2014

Statistical Discourse Analysis of Online Discussion: Informal cognition, social metacognition, and knowledge creation

M. M. Chiu

N. Fujita
University of Windsor, nfujita@uwindsor.ca

Statistical Discourse Analysis of Online Discussions: Informal Cognition, Social Metacognition and Knowledge Creation

Ming Ming Chiu
University at Buffalo,
State University of New York
564 Baldy Hall
Buffalo, NY, 14260-1000, USA
mingchiu@buffalo.edu

Nobuko Fujita
University of Windsor
401 Sunset Avenue
Windsor, ON, N9B 3P4, Canada
nfujita@uwindsor.ca

ABSTRACT
To statistically model large data sets of knowledge processes during asynchronous, online forums, we must address analytic difficulties involving the whole data set (missing data, nested data and the tree structure of online messages), dependent variables (multiple, infrequent, discrete outcomes and similar adjacent messages), and explanatory variables (sequences, indirect effects, false positives, and robustness). Statistical discourse analysis (SDA) addresses all of these issues, as shown in an analysis of 1,330 asynchronous messages written and self-coded by 17 students during a 13-week online educational technology course. The results showed how attributes at multiple levels (individual and message) affected knowledge creation processes. Men were more likely than women to theorize. Asynchronous messages created a micro-sequence context; *opinions* and *asking about purpose* preceded *new information*; *anecdotes*, *opinions*, *different opinions*, *elaborating* ideas, and *asking about purpose or information* preceded *theorizing*. These results show how informal thinking precedes formal thinking and how social metacognition affects knowledge creation.

Categories and Subject Descriptors
Knowledge.3.1 [**Computer Uses in Education**] Collaborative learning

General Terms
Human Factors.

Key Words. Statistical discourse analysis, informal cognition, social metacognition, knowledge creation

1. INTRODUCTION
The benefits of online discussions have increased both their uses and records of their uses, which allow detailed analyses to inform design and to improve their productivity. Unlike face-to-face talk, students on asynchronous, online forums can participate at different places and times and have more time to gather information, contemplate ideas, and evaluate claims before responding, resulting in superior decision-making, problem solving, writing and *knowledge creation* (KC, [35][42][19]). The nascent field of learning analytics seeks to understand and optimize learning and the online learning environment in which it occurs [1]. Using online forum data, studies using aggregate counts show how specific actions (e.g., "why" or "how" questions, explanations, evidence, summaries) are related to KC [31][32][45].

While aggregate counts provide descriptive summaries, they do not fully utilize the information relating to the time and order of collaboration and learning processes [39], or capture the sequential data needed to test KC hypotheses about how group members' actions/posts/messages are related to one another [12].

In contrast, discourse-centric learning analytics go beyond surface measures to investigate the quality of the learning process, specifically the rhetorical dimensions, to improve discourse for deeper learning and learning design [17]. In a similar vein, analyses of sequences of messages can illuminate the relationships among processes that contribute to knowledge creation 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 types of target messages (e.g., *theorizing*). These results can help us understand the temporal and causal relationships among different types of messages or message sequences that aid or hinder knowledge creation. We show how statistical discourse analysis (SDA, [13]) can model these sequences to test these KC hypotheses. To explicate SDA, we introduce data [18] 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*) increase the likelihoods of *new information* or *theoretical explanations* in subsequent messages. This example shows how SDA might be fruitfully applied to large datasets (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-

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA
ACM. ACM 978-1-4503-2664-3/14/03...$15.00.
http://dx.doi.org/10.1145/2567574.2567580

based research study [18]. 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 [3][4].

2.1 Participants

Participants were 17 students (12 females, 5 males) (see Table 4). They ranged in age from mid-20s to mid-40s. Five were students in academic programs (4 M.A., 1 Ph.D.); 12 were students in professional programs (9 M.Ed., 3 Ed.D.).

2.2 Procedure

Students were encouraged to engage in progressive discourse through three interventions : a reading by Bereiter [4], 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 [46] to help elementary school teachers and preservice teachers improve their face-to-face collaborative discussion. 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 [4]; *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 refer to in this paper as "messages" (see Figure 1).

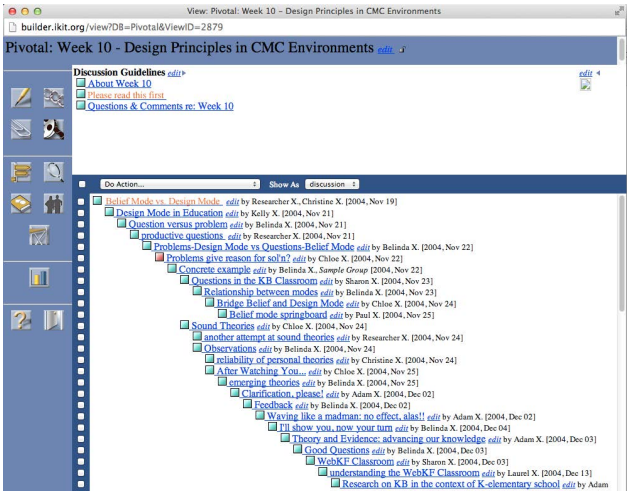

Figure 1. KF view showing thread structure of messages.

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 (see Figure 2).

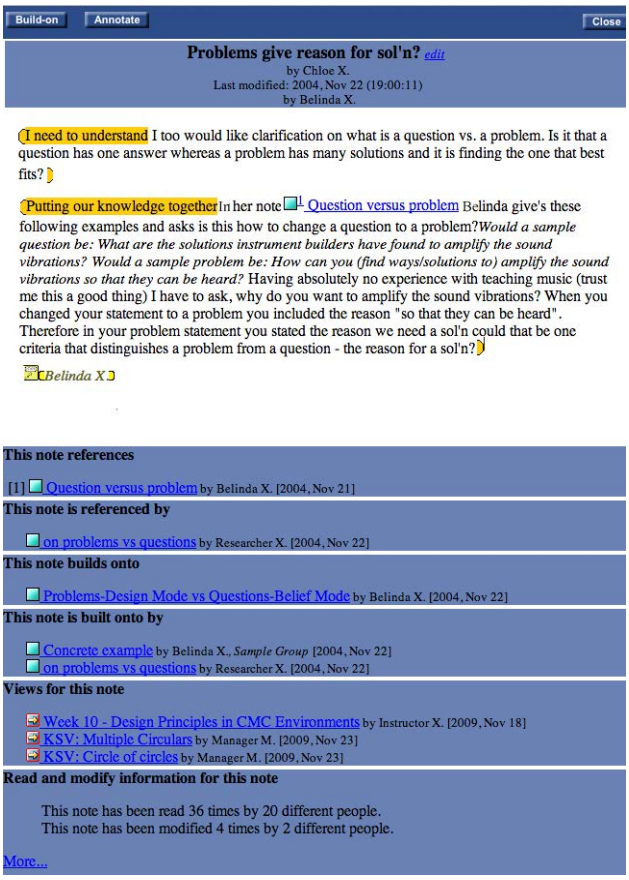

Figure 2. KF Message with scaffold supports, link, annotation, and other information.

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 in the 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 the student designers of the scaffolds to emphasize the socio-cognitive dynamics of "improvable ideas," one of the twelve knowledge building principles [40] 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.

Table 1. Knowledge Forum Scaffolds and Scaffold Supports Used in Iteration 2

Scaffolds		
Cognition	**Social Metacognition**	**Dependent variables**
Opinion (I think knowledge building takes a long time.)	*Ask for explanation* (I need to understand why knowledge building has to take a long time.)	*Theorize/Explain* (My theory of the time needed for knowledge building is based on its sequence of parts …)
Elaboration (I think knowledge building takes a lot of smaller steps.)	*Ask about use* (Why do we need to understand how much time knowledge building takes?)	*New information* (Scardamalia and Bereiter's [39] study showed that computer supports can support knowledge building in classroom learning communities.)
Anecdotal evidence (Last week, our class took over an hour to come up with a good theory.)	*Different opinion* (I don't think knowledge building has to take a long time. It might depend on the people.)	

3. HYPOTHESIS

We test whether recent cognition or social metacognition facilitate new information or theoretical explanations [10][34]. 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 knowledge creation. As students propose integrative theories that explain more facts, they create knowledge through a process of explanatory coherence [43]. Hence, new information and theoretical explanations are suitable target processes to serve as dependent variables in our statistical model.

Researchers have shown that many online discussions begin with sharing of opinions [23]. Students often activate familiar, informal concepts before less familiar, formal concepts [9]. During a discussion, comments by one student (e.g., a key word) might spark another student to activate related concepts in his or her semantic network and propose a new idea [37]. When students do not clearly understand these ideas, they can ask questions to elicit new information, elaborations or explanations [25]. Also, students may disagree (different opinions) and address their differences by introducing evidence or explaining their ideas [27]. Whereas individual metacognition is monitoring and regulating one's own knowledge, emotions, and actions [24], *social metacognition* is defined as group members' monitoring and controlling one another's knowledge, emotions, and actions [16]. 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) increase the likelihoods of new information or theoretical explanations in subsequent messages. See Table 2. To reduce omitted variable bias, additional individual and time explanatory variables were added. For example, earlier studies showed that males were more likely than females to make claims, argue, elaborate, explain, and critique others [34].

Table 2. Hypotheses regarding the effects of classroom problem solving processes on the outcome variables *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 parentheses indicate expected relationship with the outcome variables: positive and supported [+], hypothesized but not supported [ns +]).

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 [38].

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 [36]
• Nested data (Messages within Topics)	• Multilevel analysis (Hierarchical linear modeling [5][18])
• Tree structure of messages (Λ)	• Store preceding message to capture tree structure
Dependent variables	
• Discrete variable (yes/no)	• Logit / Probit

• Infrequent variable	• Logit bias estimator [28]
• Similar adjacent messages ($m_3 \sim m_4$)	• I^2 index of Q-statistics [26]
• Multiple dependent variables (Y_1, $Y_{2, ...}$)	• Multivariate outcome models [18]
Explanatory variables	
• Sequences of messages (X_{t-2} or $X_{t-1} \rightarrow Y_t$)	• Vector Auto-Regression (VAR, [27])
• Indirect, multi-level mediation effects ($X \rightarrow M \rightarrow Y$)	• Multilevel M-tests [34]
• False positives (Type I errors)	• Two-stage linear step-up procedure [1]
• 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 (vs. different topic folders) can underestimate the standard errors [20]. To address this issue, SDA models nested data with a multilevel analysis [20][6].

Unlike a linear, face-to-face conversation in which one turn of talk always follows the one before it, an asynchronous message in an online forum might follow 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).

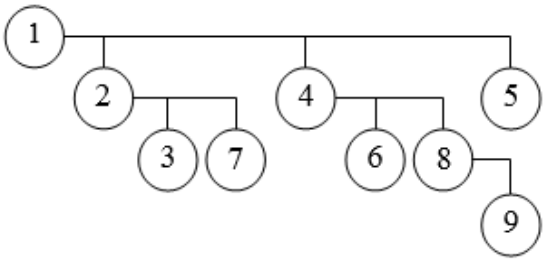

Figure 3. Tree structure showing how nine messages are related to one another.

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 followed message #4 (not #5). To capture the tree structure of the messages, we identify the immediate predecessor of each message. Then, we can reconstruct the written reply structure of the entire tree to identify any ordinal predecessor of any message. Patterns of reading behavior may be irregular across threads and are thus more difficult to trace.

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 occurs in a conversation or it does not) rather than continuous (e.g., test scores), so standard regressions such as ordinary least squares can bias the standard errors. To model discrete dependent variables, we use a Logit regression [29]. As infrequent dependent variables can bias the results of a Logit regression [30], we estimate the Logit bias and remove it [30].

As adjacent messages are often more closely related to one another more than messages that are far apart, failure to model this similarity (serial correlation of errors) can bias the results [29]. An I^2 index of Q-statistics tested all topics simultaneously for serial correlation of residuals in adjacent messages [28]. 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 modeling the outcome variable *theory*, add whether it occurs in the previous message [*theory* (–1)] [15]; see paragraph below on vector auto-regression.

Multiple outcomes (*new information*, *theorizing*) can have correlated residuals that can underestimate standard errors [20]. 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 [20].

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, [29]) 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 [36]. 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 [2].

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 estimated data).

4.1 Analysis Procedure

After MCMC-MI of the missing data (less than 1%) 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 modeled two process variables in students' messages (*New Information* and *Theorizing*) with SDA [11].

$$\mathbf{Process}_{ymt} = \beta_y + \mathbf{e}_{ymt} + \mathbf{f}_{yt} \quad (1)$$

For $\mathbf{Process}_{ymt}$ (the 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 $\mathbf{e}_{mt}$ and $\mathbf{f}_t$ respectively. As analyzing rare events (target processes occurred in less than 10% of all messages) with Logit/Probit regressions can bias regression coefficient estimates, King and Zeng's [30] bias estimator was used to adjust them.

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, [29]).

$$\mathbf{Process}_{ymt} = \beta_y + \mathbf{e}_{ymt} + \mathbf{f}_{yt} + \beta_{ydt}\mathbf{Demographics}_{ymt} + \beta_{yst}\mathbf{Schooling}_{ymt} + \beta_{yjt}\mathbf{Job}_{ymt} + \beta_{yxt}\mathbf{Experience}_{ymt} + \beta_{ypt}\mathbf{Previous}_{ymt} \quad (2)$$

Next, schooling variables were entered: *doctoral* student, *Masters of Education* student, *Masters of Arts* student, and *part-time* student (**Schooling**). Then, students' job variables were entered: *teacher*, *post-secondary teacher*, and *technology* (**Job**). Next, 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) (**Previous**). The attributes of the message two responses ago along the same thread (-2) were entered, then, those of the message three responses ago along the same thread (-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 **Previous** message (e.g., evidence; $\beta_{ypt} = \beta_{yt} + f_{yj}$) differed significantly ($f_{yj} \neq 0$?), then a moderation effect might exist, and their interactions with processes were included.

The multilevel M-test [36] 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) was reported as the increase or decrease (+TE% or –TE%) in the outcome variable [29]. 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 [2]. An I^2 index of Q-statistics tested messages across all topics simultaneously for serial correlation, which was modeled if needed [21][28][33].

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 modeled 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 [22] 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:

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

For a large model of 20 explanatory variables with a small expected R^2 of 0.10, the required sample size is 91 messages: $= 8 \times (1 - 0.10) / 0.10 + 20 - 1$. 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 units must be interpreted cautiously.

5. RESULTS

5.1 Summary Statistics

In this study, seventeen students wrote 1,330 messages on 13 domain-based, not procedural, topics (e.g., History of CMC, Different CMC Environments), organized into folders in the forum. Length of messages were not normalized. Students who posted more messages on average than other students had the following profile: older; enrolled in Masters 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%), and none of the above (83%). (As some messages included more than one of these attributes, these percentages do not sum up to 100%.)

5.2 Explanatory Model

As none of the second level (topic) variance components were significant, a single-level analysis was sufficient. 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% more likely in the next message. After a question about use three messages before, new information was 10% more likely. Together, these explanatory variables accounted for about 26% of the variance of new information. See Figure 4.

5.2.2 Theorize

Gender and attributes of previous messages were significantly linked to theorizing. Men were 22% more likely than women to theorize. Demographics accounted for 5% of the variance in theorizing.

Attributes of earlier messages up to three messages before were linked to theorizing. After an explanation or an elaboration, theorizing was 21% or 39% more likely, respectively. If someone asked about the use of an idea, gave an opinion or gave a different opinion two messages before, theorizing was 21%, 54%, or 12% more likely, respectively. After anecdotal

evidence three messages before, theorizing was 34% more likely. Altogether, these explanatory variables accounted for 38% 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 was unlikely.

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 PhD or an EdD program.
Masters of Art	0.22	22% were enrolled in MA program.
Masters of Education	0.50	50% were enrolled in MEd 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	PhD students posted 20% of all messages.
Masters of Art	0.33	MA students posted 33% of all messages.
Masters of Education	0.47	MEd 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 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. The other 83% of messages shared personal experiences and unsubstantiated opinions rather than engaging in progressive knowledge creation.

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

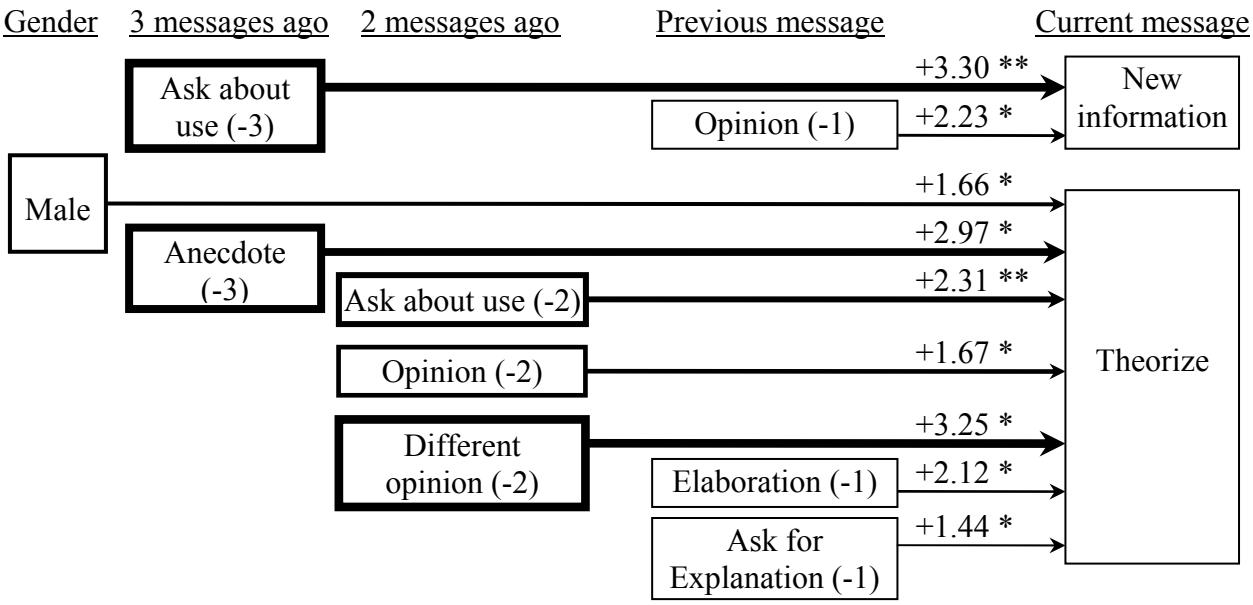

Figure 4. **Path diagram for *New information and Theorize*. Thicker lines indicate stronger links. *$p < .05$, **$p < .01$, ***$p < .001$.**

6. DISCUSSION

During asynchronous, online discussions, students have more time to gather information, contemplate ideas, and evaluate claims, so they often display higher levels of knowledge creation than during face-to-face discussions [26][35][42]. 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 statistical discourse analysis might be fruitfully applied to large datasets (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 behaviors [8], 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 [44]; 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 views that familiar, informal cognition is often activated before more formal cognition [9] and that the former can facilitate the latter through spreading activation of related semantic networks both in the individual and among group members [37]. 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 use of a particular idea had the largest effect on inducing more new information, showing their power to influence other's behaviors, which is consistent with Bereiter and Scardamalia's [5] conceptions of "design mode" teaching and earlier research (e.g., [8]). Furthermore, both types of questions elicited more theorizing, which is also consistent with earlier studies (e.g. [34]). These results suggest that educators can design instruction to give students autonomy or "collective cognitive responsibility" [40][47] so that students can create their own learning goals (or at least subgoals) 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 dis-equilibrium research showing that disagreements provoke explanations (e.g., [14]). 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

As the large data set includes participant-coding of their messages, SDA offers the potential for semi-automatic analyses that integrates multiple analyses encoded into computer programs on large data sets such as the online discussions of massive open online courses (MOOCs). If participant coding yields sufficiently similar categories of codes (an open and a valuable research area), the codes can be entered into SDA-

encoded computer programs, and users can test explanatory models.

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. First, data issues include missing data, nested data and the tree structure of online messages. Second, 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, 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.

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 behaviors in addition to their writing behaviors.

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 behaviors. Specifically, men were more likely than women 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.

8. REFERENCES

[1] 1st International Conference on Learning Analytics and Knowledge, Banff, AB, February 27-March 1, 2011, https://tekri.athabascau.ca/analytics/.

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

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

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

[5] 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.

[6] Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. London: Sage.

[7] Chen, G., & Chiu, M. M. (2008). Online discussion processes. *Computers & Education, 50*(3), 678-692.

[8] Chen, G., Chiu, M. M., & Wang, Z. (in press). Social metacognition and the creation of correct, new ideas: A statistical discourse analysis of online mathematics discussions. *Computers in Human Behavior*.

[9] 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.

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

[11] 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, US: Nova Science.

[12] 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.

[13] 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.

[14] 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.

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

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

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

[18] 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

[19] 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.

[20] Goldstein, H. (1995). *Multilevel statistical models*. Sydney: Edward Arnold.

[21] Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel models with applications to repeated measures data. *Statistics in Medicine, 13*, 1643-1655.

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

[23] 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.

[24] 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, Connecticut: Information Age.

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

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

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

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

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

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

[31] 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.

[32] 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.

[33] Ljung, G., & Box, G. (1979). On a measure of lack of fit in time series models. *Biometrika, 66*, 265-270.

[34] 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.

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

[36] 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.

[37] 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.

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

[39] 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*, 239-257.

[40] 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.

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

[42] 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.

[43] Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 1989*(12), 435-502.

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

[45] Wise, A., & Chiu, M. M. (2011). Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning 6,* 445-470.

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

[47] 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.