Statistical Discourse Analysis: A Method for Modelling Online Discussion Processes

Ming Ming Chiu [1] **and Nobuko Fujita** [2]

[1]Purdue University, USA
[2]University of Windsor, Canada
mingmingchiu@gmail.com

ABSTRACT: Online forums (synchronous and asynchronous) offer exciting data opportunities to analyze how people influence one another through their interactions. However, researchers must address several analytic difficulties involving the data (missing values, nested structure [messages within topics], non-sequential messages), outcome variables (discrete outcomes, rare instances, multiple outcome variables, similarities among nearby messages), and explanatory variables (sequences of explanatory variables, indirect mediation effects, false positives, and robustness of results). We explicate a method that addresses these difficulties (Statistical Discourse Analysis or SDA) and illustrate it on 1,330 asynchronous messages written and self-coded by 17 students during a 13-week online educational technology course. Both individual characteristics and message attributes were linked to participants' online messages. Men wrote more messages about their theories than women did. Moreover, some sequences of messages were more likely to precede other messages. For example, opinions were often followed by elaborations, which were often followed by theorizing.

KEYWORDS: Statistical discourse analysis, informal cognition, social metacognition

1 INTRODUCTION

The advantages of online discussions over face-to-face discussions have led to extraordinary growth in both online courses and online discussion data. In traditional classrooms, students talk face-to-face at the same time and in the same place. In contrast, students engaged in online discussions can participate from different places (and at different times in the case of asynchronous discussions). By facilitating participation from different places, online discussions enable a broader range of people to interact with one another. With 6.7 million students taking at least one online course and an enrolment growth rate of 9.3 percent, online courses and programs continue to grow at faster rates than those of higher education overall (Allen & Seaman, 2010).

Learning in online courses can be as effective as learning in traditional classrooms, and perhaps significantly more so when they are well designed and well implemented (Tallent-Runnels et al., 2006). Elements of effective online course design includes establishing a strong social presence (Rourke, Anderson, Garrison, & Archer, 1999), creating an online learning community (Palloff & Pratt, 2007), and

facilitating interactivity through a clear course structure and structured learning activities (Kanuka, Rouke, & Laflamme, 2007). Well-designed online learning activities encourage students to contribute thoughtful messages to asynchronous online discussions for deeper learning and knowledge creation.

Online courses also provide several advantages over face-to-face courses. By allowing participants to interact at different times, online participants have more time than those in face-to-face conversations to gather information, contemplate ideas, and evaluate claims before responding; as a result, they often display higher levels of decision making, problem solving, and writing (Hara, Bonk, & Angeli, 2000; Luppicini, 2007; Tallent-Runnels et al., 2006). During higher quality discussions, students explain and synthesize ideas more often, so they typically learn more (Clark & Sampson, 2008; Glassner, Weinstoc, & Neuman, 2005). The explosion of online discussions leaves extensive data traces, unlike most face-to-face discussions (unless they are taped). This wealth of data creates exciting opportunities to understand productive versus unproductive discussions, their respective causes, and the potential for effective interventions — central issues in the emerging field of learning analytics (LAK, 2011; Siemens, 2013).

Siemens (2011) defined learning analytics as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs." This definition highlights the bridging function of data analysis, to understand and optimize the learning process, in various research traditions (including computer-supported collaborative learning, academic analytics, and educational data mining; Haythornthwaite, de Laat, & Dawson, 2013). We draw on and extend learning analytics techniques from the computer-supported collaborative learning tradition to investigate learning interactions through online discussions.

Earlier studies have applied content analysis to data from online forums to explore how specific actions (e.g., "why" or "how" questions, explanations, evidence, summaries) might be related to individual learning (Lee, Chan, & van Aalst, 2006; Lin & Lehman, 1999). While such aggregate counts provide descriptive summaries, they do not fully utilize the information relating to the time and order of collaboration and learning processes (Reimann, 2009), or capture the sequential data needed to test hypotheses about how group members' actions/posts/messages are related to one another (Chiu, 2008a). Furthermore, content analysis is a time-consuming, laborious, and reductive analysis that may not be practical for large data sets or for classroom intervention studies (Fujita, 2013).

Complementing this content analysis, learning analytics researchers from computer-supported collaborative learning have applied techniques such as social network analysis to examine networks among learners in online discussion environments over time (Haythornthwaite, 2001; de Laat, Lally, Lipponen, & Simons, 2007). Describing and understanding patterns of interaction that form among two or more learners may help identify productive versus unproductive discussions, but on its own, social network analysis does not support analysis of the progression of different types of cognition in the discourse (Teplovs & Fujita, 2013).

Meanwhile, discourse-centric learning analytics goes beyond surface measures to investigate the quality of the learning process, specifically the rhetorical dimensions — rhetorical roles and rhetorical moves — to improve discourse for deeper learning and learning design (De Liddo, Buckingham Shum, Quinto, Bachler, & Cannavacciuolo, 2011). This computational approach draws on discourse analysis, argumentation theory, and learning dialogue visualization to examine the learning process in online discussions. It introduces two main discourse elements: 1) *post type*, expressing the rhetorical role of the post in the conversation; and 2) *semantic connection*, expressing the rhetorical move that the author of the post wanted to make towards a particular post or participant. Discourse-centric learning analytics can thus provide learning analytics on these two discourse elements for individual learners and groups of learners. It can answer questions about learners' attention, rhetorical attitude to discourse contributions, topics distribution, and social interactions. In particular, making explicit the rhetorical role that a learner contributes to a discussion (e.g., identifying a problem, challenging a viewpoint, contributing new data) allows researchers to analyze the semantic connections between posts and their authors' practices in a computational way.

In a similar vein, analyses of sequences of messages can illuminate the relationships among processes that contribute to new ideas or theorizing by testing whether some types of messages (e.g., *asking for an explanation*) or sequences of messages (*different opinion* followed by *asking for explanation*) often precede them. These results can help us understand the temporal and causal relationships among different types of messages or message sequences that aid or hinder learning. We show how statistical discourse analysis (SDA; Chiu, 2008b) can model these sequences to test these hypotheses. To explicate SDA, we introduce data (Fujita, 2009) and hypotheses to contextualize the methodological issues. Specifically, we test whether three types of cognition (*informal opinion, elaboration,* and *evidence*) and three types of social metacognition (*ask for explanation, ask about use,* and *different opinion*; Chen, Chiu, & Wang, 2012) increase the likelihoods of *new information* or *theoretical explanations* in subsequent messages. This example shows how SDA might be fruitfully applied to large data sets (e.g., massive online open courses, MOOCs) as a vital learning analytics tool.

2 DATA

In this study, we examine asynchronous, online forum messages written by students in a 13-week online, graduate, educational technology course delivered using Web-Knowledge Forum (KF). These data are the second iteration of a larger design-based research study (Fujita, 2009). Data sources included questionnaire responses, learning journals, and discourse in KF. One of the authors participated in the course both as a design researcher collaborating closely with the instructor and as a teaching assistant interacting in course discussions with students. The goals for this study were twofold: to improve the quality of online graduate education in this particular instance, and to contribute to the theoretical understanding of how students collaborate to learn deeply and create knowledge through progressive discourse (Bereiter, 1994, 2002).

2.1 Participants

Seventeen students (12 females, 5 males) participated (see Table 4). Their ages ranged from mid-20s to mid-40s. Five were students in academic programs (4 M.A., 1 Ph.D.), and 12 were students in professional programs (9 M.Ed., 3 Ed.D.).

2.2 Procedure

The instructor encouraged students to engage in progressive discourse through three interventions: a reading by Bereiter (2002), classroom materials called Discourse for Inquiry (DFI) cards, and the scaffold supports feature built into KF. The DFI cards were adapted from classroom materials originally developed by Woodruff and Brett (1999) to help elementary school teachers and pre-service teachers improve their face-to-face collaborative discussions. The DFI cards model thinking processes and discourse structures to help online graduate students engage in progressive discourse in KF. There were three DFI cards: *Managing Problem Solving* outlined commitments to progressive discourse (Bereiter, 2002); *Managing Group Discourse* suggested guidelines for supporting or opposing a view; and *Managing Meetings* provided two strategies to help students deal with anxiety. The cards were in a portable document file (.pdf) that students could download, print out, or see as they worked online.

KF, an extension of the CSILE (Computer Supported Intentional Learning Environment), is specially designed to support knowledge building. Students work in virtual spaces to develop their ideas, represented as "notes," which we will call "messages" in this paper (see Figure 1). KF offers sophisticated features conducive to learning analytics that are not available in other conferencing technologies, including "scaffold supports" (labels of thinking types), "rise-above" (a higher-level integrative note, such as a summary or synthesis of facts into a theory), and a capacity to connect ideas through links between messages in different views. Students select a scaffold support and typically use it as a sentence opener while composing messages; hence, they self-code their messages by placing yellow highlights of thinking types in the text that bracket segments of body text (see Figure 2).
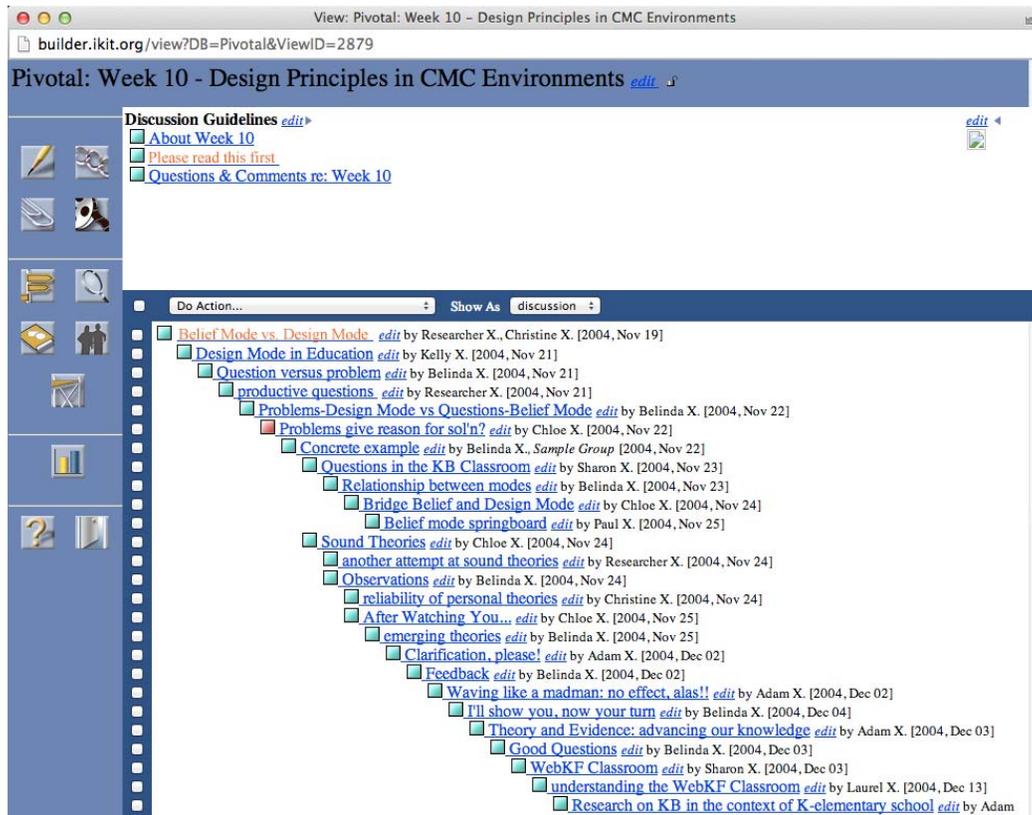
Figure 1: KF view showing thread structure of messages.

At the beginning of the course, only the Theory Building and Opinion scaffolds built into KF were available. Later, in week 9, two students designed the "Idea Improvement" scaffolds (e.g., what do we need this idea for?) as part of their discussion leadership (see Table 1). The Idea Improvement scaffolds were intended by their student designers to emphasize the socio-cognitive dynamics of "improvable ideas," one of the twelve knowledge building principles (Scardamalia, 2002) for progressive discourse. In this study, we focus our analysis on tracing messages with scaffold supports that build on or reply to one another. Types of scaffold supports relevant to our hypotheses are organized and renamed (*italicized*) in terms of cognition, social metacognition, and dependent variables.

Figure 2: KF Message with scaffold supports, link, annotation, and other information.

For analytic purposes, the KF scaffold supports enable self-coding in addition to external coding. Rather than relying on external coders' interpretations of participant intent for each message, participants identify their intents by self-coding their message with their scaffold supports. However, KF scaffold users might not understand the meanings of the scaffold supports in the same way as expert external coders. To address this problem, 56 segments of student discourse containing a scaffold support were randomly selected from the sample to check to see if a neutral observer could predict the scaffolds that

students used in the database. The scaffold support that the participants used was omitted from the text, and another graduate student was asked to guess the appropriate scaffold support based on the discourse processes reflected in the text. These graduate students correctly predicted the scaffold support 79 percent of the time.

Table 1: Knowledge forum scaffolds and scaffold supports used in iteration 2

Scaffolds

Cognition	Social Metacognition	Dependent variables
Opinion	*Ask for explanation*	*Theorize/Explain*
I think knowledge building takes a long time.	I need to understand why knowledge building has to take a long time.	My theory of the time needed for knowledge building is based on its sequence of parts…
Elaboration	*Ask about use*	*New information*
I think knowledge building takes many smaller steps.	Why do we need to understand how much time knowledge building takes?	Scardamalia and Bereiter's (1994) study showed that computer supports can support knowledge building in classroom learning communities.
Anecdotal evidence	*Different opinion*	
Last week, our class took over an hour to come up with a good theory.	I don't think knowledge building has to take a long time. It might depend on the people.	

2.3 Data Extraction

KF uses a database called a tuplebase based on the Zoolib cross-platform open-source library. Data were extracted from the KF using a JavaScript Object Notation (JSON) interface. First, a python script was run to extract all of the view links from the relevant course week discussion "views" or folders, followed by all of the "notes" or messages contained in those virtual spaces. Each message was identified by a number and information such as the message title, textual content, authorship, a list of participants who have read the note. The scaffolds indicating message types were made available for analysis. Second, the data were exported as a widely adopted software file: a comma separated value (CSV) file output (Teplovs, 2013).

3   HYPOTHESES

As shown in Table 2, we tested whether recent cognition or social metacognition facilitated new information or theoretical explanations (Chiu, 2000; Lu, Chiu, & Law, 2011). Introducing new information and creating theoretical explanations are both key processes that contribute to knowledge building discourse. New information provides grist that theoretical explanations can integrate during discourse to yield learning. As students propose integrative theories that explain more facts, they create knowledge through a process of explanatory coherence (Thagard, 1989). Hence, new information and theoretical explanations are suitable target processes to serve as dependent variables in our statistical model.

Table 2: Hypothesized effects of online processes on *new information* and *theorizing*

Explanatory variables	Dependent variables	
Cognition	New information	Theorizing
Opinion	+	+
Elaboration	ns	+
Anecdotal evidence	ns	+
Social metacognition		
Ask about use	+	+
Ask for explanation	ns	+
Different opinion	ns	+

Symbols in indicate expected relationships with the outcome variables: positive and supported [+], hypothesized but not supported [ns].

Researchers have shown that many online discussions begin with sharing of opinions (Gunawardena, Lowe, & Anderson, 1997). Students often activate familiar, informal concepts before less familiar, formal concepts (Chiu, 1996). During a discussion, comments by one student (e.g., a keyword) might spark another student to activate related concepts in his or her semantic network and propose a new idea (Nijstad, Diehl, & Stroebe, 2003). When students do not clearly understand these ideas, they can ask questions to elicit new information, elaborations or explanations (Hakkarainen, 2003). In addition, students may disagree (offer different opinions) and address their differences by introducing evidence or explaining their ideas (Howe, 2009).

Specifically, we tested whether three types of cognition (informal opinion, elaboration, and evidence) or three types of social metacognition (ask for explanation, ask about use, and different opinion) increased the likelihoods of new information or theoretical explanations in subsequent messages. Whereas individual metacognition is monitoring and regulating one's own knowledge, emotions, and actions (Hacker & Bol, 2004), *social metacognition* is defined as group members' monitoring and controlling one

another's knowledge, emotions, and actions (Chiu & Kuo, 2009). To reduce omitted variable bias, additional individual and time explanatory variables were added. For example, earlier studies suggest that males and females interact differently (Lu et al., 2011).

4   ANALYSIS

To test the above hypotheses, we must address analytic difficulties involving the data, the dependent variables and the explanatory variables (see Table 3). Data issues include missing data, nested data, and the tree structure of online messages. Difficulties involving dependent variables include discrete outcomes, infrequent outcomes, similar adjacent messages, and multiple outcomes. Explanatory variable issues include sequences, indirect effects, false positives, and robustness of results. SDA addresses each of these analytic difficulties, as described below.

SDA addresses the data issues (missing data, nested data, and tree structure of online messages) with Markov Chain Monte Carlo multiple imputation (MCMC-MI), multilevel analysis, and identification of the previous message. Missing data (due to uncoded messages, computer problems, etc.) can reduce estimation efficiency, complicate data analyses, and bias results. By estimating the missing data, MCMC-MI addresses this issue more effectively than deletion, mean substitution, or simple imputation, according to computer simulations (Peugh & Enders, 2004).

Table 3: Statistical Discourse Analysis strategies to address each analytic difficulty

Analytic difficulty	Statistical Discourse Analysis strategy
Data set	
• Missing data (0110??10)	• Markov Chain Monte Carlo multiple imputation (Peugh & Enders, 2004)
• Nested data (Messages within Topics)	• Multilevel analysis (Snijders & Bosker, 2012)
• Tree structure of messages (Λ)	• Store preceding message to capture tree structure (Chen, Chiu, & Wang, 2012)
Dependent variables	
• Discrete variable (yes/no)	• Logit/Probit (Kennedy, 2008)
• Infrequent variable	• Logit bias estimator (King & Zeng, 2001)
• Similar adjacent messages ($m_3 \sim m_4$)	• I^2 index of Q-statistics (Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006)
• Multiple dependent variables ($Y_1, Y_{2, \ldots}$)	• Multivariate outcome models (Snijders & Bosker, 2012)
Explanatory variables	
• Sequences of messages	• Vector Auto-Regression (VAR; Kennedy, 2008)

$(X_{t-2} \text{ or } X_{t-1} \rightarrow Y_t)$

• Indirect, multi-level mediation effects $(X \rightarrow M \rightarrow Y)$	• Multilevel M-tests (MacKinnon, Lockwood, & Williams, 2004)
• False positives (Type I errors)	• Two-stage linear step-up procedure (Benjamini, Krieger, & Yekutieli, 2006)
• Robustness	• Single outcome, multilevel models for each outcome
	• Testing on subsets of the data
	• Testing on original data

Messages are nested within different topic folders in the online forum, and failure to account for similarities in messages within the same topic folder (versus different topic folders) can underestimate the standard errors (Snijders & Bosker, 2012). To address this issue, SDA models nested data with a multilevel analysis (also known as hierarchical linear modelling; Snijders & Bosker, 2012).

Unlike a linear, face-to-face conversation in which one turn of talk follows the one before it, an asynchronous message in an online forum often follows a message written much earlier. Still, each message in a topic folder and its replies are linked to one another by multiple threads and single connections in a tree structure. See Figure 3 for an example of a topic message (1) and its 8 responses (2, 3, ... 9).
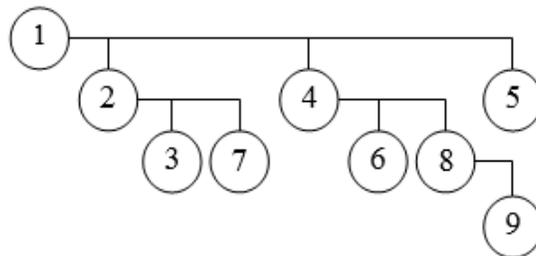
Figure 3: Tree structure showing how nine messages are related to one another.

These nine messages occur along three discussion threads: (a) 1 → 2 (→ 3; → 7), (b) 1 → 4 (→ 6; → 8 → 9) and (c) 1→ 5. Messages in each thread are ordered by time, but they are not necessarily consecutive. In thread (b) for example, message #6 responds to message #4 (not #5). To capture the tree structure of the messages, we identify the immediate predecessor of each message (Chen et al., 2012). Then, we can reconstruct the written reply structure of the entire tree to identify any predecessor of any message.

SDA addresses the dependent variable difficulties (discrete, infrequent, serial correlation, and multiple) with Logit regressions, a Logit bias estimator, I^2 index of Q-statistics, and multivariate outcome analyses. The dependent variables are often discrete (a justification either occurs in a conversation or it does not;

yes versus no) rather than continuous (e.g., test scores). As a result, applying standard regressions, such as ordinary least squares, to discrete dependent variables can bias the standard errors. To model discrete dependent variables, we use a Logit regression (Kennedy, 2008). As infrequent dependent variables can bias the results of a Logit regression, we estimate the Logit bias and remove it (King & Zeng, 2001).

Adjacent messages are often more closely related to one another than messages that are far apart, and failure to model this similarity (serial correlation of errors) can bias the results (King & Zeng, 2001). An I^2 index of Q-statistics tests all topics simultaneously for serial correlation of residuals in adjacent messages (Huedo-Medina et al., 2006). If the I^2 index shows significant serial correlation, adding the dependent variable of the previous message as an *explanatory* variable often eliminates the serial correlation (e.g., when modelling the outcome variable *theory*, add whether it occurs in the previous message [*theory* (–1)] (Chiu & Khoo, 2005); see paragraph below on vector auto-regression).

Multiple outcomes (*new information*, *theorizing*) can have correlated residuals that underestimate standard errors (Snijders & Bosker, 2012). If the outcomes are from different levels, separate analyses must be done at each level, as analyzing them in the same model over-counts the sample size of the higher-level outcome(s) and biases standard errors. To model multiple outcomes properly at the same level of analysis, we use a multivariate outcome, multilevel analysis, which models the correlation between the outcomes (*new information*, *theorizing*) and removes the correlation between residuals (Snijders & Bosker, 2012).

Furthermore, SDA addresses the explanatory variable issues (sequences, indirect effects, false positives, robustness) with vector auto-regression, multilevel M-tests, the two-stage linear step-up procedure, and robustness tests. A vector auto-regression (VAR; Kennedy, 2008) combines attributes of sequences of recent messages into a local context (*micro-sequence context*) to model how they influence the subsequent messages. For example, the likelihood of *new information* in a message might be influenced by attributes of earlier messages (e.g., *different opinion* in the previous message) or earlier authors (e.g., *gender* of the author of the previous message).

Multiple explanatory variables can yield indirect, mediation effects or false positives. As single-level mediation tests on nested data can bias results downward, multi-level M-tests are used for multilevel data — in this case, messages within topics (MacKinnon et al., 2004). Testing many hypotheses of potential explanatory variables also increases the likelihood of a false positive (Type I error). To control for the false discovery rate (FDR), the two-stage linear step-up procedure was used, as it outperformed 13 other methods in computer simulations (Benjamini et al., 2006).

To test the robustness of the results, three variations of the core model can be used. First, a single outcome, multilevel model can be run for each dependent variable. Second, subsets of the data (e.g., halves) can be run separately to test the consistency of the results for each subset. Third, the analyses can be repeated for the original data set (without the MCMC-MI estimated data).

4.1 Analytic Procedure

After MCMC-MI of the missing data (less than 1 percent) to yield a complete data set, each online message's preceding message was identified and stored to capture the tree structure of the messages. Then, we simultaneously modelled two process variables in students' messages (*new information* and *theorizing*) with SDA (Chiu, 2001).

$$\textbf{Knowledge\_Process}_{ymt} = \beta_y + \textbf{e}_{ymt} + \textbf{f}_{yt} \qquad\qquad (1)$$

For **Knowledge_Process**$_{ymt}$ (the knowledge process variable y [e.g., *new information*] for message m in topic t), β_y is the grand mean intercept (see Equation 1). The message- and topic-level residuals are **e**$_{mt}$ and **f**$_t$ respectively. As analyzing rare events (target processes occurred in less than 10 percent of all messages) with Logit/Probit regressions can bias regression coefficient estimates, King and Zeng's (2001) bias estimator was used to compute and remove this bias.

First, a vector of student demographic variables was entered: *male* and *young* (**Demographics**; see Equation 2). Each set of predictors was tested for significance with a nested hypothesis test (χ^2 log likelihood; Kennedy, 2008).

$$\textbf{Knowledge\_Process}_{ymt} = \beta_y + \textbf{e}_{ymt} + \textbf{f}_{yt} + \beta_{ydt}\textbf{Demographics}_{ymt} + \beta_{yst}\textbf{Schooling}_{ymt}$$
$$+ \beta_{yjt}\textbf{Job}_{ymt} + \beta_{yxt}\textbf{Experience}_{ymt} + \beta_{ypt}\textbf{Earlier\_Action}_{ym(t-1)}$$
$$+ {}_{ypt}\textbf{Earlier\_Action}_{ym(t-2)} + \gamma_{ypt}\textbf{Earlier\_Action}_{ym(t-3)} \ldots \qquad (2)$$

Next, schooling variables were entered: *doctoral* student, *Master of Education* student, *Master of Arts* student, and *part-time* student (**Schooling**). Then, students' job variables were entered: *teacher*, *post-secondary teacher*, and *technology* (**Job**). Afterwards, students' experience variables were entered: *KF experience* and *number of past online courses* (**Experience**).

Then, attributes of the previous message were entered: *opinion* (-1), *elaboration* (-1), *anecdote* (-1), *ask about use* (-1), *ask for explanation* (-1), *different opinion* (-1), *new information* (-1), *theory* (-1), and *any of these processes* (-1) (**Earlier_Action** $_{ym(t-1)}$). The attributes of the message two responses ago along the same thread (-2) were entered (**Earlier_Action** $_{ym(t-2)}$), then, those of the message three responses ago along the same thread (-3) (**Earlier_Action** $_{ym(t-3)}$), and so on until none of the attributes in a message were statistically significant.

Structural variables (**Demographics**, **Schooling**, **Job**, **Experience**) might show moderation effects, so a random effects model was used. If the regression coefficients of an explanatory variable in the **Earlier_Action** message (e.g., evidence; $\beta_{ypt} = \beta_{yt} + \textbf{f}_{yj}$) differed significantly ($\textbf{f}_{yj} \neq 0$?), then a moderation effect might exist, and their interactions with processes were included.

The multilevel M-test (MacKinnon et al., 2004) identified multilevel mediation effects (within and across levels). For significant mediators, the percentage change is 1 — (b'/b), where b' and b are the regression coefficients of the explanatory variable, with and without the mediator in the model, respectively. The odds ratio of each variable's total effect (TE = direct effect plus indirect effect) is reported as the increase or decrease (+TE% or –TE%) in the outcome variable (Kennedy, 2008). As percent increase is not linearly related to standard deviation, scaling is not warranted.

An alpha level of .05 was used. To control for the false discovery rate, the two-stage linear step-up procedure was used (Benjamini et al., 2006). An I^2 index of Q-statistics tested messages across all topics simultaneously for serial correlation, which was modelled if needed (Huedo-Medina et al., 2006).

4.1.1 Conditions of Use

SDA relies on two primary assumptions and requires a minimum sample size. Like other regressions, SDA assumes a linear combination of explanatory variables (nonlinear aspects can be modelled as nonlinear functions of variables [e.g., age^2] or interactions among variables [*anecdote* x *ask about use*].) SDA also requires independent residuals (no serial correlation as discussed above). In addition, SDA has modest sample size requirements. Green (1991) proposed the following heuristic sample size, *N*, for a multiple regression with *M* explanatory variables and an expected explained variance R^2 of the outcome variable:

$$N > (\{8 \times [(1 - R^2) / R^2]\} + M) - 1 \qquad (3)$$

For a large model of 20 explanatory variables with a small expected R^2 of 0.10, the required sample size is 91 messages: = 8 × (1 − 0.10) / 0.10 + 20 − 1. Less data are needed for a larger expected R^2 or smaller models. Note that statistical power must be computed at each level of analysis (message, topic, class, school … country). With 1,330 messages, statistical power exceeded 0.95 for an effect size of 0.1 at the message level. The sample sizes at the topic level (13) and the individual level (17) were very small, so any results at these levels must be interpreted cautiously.

5   RESULTS

5.1 Summary Statistics

In this study, seventeen students wrote 1,330 messages on 13 domain-based topics (e.g., history of computer-mediated communication [CMC], different CMC environments), organized into folders in the forum. The length of messages was not normalized. Students who posted more messages on average than other students had the following profile: older; enrolled in Master of Arts (MA) programs; part-time students; not teachers; worked in technology fields; or had KF experience (*older*: m = 47 vs. other m = 37 messages; *MA*: 64 vs. 36; *part-time*: 47 vs. 27; *not teachers*: 55 vs. 36; *technology*: 54 vs. 39; *KF*: 44 vs. 32). Students posted few messages with the following attributes (see Table 4, panel B): new information (1%), theory (4%), opinion (5%), elaboration (2%), anecdotal evidence (1%), ask for explanation (9%), ask about use (2%), different opinion (1%). Eight-three percent of the messages had

none of the above attributes. (As some messages included more than one of these attributes, these percentages do not sum up to 100.)

Table 4: Summary statistics at the individual level (panel A) and message level (panel B)

A. Individual Variable (N = 17)	Mean	Description
Man	0.28	28% of participants were men; 72% were women.
Young (under 35 years of age)	0.50	Half of the participants were under 35 years of age.
Doctorate	0.22	22% were enrolled in a Ph.D. or an Ed.D. program.
Master of Arts	0.22	22% were enrolled in an M.A. program.
Master of Education	0.50	50% were enrolled in an M.Ed. program.
Part-time Student	0.78	78% were part-time students; 22% were full-time.
Teacher	0.67	67% worked as teachers.
Post-Secondary Teacher	0.28	28% taught at the post-secondary level.
Technology	0.22	22% worked in the technology industry.
Knowledge Forum (KF)	0.83	83% had used KF previously.
Past Online Courses	2.89	Participants had taken an average of 2.89 online courses. SD = 2.74; Min = 0; Max = 8.

B. Message Variable (N=1330)	Mean	Description
Man	0.26	Men posted 26% of all messages; women posted 74%.
Young (under 35)	0.44	Young participants posted 44% of all messages.
Doctorate	0.20	Ph.D. students posted 20% of all messages.
Master of Arts	0.33	M.A. students posted 33% of all messages.
Master of Education	0.47	M.Ed. students posted 47% of all messages.
Part-time Student	0.86	Part-time students posted 86% of all messages.
Teacher	0.57	Teachers posted 57% of all messages.
Post-Secondary Teacher	0.23	Post-secondary teachers posted 23% of all messages.
Technology	0.28	Those working in technology posted 28% of all messages.
Knowledge Forum (KF)	0.87	Those who used KF before posted 87% of all messages.
Past online courses	3.35	SD = 2.21; Min = 0; Max = 8. The average number of author's online courses, weighted by number of messages.
New information	0.01	1% of the messages had at least one new piece of information.
Theorize	0.04	4% of the messages had theorizing.

Opinion	0.05	5% of the messages gave a new opinion.
Elaboration	0.02	2% of the messages had an elaboration of another's idea.
Anecdotal evidence	0.01	1% of the messages gave evidence to support an idea.
Ask for explanation	0.09	9% of the messages had a request for explanation.
Ask about use	0.02	2% of the messages had a request for a use.
Different opinion	0.01	1% of the messages had a different opinion than others.
Any of the above processes	0.17	17% of the messages had at least one of the above features.

Note: Except for past online courses, all variables have possible values of 0 or 1.

5.2 Explanatory Model

The two-level variance component analysis showed no significant variance at the second level (topic), so we proceeded with a single-level analysis. All results discussed below describe first entry into the regression, controlling for all previously included variables. Ancillary regressions and statistical tests are available upon request.

5.2.1 New Information

The attributes of previous messages were linked to new information in the current message. After an opinion, new information was 7 percent more likely in the next message. After a question about use three messages earlier, new information was 10 percent more likely. Together, these explanatory variables accounted for about 26 percent of the variance of new information (see Figure 4 and Table 5).

5.2.2 Theorize

Gender and attributes of previous messages were significantly linked to theorizing. Men were 22 percent more likely than women were to theorize. Demographics accounted for 5 percent of the variance in theorizing. Attributes of earlier messages (up to three messages earlier) were linked to theorizing. After an explanation or an elaboration, theorizing was 21 percent or 39 percent more likely, respectively. If someone asked about the use of an idea, gave an opinion, or gave a different opinion two messages earlier, theorizing was 21 percent, 54 percent, or 12 percent more likely, respectively. After anecdotal evidence three messages earlier, theorizing was 34 percent more likely. Altogether, these explanatory variables accounted for 38 percent of the variance of theorizing.
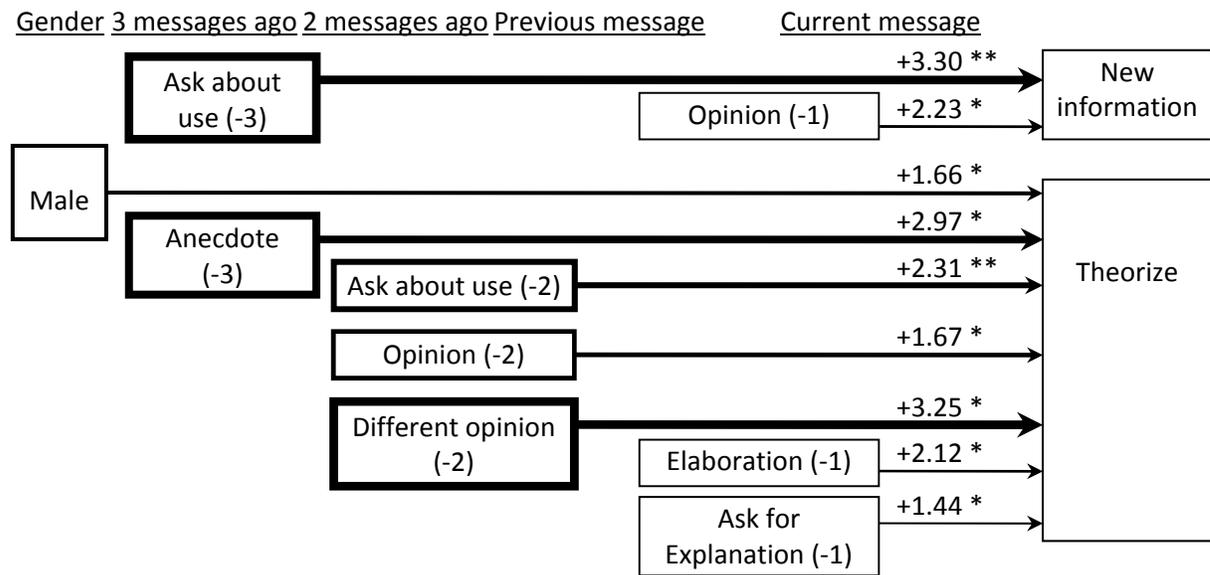
Other variables were not significant. As the I^2 index of Q-statistics for each dependent variable was not significant, serial correlation of errors was unlikely.

Table 5: Statistical discourse analysis results modelling New information and Theorize

Explanatory variable	Model 1		Model 2		Model 3		Model 4	
			New information					
Opinion (-1)			2.565	*	2.565	*	1.618	
			(1.025)		(1.025)		(1.066)	
Ask about Use (-3)							3.186	**
							(1.070)	
Explained variance	0.000		0.119		0.119		0.257	
			Theorizing					
Male	1.330	**	1.362	*	1.615	*	1.588	*
	(0.493)		(0.536)		(0.709)		(0.702)	
Ask for explanation (-1)			1.642	**	1.463		1.427	
			(0.586)		(0.767)		(0.757)	
Elaboration (-1)			2.366	**	2.037	*	1.709	
			(0.830)		(1.025)		(1.034)	
Purpose (-2)					2.059	*	2.327	**
					(0.869)		(0.868)	
Different opinion (-2)					3.567	**	3.164	*
					(1.245)		(1.249)	
Opinion (-2)					1.483	*	1.742	*
					(0.712)		(0.705)	
Evidence (-3)							2.890	*
							(1.133)	
Explained variance	0.051		0.148		0.314		0.367	

Note: Each regression model included a constant term.

$*p < .05, **p < .01$

Gender 3 messages ago 2 messages ago Previous message   Current message

Ask about use (-3) → +3.30 ** → New information

Opinion (-1) +2.23 * → New information

Male +1.66 * → Theorize

Anecdote (-3) +2.97 * → Theorize

Ask about use (-2) +2.31 ** → Theorize

Opinion (-2) +1.67 * → Theorize

Different opinion (-2) +3.25 * → Theorize

Elaboration (-1) +2.12 * → Theorize

Ask for Explanation (-1) +1.44 * → Theorize

Figure 4: Path diagram for *New information* and *Theorize*. Thicker lines indicate stronger links
***p* < .05, ***p* < .01**

6 DISCUSSION

During asynchronous online discussions, students have more time to gather information, contemplate ideas, and evaluate claims, so they might display higher levels of learning than during face-to-face discussions (Hara et al., 2000; Luppicini, 2007; Tallent-Runnels et al., 2006). Extending this research beyond aggregate attributes of *separate messages,* this study examined the *relationships among messages* with statistical discourse analysis. Both individual characteristics and the micro-sequence context of recent messages' cognition and social metacognition affected the likelihoods of subsequent new information and theorizing. This study suggests how statistical discourse analysis might be fruitfully applied to large data sets (e.g., massive online open courses, MOOCs) as a vital learning analytics tool.

6.1 Gender

Past studies of primary and secondary school students had shown that individual differences in gender accounted for little of the variance in discussion behaviours (e.g., Chen et al., 2012), but this study showed that these men were more likely than these women to theorize. Future studies with larger samples can test the generality of this result.

6.2 Micro-sequence Context of Recent Messages

Beyond the effects of individual characteristics, both cognitive and social metacognitive aspects of recent messages showed micro-sequence context effects on subsequent messages. These results showed that asynchronous messages are more than simply lists of individual cognition (Thomas, 2002); instead, these messages influence and respond to one another.

Informal cognition (opinions, elaborations, anecdotes) often preceded formal cognition (new information, theorizing). After a message containing an opinion, messages containing new information and theorizing were more likely to follow. Anecdotes and elaborations were also more likely to be followed by theorizing. Together, these results are consistent with the view that familiar, informal cognition is often activated before more formal cognition (Chiu, 1996) and that the former can facilitate the latter through spreading activation of related semantic networks both in the individual and among group members (Nijstad et al., 2003). This order of informal cognition before formal cognition also reflects the social nature of knowledge building discourse; individuals share their informal experiences, which group members consider, reshape, and integrate into formal, public, structured knowledge. For educators, these results suggest that students often share their ideas informally, and teachers should encourage students to use one another's ideas to create formal knowledge.

Social metacognition, in the form of questions and different opinions, also affected the likelihoods of new information and theorizing. Reflecting students' knowledge interests, their questions identify key goals and motivate knowledge building. Questions asking about the use of a particular idea had the largest effect on inducing more new information, showing their power to influence other's behaviours, which is consistent with Bereiter and Scardamalia's (2006) conceptions of "design mode" teaching and earlier research (e.g., Chen et al., 2012). Furthermore, both types of questions elicited more theorizing, which is also consistent with earlier studies (e.g., Lu et al., 2011). These results suggest that educators can design instruction to give students autonomy or "collective cognitive responsibility" (Scardamalia, 2002; Zhang, Scardamalia, Reeve, & Messina, 2009) so that students can create their own learning goals (or at least sub-goals) and ask questions to motivate themselves and their classmates to build knowledge that is meaningful to them. Lastly, a different opinion had the largest effect on a subsequent theory, consistent with past disequilibrium research showing that disagreements provoke explanations (e.g., Chiu & Khoo, 2003). Together, these results suggest useful prompts that a teacher might encourage students to use during online discussions, for example through brief cue cards or direct teacher questioning.

6.3 Statistical Discourse Analysis

This study showcases a methodology for analyzing relationships among individual characteristics and non-linear asynchronous messages during an online discussion. Such analyses must address analytic difficulties involving the data, the dependent variables, and the explanatory variables. Data issues include missing data, nested data, and the tree structure of online messages. Difficulties involving dependent variables include discrete outcomes, infrequent outcomes, similar adjacent messages, and multiple outcomes. Lastly, explanatory variable issues include sequences, indirect effects, false positives, and robustness of results.

SDA addresses each of these analytic difficulties as follows (see Table 3). First, SDA addresses the data issues (missing data, nested data, and tree structure of online messages) with Markov Chain Monte Carlo multiple imputation (MCMC-MI), multilevel analysis, and identification of the previous message. Second, SDA addresses the dependent variable difficulties (discrete, infrequent, serial correlation and multiple) with Logit regressions, a Logit bias estimator, I^2 index of Q-statistics, and multivariate outcome analyses. Lastly, SDA addresses the explanatory variable issues (sequences, indirect effects, false positives, robustness) with vector auto-regression, multilevel M-tests, the two-stage linear step-up procedure, and robustness tests.

Furthermore, the nature of this data set offers the possibility of automated statistical analyses. To create this data set, participants selected scaffolds to support composing of their messages, which essentially provides user-perspective codes for each message at the time that they composed them (Fujita, 2013). Whether a scaffold is used in the same way by different users or by the same user over time requires further study. If there is sufficient inter-coder reliability for these scaffolds however, computer programs can automatically compute both summary statistics and run statistical analyses to test explanatory models on these data. Specifically, SDA can be encoded into a computer program and researchers (or teachers or students) can choose among the possible variables to test whether they account for differences in a set of outcome variables. Such analyses can identify both the degree to which the explanatory model fits the data and the specific instances unexplained by the model. Such analyses of large data sets, such as massive open online courses (MOOCs), can help researchers build appropriate theoretical models of discussions and help educators decide which students, or groups of students, are struggling and require assistance.

6.4 Limitations

This study's analytic categories and data might sharply limit the utility of its results for other students, groups, activities, and contexts. These mostly dichotomous analytic categories are a first step toward a more comprehensive set of categories. Furthermore, the sample sizes of students (17) and courses (1)

can be expanded in future research. Future research can also model actual time and students' reading behaviours in addition to their writing behaviours.

7   CONCLUSION

This study extends the online discussion research beyond aggregated attributes of *separate messages* to *relationships among messages* by showcasing how statistical discourse analysis can model these relationships. The results showed that both individual characteristics and the micro-sequence context of recent messages' cognition and social metacognition affected the likelihoods of subsequent new information and theorizing. Unlike past studies of students, this exploratory study with a few students suggests that gender in adults might account for substantial differences in online behaviours. Specifically, men were more likely than women were to theorize. Rather than simply being lists of individual cognition, asynchronous messages create a micro-sequence context that affects subsequent messages. Informal cognition (opinions, anecdotes, elaborations) facilitates more formal cognition (new information and theoretical explanations). Meanwhile, social metacognition, in the form of questions and different opinions, had the strongest effects on subsequent new information and theoretical explanations.

ACKNOWLEDGEMENTS

REFERENCES

Allen, I. E., & Seaman, J. (2010). *Class differences: Online education in the United States, 2010.* Needham, MA: Sloan Consortium.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93*, 491–507.

Bereiter, C. (1994). Implications of postmodernism for science, or science as progressive discourse. *Educational Psychologist, 29*(1), 3–12.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bereiter, C., & Scardamalia, M. (2006). Education for the knowledge age: Design-centered models of teaching and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology*, 2nd ed. (pp. 695–713). Mahwah, NJ: Lawrence Erlbaum.

Chen, G., Chiu, M. M., & Wang, Z. (2012). Social metacognition and the creation of correct, new ideas: A statistical discourse analysis of online mathematics discussions. *Computers in Human Behavior, 28*(3), 868–880.

Chiu, M. M. (1996). Exploring the origins, uses and interactions of student intuitions: Comparing the lengths of paths. *Journal for Research in Mathematics Education, 27*(4), 478–504.

Chiu, M. M. (2000). Group problem solving processes: Social interactions and individual actions. *Journal for the Theory of Social Behavior, 30*(1), 27–50.

Chiu, M. M. (2001). Analyzing group work processes: Towards a conceptual framework and systematic statistical analyses. In F. Columbus (Ed.), *Advances in psychology research,* vol. 4 (pp. 193–222). Huntington, NY: Nova Science.

Chiu, M. M. (2008a). Effects of argumentation on group micro-creativity: Statistical discourse analyses of algebra students' collaborative problem solving. *Contemporary Educational Psychology, 33*, 382–402.

Chiu, M. M. (2008b). Flowing toward correct contributions during group problem solving: A statistical discourse analysis. *Journal of the Learning Sciences, 17*(3), 415–463.

Chiu, M. M., & Khoo, L. (2003). Rudeness and status effects during group problem solving: Do they bias evaluations and reduce the likelihood of correct solutions? *Journal of Educational Psychology, 95*, 506–523.

Chiu, M. M., & Khoo, L. (2005). A new method for analyzing sequential processes: Dynamic multi-level analysis. *Small Group Research, 36*, 1–32.

Chiu, M. M., & Kuo, S. W. (2009). From metacognition to social metacognition: Similarities, differences, and learning. *Journal of Education Research, 3*(4), 1–19.

Clark, D. B., & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. *Journal of Research in Science Teaching, 45*(3), 293–321.

de Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning, 2*, 87–103.

De Liddo, A., Buckingham Shum, S., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics. First International Conference on Learning Analytics and Knowledge, Banff, AB, 27 February–1 March 2011. Retrieved from http://oro.open.ac.uk/25829/1/DeLiddo-LAK2011.pdf

Fujita, N. (2009). *Group processes supporting the development of progressive discourse in online graduate courses.* Unpublished Doctoral Dissertation, University of Toronto, Toronto, ON. Retrieved from http://hdl.handle.net/1807/43778

Fujita, N. (2013). Critical reflections on multivocal analysis and implications for design-based research. In D. Suthers, K. Lund, C. Rose, C. Teplovs, & N. Law (Eds.). Productive multivocality in the analysis of group interactions (pp. 435-455). New York: Springer.

Glassner, A., Weinstoc, M., & Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *British Journal of Educational Psychology, 75*, 105–118.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499–510.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431.

Hacker, D. J., & Bol, L. (2004). Metacognitive theory. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited,* vol. 4 (pp. 275–297). Greenwich, CT: Information Age.

Hakkarainen, K. (2003). Emergence of progressive-inquiry culture in computer-supported collaborative learning. *Learning Environments Research, 6*(2), 199–220.

Hara, N., Bonk, C. J., & Angeli, C. (2000). Content analysis of online discussion in an applied educational psychology course. *Instructional Science, 28*, 115–152.

Haythornthwaite, C. (2001). Exploring multiplexity: Social network structure in a computer-supported distance learning class. *The Information Society, 18*(5), 385–401.

Haythornthwaite, C. , de Laat, M. & Dawson, S. (2013). Introduction to the special issue on learning analytics. *American Behavioral Scientist, 57*(10), 1371–1379.

Howe, C. (2009). Collaborative group work in middle childhood. *Human Development, 52*(4), 215–239.

Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis. *Psychological Methods, 11*, 193–206.

Kanuka, H., Rourke, L., & Laflamme, E. (2007). The influence of instructional methods on the quality of online discussion. *British Journal of Educational Technology, 38*(2), 260–271.

Kennedy, P. (2008). *Guide to econometrics*. Cambridge: Wiley-Blackwell.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*, 137–163.

LAK (2011). First International Conference on Learning Analytics and Knowledge, Banff, AB, 27 February –1 March 2011. Retrieved from https://tekri.athabascau.ca/analytics/

Lee, E., Chan, C., & van Aalst, J. (2006). Students assessing their own collaborative knowledge building. *International Journal of Computer-Supported Collaborative Learning, 1*(1), 57–87.

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment. *Journal of Research in Science Teaching, 36*, 837–858.

Lu, J., Chiu, M., & Law, N. (2011). Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior, 27*, 946–955.

Luppicini, R. (2007). Review of computer mediated communication research for education. *Instructional Science, 35*(2), 141–185.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128.

Nijstad, B. A., Diehl, M., & Stroebe, W. (2003). Cognitive stimulation and interference in idea generating groups. In P. B. Paulus & B. A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 137–159). New York: Oxford University Press.

Palloff, R. M., & Pratt, K. (2007). *Building online learning communities: Effective strategies for the virtual classroom*. San Francisco, CA: Jossey-Bass.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research. *Review of Educational Research, 74*, 525–556.

Reimann, P. (2009). Time is precious: Variable and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 4*(3), 239–257.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (1999). Assessing social presence in asynchronous text-based computer conferencing. *Journal of Distance Education, 14*(2), 50–71.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Chicago: Open Court.

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

Siemens (2011). *First International Conference on Learning Analytics and Knowledge Call for Papers*. Retrieved from https://tekri.athabascau.ca/analytics/call-papers

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380–1400.

Snijders, T. A. B., & Bosker, R. (2012). *Multilevel analysis*. London: Sage.

Tallent-Runnels, M. K., Thomas, J. A., Lan, W. Y., Cooper, S., Ahern, T. C., Shaw, S. M., & Liu, X. (2006). Teaching courses online. *Review of Educational Research, 76*(1), 93–135.

Teplovs, C. (2013, February 22). *Analytics for Knowledge Creation (A4KC) Part II: Zoolib tuplebase queries*. Retrieved from http:// http://christeplovs.com/2013/02/22/a4kc1/.

Teplovs, C. & Fujita, N. (2013). Socio-dynamic Latent Semantic Learner Models. In D. Suthers, K. Lund, C. Rosé, C. Teplovs & N. Law, (Eds.), *Productive multivocality in the analysis of collaborative learning* (pp. 382–396). New York: Springer.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435–502.

Thomas, M. J. W. (2002). Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning, 18*, 351–366.

Woodruff, E., & Brett, C. (1999). Collaborative knowledge building: Preservice teachers and elementary students talking to learn. *Language and Education, 13*(4), 280–302.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge building communities. *Journal of the Learning Sciences, 18*(1), 7–44.