodology where a model was first trained on all but one student, and then its performance was evaluated over the segments of the remaining student. I did this for each student and then averaged across students to compute a performance of 74.26% accuracy. I then validated the model by using the human-coded numbers for each student to compute an average activity level, then made a similar computation using predicted values from the cross-validation experiment. When I correlated the average activity levels for each student based on human codes with those based on the automatic codes, I achieved a correlation coefficient of 0.97, indicating that I can achieve a reliable estimate of activity level using a machine learning model. I then trained a model using all coded data, which was used subsequently to code the data in the correlation analysis presented later in the chapter.

Table 6.1: Example predictions of speech activity[1]

(sec)	Student 1	Student 2	Student 3	Student 4
0~9	3 (3)	0 (0)	0 (0.5)	1 (0.5)
10~19	3 (3)	0 (2/3)	1 (1)	0 (2/3)
20~29	3 (3)	2 (1)	2 (1)	1 (1/3)
30~39	3 (3)	1 (1.5)	0 (1)	0 (0.5)

[1]Numbers in parenthesis indicate the smoothed values.

I applied this trained model to a separate set of speech data from that used to build the models for predicting student activity. For the test data, two of the five meeting recordings that were submitted in phase two (the class was divided into three phases) were randomly selected for each student, then averaged to yield one score per student. Four students never turned in any recordings, so 18 students' recordings were segmented into 10 second segments. The length of each recording differed due to differences in meeting lengths. The number of segments ranged between 7 minutes 30 seconds, on the one hand, to 2 hours 19 minutes 50 seconds in length, on the other (45 to 839 ten second segments), averaging 47 minutes in length (282 segments). Using the speech model from above, student recordings were then assigned value corresponding to an amount of talk. Example predictions are shown in table 6.1. Here we see that Student one is the dominant speaker for all three segments shown. Students two and three start to contribute more substantially during the third segment, and student four only back-channels.

Average activity level

Using these predictions of activity level, an average activity level was calculated for each student. Average activity level is an approximate measure of the amount of talk that the student contributed during group meetings. Because activity level was coded according to the amount of speech from the student, average activity level is directly proportional to their amount of talk. To calculate average activity level for each student, I took the average of the predictions across the 10 second segments of that student's speech. For instance, the amount of talk for student three in the first 40 seconds of the meeting would be (0+1+2+0)/4 = 0.75. By calculating the amount of talk, one can see which student contributed most in terms of proportion of the meeting any student spent talking. Using the data from table one, one can see that student one talked most, student two and student three spoke the same amount, and student four spoke the least.

Amount of overlap

Using the same predictions of activity level per segment, an amount of overlap index was calculated for each student. Overlap is defined as the activity level of group mates when another group mate is actively talking.

In order to compensate for some error in coding activity level, I first smoothed the predictions of activity level by averaging the activity level prediction of a segment with those of the segment before it, and with the segment after it. The smoothed scores were then real values between 0 and 3. The smoothed values of each segment are displayed in parenthesis in table 6.1. I then applied a threshold to determine what segments would be treated as segments during which a student was speaking. The threshold for each meeting was computed as the average of all activity levels over all smoothed segments in whatever meeting. For each of the 10 second segments, I compared the student's smoothed activity level to the threshold. Therefore, if a student's smoothed activity level in the given segment was above the threshold, that student was speaking during that segment.

For example, consider the segments in Table 6.1. The threshold for all the data points would be the average of all smoothed data points, which are the numbers in parenthesis. Thus the threshold would be (3+3+3+3+0+2/3+1+1.5+0.5+1+1+1+0.5+2/3+1/3+0.5) / 16 = 1.29. Comparing each smoothed activity level segment to this threshold value, one can see that, all of student one's smoothed activity levels are above the threshold of 1.29, as well as the fourth segment of student two. Next, for the segments where activity level was larger than the threshold, which are segments where the student was talking, I computed the average smoothed activity level of the other meeting participants during that segment. This is the overlap score for that student, and it indicates the prevalence of other group members talking at the same time this student is talking. For example, using the numbers in table two, the amount of overlap for the fourth segment of student two would be the average activity level of other three students, (3+1+0.5) / 3, or 1.5. Finally, after computing amount of overlap for all the 10 second segments, an average of the amount of overlap is computed over all segments in a given meeting to yield one overlap score per student.

6.3. Findings

Table 6.2 presents correlations between the observer assigned scores and three types of scores; 1) instructor assigned grades; 2) instructor assigned scores; and 3) Speech activity scores predicted by machine learning. The correlations were compared along the five dimensions identified from study one. Unfortunately, I could not compute correlations for the Group Dynamics dimension due to insufficient variability in the instructor ratings. Note that this analysis focuses on individual level assessments since the indices extracted from the speech are from individual recordings.

I first computed a Pearson's correlation between the observer and instructor ratings for each student. The correlations were computed along the five dimensions identified from study one and their average score. The correlation for the dimension of teamwork could not be computed due to insufficient variance. The average was computed by taking an average value of the five dimensions for each student before calculating the correlation. These numbers are reported in the first row of Table 6.2. Surprisingly, the correlations are consistently below .3, sometimes essentially 0, and other times are even negative. The second row of the table shows correlations between the instructor's official grades and the observer ratings for those students on each dimension. Again, the correlations are not too strong. Only the Knowledge Sharing dimension showed a substantial improvement in predictive power over the more specific instructor observation scores that were assessed for each dimension. This result is consistent with other types of correlations reported in study two (section 4.3). In particular, study two presented the correlations between student score and observer and instructor scores across the five different dimensions of group processes. Here, we see that the correlations between instructor and observer are low as well. Therefore, in addition to the evidence shown in study two, one can see that instructors views on each of the assessment categories do not line up very well with those who directly observed student group meetings.

Table 6.2. Correlation between observer ratings and other measures

	Observer Goal	Observer Progress	Observer Knowledge Sharing	Observer Labor Division	Observer Average
Instructor scores	.032	-.347	-.021	.222	.292
Instructor grade	-.032	-.169	.366	.211	.273
Speech activity	.514	-.031	.309	.351	.447

Finally, I computed a correlation between the level of speech activity indicator computed from the meeting recordings and the observer ratings. These numbers are reported in the third row of Table 6.2. Notice that, despite being an automatically computed indicator, this was the best performing indicator in the Goal Setting, Division of Labor, and Average categories. For Knowledge Sharing, it was not substantially different from that of Instructor grade. The only place where it under-performed is on the Progress dimension, where its correlation with observer ratings is essentially 0. These correlation scores imply the potential of even a very simple automatic assessment from speech to be an additional source of information that instructors can use in evaluating students.

In addition to speech activity scores, I also computed average activity level and overlap scores. These two quantities are marginally correlated (r = .46, p<.1). Thus it is not surprising that when I computed the correlations with these two values, with the average instructor scores, they were both significantly and equally correlated. The correlation between the average activity score and instructor score was (r = .51, p< .05), and the correlation between the overlap score and instructor score was (r =.54, p<.05). However, ideally instructors should discriminate between these two.

In order to evaluate the difference between these two scores, I first divided the students into two sets, according to how well the average instructor score correlated with the average observer score. I did this by first computing the residual from the regression between average observer score and average instructor score. The absolute value of these residual scores can be interpreted as a measure of how far off the instructor's scores were. There is some noise in this measure since one cannot claim that observer scores are the ground truth. However, as seen in study two, observer scores are not as subject to the fundamental attribution error and the halo effect as the instructor scores. Next, I divided students into two sets using a median split on the absolute value of the residual scores; I did this in order to identify those students whose instructor scores were the farthest off. Being able to distinguish students in this set from the other students would be valuable, since it would offer instructors insight into how they can improve the accuracy of their assessment by strategically investing their time to gain more first-hand exposure to the behavior of specific students.

One reason that instructors should distinguish between the average activity score and overlap score is because some students speak up a lot in meetings may be taking a lot of initiative and leadership, but if their overlap score is also high then it may indicate that their team mates do not see them as contributing something valuable to the discussion that is worth stopping and listening to. As supporting evidence for this interpretation, the average observer scores correlate marginally with average activity score (r = .44, p<.1) and do not correlate at all with overlap score (r = .03, n.s.). Since instructors don't have the opportunity to attend all of the group meetings, they do not have the opportunity to observe how a student's teammates are responding to his contributions. If the presence of overlapping speech is a key indicator of students for whom the instructor's assessment is likely to be far off, then one would expect that students in the set where assessments were estimated to be far off should have higher than average overlap scores. Indeed, that is what I found (F(1,10) = 6.95, p<.05, effect size .43 s.d.). And no difference between these two subsets of students in terms of average activity level (F(1,10) = .75, n.s.) was found. Thus the students who the instructor can accurately assess are no different from those that are problematic in terms of their contribution in group meetings; however, the way those students are received by their group mates is different.

6.4. Discussion

Recent advances in automatic collaborative learning process analysis (Rosé et al, 2008) brings the vision of developing a tool to support automatic group assessment within practical reach. This technology has been shown to be capable of detecting important conversational events that are indicative of successful group learning, in highly controlled settings, over short periods of time. However, I have used data collected from real world classroom environment to detect student activities. In that sense, this work presents important foundational explorations into how this basic technology can be adapted and applied to a larger problem, in the more realistic and less controlled setting of an engineering design course.

In that regard, this chapter presented a technique for predicting the amount of speech activity from recordings collected of "real" group meetings. My results suggest that quantities such as amount of speech activity, which can be computed directly from recorded speech, are useful for making assessments of group work. In particular, my findings show that the automatically extracted indices correlate better than the instructor ratings with the objective observer's ratings. In addition, I have shown the potential for using both the amount and style of speech to provide instructors with a look into student's behavior in group meetings, information that busy instructors did not have access to in the first place.

I have only begun to scratch the surface in terms of what can be detected in speech recordings collected from group meetings. The main take away from this investigation is that rudimentary assessments, such as level of activity and basic turn taking, can be made from speech fairly easily. However, while they provide some indication of the types of group processes that are of interest to instructors, they only tell us a small amount of what an instructor needs to know. For example, in Table 6.2 we see that at best they explain 25% of the variance in human assigned scores (i.e., on the Goal Setting dimension). Further work can also be done to identify other quantities apart from the amount of speech activity, which can be extracted from the speech and that might be useful for project course instructors or facilitators in the future.

CHAPTER 7

Assessing idea co-construction from speech[4]: Study 8

In studies one and two (chapters three and four), the importance and value of the assessment of the knowledge co-construction process was verified. A later study (study five; chapter five), showed the important role of one type of the knowledge co-construction process, namely the idea co-construction (ICC) process. In this chapter, I develop technology to automatically detect where ICC contributions are occurring within group interactions. This work builds on the speech processing technology used in chapter six (study seven). More specifically, I make use of features that can be extracted from speech as in study seven, such as pitch and energy level. However, in addition to these prosodic features, I also extracted phoneme features, which suggest aspects of speech content without attempting to directly extract the content from the speech.

[4] This work was published in Gweon, G., Raj, B., Rosé, C. P., Agrawal, P., Udani, M. (2011). The automatic assessment of knowledge integration processes in project teams. In Proc. Computer Supported Collaborative Learning. P.462-269. Best Student Paper Award.

7.1. Motivation & design

The research goal of the work presented in this chapter is to develop a speech processing technique capable of making a three way distinction between: 1) contributions to a conversation that do not make reasoning explicit (“non-reasoning”); 2) contributions that display reasoning and represent a new direction in the conversation that does *not* build on prior contributions (“externalizations”); and 3) ICC contributions that display reasoning that builds on prior contributions.

Automatic analysis of the ICC process is not a new direction in the CSCL community; however, prior published work has largely been related to the automatic processing of text, such as newsgroup style interactions (Rosé et al., 2008), chat data (Joshi & Rosé, 2007), and transcripts of whole group discussions (Ai et al., 2010). A key feature enabling high accuracy of recognition is the ability to measure content similarity between a contribution from one speaker and the contributions from other conversational participants, which occur within the same topic segment earlier in a conversation. For example, Rosé and colleagues (2008) report the same concern in a classification task with a coding scheme related to transactivity. The authors’ coding scheme here was based on Weinberger & Fischer’s (2006) notion of transactivity, which is what I adopted in defining a “reasoning statement” used in ICC coding (details of my coding scheme are presented in section 7.2.). By adding a single feature representing content similarity with prior contributions within the same thread from other participants to a baseline feature space, and by keeping all other aspects of the modeling technique constant, they produced an increase in agreement with human coding from 0.5 Kappa to 0.69 Kappa.

The unique contribution of the work I am presenting here is that it is not applied to text but to recorded speech. Although the speech data used in this study has been transcribed prior to the annotation process, the automatic analysis technique I describe does not use the transcriptions as input. Rather, the speech signal is first processed using basic audio processing techniques in order to extract features from segments of speech, which are then used for classification using a machine learning model. One might assume that the most straightforward approach would be to use speech recognition technology to transform a speech recording into an automatically obtained transcript and then to simply apply a model such as the one developed by Ai and colleagues (2010), which was applied to transcriptions of face to face interactions. However, even state of the art speech recognition is not advanced enough to use raw speech from meetings (Stolcke, Friedland, & Imseng 2010). Therefore, despite the great potential value in automatic transactivity analysis directly from speech (Joshi & Rosé 2007), it remains to be seen what level of accuracy is possible just from the speech signal itself.

The technique I evaluate in this chapter is related to prior work on speech processing for other classification tasks. There has been some prior research on automatic assessment of group interactions in the CSCL community focusing on speech as input (DiMicco et al., 2004; Gweon et al., 2009). For example, in chapter six (study seven), we saw that the amount of speech activity can be estimated from recordings of group meetings (Gweon et al., 2009). However, previous work was more focused on the amount of contribution from each speaker rather than on anything specific related to the nature of individual contributions. In contrast, some prior work has focused on the nature of conversational contributions in the language technologies community (Ang, et al., 2002; Kumar et al., 2006; Liscombe, et al., 2005; Ranganath, et al., 2009) All of this work makes use of signal processing techniques that are able to extract basic acoustic and prosodic features as reviewed in section 2.3.

Acoustic and prosodic features are frequently associated with intuitive interpretations that make them an attractive choice to use in baseline techniques for stylistic classification tasks. For example, higher pitch might indicate that the speaker is excited about his idea. Such interpretations are grounded in sociolinguistic work related to the way that speech style specifically (Coupland, 2007; Eckert & Rickford, 2001; Jaffe, 2009) and language style more generally (Fina et al., 2006) reflect both intentional and subconscious aspects of the ways a speaker positions him or herself within an interaction at multiple levels. These recent accounts build on decades of work beginning with Labov's research into speech characteristics that signal social stratification (Labov 1966) and Giles' work developing Social Accommodation Theory (Giles 1984), which describes how speech characteristics shift within an interaction, and how these shifts ought to be interpreted. Based on a simple interpretation of this work, I hypothesize that hidden within the speech signal are features that enable prediction of social meaning. The Ranganath work cited above related to detection of flirting supports this view. For example, it is possible to argue that, while the essence of transactivity is related to content level distinctions, it also has a social interpretation, and therefore might be detectable from speech as well. Consider that externalizations position students as intellectual leaders within a conversation. However, if true leadership requires that the leader be received as such by the other group members, and ICC contributions indicate that reception, then the occurrence of ICC contributions say something about the relationship between speakers. One can then expect that stylistic features that predict positive reception between conversational participants may also predict ICC. The easiest way to begin this research would be to begin with the types of features used in prior work, detecting social aspects of conversations from speech, such as flirting. In this study, I refer to those features as "prosodic" features. However, in addition to prosodic features, I also collected phoneme features, which give cues to the content of speech without looking at the actual words. Below, I present the specific procedure I followed.

7.2. Procedure

My technical approach consists of four main stages:1) collecting the audio data; 2) preparing the audio data by transcribing and segmenting the data; 3) manual assessment; and 4) automatic assessment that includes extracting features and applying machine learning.

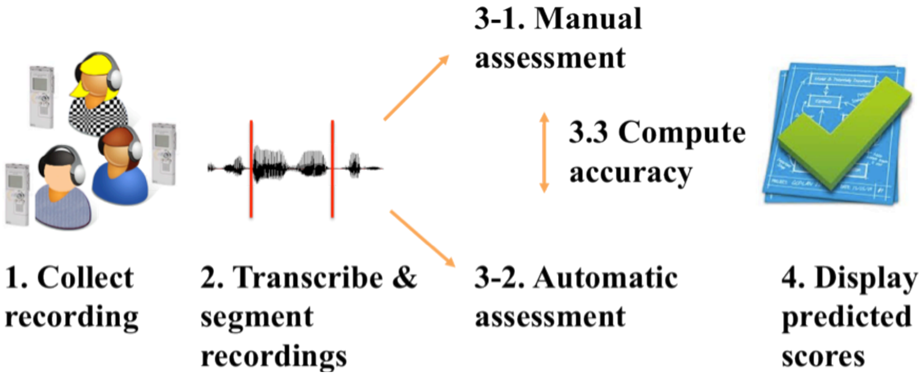

3-1. Manual assessment
3.3 Compute accuracy
1. Collect recording
2. Transcribe & segment recordings
3-2. Automatic assessment
4. Display predicted scores

Figure 7.1. Overview of the 4 state assessment process.

An overview of this process is presented in Figure 7.1. In this section I describe each step in more detail.

Collecting the audio data

The corpus was collected in a laboratory setting while students worked faceto-face in groups of three. In this chapter, I focus on a subset of data that has already been collected, transcribed, and annotated. Students here designing a contraption to protect an egg when falling from a distance of two stories, or two roughly two flights of stairs. This task involves applying a variety of principles from physics. The data set is from a 30-minute discussion portion of each 3-student group work session, wherein participants were to design and build a device meant to hold the egg together after it fell. In order to collect “clean speech” with each student on a separate channel, each student wore a directional microphone. This made it possible to clearly identify the main speaker from an audio file, as well as crosstalk from other members.

Transcribing and segmenting the audio data

For each audio file, the main thirty-minute discussion sessions were transcribed and manually segmented for further analysis. A total of 8 meetings were collected, transcribed, and segmented into 4361 parts according to the following two rules:

1. Begin a segment when the main speaker starts talking. If there is silence at the beginning of the file when the main speaker is silent, this means that there will be an “empty” segment in the beginning.

2. A segment should contain the main speaker’s continuous speech. If there is an interruption (silence or crosstalk) that lasts for more than 1 second, a new segment should be created. When you create a new segment, there should be two boundaries – one that marks the end of the main speaker’s first utterance, and another that marks the start of the next utterance after the pause.

Manual assessment

In this section, I present a brief description of how I operationalized the ICC process. The basic elements and the structure of the operationalization process were introduce earlier in study five (section 5.2). Here, I only present an overview of the process with examples from the current corpus as a reference. Recall that the operationalization is a two-step coding process. The first step is to mark a distinction between non-reasoning statements and reasoning statements. The second step focuses on the distinction between reasoning statements that represent new directions within a conversation (i.e., externalizations) from those that build on prior contributions (i.e., ICC contributions).

To detect an instance of reasoning, consider the following example segment. The segment shows a conversation with instances of displayed reasoning highlighted using italics.

> s1: *i think we'll need only one rubber band because the rubber band is circular. We can just break it off right*
>
> s3: oh yeah. that's a good idea
>
> s2: See what are the weights
>
> s1: *it is some significant difference*
>
> s2: *Yeah this is heavier. So this could be on top*
>
> s3: *yeah cause if we did that then that would fall on the bottom, right? It might do some spinning*

A simple way of thinking about what constitutes a reasoning display is that it contains two or more concepts connected by a relationship. The basic premise is that a reasoning statement should reflect the process of drawing an inference or conclusion through the use of reason. Reasoning statements stand in contrast to mere information sharing statements, which can be thought of as sharing rote knowledge. Table 7.1 shows the five types of concepts, which are the building blocks of a reasoning statement. For each concept, the definition and an example are illustrated as well. These examples are from the current dataset, where students are discussing the best approach to build an egg holder. Note that the "system" in this case is the egg holder, plus any materials that are available for use in building it.

Table 7.1. Definition and examples for the 5 concepts.

Type	Definition	Example
Theoretical concept	principles (i.e. physics principle) and theories	when an object is falling, the force of impact when it hits the ground can be decreased by slowing down the speed.
Prior Knowledge	information based on common sense	Using a small amount of tape would not be enough to hold two bowls together
Physical system properties	elements and characteristics of elements that are available for the system	paper bowl is round, straws are flexible
Emergent system properties	characteristics of elements that appear in a process	stability of an egg holder which emerges as a result of using certain materials
Goal	general believes or perspectives, anything associated with strong expectations related to points of view	aesthetics of an egg holder; i.e., trying to make the egg holder aesthetically pleasing

The other important part of determining whether a statement contains reasoning is the relationship between multiple concepts. For example, a simple list of concepts (e.g., this cup is round, and it is also white) is information sharing, and not reasoning. The two types of relationships that signal a reasoning process are: 1) Compare and contrast; and 2) Cause and effect.

- Compare and contrast, tradeoff: When the speaker compares two concepts, the speaker is making a judgment, which involves thinking about how two concepts are related to another.

    - The speaker compares two materials ("that" & "rubber band") for his solution: *"I am thinking that might work* <u>*better than*</u> *a lot of rubber bands."*

- Cause and effect: When the speaker uses a cause-and-effect relationship, the process involves establishing a relationship between two concepts through a reasoning process. The general relation in this category is "doing x helps you achieve y." There are three main types of causal relationship a) cause and effect b) in order to c) analogy. Examples for each of the three types are illustrated below.

    - Let's do A because of B: *"Let's use bubble wrap* <u>*because*</u> *it cushions the fall."*
    - Let's do A in order to achieve B: *"Let's use rubber bands* <u>*for*</u> *tying the bag onto the bowl."*
    - When a speaker makes an analogy, he is making a link due to the similarity between two concepts. Some of the keywords that signal analogies are "like," and "as": *"Oh, you're trying to use the bowl* <u>*as*</u> *a parachute."*

After determining whether a given statement contains reasoning, the next step is to distinguish between externalizations (new topic) and ICC contributions (building on prior contributions). Take the sample conversation I used earlier to illustrate a reasoning contribution. The lines marked with an (E) at the end is a contribution are categorized as externalizations, the ones marked with an (I) are ICC contributions. The first statement by s1 is an externalization since s1 starts a new topic, thus their contribution is not building on a prior contribution. Subsequent reasoning contributions in this discussion are coded as ICC because they each build on statements that directly precede them.

s1: *i think we'll need only one rubber band because the rubber band is circular. We can just break it off right (E)* s3: oh yeah. that's a good idea.

s2: See what are the weights

s1: *it is some significant difference (I)*

s2: *Yeah this is heavier. So this could be on top (I)*

s3: *yeah cause if we did that then that would fall on the bottom, right? It might do some spinning. (I)*

Two coders were trained using a manual that describes this operationalization. In a final evaluation of reliability for reasoning process, the kappa agreement was 0.72 between two coders over all data. After calculation of the kappa, disagreements were settled by discussion between the two coders. For the coding manual for detecting instances of ICC and externalization, the kappa agreement was 0.70.

Automatic assessment

After segmenting the data into units, the next stage involved transforming each segmented unit into a set of feature-value pairs. For the feature set, three types of features were extracted: 1) Acoustic features; 2) Phoneme features; and 3) Auxilary features. All three feature sets reflect *how* the words were spoken rather than the semantic content of the words.

Acoustic features capture certain structural aspects of speech; for example, amplitude, pitch, and energy. More intuitively, these features reflect the intensity and energy level of a given speech segment. For instance, a higher value of amplitude means a higher volume from the speaker. If there is variation in the amplitude, this indicates that the speaker's volume varied over time. I collected 4 amplitude features, which are the mean value of amplitude over the whole segment, as well as the mean, median, and variance of the one second windows in a given speech segment. Similarly, I extracted four pitch and four energy features: pitch/energy of the overall segment, mean, median, and the standard deviation of pitch over one second windows in a given segment. The pitch features were extracted using the YIN algorithm (De Cheveigné & Kawahara, 2002). In addition to these 12 features, I also used 28 of 40 Mel Frequency Cepstral Coefficients (mfcc). The initial 40 mfcc features are the result of applying a set of 40 standard filters, which are available as part of VoiceBox Matlab Toolbox (Voicebox, 2010). The mfccs are standard acoustic features that are commonly used in speech processing. They reflect the distribution of energy level in the given speech. Because using all these 40 features would capture somewhat redundant information, I took the top 28 features using principal component analysis (PCA). The decision to take the top 28 features was based on a rule of thumb that this number of features is sufficient for a variety of speech classification tasks of a roughly similar nature.

Phoneme related features are based on English phonemes, which are the smallest building block of sound in English that carries linguistic meaning. For instance, the phoneme that distinguishes the words tip and dip are the [t] and the [d] phonemes. Sphinx (CMUSphinx 2010), a speech recognition system developed at Carnegie Mellon University, identifies 48 phonemes in the English language. Thus, I used the 48 phoneme probabilities as part of the feature set. Using phonemes could capture certain aspects of content that would reflect the coding process used by human annotators or the structure of the language data. For instance, according to my operationalization of a reasoning statement, cause and effect relationships can be used to causally connect two concepts. Certain words, such as “because” or “for”, are often used in cause and effect relationships. Therefore, phonemes such as [b] or [f] may occur frequently in statements that contain reasoning contributions. In addition to the phonemes, a phoneme-count feature and phoneme rate were computed. The phoneme-count feature shows the total number of phonemes, which tells us how much the speaker spoke in the given segment. The phoneme rate feature is the number of phonemes divided by length of the segment, and it provides us with an estimate of how fast a person spoke.

In addition to the acoustic and phoneme features, other features were computed, namely duration of the segment, a speaker feature, and a feature that reflects stylistic language matching. The duration of the segment was the length of the given segment in seconds. The speaker feature was a binary feature, 0 if the speaker of the given segment is same as the speaker of the last segment, 1 otherwise. For the feature that reflects the stylistic language matching, I computed the Kullback-Leibler distance between phoneme probabilities, which is a measure of how different two distributions are from one another.

Once all features are extracted, I used the Adaptive Boosting machine learning algorithm (Freund & Schapire, 1995) to train a predictive model, and evaluated whether it was possible to automatically assign segments of speech as containing a “non-reasoning/ externalization/transactive” contribution with high enough accuracy to be useful. The Adaptive Boosting algorithm was designed to be resilient to noisy data and outliers because of the way it trains a model over multiple iterations, and the instances that are misclassified in early iterations receive more attention in the subsequent rounds through a reweighting mechanism.

7.3. Findings

In order to avoid the evaluation results being inflated due to overlap in speakers between train and test sets, I separated the data into two sets, each with a distinct set of students; specifically, a training set for building a model, on the one hand, and a test set for testing the accuracy of the model, on the other hand. Given the limited amount of data, I adopted a 10-fold cross validation methodology. This methodology involves averaging the performance obtained for each of the ten test sets. For each test set, 1/10 of the data is set apart as test data, and the remaining 9/10 of the data is used to build a model.

The results of the machine learning experiments are shown in table 7.2. In addition to the baseline, recall, precision, and f-score, the table also lists the top three most predictive features used for the prediction. Although the recall and precision rates may not seem very high, they are a significant improvement over only predicting the majority class.

For all three types of prediction, duration of segment was the top indicator. This result supports the idea that if a contribution from a group member contains reasoning, or ICC, or is the start of a new topic, it may be of longer duration because the speaker needs more time to express his thoughts. In particular, the second most highly ranked feature used for predicting the ICC process, phoneme “B,” could be due to the use of many words, such as “but,” or “because,” when a speaker is engaged in the ICC process.

Table 7.2. Machine learning experiment results.

Prediction	Baseline F-score	Recall (%)	Precision (%)	F-score (%)	Feature 1	Feature 2	Feature 3
Reasoning?	0.20	0.63	0.51	0.56	Length (27.8%)	Phoneme rate (6.5%)	12th PCA (5.2%)
ICC?	0.12	0.72	0.24	0.35	Length (17.8%)	Phoneme 'B' (18.2%)	12th PCA (6.7%)
Externalization?	0.08	0.70	0.22	0.32	Length (35.2%)	2nd PCA (4.7%)	9th PCA (4.3%)

7.4. Discussion

I began an investigation of the application of speech technology to idea co-construction detection using a straightforward application of prior work that detected social aspects of conversations from speech; e.g., such as flirting. This work shows promise that using machine learning, the classification of a statement as reasoning/non-reasoning is feasible, even with limited training data. However, results at distinguishing ICC contributions from others are still weak, especially with respect to precision.

Thus, there is more work to do to understand how ICC is manifest in speech, and in the next section I present an approach that uses more sophisticated adaptations of sociolinguistic work. Existing literature on sociolinguistics of speech style emphasize social interpretations of stylistic shifts within an interaction (Eckert & Rickford, 2001). For example, Social Accommodation Theory (Giles 1984) emphasizes the important function of stylistic convergences between speakers within an interaction. This work suggests that more complex features, computed over patterns of the types of acoustic and prosodic features used in this study may be more conducive to high levels of accuracy.

CHAPTER 8

Incorporating ideas from sociolinguistics to predict the ICC process from speech[5]:Study 9

The goal of this chapter is to develop a speech processing technique capable of predicting idea co-construction (ICC) contributions in speech. Recall that an ICC contribution is one where the reasoning is explicit, and that the reasoning builds on a prior reasoning statement within the discussion. As mentioned earlier, the automatic analysis of deep knowledge integration processes is not itself a new direction in the CSCL community. However, the unique contribution of my work is that it is one of the first approaches to detect the ICC process in speech; it is also the first that does so by leveraging insights from sociolinguistics. In this study, I use speech style accommodation to measure social factors that impact group interactions.

8.1. Motivation & design

In state-of-the-art approaches to applying machine learning technology to speech data, the speech signal is first processed using basic audio processing techniques, including the earlier work on detecting ICC from speech introduced in study 8. The signal is processed in order to

[5] Part of this work will be published in Gweon, G., Jain, M., Raj, B., Rosé, C. P. (2012). Predicting Idea Co-Construction in Speech Data using Insights from Sociolinguistics. In Proc. International Conference of the Learning Sciences (to appear).

extract features from segments of speech, which are then used for classification using a machine learning model. Specifically, in study eight, I used the acoustic and prosodic features typically used for predicting emotion in speech (Ranganath et al., 2009; Ang et al., 2002; Kumar et al., 2006; Liscombe et al., 2005). This research makes use of signal processing techniques that are able to extract basic acoustic and prosodic features; for example, variation and average levels of pitch, intensity of speech, or the amount of silence and duration of the speech. However, previous work has not considered the way the values of these features change over the course of an interaction, nor how that change itself is meaningful. It is social interpretations of changes in speech style characteristics that is the crux of my contribution in this chapter.

Acoustic and prosodic features are frequently associated with intuitive interpretations, and this makes them an attractive choice to play a role in baseline techniques for stylistic classification tasks. For example, increased variation in pitch might indicate that the speaker wants to deliver his ideas more clearly. Likewise, volume and duration of speech may signal that a speaker is explaining his ideas in detail, presenting *his* point of view about the subject matter. The difference between their work of is that theirs is based on insights about the ways of speech style specifically (Coupland, 2007; Eckert & Rickford, 2001; Jaffe, 2009) and language style more generally (Fina et al., 2006) that reflects both intentional and subconscious aspects of the way a speaker positions him or herself within an interaction at multiple levels. These recent accounts build on decades of research beginning with Labov's work on speech characteristics that also signal social stratification (Labov 1966) and Giles' work developing Social Accommodation Theory (Giles 1984), which describes how speech characteristics shift within an interaction, and how these shifts are best interpreted.

While the essence of transactivity is related to content level distinctions, it also has a social interpretation. For example, Azmitia and Montgomery (1993) have demonstrated that friends exhibit higher levels of transactive conversational moves than pairs who are not friends, moves that are operationalized in a way similar to ICC. Furthermore, it makes sense to consider than to build on a partner's reasoning, one must be attending to the partners reasoning in the first place, and deem it worth referring to in the articulation of one's own reasoning. My hypothesis is that we can expect a speech style accommodation to reflect these social processes, and thus predict the prevalence of ICC.

Defining Speech Style Accommodation

"Speech Style Accommodation" has its roots in the the social psychology of language, where the many ways that social processes are reflected through language, and conversely how language influences social processes, are targeted for investigation (Giles & Coupland, 1991). As a first step towards leveraging this broad range of language processes, I refer to one very specific process, which has been previously been referred to as "entrainment," "priming," "accommodation," or "adaptation" in other computational work (e.g., Levitan, Gravano, & Hirschberg, 2011). Specifically I refer to those moments that conversational partners shift their speaking style within an interaction to become more or less similar to their conversational partner.

"Accommodation" refers to the process of speech style convergence within an interaction. Stylistic shifts may occur at a variety of levels of speech or language representation. For example, much of the early work on speech style accommodation focused on regional dialect variation, and specifically on aspects of pronunciation, such as the occurrence of post-vocalic r in New York City, that reflected differences in age, regional identification, and socioeconomic status (Labov, 2010). Distribution of backchannels and pauses have also been the target of prior work on accommodation (Levitan et al., 2011). These effects may be moderated by other social factors. For example, Bilous and Krauss (1988) found that females accommodated to their male partners in conversation in terms of average number of words uttered per turn. Hecht, Boster, and LaMer (1989) also reported that extroverts are more listener adaptive than introverts, and so extroverts converged more in their data.

Social interpretation of Speech Style Accommodation

It has long been established that, while some speech style shifts are subconscious, some speakers may also choose to adapt their way of speaking to achieve social effects within an interaction (Sanders, 1987). One of the main motives for accommodation is to decrease social distance. On a variety of levels, speech style accommodation has been found to affect the impression that speakers give within an interaction. For example, Welkowtiz and Fledstein (1970) found that when speakers become more similar to their partners, they are liked more by partners. Another study by Putman and Street (1984) demonstrated that interviewees who converge to the speaking and response rates of their interviewers are rated more favorably. Giles and colleagues (1987) found that more accommodating speakers were also rated as more intelligent and supportive by their partners. Conversely, social factors in an interaction affect the extent that speakers engage in the first place, and some times chose not to engage in, at all. For example, Purcell (1984) found that Hawaiian children exhibit more convergence in interactions with peer groups that they like more. Bourhis and Giles (1977) found that Welsh speakers, while answering to an English surveyor, broadened their Welsh accent when their ethnic identity was challenged. Scotton (1985) also found that few people hesitated to repeat lexical patterns of their partners to maintain integrity.

Computational models of speech style accommodation

Prior research has attempted to quantify accommodation computationally by measuring similarity of speech and lexical features either over full conversations or by comparing the similarity in the first half and the second half of the conversation. For example, Edlund and colleagues (2009) measure accommodation in pause and gap length, using measures such as synchrony and convergence. Levitan and Colleagues (2011) found that accommodation is also found in special social events in conversation such as backchannels. They show that speakers in conversation tend to use similar kinds of speech cues, such as high pitch at the end of utterance, to invite a back channel from their partner. In order to measure accommodation on these cues, researchers usually compute the correlation between the numerical values of the cues used by partners.

When stylistic shifts focus on specific linguistic features, then measuring the extent of the stylistic accommodation is simple because a speaker's style may be represented within a one or two dimensional space, and its movement can then be measured precisely within this space using simple linear functions. However, the rich sociolinguistic literature on speech style accommodation highlights a much greater variety of speech style characteristics that could be associated with social status. Unfortunately, within any given context, the linguistic features that have these status associations, generally referred to as “indexical” features, are only a small subset of all the linguistic features that are being used by a speaker in some way. Furthermore, which features carry this indexicality are always specific to a context. So separating the socially-meaningful variation from variation in other linguistic features occurring for *other* reasons can be like searching for a needle in a haystack. To meet this challenge, accommodation is measured with Dynamic Bayesian Networks (DBNs) in our work.

The unsupervised Dynamic Bayesian Network Model allows one to model speech style accommodation without specifying the targeted linguistic features (more details on this model can be found in the methods section Step 3.2). Because accommodation reflects social processes that extend over time within an interaction, one may expect a certain consistency of motion within the stylistic shift. The insight behind this unsupervised modeling approach is that one can leverage the social reasons behind accommodation in a way that better captures the inherent structure within the speech. Specifically, one can leverage this consistency of style shift to identify socially-meaningful variation, without specifying ahead of time what particular stylistic elements are the focus.

8.2. Procedure

My hypothesis is that we can model complex social signals, namely ICC, by leveraging insights from sociolinguisitcs. I investigated my hypothesis at two different levels: (1) at the contribution level, by conducting machine learning experiments; and (2), at the session level, by conducting correlational analysis. The setup for predicting the idea co-construction process is outlined in Figure 8.1. The result of both contribution level and session level experiments (step 4.1. and 4.2) are presented in section 8.3.

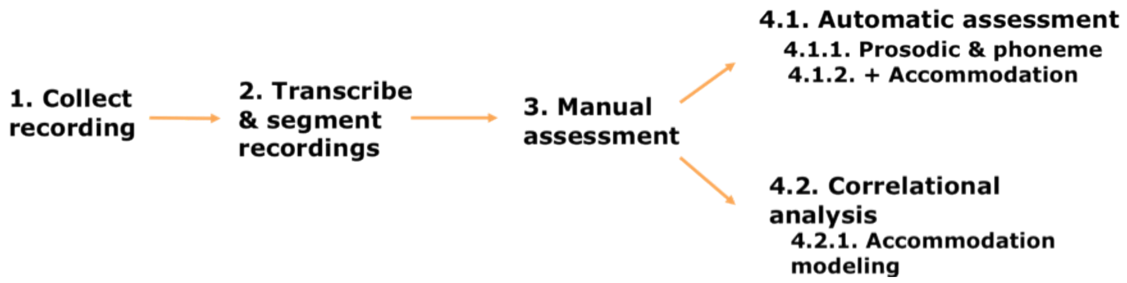

1. Collect recording
2. Transcribe & segment recordings
3. Manual assessment
4.1. Automatic assessment
4.1.1. Prosodic & phoneme
4.1.2. + Accommodation
4.2. Correlational analysis
4.2.1. Accommodation modeling

Figure 8.1. The Process Overview.

Step 1. Data collection using speech recorders

This corpus is taken from face-to-face debate discussions collected as part of research on arousal and learning (Nokes et al., 2010). The study was conducted in a laboratory setting where pairs of participants were engaged in a debate wherein they took opposing sides on a controversial topic. The specific task that the participants were asked to discuss was the cause of the decline of the Ottoman Empire, which has prompted some controversy among historians. One side of the debate emphasizes factors internal to the Empire, while the other side emphasizes external factors. Each of the participants were provided with background materials that support the idea of an internal or external cause, and were then asked to argue for their side. Each debate lasted eight minutes. The experiment had two conditions in terms of conversation patterns: blocked and freeform. In the freeform condition, the two speakers could talk freely for the duration of eight minutes. In the blocked condition, each speaker was given a chance to speak for two minutes in turn, resulting in two turns per speaker during the eight minutes.

Participants were male undergraduate students, between the ages of 18 and 25. In prior studies, it has been shown that accommodation varies based on gender, age and familiarity between partners. Because this corpus controls for most of these factors, it is appropriate for this experiment. Furthermore, because the participants did not know each other before the debate, one can assume that if accommodation occurred, it was only during the conversation.

In order to collect clean speech with each student on a separate channel, each student wore a directional microphone. Although it was possible to clearly identify the main speaker from the audio file, crosstalk, which is the other participants' voice, could still be heard in the background. A total of 66 sessions (with 132 participants) were collected and used for further analysis.

Step 2. Transcribing and segmenting the recorded data

For each audio file, each eight-minute discussion sessions were transcribed and manually segmented for further analysis. A total of 66 meetings were collected, transcribed, and segmented in 5490 parts, according to the following three rules:

- Independent-clause rule: a segmentation boundary should be placed as soon as an independent clause is identified.
- Dependent-clause rule: a sentence that cannot stand alone should be unitized with either the preceding or following unit.
- Analyze-from-beginning rule: sentences should be analyzed from the beginning of the sentence towards the end; i.e., match the subject of the sentence with the closest verb.

Step 3. Manual assessment (computing the amount of Idea co-construction)

Recall that a contribution containing idea co-construction (ICC) process is one where reasoning is made explicit, and that reasoning builds on a prior reasoning statement. Determining whether a sentence contains a reasoning statement is quite subjective – especially in the conversational sentences. In order clarify the concept of reasoning, I presented a two-step coding process in studies five and seven. A more thorough explanation of the coding process is presented in study five (section 5.2). Here, I present a brief description of the operationalization of ICC process.

The first step of the ICC process is to distinguish between non-reasoning statements and reasoning statements. The second step focuses on the distinctions between reasoning statements that represent new directions within a conversation (i.e., externalizations) and those that also build on prior contributions (i.e., ICC contributions). In addition to these two steps, an additional layer of coding was conducted for this corpus. A statement could be based on the speaker's own contribution (self ICC) as well as the other person's (social ICC). This third layer of coding was based on this distinction.

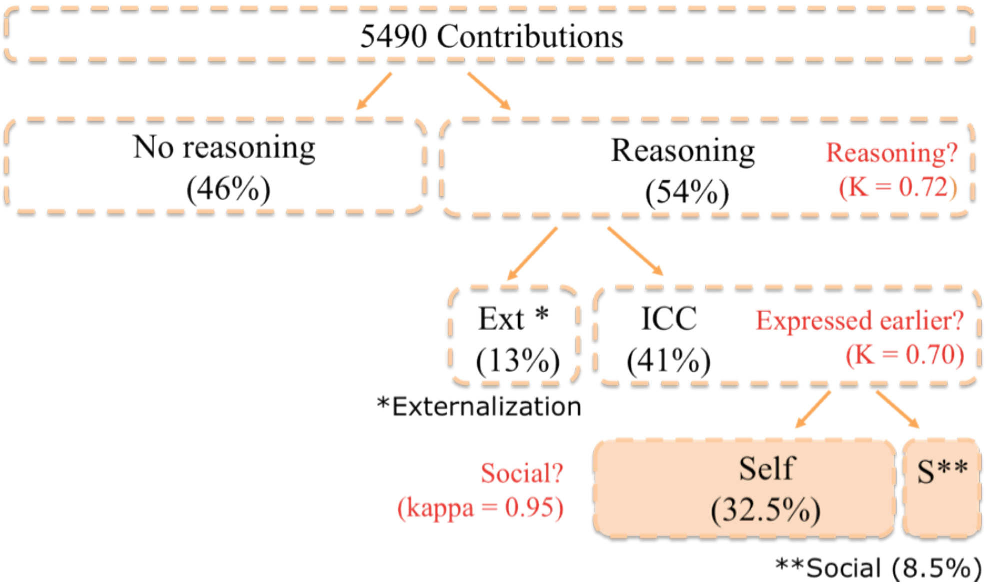


Figure 8.2. Three step coding process

To detect reason, I first looked for relationships between two or more concepts. This first concept could be prior knowledge, or one of the facts provided to the participants. The presence of multiple concepts in a statement by itself does not determine whether a statement contains reasoning; rather, the relationship between multiple concepts is itself the determining factor. For example, a simple list of concepts (e.g., population decreased) is information sharing, not reasoning. I identified two types of relationships that signal a reasoning process: 1) Compare and contrast; and 2) Cause and effect.

1. Compare and contrast, tradeoff: When the speaker compares two concepts, the speaker makes a judgment, which involves thinking about how two concepts are related to each other.
   - The speaker compares two time periods ("at the time" and "today"): *"At the time, if you look at the technology, it wasn't that advanced as we have today."*
   - When a speaker makes an analogy, he makes a link because of a similarity between two concepts. *"Outside powers were like the match lighting the fire."*

2. Cause and effect: When the speaker uses a cause-and-effect relationship, he establishes a relationship between two concepts through a reasoning process. The general relation in this category is "doing x helps you achieve y." Examples are illustrated below
   - A because of B: *"They forced the Empire to be economically dependent because they set up trading posts and banks."*
   - A in order to achieve B: *"Great Britain came in and introduced capitulations to control schools and health systems."*

Table 8.1 shows a segment of conversation from the corpus used in this study. The fourth column indicates whether the given contribution contains reasoning ("R") or no reasoning ("N"). The last column of the table is marked as either an externalization (E), or as an idea co-construction (I) for the statements marked as (R). The first statement by B is an externalization, since B starts a new topic; thus this contribution is not building on a prior contribution. Subsequent reasoning contributions in this discussion are coded as (I) because they each build on statements that directly precede them.

Table 8.1. Sample contribution

Line	Speaker	Contribution	R/N	I/E
14	B	I think that the economic downfall of the Ottoman Empire was due to internal problems because of the first World War uh, and other civil wars going on uh, beforehand which took place over the hundreds and thousands of years that people have been in that area.	R	E
15		Um, this lead to, these wars lead to population problems.	R	I
16		Uh, people were either being killed or they couldn't farm,	N	N
17		and if you can't farm, you can't feed people	R	I

This coding process was learned by two coders, initially trained using a manual that describes the above operationalization of reasoning displays and ICC, along with an extensive set of examples.

After each coding session, coders discussed disagreements and refined the manual as needed. Most of their disagreements were due to the interpretation of what the students meant rather than with the definition of reasoning itself. Therefore, later efforts focused more on defining how much the context of a statement could be brought to bear on its interpretation. In a final evaluation of reliability for reasoning coding, our kappa agreement was 0.72 between two coders over all of the data. After calculation of the kappa, disagreements were settled by discussion between the two coders. For detecting instances of ICC and externalization, the coding yielded a kappa value of 0.7. For the distinction between self oriented and social ICCs, the kappa value was 0.95.

Because the accommodation scores were computed for each eight minute session, I computed a comparable score by adding up the number of idea co-constructions for each session. This resulted in an average score of 36. The minimum score for a session was 22, the maximum score was 60.

Step 4.1. Automatic assessment

Two types of automatic assessment was conducted. The first was a replication of study eight (chapter seven) where prosodic and phoneme features were collected as features for a machine learning experiment. All the features reflect "how" the words are spoken rather than the semantic content of the words. The second type of automatic assessment was to add the accommodation score computed in step 4.2 as an additional feature. The hypothesis was that adding the accommodation score would improve the prediction of ICC compared to only using the prosodic and phoneme features.

Step 4.2. Accommodation modeling (measuring amount of accommodation)

This study evaluates measuring speech style accommodation as a way to predict the prevalence of ICC in debate-style conversations. In my joint work with Jain and colleagues (2012), we used Dynamic Bayesian Networks (DBNs) to measure amount of accommodation. The models were learned in an unsupervised fashion. What we are specifically interested in is the manner that the influence of one partner on the other is modeled. What is novel in Jain's approach is the introduction of the concept of an "accommodation state," or relational gestalt variable, which essentially models the momentum of the influence that one partner is having on the other partner's speaking style. This variable allows us to structurally represent the insight that accommodation occurs over time as a reflection of a social process, and thus has some consistency in the nature of the accommodation within some time span.

One problem with some of the work described earlier in section 8.1 (Computational models of speech style accommodation) is that it can seem to assume that one speaker influences their partner's style directly, and within an instant, as the floor shifts from one speaker to the next. But those approaches do not model any consistency in the manner in which the accommodation occurs. The major advantage of modeling consistency of motion within the style shift over time is that it provides a sign post for identifying what style variation within the speech is most salient with respect to social interpretation within a specific interaction. Therefore, the model may remain agnostic, and may thus be applied to a variety of interactions that differ with respect to those stylistic features that are salient. Jain developed six different models of increasing complexity, differing in the assumptions that are made about how one speaker's style influences the other speaker's. The technical details and inner workings of the model, as well as a validation that it is able to measure speech style accommodation, are described in a separate paper (Jain et. al, 2012).

Speech style information is reflected in prosodic features such as pitch, energy, and speaking rate. In this study, I leverage several of these prosodic speech features to quantify accommodation using Dynamic Bayesian Networks (DBNs). To extract features, speech from each session was segmented into a window of 50ms, with adjacent overlapping windows of 40ms. From each window, a total of 7 features were computed using the OPENSmile toolkit (OpenSmile, 2011). These features were voice probability, harmonic to noise ratio, voice quality, and three different measurements for pitch (F_0, $F_0{}^{raw}$, $F_0{}^{env}$), and loudness. Frame size within each window was set to 50 milliseconds, and frame step was set to 10 milliseconds. Next, a 10 bin histogram of feature values were computed for each of these features, which was then normalized to sum to 1.0. Each normalized histogram represents both the values and the fluctuation of the associated feature. For example, a histogram of the loudness feature captures the variations in the loudness of the speaker within a turn.

The model was run five times. To get a stable accommodation score, I used the average of the top three numbers that captures speech style accommodation best. To evaluate speech style accommodation, the performance of each model was determined based on how well it distinguished real conversations from constructed (fake) conversations. Because in real conversations speakers are expected to accommodate to each other's speaking style, a "good" model should capture the accommodation that occurs in real conversations, but not in constructed conversations. The possible range of accommodation scores ranged from 0 to 1, where 0 means there was no accommodation between the two speakers in a given conversation. The average accommodation score was 0.43.

8.3. Findings

Results from both the machine learning experiments (contribution level) and the correlational analysis (session level) are presented in this section.

Results from Step 4.1. Automatic assessment

I ran machine learning experiments to try to predict whether each clause is an instance of idea co-construction. Here, the amount of accommodation measured at the "contribution" level was the unit of analysis used for idea co-construction coding.

Table 8.2. Machine learning experiment using prosodic & phoneme features

Prediction	% correct	Baseline F score	F score	kappa
Reasoning?	68.74%	0.38	0.68	0.36
ICC?	64.3%	0.44	0.64	0.23
Social ICC?	91.55%	0.88	0.88	0
4 way	54.53%	0.29	0.48	0.24

I first extracted the prosody and phoneme features collected in study eight (chapter ten) to examine the accuracy of machine learning prediction. Table 8.2 presents these results. Next, I added the accommodation score as an extra feature to see if this feature could improve prediction. However, as seen from table 8.3, there was no significant difference.

Table 8.3. Using prosodic, phoneme & accommodation features

Prediction	% correct	Baseline F score	F score	kappa
Reasoning?	69.71%	0.38	0.70	0.39
ICC?	64.55%	0.44	0.64	0.25
Social ICC?	89.76%	0.88	0.85	0
4 way	56.5%	0.29	0.5	0.25

In hindsight, this lack of improvement made sense because of the unit of analysis. The newly added "accommodation feature," which is computed using the model described in section 8.2 (step 4.2), is based on models that capture accommodation at a session level rather than at contribution level. However, machine learning experiments were conducted at the contribution level. The accommodation scores mark instances when styles between two speakers match. In a conversation, speech styles between speakers may match for several consecutive turns. However, the place where accommodation *begins* is not distinguishable with this approach because accommodation will occur somewhere within the range where the styles match. Since we are trying to detect turns that contain ICC contributions, many of these turns that reflect style accommodation will not be instances of ICC. Adjacent turns, next to ICC contribution, will also reflect style accommodation. Therefore the contribution level accommodation feature did not turn out to be useful.

<u>Results from Step 4.2. Correlational analysis</u>

When we consider ICC at the session level, we do not have the problem just discussed in connection with ICC. Rather than trying to locate the exact position of ICC statements, what we are predicting is the prevalence of ICC. It makes sense to believe that extent of accommodation says something about the effort participants in a conversation are likely to be making towards building common ground. Sixty six sessions were available for analysis. This amount of data is not enough to run machine learning experiment. Thererfore, for the session level, I ran a correlational analysis. My hypothesis was that I would see a significant positive correlation between the prevalence of ICC and the amount of accommodation (step 4.2). In addition, I hypothesized that there will not be a significant correlation between amount of accommodation and reasoning/ ICC because reasoning and ICC are not social in nature. These hypotheses were tested by running correlational analyses. Over the 77 discussion transcripts discussed above, I counted the number of contributions that were coded as ICC, and labeled this sum Prevalence of ICC. Similarly, I labeled the score computed by the network model as Accommodation score.

Table 8.4. Correlational analysis in freeform condition

Model #	Acc & Reas	Acc & ICC	Acc & Social
1	0.15	0.11	0.36*
2	0.10	0.07	0.35*
3	0.12	0.07	0.30*
4	0.18	0.12	0.37*
5	0.18	0.13	0.36*
6	0.15	0.10	0.33*

The results of my correlational analyses are on Table 8.4. shows using the data from the freeform condition. Recall that the participants could talk freely during the eight minute session, thus this is a natural form of conversation. The rows in the table lists numbers from the six different models, each ordered by complexity. The second column shows correlations between the accommodation score and the amount of reasoning. The third column presents correlations between accommodation score and the amount of ICC. None of the correlations are significant in either the second or third column. However, as predicted, the correlation between the accommodation score and the amount of social ICC is significant, as shown in the last column of table 8.4. This correlational analysis shows that the accommodation score is capturing the social aspects of the conversation.

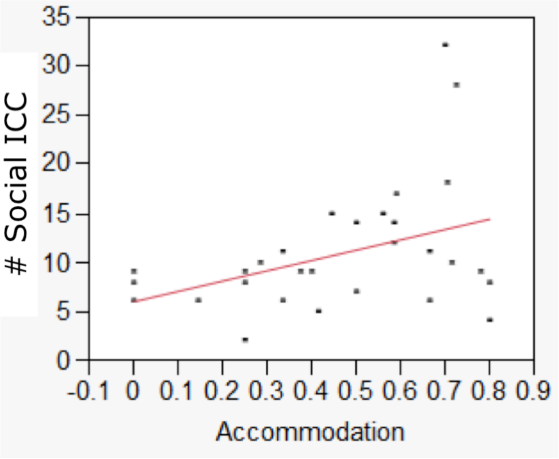


Figure 8.2. Correlation between Other transacts & accommodation

Figure 8.2 shows a graphical representation of the relationship between the accommodation score and social ICC in the best model (#4). The y-axis shows the social ICC scores and the x-axis shows the accommodation scores. The predictive linear regression line fits better at the lower end of the spectrum. Thus a nonlinear model may have fit the data better. However, due to small sample size, I used a linear model. The result shows promise in that, in the best model, accommodation can explain about 13% of the variance in the data ($r = 0.37$, $r^2 = 0.13$, $p < .05$).

Table 8.5 shows the corresponding correlations from the blocked condition, which is an unnatural form of conversation wherein speakers talk for 2 minutes each, without any interaction. As in table 8.4, the rows correspond to the six accommodation models, increasing in complexity as the model number increases. As before, there is no significant correlation for the accommodation and reasoning/ICC.

For the accommodation and social ICCs, the predictive value of speech style accommodation in these unnatural conversations is weaker than in the more natural condition. The best r value is 0.24, which is computed from model #4. This is the only model with a significant correlation, and it is a model that includes a modeling of the relationship between speaker's speaking styles.

A weaker correlation is expected because ICC is not expected to occur as frequently when a speaker is talking for the given two-minute block of time. In this condition, each speaker is only given two chances to speak. Therefore, unlike in a natural conversation, where the feedback loop is tighter, when speakers are focused on their own ideas for a block of time, they are less sensitive to the other speaker's style.

Table 8.5. Correlational analysis in blocked condition

Model #	Acc & Reas	Acc & ICC	Acc & Social
1	0.02	0.16	-0.01
2	-0.05	-0.01	-0.12
3	-0.04	0.06	-0.17
4	-0.01	0.04	-0.24*
5	-0.05	-0.03	-0.19
6	-0.02	-0.01	-0.21

In addition, all the correlations are negative in value which is the opposite relationship compared to the freeform condition. In other words, speech divergence is what predicts social ICCs in these unnatural conversations. One possible explanation for this divergence may be explained by the nature of the task. For this task, the speakers are not in a cooperative working environment, but are debating with each other, arguing for two opposing perspectives. In order to get their idea across in a short time, they might emphasize the differences in styles, because they are arguing for opposing ideas. Looking at the data, I found anecdotal evidence that social ICCs tended to occur in the blocked condition mainly when the participants were particularly antagonistic towards each other.

8.4. Discussion

In this chapter, I presented my work toward an automatic detection of idea co-construction contributions in speech data. Our need for a tool that predicts levels of ICC contribution has been demonstrated in the earlier studies, where I investigated the needs of instructors who teach project courses (studies one and two). The goal of this study was to develop technology to address these needs. To this end, I have demonstrated the feasibility of predicting ICC by modelling the stylistic convergence of speech. This research shows promise in that the model generated a predictor of the amount of idea co-construction (r = 0.37).

My investigation into the application of speech technology on idea co-construction detection was begun by using a straightforward application of prior work that detected the social aspects of conversations from speech, such as instances of flirting. Other work on sociolinguistics of speech style emphasize social interpretations of stylistic shifts within an interaction (Eckert & Rickford, 2001). For example, Social Accommodation Theory (Giles 1984) emphasizes the important function of stylistic convergences between speakers within an interaction. This work suggests that more complex features computed over patterns, of the types of acoustic and prosodic features introduced in this chapter, may be more conducive to higher levels of accuracy. For future work, more sophisticated adaptations of sociolinguistic work might suggest follow-up techniques. For example, I plan to investigate sequencing and timing rather than just quantity as adopted by Kapur and colleagues (2009). In terms of data, collecting and annotating audio data from additional meetings, as well as other contexts, would validate these results further and test its generality across a wider variety of student groups and contexts.

Many things could be done to improve the performance of the automatic assessment work featured in this dissertation. Currently, I am using surface level speech features along with simple style matching. However, more complex types of style matching can be implemented to improve performance. We already know that reliable detection of idea co-construction is possible using machine learning technology when applied to text based interactions (Rosé et al.,

2008) and transcriptions of face-to-face group interactions (Ai et al., 2010). Therefore, adding more semantic information from the content of those speeches could also improve performance.

CHAPTER 9

Conclusions

The main goal of my dissertation is to support discussion processes in collaborative group work, especially in environments where innovative solutions to difficult engineering design problems are generated. To support group work processes in project-based learning environments, I started by investigating the type of assessments that instructors value most in providing guidance to student groups (study one). Next, in study two, instructor needs and current assessment practices were evaluated in a classroom study using the five types of assessment dimensions identified in study one. Based on the feasibility of performing the identified assessments by human annotators, and the need for automatic assessment, I focused on automatically tracing one of the assessment categories (the knowledge co-construction process). However, before addressing the technical challenges, the importance of idea co-construction was evaluated (study five).

Ultimately, my vision was to automatically trace the five pairs of assessment categories identified in study one using machine learning technology. In this dissertation, I have presented methods for automatically assessing predictors that are indicative of five group processes using speech (study seven). Given my initial success, the final two studies examine methods for enabling the monitoring of idea co-construction (studies eight and nine). The results from studies eight and nine show promise for using speech to detect the ICC process.

My dissertation work makes contributions to the fields of human computer interaction, computer supported cooperative work, cognitive and educational psychology, communications theory, conversational analysis, and computational linguistics. The contributions span theoretical, methodological, and practical arenas, and a more detailed discussion about my contributions are given in sections 9.1, 9.2 and 9.3. Lastly, I present possibilities for extending the research presented in this dissertation, as well as more general research directions that are of interest (9.5).

9.1. Instructor needs in problem based learning environments – studies 1 & 2

I started out this research with the simple observation that group assessment in project courses is far from ideal. This was confirmed in my interview study (study one), where instructors reported their concerns and dissatisfaction with the level of insights they have into the teams they oversee. Five pairs of group processes were identified, which the instructors of project group courses value in evaluating and guiding student teams. I further investigated the extent of the issues identified through self-reporting in the interview study, in the context of a classroom study in an actual project course (study two). In the classroom, I collected quantitative evidence that showed there was a mismatch between the assessments of instructors and assessments of lay observers who were sitting in on group meetings. A mismatch by itself, in this context, would not necessarily show a deficit on the part of the instructors. However, the fact that instructors' assessments along different dimensions were highly correlated, when they were meant to be differentiated, is a concern. This did provide some evidence that instructors are having trouble evaluating group processes identified in study one.

The research questions and contributions addressed in each of the studies are:

3. Study one: What do instructors want students to do and learn in project-based learning environments?
    - The formalization of assessment criteria in the context of project courses (theoretical).

4. Study two: How much do instructors really know about what is going on during group projects?
    - The evidence that instructors could benefit from assessment support (theoretical).
    - The instrument for measuring instructor and observer perspective along the five pairs of assessment categories (methodological).

Other limitations exist in my work due to the nature of the data. The interview study was conducted in a single university, Carnegie Mellon University, which is located in the eastern part of the United States. In addition, the data for the classroom study was collected from an engineering course where students worked on learning the process of conducting a group project. Therefore, the findings may not transfer to a different cultural context. The findings also may not generalize to other types of project courses, to a different domain, or to a context where students are evaluated more on the result of the project rather than on its process. Data collection in another type of project course, and from different areas, is still needed in order to test whether these results can be generalized further.

9.2. Identifying conditions that support the knowledge co-construction process – study 5

While the first two studies focused on a wide variety of group processes, I chose to focus my subsequent work on the knowledge co-construction process. This process was one of the assessment categories identified as important to instructors in study one, and problematic from an assessment standpoint, both based on self-report from study one and quantitative analysis from study two.

In study five, I chose to take this research in a new direction, namely I sought to form a bridge between the fields of computer-supported collaborative learning and collaborative work. In study five, I demonstrated a positive correlation between idea co-construction and one important group outcome; namely, knowledge transfer, which is associated with group efficiency.

The research questions and contributions given in study five are:

- Study five: Are students more likely to transfer knowledge when they engage in the idea co-construction processes?
    - The evidence that the idea co-construction (ICC) process is positively associated with knowledge transfer (theoretical).
    - The reinterpretation of ICC from a social perspective (theoretical).
    - The measuring ICC for a given engineering design task (methodological).

The results of the studies have a number of limitations that restrict the findings to the these tasks and this environment. In particular, study five, which I used to establish the idea co-construction, its role, and its affect on various types of group work outcomes, places an important limitation on the generalizability of the results. I selected a relatively simple task that could be completed within a single session. Thus results might differ if groups are working on a more complex task that requires in-depth expertise or multiple meetings. In addition, the generalizability of the result to a real-world task, that is, one outside of a laboratory environment, should be further studied. Additional limitations include the diversity of participants that we recruited for the study; e.g., gender composition, culture, and language.

9.3. Automatically monitoring group processes – studies 7, 8 & 9

In the technical portion of this dissertation (studies 7 ~ 9), I developed techniques that take text and raw speech as input and provide assessments as output. I worked on a variety of machine learning models that can be used to make these kinds of assessments, which can then be used to provide reports to instructors.

The research questions and contributions addressed in each of the studies are:

- Study seven: To what extent can the rudimentary features extracted from speech recordings of collaborative discussions be used to predict various group processes, including the process of knowledge co-construction?
    - A method for addressing the problem that respects privacy and compensates for shortcomings in accuracy of state-of-the-art speech recognizers by using prosody of speech rather than content.
    - The automatic assessment of speech activity, average activity level, and overlap for individual students using speech recordings (technical).

- Study eight: To what extent can the rudimentary features extracted from speech recordings of collaborative discussions be used to predict where idea co-construction is occurring in those discussions?
    - The advance of automatic analysis of idea co-construction by identifying important aspects of speech dynamics that can have predictive value (theoretical).
    - The automatic assessment of the idea co-construction process (technical).

- Study nine: Can the insights from sociolinguistics be used to improve the prediction of where idea co-construction occurs in collaborative discussions?
    - The computationalization of theories from pragmatics and sociolinguistics (theoretical).
    - The connection between “language technologies” and “sociolinguistics” communities.

One source of limitation was in the speech data collection. Because speech was collected in real meetings rather than in laboratories, the quality of speech was not the best for extracting speech features. For instance, one could sometimes hear voices of other members of the group in the recording of a particular student. Additionally, because I asked students to turn in recordings (study seven), data from several students went missing when students failed to follow through. In future data collection efforts, these problems can be minimized by asking students to wear microphones not too far away from the mouth, and by having the research staff upload recordings so that we do not loose data.

9.4. Extending current research

The research goal of this dissertation is to support discussion processes that occur in project-based group work environments. Based on the demonstrated need and importance of group processes identified in such an environment, I presented an approach for automating assessments using conversational data collected during group work. Given that automatically assessing five different valuable processes is beyond the scope of this Ph.D. dissertation, I narrowed the scope of my research to an examination of the process of idea co-construction, which is itself a type of knowledge sharing process. What is interesting and unique about the automated process examined in this dissertation is that non-content (prosodic and phonological) features are used to detect a complex process that is not obviously related to these prosodic speech features. This approach was motivated by considering the social implications that underlie conversations, which falls under the domain of sociolinguistics. Indeed, the results presented in the last chapter (study nine) show promise that, by capturing the social nature of conversation, one can also capture certain aspects of the idea co-construction process.

Given my initial success, there are multiple directions that could be pursued in my subsequent research. The first would be to assess the other four types of group processes identified as valuable in study one; namely, goal setting, progress, participation, and teamwork. The framework presented for assessing knowledge sharing does not use content to predict the occurrence of knowledge sharing instances. Not using words as the sole object of study has two advantages: 1) The approach is not domain specific; and 2) The approach can be used to assess group processes that are more social in nature.

In terms of increasing the accuracy of assessment for ICC/knowledge sharing, additional methods for capturing stylistic accommodation could also be used. For example, one could see how each prosodic feature, such as pitch, differs from the average value of the feature that would be typical for the context. Similarly, one could examine the shifts of prosodic features themselves, regardless of convergence or divergence, and that could be used to predict ICC. In both of these cases, the marked nature of the value of the feature could signal a social interpretation.

9.5. Future research directions

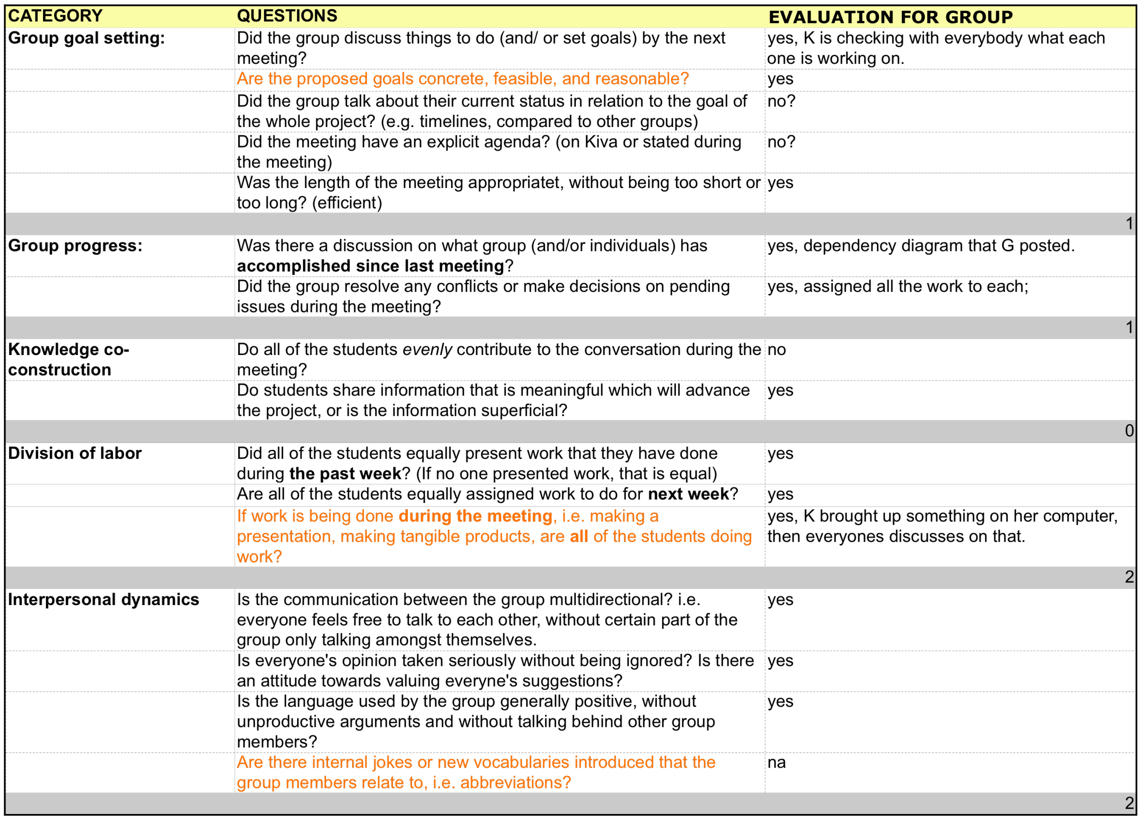
My general research interest is in better understanding and supporting group processes that impact collaboration. I believe that my work has many possibilities for scientific exploration and in designing systems that could have real world impact. So far, I have looked at processes that impact learning or knowledge transfer, which is related to learning. Two research questions that I would like to pursue further are: (1) How can we represent speech so that patterns that indicate social processes from conversations are learnable by machines?; and(2), How can we use non-verbal behaviors to build a technical infrastructure that supports collaborative work?

Given the advancement in computing power and algorithms, researchers are already examining ways to support social behavior through computational modeling. The research presented in this dissertation, which uses machine learning to predict whether a contribution contains ICC is just one example. Machine learning technology allows computers to classify data automatically. Thus machine learning is a great technology to leverage the plethora of data that can be made available to us in this modern day and age. However, machine learning is useful *only* if appropriate algorithms are provided. Identifying meaningful patterns requires deep insight into data, which I have tried to gain via conversational analysis. Conversational analysis can be used to predict other types of social processes, too. In particular, I am interested in social processes that impact human well being, such as those that build confidence or provide emotional support.

In my work thus far, I have leveraged speech data to build a technical infrastructure that supports collaborative work. More specifically, I used non-content aspects of speech to predict the idea co-construction process (Gweon, et al., 2011b; Gweon, et al., 2009b). My approach was based on previous research that established the presence of social cues in non-content aspects of speech. Similarly, other non-verbal behavior, such as gaze, gestures, or facial expressions have implications on human social behaviors. Ultimately, I want to leverage speech data and non-verbal cues to build technical systems that can support collaborative work.

APPENDIX A

Instrument for measuring the five group processes from study two



CATEGORY	QUESTIONS	EVALUATION FOR GROUP
Group goal setting:	Did the group discuss things to do (and/ or set goals) by the next meeting?	yes, K is checking with everybody what each one is working on.
	Are the proposed goals concrete, feasible, and reasonable?	yes
	Did the group talk about their current status in relation to the goal of the whole project? (e.g. timelines, compared to other groups)	no?
	Did the meeting have an explicit agenda? (on Kiva or stated during the meeting)	no?
	Was the length of the meeting appropriatet, without being too short or too long? (efficient)	yes
		1
Group progress:	Was there a discussion on what group (and/or individuals) has **accomplished since last meeting**?	yes, dependency diagram that G posted.
	Did the group resolve any conflicts or make decisions on pending issues during the meeting?	yes, assigned all the work to each;
		1
Knowledge co-construction	Do all of the students *evenly* contribute to the conversation during the meeting?	no
	Do students share information that is meaningful which will advance the project, or is the information superficial?	yes
		0
Division of labor	Did all of the students equally present work that they have done during **the past week**? (If no one presented work, that is equal)	yes
	Are all of the students equally assigned work to do for **next week**?	yes
	If work is being done **during the meeting**, i.e. making a presentation, making tangible products, are **all** of the students doing work?	yes, K brought up something on her computer, then everyones discusses on that.
		2
Interpersonal dynamics	Is the communication between the group multidirectional? i.e. everyone feels free to talk to each other, without certain part of the group only talking amongst themselves.	yes
	Is everyone's opinion taken seriously without being ignored? Is there an attitude towards valuing everyne's suggestions?	yes
	Is the language used by the group generally positive, without unproductive arguments and without talking behind other group members?	yes
	Are there internal jokes or new vocabularies introduced that the group members relate to, i.e. abbreviations?	na
		2

Table A.1. Group assessment form

CATEGORY	QUESTIONS	EVALUATION FOR EACH STUDENT				
		K ☺	R ☺	G ☺	W ☺	A ☺
Personal goal setting:	Is the student suggesting next steps (plans, high level steps) for himself or for the team?	yes, let's upload all then put them into MS word;asked A to do a design flow diagram on Monday; keeping track of what everyones' working on.; briefing the things to do for G.	no, speech and contact recognition + partially communication standard (asked by K, if he wants to be in) then he doesn't care about it. More question about time; Wizard of Oz interface? -- or Rosie.	no, ??	yes-> no, ? + communication standard; ?? Getting prescription?	no, avatar+design flow diagram
	Are the proposed steps concrete, feasible, and reasonable?	yes, explain what's the design flow diagram to A	na	na	na	na
	Did the student state (know about) his responsibility for the next meeting?	yes, went over to check the list of items that are assigned each student in the team.	yes	yes	yes	yes, designing Avatar
		2	-1	1	2	1
Personal progress:	Were there items that he finished (exclude relaying information as a liason) during the past week?	yes	no	no	no, ???	no
	Did the student finish **all of** his responsibility? (yes, if student has no responsibility)	no->yes, we'll work on state diagram that we haven't finished last night.(w/ whom)	yes	yes	yes	yes
	Did the student finish his responsibility **on time**? (answer only if students had responsibility)	na	na	yes?, dependency diagram	na	na
		0	0	0	0	0
Knowledge contribution	Did the student present **new ideas and solutions** for problems being discussed during the meeting?	yes, don't care about either of one. Instead, gesture over joystick has been developed. We might use that.	no	no->yes	no->yes	no->yes
	Did the student present any **concerns** (not clarifying questions) during the meeting?	yes	no	no	yes, what about 3G card vs. wi-fi?	no->yes
	Did the student volunteer to use his professional skill to solve a problem being discussed?	yes->na, explain ? On the whiteboard what I'm picturing about state diagram.	na	na	na	na
		2	-2	-2	0	-2
Participation	Did the student come to the meeting on time? / left the meeting on time?	yes	no	no	yes	yes
	Did the student speak during the meeting?	yes	yes		yes	yes
	Does the student seem engaged in the meeting by giving full attention to the meeting itself? For example, NOT playing games or NOT checking email constantly on a computer.	yes	no, doesn't want to make eye contacts with any others.(doesn't seem that he's making a note)	yes	yes	yes
	Did the student volunteer for work?	na, asked everyone if he wants to join any of other works.	na	na	na	na
		2	-1	1	2	2
Team Player	Is the student addressing the whole team instead of addressing part of the team?	no	yes	yes	yes	yes
	Did the student respect others opinion by allowing them to speak/ respond?	yes	yes	yes	yes	yes
	Is the student making productive argument without talking negatively?	yes	yes	yes	yes	yes
		2	2	2	2	2

Table A.2. Group member assessment form

APPENDIX B

Coding manual for measuring ICC (Origami) from study five

Reasoning Coding Manual

Our goal is to detect instances of reasoning statements. Because the concept of reasoning is somewhat abstract, this section of the manual presents guidelines for detecting instances of it.

Reasoning Contribution

In order to be considered to be a reasoning statement, the statement should combine two concepts with a relation. There are five types of concepts, which are detailed in section **<CONCEPT>.** For the origami task, the most frequently used concepts are a type of "theoretical concept," which are the two different methods that the students learned (X/Z's method, and Y's method). There are two types of relationships, which are detailed in the current section. R1. Compare and contrast, R2. Cause and effect. Sentences marked [R] are reasoning statements, sentences marked [N] are not reasoning statements.

Headings & Examples in BLUE are [R].

Headings & Examples in RED are [N].

<R1.Compare and contrast, juxtaposition, tradeoff>

We are interested in reasoning statements that are used to determine whether the team will use the current method, or the other team's method. Therefore, often a reasoning statement contains compare/contrast.

- 756_X: Is that the end? [N, Asking it's end of Y's method]
  756_Y: Yeah [N]
  756_X: No wonder they had so many. [R, *X compares Y's & his method* and reasons that *because* their method is shorter, Y's group produced more boats]

<R2. Causal>

Some contributions contain causal reasons to support the speaker's argument. Such statements are often signaled by used of words such as "because," "in order to," and "like."

- 754_Y: It might be better [N]
  754_Y: now that I've figured out what I am supposed to do. [R, Y says X/Z's method might be *better compared to* Y's method *because* Y has learned the new method. note that the previous line is N because the reasoning statement is not completed until this second line]

NOTE: Directions, or questions about directions, are not counted as reasoning.

- 758_Z: You fold diagonally, so the square goes like that. [N, directions for steps involving X/Z's method]
- 747_Z: Like a square base on each side? [N, asking about directions for the method]
- 741_Y: Maybe when I'm almost done with mine, you can pass it over [N, instruction on when to pass over the boat, doesn't involve "why"/ reasoning]
- 640_X: So go ahead and do your step and I guess we'll see what comes out [N, direction]

BUT interpretation of directions would be R.

- 740_Y: What do you guys lose points for? [N]
  740_X: Just if we made them too fast or tried to finish them in the last two seconds. [R, interpretation of instruction]

<General rules>

1. Be conservative: If you cannot understand what the speaker is saying, be conservative. In other words, if the interpretation of a contribution is not clearly a reasoning statement, then you should NOT mark the segment as a reasoning statement.

- 751_Z: You take two corners … [N]
  751_X: Like that, exactly. [N, not clear if this was used as a monitoring action (which would be N), or a judgment on whether or not the method used was correct (which would be R)]
- 748_Z: I wonder which one is faster? [N]
  748_X: We got through three and four, and he got through nine, so … [R, "his method is faster", is implied]

748_Z: Maybe we should just do this thing [N, not clear what "this thing" is. Is it the unfinished boat? X/Z's method? Y's method?]

- 748_Y: It's a sailboat (BB), but it's not the best [N, not sure what Y is trying to say. Y should follow up with what to do as a result of BB not being the best for this statement to count as reasoning]
- 746_Y: I don't know what the last step is [N]
  746_X: She has to make sure it looks perfect [N, not sure why X said this]

2. Completion given context: If the sentence is not completed, but it is clear what the speaker is saying given the context, then mark the segment as a reasoning statement.

- 748_Z: I wonder which one is faster [N]
  748_X: We got through three and four, and he got through nine, so… [R, "his method is faster", is implied]
- 751_Y: She hasn't called 1 minute yet so.. I think there's still.. [R, "there's still time" is implied clearly from context. Y is providing reason for why they should keep doing their work]

3. Repeated reasoning statement: If the reasoning statement is same as a previously stated reasoning statement, it should not count again. Such reasoning statement should be almost verbatim, and intended as a repeat.

747_X: I think you learned a different way ... [R]
747_Z: You learned a different way [N, already said above]

- 746_Y: It was just easier on the other side [R]
  746_Y: But that's okay [N]
  746_Y: We got the easier part [N, repeated reasoning]
- 754_X: That's different because they taught us with shutters, so ... [R]
  754_Y: Shutters? We were taught differently ... [N, repeated reasoning]
  754_Y: I think we did it differently. Obviously [R, Here, although Y is also talking about noting the difference between two methods, the statement is not verbatim, and so "obviously" signals that it's not an exactly like the earlier statement]

NOTE: Note that the examples below would NOT count as repeat.

- 748_Y: Totally different [R]
  748_X: Is it? [N]
  748_Y: Yeah [R, this is not a repeat because it's not verbatim. Furthermore, in order to answer this question Y had to evaluate two different methods]
  758_Z: Oh wait, you're folding it… I guess they learned how to do it differently? [R]
  758_X: Oh, I think we are doing a different thing [R, "Oh" signals realization, not quite verbatim]

<Typical places when reasoning occurs in the Origami task>

Given the domain of comparing origami methods, the two main places where reasoning occurs is: 1) the realization that the two methods are different; and 2) when comparing two methods/steps and noting the similarity/difference. Here are some examples:

1. The realization that the two methods are different.

- 758_Z: Oh wait, you're folding it. I guess they learned how to do it differently? [R, this statement was made as the result of comparing the two methods]
  738_Z: Interesting [N, too vague, not sure if this was said as a result of noting difference or not]

2. Comparing two methods/steps and noting the similarity/difference

- 747_X: Now I have to go faster [R, in order to say "faster" you need to compare two methods]
- 751_Z: Your folds are better than hers [N, "better" is too vague. If he specified why e.g. faster, prettier, it's R]
  751_Z: It's easier to do yours [R, in order to say "easier" you need to compare two methods, also the previous line was used as the reason why it's easier]

If the speaker has to evaluate whether the step was done correctly, then it's R. However, if the speaker is only confirming what the next step is then it's N (refer to **NOTE** section below)

- 754_Y: How far am I supposed to go, do I fold it then like this? [N]
  754_X: Yeah. And then fold the other side. [R, "yeah" is a statement made as a result of noting the similarity between his suggested method and what Y did. Note that Y asked for a critical evaluation]
  754_Y: I think we did it differently, obviously [R, Y notes that there are two different methods]
- 756_Z: Does it matter how far up you go? [N]
  756_Y: Just probably up a little, like enough to have a base [R, correction. Needs to evaluate how far up you should go.]
- 747_Y: And then you just fold the bottom up like this, just halfway up, that's perfect [R, that's perfect is an evaluation]
- 740_X: Did he have the same folds or what? [N]
  740_Z: No [R, answering this question requires the evaluation of whether both methods have "same folds"]

NOTE: Noting the similarity/ difference must involve critical evaluation/ judgment. Acknowledging the status quo, confirming instructions, or monitoring statements are not examples of reasoning.

<acknowledging status quo>

- 751_Y: So we're making different sailboats [R, noting the difference in methods]
  751_X: Yeah, there's a little trick they ran [N, acknowledgement, Y did not ask for a critical evaluation/judgment]

- 751_Y: Yours is a lot more complicated [R, noting the difference in methods]
  751_Z: Yeah [N, acknowledgement, Y did not ask for a critical evaluation/judgment]

  <confirming instructions>

- 748_Y: So that's what I need to do, and pass it to you? [N]
  748_Z: Yeah [N, confirming the instruction, Y asked for instructions, NOT an answer that requires a critical evaluation/judgment]
- 756_Y: Push it in like this? [N]
  756_Z: Yeah [R, in order to answer Y, Z had to evaluate the extent that the paper should have been pushed in to make a proper boat]
  756_Z: They gotta keep touching?
  756_Y: Yeah [N, confirming instruction. In order to answer Z, Y needed a visual inspection, but did not require evaluation of the method Z was using]
- 748_X: So I do the same step right? [N]
  748_Y: Yeah, you do the same step [N, instruction confirmation]
- 747_Z: And now bring it together? [N]
  747_Y: Yeah [N, Previous question was a confirmation question]

  <Monitoring action, observations>

- 751_X: And then the bottom part ... just like that [N, monitoring actions. Does not involve critical evaluation]
- 751_X: These folds are fast to make [observation. If it was "faster," it would involve comparing two different methods, so it's R]
  751_X: So we don't have to accumulate any [R, previous line was used as the reason why they don't have to accumulate]
- 748_Y: That's pretty fast [N, observation]
- 745_Z:Oh it's a 2 sail [N, observation]

3. Deciding whether to make more boats once the one minute warning is given.

If it's clear that the decision to make an additional boat was considered only in the context that there's one minute left, mark it [R].

- 751_Y: She hasn't called 1 minute yet so ... I think there's still ... [R, "there's still time" is implied clearly from context. Y is providing reason for why they should keep doing their work]
- 751_X: Wow, we really could have done another one I think [R, "given that we had this much time" is implied from the context, clearly shows X considered the feasibility of making another boat given the time]
- 747_Y: Let's try to get one more [R, "given the time" is implied]

- 746_X: Should I make another one? [N]
  746_Y: What do you think? [N]
  746_Y: I could probably do another [R, “given the time”]
  746_X: No, I’ll stop [R, “give the time”]
- 740_Y: Two more [R, This is said when one minute warning is given]

If it’s unclear whether the reply was based on a reasoning, follow the general rule (Be conservative) and mark [N]

- 751_X: I don’t know, should we do another one? [N]
  751_Z: Go for it. [N, not clear if this was a response made without a judgingwhether it’s possible to make another one (which would be N), nor did they consider the feasibility of making another one (that would be an instance of R))
- 745_X: Should I fold another or not? [N]
  745_Y: Yeah, quickly [N, ambiguous if Y is saying that X should fold another one OR saying that Y said it to facilitate process (i.e., yeah ... whatever ... keep on folding quickly) …]

<CONTEXT>

For each reasoning statement, we need to locate two concepts and a connection between them as defined on page two of this manual. In some cases, one of the concepts is not explicitly mentioned in the current contribution, but is mentioned in a previous contribution that is still “salient.” A concept is “salient” if one can find it within a current context. The following paragraph defines what a context is for a given contribution.

<Since each session is only 4 minutes long, we can consider the whole 4 minutes as context for the origami task. Therefore, within the 4 minute window, every supporting argument counts as reasoning>

- 741_Y: We did it a lot simpler [R]
  741_Y: Can we change our strategy? [N, the reasoning part is previous line, this is suggestion]
  741_Y: Actually it’d be kind of easy to teach [R, a supporting reason for “changing the strategy”]

<CONCEPT>

ON TASK/DOMAIN CONCEPT: Domain concepts are concepts that are used to solve the design problem, e.g. designing the egg holder. These concepts can be any of the five types described below.

1. **Theoretical concepts**: Theoretical concepts are principles that students can apply in decision making; e.g., knowledge about the origami steps.

2. **Prior Knowledge** : Prior knowledge is pieces of information that help decision making based on common sense.

3. **Physical System properties**: Elements or characteristics of the elements that are available for the system; e.g., paper is difficult to fold

4. **Emergent System properties**: characteristics that appear in the process of doing origami; e.g., the properties of the boat.

5. **Goal**: The "label" is a commonly used term that represents a set of beliefs, a general perspective, or anything that is associated with strong expectations related to points of view. Label incorporates unstated opinions or perceptions. For origami, it's efficiency, or making as many boats as possible.

<OUT OF DOMAIN>

Any contribution that is not related to the design activity would be considered out of domain.
Activities that are clearly not part of doing the design exercise are off task. For example, answering a phone call, dinner suggestions, or asking about the payment for the study are all out of domain.
In addition, statements that do not directly contribute to solving the design problem should also be considered out of domain; e.g., coordinating for the next meeting time, telling jokes, expressing feeling, or asking for clarification of the instructions for the study would all be off domain.

Idea co-construction (a.k.a. Transactivity)

Our goal is to detect instances of idea co-construction contributions, which characterize the knowledge integration process in conversation during group work. Idea co-construction occurs when individuals use their knowledge to operate on the previous reasoning of their partner, or to clarify their own ideas. Such a consideration of another perspective (of the other's or of the self) is what makes idea co-construction statements valuable. We believe that these contributions are important in creating new knowledge, and thus developed this manual to detect idea co-construction.

Headings & Examples in BLUE are [I].

Headings & Examples in RED are [N].

Idea co-construction Contribution

Look at the current contribution and see if the current contribution is related to a previous contribution.
In order be labeled as an idea co-construction statement, the current statement should: 1) contain a reasoning statement; and 2), be related to a previous statement. There are **3 main steps** in determining whether a sentence

should be coded as an instance of idea co-construction. Note that sentences marked [I] are idea co-construction statements, and ones marked [N] are NOT idea co-construction statements.

The following examples contain REAS that can be labeled as [I] based on steps one, two, and three, which are listed below.

- 756_X: Is that the end?
  756_Y: Yeah
  756_X: No wonder they had so many. [I, the topic is that the other method is more efficient, previous statement is 2 lines above, which confirmed that the other method is indeed shorter and so quicker]

STEP1. Because idea co-construction statements should contain a reasoning statement, you only need to consider the statements coded as [R] as candidates for [I]. When considering the line marked R, first look at why that line was marked R (we will refer to that segment as REAS from here on). For steps two and three below, you should look at REAS rather than the whole segment marked R.

STEP2. For REAS part of the segment, see if the sentence is on-topic. Namely, the discussion at hand should be related to what we are interested in, i.e., to methods that achieve the goal of the origami task most efficiently, which is making as many boats as possible in the given time. There are several topics that frequently occur for the origami task in this regard. The three categories (2a, 2b, 2c) are on topic, whereas categories (2d, 2e) are off topic.

<Step2. On topic> On topic statements have potential to be [I]. If on-topic, go to STEP 3 to determine whether statement is I/N.

2a. Difference in the two methods. Includes discussion on whether a method is faster, easier, efficient, or more complicated. This comparison should relate to whether or not you should adopt one method over the other. Note that a simple occurrence of the word “faster” can be used on contributions that do NOT relate to determining which method is better (i.e., “Now I have to go faster”).

2b. How they can gain more points. Includes discussion of whether or not there is enough time to make more boats, or if there is enough time to discuss alternative methods

2c. Interpretation of the instruction, method

- 756_Z: Does is matter how far up you go?
  756_Y: Just probably up a little, like enough to have a base [I]

<Step2. Off topic> If off-topic, mark the statement as [N].

2d. Reasoning about instruction (without interpretation)

- 747_X: I think you learned a different way

  747_X: Oh my god

  747_X: Now I have to go faster [N, reasoning about instruction (noting they only have limited amount of time). If it's clear that this was based on X's first contribution "different way," it would be I. However, it is unclear]
- 746_X: At least you got good at it fast

  746_Y: Yeah

  746_X: We had longer to learn it so ... [N, this was part of instruction so no interpretation is involved]

2e. Contributions not related to determining which method to adopt.

- 746_Y: I just wasn't sure if you had some sort of insight into this.

  746_X: No,

  746_X: It just gets faster the more you do it. [N, although X is answering Y's question with elaboration, X is not using his answer to support what method to adopt]

STEP3. If REAS is on-topic, determine whether it is building on **a previous statement** [I] OR a **new topic** [N].

<Step3. building on previous statement> Contributions built on previous statements are considered to be [I].

Example of reasoning contribution that is built on previous statement.

- 736_Z: Don't make any more than one0

  736_Z: Our quality is not so good [I, second sentence is reasoning on it's own (evaluative statement). Plus, it builds on previous sentence]
- 736_X: Do you think yours is easier?

  736_Y: Yeah, it's only a triangular sail. [I, Y is saying "mine is easier because it's only a triangular sail". This reasoning is based on the question on the previous line]

Some of the keywords provide "hints" as to whether the topic is related to a previously stated contribution. E.g. that, this, so

- 754_X: That's different, because they taught us with shutters, so [I, "that" is referring to previous discussion]
- 754_X: "Does that mean I can improvise too?" [I, "that" is referring to a previously stated contribution]
- 751_Z: This is not what we do [I, "this"]
- 751_Y: So, we're making different sailboats [I, "so"]

The first part of contribution could count as a "previous" sentence if the reasoning statement was only based on the second part of a contribution.

- 754_X: I think I could fold two at once or something. Like every time it got to you, you could do quicker with it. [I, the first sentence is the "previous" sentence]
- 751_X: Like three and a half. One was kind of messed up
  751_Z: Yeah we made four but then we had one extra that we didn't get to finish. It was like right at the minute so… [I, the second sentence builds on first]
- 748_Z: No, don't fold it like that. Turn this a different way [I, the second sentence builds on first]

C.F. In the following case, although the second line is R, it is R only when combined with the previous line. Therefore, you cannot take the previous line as a "previous contribution".

- 746_X: I think we're just going to have to help her
  746_X: Otherwise, we're just sitting here [N]

<Step3. New topic> If new-topic, mark the statement as [N].

Bringing up a new idea/topic. Because idea co-construction statements should be related to a previous contribution, brining up a new idea/topic would NOT be considered [I]. However, note that we defined the context as the whole four minutes of discussion. Therefore, be sure to **check the whole context to see whether the idea is related to a previously stated contribution.**

- 758_Y: I don't think you should make another one [N, this is the first time Y talks about making another boat given the one minute warning]
- 751_X: Wow, we really could have done another one I think [N, new topic]
- 751_X: It's weird
  751_X: Because didn't she show the boats in the room to everyone? [N, why they are making a different boat is a new idea]
- 747_Y: Am I faster or slower?
  747_X: I think you learned a different way [N, new topic, because X is not directly addressing Y's concern]

<General rules>

General rules are identical to the ones used for determining reasoning statements. In summary, the three general rules are 1. Be conservative 2. Exhaust the given context 3. Make sure it builds on a previous statement.

Some examples are provided for reference.

1. Be conservative

758_Z: Oh wait, you're folding it … I guess they learned how to do it differently? [N, start of a new topic]

758_X: Oh, I think we are doing a different thing [N, although this topic was introduced earlier, the use of "Oh" makes it ambiguous since it seems like X just realized it despite the earlier discussion]

<Context>

For the context, you should consider the whole four minute session as the context, which is identical to the rule used for coding reasoning.

- 751_Y: Yours is a lot more complicated [I, The context is the whole four minute session, and this observation was based on interaction thus far]

<Typical places when [R] ≠idea co-construction in the Origami task>

1. Answering simple yes/no questions. Answering simple yes/no questions were marked as reasoning when such answers involved critical judgment/ evaluation. However, such answers do not add any new content that's useful in bringing new ideas to the discussion or creating new knowledge. Therefore, such statements are not idea co-construction, and are thus marked [N].

- 758_Y: Like this?

  758_X: Mm-hmm [N, answering simple question is not idea co-construction because you are not adding new info]
- 756_Y: Isn't this how you're supposed to do it? (as Y does makes the second fold)

  756_X: We had a different thing. [N, answering a question without adding new content. This is essentially answering "no"]
- 756_X: I'm used to you guys being so slow

  756_Z: I know, it is slow the other way, [I, in addition to agreeing to previous line, he is also comparing their method to the other method. So this is more than simple a simple case of yes/no]

APPENDIX C

Coding manual for measuring ICC (Eggdrop) from study eight

Our goal is to detect instances of transactivity, which characterizes the knowledge integration process as it takes place in conversation during group work. Transactivity is well studied in the domain of educational psychology and computer-supported collaborative learning. Transactive contributions are arguments that are constructed based on a piece of knowledge. According to Teasly, transactivity occurs when individuals use their knowledge to operate on the reasoning of their partner, or to clarify their own ideas. We believe such transactive contributions are important in creating new knowledge, and thus developed this manual to detect transactive contributions.

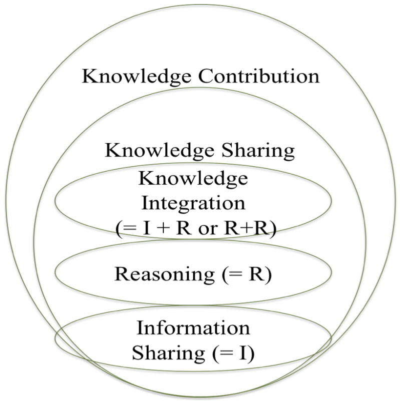

Knowledge Contribution
Knowledge Sharing
Knowledge Integration (= I + R or R+R)
Reasoning (= R)
Information Sharing (= I)

Figure C.1. Types of knowledge sharing

Transactivity does not occur when individuals simply share knowledge. Instead, transactivity characterizes a certain type of knowledge sharing, namely knowledge integration. Knowledge integration is defined as the process of reconciling multiple perspectives. Each perspective that contributes to knowledge integration can be represented as a reasoning statement (R) or an information sharing statement (I) . The relationship between a reasoning statement (R), and information sharing statement (I), and a knowledge integration statement is shown in figure C.1.

Our main goal is to detect transactive statements, which characterizes knowledge integration statements. From Figure C.1, you can see that a knowledge integration statement has at least one reasoning statement (a reasoning statement plus an information sharing statement (I + R) or two reasoning statements (R+ R)). Because the concept of reasoning is somewhat abstract, this manual also includes our definition of reasoning.

<Domain Scenario>

Three participants are asked to build an egg holder. The egg holder will contain an egg, and should protect it from breaking when dropped from a height of two stories . When participants arrive for the experiment, they are asked to spend five minutes doing an individual brainstorming exercise. Depending on the condition they are in, they are given either two analogies (Heuristic condition) or principals (Functional condition) that would help them with the exercise. Participants are asked to list different methods that can be used in the design for each analogy/principal. In addition they are asked to come up with two additional analogies/principals on their own, along with corresponding methods to realize the analogies/principals.

After the individual brainstorming session, the team is given an additional 30 minutes to design and build the egg holder. The coding for reasoning is done on data collected during the main 30-minute discussion session.

Two analogies given in heuristic condition:

- To preserve egg integrity, think of analogies, such as helmets, that are used to protect fragile objects
- To improve the accuracy during drop, think of analogies, such as wings, that are used to control the path of moving objects.

Two principles given in functional condition:

- To preserve the integrity of the egg, you can decrease the force on the egg as it hits the ground.
- To improve the accuracy during drop, you can minimize the drift of the egg holder by manipulating the center of gravity of the carrier.

<Unit of analysis/Segmentation>

In order to analyze the interactions/conversation, the conversational data needs to be first segmented into a smaller unit for analysis. The following rules were used for segmenting the data, each segment will be the unit of analysis.

Note that speaker refers to the "main speaker" of the audio file.

1. Begin a segment when the speaker starts talking. If there is silence at the beginning of the file when the main speaker is silent, this means that there will be an "empty" segment at the beginning.
2. A segment should contain a speaker's continuous speech. If there is an interruption (silence or crosstalk) that lasts for more than one second, a new segment should be created. When you create a new segment, there should be two boundaries – one that marks the end of the speaker's first utterance, and another that marks the start of the next utterance after the pause.

For each unit, the coder should determine whether the segment is transactive or not (p. 3). If a segment is transactive, the coder should decide the type of transactive contribution along the three dimensions that was used by Berkowitz & Gibbs (Please refer to the categories of transactive contribution on p. 5)

line	person	Transcription	cat	notes
5	S1	We can’t use the paper bag because the paper bag is going to be pretty heavy	n	
6	S2	Yeah	n	
7	S3	How about a zip lock bag? If you put the egg inside a zip lock bag then maybe using the rubber band we can hang it through the paper bowl as top of the parachute?	e	Extension, Using materials as a parachute contains reasoning

Table C.1. Sample coded 30 second segments

<General rules>

- If the audio is not transcribed and you cannot clearly hear what the speaker is saying, or if you cannot understand what the speaker is saying, be conservative. Do NOT mark the segment as transactive.

TRANSACTIVE CONTRIBUTION

Look at the current contribution and see if it’s linked to a previous contribution within the context (for the definition of “context” see p. 7)

In order to be transactive: 1) the current statement should contain a reasoning statement and be linked to a previous statement; OR 2) the current statement should compose a reasoning statement when combined with a previous statement. A reasoning statement consists of two concepts and a relationship. (For a more detailed definition of reasoning refer to p. 8).

1. If the current statement contains a reasoning statement, there should be a “link” that explicitly shows that the current statement is based on a previous statement for the current statement to be considered transactive. This link should contain a clear reference to a concept in the previous contribution. A clear reference may be a pronoun (it, that), a repeated noun (safety pin), or an acknowledgement (yes, okay).

e.g.*e.g.m2_s1:241:well well the straw is the same price as the rubber band*
m2_s3:242:and that's not bad -> [transactive reasoning, “that” refers to price, so there is a clear reference to concept in previous contribution (= price). “The price is not bad” is a reasoning statement whose relationship is comparison.]

2. If the current contribution by itself is not a reasoning statement, it may be combined with a previous contribution and yield a reasoning statement.

a. When composing the reasoning statement (using current and previous contributions), only use words that exist in the conversation.

e.g. m2_s3:82: but by being inside the bowl, would that make the bowl just tip over?
m2_s1: 83: no, it should not because we have some counter balance -> [transactive reasoning, this sentence can be expanded just using the words in the conversation. It is equivalent to "no the bowl should not tip over, because we have some counter balance."

b. As a counter example to 1a., do not "invent" words that are not in the contribution to make a reasoning statement

e.g. m2_s1:53: because it (bowl) has such a good surface area... the air won't just like flop and then turn it inside out which is what would
m2_s2:55: okay, then how are we going to cushion the egg? -> [N, Just using the words that is in the conversation, these two sentences cannot be combined to form a reasoning statement. Do not invent words to link line 53 & 55. For example, "So if the bowl doesn't flop and turn inside out, then how are we going to cushion the egg?"]

c. Although there may be many previous contributions that can be used to form a reasoning statement from the current statement, you should use the most recent previous contribution that's feasible.

REASONING CONTRIBUTION

In order to be considered to be a reasoning statement, the statement should combine two concepts with a relation.

1. Concepts: There are five types of concepts. (Refer to p. 8 to see the types).

   a. The two concepts should be explicit (=written). The concepts can be nouns (safety pins) or pronouns (that, it).

      e.g. m2_s1:393: let's order then to start building because we don't have enough time -> [R, first concept is "start building". Second concept "we don't have enough time". Both are explicitly written]

b. If the concept can be clearly implied, but nevertheless omitted, do not count it as a concept.

e.g. m2_s3:402: We could work with something inexpensive like a paper bag if we wanted to add some shred, or something to kind of absorb…"-> [N, the second concept "the impact", which should come at the end of the sentence is not explicitly mentioned. Although, one can infer that the missing word is "impact", we only consider explicit concepts]

2. Relationship: There are 2 types of relationships. (Refer to p. 9 for the types of relationships).

a. Explicit Relationship: The relationship should be explicit (=written). For example, "if … , then …" "… or …" " … because …"

e.g. m2_s1:67: Oh yeah, there too
m2_s2:68: Cause that's all gonna be extra weight [transactive reasoning, the relationship is causal. Used the word "Because"]

b. Special case (AND): A same word can be used as a filler sentence or to specify a relationship between two concepts. The word "and" is one such word. If you can substitute the concept in the first sentence with the new concept in the second sentence, then the word is used as a relationship, not as a filler.

e.g. m2_s3:284: Okay. So are we still going to use rubber bands? because if so then the rubber bands can do double duty of holding the ends
m2_s1: 285: And cushioning [transactive reasoning, equivalent to "because if the rubber bands can do the double duty of cushioning." The additional concept "cushioning" is introduced into the current contribution and makes for a reasoning statement that is a different contribution than the one in previous contribution]

e.g.m2_s1:241:Well well the straw is the same price as the rubber band
m2_s3:242:And that's not bad. so we could buy a whole bunch of them [transactive reasoning]
m2_s1:243:Yeah
m2_s1:244:Yeah
m2_s3:245:And sort of build up these little legs all over the bottom of the bowl, or the top even [N, the current contribution + previous contribution is "we could build up these little legs…" this new composed sentence does not contain reasoning]

CATEGORIES FOR TRANSACTIVE CONTRIBUTION

Once you determine that a statement is reasoning, one should determine which transact category the statement belongs to. Berkowitz & Gibbs has four main dimensions: 1) Primary Focus (Ego/ Alter/ Dyad); 2) Mode (Non-competitive/Competitive); 3) Type (Elicitational/Operational/Representational); and 4) Style (Interrogative/Declarative).

Table C.2 lists the original 18 categories that were defined by Berkowitz & Gibbs (1983). Some of the categories were not found in our corpus (crossed out). Therefore, we transformed the original 18 categories to 10 categories along the three dimensions of primary focus, mode, and type.

	Non-competitive[2]	Competitive[2]
Ego[1]	~~Feedback request (E)~~[3] Clarification (O)	Competitive Clarification (O) = competitive clarification + refinement[5]
Alter[1]	Paraphrase (R/ ~~E~~)[3] ~~Justification Request (E)~~[3] Extension (O) = completion + extension[5]	Competitive Paraphrase (R/ ~~O~~)[5] Reasoning Critique (O) = contradiction + reasoning critique + competitive extension + counter consideration[4]
Group[1]	Dyad paraphrase (R) = juxtaposition+dyad paraphrase[1] Integration (O)	Competitive Juxtaposition (R) Comparative Critique (O)

Table C.2. 10 new transact categories

	Non-competitive[2]	Competitive[2]
Ego[1]	C: Clarification (O) “No, what I am trying to say is the following”	CCL: Competitive Clarification (O) “My position is not X” “Let me refine my position”
Alter[1]	P: Paraphrase (R) “Let me paraphrase your position” E: Extension (O) “I can complete or extend your argument”	CP: Competitive Paraphrase (R) “here’s a paraphrase of your reasoning that emphasizes its weakness” RC: Reasoning Critique (O) “There is a flaw of your reasoning”

Group[1]	DP: Dyad paraphrase (R) “Your position is X mine is Y” “Our shared position is Z” I: Integration (O) “We can combine our position to a common view” “Here’s a general premise common to both views”	CJ: Competitive Juxtaposition (R) “I will make a concession to your position, but also reaffirm my position” CCR: Comparative Critique (O) “My argument is better than yours because … ”

Table C.3. Sample contribution for 10 categories

1. Primary Focus (Ego/Alter/Group)

   Berkowitz & Gibbs categorized transacts according to the owner of the position being discussed. This distinction is important because you want to know whose contribution you’re operating on.

   The original categories were Ego/Alter/Dyad or EgoAlter. We are modifying the last category of “Dyad” to “Group” because we are conducting analysis on group conversation instead of dyads. We are not going to distinguish between Dyad or EgoAlter because, for our purposes of deciding the owner of the position, the important thing is that the contribution has perspectives of more than one member. Whether or not the contribution consists of two different views or two combined views is less meaningful.
2. Mode (Non-competitive vs. Competitive)

   The distinction between competitive vs. non competitive is important because we want to know whether there was a conflict.
3. Type (Operational/ Representational)

   Our goal is to analyze the reasoning that has already been expressed, not the instances when someone is prompting for a reasoning statement. Therefore, we are not coding any elicitational statements and have removed that category. Even in Berkowitz & Gibb’s later version (1983), the removed the category of E.

   We are keeping the distinction between operational vs. representational. The operational needs a clear critique or transformation of a statement, whereas representation merely re-presenting the reasoning, such as in paraphrase. This distinction is important because we would like to know whether the reasoning is a mere paraphrase or if it contains additional concepts.
4. Style

   In the original categories given by Berkowitz and Gibbs, elicitational statements are considered interrogative, while the operational/representational are considered declarative. Because we are not coding for elicitational statements, the style dimension is not needed for our purpose.Removing this dimension allows for the merging of some categories. For example, contradiction, reasoning critique, competitive extension, and counter consideration are all considered “reasoning critique.”

5. Additionally merged categories

   Any additional categories that do not differ along the three main dimensions (primary focus, mode, type) are merged to form a single category. For example competitive clarification and refinements are merged.

   Berkowitz & Gibbs categorized completion as alter, non-competitive, and either representational or operational. Distinguishing between representational or operational for a completion would require a very subjective judgment since it would require mind reading to figure out what the alter intended to say. Therefore, we categorize completion as O, giving the ego credit for the reasoning stated in order to finish a sentence. Thus we can combine completion and extension into one category because they share the same condition along all 3 dimensions.

   Berkowitz & Gibbs categorized competitive paraphrase as alter, competitive, and either representational or operational. We categorize this simply as representational because if the competitive paraphrase is operational, it could be regarded as reasoning critique. For our purpose, whether the statement contains an operational statement *is* the deciding factor, and whether or not the speaker used paraphrase as part of the reasoning is less important.

CONTEXT

For each reasoning statement, we need to locate two concepts and a connection between them as defined on page two of this manual. In some cases, one of the concepts is not explicitly mentioned in the current contribution, but is mentioned in a previous contribution that is still "salient." A concept is "salient" if one can find it within a current context. The following paragraph defines what a context is for a given contribution.

- Start of context: Context for a given contribution begins at the current speaker's previous "contentful" contribution This means the current speakers previous contribution that are merely doing back channeling such as "yeah,", "uh huh" does not count as a contribution (Example one). In addition, when there is only back channeling by other speakers between the current speaker's current and last contribution, you should merge the current and last contribution into one contribution (Example two). Thus, the start of the contribution in this case would be the previous contribution's previous contribution (Example three)
- End of context: The current contribution.

Example one. Context for s3's contribution in line 224.

line#	student	Contribution	Notes
221	m2_s3	To the bottom ... yeah ... umm ... okay cause the straw is really inexpensive. And we can get a whole bunch of little slices	Previous content of contribution
222	m2_s2	Yeah	
223	m2_s2	Yeah i didn't notice the straw before. I was thinking just something to	
224	m2_s3	Yeah. cause we could get a lot.	Current contribution

Example two. Context for s1's contribution in line 4 (Note line 1, 2 & 4 are all by s1, but line 2 is ignored b/c it's back channeling).

line#	student	contribution	Notes
1	m2_s1	Yeah I have a pin over here and I have a pin over here	s1's previous contentful contribution
2	m2_s1	Yeah	Back channeling, ignore
3	m2_s2	Are you sure?	
4	m2_s1	We'll we'll attach the ... We'll use ...	Current contribution

Example three. Context for s1's contribution in line 399. Note that the other speaker's back channeling is also ignored in defining context, in addition to the current speaker's back channeling.

line#	student	contribution	Notes
393	m2_s1	Yeah I have a pin over here and I have a pin over here	s1's previous contentful contribution
394	m2_s3	So. It's this.	
395	m2_s2	I don't see the egg staying in that at all. I don't I don't I don't think well I don't see it staying in it.	
396	m2_s3	Hmm. Well we're gonna tape it.	
397	m2_s1	We'll, we'll attach the ... we'll use ...	Merge with line 399, so also count as current contribution
~~398~~	~~m2_s3~~	~~Yeah~~	Backchanneling doesn't count as contribution
399	m2_s1	We'll attach the egg with the masking tape right? All around.	Current contribution

CONCEPT

ON TASK/ DOMAIN CONCEPT: Domain concepts are concepts used to solve design problems; e.g., designing the egg holder. These concepts can be any of the five types described below.

1. **Theoretical concepts**: Theoretical concepts are principles that students apply in decision making; e.g., decrease force, minimize drift.

- The speaker uses the theoretical concept of "slowing down the egg to decrease the force of impact" (by cushioning).
  s1: "Yeah one concern I do have is that the egg's just free falling. So instead of cushioning, we're just slowing down the speed of the egg in general, but if we could cushion it yeah. Oh yeah we could stuff the bag with napkins. [REAS]

2. **Prior Knowledge** : Prior knowledge is pieces of information that helps decision making based on common sense.

<positive examples>

- The second speaker used prior knowledge about crepe paper and scotch tape and concluded that using those materials would not help.
  m3_s2:59: Do you think we'll able to put the clay around it? Or we could use crepe paper and scotch tape
  m3_s1:62: No that won't help [REAS]

<negative examples>

Prior knowledge can be objective in nature such as prior knowledge about crepe paper not being suitable for the task. However, prior knowledge should not be based on emotion.

- s1: Let's do that, because I like it [N, "I like it" is not a valid concept, out of domain]

3. **Physical System properties**: Elements and characteristics of the elements that are available for the system; e.g., cup, cup is green, bowl is round, straws are flexible, fur is soft.

- The action clearly indicated the cup flipping, which is based on the physical property of the cup.
  s1: So that it doesn't do this *(actions the cup flipping)* [REAS]
- Bubblewrap is an element, so it's a concept by itself. "Impact of the fall" is the second concept, and it's a physical system property. Relationship is causal: using bubblewrap, results in cushioning the impact of the fall.
  s1: bubblewrap cushions the fall [REAS]

4. **Emergent System properties**: Characteristics that appear in the process; e.g., properties of the egg holder.

- The stability of the egg holder is a property that emerged as a result of the combination of materials that the students used to build the egg holder.
  s1: Egg holder is stable.
- The first concept: "whole bunch of straws" are physical properties. But the second concept "build little legs all over" is an emergent property of the straws. The new role of the straws as "little legs" is a new property that emerged from out of the design process. Relationship is causal, or "in order to ..."
  m2_s3:226: So we could buy a whole bunch of them (straws). And sort of build these little legs all over the bottom of the bowl, or the top even [REAS]

5. **Goal**: The "label" is a commonly used term that represents a set of beliefs, a general perspective, or anything that is associated with strong expectations related to points of view. Label incorporates unstated opinions or perceptions. For the eggdrop study it is the four goals (egg integrity, economical use of material, accuracy during drop, and aesthetics)

- The speaker uses the goal of economical use of the material.
  s1: Because we want to save money

OUT OF DOMAIN *:* Any contribution that is not related to the design activity would be considered out of domain. Activities that are clearly not part of doing the design exercise are off task and out of domain. For example, answering a phone call, dinner suggestions, or asking about the payment for the study would be considered out of domain. In addition, statements that do not directly contribute to solving the design problem should also be considered out of domain. For example, coordinating for the next meeting time, telling jokes, expressing a feeling, or asking for clarification on the instructions of the study would be off domain.

- The speaker complements their egg holder
  s1: "It (egg holder) is fancy~" [N, joking]
- The speaker claims that they should get more time due to an experimenter error.
  s1: "Shouldn't we get more time because you guys messed up our design?" [N, instruction]
- The speaker simply asks about the time. This is not directly related to the task of building the egg holder.
  s1: What's the time remaining? [N, instruction]
  - c.f. However, if the speaker uses the timing information in order to make a suggestion on what to do next in terms of building the egg holder, that would be reasoning. (first concept: five minutes left, second concept, buy materials)
    s1: We only have 5 minutes left, so let's start buying materials. [REAS]

RELATION

Two concepts should be connected with a relation in order for the statement to be considered an instance of reasoning There are two types of relations: 1) Compare and contrast; and 2) Cause-and-effect.

1. juxtapositions, compare and contrast, tradeoff: When the speaker compares two concepts, the speaker makes a judgment. Such a judgment involves a reasoning process.

- Keywords, such as “more” indicates that the speaker is making a comparison.
  ChatExamples-1.xls (35): "Well, single cycle is more friendly to the environment …” [REAS]
- The speaker contrasts “low q” with “cost of thermal efficiency.” Note that “lower” and “cost” also imply comparisons.
  Set5b.xls (96): “I got a lower q than you, but it cost my thermal efficiency a lot” [REAS]
- The speaker compares two materials (“that” and “rubber band”) for his solution.
  m2_s2:184: I am thinking that might work better than a lot of rubber bands [REAS]
- The speaker compares the price of straw to the rubberband, and uses that comparison as a basis for making a design decision.
  m2_s1:225: The straw is the same price as the rubber band [REAS, comparison]
  - c.f. However, if the speaker just lists the price of materials, there is only one concept. Nothing is being compared, so it’s not reasoning.
    s1: Straws are 1 mues each [N, instruction].
- The speaker compares two alternatives. The contribution takes the form of A or B.
  m3_s2:59: Do you think we’ll be able to put the clay around it? Or we could use crepe paper and scotch tape. [REAS]
- In the last line, s1’s solution takes the form of A or B, where B is the alternative solution to six pins.
  m2_s3:381: If we do a triangle then we won’t have to use six right? [REAS]
  m2_s1:382: But then it’s not gonna be secure [REAS]
  m2_s1:383: Oh yeah, instead of a triangle, we can do one rubber band like this and one rubber band like this. Or with two safety pins only [REAS. A or B]

2. Causal: Causal relationships link two concepts by using a cause and effect connective. The general relation is doing *x* helps you achieve *y*. There are three main types of causal relationship: a) cause and effect; b) in order to; and c) analogy

a) cause and effect: A occurred because of B. Let’s do A because B.

- Speaker supplies a possible explanation based on speaker's past experience, using “because” clause.
  ChatExamples-3.xls (29): I think we should do a reheat, [because] lots of people seem to be doing that [REAS]
- The two concepts are A:“power is high” and B: “efficiency is low”. The relationship is causal since because of A, B occurred.
  Set6b.xls (70) High power uses low efficiency
- The following three contributions take the form of doing *x* will result in *y*. This is cause and effect.
  s1: Increasing heat input to the cycle, increases efficiency [REAS]
  ChatExamples-1.xls (35): "... So if we do that [then] I'll concede to nuclear as the fuel [REAS]
  ChatExamples-6.xls (90): Haha, Pmin is 10 at s5 now and my efficiency is 48% [REAS]
- First concept: bubblewrap, second concept: the impact of the fall (physical system property), the relationship is causal; i.e., using bubble wrap results in cushioning the impact of the eggs fall.
  s1: Bubblewrap cushions the fall [REAS]

b) In order to: Do *a* in order to achieve *b*. association (use … in), compatability

- First concept: using rubber band, second concept: hanging the bag, relation: causal
  m2_s1:24: Rubber bands used for hanging the bag [REAS]

c) Analogy: When a speaker makes an analogy, he makes a link due to the similarity between two concepts. Some of the keywords that signal analogies are “like,” or “as.”

- s2 Figured out why the others were making a suggestion of using the bowl by using an analogy between “using the bowl” and “parachute.”
  m2_s2:49: Oh, you’re trying to use the bowl as a parachute [REAS]
- s3 Makes a connection by using an analogy “like a fire” because “do something like that.”
  m2_s2:299: Yes we’ll we’ll do something like that yeah
  m2_s3:303: So like a little bungy, okay [REAS, analogy]
- s3 makes a connection by using an analogy “like a fire” because s2’s description.
  m2_s2:421: It’s if you make it at least three pieces and you weave it together
  m2_s3:424: Like a fire [REAS, analogy]

NO RELATION

This section contains sample sentences that contain 2 concepts, but not a legitimate relation.

- Do not get confused by the word “for” in the following sentence. The speaker is saying “this cup is free”, where the first concept is “cup,” and the second concept is “free.” But the relationship between the two concepts is neither compare/contrast nor causal.

"We have this cup for free -> [N, observation]

- c.f. The speaker has a suggestion in addition to the observation. The suggestion is one concept, the observation is another concept. The relationship is causal.

  "We have this cup for free, why don't we use it? [REAS]

NOTES ON LOCATING TWO CONCEPTS

In some cases, it may not be clear whether there are two concepts. Here is a list of things you should consider when trying to locate two concepts in a contribution.

1. Contextual Connection: If the two concepts are not explicitly mentioned in the current contribution, you should look within the context to see if you can find another concept. **Refer to <Context> section below (p. 6) to see how a context for a given contribution is defined before you continue reading this manual.** Contextual connection can be found when speakers uses: a)anaphora; b)ellipsis; or c) completion.

<positive examples>

In conversations, speakers want to make economical use of their words while maintaining textual coherence. Speakers achieve this by using various elements such as: a) anaphora, or b) ellipsis. Because understanding such utterances requires contextual information, we look for concepts that may not be explicit in the current contribution, but that are mentioned explicitly in the context instead.

a) Anaphora: Anaphora reveals broader interpersonal and interactive relations between the components of an utterance (Lyons, 1981). Common types of anaphora are pronouns, which are used to substitute common/proper nouns; e.g., "it," "that". Pronouns are also used for co-reference, which helps to achieve textual coherence by relating two contributions. (Dressler, W.U. & R. de Beaugrande, 1981).

The second speaker uses "that," which refers to s1's contribution. Here, the use of the pronoun "that" is one of the concepts in s2's contribution. This is an anaphoric relation. The other concept is "work," which the speaker knows from prior experience. The relationship between the two concepts is causal. Given the size of the cup, it will work for our purpose.

- s1: So, that's the size of our cup [N, observation]
  s2: That will work [REAS]

b) Ellipsis is the absence of particular elements of an utterance, which can easily be comprehend within the context, mostly from what has been said in previous sentences. The speaker can use ellipsis to abbreviate noun or verb phrases, or even whole clauses.

- In the conversation below, s2 does not explicitly mention one of the concepts, "to build a parachute," because it is already mentioned by s3 previously.
  m2_s3:12: So we've talked about putting some kind of a parachute on. Is that something that we could reasonably build? Or ...
  m2_s2: I don't think there's anything here that's big enough and light enough [REAS]
- The speaker abbreviated "around the egg" at the end of the sentence. So the two concepts are "clay around the egg," and "cup is around the egg," in a relation of comparison.
  s1: My idea was just to have the clay around the egg, like the clay would be around the egg much as the cup is [REAS]

c) Completion. One can also find two concepts by considering context when a speaker completes another speaker's contribution. Here, line 37 is reasoning because s3 completed the sentence to make the second concept explicit.

- m2_s1:36: Hmm, then we can use a safety pin to limit the, the
  m2_s3:37: The amount of space in the bag? [REAS, causal]
  m2_s1: 38: Yes
  m2_s3: 39: Okay, so we can keep it tight? [N, doesn't make sense given the context, which starts in line 37]

<negative examples>

Note that a contribution is not reasoning when the other concept is mentioned outside of context (refer to the definition of context in p. 6). This restriction is applied so that different coders would consider the same amount of information that is salient in a given context. In the contribution by s2 below, one of the concepts was explicitly mentioned earlier in the conversation. However, the earlier mention of the concept is not shown in the conversation below because it was outside of the context.

- m2_s2:468: Well let's ...
  m2_s2:469:Let's see now that the, how much. Okay it looks like we need to be about here [N]

2. Explicitly mentioned: The concepts should be explicitly mentioned and make sense. If one of the concepts is implied in the context, but the speaker didn't finish the sentence and you cannot find the words to complete the sentence within the context, then it is NOT reasoning. The purpose of this rule is to prevent coders from reading too much into lines.

<negative examples>

- s2 did not complete the sentence. Although given the domain whatever is probably referring to the word "impact," we don't see the word explicitly mentioned anywhere in the context, so we cannot count it as a concept.

  m2_s3:Hmm ... like on top here or ...?

  m2_s2:196: Yeah, on top of these things. So they absorb some of the. Whatever [N, didn't complete sentence.]
- One might reason that because they don't have much money to buy tape, they have to be careful with tape. But this was not explicit enough in the context. Don't read too much into the lines.

  m2_s3:333: We only have thirty. Yeah that should give us enough for tape.

  m2_s3:337: Yeah, we'd have to be really careful with our tape. [N]
- First concept: Styrofoam bowl. Here, "help" is not considered to be a concept because it's too ambiguous. In order to be a concept, the speaker should specify "help in doing *x*."

  m2_s1:08: I think a Styrofoam bowl would help -> [N, 1 concept]
- First concept: "absorbs the impact". "Something" is not a concept because the speaker did not explicitly say what something refers to. Given the context, one can infer that "something" refers to "the eggholder," but the rules say that a speaker has to "explicitly mention" the concept.

  m2_s2:81:I'm thinking we might be better off just building something that absorbs ...

  m2_s1:83: The impact? [N, completed a sentence, 1 concept]

3. New contribution: The concept has to be a new contribution. If the same idea is mentioned again within the context, do not count it as new reasoning unless it's phrased differently. Refer to the two positive and two negative examples below.

<positive examples>

- In line 257, "it" refers to "the work" in line 251. This is a new reasoning because the number three has a meaning. In line 251, 4~5 pins were suggested, but in line 257 s1 suggests three, specifically after considering what would be required for their design to be successful.

  m2_s1:251: But I think safety pins like four of five of them could do the work [REAS]

  ..

  m2_s1:257: Three should do it right? [REAS]
- In line 349, s3's contribution is the same meaning as s2's contribution when she says "bands will give." They are also in the same context. However, it is expressed in a different way, so it is a new reasoning.

  m2_s2:348: I am wondering from that height, whether the impact is gonna overcome ... I can see it like giving with that much of a ... [REAS]

  m2_s3:349: That the bands will stretch? [REAS]

<negative examples>

The following examples show contributions that are not phrased differently enough to be counted as a new concept.

- Speaker s3's reasoning was elaborated explicitly and is a new contribution, so his statement counts as reasoning. However, s1's contribution is not new, and it is not phrased differently enough to be considered new reasoning. In this case, s1's contribution is not reasoning because it is not a new contribution within the context.
  m2_s3:58: If we could cushion it. Oh yeah we could stuff the bag with napkins [REAS]
  m2_s1:59: How about a napkin? [N, not new contribution]
- Again, in line 224, s3 mentions the same idea that he already mentioned in line 221. In addition, both contributions are phrased similarly. Therefore, his first contribution in line 221 counts as reasoning, and the contribution in line 224 doesn't count.
  m2_s3: 221: To the bottom ... yeah ... umm ... okay cause the straw is really inexpensive. And we can get a whole bunch of little slices [REAS]
  m2_s2: 222: Yeah
  m2_s2: 223: Yeah i didn't notice the straw before. I was thinking just something to ...
  m2_s3: 224: Yeah. cause we could get a lot [N]

4. Sentence in question format: Some times, a sentence in the format of a question can still contain two concepts when you consider the context.

<positive examples>

- The speaker had to make a judgment on the amount of bubble wrap and the size of the cup in order to ask this question. Therefore, the two concepts that are explicit in this sentence is first concept: so much bubble wrap (amount), and second concept: cup is small. Relation: causal.
  m3_s2:112: Do you think it will take so much bubble wrap in a small cup?
- The speaker used two pronouns "that" to refer to each concept. The relationship is causal; e.g., "in order to be sufficient."
  m3_s1:207: But would that be sufficient for that?

<negative examples>

In some cases, the speaker will simply ask a question such as "why," or "why not," or "why's that." Such questions only contain elicitation. Asking a why question implies that there may be some reasoning occurring in the subject's mind, but it is not explicitly described. Therefore, you will not be able to find two concepts.

- The speaker asks for reasons from the other speaker. There is only one concept.
  s1: “We might still need the straws”
  s2: “why?” -> [N, why]
  - c.f. It is important to note that why questions with further reasoning or explanation can be considered reasoning when the contribution contains two concepts. Here the speaker asks a question to understand/clarify the other person's position. First concept: “need straws.” Second concept: “make egg holder heavier.” The relation is causal.
    s1: "Why do we specifically need the straws? To make the egg holder prettier?" [REAS]

In some cases, the speaker will ask a question such as “how,” or ask any of the w-questions (who, whom, whose, which, what). Again, you should look for two concepts in such sentences. The examples below contain sentences that contain only one concept.

- The speaker is asking for a method to prevent egg from falling quickly. This is analogous to the suggestion “we have to keep it from falling really quickly.” There is only one concept here because the speaker did not say *why* he wants to accomplish his suggestion.
  m3_s3:124: So how do we keep it from falling really quickly
- Similar to the example above, there is only one concept; i.e, hold the plastic bag onto the wrap.
  m3_s2:155: How are you gonna hold the plastic bag on to the wrap?

5. Judgment: A reasoning sentence requires the speaker to make a judgment in order to make a contribution. Some times, there are specific indicators that signals propositional (reveal’s speaker’s intent) content; e.g., I believe.

<positive examples>

- In order to generate this sentence, s1 had to make a judgment using two concepts. First concept: size of "it." Second concept: cup is small. The relation is causal; it is small enough. Therefore, it can fit inside the cup.
  m2_s1:46: I believe it could fit inside in small cups. [REAS]
  - c.f. However, when the sentence only describes a property of an object, it is not reasoning. It is a fact. There is only one concept, which is the property of an object.
    s1: “[After observation] It fits inside small cups” [N]

<negative examples>

A speaker has to make a judgment when asserting agreement/disagreement regarding an idea. However, an agreement by itself is not considered to be a reasoning statement unless it contains two explicit concepts. In line 66, you can see from context that s1 agrees they need safety pins for tethering objects. But he didn’t say why he agrees, so you cannot find two concepts; therefore, it is not reasoning

- m2_s1:63: I think we’ll need safety pins for tethering them over here and over here [REAS]
  m2_s3:64:Yeah

m2_s3:65: Okay, what about down here?

m2_s1:66: Oh yeah, there too [N, 1 concept]

ONLY 1 CONCEPT even with the context (observation, ideas for design without explanation, concept definition)

1. Observation/ Instruction: Speakers sometimes comment on a phenomenon, system property of an element, or rules for the design exercise. If there is only one concept in the contribution, it is not reasoning.

- Oh you're trying to close the cup -> [N, 1 concept, observation]
- The design is also important, the aesthetics -> [N, 1 concept, instruction]
    - Cf. Let's color this because the design is important -> [REAS, 1st concept: let's color this, 2nd concept: design, which is one of the goals]
- Would it be strong enough? -> [N, 1 concept (if there is nothing else in the context)]

2. Suggestion/ideas for design without explanation: When speakers make a suggestion without providing a reason, often there is only one concept; i.e., the suggestion itself.

- The first concept is "threading further." There is no more concept. Therefore, it's not reasoning.
  s1: Let's see if I can thread it bit further -> [N, 1 concept]

3. Definition

- A definition by itself is not reasoning because there is only one concept, the definition itself. Here, the speaker just provides an explanation of what sensor modes are/
  "Sensor modes are …" -> [N, definition]

APPENDIX D

Coding manual for measuring ICC (Ottoman) from study nine

FACTOR

	Types of Reasoning/Explanations	Example	Fac	Explanation	Line in Context
F1	**Keywords - If the keywords (or concepts equivalent to keywords) are mentioned, that is enough to count as a factor. The exception to this rule is the name of a country (e.g. German)**	**arab unrest, nationalism, decentralization of power, inefficiency of political system, economic weakness, population problems, abolition of slavery, inability to match technology, lack of integration between Turks & others**		lack of integration amongst different nationalities is not if8	
		shift of balance of power, economic weakness, capitulations, manipulation of minorities			
		The first being the shift of power and balance within the European community.	R	Listing a factor listed in the fact sheet.	P17&18:6
		A lot of the um what was I gonna say, the reasons for…I think that they had a weak economy	ef2	"weak economy" is the keyword for ef2 (economic weakness)	P31&32:36
		Um, Russia was the third one that was mentioned.	N	name of a country is not enough for a keyword	P33&34:13
		And all the Greeks and the Balkans would want their own,	N	Greeks, not a keyword. And the person didn't finish saying what they want. So it's not enough to be a factor	P25&26:51

		but internal factors are the main problem because, something like population.	N	population. If he said "population problem" it's a keyword. But just population is too vague	P11&12:70
		The Ottoman Empire's economy was mostly based in agriculture,	N	agriculture	P11&12:34
		The uh there was just so many problems internally that uh des, the economy was one of them	N	economy	P31&32:10
		It's a good trade route um in the Mediterranean.	N	Mediterranean is only mentioned in ef5, but it should have been mentioned with Russia wanting it	P33&34:12
		I think a lot of the same issues, uh technology uh not as...	N	should say "inability to match technology"	P33&34:54
		When actually one of the main problems was the different rules	N	what about the different rules? Doesn't match any of the key words, so can't stand alone	P37&38:34
F2	**can map to a specific factor - If a "concept" is discussed, see if you can map the concept to a specific factor. Exception to this rule is when the "concept" is a factor**	and they were becoming so dependent on Europe that it ended up creating debt because European Europe was um le lending money to the Ottoman Empire	ef2	debt is only mentioned in ef2	P39&40: 34
		Russia that was mainly um territory.	N	we said this is N originally, but it seems to be ef5	P33&34:39
		There was no military to uh keep everyone in line	if3	concept mentioned only in if3	P31&32:48
	<factor>	France uh basi, uh Fr France, all together they undermined the, Europe on a whole undermined the economy of uh of the Ottoman Empire.	ef2	"economic weakness" is a factor own it's own, so F2 doesn't apply. Note that although speaker listed a specific country, his point was europe as a whole as signaled by "all together"	P23&24:42
	<counter> : cannot map to a specific internal/ external factor	Not only that, first off, Europe had a huge part in this	N	Not complete reasoning yet, b/c you cannot tell which factor (listed in the fact sheet) this argument corresponds to	P11&12:18

		Um for thousands of years uh Islam's uh er I mean sorry, Muslim's, Jews and Christians have been fighting each other in this area.	N	Can't tell which internal factor this reasoning corresponds to yet. Internal fighting of some sort is mentioned in if1, if2, & if4	P7&8:34
F3	**specific factor - If a concept is mentioned, but they are both factors own their own (i.e. economic weakness vs. British economic influence), go with the more specific factor (i.e. British)**	Economically the British, the French, they were um controlling a great deal of the um the economics in the Ottoman Empire.	ef7, ef8	British, French are more specific than Europe	P27&28: 6

REASONING

	DOMAIN	Two students are debating whether the cause of the Ottoman Empire fall is internal or external.	
	PROCEDURE	1. Label factors: This is like doing topic segmentation. Every time there's a new factor being discussed, label the conversation according to the factors listed in Labeled_IF.docx and Labeled_IF.docx.	
		2. Code for reasoning: determine R/N according to the rules listed below in this manual	
	Typical places for reasoning in this corpus	2A> If the current line is the start of a new internal/ external factor listed in the fact sheet, it should be probably labeled reasoning (unless it's an unfinished reasoning statement OR you can't tell which factor it corresponds to OR it's factors that are applicable for both internal/external -> Look @R3 for details)	
		2B>Relationship types: causal (R1), compare/ contrast (R2),	
		3. Code for transactivity: If the current line is R, you should further determine whether it is T (Transactivity) or E (Externalization). For detailed rules, refer to sheet labeled "transactivity" in this manual	
	GENERAL RULES FOR REASONING CODING	Facts that are causal count as reasoning because they are explained with relation to the smaller facts	
	Be conservative	When there is a way to interpret the segment as both R and N, be conservative and mark the unit as N	
	Distribution of explicit marker in a compound sentence	If an explicit marker can be distributed to multiple parts of a compound sentence, then apply the marker. Refer to <distribution> section in R1/ R2 for examples	

	Blue- Example of Reasoning				
	Red- Example of Non reasoning				
	Types of Reasoning/Explanations	**Example**	**R/N**	**Explanation**	**Line in Context**
R1	**Causal Relationship - Some contributions contain causal reasons to support the speaker's argument. Such statements are often signaled by the use of words such as "because", "in order to", "like", "so"**	and so they forced the Ottoman empire to, um, be economically dependent on them because they set up, like, trading posts and other things within the country, just for their own benefit.	R	"because" indicates causal relationship	P3&4: 20
	<causal verbs> : Verbs that signal causal relationships such as "create", "to become", "change"	Nationalism always creates divisions between the groups within them,	R	Create is a causal verb, explaining a mechanism	P3&4: 8
	"create"	but um the European powers coming in really created a lot of dissention, especially among Christians.	R	Create is a causal verb, explaining a mechanism	P1&2: 16
	"to"	So like the slaves would be raised since they were younger to be in the military and to grow up and lead the military,	R	"to be", "to grow up"	P1&2:28
	"change"	You had Great Britain coming in and changing procedures and introducing capitalization, where they controlled like schools and health and things like that to	R	"Change" is a causal verb	P5&6: 87
	"to"	They fought to try to get, uh, whatever language they speak.	R	"to" is "in order to"	P151&!52 : 23
	<implied causal relationship with an explicit marker> : "even though, "and"	They had the sense of just, even though there might have been different cultures in the country, they ruled under one government that they all believed in	R	"Even though" presupposes that Y implies not X. In this case the speaker presupposed that "having different country -> rule under different government."	P15&16:72
		and like they didn't really feel connected to the Ottoman Empire.	R	"and" is used like "as a result" or "so", which is an explicit marker for implied causal relationship	P3&4: 12
	"with"	It's hard to maintain such a large Empire with a diminishing population.	R	"with" is used as "when"	P151&152 :76

	<answering questions> :When the contribution a direct answer to a question, and it forms a reasoning in combination with the question that was asked, then it's R	Uh, and influx of disease brought in by (B interrupts)	R	This is in response to the question in line 73 "what caused the high mortality rate". Line 74 is clearly a response to this question, essentially saying "influx of diseased caused the high mortality rate". This is R	P11&12:74
		They had…The didn't, they didn't put it to good use, I would say.	R	In response to question in line 86&87, this sentence is saying "their technology was not built up because they didn't put resources to good use"	P11&12:88
	<counter - implied causal without a marker> : According to the "be conservative" general rule, we are going to mark any implied causal relationship without an explicit marker as N.	They hired farmers	N	this is can be interpreted as being a result of the previous line (30) "they had to hire farmers". Thus this is implied causal relationship. However, there is no explicit marker, so "N".	P149&150:31
		They couldn't, they didn't have any way to defend against the Europeans.	N	This could be interpreted as either causal from previous line (33) or start of a new fact linking to the next line (35).Either way, there is no explicit marker. So "N"	P149&150:34
	<counter - not used as a causal verb> : Verbs that typically signal causal relationship may not be used as a causal verb depending on the context	and they kind of had a bunch of separate states that it was divided into that had like sub rulers who all created separate laws and separate economic, um, like, systems.	N	typically "create" is a causal verb. However, here the rulers "created" laws by the nature of who they are.. So this is not causal	P141&142:67
	"to"	The French tried to defend or persuade the English, the Catholics, the Protestants, the, um.	N	In the corpus this is "R" because it's a start of a new factor. However, this example was added to illustrate a case when "to" is NOT used as a causal verb. "try to" is not a causal verb	P151&152:57

	<counter - filler> : Words that typically signal causal relationship may be used as a filler depending on the context	So all these countries were, like interested in the Middle East and india for trading.	N	"So" is used as a filler in this conext	P3&4: 72
R2	**Comparing and Contrasting - Comparing the facts counts as reasoning when it highlights the differences or similarities that support the speaker's argument. Signaled by words such as "but", "however"**	but they were dependent on other countries.	R	"but" indicates contrasting relationship	P3&4: 56
		Yeah, but they also, uh, abolished slave trade	R	"but" indicates objection to the previous contirbution. Note that the speaker also supports his objection by citing a reason; "abolished slave trade"	P1&2: 26
		and that was like the main, the main, like, like the president of the Ottoman Empire you could say.	R	comparing sultan to the president, explicit marker is "and"	P1&2:7
	same	Now the same would apply to all the different nationalities	R	"same" is an explicit marker used for comparison	P15&16:42
	as well, better	I mean, Pennsylvania, where we live now, uh, it had a lot of struggles as well	R	"As well" is an explicit comparison	P5&6:61
	largest, bigger, biggest	For the Ottoman Empire um, the second largest group in Ottawa was the Arabs.	R	"second largest" is a comparison. Similar to "the biggest" in P11&12: 8	P13&14:14
	<contrast within same sentence> : sometimes two contrasting concepts are presented in a same contribution without the use of an explicit marker. This is R.	At that time if you look at the technology, it wasn't that advanced that we have today,	R	comparison, "at that time" vs. "today" is intended as a comparison	P15&16:56
		and they convinced the Ottoman Empire to have the people from the different countries follow the laws of their countries, not the Ottoman's.	R	contrast between "ottoman empire & their country"	P17&18:26
	<challenge> : challenges the other speaker in the format of a question	and why wouldn't the collapse have happened sooner than that?	R	speaker is saying "but it didn't collapse". This contribution uses question format to indicate objection	P1&2:12

		Yeah but what do you think caused the high mortality rate?	R	speaker is challenging the other speaker.	P11&12:73
	<asking for clarification> : Asking for clarification of what the other person said is N. Note that unlike challenge questions, clarification questions typically cannot be rephrased in the format of an argumentive statement	Are you saying that they chose not to or that they? (A interrupts)	R	This is asking for clarification question. If you rephrase as "You are saying they chose not to or they chose to.", you can see that the speaker is not making an argument for one way or the other so it's not a challenge	P11&12:92
	<analogy> : analogies and use of idioms to support an argument counts as R	Outside powers were only the icing on the cake;	R	"icing on the cake" is an idiom, so R	P149&150:63
	<distribution rule> : If an explicit marker can be distributed to multiple parts of a compound sentence, then apply the marker.	and it was behind times, and different things.	R	this is a compound sentence along with the previous line "but it was …". This is a compound sentence with an explicit marker at the start of this line ("and"). Using "distribution rule", you can apply "but" from the previous sentence. Thus the sentence becomes, "but it was behind times..". So this is a contrast.	P151&152: 7
	<counter - implied compare/ contrast without explicit marker>	They could've avoided them.	N	contrast, implicit "but" at the beginning of the sentence can be assumed. However since there is no explicit marker, it's N	P17&18:74
		It's relevant today,	N	doesn't say "how" it's relevant, and there's no explicit marker	P5&6:9
		it was relevant before the collapse of the Ottoman empire,	N	doesn't say "how" it's relevant, and there's no explicit marker. "before" signals timing, not a comparison	P5&6:10

	<counter - not used as a compare/ contrast> : Words that typically signal compare/contrast relationship may not be used as a comapre/contrast depending on the context	I had ice-cream and coffee as well	N	"as well" is addition, not compare/ contrast	
R3	**Relevance to Ottoman Fall - Given the objective of the task, any statement that established the relationship to the fall of the Empire should be marked as reasoning**	Well I think I mean to a certain extent obviously the the internal factor that the Ottoman Empire was so diverse is a problem	R	"so diverse" is a reasoning given for the fall ("problem"), it's not clear which factor it corresponds to, but it links to the fall explicitly.	P37&38:43
	<factor listed in fact sheet> : If the speaker lists an internal/ external factor listed in the fact sheet, this is because the speaker is making an argument for internal/external cause	Also, there was a rise in nationalism in the country.	R	Listing another factor that is an internal factor, although the speaker didn't tie it back explicitly to the fall	P13&14:21
		The first being the shift of power and balance within the European community.	R	Listing a factor listed in the fact sheet.	P17&18:6
	<incorrect facts> :If incorrect fact is used, but you can clearly identify which fact it is, then it's R. This is consistent with a more general rule which states, "even if it's a false statement, if it's in the form of a reasoning, then it's reasoning"	Because Germany, they fought against the Ottoman Empire in nineteen fourteen,	R	It's Russia, not Germany. But in this case, it's clear that the speaker made a mistake and the facts map to ef5. So it's R	P149&150:59
	<counter> : not complete reasoning yet	First of all, uh the Ottoman Empire st, stayed, survived over centuries,	N	Starting to make an argument, but not a complete "Reason" yet.	P15&16:9
	<counter> : cannot map to a specific internal/ external factor	Not only that, first off, Europe had a huge part in this	N	Not complete reasoning yet, b/c you cannot tell which factor (listed in the fact sheet) this argument corresponds to	P11&12:18
		The Ottoman Empire's economy was mostly based in agriculture,	N	Can't tell which internal factor this reasoning corresponds to yet.	P11&12:34

		but internal factors are the main problem because, something like population.	N	didn't say "what" about population. If the speaker said "population problem" it would be R	P11&12:70
		Um for thousands of years uh Islam's uh er I mean sorry, Muslim's, Jews and Christians have been fighting each other in this area.	N	Can't tell which internal factor this reasoning corresponds to yet.	P7&8:34
	<counter> : factor that is in fact sheet, but not listed as an internal/ external factor	Um when Europe after World War one wanted to divide the land between the victors, they did it sort of arbit, arbitrarily and hastily without really consulting the people living there,	N	It's factor i3 listed in the fact sheet, but this fact was given for both internal & external. Therefore, just listing this factor without an explicit reasoning for internal/ external is not R.	P7&8:41
		But anyway, they empire should've was too weak with internally to actually deal with the issues you you the,	N	if0, we don't count as a factor, only if1~8. if0 is too general, and overlaps for both internal & external	P27&28:35
	<counter> : for the factors that are not listed in the fact sheet, the speaker should make an explicit link back to the fall. Otherwise, coders have to guess whether it's a fact listed to back up the internal/ external factor so it becomes ambiguous	They didn't really have the sense of 'let's work for our country.'	N	there is no explicit link back to anything	P141&142:34
		They ever since, pretty much throughout ever since the Turks conqured it, um the centuries before, after the fall of Constantinople they looked internally for most, pretty much with everything they didn't look outside.	N	there is no explicit link back to anything	P37&38:53
	Types of NON-Reasoning/Explanations	**Example**	**R/N**	**Explanation**	**Line in Context**
N1	**Instruction - Although the sentence appears to be reasoning, if it's due to instruction of the task, then it does not count as R**	Okay, so in the case of the fall of the Ottoman Empire, its obvious to me that it was internal problems that caused the fall.	N	Instruction of this task was to argue for internal/external causes	P11&12:5

		Well, I agree with what you're saying how it was, they uh, fell from within,	N	Instruction of the other person was to argue for internal i.e. "fell from within"	P147&148:84
		you can only really think of them as um being brought about by certain factors outside of the Empire itself.	N	Although the speaker didn't use the word "external", the speaker is basically saying the cause is "external" & nothing more.	P153&154:6
		but, however I feel that external powers are a little bit more, external forces are a little bit more detrimental to the fall of the Ottoman Empire.	N	He's saying nothing more than what's in the instruction. Even if he said "European powers", external powers = european powers for our corpora context, so it's N	P11&12:17
		It's because they don't have their own ways of uh goodness gracious external factors may play a big role	N	instruction	P11&12:69
		Well, the fall of the Ottoman Empire can really be contributed to the outside, the outside factors of the European powers for the simple fact that the Ottoman Empire is one Empire <THROAT>	N	listing "european powers" is synonymous to saying "external factors"/ "external powers". So that's not enough. The specific factor that he started to elaborate is incomplete	P13&14:4
		And pretty much that's why the internal arguments pretty much destroyed it.	N	instruction	P13&14:118
	<counter - R3> : lists additional factors other than internal/ external	So one of the things that uh, I think would be a huge contributor to what caused the fall of the Ottoman empire, in terms of internal, uh, faults would be the rise of nationalism.	R	listed "rise of nationalism" in addition to "internal"	P3&4:7
	<counter - used as argument> : When the speaker uses other person's contribution to argue for his own {internal/ external} aspect, then it's R. Typically it's in the form of "xxx is actually internal/ external"	but that's only based off (P B interrupts) of internal factors.	R	saying more than "it's because of internal factors", he is saying external factors are "based on" internal factors.	P11&12:56

		Internal uh internal problems that were brought about because of external factors.	R	saying that the other speakers internal evidence were actually because of external factors..	P23&24:40
N2	**Definition - Definition/ Explanation of terms are N**	need example from corpus…			
		it's (imperialism) how many different countries you can take over and extending your empire as far as possible.	N	Defining "imperialism". Although it's an inaccurate definition, this is what the speaker thought imperialism is	P5&6:20
	<asking for elaboration/ definition> Asking for simple elaboration or definition in a format of a question is N	Um, what were your other points?	N	Asking for repeat of information	P143&144:36
N3	**Repeat/rephrase - Reasoning should only be labeled as such when it is the first occurence. Same concepts should not be marked as reasoning even if the formati is reasoning**	And um, but the Turks ruled all.	N	exactly same information was said in the previous line (17)	P151&152:18
		so that they could get the Asia …	N	If no additional information is being conveyed, even if the format is reasoning, it should be marked N.	P143&144:49
		and this caused a lot of fighting	N	it's a repeat, although it contains reasoning within the contribution	P7&8:58
		They essentially, when they abolished it, they made their economy less strong.	N	this is not adding any more information than the previous line	P21&22:71
	<bringing up same subject> after a while without additional information is repeat	need example			
	<we got rid of negating the opposite>				

N4	**Incomplete/ incoherent sentence**	Population problems because, due to a war going on-	N	Incomplete sentence	P3&4:30
		Well, the fall of the Ottoman Empire can really be contributed to the outside, the outside factors of the European powers for the simple fact that the Ottoman Empire is one Empire	N	After reading the following line, you know that at this point, the person is not done with his reasoning. So this is incomplete	P13&14:4
		The Ottoman's might have won if such and such a reason,	N	This sentence is incomplete. Specific "reason" is not listed	P19&20:59
		So, with that regard, they really had no incentive,	N	no incentive for what? Incomplete	P1&2:56
	<counter> : incomplete, but can identity reasoning within the incomplete sentence	so - but also, um, the Ottoman Empire became pretty economically weak because Europe just came in and ..	R	This sentence is not complete, but you can see causality.. E.g. between week economy & europe	P141&142:64
N5	**Affect - positive or negative feelings by the speaker, e.g. happy/sad, confident/anxious, interested/boared**				
N6	**Abstration/ Generalization - When the speaker reinterprets a situation by using abstration/ generalization with given facts, it is R.**	So really, they had the states,	N	abstraction of previously stated contributions	P5&6:108
N7	**Application - Going from a general principle to a specific detail**	and all of these groups were getting economic and all sorts of aid there	N	European nations were backing up minority group by giving "economic aid".	P9&10:18
N8	**Interpretation- sentence using judgement or appreciation is N**				
	<Judgement> :Attidue towards {human/ country's} behaviour, which the speaker may admire/ criticize/ praise/ condemn. Such behaviours may get formalized by rules & regulations by church & state. E.g. How special (lucky, fashionable), capable (powerful, insightful), dependable (careful, flexible), honest (truthful, descrete), ethical (good, polite)	It just seemed that everyone was really only interested in the self-interest, not so much as the whole empire.	N	speaker made a judgment about "everyone" It's a conclusion drawn based on facts	P1&2:58

	<Appreciation> :Attitude towards value of things, which the speaker may evaluate/ criticize. Such values may get formalized by awards/ assessments. E.g. Impact (fascinating, boring), Likeable (good, ugly), Balance (harmonious, consistent), Complexity (simple, unclear), Value (innovative, dated)	and like, it's hard for a country to come back after Russia invaded them and had the three year war, starting in nineteen fourteen	N	speaker said it's "hard" based on facts (Russia invasion, war)	P3&4: 33
		and it was impossible to keep peace,	N	"impossible" is a conclusion drawn based on facts	P1&2: 48
		There was many rulings trying to fuse as one that just couldn't last.	N	"couldn't last" is a conclusion drawn based on facts	P1&2:10
N9	**Agreement/ Objection - Simple agreement or objection without specifying what the speaker is agreeing/ objecting to is N**	But you can't say that.	N	"but" indicates objection, yet this contribution does not contain support for his objection.	P1&2: 41
		We're not speaking just before.	N	this is disagreeing to the previous contribution, nothing more	P11&12:79

ICC

	GENERAL RULES	In order to determine whether a sentence is "transactive", first determine when the speaker lists a "new factor" (reason). Possible factors are listed in the handouts (8 external & 9 internal). When the speaker starts discussing a new factor, we should mark that segment as "E"(externalization), if not "T"(transactive)			
		Note that FACTOR is different from TOPIC. Factor is one of the 8 internal/ external factors listed in the handouts. Topic can be same across factors. E.g. talking about European invasion (topic), listing 2 separate reasons Germany and France (2 factors)			
	Blue- Example of Transactive statements				
	Red- Example of Externalization				

	Types of Transactivity/Explanations	Example	T/E	Explanation	Line in Context
T1	**Continuation of previously stated factor : Previous stated factor has to be marked as R, because transactivity is building on previously stated reasoning.**	um and they didn't want to annex the whole empire because they knew that there would be fighting	T	t, still about power balance	P17&18: 9
		Well my point was that the only reason why the Europeans didn't take over the Ottoman Empire was because of war amongst themselves, which eventually caused frankly the collapse of the Ottoman Empire.	T	t, "my point was" indicates that this is a continuing point from previous contribution. The previous contribution was actually a while back, but it's still T	P19&20:62
	<speaker change> : even if the speaker changes, if it continues from the previous factor, then it's T.	Yeah, and once those three countries split apart, they just, they started to clash with each other because they split apart.	T	continuation of same factor ef0, without the addition of another one by speaker B	P147&148:94
	<explicit reference> : even if it's start of a new factor, if it clearly builds on the previous contributions, it should be T e.g. "this", or repetition of words from previous contribution	and this hurt their economy because, like, their agriculture, they couldn't farm as much because everyone was out fighting,	T	"and this" signals an explicit reference	P143&144: 23
		and these, uh, countries, were all so embroiled in rapid industrialization which caused them to uh, have a great, uh, economic power that they were able to sort of press on to the Ottoman Empire, creating a very weak and dependant state.	T	"and these countries" is an explicit reference	P153&154:9
		I, like I said before, I think the reason for this newfound nationalism that they had for their old country, was because of the a European nations taking uh action and saying we support so and so minority group.	T	"like I said before" is an explicit reference	P9&10:93

	mentioning same topic	If we want to talk about violence and economic problems, then we can't not talk about, um, the different forms of Empire at this time.	T	"violence and economic problems" was mentioned in line 30, although not in the exact same words. However, it is clear that they are equivalent to the topic mentioned in line 30 (economic weakness & fighting)	P153&154:34
		The uh, the rise of nationalism uh of of different cultural and religious groups predates the uh uh the arbitrary break up um by the European powers after World war one,	T	"arbitrary breakup" is the same topic they are on. Note that the topic occurs towards the end of the sentence, whereas usually the topic is in the beginning	P7&8:50
	<can't interpret explicit reference> : when there is an explicit reference, such as "this" or "that", but you cannot interpret that explicit marker, then it's E	and that was even um worsened by the fact that whenever the uh slave trade was abolished, all these slaves were in the military	E	can't tell what the speaker is referring to when he says "that"	P9&10:36
	<explicit "objection"> : even if it's start of a new factor, if it is used as an objection to the previous contributions, it should be T e.g. "but", "even though"	Um but they were able to take over the Ottoman Empire because they had no economy to speak of	T	New topic of if4, but as indicated by "but" this is a contrast on previous sentence	P7&8:69
		And also within the, even with the war, the Ottoman Empire had it's own economic weaknesses, um because they were fighting they really couldn't grow the crops or any kind of export.	T	"even with the war" shows that this contribution is a contrast on previous discussion about war	P1&2:31
		Well, uh, the internal fighting was actually caused by the uh, the external groups.	T	"actually" indicates objection (used as a contrast)	P147&148:18
		Well, uh, even worser than listening to their own sultan, they followed rules, the laws of countries that were not part of the Empire, of external countries.	T	"Even worser" is used as a contrast	P147&148:73

		Well it's the fact that Great Britain had a commercial dominance in in the Ottoman Empire.	T	"well" is used as an explicit marker for objection in this context	P13&14:63
	even if	Well even if they were in no shape um it's definitely true that their eco economically they were weak.	T	"even if" is used as a contrast	P21&22:38
	<explicit addition> : unlke a contrast, additions are not transactive (since that would make all contributions trasactive)	Well, a lot of it was also due to the Arab unrest, which wasn't mentioned at all, like.	E	"also" is a word that indicates an explicit link to a previous contribution. However, this is an "addition". So this is E	P147&148:24
		For instance, France ran most of its banks and lent it a lot of money.	E	"for instance" is a word that indicates an explicit link to a previous contribution. However, this is an "addition". So this is E.	P151&152:11
		but also that just goes to show the decentralization of power that was occurring at the time because they were you know so unstable domestically that uh it caused them to also to not only be less you know more inferior abroad	E	"but" is not used as an objection. Here, "but also that" is used together to signal an explicit addition.	P9&10:56
		Right well playing off of that it just also goes to show that that the inefficiency of the Ottoman political system contributed to that because that could have been completely controlled	E	"playing off of that" is an explicit addition in this context	P9&10:36
T2	**Repetition of the last turn :** If the speaker picks up exactly from where he left on this previous turn (as signaled by almost verbatim words), then it's T	Um. So, they produced different economic weaknesses by setting up different things.	T	Line 13 (the last contribution) ends with exactly the same words	P151&152:31

	Types of Externalization/Explanations	Example	T/E	Explanation	Line in Context
E1	**First instance of a new factor : When a new factor that hasn't been discussed previously is introduced, mark as E.**	and because of all the political and um fighting unrest in the ottoman empire, a lot of people went and like fought in wars	E	e. b/c it's the first time it becomes reasoning, 16 is not independent from 17	P17&18: 17
		Yeah, but they had no…There was Arab unrest	E	typically "but" signals that this sentence is built on a previous contribution. However, in this case it was a false start. The contribution starts a new topic, instead of building on a previously stated fact.	P1&2:17
		and when they abolished the military slaves, free men had to go work as soldiers and be in the military which led to diminished crop harvest because people weren't home to work on their crops, and to do their jobs like they did before.	E	start of a new topic, no causality with previously mentioned topics	P141&142:69
		and because of these ethnic diversities they each wanted a a nationalism.	E	This is the first instance of factor if2.	P21&22:12
	<no explicit link> If there is no explicit link, we assume that the speaker is switching his line of reasoning to another factor.	and then Russia just wanted to get like a sea route to Asia,	E	Doesn't have an explicit link	P143&144:5
		and there was so much internal fighting that the military had to spend, like, thousands of millions of dollars on just dealing with the internal fighting which just killed the economy itself.	E	Although they are talking about "internal fighting" (same topic), the speaker is listing a new factor. In addition, there is no explicit link.. So this is E	P147&148:14

	<new factor not in fact sheet> : Sometimes the speaker introduces a new topic that is not in the fact sheet. They are E	and thus, made it easy for external factors to, um, bring it down.	E	Speaker talks about a new factor in line 13. Line 13&14 were labeled as N. So the first line that's marked as R with this new topic gets E	P153&154:15
E3					
E2	**Re-introducing a factor: new info : If a factor that's been mentioned previously has been reintroduced after another factor, look at that segment of conversation. If there is new information being added, then it's E.**	Like, Russia actually declared war on the Empire because they wanted the Baltic states and the straights that they had	E	Although ef5 was mentioned in line 82, ef6 was the main topic in between. So re introducing ef5 is considered E	P151&152:86
	<re-introducing a factor: repeat info>: If a factor that's been mentioned previously has been reintroduced after another factor, look at that part of the conversation. If there is NO new information being added, it's T	>>> this case should never happen, because this should have been marked as "N", not "R"			

FACT SHEET

	General
	Current strife in the Middle East can be better understood by learning about history of Ottoman Empire, which was ruled by Ottoman Turks.
	Ottoman Empire spanned three continents, lasted for six centuries, until end of WWI.
	European powers greatly influenced the politics of Middle East; e.g., fighting between Kurds & Iraquis, Kurds and Turks, Israilis and Palestinians, Iran and Iraq.
	In 1922, the Ottoman Empire was divided between Great Britain, France and Turkey.
	Great Britain & France were designated "protectors" of most of Middle East.
	Boarders were arbitrary and hastily decided without regard to the wishes of people living in the areas.
	Internal factors
i1	Arab unrest: Arabs second largest after Ottoman Turks. Arabs were respected because prophet Mohammed was an Arab. Arabs were privileged, but had to speak Turkish. Arabs wanted Turkish to be official language.
i2	Rise of nationalism: Each ethnic group began to be openly proud of their heritage and wanted to create their own nation. Different ethnic groups inhabited and fought over same land.
i3	Decentralization of power and inefficiency of the Ottoman political system: Empire didn't have military power to enforce laws. Sultan was a figurehead. Many communities were ruled by their own local ruler and by local laws
i4	Economic weakness: Economy was drained by constant fighting within the Empire. Fighting led to decrease in agriculture which in turn weakened the economy. People could not tend their crops, and the economy suffered as a result.
i5	Population problems: Mortality rate was high due to poor health-care, disease, starvation and civil war. Great loss of able-bodied people.
i6	Abolition of slave trade: Slavery was pervasive in Ottoman society. Slaves were treated as part of the family in Islam laws, but some suppressed slavery. After abolishment of slavery, military slaves were replaced by freeman, who were corrupt and had poor military skills
i7	Inability to match Europe and the West technologically: This made it difficult to communicate and travel effectively. The empire has a disadvantage in WWI.
i8	Lack of integration between Turks and non-Turks: Ruling class made no attempt to integrate conquered people. Certain regions had no economic links to Istanbul.
	External factors
e1	Shift of power within the European community: Ottoman territory was desirable b/c Empire served as gateway between Europe and Asia. At first, European countries preferred to leave Ottoman government in place, but render it submissive. Europeans didn't want to threaten peace by fragmenting the Empire. However, with WWI, it became apparent Ottoman Empire would be split up among the victors, so European powers began to take offensive approach to hasten its decline.

e2	Economic weakness: Europe had very high involvement in the economic affairs of Empire. Private businessmen from Europe also weakened the economy by taking jobs and money away from Ottoman subjects.
e3	Capitulations
e4	Europe's manipulation of minorities within the Empire: Each European power claimed to support the rights of specific ethnic groups that were non-Islamic. The Ottoman Empire tried to maintain its power on traditional Islamic religious grounds and thus alienated its non-Moslem subjects. The European Powers were accepted by the non-Moslems, who felt that they had enough support to denounce the Empire.
e5	Russia: Declared war on the Empire to gain control over Straits of Dardenelles.
e6	Germany: Increased the Empire's military and economic dependence. Germans expected Ottoman Empire to be an ally in WWI.
e7	France: Weakened the Empire both politically and economically. France wanted Syria as a territory. They also had high positions in Imperial Ottoman bank, as well as railway, utility, and coal.
e8	Great Britain: Wanted to controls parts of the Empire that were closest to India. Promoted commercial and economic interests.

APPENDIX E

Challenging theories of feedback from intelligent tutoring[6]: study 3

The study presented here is motivated by the need for design principles that would encourage the knowledge co-construction process in collaborative problem-solving environments. In studies presented in the body of this dissertation, I have conducted research in the context of project-oriented courses where the interaction is face-to-face. In the following studies three and four (Appendix E and F), investigation occurs in the context of tutoring environments, where students interact through computer based chat interfaces. In both types of context, common types of problems must be addressed to support collaborative learning. More specifically, with respect to knowledge sharing, students should be encouraged to share knowledge and provide help to each other. In addition, supporting the instances of reasoning articulation, or engaging in explanation behavior, are important behaviors that should be encouraged in collaborative learning environments. The state-of-the-art in intelligent tutoring technology has been optimized for success in an individual learning scenario. Therefore, many interaction design issues may need to be revisited to achieve success in a collaborative learning setting with regards to knowledge co-construction behaviors.

[6] This work was published in Gweon, G., Rosé, C., Albright, E., Cui, Y. (2007). Evaluating the Effect of Feedback from a CSCL Problem Solving Environment on Learning, Interaction, and Perceived Interdependence. In Proc. Computer Supportive Collaborative Learning.

One issue related to knowledge co-construction is the design of feedback from intelligent tutoring systems. At the time of this research, state-of-the-art intelligent tutoring systems provide immediate feedback, involving indications of correct versus incorrect problem solving actions, as well as solicited versus unsolicited hints during problem-solving sessions. However, it is not clear whether such feedback from the intelligent tutoring environment will be helpful or harmful in a collaborative learning setting. This chapter investigates the hypothesis that the typical state-of-the-art intelligent tutoring system feedback (immediate feedback) interferes with collaborative learning because it can be treated as a replacement for interaction between students. In contrast, a delayed form of feedback may encourage processes such as the knowledge co-construction process, which are supposed to be encouraged in collaborative learning. If this hypothesis is correct, it would predict that the benefit of the collaboration will be reduced when immediate feedback is present in a collaborative learning environment.

E.1. Motivation and Design

For decades a wide range of social and cognitive benefits have been extensively documented in connection with collaborative learning. Based on Piaget's foundational work (1985), one can argue that a major cognitive benefit of collaborative learning is that when students bring differing perspectives to a problem solving situation, then the interaction causes the participants to consider questions that might not have occurred to them otherwise. This stimulus, which includes knowledge sharing and co-construction, could cause them to identify gaps in their understanding, which they would then be in a better position to address. This type of cognitive conflict has the potential to lead to productive shifts in student understanding. Related to this notion, other cognitive benefits of collaborative learning focus on the benefits of engaging in teaching behaviors, especially deep explanation (Webb, Nemer, & Zuniga, 2002). Vanlehn (1999) has argued that deep explanation, even in an individual learning setting, has the potential to lead to cognitive conflict and subsequently to learning. Other work in the computer-supported collaborative learning community demonstrates that interventions that enhance argumentative knowledge construction, wherein students are encouraged to make their differences in opinion known in explicit collaborative discussion, enhances the acquisition of multi-perspective knowledge (Fischer et al. 2002). Furthermore, based on Vygotsky's seminal work (1978), we know that when students who have different strengths and weaknesses work together, they can provide support for each other that allows them to solve problems that would be beyond their reach if they were working alone. This makes it possible for them to participate in a wider range of hands-on learning experiences. It is in connection with this Vygotskian model of collaborative learning that I see a conflict with the design of feedback, sometimes called scaffolding, which is the hallmark of the state-of-the-art in intelligent tutoring technology, and is based on the same principles, and thus designed to meet the same needs.

Therefore, I hypothesize that, compared to the immediate feedback, which is the typical state-of-the-art feedback from intelligent tutoring systems, the delayed feedback would be more beneficial in collaborative learning environments. The delayed feedback is based on a long line of investigations into the use of "worked-out" examples for instruction. My hypothesis predicts that the presence of typical intelligent tutoring style feedback in a collaborative problem solving environment will reduce the amount of student interaction. Furthermore, a reduction in collaborative interaction may then lead to a reduction in the exchange of alternative perspectives on problem solving, and thus also interfere with the benefits of collaboration from the Piagetian perspective.

While these cognitive benefits of collaborative learning are valuable, they are not the only positive effects of collaborative learning. In fact, the social benefits of collaborative learning may be even more valuable for fostering a productive classroom environment. By encouraging a sense of positive interdependence between students, where students see themselves both as offering help and as receiving help from others, collaborative learning has been used as a form of social engineering to address conflict in multi-ethnic, inner-city classrooms (Sharan 1980). Some examples of documented social benefits of successful collaborative-learning interactions include increases in acceptance, the liking of others from different backgrounds, including an identification with, and commitment to, participation in a learning community, as well as improvements in motivation and an aptitude toward long term learning (Sharan 1980). These social benefits of collaborative learning are closely connected with the Vygotskian foundations of collaborative learning because the positive interdependence that is fostered in it is related to the exchange of support, or scaffolding, that I hypothesize will be replaced with scaffolding offered by the environment, wherein typical intelligent tutoring technology is used.

In the remainder of the chapter I further explore foundational issues related to the design of an environment that supports collaborative problem solving with intelligent tutoring technology. I present an empirical investigation wherein I experimentally contrast collaboration in two feedback configurations, one that is identical to state-of-the-art intelligent tutoring technology, which I refer to as the Immediate Feedback condition, and one that is based on a longline of investigation of "worked out" examples and their use for instruction, which I refer to as the Delayed Feedback condition. I investigate the impact of this experimental manipulation on perceptions of collaboration with a questionnaire, with evidence of learning from tests and quizzes, and through a qualitative analysis of the collaborative problem solving processes from coded chat logs, collected during the collaborative problem solving sessions. Ulitmatley, the data does not support the strong form of my initial hypothesis. Rather, I find a gender by condition interaction, where male students prefer and benefit more from the Immediate Feedback condition and female students prefer and benefit more from the Delayed Feedback condition. I conclude with a discussion of design implications and plans for future experimentation.

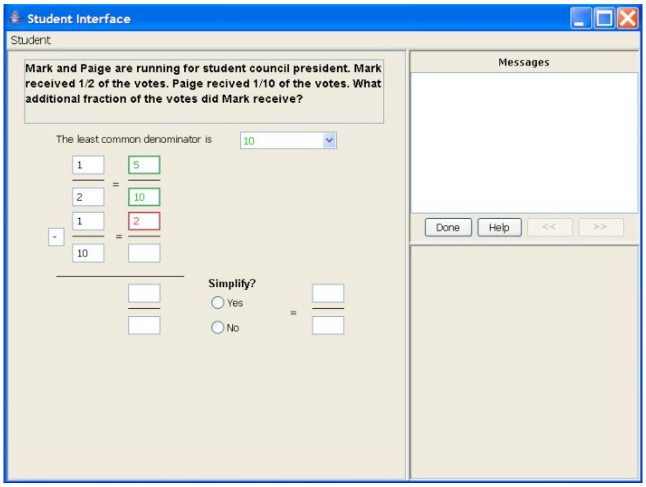
E.2. Procedure

In this section I discuss the experimental infrastructure used to conduct an investigation, both in terms of the technology that was used and in how the lab, where the students worked, was setup. My current infrastructure was built with the Cognitive Tutor Authoring Tools, which support quick authoring of Cognitive Tutor style problem solving systems (Koedinger et al., 1997). Other development work related to supporting collaborative learning in connection with Cognitive Tutors is found in (Walker, 2005). As mentioned, the purpose of this study is to explore issues related to the design of problem solving feedback offered by an environment during collaborative problem solving. The infrastructure used in this study is a simple extension of the typical structured problem solving interfaces that are characteristic of Cognitive Tutors, or other tutors that are part of the knowledge tracing tutor tradition. This infrastructure is designed to support experimentation with alternative feedback designs, keeping all other aspects of the student's experience constant across conditions.

In this study, I am contrasting two designs for feedback from the learning environment, namely "immediate feedback" and "delayed feedback." These alternative feedback paradigms have been experimentally contrasted in individual learning settings in the past (e.g., Bjork, 1994,;Nathan, 1998). Typically, immediate feedback consists of what is called "flag feedback," which signals to students after each problem solving action whether it was correct or not, and offers hints on demand, hints that are typically arranged in hint sequences, beginning with less directive hints, and ending with more directive hints. In a delayed feedback setting, flag feedback is typically withheld so that students must use their own self-monitoring skills to detect their errors. Furthermore, hints may be withheld altogether or changed in nature so as not to be as focused narrowly on the correct solution path, so that students have a greater responsibility for keeping themselves on track. In this study, both flag feedback and hints were withheld from students in the Delayed Feedback condition. Instead, when students decided that their solution as complete, they submitted the solution and were only then presented with a fully worked up version of their problem, with some explanation about how the solution was constructed. In order to control for information access between conditions, the instructional content in the explanation was constructed by concatenating the content encoded in the hints that students had access to in the Immediate Feedback condition.

Based on prior work, we know there are trade-offs between immediate and delayed feedback for individual learning, especially regarding efficiency and retention (Bjork, 1994; Nathan, 1998). Studies have shown that immediate feedback is more efficient because students are never allowed to stray too far from the correct solution path. Therefore, a shorter amount of time is required to solve each problem; in practice, students thus solve more problems (Corbett & Anderson, 2002). Yet, other studies show that students get a deeper understanding of material in a Delayed Feedback setting since they have time to reflect on their errors, and also because they have the opportunity to develop self-monitoring. This has been shown in cognitive tasks such as learning genetics (Lee, 1992), as well as in motor tasks such as learning arm movement motions (Schmidt & Bjork, 1992). Most state of the art intelligent tutoring systems such as Cognitive

Tutors have adopted an Immediate Feedback approach because, in practice, a greater degree of efficiency leads to higher learning gains in an individual learning scenario, and this is because of the relatively large numbers of problems students are able to work through. However, I conjecture that the optimal problem solving feedback design in a collaborative learning setting may be different.



Mark and Paige are running for student council president. Mark received 1/2 of the votes. Paige recived 1/10 of the votes. What additional fraction of the votes did Mark receive?

The least common denominator is 10

Figure E.1. Problem solving interface for Immediate Condition.

As the students worked in through a lab session, their computer's display was composed of two side-by-side panels. In the panel on the lefthand side of the screen, displayed in Figure 5.1, was the problem solving interface. In the Immediate Feedback interface, the Cognitive Tutor program provided feedback as students solved a problem. In the delay feedback interface, students receive feedback from the cognitive tutor only after they submit their answers to a whole problem. In case of Immediate Feedback, corrected answers were given after students typed in an answer for the boxes in the interface. If the answers are correct, the box turns green; otherwise, it turns red. Students can click on the "help" button to receive successive hints. When they are finished, they click on the button "Done" to move on to the next problem. For Delayed Feedback, students agreed on when their answer was complete, then clicked on the "Submit" button. The program then gave corrected feedback if an incorrect solution was submitted. In both conditions, the program presents the students with a new problem only after they submit a correct solution.

Using RealVNC's Virtual Network Computing (http://www.realvnc.com/), the panel shown in Figure E.1 was shared between the screens of the respective computers of collaborating pairs, so that they were both free to contribute to the evolution of their joint solution. In the panel on the right hand side of the screen was an MSN messenger window where students could chat about their problem. The arrangement of the lab where this study was conducted was such that each student was sitting at his own computer, in such a way that collaborating pairs could not easily talk face-to-face since, in all cases, there was a row of desks with computers in between that student's row and the row where the partner was seated. The students did not know who their partner was, or where they were seated. The purpose of this arrangement was to encourage communication through the chat interface so that it could easily be recorded and processed online during collaboration.

In order to test my hypothesis, I conducted a two condition within-subject experiment where I manipulated the characteristics of the feedback and hints offered by the environment. In both conditions, students work in pairs to solve math problems such as the addition of fractions, multiplication, subtraction, and division problems. In the control condition, students received immediate feedback from the cognitive tutor (Immediate Feedback condition), whereas in the experimental condition students get delayed feedback (Delayed Feedback condition).

The experimental procedure extended over four school days, with the experimental manipulation taking place during days two (i.e., Lab Day One) and three (i.e., Lab Day Two), which I refer to as the first and second lab day since the students worked together in pairs in a computer lab at their school. The fourth day of the experiment was separated from the third day of the experiment by a weekend. I used two different units of material, where each was experienced by each pair in only one condition or another; this way I could distinguish learning that resulted from work in one condition from learning that resulted from work in the other condition. The two units were the addition of fractions and subtraction (AddSub) and fraction multiplication and division (MultDiv). I also counter-balanced the order of the units and conditions to control for ordering effects (Table E.1).

On the first day of the four day study, students took a pretest, which lasted for 30 minutes, to assess how much they already knew about the subject matter. I also provided a short collaboration training manual, where the teacher gave an example of good, collaborative conversation. In addition, pairs of students were formed by the instructor. Teams remained stable throughout the experiment. The students were instructed that the teams would compete for a small prize at the end of the study based on how much they learned, and how many problems they were able to solve together correctly. The second and third days were lab days where the students worked with their partner on one of the units, in one of the conditions. On each lab day they worked through a different unit, in a different condition, from what they were working on the previous day. Each lab session lasted 35 minutes. At the end of each lab period, the students took a short quiz that lasted about 10 minutes. At the end of the first lab day only, students also filled out a short questionnaire to assess their perceptions of the help they received, were offered, or thought they benefited from during the collaboration. On the fourth experiment day, which was two days after the last lab day, they took a post test, which was isomorphic to the pre test, and was used for assessing retention of the material.

Table E.1. The experimental setup

	Pairs	Lab day 1	Lab day 2
Class 1	1~4	AddSub, Imm	MultDiv, Delay
	5~8	MultDiv, Delay	AddSub, Imm
Class 2	9~11	AddSub, Delay	MultDiv, Imm
	12~15	MultDiv, Imm	AddSub, Delay

Thirty sixth grade students from a suburban elementary school participated in the study. The students were from two different classes taught by the same teacher, with 16 students in the first class and 14 students in the second class. Students were arranged into arbitrary pairs by their instructor. Students were not told who their partner was. There was a mixture of mixed-ability and homogenous ability pairs. Furthermore, out of 15 pairs who participated in the study, 12 of them were mixed gender pairs, 2 of them were all female pairs, and one of them was an all male pair. Because only a small number of pairs were homogeneous gender pairs, one cannot draw any conclusions from this data about the relative merits of mixed gender versus homogeneous gender pairs. Furthermore, we cannot distinguish between gender effects that are specific to mixed gender pairs, versus gender effects that are independent of group composition.

The materials for the experiment consisted of the following:

- A mathematics tutoring program. The two mathematics chapters were fraction addition and subtraction and fraction division & multiplication.
- Two extensive isomorphic tests (Test A and Test B) were designed for use as the pretest and the posttest. These tests each consisted of 16 near transfer and eight far transfer problems, balanced between the two units of material. Likewise, I had Quiz A and Quiz B, which were designed to be isomorphic to a subset of the pre and posttests. Thus, quizzes are shorter versions of the tests, administered after each lab day. Thus, I was able to use pre to posttest gains as a measure of retention (since there was a two day lag between the last lab day and the posttest day).
- A Questionnaire. As a subjective assessment of socially-oriented variables, I used a questionnaire with eight questions related to the perceived problem solving competence of themselves and their partner, their perceived benefit, their perceptions of the help they received, and what they thought the help provided. Each question consisted of a statement such as "The other student depended on me for information or help to solve problems," and an 11 point scale that ranged from -5, labeled "strongly disagree," to +5, labeled "strongly agree."

E.3. Findings

I began the analysis by investigating the socially-oriented variables measured by means of the questionnaire, specifically the perceived problem solving competence of self and partner, perceived benefit, perceived help received, and perceived help provided. Neither of the experimental conditions maximized all of these outcome variables for both genders. Instead, I observed a consistent gender by condition interaction across perceived benefit, perceived help received, and perceived help offered, although it is only significant in some cases and marginal in others. Specifically, in the Delay condition boys rated themselves as offering more help and receiving less help, as well as benefiting less from help, whereas the pattern was the opposite for girls, although the effect was not as strong.

Consistent with prior work investigating the well known gender gap in math achievement for middle school children, I found a main effect of gender, whereby boys rated themselves on the questionnaire as being more competent problem solvers $F(1,29) = 5.01$, $p < .05$, effect size .7 s.d., although there was no significant difference in grade so far in the class reported by their teacher $F(1,29) = 0.46$, p = n.s. There was, however, a significant difference in pretest scores, whereby boys scored higher than girls $F(1, 29) = 6.13$, $p < .05$, effect size 1.2 s.d., thus demonstrating that boys came into the experiment with more prior knowledge about the specific material covered. In terms of perceived benefit from the collaboration, boys rated themselves as benefiting significantly less than girls did $F(1,29) = 2.15$, $p < .05$. As mentioned, there was a significant interaction with condition such that the difference is only significant in the Delay condition $F(1,29) = 4.63$, $p < .05$, effect size 2.5 s.d. This effect did not seem to be related to the relatively higher pretest scores of boys since there was no significant correlation between perceived benefit and either the pretest score of the student or that of their partner. Related to perceived help provided I also found a significant gender by condition interaction $F(1,29) = 4.84$, $p < .05$. Specifically, girls' ratings of the extent that they offered help was significantly lower than that of boys, but only in the Delay condition. There was a corresponding marginal gender by condition interaction $F(1,29) = 2.62$, $p = .1$ whereby girls' ratings of the extent that they received help was higher in the Delay condition, whereas the opposite was the case for boys.

The learning gains analysis is consistent with the interaction between gender and condition observed on the questionnaire, and it offers some weak evidence in favor of the Delay condition on learning overall. There was no measurable gain on far transfer items either within conditions, or over the whole population, and thus I suspect that the far transfer items may have been too difficult for these students. Therefore, I only consider learning on near transfer items for the remainder of the analyses to distinguish between conditions.

The first focus is on immediate learning. For the measure of immediate learning, I measured learning gains that occurred locally within lab session. Recall that the pre and posttest were more extensive than the two quizzes, but also contained a section that was isomorphic to the quizzes so as to enable a consistent measure of growth in understanding of the material over the four days of the experiment. The posttest for each lab session was a quiz administered on the day of the session. For the first lab session, the pretest was the score on the subset of the pretest from day one of the study, which was isomorphic to the quizzes. The pretest for the second lab day was the quiz score from the first lab day. I only considered data from the 12 out of 15 pairs whereboth students were present for the pretest and both lab days.

For this analysis I used an ANCOVA model with the posttest score as the dependent variable, condition, pair nested within condition, unit of material, time point, and gender as independent variables, and used the pretest score as the covariate. The purpose of this ANCOVA design was to control for all factors that may have accounted for performance differences on the test, such as what units of material the students had been exposed to, when the test was administered, and gender (since I observed gender effects in the data). There was a marginal effect of pair on learning gains F(11,32) = 1.94, p = .07, but no effect of unit of material (i.e., AddSub versus MultDiv) or time point (i.e., lab session 1 versus lab session 2). We see a marginal crossover interaction between gender and condition on near transfer items, such that there was a trend for girls to learn more on average than boys in the Delay condition, and for boys to learn more on average than girls in the Immediate condition F(1,32) = 3.43, p = .07. While it was true that boys came into the experiment with higher pretest scores, I did not find a significant or even marginal aptitude-treatment interaction that might provide an alternative explanation for the gender by condition interaction on learning.

Because the strongest evidence presented thus far is for the Delay condition to be experienced negatively by boys, and only marginally significant evidence in favor of the benefit of the Delay condition for girls, one might argue that the data suggests that the most reasonable implication of these results would be to choose the Immediate feedback condition for all students. However, on the retention test, there was only a significant pre to posttest gain in the Delay condition. For this analysis, it is necessary to separate the test questions into subsets related to each unit because each student learned each unit of material in a different condition, in order to measure learning per condition. If a student learned the AddSub unit in the Delayed Feedback condition, then that student's pre and posttest score for the Delayed Feedback condition would be the score on the part of the pre and post tests that were related to AddSub, and the corresponding portions of the tests related to MultDiv would be that student's pre and post test scores for the Immediate Feedback condition. If a student was absent, I dropped the analysis data from segments of material that student was absent for. One student did not take the post test, and three students were absent on the second lab day, one in the Immediate Feedback condition and two in the Delayed Feedback condition. Thus I have 56 pairs of scores: 29 for the Immediate condition, and 27 for the Delay condition.

I computed the significance of the pre to posttest difference using 2-tailed paired t-tests. Note that this analysis controls for pair effects and gender effects since all comparisons are for scores pertaining to an individual student. As mentioned, the difference was significant in the case of the Delay condition $t(26) = 1.58$, $p < .05$, but not in the case of the Immediate condition $t(28) = 2.27$, $p = .12$. This is consistent with the findings from other studies insofar as delayed feedback fosters a deeper understanding of the material and would thus be beneficial for retention of the material.

The student chat logs contain rich data on how the student's collaborative problem-solving process transpired. I conducted a qualitative analysis of the conversational data recorded from MSN messenger to better illuminate the findings from the tests and questionnaire data. Based on the analysis of the questionnaire data, I expected to find that boys offered more help in the Delayed Feedback condition, but received more help in the Immediate Feedback condition, and I expected the opposite would be the case for girls. However, I found some surprising relationships between chat behavior and questionnaire data, on the one hand, and between the more straightforward relationships in patterns of chat data and how much the students learned. Specifically, I find that the condition where students offer more help is also that condition where they perceived to benefit and learn more.

In order to make the sometimes cryptic statements of students clearer during analysis, and also to provide an objective reference point for segmenting the dialogue into meaningful units, I merged the log file data recorded by the tutoring software, r using time stamps for alignment. I then segmented the data into episodes using the log files from the tutoring software as an objective guide. Each episode was meant to include conversation pertaining to a single problem-solving step as reified by the structured problem-solving interface. All entries in the log files recorded by the tutoring software refer to the step the action and is associated with any hints or other feedback provided by the tutoring software.

I approached the design of the coding scheme with some questions in mind. For example, I wanted to investigate how many times each student requested help in each condition. Furthermore, I wondered how their partners responded to their help requests. A preliminary cursory analysis of the MSN messenger logs revealed that frequently students requested help but did not receive any response from their partner. I also observed signs of frustration between students, and some cases where students explicitly refused to help one another. Because my focal questions all pertain to issues related to help seeking and help provision, I designed a coarse grained coding scheme to identify the regions of the integrated log files where this help seeking and help providing behavior is found. In the future, I may code additional types of behaviors or make finer grained distinctions. My current coding scheme has five mutually exclusive categories, namely (R) Requests received, (P) Help Provision, (N) No Response, (C) Can't Help, and (D) Deny Help. Along with the "other" category, which indicates that a contribution does not contain either help seeking or help providing behavior, these codes can be considered exhaustive. A sample of coded dialogue is found in Table 1, where the second and third columns contain the assigned codes. Each column is associated with a single conversational participant.

The first type of conversational action I coded was Help Requests (R). Help Requests are conversational contributions; e.g., asking for help on problem solving, asking an explicit question about the domain content, or expressing confusion or frustration. Not all questions were coded as Help Requests. For example, there were frequent episodes where students discussed coordination issues, such as whether the other student wanted to go next, or if it was their turn, and these questions were not coded as help requests for the purpose of addressing the research questions. Adjacent to each coded help request, in the column associated with the partner student, I coded four types of responses. Help provisions (P) are actions that attempt to provide support or substantive information related to the other student's request, regardless of the quality of this information. These actions are attempts to move toward resolving the problem. Can't Help statements (C) are responses where the other student indicates that he or she cannot provide help because he or she doesn't know what to do either. Deny Help (D) statements are where the other student responds in such a way that it is clear that he or she knows the answer but refuses to stop to help the other student. For example, "Ask [the teacher], I understand it" or "Hold on [and the other student proceeds to solve the problem and never comes back to answer the original question]" are type D statements. And finally, no response (N) are statements where the other student ignores help requests completely.

Each log file was coded separately by two coders, who then met and resolved all conflicts. Using consensus coding, I then tabulated the number of occurrences of each code, in each condition, associated with each gender. An example of one such interaction is displayed in Table E.2. Here students take turns working out parts of a math problem (line 92, 105, 103). When help is requested, the other student provides an answer with some explanation (line 97). Such successful interactions where students benefit from the help of their partner, and also see themselves as contributing to the success of their partner, promote feelings of positive interdependence between students (Sharan, 1980). In Table E.3 I display the average counts of actions within a single problem solving session. I tabulated the codes from the perspective of each student. Therefore, for each student I obtained a count for help requests made during the associated session as well as help requests received. I also noted how many problems were solved by that student working with his or her partner during the associated lab session, as well as how many conversational segments there were in the integrated log file.

Table E.2. Example Coded Conversation.

Line	S23	S24	speaker: content
92			s24: ur turn
93			s23: k
94	R	P95	s23: is it 1/20?
95			s24: no it is 4/20
96	R	P97	s23: y?
97			s24: cause to get 5 to 20 you need to multiply it by 4 and what you do to the bottom you must do to the top
98			s23: oooooo
99			s23: IM SO SRY
100			s23: :$
101			s24: thats ok
102	R	P103	s23: i feel like a dope
103			s24: :D
105			s23: your turn
106			s24: k
107	P108	R	s24: you have to subtract right
108			s23: yea
110			s24: k
113	P114	R	s24: do you want to do the simplify it
114			s23: Sure
137	C138	R	s24: whats wrong with it
138			s23: idk [I don’t know]

**For simplicity, portions of the integrated log file related to the interaction with the problem solving interface have been removed*

Table E.3. Average numbers (and standard deviation) of coded categories per session.

	Males Immediate (7)	Females Immediate (9)	Males Delay (8)	Females Delay (6)
Problems Solved	10.0 (7.19)	6.0 (5.6)	5.0 (2.4)	4.7 (2.8)
Segments	21.7 (11.6)	17.1 (11.2)	15.1 (4.4)	16.8 (3.6)
(R) Requests Received	5.6 (3.3)	2.4 (1.2)	4.6 (3.5)	6.5 (4.5)
(P) Help Provision	3.3 (1.9)	0.6 (0.7)	2.0 (1.9)	3.3 (3.1)
(N) No Response	1.7 (1.4)	1.3 (1.1)	2.2 (1.6)	2.2 (3.9)
(C) Can't Help	0.3 (0.5)	0.6 (0.7)	0.1 (0.4)	0.7 (0.8)
(D) Deny Help	0.3 (0.8)	0 (0)	0.3 (0.7)	0.3 (0.8)
R Given	2.6 (0.8)	4.8 (3.4)	5.4 (4.4)	5.5 (3.5)
P Received	1.0 (1.0)	2.3 (2.3)	2.5 (3.7)	2.7 (1.8)
N Received	1.1 (.9)	1.8 (1.4)	2.1 (3.4)	2.3 (1.5)
C Received	0.4 (.8)	0.4 (0.5)	0.5 (0.7)	0.2 (0.4)
D Received	0 (0)	0.2 (0.7)	0.25 (0.7)	0.3 (0.8)

**Note that statistical comparisons in this chapter are presented both in terms of raw numbers and proportions.*

As a manipulation check, after I tabulated the number of occurrences of each code in each integrated log file, I first checked to see whether there was a significant effect of condition on patterns of occurrence in the codes. For this analysis, each count pertained to a single lab session, but I used data from both lab sessions. There was a marginal main effect of condition on number of problems solved F(1,44) = 3.49, p = .07, and a significant main effect of condition on number of segments $F(1, 44) = 9.45$, $p < .005$, with no interaction with gender. The larger average number of problems solved, and larger average number of segments was found, in the Immediate Feedback condition. This is to be expected based on prior findings that immediate feedback increases problem solving efficiency. While there was a significantly larger number of conversational segments in the integrated logs from the Immediate Feedback condition, the proportion of segments that contained a help request were not stable across conditions. Thus, there was no significant main effect of condition on raw numbers for either help requests received or offered. There was, however, a significant gender by condition interaction on raw number of requests received $F(1,42) = 4.79$, $p < .05$, and a marginal gender by condition interaction on both help requests given and help requests received when the raw counts are normalized by number of segments: $F(1,42) = 3.62$, $p = .06$ and $F(1,42) = 3.10$, $p = .09$ respectively. In all cases there was no significant or marginal gender effect except in the Immediate feedback condition, where males received more requests than females, as well as participated in a higher proportion of discourse segments where they received a request than females did. In contrast, females participated in a higher proportion of segments where they made more requests than males did.

Taking into consideration that the majority of collaborating pairs were of mixed gender pairs, this analysis suggests that in the Immediate Feedback condition we find an asymmetric collaboration pattern where males appear as help providers and females appear as the help receivers. To further investigate this finding, I compared counts of response types across conditions, normalized by the number of requests. Data from transcripts where no requests were received were dropped from this analysis. There was a significant main effect of condition on number of Can’t Help responses such that a larger proportion of requests were met with a Can’t Help response in the Immediate Feedback condition than in the Delayed Feedback condition, with no interaction with gender $F(1,42) = 4.86$, $p < .05$, effect size 1.5 standard deviations. This suggests that the nature of help requests may have been different in the two conditions. The coarse-grained coding of the collaborative behavior does not allow one to further address the question of what caused this difference at this time.

For the other three response types, we see a significant gender by condition interaction but no main effect of condition: Help Provision F(1,40) = 4.84, p < .05; Deny Help F(1,40) = 3.96, p < .05; No Response F(1,40) = 4.91, p < .05. For girls, the proportion of Help Provision and Deny Help responses is lower in the Immediate Feedback condition than in the Delayed Feedback condition, but higher for No Response responses. The pattern is almost the opposite for boys, where proportion of Deny Help responses remains stable between conditions, but the proportion of No Response responses is lower in the Immediate Feedback condition than in the Delayed Feedback condition, and the proportion of Help Provision responses is higher in the Immediate Feedback condition than in the Delayed Feedback condition. Thus, the asymmetric collaboration pattern reverses directions between conditions when we examine responses to help requests. Whereas girls offered more help in the Delayed Feedback condition, boys offered more help in the Immediate Feedback condition.

I examined the relationships between patterns of occurrence of those codes in the collaborative process and the quantitative social and cognitive outcome measures that came from the questionnaire data, as well as from the tests and quizzes. These findings are described in the following two sections. My purpose has been to inform the design of collaborative learning environment with features that will enhance positive interdependence between students, as well as facilitating learning. However, based on the questionnaire data, neither of my conditions consistently maximized all three of the socially-oriented dependent variables; namely, perceived benefit, perceived help received, and perceived help offered. The surprising finding is that it appears that girls perceive themselves as benefiting more and receiving more help in the condition where they were actually offering more help; conversely, boys see themselves as receiving more help and benefit in the condition where they are offering more help. Specifically, what I found was a male preference for the Immediate Feedback condition and a female preference for the Delayed Feedback condition, such that girls generally perceived themselves as receiving more help and more benefit in the Delayed Feedback condition, whereas the pattern was the opposite for boys. In terms of perceived help offered, there was no difference between how girls and boys rated themselves in the Immediate Feedback condition, but girls rated themselves as offering significantly less help in the Delayed Feedback condition than boys did. As mentioned, what I observed based on the corpus is that girls responded to a higher proportion of help requests with a substantive answer in the Delayed Feedback condition, whereas boys responded to a higher proportion of help requests with a substantive answer in the Immediate Feedback condition.

One possible explanation for perceiving more help where one is in fact offering more help is that the act of offering help is an instructionally beneficial activity. When students engage in this activity, they perceive themselves as receiving help and they benefit because they are learning. Recall that with the quantitative analysis, we observed that girls learned more in the Delayed Feedback condition where they are offering more help. In contrast, boys learned more in the Immediate Feedback condition where they are offering more help. As further evidence of this connection, we see a significant correlation between total number of Help Provision responses and learning when we compute a multiple regression with pretest score and with the number of Help Provision responses as independent variables, and with the posttest score as the dependent variable (R^2=.84, p = .001, N=30), as well as a significant gender by condition interaction on total number of Help Provision occurrences that mirrors the earlier analysis, with respect to proportion of Help Provision responses F(1,26) = 7.79, p=.01. A Bonferroni posthoc analysis reveals a marginal difference between number of Help provision statements made by girls in the Delayed Feedback condition, and that of the Immediate Feedback condition (effect size .89 standard deviations), and a marginal difference between number of Help provision statements made by boys in the Immediate Feedback condition, and by girls in the Immediate Feedback condition (effect size .86 s.d.).

E.4. Discussion

I have investigated the hypothesis that the presence of a typical intelligent tutoring system would interfere with collaboration and dampen its positive effects in a collaborative problem-solving interface. While my data does not support the strong version of my hypothesis, we are left with the challenge of reconciling the dichotomous needs and preferences of girls and boys. What is interesting here is that the evidence from this study challenges the strongly supported view that immediate feedback during problem solving is always beneficial. One can see that the social context of students working together affects the way that feedback functions. Thus, the wholesale application of findings from studies of individual learning with feedback technology is not warranted for collobrative group work.

Further experimentation is required to identify a satisfactory solution. One obvious follow-up study is to replicate the design from this study, except to use only homogeneous gender pairs rather than mixed-gender pairs. This would allow us to separate gender preferences that are specific to mixed-gender pairs from those that are more generally gender based. Further analysis of the data from this investigation might yield additional insights that would allow us to identify other possible ways of reconciling the different needs and preferences of girls and boys. For example, while we have evidence that the experimental manipulation lead to increases in productive behavior for learning in one condition for boys and the other for girls, we do not know why they responded more positively to different conditions. There may be deeper differences in the interaction styles characteristic of each feedback condition that are obscured by the coarse-grained analysis of the data. Future work on a deeper analysis of the conversational data would yield new insights.

I started with a question related to the extent that immediate feedback, from a problem solving environment, might interfere with collaboration between students. The results indeed challenge a wholesale adoption of design principles developed through empirical studies of individual learning with technology to collaborative contexts. The picture turns out to be quite complex, and the study raises more questions than it answers. We now move on to the series of studies presented in chapter six , which shows that despite the findings from study three, I was still able to have a positive effect on collaborative problem solving through the dynamic use of feedback, which targets collaborative processes rather than helps on problem solving per se.

APPENDIX F

Motivating a paradigm for dynamic support for collaboration[7]: Study 4

In the previous chapter, I investigated the effect of immediate versus delayed feedback conditions in a collaborative learning context. The choice of these conditions was based on the need for developing supportive environments that encourage the knowledge co-construction process. In this chapter, I study other collaborative learning configurations that may encourage the knowledge co-construction processes, namely group compositions, or more specifically, the partnering student's characteristics (engagement level and competence) and the existence of prompts from the tutoring system. Before examining these two variables, the chapter also discusses a preliminary study that was conducted to examine whether the tutoring system is more effective when students collaborate during sessions than when they solve problems on their own. These three studies were conducted using an experimental procedure and infrastructure that was innovative insofar as it allowed for the control of one side of the collaboration by using a confederate student.

[7] This work was published in Gweon, G., Rosé, C., Carey, R., Zaiss, Z. (2006). Providing Support for Adaptive Scripting in an On-Line Collaborative Learning Environment. In Proc. CHI. Honorable Mention Award (Top 5%)

F.1. Motivation & design

One innovative aspect of the work presented in this section is its experimental paradigm, which provides a highly controlled way to examine mechanisms where one learner's behavior influences their partner's behavior and learning. This was accomplished by pairing real students with confederate peer learners who were staff members on the research team behaving in a highly prescribed manner. By holding the behavior of one member of a dyad constant within conditions but varied systematically across conditions, we can measure the causal effect of the variables that we manipulate. Furthermore, this approach allows us to observe the interaction between both typical and unusual combinations of the variables we manipulate. While this approach lacks the high degree of external validity found in more naturalistic observations of collaborative learning interactions, it provides complementary insights not possible within that framework. Confederate peer learners were used in the two latter studies reported in this chapter. In the first study, where I contrast solitary problem solving (SOL) and naturalistic peer problem solving (P2P), no confederate peer learners were necessary. An identical experimental procedure and infrastructure were used across all three studies.

F.2. Procedure

The experimental procedure consisted of five phases, composed of three test phases alternating with two instructional phases. The experimental manipulation took place during phase four, which was an instructional phase.

I strictly controlled for time in all phases. During the pre-instructional testing phase (phase one), students filled out a consent form, took a pretest to assess their domain specific knowledge (for 15 minutes), and read the instructions for the first instructional phase. During the first instructional phase (phase two), which was a human tutoring phase lasting 45 minutes, students received tutoring on the general concept of differentiation as well as seven specific rules of differentiation from a human tutor. Although requiring students to learn independently in the first instructional phase would have been closer to what students face in real on-line environments, I needed to provide students with some common ground quickly for the purpose of the short term lab study. The tutor was blind to the student's condition and adhered to a rigid schedule for covering all of the content in a consistent way across students. During the mid-instructional testing phase (phase three), students took a short middle test to assess their learning during phase two (for 10 minutes). They also read the instructions for the second instructional phase. The second instructional phase (phase four), was a problem solving phase where students worked through as many of 12 multi-step derivation problems as possible during the allotted 35 minutes. Finally, in the post-instructional phase (phase five), students took the post-test (for 15 minutes) and filled out a questionnaire.

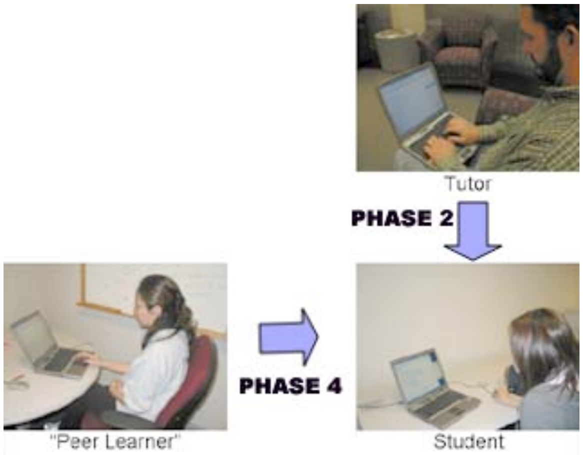


Figure F.1. The student participant and tutor work together during Phase 2. The student participant and confederate peer learner work together during Phase 4.

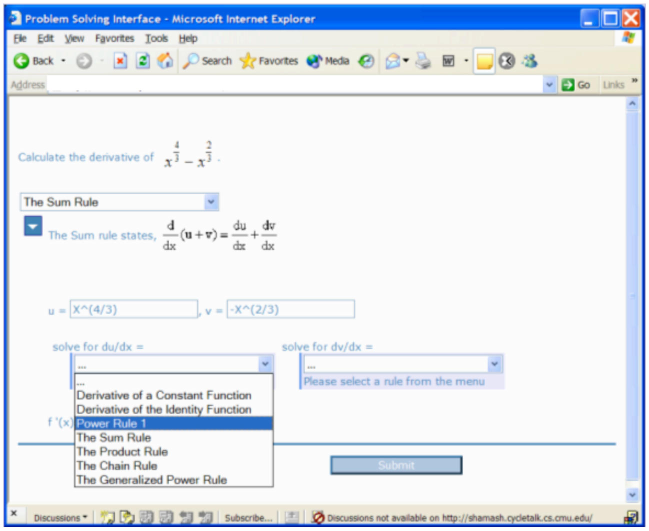
The experimental setting is displayed in Figure F.1. The student participant, tutor, and confederate peer learners were all located in separate rooms. The tutor and the peer learner roles were each played by two of the members on the research team each time. All students were told that their participation was part of a contest up front in all three studies. Also, the role of the student participants in all collaborative conditions across all studies was the same. Pairs working together interacted with a shared web-based problem solving interface using RealVNC software. The interface is described in greater detail below under the "Problem Solving Interface" section. They communicated with one another using MSN Messenger. During the tutoring phase (phase two), time stamped logs of chat behavior were recorded. During the problem solving phase (phase four), submitted solutions, points assigned for each problem, and all chat behavior were collected in time-stamped logs. In addition, all activity with the problem solving interface was recorded using Camtasia Studio software made by TechSmith Co. In the SOL condition, students worked alone during phase four using the same web based interface. Their interactions with the interface were also recorded using Camtasia Studio. They inserted think aloud comments in the MSN Messenger interface as well.

The materials for the experiment consisted of the following:

- An eight page web based lesson designed in collaboration with a calculus instructor from the Math department at Carnegie Mellon University. This lesson that focused on derivatives provided material for the tutor and student to work through during phase two. It consisted of an overview and individual units on each of seven specific rules of derivation. Each unit consisted of some explanation of the rule and an example problem for the student to work through using a structured problem solving interface.
- 12 on-line problem solving exercises for Phase four, each requiring the applications of multiple derivation rules.
- 2 extensive tests (Test A and Test B) were used for the pre-test (in phase one) and the posttest (in phase five). These tests each consisted of seven algebraic manipulation problems, seven simple calculus problems to test knowledge of each specific differentiation rule, and six complex calculus problems requiring both multiple rule applications and algebra. In order to maintain consistency of content coverage and difficulty across tests, each problem on test A had an isomorphic problem on Test B, which required the use of the same skills. To further control for test difficulty and coverage, I counterbalanced the order of the tests. In phase three, students took a middle test with eight simple calculus problems, isomorphic to the second section of tests A and B, and three complex calculus problems requiring multiple rule applications.
- Instructions for phase two were provided on paper. The instructions before the first instructional phase were identical for all conditions except that students in the solitary learning condition in the first study were told that they were preparing to solve problems independently,

whereas students in other conditions were told they were preparing to solve problems with a peer.

- Instructions for phase four were again identical for all but the SOL condition. The instructions for all but the SOL condition began with the following:
  - "During this portion of the experiment, you and another student will be working together to complete Calculus problems based on the rules you learned in part one. You will have 35 minutes to complete as many problems as you can (up to 12). Both students will be manipulating the webpage that you see on the screen. Each problem is worth up to two points. In order to get two points on a problem, both students must contribute equally to the problem solving and the solution must be correct. A correct solution where one student does the most problem solving will only be worth one point."
- The instructions for SOL were identical except that all mention of a peer problem solver and division of labor were deleted since these are not relevant for solving problems alone. Thus, in the SOL condition, students were told they would receive two points for a correct solution and 0 otherwise.


Problem Solving Interface - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address Go Links
Discussions Subscribe...
Discussions not available on http://shamash.cycletalk.cs.cmu.edu/
Submit

Calculate the derivative of $x^{\frac{4}{3}} - x^{\frac{2}{3}}$.

The Sum Rule

The Sum rule states, $\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$

u = X^(4/3), v = -X^(2/3)

solve for du/dx = ...

...
Derivative of a Constant Function
Derivative of the Identity Function
Power Rule 1
The Sum Rule
The Product Rule
The Chain Rule
The Generalized Power Rule

solve for dv/dx = ...

Please select a rule from the menu

f '(x)

Figure F.2. Structured problem solving interface.

All on-line problem solving was done using a structured problem solving interface designed for solving differentiation problems (Figure F.2). Students first select a rule from a menu. Based on their selection, some explanation about the rule and slots to fill in were presented to the student. In some cases, additional menus were presented, allowing for embedded rule applications.

No feedback was provided by the system based on the students' selections from the menu or from entries in the text input boxes during the problem-solving process. When the student or pairs of students were satisfied with their solution, they submitted it. If it was incorrect, they were then shown their incorrect derivation next to the correct one as a worked example including both the derivation and some explanation. The purpose was for them to compare and see how the problem should have been worked out and where their mistake occurred. See Figure F.3 for an example. This approach of offering a correctly worked out version of the problem as feedback was motivated by a long and well established tradition of learning from worked examples. Presenting students with correctly worked examples has been demonstrated to be highly effective for learning, even more effective than problem solving at early stages of skill acquisition (Atkinson, 2003; Renkl & Atkinson, 2003; Renkl, 2002; Schworm & Renkl, 2002).

PeerLearner: okay.. i think that's the answer, what do you think?
RealStudent: I have no idea
PeerLearner: so let's try submitting then
...
PeerLearner: damn it...we got it wrong again... i think we almost had it, right? we got product rule
RealStudent: shall we move on?
PeerLearner: but do you understand this?
RealStudent: hell no
PeerLearner: cause we're going to keep getting things wrong if we don't understand... so should we study this a little maybe?
RealStudent: sounds like a plan :)

Figure F.3. Dialogue between a confederate peer learner and a student participant when comparing their submitted faulty solution with the correct solution displayed next to it

In the case of a correct submission, the students moved on to the next problem. There was no need to compare their solution with an ideal solution if their solution was correct. At all times their current score was displayed next to an unchanging Highest Score.

F.3. Exploring the benefits of collaboration under non-ideal circumstances – study 4.1

In classroom settings, where collaborative learning has been used successfully, the teacher plays an important supportive role in facilitating productive student interactions (Meloth and Deering, 1999). However, when this support is absent, it is not clear whether collaborative learning will be beneficial because the students are still in the process of learning the materials. Therefore, the support that the students offer each other is imperfect. Nevertheless, I hypothesized that even in the absence of a teacher/ facilitator students would benefit from solving problems with a peer if they were rewarded for cooperating with one another. Although students in this study did receive some faculty support in the form of tutoring during phase two, the collaborative problem solving phase in which the experimental manipulation took place was unsupported. The students in the studies believed they were participating in a contest in which they would be rewarded both for their correct problem solving behavior as well as for the extent to which they kept their distribution of labor equal. This competition scenario (where dyads cooperated with one another, but competed with other dyads) is a typical collaborative learning configuration that provides a light form of non-invasive support for collaboration. The purpose of Study 4.1 was to measure the effect of collaborating with a peer on calculus problem solving, but in the context of light support.

F.3.1. Experimental manipulation

The experimental manipulation, which occurs during phase four of the experimental procedure discussed above, consisted of Solitary problem solving (SOL) and Naturalistic Peer-to-Peer problem solving (P2P). In the SOL condition, students solved problems alone during phase four, whereas in P2P students solved the same problems, but in pairs. Note that in contrast to the collaborative problem solving conditions in the second two studies, in the P2P condition, both students are student participants, not experimenters.

21 participants for whom I measured learning were undergraduates or administrative personnel at the Carnegie Mellon University. They were randomly assigned to the two conditions. 12

students were assigned to the P2P condition in six pairs, four of which were same gender pairs. Nine students were assigned to the SOL condition.

F.3.2. Findings

Using an ANCOVA with Post-test score as the dependent variable, Condition (SOL versus P2P) as the independent variable, and Pre-test score as a covariate, I verified that students in the P2P condition learned more than their peers in the SOL condition F(1,18) = 6.0, p<.05, MSE = 5.64, effect size = 1.1 standard deviations. I did not use the mid-test score as a covariate along with pretest score because it was not reliably correlated with post-test score with this population of students after I first factored out the effect of pretest score. Note that this is not a methodological problem because the experimental procedure up until the mid-test was identical across conditions. Although both high and low pretest students benefited from collaboration, there was a trend for high pretest students to benefit more than low pretest students. The gap between gain in the solitary condition and in the collaborative condition widens as pretest score increases. Specifically, in the two pairs where high pretest students were paired with very low pretest students, the high pretest student gained substantially more than predicted based on their pretest score.

F.3.3. Discussion

These results are important because they demonstrate, in a highly controlled setting, the value of collaboration in problem solving despite the fact that students are fallible. Students contribute both correct and incorrect problem solving actions, advice, and feedback. Nevertheless, the interaction is beneficial although the degree of benefit may differ. In the second study, I systematically explore the impact of the erroneous contributions made by peer problem solvers. In the subsequent study, I explore a mechanism that we can use to improve the effectiveness of collaboration using an adaptive computer based support mechanism.

F.4. Exploring the interaction between engagement and competence – study 4.2

While the first study demonstrated that randomly assigned pairs of students collaborated with one another in a way that lead to significantly more learning than a solitary problem solving control condition, this initial success led us to ask what strategies for matching students with learning partners would produce the optimal conditions for learning. I chose engagement and ability level as variables to manipulate since they are directly related to the amount of knowledge co-construction that students can contribute in learning sessions. In addition, these are standard variables used to measure student performance, specifically correctness of solutions and evenness of distribution of labor. There are many reasons to believe these variables might interact with one another. For example, while we observed a benefit for collaboration in the first study, even in the face of errors contributed by students working together, there is reason to believe that, as errors are contributed with much higher frequency, they would become a hindrance and a distraction.

F.4.1. Experimental manipulation

The experimental design for the second study was a 2X2 factorial design in which I varied two factors describing characteristics of a scripted confederate peer problem solver, namely Lazy(LA)/Engaged(EN) referring to the frequency of the confederate problem solver's contributions, and High(HI)/Low(LO) referring to the accuracy of the confederate peer learner's contributions.

During this phase of the experiment, one member of the research team acted as a confederate student and another experimenter kept track of score, timing, and distribution of labor in order to ensure that all students within the same condition were treated in a consistent way. The confederate student behaved according to the following rules:

- LA/EN: In the Lazy condition (LA), the confederate student contributed to solving the problem either by offering part of the solution in the chat window or by performing an action in the problem solving interface every 45 seconds. In the Engaged condition (EN), the confederate peer learner contributed every eight seconds.

- HI/LO: In the High performing condition (HI), the confederate student provided only correct contributions. In the Low performing condition (LO), the confederate student provided incorrect contributions 2/3 of the time. 2/3 was chosen after some pilot testing since it seemed unrealistic for even a low performing student to get incorrect answers 100% of the time.

36 university students and staff participated in the study, and were randomly assigned to the four conditions.

F.4.2. Findings

As predicted, I found a significant interaction effect using an ANCOVA with Post-test scores as the dependent variable, LA/EN and HI/LO as factors, and pretest and midtest scores as covariates $F(1,30) = 7.47$, $p < .05$, MSE= 7.41. In a post-hoc analysis using a Bonferroni test, the students in the Engaged High performing condition achieved significantly higher post-test scores than the students in the Engaged Low performing condition, $p < .05$. There was a marginal trend in favor of Lazy Low in comparison with Lazy High $p < .1$. Thus, within the Lazy condition, Low performing partners were marginally more effective while in the Engaged condition, Low performing partners were significantly worse.

The strongest predictor of student learning was the number of correct problems the pairs managed to submit during the problem solving phase (CorrectProb). I computed this with a linear regression between CorrectProb and Post-test score with effect of Pre-test score factored out. R-squared=.70, $p<.001$, N=36. There was a main effect of the HI/LO factor on the number of correct solutions contributed, with the effect of Pre-test and Mid-test scores used as covariates, $F(1,30) = 49.1$, $p < .001$, MSE=.93, effect size = 2.4 standard deviations. This makes sense since errors contributed as part of the problem solving process must be corrected in order to submit a correct solution. Thus, errors cause more work to be required for a correct solution, which is more work and thus takes more time. On the other hand, errors may be left uncorrected, in insofar as the problem solving may not take more time, but the solution that is submitted will not be correct, and thus will not increment the number of correct solutions.

Based on the above reasoning, a reduction in number of problems submitted was predicted. Thus, one would predict that Low performing confederate peer learners would be less effective as learning partners since their errors slow down the rate at which correct problems are submitted. With this in mind, it is surprising that students in the Lazy Low condition performed marginally better (rather than significantly worse) than the students in the Lazy High condition. Furthermore, an ANCOVA with post-test as the outcome measure, LA/EN and HI/LO as factors, and pre-test and CorrectProb as covariates, I found a significant crossover interaction effect explaining an additional 4% of the variance that provided some weak evidence that the errors contributed by the fake peer learners sometimes had a positive effect on student learning. $F(1,30) = 4.96$, $p<.05$, $MSE=10.68$. Student participants paired with Lazy Low performing confederate peer learners learned more than would be predicted based on their pretest score and how many correct problems they managed to submit alone.

F.4.3. Discussion

The results from the second study offer limited evidence for the instructional value of student exposure to errors. Nevertheless, the results suggest that students who are high engagement, but low in ability level, are dangerous learning partners. Working with high engagement, low ability level confederate peer learners was less effective for learning than any of the other conditions. It was significantly worse than working with high engagement, high ability level confederate peer learners. Thus, one goal for the design of an adaptive support for effective collaborative problem solving would be to slow down high engagement, low performing students so that they won't produce a harmful level of erroneous problem solving behavior that might confuse, distract, or hinder their peer. Furthermore, I observed a disturbing lack of teaching behavior in the conversational logs from the first two studies, thus demonstrating a need for support in this area. In study three, I evaluate the effectiveness of an adaptive support mechanism whereby prompts are strategically offered to students when either of these two needs are evidenced in the collaboration.

F.5. Evaluating the impact of adaptive collaboration support – study 4.3

I hypothesize that prompts that are offered only when deemed absolutely necessary will have more of an influence on conversation over time. For example, several studies have evaluated the impact of providing a social script that encourages productive consensus building behavior related to knowledge co-construction such as transactivity (Weinberger et al., 2004; Weinberger et al., 2005). Such conversational behavior is accomplished by assigning students to roles (i.e., case analyst or constructive critic) and providing prompts that target particular ways in which contributions may relate to each other, for example "We have not reached consensus concerning the following points:". Rather than providing this prompt each time students formulate a contribution, as is the current, non-adaptive approach, a more adaptive approach would be to offer this prompt only in cases where non-productive forms of consensus building are detected, for example, where students fall into a pattern of quick consensus building rather than discussing the reasons for their differing points of view. Recent work demonstrates that patterns such as this can be detected with a high degree of reliability in collaborative discourse (Domnez et al., 2005). Nevertheless, the potential disadvantage of this adaptive scripting approach is that students receive much less scaffolding overall. Thus, it is necessary to experimentally verify whether this dramatically reduced level of scaffolding will be sufficient to yield a noticeable effect on behavior and learning.

Applying supportive scripting in an online learning context, where we are concerned about student interactions in the environment over an extended period of time, raises new questions not yet previously explored in the literature on scripted collaboration from the computer-supported collaborative-learning community. While previous approaches to scripting vary along numerous dimensions, previous approaches to scripting were all static, one-size-fits-all approaches that were not sensitive to what was actually happening in the interactions. This can lead to over scripting (O'Donnell, 1999) or interference between different types of scripts (Weinberger et al., 2005). I hypothesize that over long periods of time students will begin to ignore the prompts that scripts are composed of, if they students them as not adapted to what is actually happening in the conversation.

F.5.1. Experimental manipulation

The experimental design for the third study was a 2X2 factorial design in which I varied one factor relating to characteristics of a confederate peer problem solver and one characteristic relating to adaptive collaboration support. Specifically, High(HI) versus Low(LO) was a replication from the previous study. Prompt (PR)/No Prompt(NP) referred to the presence or absence of adaptive collaboration support in the form of prompts. In all conditions, the confederate peer learner in this study followed the rules for Lazy (LA) peer learners from the previous study. Prompting was offered in one of four cases outlined below. The prompts given in each case were canned text, worked out in advance so that they were presented the same way each time. Each one relates either to curbing frequency of contribution of high engagement student participants or eliciting reflection and explanation from the student participant. The prompts were not meant to change the role of the student participant, but to encourage behavior for instructionally beneficial collaboration.

In the prompt condition, students were told that automated prompts would appear on their screen to support their collaboration, but not on the other student's screen. The list of circumstances under which students received prompts is found below. The exact prompts associated with these circumstances are listed in Table F.1.

- #1: Vague Help Offered - Confederate makes incorrect problem-solving action and participant offers vague / incomplete help (e.g., "That's wrong").
- #2: Answer Corrected but No Help Offered - Confederate makes incorrect problem-solving action and participant changes to correct problem-solving action.
- #3: Participant Working Independently - Participant makes five correct problem-solving actions without confederate contribution.
- #4: Insufficient Review of Incorrect Answer - Participant wants to move on from wrong answer page before 2 minutes have elapsed.

Table F.3. Prompts.

Case	Prompts
1	The other student would benefit from more explanation. Please elaborate on your correction.
2	Help the student understand your correction. The other student seems to be struggling with this section of the problem. Please offer your assistance.
3	Please be sure you are working with the other student to solve the problem. It seems like the other student has not contributed lately. Why don't you see if they need help?
4	It seems like you are moving on before understanding your errors. Please spend more time reviewing this page. Does the other student understand the errors made on this problem? Please share your understanding of this page with the other student.

There were minor differences in the instructions in study 4.3 compared to study 4.2. The student participants in all conditions were told as part of the instructions prior to phase two (with the tutor) that the other student would not receive tutoring. They were told that they should prepare to work with and teach the other students if they need help during the problem solving that occurs in phase four. This was reiterated in the second instruction sheet before the peer learning session in phase four. Students were also told that they would receive a bonus if the other student's score improves in the post test.

40 university students and staff participated in the study, randomly assigned to the 4 conditions.

F.5.2. Findings

Overall, the results from this experiment offer evidence in favor of the effectiveness of adaptive support for improving student behavior and learning. They also point towards specific ways in that the design for adaptive support should be modified in order for it to be more effective. Here, I will first explore the effect of the prompts on student behavior in depth. I will then examine the effect on learning.

I first evaluated whether the prompts offered to students had a significant effect on their behavior. Remember that the prompts were primarily for two purposes: namely, to regulate the frequency of contribution of students, and to increase the amount of teaching behavior students offered. To evaluate whether the prompts were effective for regulating the frequency of contribution of students, I first analyzed trends in change of distribution of labor over time in the problem solving logs for each student participant. I looked at the number of contributions made by student participants and confederate peer learners for each problem solution submitted. From this, I computed for each problem submitted a LaborDistribution score between 0 and 1 indicating how different from equal the distribution of labor was, with 0 being the best and 1 being the worst. This was computed by the following formula, where PLC indicates the number confederate peer learner contributions and SPC indicates the number of student participant contributions:

$$2 \text{ X Abs } (.5 – (PLC / (PLC + SPC)))$$

I then computed for each student an improvement score indicating the extent that the distribution of labor became more equal during problem solving. I did this by subtracting the LaborDistribution score of the final problem submitted with that of the first problem submitted. Positive values indicate an improvement in distribution of labor, whereas negative values indicate the opposite.

On average the LaborDistribution scores in the no prompt conditions remained stable over time, whereas in the prompt conditions where students received prompts there was improvement over time and a significant correlation between amount of improvement and number of prompts received (R-squared=.27, P< .05, N=20). Students remained out of the danger zone in the experimental condition, with an average LaborDistribution score of .32. On average only one prompt related to distribution of labor was required over the entire collaborative problem solving session, although some students received as many as three. Because not many prompts related to distribution of labor were required, there was no significant effect of the prompting manipulation on average distribution of labor between conditions. Nevertheless, based on the significant correlation between number of prompts and improvement in LaborDistribution score, I conclude from this that prompts are effective in manipulating student behavior.

The effect of the explanation-oriented prompts was more obvious upon inspection. The prompts had a local effect on explanation behavior in that I saw students attempt in all cases to follow the instruction offered in the prompt.

Thus, prompts had a positive effect on student behavior in the intended direction, and offered evidence in favor of my design for adaptive support. However, I also observed some negative effects of prompts on student performance that also negatively interacted with student learning within the LO condition. This finding led me to revise the design for adaptive support. In particular, there was a non-significant trend for distribution of labor prompts to reduce the number of correct problems solved within the LO conditions. Although the effect was not significant, the added noise in terms of number of correct problems submitted obscured the difference in learning between the PR and NP conditions. Remember that there was a large and statistically significant correlation between number of correct problems submitted and student learning. When I factor out this effect by including correct problems submitted as a covariate in an ANCOVA comparing pre to posttest gains of students in the PR condition to students in the NP condition, we see a significant benefit for prompting on student learning. $F(1,39) = 4.12$, $p < .05$.

In future iterations I plan to modify the distribution of labor prompts so that the ideal distribution of labor is dependent upon the relative ability levels of the two students. Upon reflection, it does not make sense to encourage students who contribute errors 2/3 of the time to take an equal role in the problem solving. However, completely discouraging their involvement is not recommended. So we need to further explore how to balance the concern over maximizing the number of correct problems submitted with optimizing the balance of engagement between partners.

Further analysis of the learning gains reveal additional insights for appropriate matching of students with optimal learning partners. I found a significant aptitude-treatment interaction, showing that High performing peer learners become less effective as learning partners as student pretest scores increase while Low performing peer learners become more effective as partners as student pretest scores increase. F(1,39) = 5.97, p < .05, N=40. The difference in effectiveness between High performing peer learners as partners versus Low performing peer learners as partners for high pretest students is in favor of Low performing peer learners, but as in the previous study, the difference is only marginal. However, the lack of significance could simply be due to a Type II error. Additional data is needed to test this. I did not find this aptitude-treatment interaction in my previous study because the range of pretest scores was much higher in the first study.

F.5.3 Discussion

The results from this third study are particularly interesting from the standpoint of supporting effective collaborative learning. First, these results demonstrate that an approach to automatic strategic prompting based on patterns in collaborative discourse have a significant impact on learning. Prior work has demonstrated excellent results automating the application of sophisticated, multi-dimensional coding schemes for characterizing the collaborative learning process to naturally occurring collaborative learning data (Donmez et al., 2005). Thus I believe the goal of adapting this technology for use in creating an environment that automatically offers students this form of strategic collaboration support is within our reach. Furthermore, the results from my investigations yield insights that can be used in matching students for effective learning together. If more data renders the difference in effectiveness of High Performing versus Low Performing confederate peer learners as learning partners, we can also use these results to motivate the design of effective pedagogical agents that are tailored to the competence of the students who will use them together as virtual learning partners.

F.6. Towards Dynamic Support for Project Teams – discussions for study 4.1, 4.2, 4.3

In this chapter, I studied different conditions in collaborative learning environments that might foster more knowledge co-construction. The last study in particular explored the potential for dynamic support in collaborative learning environments. Although the rest of my dissertation focuses on other aspects of the knowledge co-construction process and automatically assessing those processes, this vision for the development of dynamic support for collaborative learning has gained support in collaborative learning community. For instance, a symposium at ICLS 2008 (Rummel & Weinberger) and a workshop at ITS 2010 (Walker et al.) was held with regards to this topic. In addition, other PhD. dissertations have investigated issues related to dynamic support in more detail. Walker (2010) conducted extensive research related to support of helping behavior in peer tutoring in a problem solving context, which addresses similar questions to those discussed in study 4.2 above. Kumar and colleagues (Kumar et al., 2007; Chaudhuri et al., 2008; Chaudhuri et al., 2009; Kumar et al., 2010; Ai et al., 2010; Kumar & Rosé, in press) took the results from the Wizard-of-Oz studies discussed in Section F.3 and created a fully automatic architecture, where feedback from conversational agents could be offered to student groups working on a variety of small scale engineering design tasks.

What is in common between these two lines of research and the work I pursue in the rest this dissertation is the need for automatic tracking of group processes through machine learning technology (studies six, seven, eight, and nine). What new challenges come with this focus is an emphasis on collaborations that focus on more open-ended problems that frequently cannot be addressed within a single work session, and the need for assessment of group processes through speech rather than through text based interaction, since most of the work is conducted either face-to-face or in computer mediated contexts involving voice or video.

APPENDIX G

Automatic assessment of group processes using text from message boards: study 6

This chapter presents a preliminary study conducted for the automatic assessment of group processes using text data from discussion boards. The knowledge co-construction process is monitored to predict an individual student's productivity. In my first attempt at the automatic assessment of the knowledge co-construction process presented in this chapter, I used rudimentary features extracted from message board discussions, collected from an engineering project course over the course of a semester. The machine learning experiment results showed promise in revealing student knowledge sharing episodes. However, the text based discussions did not have many instances of knowledge sharing where students were engaged in tight interactions that built on each other's ideas, as expected when people engage in ICC. Therefore, I studied conversational speech data collected from face-to-face meetings in my later studies presented in the body of this dissertation (studies seven, eight, and nine).

G.1. Motivation and deigns

Given that the process assessment categories are observable and traceable by human annotators as presented in study two (chapter four), my next research challenge is identifying methods of automatically tracking those processes and displaying them to instructors, so that instructors can gain more insight into group processes. Joshi and Rosé (2007) found that machine learning techniques applied to chat logs from collaborative learning discussions were more accurate at ranking how well student groups learned together than humans observing the complete chat transcripts. Based on prior work such as this, I expect that it would be feasible to use machine learning technology to track important group processes by leveraging the data collected during the class.

My first attempt at automatic assessment of project course students began with data collected during the Spring of 2006 in the Rapid Prototyping of Computer Systems (RPCS) course that provided the context for the classroom study I conducted in Spring of 2007, described in an earlier chapter as study two. The RPCS course is a project oriented class where students work in teams. Among various communication media is the groupware system that is in the form of a threaded discussion environment where students post messages for communication. This groupware environment is known as the Kiva, which is used by the students in the PRCS course to coordinate their efforts (http:/kiva.ices.cmu.edu) (Finger et al.*,* 2006). The Kiva is a web-based, asynchronous collaboration tool that was first prototyped by the students in the RPCS course in 2003 under the auspices of an NSF CRCD Grant. The core interaction of the Kiva combines aspects of both email and bulletin boards to keep threaded discussions intact. Students can post documents, diagrams, conversations, meeting notes, notes to self, task assignments, and so on. The discussion pages are designed to feel like a chat session in which students respond easily to one another. For the rapid prototyping course, the instructors of the course have incorporated a *worklog* for students to track time spent, reflect on work, and plan for the coming period. Time and task can be consolidated by team and individual. Periodically, the instructors post reflective design questions in the weekly log. The group correspondence in the Kiva provides the instructors with important insights into group functioning.

While the Kiva captures data that could be valuable to instructors for gaining insights into group functioning, the sheer volume of correspondence is far too great for an instructor to keep up with. Typical Kivas have many thousands of posts organized into hundreds of threads. For example, the Spring 2006 RPCS course had 692 topic threads, each with an average of about 10 posts per topic. The students posted 1,244 files and they were still posting even after the class was officially over.

From an informal analysis of the conversations on the Kiva, it is apparent that most team discussions take place through the Kiva rather than through email; if a private email conversation results in something the whole team should know, the conversation is posted to the Kiva. A separate database is created for each class or research group that requests a Kiva, but within a Kiva, all members have access to essentially all the data. One indication of the success of the software is that 14 of the 20 class Kivas that have been created were requested by faculty because students in their class demanded to use it, rather than Blackboard, the official Carnegie Mellon course management software. Thus, in practice, this valuable resource goes unused. Part of my technical goal is to leverage this resource by using technology to make an assessment of student productivity from their conversational behavior in this online environment.

I began the analysis of this data using patterns of typical information requests extracted from a separate corpus of collaborative learning interactions. Using those patterns as queries, I was able to automatically extract 108 information requests such as, "Um I dont know if it just me but I did see a GPS sensor section in there and I posted stuff on GPS on Wed. Am I missing something?" Roughly half of these information requests were related to team coordination. However, the other half were more substantially related to design issues or basic skills required to carry out the work. I then examined the threads where these information requests were posted to see how team-mates responded. I found that a substantive response was forthcoming in only 73 of these cases, or roughly 68%. In other cases, I found no evidence of a response to an information request, and occasionally students received a dismissive response. For example, here is a case where a student posted a substantive information request:

> I've looked at the spreadsheet that [a team mate] posted above and I have a question about my part. I know that I am in charge of the math models but I really don't know much about the dashboard computations (distance, energy used by appliances, cars generating pollution, energy produced by solar power). Is it possible if any of you can help me in the beginning so that I can complete this part?

Rather than offering the help this student would have needed to be able to gain the skills to do this part of the work, the team reassigned the task to a different team member who already possessed this expertise. If the instructor had been aware of this incident, this would have been an excellent opportunity to step in and encourage students to take an approach that would maximize knowledge co-construction between students so that students can learn from each other. Therefore, in this study, I collected “features” that signal the knowledge co-construction process. For example, I looked at the activity level on the class discussion board such as number of new posts or number of different words used as a proxy for measuring the amount of knowledge co-construction process. My hypothesis is that a proxy has predictive power for forecasting the productivity based grades that instructors assign throughout the semester. If this prediction proves to be useful, then an automatic assessment system, which uses this technology, can be built to provide instructors with a summary of the knowledge co-construction process throughout the semester.

G.2. Procedure

My technical approach is to use machine learning to build models that can identify meaningful patterns in the conversational behavior of students. It can then use these to form an assessment of the student’s course behavior and thus provide a useful report to the instructor. Machine learning algorithms can learn mappings between a set of input features and a set of output categories or numeric predictions (such as a regression model). They do this by examining a set of hand coded “training examples” that exemplify each of the target categories. The goal of the algorithm is to learn rules by generalizing from these examples in such a way that the rules can be applied effectively to new examples. Once the input features have been set, a large number of training examples are then typically hand coded. Toolkits such as Weka (Witten & Frank, 2005) provide a large number of alternative machine learning approaches. This procedure I followed is presented in figure G.1 below.

I ran the evaluation with date from the RPCS spring 2006 course. Productivity grades are assigned three times in the semester and are an important means of communication between instructor and students about standing within the course. In this study, I investigated how evidence of student productivity extracted from student on-line behavior from discussion boards correlates with instructor judgment of productivity based on the instructor's human understanding of the work logs, observations in class, assessment of group product, and insights gained from weekly meetings.

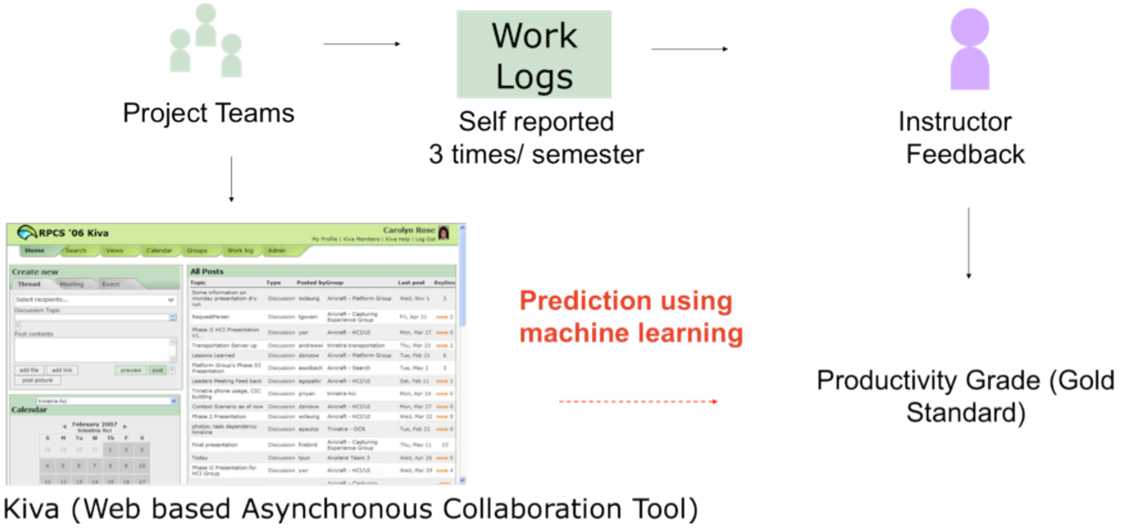


Figure G.1. Procedure

Measuring student online behavior from discussion boards

I first collected data from Kiva to predict productivity grade using machine learning technology. This process is indicated by the red arrow in figure G.1. Using the discussion board data, I collected a total of 1157 posts. I divided the data into collections of posts posted by an individual student on each specific week of the course. Altogether I assembled 510 data points from a total of 34 students during the 15 weeks of class. From this raw data, I extracted a set of seven features, referred to as feature set A~G, to use for features sets in conducting machine learning experiments. The baseline performance was based on using feature set A only.

The following is the list of feature sets A ~ D:

- Set A: Five meta features that was cited in previous work in the computer supported collaborative learning community as shallow measures of productivity (Kay, et. Al 2006, Kreijns & Kirschner, 2004); namely, they were: 1) phase in the semester; 2) week number; 3) number of posts; 4) average length of post; and5) number of attached files .

- Set B: Three other meta level features: 1) number of active days for the person; 2) average response time to post a reply; and 3) number of new posts.
- Set C: Unigram (i.e., single word stems), and bigrams (i.e., pairs of word stems occurring adjacent in the text) using TagHelper tools (Donmez et al., 2005; Rosé et al., 2008), which can be downloaded from http://www.cs.cmu.edu/~cprose/TagHelper.html. These features were both stemmed. I did not include words that occurred less than five times in the corpus or typical function words such as prepositions or determiners.
- Set D: Pennebaker's (2001) linguistic inquiry and word count (LIWC).

Next I used the support vector machine (svm) regression model included in the Weka toolkit (Witten & Frank, 2005) to build a predictive model. The SVM regression algorithm is a type of regression learner that learns from data how to weight various features that are provided. I began by extracting multiple features from the collections of posts that are thought to be relevant. Next, I labeled instances, which are vectors of such feature values, with the instructor assigned grades associated with the week the data is from. The regression learning algorithm then learnsedweights for various features depending on their predictive value with respect to the instructor assigned scores. Predictions can then be made for vectors of feature values by evaluating the learned linear function.

Instructor assigned productivity (gold standard)

Because the instructors assigned students grades three times throughout the semester, the grade associated with an individual week was the grade assigned to the student for the segment of the course that that week was part of. Since student productivity may vary from week to week throughout the semester, this target assessment may be somewhat noisy. However, it is the best objective measure I had for individual student productivity in the course.

I based the gold standard of predicting the productivity set by the instructors of the course, which was used to assign grades to the students. The grades were assigned based on the amount of work students self-reported in the work logs, which were required three times during the whole course of semester from the students. In addition, I incorporated the feedback on students from the instructor of the class. The productivity is rated on a five point scale, one being most productive. The instructor identified students who were exceptionally strong. Not surprisingly, all of them received an "A" in the class. The standard I used to assign the productivity is shown in table G.1.

Table G.1. Gold standard for productivity using instructor assigned grades.

Productivity	Standard
1	Grade "A", AND instructor identified as superior student
2	Grade "A"
3	Grade "A-" or "B+"
4	Grade "B"
5	Grade "C" OR instructor/ peer students identified as a poor student

G.3. Findings

In this section, I present the models that were built and tested based on the message board data from the Rapid Prototyping of Computer Systems (RPCS) spring 2006 course. I predicted students' productivity based on the messages posted through this medium.

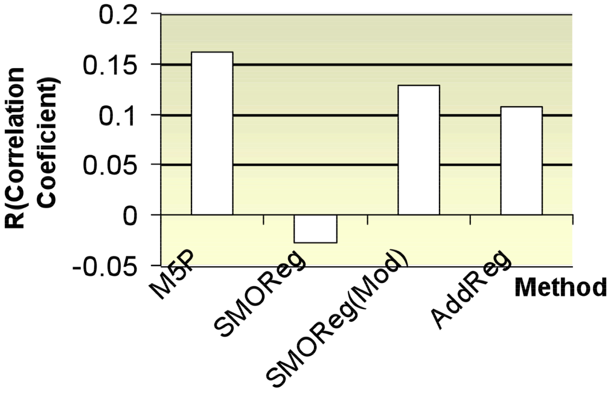
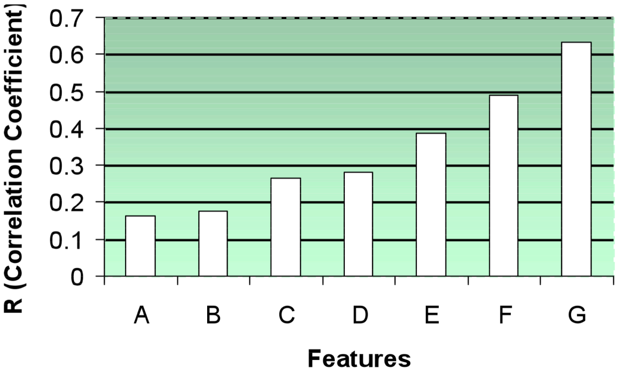


Figure G.2. Baseline results using various machine learning algorithms

I first used feature s of set A to measure the baseline using three classifying algorithms: M5P, SMO regression, and Linear regression. The result is shown in Figure G.2. The best performance on the baseline was 0.16 using the M5P algorithm.

Next, I used feature sets B, C, and D, in addition to the baseline feature set A. (Figure G.3), to show the machine learning results. The best performance achieved was a correlation coefficient of .63 with the full set of features and the smoothed predictions from the additive regression model, which is a substantial and statistically significant improvement over the optimistic baseline performance of 0.16.



- feature set A (Baseline)
- feature set A~B
- feature set A~C
- feature set A~D
- feature sets A~D using optimized SMO
- feature sets A~D using boosting
- feature sets A~D using smoothed value of predicted productivity

Figure G. 3. Machine learning result

Using a methodology where I trained on part of the data but test on the remainder of the data, I built a model that can make reasonably accurate assessments of productivity (r = 0.63). Interestingly, some of the most predictive linguistic features were social in nature. For example, I observed that words such as “Thanks,” “Hi” and “Please” ranked among the top attributes in the feature space. This might suggest that more polite and friendly group members make the group a more attractive environment to work in, and thus students exhibit less social loafing, as implied by the Collective Effort Model by Karau and Williams (1993), and are consequently more productive. This preliminary result shows promise that an automatic analysis of online communication behavior of groups can provide instructors with valuable early warning signs that some groups, or certain students within groups, require additional instructor support.

G.4. Discussion

The results from this study shows that shallow indicators that have been used in prior state-of-the-art work such as number of posts, number of attachments, and length of posts have no significant correlation with an instructor assigned productivity grade in the data (r = 0.16)(Kay et al., 2006a; Kay et al., 2006b; Kreijns and Kirschner, 2004). However, using a model constructed with the TagHelper toolset developed by Donmez and her colleagues (2005), I built a model that *can* make a prediction about student productivity based on data that was not used in training the model (r = 0.63).

My evaluation demonstrates that linguistic features extracted from on-going conversations significantly improve the accuracy of a model trained to predict an individual's level of productivity over time, when compared with a baseline model trained with more intuitive features used in prior work. These early investigations showed promise that meaningful automatic assessment could be conducted from the text posted to a groupware environment such as the Kiva. However, I was disappointed to find that very few substantive project discussions were conducted in this groupware environment. More valuable assessments could be made from the discussions that occur in face-to-face group meetings, which capture most of the important decisions made during meetings. In addition, speech as a data source is a more natural form of interaction than text. Speech contains non-verbal cues that provide information about social aspects of interactions. Therefore, in subsequent work, I use speech from group meetings as a data source (studies seven, eight and nine).

References

Ai, H., Kumar, R., Nguyen, D., Nagasunder, A., Rose, C. P. (2010). Exploring the Effectiveness of Social Capabilities and Goal Alignment in Computer Supported Collaborative Learning, in Poc. ITS.

Atkinson, R. (2003). Transitioning From Studying Examples to Solving Problems: Effects of Self-Explanation Prompts and Fading Worked-Out Steps, Journal of Educational Psychology, 95, 4.

Adams, R.S., Turns, J., Atman, C.J. (2003). "Educating Effective Engineering Designers: The Role of Reflective Practice." *Design Studies* 24, no. 3: 275-294.

Anderson, J.R., Boyle, C.F., Corbett, A., Lewis, M.W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, 42, (1990), 7-49.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proceedings of International Conference on Spoken Language Processing* (*ICSLP 2002)* (pp. 2037-2039). Denver, CO.

Argote, L., & Ingram, P. (2000). Knowledge transfer: A basis for competitive advantage in firms. *Organizational Behavior and Human Decision Processes, 82*, 150-169.

Automatic Speech Recognition for meetings (2010). Available at http://www-speech.sri.com/projects/meetings/

Azmitia, M., & Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. *Social Development*, 2, 202-221

Becker, B. E., Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology*, 71, 662-671.

Beehr, T., et, al. (2001). Evaluation of 360 degree feedback ratings: relationships with each other and with performance and selection predictors. *Journal of Organizational Behavior*. 22, 775-788.

Berkowitz, M., & Gibbs, J. (1983). Measuring the developmental features of moral discussion. *Merrill-Palmer Quarterly*, 29, 399–410.

Beskow, J., Sjlander, K. (2000) “Wavesurfer–an open source speech tool”, *Proc. of ICSLP*, Beijing, p464-467.

Bilous, F. & Krauss, R. (1988). Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language and Communication*, 8(3), 4.

Bjork, R. A. (1994). Memory and metameory considerations in the training of human beings. *In J. Metcalfe and A. Shimamula (Eds). Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press. pp, 185-205

Bourhis, R. & Giles, H. (1977). The language of intergroup distinctiveness. *Language, ethnicity and intergroup relations*, 13, 119.

Burger, J. (1991). Changes in attributions over time: the ephemeral fundamental attribution error. *Social Cognition*, 9:2. 182-193.

Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., & Huber, R. (2009). Detecting real life anger. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan. April 19-24, p.4761-4764.

Byard, V. (1989). Power play: The use and abuse of power relationships in peer critiquing. *In Proc. College Composition and Communication*. Seattle, WA.

Carroll, J. S. 1978. Causal attributions in expert parole decisions. *Journal of Personality and Social Psychology*, 36. 1501-1511.

Chen, M. (2003). Visualizing the pulse of a classroom. In Proc. MM’, ACM Press, 555-561.

Chaudhuri, S., Kumar, R., Joshi, M., Terrell, E., Higgs, F., Aleven, V., Rosé, C. P. (2008). It’s Not Easy Being Green: Supporting Collaborative “Green Design” Learning, *In Proc. of Intelligent Tutoring Systems.*

Chaudhuri, S., Gupta, N., Smith, N. A., Rosé, C. P. (2009). Leveraging Structural Relations for Fluent Compressions at Multiple Compression Rates. *Proceedings of the Association for Computational Linguistics.*

Clark, H. H., & Brennan, S. A. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp.127–149). Washington, DC: APA Books.

Corbett, A. T., & Anderson, J. R. (2002). Locus of feedback control in computer-base tutoring: impact on learning rate, achievement and attitudes. In *Proceedings of CHI 2002*, ACM, 245-252.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement.* 20, 37-46.

Cook, M., Klumper, D. (1999). Metacognitive, social and interpersonal skills and aptitudes in officer performance with distributed teams. *RTO HFM workshop on "Officer Selection"*, Monterey, USA.

Cooper, W. W. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.

Coupland, Nikolas. (2007). *Style: Language variation and identity.* Cambridge, U.K.: Cambridge University Press.

Cummings, J. N., and Kiesler, S. (2005). Collaborative research across disciplinary and organizational boundaries. *Social Studies of Science*, *35*, 703-722.

Dabbs, J., and Ruback, B. (1987). Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology, 20*, pp 123-169.

Damon, W, & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Education Research*, 13, 9-19.

DiMicco, J., Pandolfo, A., and Bender, W. (2004). Influencing group participation with a shared display. *Proc. CSCW 2004*, ACM Press, 614-623.

DiMicco, J., Hoolenbach, K., and Bender, W. (2006). Using visualizations to review a group's interaction dynamics. *Proc. CHI 2006*, ACM Press, 706-711.

Dimitracopoulou, A., Hoppe, U., Dillenbourg, P. (2004). Interaction analysis supporting participants during technology based collaborative activities. *CSCL symposium, Kaleidoscope Noe*, Lausanne, October 7-9, 2004.

Drach-Zahavy, A., Somech, A. (2001): Understanding Team Innovation: The Role of Team Processes and Structures. *Group Dynamics*, vol. 5, no. 2, pp. 111-123.

Dong, A., Hill, W.A. & Agogino, A. M., (2004): A document analysis technique for characterizing design team performance, *Journal of Mechanical Design*, vol. 126. no. 3. pp. 378–385.

Donmez, P., Rose, C. P., Stegmann, K., Weinberger, A., and Fischer, F. (2005): Supporting CSCL with Automatic Corpus Analysis Technology, *Proceedings of Computer Supported Collaborative Learning,* Laurence Earlbaum Associates, pp125-134.

Dutson, A.J., Todd, R.H., Magleby, S.P. & Sorensen, C.D. (1997). "A Review of Literature on Teaching Design Through Project-Oriented Capstone Courses." *Journal of Engineering Education* 76, no. 1: 17–28.

Eckert, P. & Rickford, J. (2001). *Style and sociolinguistic variation*. Cambridge Univ Pr.

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *Proc. Interspeech*.

Faidley, J., et. al. (2000). How are we doing? Methods of assessing group processing in a problem-based learning context. In Evensen, D. H., and Hmelo, C. E. (eds.), *Problem-Based Learning: A Research Perspective on Learning Interactions*, Erlbaum, NJ, 109-135.

Falchikov, N. (1995). Peer feedback marking – Developing peer assessment. *Innovations in Education and training International*, 32, 175-187.

Feeley, T. (2002). Evidence of Halo Effects in Student Evaluations of Communication. Instruction. Communication Education, 51(3), p.225- 236.

Fina, A., Schiffrin, D., & Bamberg, M. (2006). *Discourse and Identity*, Cambridge University Press.

Finger, S., et al. (2006). "Supporting Collaborative Learning in Engineering Design," International Journal of Expert Systems and Applications.

Fischer, F., Bruhn, J., Gruesel, C & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, 12, 213–232.

Forbes-Riley, K, Litman, D. (2009) Adapting to Student Uncertainty Improves Tutoring Dialogues. AIED 2009.

FrameNet. (2010). Available on http://framenet.icsi.berkeley.edu/index.php?option=com_ frontpage &Itemid=1.

Freund, Y., Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory. p.23-37.

Fussell, S. R., et al. (1998). Coordination, Overload and Team Performance: Effects of Team Communication Strategies. *Proceeding of conference on Computer-Supported Cooperative Work, Seattle, WA, Nov 14 – 18, 1998.* ACM Press, pp. 275-284.

Giles, H. (1984). The Dynamics of Speech Accomodation, Amsterdam: Mouton.

Giles, H. & Coupland, N. (1991). *Language: Contexts and consequences.* Thomson Brooks/Cole Publishing Co.

Giles, H., Mulac, A., Bradac, J., & Johnson, P. (1987). Speech accommodation theory: The next decade and beyond. *Communication yearbook*, 10, 13–48.

Guzdial, M., Rick, J., Kehoe, C. (2001). "Beyond adoption to invention: Teacher-created collaborative activities in higher education," *Journal of the Learning Sciences*, 10(3) , 265-280.

Gweon, G., Rose, C.P, Carey, Regan., Zaiss, Z. S. (2006). Providing Support for Adaptive Scripting in an On-line Collaborative Learning Environment. *CHI 2006.*

Gweon, G. (2008). Predicting Group Behavior from Audio Recordings of Meetings. In *Proceedings of ACM-SIGCHI Doctoral Consortium.*

Gweon, G., Kumar, R. Jun, S. Rosé, C. (2009). Towards Automatic Assessment for Project Based Learning Groups, In Proc. Artificial Intelligence in Education.

Gweon, G., et. al. (2011). "The automatic assessment of knowledge integration processes in project teams." *In Proc. Computer Supported Collaborative Learning*, Hong Kong, China.

Hain, T. et. al. (2005). Transcription of conference room meetings: An investigation. In *Proc. Interspeech 2005*, Lisbon, Portugal.

Hackman, R. (1987): *The design of work teams:* In J. Lorsch (Ed)., *Handbook of Organizational Behavior.* Englewood Cliffs, NJ: Prentice-Hall.

Harvey, J., Town, J. P., Yarkim, K. (1981). How fundamental is the fundamental attribution error? *Journal of Personality and Social Psychology*, 40:2. 346-349.

Hecht, M., Boster, F., & LaMer, S. (1989). The effect of extroversion and differentiation on listener adapted communication. *Communication Reports*, 2(1), 1–8.

Heller, R. S., Beil, C., Dam, K., and Haerum, B. (2010). "Student and faculty perceptions of engagement in engineering". *Journal of Engineering Education* 99, no. 3: 253-261.

Hmelo-Silver, C. E. (2004). Problem Based Learning: What and How Do Sutdents Learn? *Educational Psychology Review* 16, 235-266.

Horwitz, S., & Horwitz, I. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33, 987.

Jaffe, A. (2009). Stance: Sociolinguistic Perspectives, Oxford University Press.

Jain, M., McDonough, J., Gweon, G., Raj, B., Rose, C. P. (2012). An unsupervised dynamic bayesian network approach to measuring speech style accommodation. 13th conference of the European chapter of the association for computational linguistics.

Jochems, W. & Kreijns, K. (2006). Measuring Social Aspects of Distributed Learning Groups, *European Educational Research Journal*, Volume 5, pp110-121.

Joshi, M. & Rosé, C. P. (2007): Using Transactivity in Conversation Summarization in Educational Dialog. *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education,* October 1 – 3, 2007. Pennsylvania, USA.

Kaiser E, Demirdjian D, Gruenstein A, Li X, Niekrasz J, Wesson M, Kumar S. (2004). A multimodal learning interface for sketch, speak and point creation of a schedule Chart. In: Proceedings of *ICMI 2004*, pp 329–330.

Kane, A. A., Argote, L., & Levine, J. M. (2005). Knowledge transfer between groups via personnel rotation: Effects of social identity and knowledge quality. *Organizational Behavior And Human Decision Processes, 96*, 56-71.

Kane, A. (2010). Unlocking Knowledge Transfer Potential: Knowledge Demonstrability and Superordinate Social Identity. Organizational Science. Vol. 21. No 3. p.643-660.

Kapur, M., & Kinzer, C. K., (2009). Productive failure in CSCL groups. In Proc, CSCL 2009. 4: 21-46.

Karau, S., Williams, K. (1993). Social loafing: A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology, 65*, 681-706.

Kay, J., Maisonneuve, N., Yacef, K. & Reimann, P. (2006). “Wattle Tree: What’ll It Tell Us?”, *University of Sydney Technical Report* 582.

Kelsey, D. et al. (2004) College students' attributions of teacher misbehaviors. *Communication Education*, 53:1, 40-55.

Kim, J., Shaw, E., Chern, G, and Herbert, R. (2007). Novel tools for assessing student discussions: Modeling threads and participant roles using speech act and course topic analysis. *In proc. AIED*, 2007.

Kochakornjarupong, D. and Brna, P. (2010). Helping Feedback-Givers to Improve their Feedback. *International* Journal *of Continuing Engineering Education and Lifelong Learning*, 20(2), 148-168.

Koedinger, K. J., Anderson, R. J., Hadley, W.H., & Mark, M.A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education.* 8, 30-43.

Kollar, I., Fischer, F., & Hesse, F. W. (2003). Cooperation scripts for computer-supported collaborative learning. In Proceedings of the International Conference on Computer Support for Collaborative Learning (CSCL) 2003, 59-61.

Kolodner, J.L., Crismond, D., Gray, J., Holbrook, J., Puntambekar, S. (1998). “Learning by Design from Theory to Practice,” *International Conference of Learning Sciences*, 16-22.

Krauss, R.M. & Fussell, S.R. (1991). Perspective-taking in communication: Representations of others’ knowledge in reference. *Social Cognition*, 9, 2-24.

Kreijns et al. (2002): The sociability of computer-supported collaborative learning environments. *Journal of Education Technology & Society,* vol. 5, no. 1, pp. 8-22.

Kruger, A. C , & Tomasello, M. (1986). Transactive discussions with peers and adults. *Developmental Psychology*, 22, 681-685.

Kumar, R., Rose, C. P., Litman, D. (2006). Identification of Confusion and Surprise in Spoken Dialogusing Prosodic Features , Proceedings of Interspeech 2006.

Kumar, R. (2011). *Socially capable conversational agents for multi-party interactive situations.*

(Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. CMU-LTI-11-013).

Kumar, R. & Rosé, C. P., (2011). Architecture for building Conversational Agents that support Collaborative Learning, IEEE Transactions on Learning Technologies, vol. 4.1, pp. 21-34.

Labov, W. (1966). The social stratification of English in New York City, Washington DC: Center for Applied Linguistics.

Labov, W. (2010). *Principles of linguistic change: Internal factors*, volume 1. Wiley-Blackwell.

Lambart, E., Sharma, A., Levy, M. (1997). What information can relationship marketers obtain from customer evaluations of salespeople? Industrial Marketing Management. 26,177-187.

Levitan, R. & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech*.

Levitan, R., Gravano, A., & Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 113–117. Association for Computational Linguistics.

Liscombe, J. Venditti, J, Hirschberg, J. (2005). Detecting certainness in spoken tutorial dialogues. In Proc. Interspeech.

de Lisi, R., & Golbeck, S.L. (1999). Implications of the Piagetian Theory for peer learning. *Cognitive perspectives on peer learning*, pp. 3-37.

Loh, B., et al. (1998). “The Progress Portfolio: Designing Reflective Tools for a Classroom Context,” *CHI98*, Los Angeles, CA, ACM Press.

Ma, J., Shaw, E., and Kim, J. (2010). Computational Workflows for Assessing Student Learning. *In Proc. Intelligent Tutoring Systems* (2), 188-197.

Madan, A., Caneel, R., and Pentland, A. (2004). GroupMedia:Distributed Multimodal Interfaces. In *Proc. Sixth International Conference on Multimodal Interfaces ICMI04*.

McGrath, J. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice-Hall.

McPherson, M., Young, S. L. (2004). What students think when teachers get upset: Fundamental attribution error and student generated reasons for teacher anger. *Communication Quarterly*, 52:4, 357-369.

Meloth, M. S., Deering, P. D. (1999). The Role of the Teacher in Promoting Cognitive Processing During Collaborative Learning, in O'Donnell & King (Eds.) *Cognitive Perspectives on Peer Learning*, Lawrence Erlbaum Associates: New Jersey.

Moreland, R. L. (1999). Transactive memory: Learning who knows what in work groups and organizations. In L. Thompson, D. Messick, & J. Levine (Eds.), *Shared cognition in organizations: The management of knowledge* (pp. 3-31). Mahwah, N.J.: Erlbaum.

Nathan, M. J. (1998). Knowledge and situational feedback in a learning environment for algebra story problem solving. *Interactive Learning Environments*, 161-180

Nokes, T., Levine, J. M., Belenky, D., & Gadgil, S. (2010). Investigating the impact of dialectical interaction on engagement, affect, and robust learning, *Proc. International Conference of the Learning Sciences.*

O 'Donnell, A. M. (1999). Structuring Dyadic Interaction Through Scripted Cooperation , in O'Donnell & King (Eds.) *Cognitive Perspectives on Peer Learning*, Lawrence Erlbaum Associates: new Jersey, 1999.

Palinscar, A.S., Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction 1*, 117-175.

Pea, R. D. (1993): Practices of distributed intelligence and designs for education. *Distributed cognitions*, pp. 47-87.

Pennebaker, J. W, et. al. (2008). *Linguistic Inquiry and Word Count.* Retrieved April 24, 2008 from http://www.liwc.net/

Phielix, C., Prins, F., & Kirschner, P. (2009). The design of peer feedback and reflection tools in a CSCL environment, In Proc. 9th international conference on Computer supported collaborative learning - Volume 1.

Phielix, C., Prins, F., & Kirschner, P. (2010). Awareness of group performance in a CSCL-environment: Effects of peer feedback and reflection, Computers and Human Behavior 26(2), pp151-161.

Piaget, J. (1985). The equilibrium of cognitive structures: the central problem of intellectual development, Chicago University Press.

Pianesi, F., et. al. (2008). Multimodal support to group dynamics. *Personal and Ubiquitous Computing*. 12, 181-195.

Pon-Barry, et al. (2006). Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. IJAIED 16.2

Purcell, A. (1984). Code shifting hawaiian style: childrens accommodation along a decreolizing continuum. *International Journal of the Sociology of Language*, 1984(46), 71–86.

Putman, W. & Street Jr, R. (1984). The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language*, 1984(46), 97–114.

Ranganath, R., Jurafsky, D., McFarland, D. (2009). It's not you, it's me: Detecting flirting and its misperception in speed-dates. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp 334-342). Singapore.

Ren, Y., Kiesler, S. & Fussell, S. R. (2008). Multiple group coordination in complex and dynamic task environments: Interruptions, coping mechanisms, and technology recommendations. *Journal of Management Information Systems*, 25, 107-133.

Renals, S., Hain, T., & Bourlard, H. (2007). Recognition and interpretation of meetings: The AMI and AMIDA projects. Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding*.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations supplement self-explanations. *Learning & Instruction,* 12, 529-556.

Rienks, R.J., Zhang, D., Gatica-Perez, D., and Post, W. (2006). "Detection and Application of Influence Rankings in Small Group Meetings." *In Proc Eighth International Conference on Multimodal Interfaces*, 2006.

Rohde, M., et. al., (2007). Reality is our laboratory: communities of practice in applied computer science. *Behavior and Information Technology*, v26 n1, 81-94.

Rose, C.P., et. al. (2007). Towards an interactive assessment framework for engineering design learning. *In proc. DETC 2007*, Las Vegas, NV.

Rosé, C. P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of

Computational Linguistics in Computer-Supported Collaborative Learning, *International Journal of Computer Supported Collaborative Learning*.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, New York: Academic Press, 10, 173–220.

Rummel, N., Spada, H., Hauser, S. (2006). Learning to Collaborate in Computer-mediated Settings: Observing a Model Beats Learning from Being Scripted, *Proceedings of the International Conference of the Learning Sciences.*

Salomon, G. (1993): No distribution without individual cognition: A dynamic interactional view. *Distributed cognitions*, pp. 111-138.

Sanders, R. (1987). *Cognitive foundations of calculated speech*. State University of New York Press.

Scardamalia, M. (2004). "CSILE/Knowledge Forum®" in *Education and technology: An encyclopedia*, ABC-CLIO, Santa Barbara, 183-192.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*, 1-24.

Schober, M. F. & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In Graesser, A., Gernsbacher, M., and Goldman, S. (Eds.) *The handbook of discourse processes* (pp. 123-164). Mahwah, NJ: Lawrence Erlbaum Associates.

Schwartz, D. (1998). The productive agency that drives collaborative learning. *In Dillenbourg, P. (Ed.) Collaborative learning: Cognitive and computational approaches.*

Scotton, C. (1985). What the heck, sir: Style shifting and lexical colouring as features of powerful language. *Sequence and pattern in communicative behaviour*, pages 103–119.

Sharan, S. (1980). Cooperative Learning in Small Groups: Recent methods and Effects on Achievement, Attitudes, and Ethnic Relations. Review of Educational Research, Vol 50, No. 2, 241-271

Shneiderman, B. (1992): Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, vol. 11, no. 1, pp. 92-99.

Smith, M., Fiore, A. (2001): Visualization Components for Persistent Conversations. *Proceeding of conference on Human Factors on Computing Systems,* Seattle, WA, March 31 – April 5, 2001. New York: ACM Press, pp. 136-143.

Soller, A., Lesgold, A. (2003) A computational approach to analyzing online knowledge sharing interaction. In Proc. AIED 2003, 253-260.

Soller, A., Mones, A., Jermann, P., Muehlenbrock, M. (2005). From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. *Int. Journal of Artificial Intelligence.* 15 (4), 261-290.

Steiner, I. D. (1972). *Group process and productivity.* New York: Academic Press.

Stokes, M. E., Davis, C. S., & Koch, G. G. (2000). *Categorical data analysis using the SAS system* (2 ed.). Cary, NC: SAS Institute Inc.

Strijbos, J. W. (2004): *The Effect of Roles on Computer-Supported Collaborative Learning.* Ph. D dissertation, Open University of the Netherlands, The Netherlands.

Stolcke, A., Friedland, G., Imseng. D. (2010). Leveraging speaker diarization for meeting recognition from distant microphones. In *Proceedings of IEEE ICASSP*, Dallas.

Suthers, D. (2006). Technology affordances for inter-subjective meaning making: A research agenda for CSCL. *International Journal of Computer Supported Collaborative Learning*, 1: 315-337

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaborations? In L. B. Resnick, C. Pontecorvo, & R. Saljo (Eds.), *Discourse, tools, and reasoning: Situated cognition and technologically supported environments*. Heidelberg, Germany: Springer-Verlag.

Tetlock, P. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*. 48:3. 227-236.

Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology,* 4, 25-29.

Tonso, K. (2006). "Teams that Work: Campus Culture, Engineer Identity, and Social Interactions," *Journal of Engineering Education*, 95, 25–37.

Tudge, J. (1992). Processes and consequences of peer collaboration: A Vygotskian analysis. *Child Development*, 63, 1364-1379.

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: An evaluation of CASCADE. *The Journal of the Learning Sciences*, 8, 71-125.

Vygotsky, L.S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press

Waibel A, Steusloff H, Stiefelhagen R. (2004). CHIL: computer in the human interaction loop. In: *NIST ICASSP meeting recognition workshop*, Montreal, Canada

Wang, et al. (2007). Thinking Hard Together: The Long and Short of Collaborative Idea Generation for Scientific Inquiry, Proceedings of Computer Supported Collaborative Learning

Walker, E. (2005) Mutual Peer Tutoring: A Collaborative addition to the algebra-1 Cognitive Tutors. Young Researchers Track at AIED 2005

Walker, E. (2010). *Automated adaptive support for peer tutoring.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. CMU-HCII-10-107).

Webb, N., Nemer, K., Zuniga, S. (2002). Short Circuits or Superconductors? Effects of Group Composition on High-Achieving Students' Science Assessment Performance, *American Educational Research Journal*, 39, 4, 943-989.

Weinberger A., Fischer F. (2006). A framework to analyze argumentative knowledge construction in computer supported collaborative learning. Computers & Education; vol 46, pp.71 – 95.

Weinberger, A., Ertl, B., Fischer, F., Mandl, H. (2004). Cooperation scripts for learning via web-based discussion boards and videoconferencing. In P. Gerjets, P. A. Kirschner, J. Elen & R. Joiner (Eds.), *Instructional design for effective and enjoyable computer-supported learning. In Proc. EARLI SIGs,* Tübingen: Knowledge Media Research Center (2004), 22-28.

Weinberger, A., Ertl, B., Fischer, F., Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning. *Instructional Science, 33*, 1, 1-30.

Weingart, L. R. (1997). How did they do that? The ways and means of studying group process. *Research in Organizational Behavior, 19*, 189-239.

Welkowitz, J. & Feldstein, S. (1970). Relation of experimentally manipulated interpersonal perception and psychological differentiation to the temporal patterning of conversation. In *Proceedings of the 78th Annual Convention of the American Psychological Association*, volume 5, pages 387–388.

Williams, K., & O'Reilly, C. (1998). Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behavior*, 20, 77-140.

Witten, I. H., Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.

Wong, J., Oh, L. M., Ou, J., Yang, J. & Fussell, S. R. (2007). Conversational grounding in dialogues between an expert and multiple novices. *Proceedings of CHI 2007* (pp. 261-270). NY: ACM Press.