Georgia State University
Digital Archive @ GSU

Middle-Secondary Education and Instructional Technology Dissertations

Department of Middle-Secondary Education and Instructional Technology

12-9-2004

Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network

Thomas Edward McKlin
tom.mcklin@ceismc.gatech.edu

Follow this and additional works at: http://digitalarchive.gsu.edu/msit_diss

Recommended Citation
McKlin, Thomas Edward, "Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network" (2004). *Middle-Secondary Education and Instructional Technology Dissertations.* Paper 1.

This Dissertation is brought to you for free and open access by the Department of Middle-Secondary Education and Instructional Technology at Digital Archive @ GSU. It has been accepted for inclusion in Middle-Secondary Education and Instructional Technology Dissertations by an authorized administrator of Digital Archive @ GSU. For more information, please contact digitalarchive@gsu.edu.

ACCEPTANCE

This dissertation, ANALYZING COGNITIVE PRESENCE IN ONLINE COURSES USING AN ARTIFICIAL NEURAL NETWORK, by THOMAS E. MCKLIN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Stephen W. Harmon, Ed. D.
Committee Chair

_____
Mary B. Shoffner, Ph.D.
Committee Member

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
William Evans, Ph.D.
Committee Member

_____
Date

_____
Karen A. Schultz, Ph.D.
Chair, Department of Middle-Secondary Education and Instructional Technology

_____
Ronald P. Colarusso, Ed.D.
Dean

AUTHOR’S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Signature of Author

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Thomas E. McKlin
1031 N. Hills Dr.
Decatur, GA 30033

The director of this dissertation is:

Dr. Stephen W. Harmon
Department of Middle-Secondary Education and Instructional Technology
College of Education
Georgia State University
Atlanta, GA 30303-3083

VITA

Thomas E. McKlin

ADDRESS: 1031 N. Hills Dr.
Decatur, Georgia 30033

EDUCATION:

Ph.D.	2004	Georgia State University Instructional Technology
M.S.	1997	Georgia State University Instructional Technology
M.S.	1996	Georgia State University Applied Linguistics
B.A.	1993	Mercer University Literature & Philosophy

PROFESSIONAL EXPERIENCE:

2002-	Georgia Institute of Technology, Atlanta, GA Research Scientist
1997-2002	Georgia Tech Research Institute, Atlanta, GA Research Scientist
1996-1997	Georgia Tech Research Institute, Atlanta, GA Graduate Research Assistant
1995-1996	Georgia State University, Atlanta, GA Graduate Teaching Assistant
1995-1997	Georgia State University, Atlanta, GA Graduate Research Assistant

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2003-Present	American Evaluation Association
2001-Present	Association for Educational Communications and Technology
1997-Present	Association for the Advancement of Computers in Education

PRESENTATIONS AND PUBLICATIONS:

McKlin, T., & Oliver, P. (2003). "The discussion list analyzer: A computational tool for evaluating online discussions." American Evaluation Association Annual Conference, Reno, Nevada.

McKlin, T., & Oliver, P. (2003). "A discussion list analyzer: Using an artificial neural network to reveal cognitive effort expressed in online discussion list messages." Ed-Media 2003, Waikiki, Haiwaii.

McKlin, T., & Oliver, P. (2002). "Cognitive Presence in online discussions: A content analysis of eCore™ courses using a neural network." Ed-Media 2002, Denver, Colorado.

Finnegan, C., Morris, L., McKlin, T., Wu, B., & Xu, H. (2002). "Research on Online Teaching and Learning." Teaching and Learning with Advanced Technologies 2002, Athens, GA.

McKlin, T. (2002). Cognitive effort in on-line learning: An artificial neural network analysis of students' online discussions. Presented and discussed online on IT-Forum.

McKlin, T. Harmon, S., Jones, M., & Evans, W. (2001). Cognitive presence in web-based learning: A content analysis of students' online discussions. AECT 2001, Atlanta, GA.

ABSTRACT

ANALYZING COGNITIVE PRESENCE IN ONLINE COURSES USING AN ARTIFICIAL NEURAL NETWORK
by
THOMAS E. MCKLIN

This work outlines the theoretical underpinnings, method, results, and implications for constructing a discussion list analysis tool that categorizes online, educational discussion list messages into levels of cognitive effort.

Purpose

The purpose of such a tool is to provide evaluative feedback to instructors who facilitate online learning, to researchers studying computer-supported collaborative learning, and to administrators interested in correlating objective measures of students' cognitive effort with other measures of student success. This work connects computer–supported collaborative learning, content analysis, and artificial intelligence.

Method

Broadly, the method employed is a content analysis in which the data from the analysis is modeled using artificial neural network (ANN) software. A group of human coders categorized online discussion list messages, and inter-rater reliability was calculated among them. That reliability figure serves as a measuring stick for determining how well the ANN categorizes the same messages that the group of human coders categorized. Reliability between the ANN model and the group of human coders

is compared to the reliability among the group of human coders to determine how well the ANN performs compared to humans.

Findings

Two experiments were conducted in which artificial neural network (ANN) models were constructed to model the decisions of human coders, and the experiments revealed that the ANN, under noisy, real-life circumstances codes messages with near-human accuracy. From experiment one, the reliability between the ANN model and the group of human coders, using Cohen's kappa, is 0.519 while the human reliability values range from 0.494 to 0.742 (M=0.6). Improvements were made to the human content analysis with the goal of improving the reliability among coders. After these improvements were made, the humans coded messages with a kappa agreement ranging from 0.816 to 0.879 (M=0.848), and the kappa agreement between the ANN model and the group of human coders is 0.70.

ANALYZING COGNITIVE PRESENCE IN ONLINE COURSES USING AN ARTIFICIAL NEURAL NETWORK

by
Thomas E. McKlin

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Instructional Technology
in
the Department of Middle Secondary Education and Instructional Technology
in
the College of Education
Georgia State University

Atlanta, Georgia
2004

Copyright by
Thomas E. McKlin
2004

TABLE OF CONTENTS

Page

List of Tables v

List of Figures vii

Abbreviations ix

Chapter

1 THE PROBLEM 1
A Brief History of Technology Use in Education 1
A Shift in Computer-Mediated Conferencing Research 3
A Shift in Communications Theory 5
Automating Content Analysis 9
Research Question 10
Purpose 10

2 REVIEW OF THE LITERATURE 13
Introduction 13
Can We Learn Online? 13
Definitions and Coding Categories for Cognitive Engagement 17
Units of Analysis 21
Content Analysis as a Guide for Instruction 23
Cognitive Structures 26
Computational Approaches to Content Analysis 28
Uses of Artificial Neural Networks in the Social Sciences 32
Summary 37

3 METHODS 40
Comparing Artificial Neural Networks to Humans 40
Research Method Steps 41
Message preparation 41
Human Content Analysis 41
Artificial Neural Network Content Analysis 42
Message Preparation 43
Transfer Message Text to a Database 44
Select Messages Used for Three Parts of the Analysis 45
Create the Tool Enabling On-line Message Coding 45
Human Content Analysis 46

Participants ... 46
Unit of Analysis ... 48
Operational Definition of Cognitive Presence ... 48
Modified Content Analysis Rubric ... 49
Coder Training ... 50
Reliability ... 51
Artificial Neural Network Content Analysis ... 52
Pilot Study ... 53
How Neural Networks Work ... 54
Numerically Describe Messages ... 57
Predictor Order ... 62
Modeling Content Analysis Decision-Making ... 63
Comparison of Models ... 65
Spelling Analysis ... 66
Comparing Human and ANN Coding Decisions ... 68
Sample ANN Analyses ... 71
Descriptive Analyses ... 71
Mean Cognitive Presence Weights ... 71
Occurrences of Each Cognitive Presence Category ... 72
Analyses Based on Course Section Variables ... 74
Beyond Descriptive Analyses ... 77
Limitations and Bias ... 78
Summary ... 82

4 RESULTS ... 83
Comparing Artificial Neural Networks to Humans ... 83
First Experiment ... 84
Human Content Analysis ... 85
Automatic Content Analysis ... 87
Modifications to the Human Content Analysis ... 90
Second Experiment ... 94
Human Content Analysis ... 94
Automatic Content Analysis ... 96
Comparison of Models ... 99
Spelling Analysis ... 104
Sample ANN Analyses ... 106
Comparison of Means ... 107
Distribution of Messages by Cognitive Presence Category ... 109
Analyses Based on Course Section Variables ... 111
Secondary analyses ... 118
Summary ... 118

5 DISCUSSION AND RECOMMENDATIONS ... 120
Comparing Artificial Neural Networks to Humans ... 120
First Experiment ... 121
The Differences Between ANN- and Human-Coded Messages ... 122
Renegade Coding ... 123

Explanation of the Differences Between ANN- and Human-Coded Messages 126
Second Experiment 132
The Differences Between ANN- and Human-Coded Messages. 133
Renegade Coding 134
Explanation of the Differences Between ANN- and Human-Coded Messages 136
Explaining the Shift in Exploration and Integration Decision Space Between Experiments 138
Question Two: Sample ANN Analyses 141
Limitations and Bias 142
Major Contributions of this Study 145
Suggestions for Future Research 146
Summary 152
References 154
Appendices 164

List of Tables

Table Page

1 Henri's Five Dimensions of the Learning Process 18

2 Cognote Model for Evaluating Online Discussion Group Messages 25

3 Number of Students Contributing Messages in Course 47

4 Interpretation of Kappa Values from Landis and Koch (1997) 52

5 Decision Logic for Aggregating Human-Coded Messages 69

6 Sample Cognitive Presence Values by Student 75

7 Number of Messages by Course Section and Topic 86

8 Agreement and Kappa Scores for Each Coder 87

9 Comparison of Coding Decisions for the First Experiment..... 89

10 Number of Messages by Course Section and Topic 95

11 Agreement and Kappa Values for Each Coder 96

12 Comparison of Coding Decisions for the Second Experiment 98

13 Synopsis of Reliability 99

14 Reliability Statistics for Topic Models 100

15 Reliability Statistics for Section Models..... 101

16 Reliability Statistics for Topic Models 102

17 Reliability Statistics for Section Models for Experiment Two 103

18 Mean Cognitive Presence Values by Student 113

19 Renegade Coding for Experiment One 125

20 Renegade Coding for Experiment Two 135

List of Figures

Figure Page

1 Shannon-Weaver model of communication. 6

2 Osgood-Schramm model of communication. 7

3 A simple ANN that classifies grapefruits. 34

4 Graphic overview of the research methods. 44

5 One artificial neuron. 55

6 A simple ANN architecture showing layers and nodes. 56

7 Message hierarchy outline. 60

8 Mean cognitive presence weight by instructor. 72

9 Mean Cognitive Presence Weight by Course Topic 72

10 Cognitive Presence Percentage by Instructor 73

11 Cognitive Presence Percentages by Topic 74

12 Cognitive presence by week. 76

13 Graphic overview of the research methods. 84

14 Mean cognitive presence value by course topic. 108

15 Mean cognitive presence weight by section. 108

16 Distribution of messages by cognitive presence category and topic. 109

17 Distribution of messages by cognitive presence category and section. 111

18 Mean cognitive presence values by week. 116

19 Number of messages by week..... 116

20 Mean Cognitive Presence Values by Thread for History 4 117

21 Cognitive Presence Category Counts for History 4, Thread 619..... 118

22 Graphic overview of the research methods..... 121

23 Decision logic within exploration and integration categories..... 140

24 Automatic content analysis research outline..... 147

Abbreviations

ALEK	Assessment of Lexical Knowledge
ANN	Artificial Neural Network
CMC	Computer-Mediated Conferencing
CSCL	Computer-supported collaborative learning
DUI	Driving Under the Influence
EDG	Electronic Discussion Group
ETS	Educational Testing Service
GMAT	Graduate Management Admissions Test
LSA	Latent Semantic Analysis
PEG	Project Essay Grader
PRW	Pattern Recognition Workbench
SEB	Scientific Epistemological Beliefs

CHAPTER 1

THE PROBLEM

A Brief History of Technology Use in Education

For the past 100 years, technology use in education has suffered cycles of initial exuberance over the promise of technology followed by disappointment over its failure to meet that promise. Saettler (1968) and Reiser (2002) outline four cycles involving film, audio-visual devices, television, and the personal computer. These efforts may offer a tool to help us break from this cycle.

In 1913, the use of films as instructional tools spawned the visual instruction movement. Thomas Edison proclaimed, "books will soon be obsolete in the schools…. It is possible to teach every branch of human knowledge with the motion picture. Our school system will be completely changed in the next ten years" (Saettler, 1968). However, Reiser also mentions a stance taken by McCluskey, one of the leaders of the visual instruction movement before WWII, who stated that the education community was not greatly affected by the growth in the visual instruction movement despite the initial exuberance over its promise. McCluskey also mentioned that more than $50 million was lost during that period, only a portion of which was due to the Great Depression (McCluskey, 1981).

During World War II, America experienced great success in using audio-visual devices to train both military and civilian adults. This success prompted one German

general to reflect that the Germans did not anticipate the speed with which America could train its service people. During World War II, the U.S. government also produced hundreds of films designed to train civilians for industry jobs, and most of the training directors agreed that this method reduced training time without compromising training effectiveness (Saettler, 1968). The success in training large numbers of diverse people during World War II renewed interest in using audiovisual devices to do the same in schools. However, the studies comparing audio-visual media to live instruction mostly revealed that students learn equally well under both treatments (Reiser, 2002). Again, the initial exuberance over audio-visual instruction was tempered by findings primarily showing no improvement over live instruction.

Like instructional film and the audio-visual movement, excitement surged again in the 1950's over instructional television. This was driven by the Federal Communications Commission's decision in 1955 to set aside 242 television channels for instructional purposes and was further driven by the Ford Foundation's donation of more that $170 million for educational programming (Reiser, 2002). Like its predecessors, however, instructional television ultimately was not widely adopted because teachers resisted the use of televisions in their classrooms; it was expensive to install and maintain the systems; the content was often not much better than a recorded lecture, and the television itself could not create a learning environment (Reiser, 2002).

Most recently, the excitement over the use of computers in education rivals the excitement of previous technologies. The Center for Social Organization of Schools (1983) indicated that by as early as 1983 computers were being used for instructional purposes in 40% of elementary schools and 75% of secondary schools. Further, the use

of internet-connected computers in schools rose throughout the 1990s. Fearing a repeat of the fate of prior technology, we must ask what is different about computer and internet technology from film, audio-visual devices, and television. Reiser mentions a few differences: Computer technology offers the ability to interact with the instructor, content, and fellow students; offers tireless feedback; presents information in a wide variety of forms; offers learners control over their environments. Further, computer technology over the past decade has converged at least two technologies, the personal computer and the internet.

Still, if computer technology is to impact education, educators must use it to provide better learning environments than traditional live instruction. Kozma (1994) implores us to capitalize on the capability of our technology to provide learning opportunities unavailable in live instruction. Current use of online discussions offers one example. Online discussions are electronic conversations among a class of students which usually happen in web-based learning environment that students access. Having an online discussion with classmates is much like having an email discussion with a friend. An online discussion is fundamentally different from a face-to-face discussion in that it offers time to reflect in order to construct a thoughtful message, it is broadcast to all class participants, multiple message contributors can make their contributions simultaneously without talking over each other, and one class participant may quote verbatim from multiple previous messages in order to construct a logical argument.

A Shift in Computer-Mediated Conferencing Research

In the state of Georgia alone, there is reason to believe that the cycle is being broken, that computer and internet technologies are being adopted within education. The

number of credit hours offered through distance learning methods jumped from 59,593 in fiscal year 2000 to 94,531 in FY 2001, an increase of 59% ("Georgia Distance Learning", 2001, December 11). Meanwhile, the International Data Corporation (IDC) estimates that the e-learning market will grow from $2.2 billion in 2000 to $18.5 billion in 2005 (Moore, 2001). Although there is no clear data on the number of students participating in online courses in which every transaction is electronic, there appears to be a migration away from courses delivered solely face-to-face to those either supplemented with or completely reliant on online discussion. This migration toward electronic learning means that the discourse from many of these learning environments is very easily captured providing an opportunity for researchers to study the process of learning in a way that has never been available before. Never before have we had access to electronic texts containing virtually every exchange made by every student for an entire term. Concurrently, our ability to use computers to process text and reveal underlying themes has steadily grown (Rife, Lacy, & Fico, 1998). The convergence of these two realities brings us to our current state in which we have numerous texts available, a growing set of analysis tools, but only recently has research begun to explain the phenomena that take place in the course of learning.

However, the analysis of discussions did not happen immediately. In the first edition of *The Handbook of Research on Educational Communications and Technology*, Romiszowski and Mason (1996) cite years of research in computer-mediated conferencing (CMC) that did not focus on the content of the discussions for analysis. They claim that "the most glaring omission in CMC research continues to be the lack of analytical techniques applied to the content of the conference transcript" (p. 443). In the

face of increased research studies focusing on the content of discussions, Romiszowski and Mason (2003) have backed off that claim in the second edition of the handbook and refer to content analysis as “one of the key areas of research in the CMC field” (p. 420). Henri (1992) was one of the first researchers to apply content analysis to analyzing online discussions. At that time, Henri lamented that “we do not yet possess a body of knowledge concerning the pedagogical characteristics of the content of computer conferences, the scenarios of how learning occurs, or the elements which give rise to learning” (p.120). In response to this absence of knowledge, Henri created a model for analyzing discussions enabling researchers to begin answering these broad and difficult questions. In the years following Henri’s seminal work, numerous content analyses have been conducted to understand the discussion scenarios that give rise to learning (see Romiszowski and Mason, 2003).

A Shift in Communications Theory

At the center of CMC research on online discussions is the message. The message is the text transmitted from one student to the course participants, and a look at the history of the message in communications theory reveals a shift in the need to understand the message itself. In 1948, Claude Shannon presented “A Mathematical Theory of Communication” which contained the following model of the communications process.

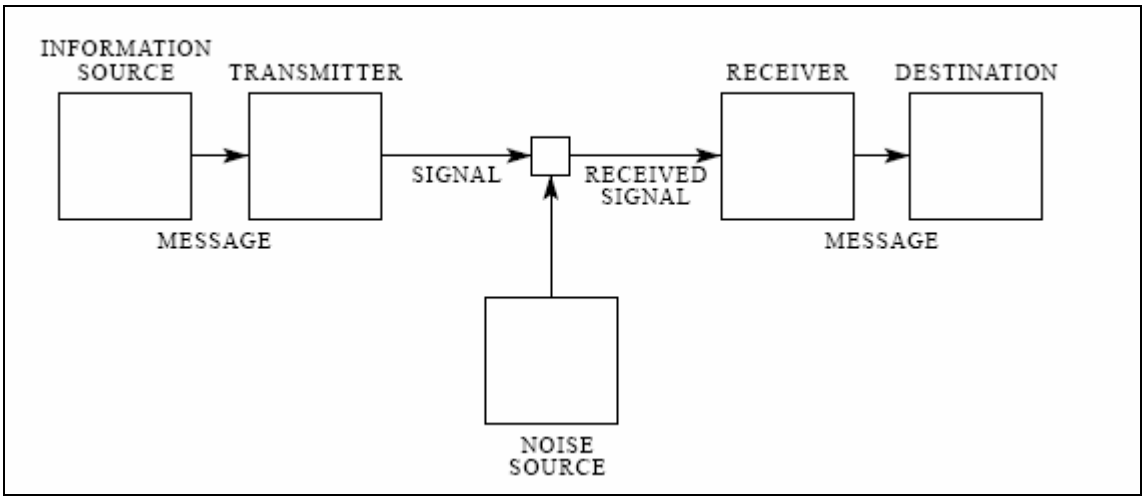


Figure 1. Shannon-Weaver model of communication.

This model, the Shannon-Weaver model, suggests that an information source generates a message which is transmitted through a signal medium and is received at a specific destination. Noise is introduced between the signal being sent and the signal being received indicating that the message may not be received with the same purity with which it was sent. Shannon's (1948) paper describes an engineering process and his model, therefore, does not account for the meaning of the message being delivered. Shannon admits, "these semantic aspects of communication are irrelevant to the engineering problem" (p. 1). For Shannon, a message is simply that which is transmitted over a piece of technology such as a telegraph or telephone.

Working independently from Shannon and Weaver at MIT, Norbert Wiener added learning to Shannon and Weaver's model in the form of a feedback loop (Wiener, 1967 cited in Griffin, 1997). His purpose for doing so was to build an anti-aircraft firing system that adjusts future trajectory based on the system's past performance (Griffin, 1997). If the system overshot the target on previous attempts, it would slightly reduce the trajectory of the next attempt and then use the feedback loop to determine if the

adjustment worked. For Wiener, communications remains an engineering problem, but the message has become a signal that can be modified and improved with appropriate feedback.

Osgood and Schramm (McQuail and Windahl, 1981) later make feedback a much more central aspect of their communications model (see Figure 2). While one sender sends a message, the receiver decodes, interprets, and encodes the message which is then decoded, interpreted, and encoded into a transformed message. Communication looks like a negotiated dance among those communicating. The Osgood-Schramm model contains the major components of the Shannon-Weaver model with the major distinction that it is circular instead of linear. In this model, a message is refined and possibly defined by those receiving and reacting to the message.

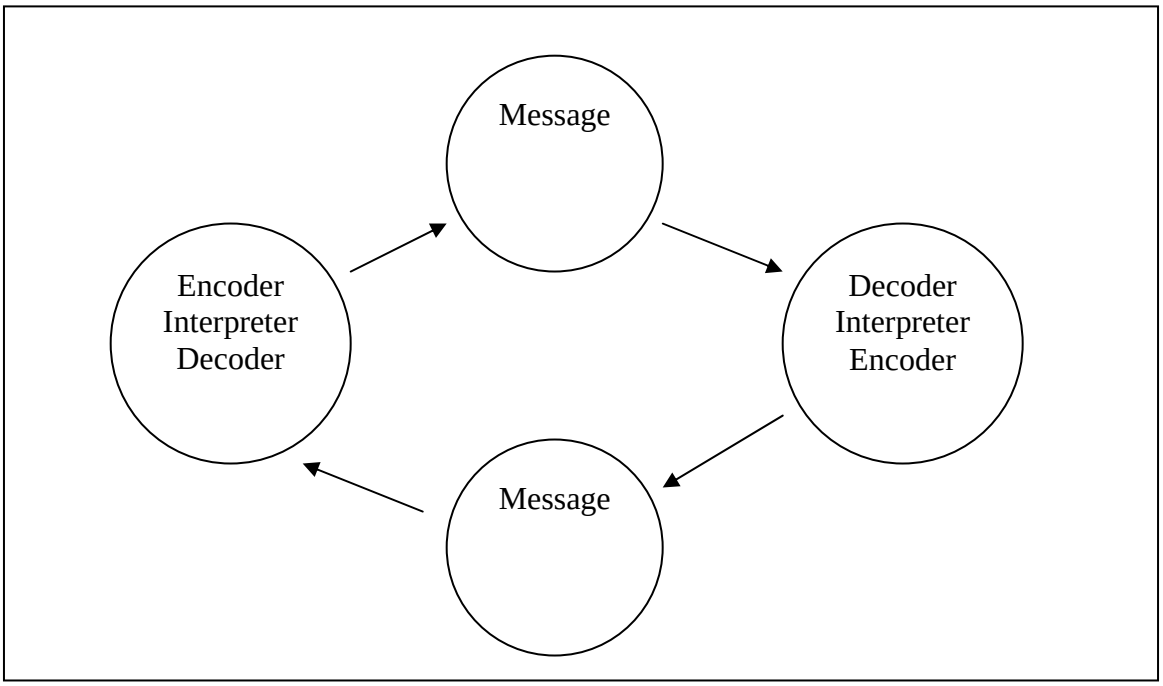


Figure 2. Osgood-Schramm model of communication.

The message in online discussions is not much different than the message described in communications theory. In this study, “message” refers to specific messages

transmitted electronically in online, educational discussion lists. Although it is a much narrower definition than that used by Shannon and Weaver in 1949 (cited in Griffin, 1997, see also Shannon, 1948), Wiener (1967) and Osgood and Schramm (McQuail and Windahl, 1981), it still contains the same characteristics. It is transmitted in the form of electronic text from an information source (the student) through a transmitter (the electronic discussion list), delivered through a receiving medium (again, the electronic discussion list) and read by other students in the class. Further, the sender receives feedback in the form of silence or direct comments from the other course participants. This feedback or the anticipation of feedback affects both the original message and subsequent messages in the conversation. In this sense, the message is not the sole product of an individual but is an expression within and affected by the larger social context of the course participants. As instructors, we want the messages among students to be thoughtful and engaging, to be both the products of and the impetus for higher-order learning. This work offers one method for measuring the quality of the messages shared among course participants in the hopes of creating learning environments rich in thoughtful and engaging discussion.

This brief history of communications models repeats a familiar theme found in CMC research. In communications theory, the meaning of a message has evolved from a place of insignificance to a place of mutually-negotiated significance. In CMC research, a lack of methods to analyze the discussion transcript, and therefore the underlying meaning of student discourse, hindered the analysis of the transcript. Content analytic approaches, however, now offer techniques to analyze the transcript.

Automating Content Analysis

While there is growing CMC research which analyzes discussion content (see Romiszowski and Mason, 2003), current techniques are resource-intensive and therefore, prohibitive. Content analytic approaches often require a trained group of coders to read the discussion list transcripts in order to reliably categorize messages. This approach offers insight into the discussions, but the demands on researchers to employ groups of raters is often prohibitive. One solution to this resource problem is to automate message classification using computer technology. Evans (2001) confirms this potential solution:

> It is no longer tenable to presume that computers cannot help content analysts discover important patterns in their data, patterns that researchers neither intended to investigate nor would have discovered without computer tools. Given recent advances in artificial intelligence, it is no longer tenable to presume that productive insights must be the exclusive province of the content analyst rather than his or her computer (¶ 25).

Garrison and Anderson (2003) agree that we should off-load the resource-intensive tasks of analyzing the discussion list messages: "The task of developing instruments and techniques for transcript analysis…is a necessary prerequisite to the empirical investigation of asynchronous, text-based computer conferencing" (p. 149).

This study builds on the content analysis research Romiszowski and Mason (1996, 2003) outline and follows the advice Evans (2001) and Garrison & Anderson (2003) offer by exploring the use of artificial intelligence to automate content analysis. This study employs an artificial neural network (ANN) as a computational decision-making tool to perform content analyses. Such a tool allows researchers to analyze every message in a discussion transcript instead of a sample and diminishes the resource barriers to conducting manual content analyses. Though a broader explanation of ANNs

and a description of their use are provided in the second chapter, a simple explanation exemplifies the process.

Consider this practical application of artificial intelligence from vision research. Imagine you are a grapefruit producer with a need to provide unbruised grapefruit to market. An ANN may be used to tell us whether a grapefruit is bruised or not. A vision system takes an electronic picture of a grapefruit, and an ANN looks at the color, shape, and size of the grapefruit to determine whether it gets placed in the “bruised” or “not bruised” category. The benefit of this system is that it is capable of analyzing more grapefruits than a human can analyze, it does so faster, and it may be more accurate. Likewise, an ANN may be applied to discussion list messages to determine whether a student applied little cognitive effort or considerable cognitive effort to create the message. Just as the vision system needs to be trained to recognize bruised and unbruised grapefruit, the discussion list analyzer must be trained to tell the difference between messages created with little effort versus those created with great effort. This research study describes how such a system is created and asks whether this system is as accurate as a human at analyzing messages.

Research Question

This research seeks to answer the question: how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans? This question has two parts. The first hypothesizes that an ANN analyzes messages with the same accuracy as a human. The second describes the information an ANN content analysis provides.

Purpose

This research project lays the groundwork for a tool that can be used by a wide variety of educators to objectively evaluate course discussion lists for cognitive effort. Information provided by such an electronic content analysis tool supports educators as they make decisions regarding the best online discussion practices and provides data by which instructors may adjust their methods, instructional involvement, and course design. Henri (1992) makes apparent the evaluative role content analysis has to play in an instructor's ability to guide learning:

> Content analysis, when conducted with an aim to understanding the learning process, provides information on the participants as learners, and on their ways of dealing with a given topic. Thus informed, the educator is in a position to fulfill his main role, which is to offer immediate support to the individual and the collective learning process. (p. 118)

Given the constraints of using content analysis techniques, there is a need for a tool to provide content analysis information automatically. This study outlines the initial phase of the construction and use of such a tool.

This research analysis and system development was conducted using messages from the University System of Georgia's eCore™ program, a distance education avenue delivering the University System of Georgia's (USG) core curriculum. It currently has over 400 enrollments and offers 12 courses in English, mathematics, communications, technology, philosophy, political science, and psychology. This effort analyzes the discussions from eCore™ courses in U.S. history and political science. The eCore™ project allows students to complete the first two years of an undergraduate degree online. eCore™ instructors, like many online instructors, currently have little capability to gain a bird's-eye view of the overall learning taking place, which limits their ability to assess student learning and to intervene if necessary. A full description of the students in the eCore™ program is provided in the third chapter.

Overall, the purpose of an automatic content analysis tool is to provide one more tool to prevent a repeat of the100-year exuberance-failure cycle. This work lays the groundwork for a measurement tool unavailable in live instruction and serves both the researcher and the instructor. For the researcher, an automatic content analysis tool provides an objective analysis of a body of discussions upon which to build theory of online learning. This tool provides objective evidence that a specific educational adjustment works or does not. It acts as a previously unavailable variable enabling numerous quantitative analyses to be performed at a relatively low cost. It also enables exploratory research by providing evidence that an increase or decrease in cognitive effort was observed which encourages the researcher to understand why such a change occurred. Before we can realize the benefits of an automatic content analysis tool, we must lay the conceptual groundwork and demonstrate how such a tool might function. The following chapters first outline the literature from instructional technology, content analysis, and artificial intelligence that, taken together, indicate the possibility for such a tool; provide a method for comparing the decision-making capability of an automatic content analysis tool to the decision-making of a set of humans; provides results from two experiments; explains those results; and finally recommends a series of next steps.

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

This literature review is a compilation of recent research illuminating the use of content analysis within educational settings to understand and inform pedagogical practices. This review focuses on but does not limit itself to online electronic discussions. Additionally, it seeks to understand: (a) whether learning can happen online; (b) how researchers define and code online learning; (c) what units researchers analyze and how content analysis may guide instruction; (d) how content analysis has been used to reveal cognitive structures, including cognitive presence; (e) computational approaches used to conduct content analyses; and (f) the uses of artificial neural networks (ANN) in social science research.

Can We Learn Online?

This section traces approximately 30 years of research in which our understanding of online learning has shifted from speculation over whether learning could ever occur online to research revealing that combined online and face-to-face learning may evoke richer learning environments than strictly face-to-face environments.

First, Rourke, Anderson, Garrison and Archer (1999) outline the debate concerning whether the types of meaningful conversation required in educational settings can happen online. The “filtered cues” argument (Short, Williams, & Christie, 1976;

Sproull and Keisler, 1986; Daft & Lengel, 1986) casts doubt that such meaningful conversations can happen online and proposes that communication environments devoid of rich nonverbal cues such as facial expressions, body movement, and eye contact have a lower ability to support social and affective interactions. These proponents of the "filtered cues" argument hold that environments such as voice mail, audio-teleconferencing, email, and electronic discussion groups foster short and pragmatic exchanges (Daft & Lengel, 1986) or uninhibited, hostile, and self-absorbed communication (Sproull and Kessler, 1986). However, opponents of the "filtered cues" argument insist that social and affective communication can and does happen in text-based, computer-mediated environments such as electronic discussion groups. Numerous studies indicate a considerable portion of electronic messages serve a social purpose such as "expressions of feeling, self-introduction, jokes, compliments, greetings, and closures" (Angeli, Bonk, & Hara, 1998), expressions of openness and solidarity (McDonald, 1998), humor and hurt feelings (Weiss & Morrison, 1998), and social interchange (Kanuka & Anderson, 1998). Opponents of the "filtered cues" argument suggest that social exchanges are not only possible but rampant throughout online educational exchanges.

Additional research suggests that on-line asynchronous discussions can enhance, even facilitate learning. Garrison, Anderson, Archer (2000) found that "the level of critical thinking" was higher in computer-mediated communication (CMC). Newman, Webb and Cochrane (1995) found that "more new ideas emerged in face-to-face seminars, and more ideas in the computer conferences were important, justified or linked together." "As compared with speaking, writing provides opportunities for students to reflect and think more deeply about what they are trying to say" (Hara, 2000), but Henri (1992) questions

whether the “value unique to CMC” with respect to learning has been established. Henri (1992) claims that the unique value of CMC with respect to learning has not yet been proven beyond anecdotal evidence.

One potential benefit of CMC is that the collaborative, public nature of the discussion may facilitate group learning. Within a traditional, classroom environment, the audience for the student’s documented reflections on the subject is generally limited to the teacher. Within asynchronous discussions, however, each student’s reflections are placed in the public space where their ideas may spawn further exploration of the subject by their peers.

Another potential benefit of CMC is the impact of personality and preferred learning style on learning. Bullen’s (1998) study of a university-level course delivered by computer conferencing found that some students prefer participation via written communication over participation in face-to-face discussions. Bullen interviewed students, and several indicated that they found classroom participation to be difficult and expressed a higher degree of comfort in the CMC environment. Bullen connects the expressed higher degree of comfort with the finding that both students and the instructor indicated that the CMC environment yielded greater student participation. However, the data related to personality and preferred learning styles remains inconclusive. One student, a self-described introvert, found the volume of communication in the online course to be overwhelming. Howell-Richardson and Mellar (1996) investigated whether “small differences in course design and in moderator behavior” influenced “the nature of participant interaction” (p. 47) and found via content analysis “that there is a much higher occurrence in inter-referentiality and building of ideas across a multi-lateral group of

participants" (p. 67). More research linking personality and preferred online learning styles is needed and can provide information on how instructors can individualize and optimize instruction in online settings.

Recent research suggests that reflection and justification are more likely to occur in asynchronous learning environments whereas collaboration, social interaction, and conflict are more likely to occur in a combination of face-to-face settings and synchronous settings. First, Picciano (1998) concludes that the asynchronous environment is distinctly different from and complements face-to-face conversational environments because the asynchronous environment provides more time for reflective practice than spontaneous interaction. Further, Pena-Shaff, Martin, and Gay (2001) claim that face-to-face interactions are more similar to online synchronous communication than asynchronous. They claim that face-to-face and synchronous discussions contain a great deal of collaboration, social interaction, and conflict (three times more conflict in synchronous than asynchronous communication), that asynchronous communication is better suited for reflection, and that synchronous communication is a better medium for quick, lively, interesting, and spontaneous discussion. Also, Sherry, Tavalin, and Billig (2000) assume a distinction between online and face-to-face conversations by highlighting the importance for instructors to use online dialogue to move beyond the recall and recitation stressed in classrooms with requests for students to explain, justify, and challenge in online discussions. By creatively employing face-to-face, synchronous, and asynchronous discussion strategies, an instructor may create an overall learning environment which does not stifle communication and which promotes shared responsibility for what others know. That is, the blended environment allows for the

creative use of strategies to lead us toward rich, diverse learning environments, and it is ultimately the instructor's responsibility to employ those strategies.

Definitions and Coding Categories for Cognitive Engagement

The beginning of this chapter outlined the filtered cues debate which indicates that "filtered" environments such as electronic discussion groups can support exchanges necessary to foster and support learning, but what specifically have researchers looked for to determine whether learning is present? Some researchers hold that critical thinking (Seaver, Smith, and Leflore, 2000; Ennis, 1987), higher order thinking (Bloom and Krathwohl, 1956), reflection (Schön, 1987), and social presence (Angeli, Bonk, & Hara, 1998) are essential before learning can become manifest. The following is an outline of a widely used framework for content analyses of online educational discussions including modifications to that framework.

Henri (1992) provides the fundamental framework for most current educational content analysis research and provides the following challenge:

> We do not yet possess a body of knowledge concerning the pedagogical characteristics of the content of computer conferences, the scenarios of how the learning occurs, or the elements which give rise to learning (p. 120).

To address this shortcoming, Henri proposes a method of analyzing content which identifies the learning processes and strategies of the learners, the results of which form guidelines for the "development of a framework to guide interventions and support the learning process" (p. 121). Henri's framework contains five dimensions elaborated in Table 1.

Table 1

Henri's Five Dimensions of the Learning Process

Dimension	Definition	Indicators
Participative	Compilation of the number of messages or statements	Number of messages Number of statements
Social	Statement or part of statement not related to formal content of subject matter	Self-introduction Verbal support "I'm feeling great…'
Interactive	Chain of connected messages	"In response to Celine…" "As we said earlier…"
Cognitive	Statement exhibiting knowledge and skills related to the learning process	Asking questions Making inferences Formulating hypotheses
Metacognitive	Statement related to general knowledge and skills and showing awareness, self-control, and self-regulation of learning	"I understand…" "I wonder…"

Newman, Web, and Cochrane (1995) build on Henri's (1992) framework in a content analysis designed to find displays of critical thinking in both face-to-face and computer conferencing seminars. They define critical thinking as "a dynamic activity, in which critical perspectives on a problem develop through both individual analysis and social interaction" (¶ 23). Specifically, these researchers combine Henri's framework

(see Table 1) with Garrison's (1992) five stages of critical thinking (problem identification, problem definition, problem exploration, problem evaluation/applicability, and problem integration) (p. 63) to analyze one semester's worth of messages from student exchanges made by students in two courses at Queen's College, Belfast. One course was comprised of computer conferencing seminars; the second was delivered face-to-face. The researchers tagged student messages for paired opposites revolving around relevance, importance, novelty, outside knowledge, ambiguities, interpretation, justification, critical assessment, practical utility, and width of understanding and found that the computer conferencing group exhibited more positive displays of each of the above categories except novelty.

Hara, Bonk, & Angeli (1998) and Angeli, Bonk, & Hara (1998) also use Henri's (1992) model as a springboard to scrutinize online discussion text, but they focus specifically on displays of online community building. Hara, Bonk, & Angeli adhere to Henri's five dimensions yet expand Henri's interaction dimension to include multiple participants, modify the cognitive dimension to reflect that the initial questions posed in an online discussion influence the quality of the ensuing discussion, and extend the metacognitive dimension to include reflection, self-awareness, regulation and planning. Angeli, Bonk, & Hara build even further on Henri's five dimensions by including the modifications above, illustrating links between online messages using interactivity graphs, and by collecting word length and number of messages.

Garrison, Anderson, & Archer's (2000) Community of Inquiry Model breaks Henri's (1992) model into three related components and expands on it through subcomponents. Specifically, Garrison, Anderson, & Archer view online discussions

through three lenses: cognitive presence, social presence, and teaching presence. Specifically, cognitive presence expands Henri's cognitive dimension by operationalizing it as communication which falls into one of five categories. Messages are either non-cognitive, triggering events, exploration messages, integration messages, or resolution messages. Further, Rourke, Anderson, Garrison, & Archer (1999) look specifically at Henri's social dimension by studying graduate level courses delivered at a distance for displays of social presence which they define as "the ability of learners to project themselves socially and affectively into a community of inquiry" (¶ 1). They further break this definition into three broad coding categories: affective response, interactive response, cohesive response, and provide operational definitions for each. Affective responses contain expressions of emotion, humor, and self disclosure; interactive responses are those in which the user continues a thread, quotes or refers to another's message, asks questions, compliments, and expresses agreement; and cohesive responses serve a social purpose such as addressing a participant by name, using inclusive pronouns (we, us, our, etc.), and expressing salutations such as greetings and closures. Social Presence is the third element in the Community of Inquiry Model. The two transcripts the researchers studied reveal a "social presence density" of 22.83 and 33.54 per 1,000 words, but the authors do not indicate the relevance of these numbers beyond their use as a numeric measure of social presence density. The researchers do not indicate whether the data represent a significant difference between the two social presence densities nor what factors may have contributed to that difference, but the data do suggest that Social Presence plays an integral part in asynchronous text-based computer conferencing. Specifically, the researchers insist that social presence density allows us to explore the

relationship between cognitive presence, teaching presence, and other indicators of participation and learning.

Units of Analysis

Rife, Lacy, & Fico (1998) define units of analysis, or units of content, as "a discretely defined element of content. It can be a word, sentence, paragraph, image, article, television program, or any other description of content based on a definable physical or temporal boundary, or symbolic meaning" (p. 58). This section outlines the units of analysis used by many online discussion list researchers and exposes the debate between syntactic and thematic units of analysis.

Henri (1992) and McDonald (1998) identify the "meaning unit" as the unit to study. Henri proposes that researchers give up purely quantitative content analyses of computer-mediated conferencing (CMC) such as word frequency counts and the number of messages from students and insists that the individual message cannot be used as a unit of measure since one message may contain multiple ideas. Therefore, Henri proposes using a thematic unit of measure and divides "messages into statements corresponding to units of meaning" (p. 126). Like Henri, Newman, Web, & Cochrane (1995) employ a method of analysis which does not restrict them to one syntactic unit of analysis. Instead, they coded each statement for the presence or absence of various indicators. The raters tagged each statement which means that the tags could overlap and could contain a single phrase, sentence, paragraph, or entire message.

Others, however, stray from Henri's (1992) use of a thematic unit because it does not yield sufficient reliability. Angeli, Bonk, and Hara (1998) use paragraphs as units of analysis. Their work is a content analysis of a graduate educational psychology course

and the researchers assume that each paragraph marks a new idea since this population should be capable of breaking their messages into paragraphs. However, when one idea was contained in two paragraphs, it was counted as two units, and the presence of two ideas in one paragraph was also considered to be two separate units. This method yields an aggregate interrater reliability score of 74.6 percent (p. 9). Rourke, Anderson, Garrison, & Archer (1999) partially reject Henri's and McDonald's use of meaning units because "they resist reliable and consistent identification" (p. 11). Instead, "the most appropriate unit [a single message] would combine the flexibility of the thematic unit, which allows coders to capture a unit in its natural form, with the reliable identification attributes of a syntactical unit" (p. 11). To clarify, a message is the body of text comprising a single student's posting to the discussion. This approach yields interrater reliability figures (calculated as percent agreement) of 0.91 and 0.95 in the teaching presence study. Though it is important to understand reliability figures resulting from different units of analysis, these figures emanate from numerous other factors besides the unit of analysis. That is, switching the unit of analysis does not guarantee high reliability figures, but there are cases in which high reliability figures have resulted from the use of syntactical units.

Fahy, Crawford and Ally (2001) "deliberately separated the unit of meaning from the unit of analysis," choosing to use several units of measure. For topical persistence, the unit of measure was the number of postings per subject thread while the unit of analysis for their Transcript Analysis Tool (TAT) Category Types was the sentence. The TAT types thus obtained were then compared to total number of student sentences and occurrences of TAT types per 1000 words. This approach yielded interrater reliability of

60% and 71% (calculated as percent agreement among three raters) on two trials and kappa values of 0.45 and 0.65 for pairs of raters on another two data samples. The researchers concluded that additional rater experience with the coding tools would probably result in increased reliability.

Content Analysis as a Guide for Instruction

Aside from the more traditional uses of content analysis, there is a body of educational research employing content analysis as teacher tools.

Misulis (1997) looks at content analysis as a tool for lesson planning, and states that there are three purposes for conducting a content analysis: First, "the content analysis helps the teacher identify what is to be learned;" second, it "facilitates continued planning of instruction;" third, it "helps the teacher select pertinent instructional methods, activities, and materials" (p. 45). Misulis refers to content analysis as a way for a teacher to focus his/her pre-planning efforts for a course.

Several studies use content analysis to identify the importance of the role of starter or facilitator within online discussions. Angeli, Bonk & Hara (1998) found that discussions which begin informally without a starter were more random and less interactive and that the starter's questions affected the quality of the students' cognitive effort. Starters in their study posed questions related to Henri's five categories: elementary clarification, in-depth clarification, inference, judgment and application of strategies. The role of a skilled facilitator was also mentioned by Garrison, Anderson, Archer (2001). The researchers noted that the message transcripts revealed a low frequency of responses related to integration and particularly resolution, categories requiring the greatest amount of cognitive effort. Several potential explanations for this

phenomenon were provided, one of which was the role of the facilitator or guide who may view his/her role as one of leading students on an exploration of a topic instead of requiring students to formally integrate what they learn or derive resolute theories based upon that integration. They also posit that the introductory nature of the course may not have lent itself to resolution, or that computer conferencing may not support resolution.

Likewise, MacKinnon (2000) and MacKinnon & Aylward (2000) use content-analysis techniques to evaluate the electronic discussion group (EDG) portion of a course at Acadia University. The course is comprised of 30 students divided into six EDG groups of five students each. The students meet in EDG groups three times throughout the term, and after each meeting, the students' contributions are graded against the following generic model based on Knight's (1990) evaluation of written reading journals (see Table 2).

Each specific interaction is a categorical code or "cognote." After two weeks of discussion, the students' text is captured, graded, and returned to the students. Not surprisingly, the researchers notice that students shift their conversation style to yield more high-grade cognotes. The researchers warn that improved communication patterns may only be a response to the grading scheme and note the emergence of two distinct groups: the first was preoccupied with the cognote system whereas the second group reported that their initial focus on the cognote system faded as their discussion progressed and believed that this formal discussion procedure had been internalized.

Table 2

Cognote Model for Evaluating Online Discussion Group Messages

Specific interaction	Grade Assigned	Code Name
Acknowledgement of opinions (evidence of participation	1	acknowledge
Question (thoughtful query)	1	Question
Compare (similarity, analogy)	2	Compare
Contrast (distinction, discriminate)	2	Contrast
Evaluation (unsubstantiated, judgment, value)	1	Evaluation
Idea to Example (deduction, analogy)	2	idea2ex
Example to Idea (induction, conclusion)	2	ex2idea
Clarification, elaboration (reiterating a point, building on a point)	2	clarify/elaborate
Cause and effect (inference, consequence)	2	Cause&effect
Off-topic/faulty reasoning (entry inappropriate)	0	Offtopic

Although a content analysis is a crucial area for research, a large body of research does not exist. Current content analysis techniques are largely manual and very labor intensive (Angeli, Bonk, Hara, 1998; Rourke, Anderson, Garrison, 1999). This appears to have impacted both the volume of research, as well as the size of the research population (Fahy, Crawford, Ally 2001; Garrison, Anderson, Archer, 2001).

Cognitive Structures

Finally, at least two educational research studies use content analysis to understand cognitive structures. First, Tsai (1999) conducted a study to understand whether students organize their scientific knowledge along empiricist (traditional) or constructivist (non-traditional) lines. To determine students' self-reported scientific epistemological beliefs (SEB), Tsai used a Chinese version of Pomeroy's (1993) bipolar agree-disagree questionnaire measuring the students' SEB and a content analysis of the students' cognitive structures. The students' cognitive structures were mapped following two treatment lessons on atomic physics. Each student was tape-recorded and a flow map of the students' narratives was diagrammed to allow researchers to acquire a complete view of the learners' cognitive structure. Each flow map was coded at the content (specifics, relations, transformations, and generalizations) and logic (defining, describing, comparing and contrasting, conditional inferring, and explaining) levels. The subcategories within these levels are ordered which means that "specifics" and "defining" are lower-order tasks whereas "generalizations" and "explaining" are higher-level tasks. Tsia's results indicate that these above-average students "tended to use relatively lower-order modes of knowledge organization and cognitive reasoning when recalling the scientific information" (p. 131). Another finding is that it took the students longer to retrieve information as the complexity level of their ideas increased. To test the reliability of Tsai's coding, a second researcher randomly selected and coded eight of the 48 flow maps. This test yielded a kappa coefficient of 0.87.

Similar to Tsai's analysis of students' cognitive structures, Domin (1999) analyzed laboratory manuals to determine which cognitive structures they encourage.

Domin (1999) hypothesizes that chemistry laboratory manuals promote more lower-order thinking than higher-order thinking and uses Bloom's taxonomy to separate these levels of cognition. This study analyzed the content of three experiments (gas laws, kinetics, and calorimetry) in ten lab manuals and looked specifically at verbs in context to determine which cognitive skill the manual requires of the student. Domin's results show that eight of the ten lab manuals require students to work at the lower levels of cognition and notes, "the laboratory manual reduces the amount of time necessary to complete a laboratory activity by providing an instructional pathway that does not require the utilization of higher-order thinking skills" (¶ 8).

The content analysis literature speaks to the complexity of Tsai's and Domin's work; it informs us of the complexity of using manifest artifacts to reveal latent constructs that brought those manifest artifacts about. Specifically, Potter and Levine-Donnerstein (1999) provide an important framework for content analysis which allows researchers to gauge the complexity of their content analysis task. They mention three types of meaning to be gleaned from content analysis: manifest content, latent patterns, and latent projections. Manifest information is the easiest to derive and reveals meaning contained within the text. Manifest content answers questions like, "how many times did the word 'yes' appear?" Latent patterns ask coders to identify patterns within content that reveal a latent construct. Coders objectively indicate the manifest parts which create a combined whole. For example, identifying formal or informal attire in a content analysis of photographs may involve multiple indicators such as the presence of a necktie and a suit against a formal setting including others wearing similar attire. Deriving latent projections, the most difficult meaning to reliably derive, means that coders use their own

schema to project meaning onto the text, to interpret text through their own social filters. A simple example of deriving latent projections would be to ask coders to identify humor in a series of stories. It would be difficult to derive a reliable, objective rubric to identify manifest indicators of humor; coders are forced to code content based on their own interpretations of humor.

Computational Approaches to Content Analysis

Certainly, recent technological innovations could advance research in content analysis of online learning environments. First, computational statistics packages have been used for some time in content analysis as in most other research methods. Evans (2001) suggests pushing the use of technology in content analysis a bit further through computer-supported content analysis. There are many computational tools to assist with human coding (e.g. NUD*ist, Atlas-TI). He also refers to Franzosi (as cited in Evans, 2001) who recommends that coding protocols be available online and that the coding itself be done online. Computers may also play a role in preprocessing content. MoCA (Movie Content Analysis) is an example of a tool that preprocesses movies by identifying scene breaks, online events (e.g. explosions), and on-screen text (e.g. signs). Finally, Evans (2001) mentions that we can reasonably expect computers to perform the actual coding that humans perform and to automatically derive coding categories. Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997; Landauer, Foltz, and Laham, 1998) is one promising approach that has been used to improve the accuracy of electronic document retrieval tools, an innovation which applies to library scientists and anyone who performs searches for documents using popular internet-based search engines. It uses a mathematical technique, singular value decomposition (Landauer, Foltz, and

Laham, 1998), to associate terms with document topics even though those terms may not appear in the same document. Landauer, Foltz, and Laham (1998) have used this technique to build a tool that paired synonyms it had never before encountered – a tool which performed with 65% accuracy on the vocabulary portion of the Test of English as a Foreign Language (TOEFL) which is "identical to the average score of a large sample of students applying for college entrance in the United States from non-English speaking countries" (p. 22). LSA may hold promise for computer-supported content analysis in that it may associate units of text with predefined categories, or its factor analytic approach may enable it to automatically derive coding categories.

Further, Evans (2001) refers to theme-based and clause-based techniques as crucial to computer-supported content analysis. Theme-based techniques, as the name implies, computationally identify themes in content. The General Inquirer (Danielson & Lasorsa, 1997) is an example of this in that it is a dictionary comprised of 12,000 words in 182 categories. The General Inquirer associates content words with General Inquirer categories, and, by matching the words in the content analysis text to words in each of the General Inquirer categories, it indicates which categories are most deeply expressed by the text. This is limited in that it focuses on decontextualized words. The General Inquirer looks only for the occurrence of words, not for their meaning or context; it does, however, do some word sense disambiguation in that it correctly associates words which fall under multiple parts of speech into their correct part of speech (i.e. it distinguishes between the verb and noun forms of the term, walk). Evans also cites clause-based techniques which seek to analyze and correlate clauses in content text. That is, the clause is the locus of meaning. There are two packages which analyze at this level and use a

neural network to do some of the processing: Map Extraction, Comparison and Analysis (MECA) and Computer-Assisted Evaluative Text Analysis (CETA). Another clause-based program, Program for Linguistic Content Analysis (PSCA), simplifies clauses for further computational analysis. For example, it has the ability to reduce an English sentence to its simplified and component parts and creates simpler sentences from the text's more complex sentences taking the form of Agent → Action → Object. For example, it may take the convert sentence A to sentence B:

> Sentence A: The angry senator from South Carolina exemplified general distrust over the way the administration handled the Gulf War Crisis.
> Sentence B: Senator distrusts administration.

Certainly some meaning is removed from the sentence, but if the converted text always falls into the same Agent → Action → Object format, then researchers can computationally analyze it, to fit agents and objects together. Evans mentions that this offers the ability to connect objects and agents even if they appear many paragraphs apart from each other. The example he provides is:

> Sentence C: X supports house bill Y (in first paragraph)
> Sentence D: House bill Y leads to unemployment (in a distant paragraph)

The software then pieces sentences C and D together to achieve:

> Sentence E: X supports unemployment

Finally, Hearst (2000) outlines the emergence of automated text grading, a practice currently employed by the Educational Testing Service (ETS) to grade student essays. Hearst first mentions Ellis Page's Project Essay Grader (PEG) which uses multiple linear regression performed on automatically extractable features of text (average word length; essay length in words; number of commas, prepositions, uncommon words) to approximate teachers' grades. PEG was created between 1966 and

1968, was the first tool to be used by ETS for automated text analysis, and generated a correlation of .78, which was almost as strong as the .85 interrater reliability statistic among human readers. Second, the Writer's Workbench was used in the 1980's as a tool to identify and extract measures of writing quality. It was not used as an essay grader per se but did provide feedback to students on the quality of their writing by providing feedback on spelling, diction, and readability. Finally, by the 1990's, the need to automatically assist in scoring essays and short answer items on the Graduate Management Admissions Test (GMAT) coupled with advances in natural language processing and information retrieval (similar to the technology used by internet search engines) led to tools that measure syntactic variety, identify sentence type, and identify topic via lexical content analysis. Still, these tools lacked the ability to identify individual arguments and to evaluate their rhetorical structure. In response to this, ETS developed tools to break an essay into its individual arguments and then to perform a vocabulary content analysis on those arguments. This work produced the e-rater, advances to PEG and latent semantic analysis (LSA). LSA is designed to go beneath the surface vocabulary to identify an essay's semantic content. Interestingly, e-rater is used by ETS to score GMAT essays, and human scorers are only brought in to resolve different scores by two e-rater models (Burstein, Marcu, Andreyev, & Chodorow, 2001).

Hearst (2000) outlines three current research topics. Assessment of Lexical Knowledge (ALEK) is a technique to detect lexical grammatical errors such as "I concentrates" which provides an inverse relationship to an essay's score. Second, Centering Theory is designed to detect rough shifts in essay topics; the more rough shifts in an essay the lower the score. Finally, current research focuses on generating

summaries to improve scoring performance. These summaries are based on lexical shifts (e.g. because, therefore, however, etc.) to generate summaries, and these lexical shifts identify the presence or absence of specific arguments.

Overall, numerous approaches have been taken to understand student-generated online text, support material, and student self reports. This analysis indicates a lack of tools and knowledge concerning the pedagogical use of online discussions to support education. Clearly, there is much yet to learn about online, student-generated text, and content analysis has a distinct, pedagogical role to play in reducing our current uncertainty.

Uses of Artificial Neural Networks in the Social Sciences

Certainly content analysis has a role to play in analyzing online discussions. This research study proposes that content analysis can be automated using an artificial neural network, that a system can be built to automatically categorize messages. With that application in mind, we must ask whether artificial neural network (ANN) research supports such tasks. Broadly, ANN research supports using ANNs for categorization tasks, but there is conspicuously little available research in which ANNs are used as a method in education research.

Artificial neural networks (ANNs) were first developed in the 1950's to both try to understand the brain and to mimic its strengths (Fausett, 1994). Fausett defines ANNs as information processing systems comprised of simple processing elements called neurons. Each neuron is connected to other neurons and each connection is associated with a weight (w). Connections with a higher weight represent strong connections whereas connections with a lower weight indicate weak connections. The weights are

critical to ANNs because they represent how the information is used by a network to solve a problem. Each neuron also has an internal function called an activation function which calculates the inputs and determines whether or not to send a signal. A neuron may receive a signal from any input neuron but may only send a signal to one neuron.

Let's go back to the bruised grapefruit problem from the first chapter. Recall that this system analyzes images of grapefruits to determine whether a grapefruit is bruised or not. For this example, there are three inputs describing the grapefruit: size, shape, and color. The neural network has been trained on data containing both descriptions and classifications of hundreds of grapefruits; the descriptions detail the size, shape, and color of each grapefruit while the classification indicates whether it is bruised or not bruised. Figure 3 shows that after training, the ANN indicates a strong relationship between color and the neuron which decides whether to categorize the grapefruit as bruised or not. The weights w_1 and w_2 are depicted with a dashed line and show a relatively weak connection while the weight w_3 is bold and shows a strong connection. The neuron, H_2, receives input from all three input neurons, applies its activation function on the three inputs, and then sends a signal to one of the two output layers. For this example, the neuron, H_2, receives input from all three input neurons but gives more weight to color. Its activation function is trained to call a grapefruit bruised if more than 10% of the grapefruit's image is brown. Imagine a case in which 50% of the grapefruit's image is brown. This information is sent to the hidden layer, the activation function sees that the color exceeds its activation threshold of 10%, and the grapefruit is classified as bruised.

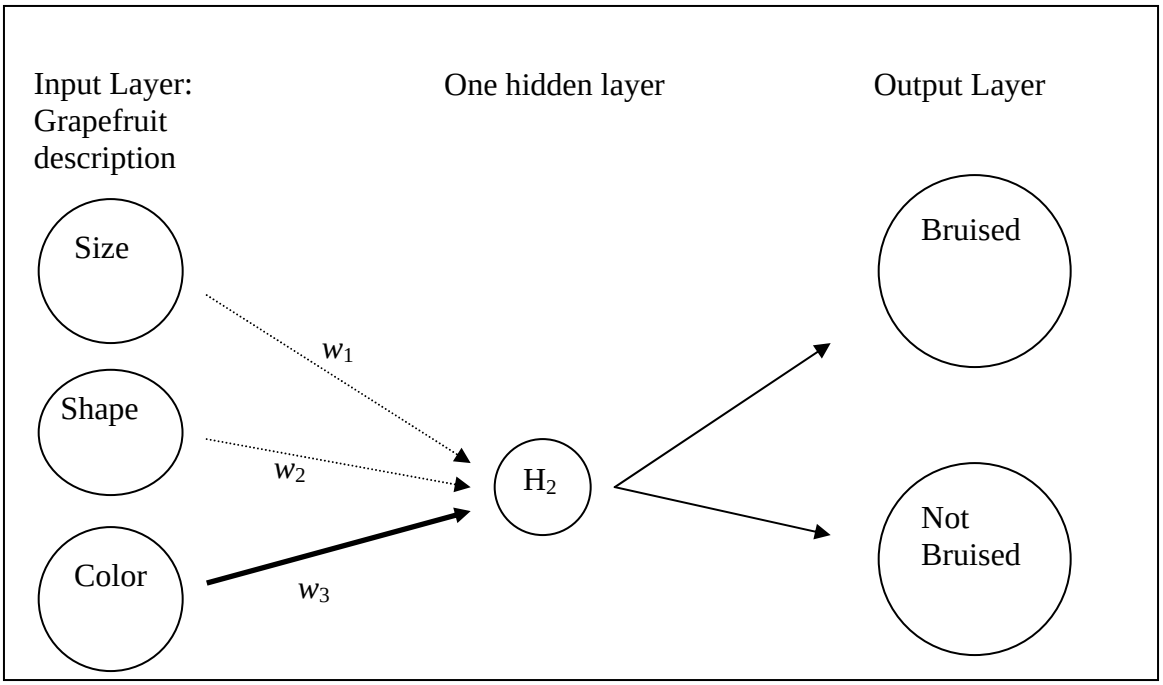


Figure 3. A simple ANN that classifies grapefruits.

Garson (1998) describes two predominant uses for artificial neural networks: prediction and classification. ANNs have been most notably employed to predict. For example, they have been employed by financial analysts to predict the end-of-year net asset value of stocks (Chiang et al., 1996), to predict the mortality of critically ill patients (Dybowski, 1996), and to predict wilderness recreation use (Pattie and Haas, 1996). They have also been used to classify or categorize information. The most notable examples of this are speech and optical character recognition. In these examples, input as either sound or images is categorized into a finite list of available phonemes or letters.

Artificial neural networks also offer an alternative to traditional linear models. Hedgepeth's (1995) study of 400 years of military combat suggests that traditional, linear statistical analyses may outperform ANN approaches where the data are clean and filtered, but that ANN approaches outperform linear techniques when the data are noisy and imperfect. Also, Garson (1998) mentions that neural models, unlike regression

models, are unaffected by the interaction of input variables, that ANNs can handle both non-linearity and interaction effects.

In their work comparing artificial neural network approaches to linear methods for tackling psychological classification and prediction problems, McMillen and Henley (2001) agree with Garson (1998) that ANNs are more suitable for some data sets. McMillen and Henley (2001) compared ANNs to discriminant analysis and logistical regression techniques in the context of classifying Driving Under the Influence (DUI) risk status among heavy drinkers using psychological predictor variables. This problem was chosen because, like many classification problems using survey data, there is often missing data and the data frequently violates the assumptions of common linear models (e.g. linearity, homoscedasticity, and intercorrelation of variables). Specifically, the researchers compared several ANNs to a regression model using 10, 12, and 14 predictor variables for high-risk and low-risk drivers. This study found that regression models using 10 and 12 predictor variables outperformed ANNs with the same number of variables. However, ANNs with 14 predictors performed as well or better (but not substantially so) than the logistical regression method. Ultimately, they side with Garson (1998) in claiming that ANNs are most useful with problematic data sets (sets with missing variables or multiple, conflicting variables) when insight and explanation are of less interest than accuracy.

This final point is a critical one. McMillen and Henley (2001) recommend employing ANNs when the researcher is more interested in accurate categorization than in insight or explanation of categorization. Smith, McKenna, Pattison, and Waylen (2001) highlight this point as well. They compared ANN techniques to structural

equation model (SEM) techniques. The major benefits of structural equation modeling is that it provides for the flexible combination of numerous multivariate techniques, provides robust goodness-of-fit statistics, is available in fairly easy to use computer programs, and identifies significant variables and their relationships with one another on human decision-making tasks. The authors recommend this technique as an intermediate step between theory speculation and fully-formed psychological theory. The major point the authors make is that SEM provides information that leads researchers to a theory explaining the relationship between the inputs and the output whereas non-linear techniques, like ANNs, do not.

Garson's (1998) outline of recent research on neural networks suggests an increasing use of ANNs in the social sciences. Financial analysis is one of the few social science domains to adopt artificial neural networks over regression models for multivariate analysis. The reason for this switch is that ANNs excel at prediction and have offered financial analysts tools that improve investment performance, predict problem credit card applicants, and calculate risk. Garson (1998) also mentions uses of ANNs in sociology. They have been used to predict violent crimes, to predict white collar crime, to model human decision-making in relation to theories of religion, and have predicted child sex abuse. Garson (1998) cites numerous uses of ANNs in political science including predicting the likelihood that students will pass the bar exam, predicting the administrative success of local school principals, and predicting the outcome of case law. Finally, Garson mentions numerous applications of ANN to psychology to support research in combat psychology, depth perception, and information

acquisition. Barring the use of ANNs to predict the performance of school principals, there is little indication that ANNs have been applied in education research.

Garson (1998) mentions that the single largest drawback to using ANNs in the social sciences is the fact that inputs cannot be traced to outputs; scientists cannot provide the decision logic that the ANN takes. Therefore, many have questioned the validity of using ANNs since no one can be absolutely certain that the answers derived from the ANN are not quirks of data. For this reason, most researchers rely on reliability statistics to explain that their network has merit; reliability, however, only tells us that the same answers will be achieved from one time to the next not that the means of deriving those answers is valid.

Overall, neural networks are appropriate for both prediction and categorization tasks, are appropriate for tasks that require accurate classification without a need to detail the relationship between inputs and outputs, may outperform linear statistical procedures for fuzzy, imperfect data, have been made accessible in the past decade through software providers who have simplified their use, and are being widely adopted in many of the social sciences.

Summary

This chapter outlines the progression of research which suggests that we not only can learn online but that learning environments can creatively blend face-to-face, synchronous, and asynchronous discussions offer educators learning environments rich in reflection, spontaneous conversation, and even conflict. But knowing that we can learn online is just the first step; we must then ask what qualities determine whether learning is occurring. Henri (1992) offered the first content analysis framework for exploring online

discussions and insists that we look at five dimensions of the discussion: participative, social, interactive, cognitive, and metacognitive. Though the model has been modified over time, it has not been rejected outright. Garrison, Anderson, and Archer (2000) created perhaps the most thorough modification and operationalization of Henri's model by breaking it into three components (cognitive presence, social presence, and teaching presence) and by expanding each component into deeper subcomponents. With this more elaborate model to guide our analysis of online discussions, we must then ask what units of the discussion are the best ones to analyze. The debate over whether it is better to use theme-based units of analysis or syntactic ones continues. Proponents of theme-based units contend that this method allows them the flexibility to capture each idea within a single discussion list posting; proponents of syntactic units reject theme-based analyses as unreliable because themes are difficult to operationalize and to consistently identify. Further, there is a growing body of literature describing how content analysis has been used in education. It has been used as a tool to guide lesson planning, to describe the importance of the discussion facilitator, to evaluate online courses, and to identify the cognitive structures underlying discussions. There are certainly many more potential applications of content analysis in education and a growing body of research indicates that computational approaches may reduce the resources currently required to conduct manual content analyses, may simplify transcripts, and may define content analysis categories. To date, Latent Semantic Analysis (LSA) has been used to associate terms with topics and the Educational Testing Service (ETS) has demonstrated a tool to automatically grade essays.

Collectively, this body of research reveals numerous needs. Among them is the need to analyze the discussion list text to describe for instructors and researchers the learning displayed in online environments. Just as important, this body of research reveals what we know. We know that it is possible to learn in electronic environments, Henri (1992) and Garrison, Anderson, and Archer (2000) provide for us a model for analyzing discussion content, we know that content analysis has informed educational practice, we know that computational power can support content analysis, and we know that artificial neural networks have provided help in making complex prediction and categorization decisions. Based on what we know from the literature, the following chapter offers a method to determine whether an automatic content analysis tool will code messages with the same accuracy as a human.

CHAPTER 3

METHODS

The literature review reveals that we can learn in online environments, that we possess models to understand the message transcript, that content analysis offers methods for analyzing the transcript, and that artificial neural networks may assist in categorization tasks. Informed by this body of research, this chapter outlines a method for determining whether an automatic content analysis tool can categorize messages as accurately as a human. Specifically, this chapter describes the methods used to answer the research question, "how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans?" This question has two parts. The first part hypothesizes that an artificial neural network (ANN) analyzes messages as well as a human. The second part describes the information expected from an ANN content analysis tool.

Comparing Artificial Neural Networks to Humans

The "Research Method Steps" section outlines the method used to address the first part of the research question, "an artificial neural network (ANN) analyzes messages as well as a human." This series of steps has three major components: message preparation, human content analysis, and artificial neural network content analysis. The steps are provided below and expanded upon afterwards.

Research Method Steps

The following steps were taken to address the first part of the research question.

Message preparation

1. Transfer message text to a database.
2. Use a systematic, random sampling technique to extract three unique bodies of messages: one for training coders (300 messages), one for reliability statistics (100 messages – all coders and the ANN tool will code this set of messages), and one that the coders will independently code (1200 messages = 200 per coder).
3. Build an online tool to allow coders to rate messages.

Human Content Analysis

4. Modify the coding rubric to match the content.
5. Select the human coders.
6. Train the coders to apply the rubric.
7. Coders code a body of training messages.
8. Compare reliability statistics among coders after each training session until the coders exceed a reliability threshold.
9. Coders use the online tool to rate their set of 300 messages (200 unique messages; 100 messages used for inter-rater reliability).
10. Build one aggregate set of human-coded message decisions. This step applies the decision logic table (Table 5) for aggregating human-coded messages.

Artificial Neural Network Content Analysis

11. Numerically describe each message. A database script parses the message body into individual words. That database script then counts the number of times each general inquirer theme is present in the message. The script generates a database table containing a count of general inquirer themes. Each General Inquirer theme category is a field in the table, and each row represents one message.
12. Determine the predictor order. A neural network model is constructed using all predictors (general inquirer categories, self-defined categories, structure categories). A database table is built in which all the fields are re-ordered from highest to lowest discrimination.
13. Build neural network models using the 1200 human-coded messages. This step requires building three types of models, overall models, topic models, and course models. To derive one overall model, 13 models were built and the best overall model was selected based on its ability to code messages similarly to the group of human coders. To derive the best topic models, 13 models for each topic (history and political science) were built and the best of these models were selected (one for each topic) based their ability to code messages similarly to the group of human coders. To derive the best course models, 13 models for each of the six courses were built and the best course models (one for each course) was selected based on their ability to code messages similarly to the group of human coders.
14. Compare the best overall model to the best topic and course model. This comparison tells us whether the overall model is more reliable than the topic or course models.

15. Using the set of 100 messages set aside for measuring inter-rater reliability, compare how the ANN categorizes the 100 reliability messages to the aggregate of human coders.

The above steps were repeated twice. The first iteration revealed flaws in the human content analysis that resulted in a lower reliability among the human coders and between the ANN and aggregate of human coders. The results section details the findings from the first iteration, outlines the modifications made to the method, and then presents the findings from the second iteration.

Figure 4 presents another way to consider the research method. Instead of following the series of activities, it looks at how the message text is transformed and ultimately categorized. A human content analysis is first performed on the message text which results in a body of categorized messages. Those categorized messages are used to train the automatic content analysis tool. Before doing so, however, the messages are numerically described. The ANN uses the numerically described messages from the human content analysis to train itself to code messages.

Message Preparation

Before any analysis can begin, the discussion list messages must be prepared. This involves three steps: exporting the messages from raw text into a relational database; separating out messages to be used for coder training, for calculating reliability, and for coding; and building an online tool allowing us to more efficiently relate message text to message codes.

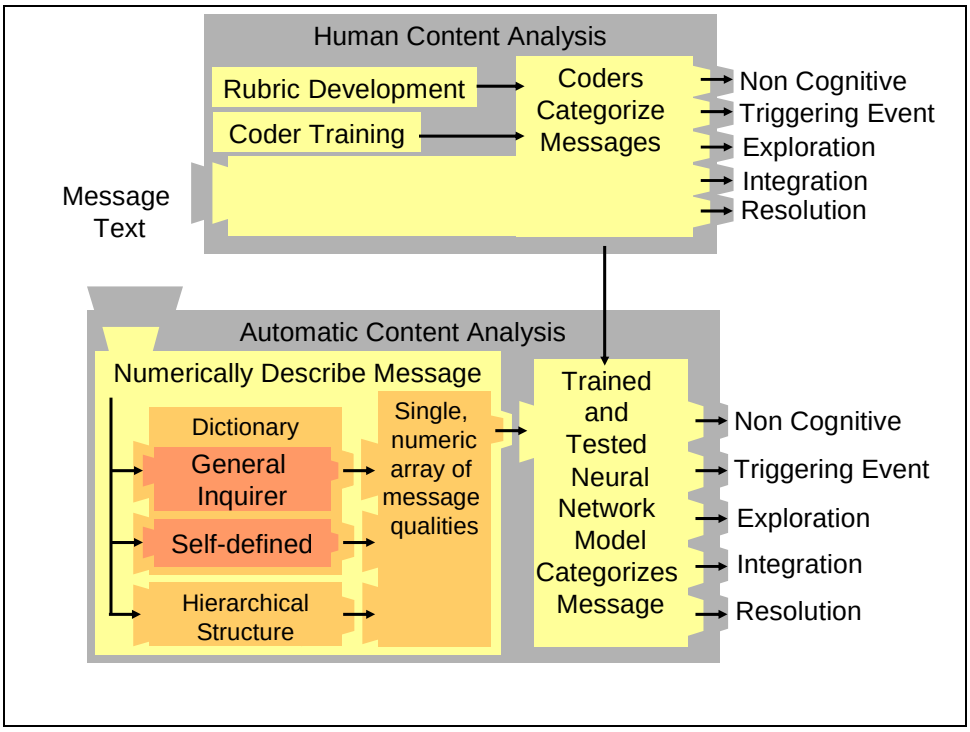


Figure 4. Graphic overview of the research methods.

Transfer Message Text to a Database

Most online learning environments, including WebCT (used in this analysis), allow administrators to export a semester of messages to a text file. For analysis purposes, identifying marks such as names, telephone numbers, addresses, and email addresses are masked using an algorithm that prevents course participants from being revealed. Once this initial change is made to the raw text file, the text file is parsed and each message including header information such as the student's masked name, message number, message number to which the current message responds, along with the time and date the message was submitted are sent to the database. One message, including its body and header information, comprises a single record in the database. For clarity's sake, a message is a single entry from an individual participating in an online discussion. This entry is comprised of the student's name, the course number, the message number,

the parent message number, the subject, the date, the instructor's name, and the message body. The only parts students generate are the message subject and body; the remaining information is generated by the course management software. For this analysis, the word "message" refers to the message body since this analysis is primarily focused on text that the student generates. Once the message database has been created, the message body and its header information (subject, name, instructor, date, message number, and parent message number) are ready for computational analysis.

Select Messages Used for Three Parts of the Analysis

Once in a database, messages are selected using a systematic, random sampling technique. Using this technique, three unique message collections are harvested. One message collection comprised of 300 messages is used for coder training. Another collection of 100 messages is used for reliability. Each coder and the trained artificial neural network code the reliability collection, and these codes are compared to determine whether the ANN codes as reliably as a set of humans. Finally, a set of 1,200 messages is harvested. The six coders rate 200 messages each from this collection. Each of the 1,200 messages will be used to train the ANN. This number of messages was chosen because the ANN requires a large number of messages (more than 1,000) in order to be properly trained and because it may have been too much to ask volunteers to code more than 300 messages (200 for the ANN and 100 for reliability) each.

Create the Tool Enabling On-line Message Coding

Finally, an online tool was constructed to allow coders to rate messages online. During training and actual coding, the coders used the tool to rate each message. The

online tool sends each coder's rating to a relational database which correlates the message to its coded value. This allows reliability statistics among coders to be easily calculated both during training and after the coders rate their body of messages. This also shows which coders have or have not completed their message coding tasks.

Human Content Analysis

This work builds on a series of content analyses described by Garrison, Anderson, and Archer (2000, 2001) who analyzed online discussions based on a community of inquiry model which splits community-based learning into three overlapping areas: social presence, cognitive presence, and teacher presence. The details of their work and method are described below. According to Rife, Lacy, and Fico (1998), content analysis is "the systematic assignment of communication content to categories according to rules, and the analysis of relationships involving those categories using statistical methods" (p. 2). Rife, Lacy, and Fico (1998) also outline the steps for performing a quantitative content analysis as defining the units of analysis, operationally defining the construct to be measured, training coders, and taking reliability measures to determine how consistently the coders have measured the construct.

Participants

The pool of participants is comprised of every student who posted a message in any section of history and political science in Georgia's eCore™ program during the Fall, Spring, and Summer terms of the 2000-2001 school year. The eCore™ program is a distance education program administered by the Advanced Learning Technologies Group of the Georgia Board of Regents. This program is designed to offer university-level core

curriculum courses to students who do not live within commuting distance to one of Georgia's state universities. According to Georgia's Advanced Learning Technology Group (ALT Distance Education Student Profile Survey, 2001), the students in eCore™ courses are a mix of traditional and non-traditional students scattered across the state with some outside the state and even outside the country. Approximately two-thirds of the students in 2001 had taken college courses within the past 12 months and about 25% are returning to school after an absence of more than 12 months. Demographically, approximately 70% of those responding to the 2001 survey identified themselves as white and 15% identified themselves as African American. The report also mentions that there have traditionally been three female students for every one male student enrolled, and slightly more than 50% of the students are married. In general, the majority of students work full-time (40 or more hours per week). Table 3 provides the number of students participating in online discussions in the six analyzed courses.

Table 3

Number of Students Contributing Messages in Course

Course Section	Number of Students
History Section 1	31
History Section 2	20
History Section 3	41
Political Science Section 1	23
Political Science Section 2	23
Political Science Section 3	24

Unit of Analysis

Garrison, Anderson, and Archer (2001) chose a syntactic, as opposed to thematic, unit of analysis in that they measure the entire message as opposed to individual paragraphs, sentences, or themes within a message. Further, they use human coders to classify messages, and their study yielded a reliability figure of kappa=0.74. To draw comparisons between this study and that of Garrison, Anderson, and Archer (2001), the same unit of analysis is used.

Operational Definition of Cognitive Presence

This work focuses on cognitive presence, which Garrison, Anderson, and Archer (2001) define as "the extent to which learners are able to construct and confirm meaning through sustained reflection and discourse in a critical community of inquiry" (p. 11). Coding decisions are made using a coding rubric from Garrison, Anderson, and Archer (2001) in which they operationalize each cognitive presence category. Each cognitive presence category is listed below along with a description of the category and a message from the study which exemplifies the cognitive presence category.

1. *Triggering Event*: a message designed to evoke a response (e.g. "In an earlier post, FirstName2 reminded us that their diet was very similar to ours. Do you think the frequency of diet related diseases in their culture was similar that in our culture?")
2. *Exploration*: a message which presents facts, feelings, ideas, suggestions, unsupported conclusions, or unsupported contradictions/disagreement (e.g. "They must have been very angry about the intrusion into their culture.")
3. *Integration*: a message which includes tentative substantiation, combination of ideas or synthesis (e.g. "The settlers must have been less austere than the author proposes.

The archeological evidence taken together with the social events described within the diaries and the town records all point towards the settlers enjoying an active social life.")

4. *Resolution*: a message that indicates commitment to a solution and includes real world applications, testing of solutions or defense of solutions (e.g. "Based on the overwhelming evidence, it is apparent that the author's account of the settlers austerity is incorrect. The settlers definitely had an active social life. This is supported by the following: the remains of several musical instruments have been found at the site. Equipment for making, storing and serving wine and ale have also been found at the site. Letters exist which describe social occasions in significant detail. Town records and diaries also include accounts of parties and social occasions. The evidence of an active social life in the settlement is overwhelming.")

Garrison, Anderson, and Archer (2001) did not include a category of messages which do not fit into any of the above categories. The pilot study revealed that numerous messages which fall out of their cognitive domain; therefore, another category, noncognitive, has been added to the above four.

5. *Noncognitive*: a message which is unrelated to the course topic, addressed course management concerns, requests technical support, or makes an external reference (e.g. "Do you have plans for Friday night?" "When I logged on last night, the server was unavailable. Did anyone else have similar trouble?")

Modified Content Analysis Rubric

In many content analyses, a rubric is used to both train the coders and to guide the coders as they make their rating decisions. The rubric should offer enough guidance to

enable the coders to code each message similarly and therefore should improve inter-rater reliability. For this analysis, the rubric from Garrison, Anderson, and Archer (2000) was modified with examples taken from eCore™ courses in history and political science (see Appendix A).

Coder Training

The content analysis was performed by six coders, each of which has either taught, administered, or taken an online course. Five of the six coders have worked directly with the eCore™ project. The coders were first trained to code online discussion messages using a rubric based on that developed by Garrison, Anderson, and Archer (2000). With the coders scattered throughout the state of Georgia, training was conducted via telephone conferences, email, and a web-based coder training tool. The coder training tool consisted of 300 messages chosen through a systematic random process that included a chronological cross-section of messages from each of the six courses. The coding process occurs in three stages. An initial meeting is scheduled with each coder to describe the project, to introduce the coding rubric, and answer general questions about coding messages using the rubric. Each coder is then asked to code the first 30 messages in the training tool. As the coders complete this set, they are contacted, provided further training based on the results for their first set of messages, and are then trained on coding instances that did not fit with the coding rubric. The coders are asked to iterate through these steps until they reach an average pairwise reliability of kappa = 0.70.

Once training is complete, the raters are asked to code three hundred messages each. Two hundred messages from each course are chosen using a systematic, random

sampling technique. That is, the first coder rates 200 messages from one section of a history course. A second coder rates 200 messages from another history course and so on. Each of these messages is unique. This is done to provide the ANN with a large set of messages on which to train. The coders must also rate a set of 100 messages systematically taken from all courses in the research study. This set of 100 messages is used to calculate inter-rater reliability statistics, and therefore, is the same set of messages for each coder. It took approximately three weeks to train all the coders, and once training was complete, it took another two weeks for all coders to rate their sets of 300 messages.

Reliability

To enable a comparison between this study and that of Garrison, Anderson, and Archer (2000), Cohen's (1960) kappa values are calculated among pairs of raters. Cohen's kappa values may be interpreted in a number of ways, and this work employs both the lenient benchmarks of Landis and Koch (1997) as well as Rife, Lacy, and Fico's (1998) more conservative benchmarks. Landis and Koch (1997) describe reliability figures in Table 4. Riffe, Lacy, and Fico (1998), however, question kappa values below 0.80 but indicate that research which is breaking new ground, a category under which this research clearly fits, often has reliability figures below the 0.80 range. Although Cohen's kappa is widely used, many recommend using multiple reliability measures. For that reason, Shrout and Fleiss' (1979) two-way random effects average measure of reliability model is also to be used as an additional reliability measure. This measure, intraclass correlation (ICC), is frequently used as a measure of inter-rater reliability, or inter-rater agreement. The second of three Shrout and Fleiss (1979) intraclass correlation models

assumes that each rater is a member of a larger subset of potential raters. In this case, each rater is a member of the larger pool of all eCore™ instructors and administrators who could possibly rate student messages. This measure is attractive because it provides a single reliability figure for more than two coders; Cohen's kappa is limited to pairwise reliability statistics.

Table 4

Interpretation of Kappa Values from Landis and Koch (1997)

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 - 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Artificial Neural Network Content Analysis

Armed with 1,200 human-coded messages, the artificial neural network (ANN) is ready to be trained. This section describes the steps taken to train and test the ANN, a series of steps leading to a single model which rates discussion list messages more accurately than the other models. First, two pieces of background information are offered. The first describes the pilot study which led to this work. The second background piece describes how artificial neural networks work. The next section

explains how and why message text is converted into numeric information. Afterwards, I describe the steps taken to arrange the predictor variables and why it is important to arrange predictor variables from those with the greatest predictive discrimination to those with the least. I then describe how the human content analysis decisions were modeled using the ANN software. Next, I describe how I compared three types of models to determine whether a single, generalizable model is the best to use. Finally, I describe how to compare the ANN coding decisions to the human coders.

Pilot Study

This work is an extension of an initial trial to determine the feasibility of using an artificial neural network (ANN) to perform a content analysis of online discussions. The pilot study (McKlin, Harmon, Jones & Evans, 2001) performed two analyses of online discussion messages obtained from a graduate Instructional Technology course on Web-Based Learning. The first analysis was performed to determine whether a neural network could be used under the best of circumstances, by correctly categorizing messages identified as most indicative of each cognitive presence category. This method resulted in a reasonably high reliability figure (kappa = 0.76) indicating that a neural network under the best of circumstances could categorize messages slightly better than humans can categorize messages under normal circumstances.

The second analysis required that the researcher code a systematic random sampling of messages and did not allow the researchers to skip all but the most indicative messages of each category. The introduction of noise generated a less well-performing model (kappa=0.31) but the analysis revealed numerous optimization techniques that could improve the reliability of the method. The present eCore™ analysis extends the

previous study in the following ways: It experiments with the creation of a cross-section/cross-course generalized neural network model and explores the feasibility of a single generic model to analyze multiple courses; it analyzes six courses instead of one; it incorporates optimization avenues discovered during the previous, single-course analysis (e.g. normalizing inputs, including structure information in the model, and including self-defined categories).

Overall, modeling the content analysis decisions of the six human coders involves four steps: transferring a semester of messages into a database for electronic manipulation (mentioned above), numerically describing each message, ordering predictors by their level of importance to the model (the strongest predictors appear first and the weaker predictors appear last), and modeling the content analysis data with ANN software to derive the best model. Given the preliminary indication that an artificial neural network could potentially and reliably categorize messages into cognitive presence categories, a second pilot was not conducted.

How Neural Networks Work

The modeling of human decision-making using artificial neural network (ANN) software does not have a strict set of procedures. Garson (1998) cautions the social scientist interested in using neural networks:

> The backpropagation model is the most common, but neural network analysis is not 'a' technique. There are many, many neural models. One could devote a lifetime to experimenting with the alternatives, optimizing them, and exploring the effects of different parameters. Ultimately, neural modeling is an art form and the social scientist who embraces it is an artist whose work is never finished, or at least, an artisan who is never sure the analysis he or she presents to the public might not be suboptimal. (p. 16 – 17)

The pilot study revealed some guidelines that will most likely apply to this study. First, the backpropagation model performed better than other models and was used in the current analysis. Garson (1998) supports this by saying that the backpropagation model is the standard by which the performance of other models is gauged (p. 41). The following is how backpropagation models work.

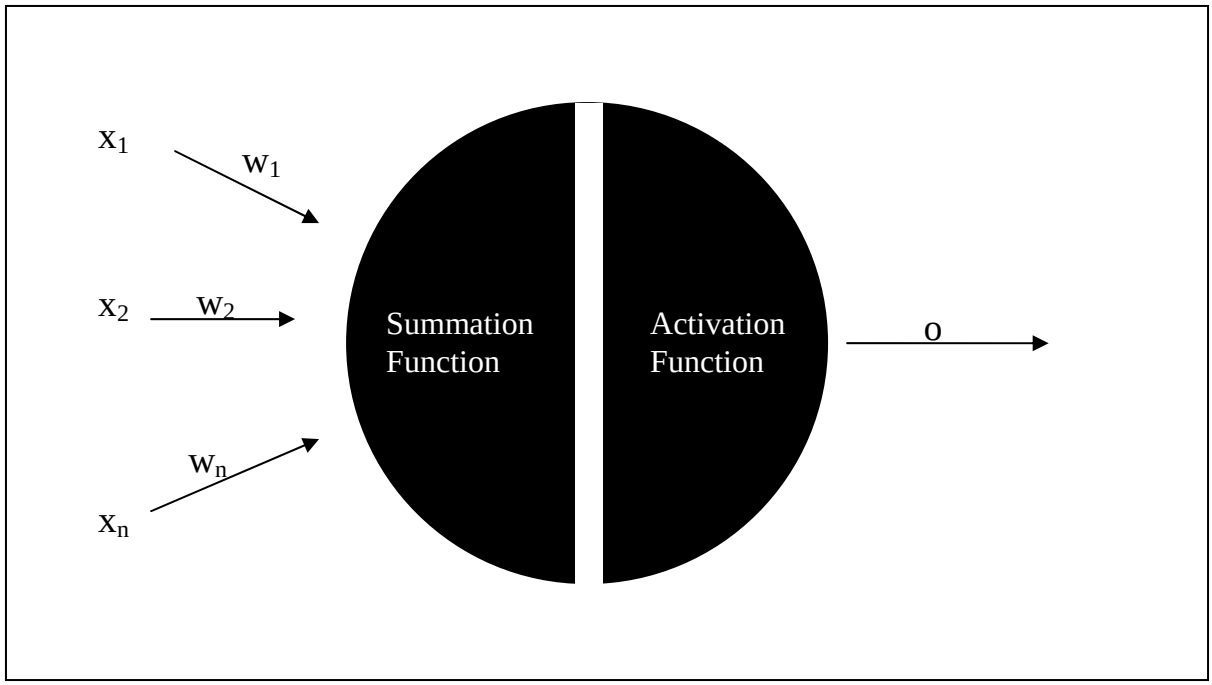


Figure 5. One artificial neuron.

Figure 5 shows a simplified three-layer network adapted from Garson (1998, p. 43). The three layers are the inputs (x_1, x_2, and x_n), the one hidden neuron depicted as a large black circle containing a summation function and an activation function, and a single output (o). Each input is multiplied by a weight (w_1, w_2, and w_n). The neuron performs two tasks. It first sums the products of each input and weight and then uses an activation function. The activation function is a transfer function which calculates the activation level of the neuron. This activation level is then compared to a threshold value. If the

activation value is above the threshold value, the neuron's output is "on" or one; if the activation value is below the threshold value, the neuron's output is "off" or zero.

One important aspect of neural networks is that they learn from a set of training data. In this case, the 1200 messages categorized by each coder serve as the training data. This data correlates a set of inputs, the numeric description of each message, to a set of outputs, one of five cognitive presence categories (see Figure 4). A neural network is trained on a data set containing inputs and outputs by first assigning a random weight between the neurons and then calculating the error which is the difference between the actual and expected results. The neural network software repeats this procedure, and each time it repeats, it adjusts the weights between neurons in an attempt to reduce the error. It continues repeating this process until the hidden neural pattern "fits" the data.

Numerous adjustments may be made to the network to improve modeling. A person may adjust the number of inputs, the number of hidden neurons, the number of hidden layers, and the number of outputs.

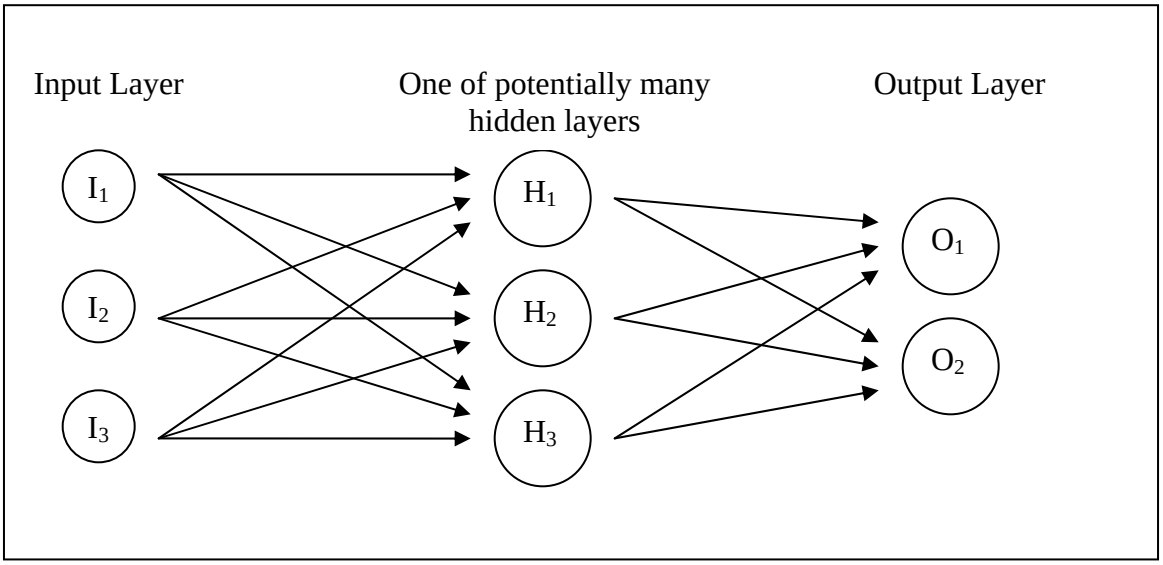


Figure 6. A simple ANN architecture showing layers and nodes.

In Figure 6, there are three hidden neurons (H_1, H_2, and H_3) in one hidden neuron layer. The arrows connecting each neuron have weights which are adjusted during training to reduce error, to reduce the difference between the actual neural network results and the desired, human-coded, results.

At first glance, a person may wish to train the neural network until the error reaches zero. Doing this, however, results in overtraining. The model has memorized the input data and will probably not generalize well to new data. To determine whether the model has been well-trained, a set of test messages is set aside to validate the model. Once a model has been trained, the model categorizes the test set of messages.

Numerous ANN software packages are commercially available. This research used three software packages. For the pilot and parts of this research, Pattern Recognition Workbench® (PRW) from Unica Technologies, Inc was used. PRW was used more out of convenience than ease of use; this software was readily available. However, the software is no longer supported by the manufacturer. For that reason and based on Garson's (1998) recommendation, two software packages from Ward Systems® were used. Neuroshell Classifier is relatively inexpensive and has a limited but simple interface focusing specifically on classification tasks. Ward Systems® Neuroshell 2 is a more expensive and robust software package allowing the user to experiment with different types of ANN models.

Numerically Describe Messages

The current state of ANN technology requires that inputs to the system be numeric. Therefore, researchers are forced to derive a strategy to numerically describe their data; in this case, I was forced to numerically describe each message text. This is

done by using a dictionary of themes and measuring the number of words or phrases in the message body which fall under each theme. The General Inquirer (see the section entitled "Dictionary" below for a fuller description) is used as the primary dictionary of themes and is extended by creating numerous self-defined categories. The messages are further described using a set of predictors to describe the placement of the message within the overall hierarchy of messages.

Dictionary

The General Inquirer is a dictionary comprised of 11,788 words in 182 categories (Danielson & Lasorsa, 1997). Each message is analyzed against each category of terms and a simple word count is taken to determine the weight of each category of terms in each message. For example, the General Inquirer category "positiv" contains the words "up, abide, and yes" meaning that the following sentence will receive a "positiv" score of two: "**YES**, I had to look **UP** to see the icon." For the present study, each message is parsed using the General Inquirer that assigns each message a numeric value for each of its 182 categories.

For this analysis, the General Inquirer has been extended by the addition of several self-defined cognitive presence categories designed to improve classification. Specifically, 37 self-defined categories were added to the list of 100 General Inquirer categories and were developed using the following process. I manually coded approximately 100 messages and reviewed them to determine what linguistic cues discriminate among cognitive presence categories, whether any messages contain words common to a single cognitive presence category but not found in other cognitive presence categories. For example, a triggering event is a message designed to elicit a response.

From this, we understand that triggering events may take the form of a question; therefore, words signaling a question like “who,” “what,” “where,” “when,” and “how” are more likely to appear in triggering events than in other cognitive presence categories like exploration messages. The goal is to derive a set of inputs which numerically describe the message text and which provide the neural network with enough information that it can assign messages to cognitive presence categories just as reliably as a human coder.

Hierarchical Structure

Five structure categories provide a description of the message within the greater structure of messages. Those five structure categories are thread number, width, depth, number child messages, and number of grandchild messages. Identifying where a message lies within the hierarchy of messages may reveal information about the cognitive presence category in which it falls. The set of messages in Figure 7 illustrates the five pieces of information gathered to describe where a message lies in the broader structure of messages.

Message no. 304:
The letter from Cortes was a primary source somewhat exaggerated. Cortes, being a conqueror and explorer, possibly felt that convincing the king that he was well respected and reveared [*sic*], capable of changing their religion and way of life, would make it necessary that the king and Spain would have to go through him to conquer Mexico. Giving him his place in history. Cortes knew of the profitable trade routes with the W.Indies, causing him to exaggerate the economy. This exageration [*sic*] could lead to profitable trade, benefiting not only Spain but himself.

> **Message no. 306: [Branch from no. 304]**
> I do agree with you. Cortes was looking for ways to benefit himself. In his letter he seems so truthful and appreciative of the beautiful sites around him. Somewhere inbetween [*sic*] the lines I feel he is looking out for himself and what he may gain from such a beautiful account of the city and its happenings. This feelings [*sic*] is given through his exageration [*sic*] of how smooth everthing [*sic*] runs.
>
> > **Message no. 307: [Branch from no. 306]**
> > I don't feel as though he were [*sic*] exaggerating at all. The Aztec civilization was trly [*sic*] flourishing in the 16th century. Just because they did not have a religion like that of the European explorers, did not mean that they were a backward people. Prior to the Aztecs, the Mayan people had calendars much more accurate than those in use by Spain, England, Portugal, and others. Their architecture was a sight to behold and it still exists to this day on the Yucatan peninsular [*sic*].
>
> **Message no. 348: [Branch from no. 304]**
> I agree with you that Cortes used this account as a strategy to win over the king. I do not think that it was as easy as he says it was to win over the people under Moctezuma's rule.

Figure 7. One section of the course transcript showing the hierarchical outline of messages.

Figure 7 shows one small section of the course transcript. Here, students are responding to the instructor's question asking whether a letter presumably written by Cortes accurately depicted what Cortes saw in Mexico. There are four messages, and the first message, message 304, is the beginning of this thread. In any given course, there are numerous threads and each new thread may be a new discussion topic. At the beginning of the course, the first few discussion threads generally address students getting to know each other or general classroom management. The final threads of a course usually concern good-byes and well-wishes. The threads in between, however, usually map to

the topics covered in the syllabus. The thread in the example is one of a few threads addressing the accuracy of Cortes' letter. The thread number may indicate what type of message it is. First threads would most likely address course management topics and would be categorized as non-cognitive.

Message width is the horizontal measurement of messages. It tells us whether a message is the first in a discussion thread, a reply, a reply to a reply, and so on. In Figure 7, message 304 has a width of one because it is the first message in this example thread. Message 306 is the first reply to Message 304; its width is two. Message 307 has a width of three because it is a reply to a reply. Message width may provide a clue as to which cognitive presence category a message falls. The first message in a thread is often a triggering event, a message culminating in a question or concern designed to spark further discussion. Replies to a triggering event may often be exploration messages, messages in which students are playing with the question or concern from the triggering event, but offering no substantiated, definitive claims. Message 348 is an example of such a message; the student agrees but offers no further substantiated claims. Presumably, messages with a greater width may be integration messages, messages which address the question or concern from the triggering event with substantiated, supported claims. Message 307 exemplifies an integration message. It addresses the triggering event with series of claims that are most likely supported by the course textbook. This student is thoughtfully putting the pieces of a substantiated argument together.

Message depth is the vertical measurement of messages. It tells us whether a message is a triggering event, the first reply to a triggering event, the second reply, and so

on. Message 306 is the first reply to message 304; therefore, it has a depth of two. Message 348 is the second reply, so it has a depth of three. Like message width, message depth may also provide clues to the cognitive presence category under which a message falls. Messages with a depth of one are usually triggering events. The first few replies to a triggering event are likely to be exploration messages and the later replies are usually made after a student has processed the first few exploration messages; therefore, a student may be more likely to incorporate and synthesize previous thoughts into a more reflective, substantiated message.

Finally, the number of children and grandchildren may provide a clue to the cognitive presence categorization of a message. In Figure 7, message 304 has two children, messages 306 and 348, and one grandchild, message 307. A message which has sparked many replies, many children and grandchildren, may be a triggering event. However, it may also be a very compelling justification which has sparked avid disagreement or agreement.

Overall, five numeric, hierarchical descriptions of the message are added to the list of General Inquirer and self-defined categories. Those descriptions are: thread number, message width, message depth, the number of children belonging to a message, and the number of grandchildren belonging to a message.

Predictor Order

Building ANN models is often a trial and error process in which a number of models are built and the best one is chosen. The strategy employed in this analysis is to construct models with different numbers of predictor variables or inputs. Since the software used for this analysis, Neuroshell Classifier, only accepts a maximum of 150

inputs, a process was put in place to choose the most discriminating 100 General Inquirer categories as predictor variables along with the 37 self-defined categories and the 5 structure categories. This results in 142 inputs into the artificial neural network software.

First, the most discriminating 100 of the 182 General Inquirer categories were selected by training the ANN to model all 1,200 messages. The ANN software package, Pattern Recognition Workbench, was used to construct the model because it accepts more than 150 input variables. Once the model was built, the ANN software associates a discrimination weight to each predictor; the higher the weight, the better the predictor variable is at assigning message input to the correct cognitive presence category. This set of predictors was sorted by each predictor's discrimination weight and the 100 most discriminating were kept.

The next step was to order all the predictor variables. This includes ordering the 100 General Inquirer predictors as well as the 37 self-defined predictors and the five message hierarchy predictors. These predictors, or inputs, were ordered from most to least discriminating using the same technique for selecting the 100 most discriminating General Inquirer predictors. That is, a model was constructed using all 1,200 messages with 142 inputs and five cognitive presence outputs (non-cognitive, triggering, exploration, integration, and resolution). After creating the model, the software produced discrimination weights for each input variable, and the inputs, or predictor variables, were ordered with the most discriminating first and the least discriminating last.

Modeling Content Analysis Decision-Making

At this point, we are ready to begin creating ANN models. We possess a set of training data, 1,200 human-coded messages. We have numerically described that data so

that it can be read by the ANN software, and the inputs have been ordered so that the most discriminating inputs are first. The data are also separated into a set of inputs, 142 predictors, and outputs, one of five cognitive presence categories. The data are fed into the ANN software. The software then asks that a set of test messages be separated from the set of 1,200 messages in order to test its accuracy and provides the capability for doing so. The software then asks the user to define the inputs and outputs. Following Garson's (1998) advice earlier in this chapter, a backpropagation ANN model was chosen. The ANN software was instructed to train itself on the data provided, and the first model with 142 inputs and 5 outputs was generated. It takes between 5 and 20 minutes for the ANN software to generate a model. Once the model is complete, a report is provided showing how well the model categorizes the 100 messages set aside for testing. If the model performs well, usually meaning that it correctly classifies more than 70% of the test set correctly, then the model is tested against the set of 100 reliability messages. If it categorizes that set of messages well, again usually more than 70% correctly classified, the model is retained and Cohen's (1960) kappa is calculated between the ANN model and the aggregate of human coders.

After the first model with 142 inputs and 5 outputs is built, tested, and retained if necessary, the second model is created. The second model uses ten fewer inputs. Recall that the input categories are ordered from most discriminating to least discriminating. The ten inputs removed from the model are the last ones, the ones which discriminate less well than the others. A model of 132 inputs and 5 outputs is then constructed and the steps to test that model are repeated. This entire process is repeated 13 times until the final model with just 22 inputs is constructed, tested, and retained if necessary.

The above series of steps is applied to create a single ANN model trained on messages from both course topics, history and political science, and from all six sections, three history sections and three political science sections. This model is referred to as the full model because it incorporates messages from both topics and all sections. The series of steps used to determine the best full model, is also applied to determine the best topic model, one model for history and one for political science, and the best section model, one model for each of the six courses analyzed. A fuller description of this process is provided in the next section, "Comparison of Models."

Comparison of Models

Before moving on to look at sample analyses, the accuracy of the full model was tested against that of the topic and section models. This exercise shows whether we may continue along the most efficient path of creating one model from both history and political science or whether separate topic or section models should be developed. There is reason to believe that the model is defined by linguistic cues not specific to any single topic; for example, the linguistic cues that predict that a message belongs to the exploration category are the same cues no matter what the topic. To verify this, a brief comparison of models is performed. The full model is an ANN constructed from the 1,200 human-coded messages from all six sections of history and political science. The topic models are two models constructed from 600 messages in each topic, history and political science. The section models are six models built from 200 messages in each of the six sections, three history sections and three political science sections.

In order to build each model, a test set and a training set are created. The ANN model is built using the training set and tested for accuracy using the test set. For the full

model containing 1,200 coded messages, a model is created using a training set of 1,100 messages and a test set of 100 messages. To create the topic models, the 600 messages in each topic are divided into a training set of 500 messages and a test set of 100 messages. For the section models, the 200 human-coded messages from each section are divided into a training set of 150 messages and a test set of 50 messages. Most importantly, each model is compared on its training set reliability.

For each model in this analysis, a systematic process for deriving the best model in each comparison category was used. This process required setting the model parameters, defining a training set of messages and a test set of messages, building the model, then testing the model to determine what percentage of test set messages were correctly categorized. Thirteen models were created for each comparison category. That is thirteen models were created using the full set of 1,200 messages and the best of these models was kept. Thirteen models were then created using the 600 history messages, and the best of these models was kept; thirteen models were created using the 600 political science messages, and the best of these models was kept. The method for choosing thirteen models is as follows. For each message, there are 142 predictors (see "Predictor Order"). The first of thirteen models uses all 142 predictors; the second model uses the top 132 predictors; the third uses the top 122 predictors; the fourth uses the top 112 predictors; and so on to the thirteenth model containing only 22 predictors.

Spelling Analysis

Since the ANN model is build from the presence or absence of words in a discussion list message, it is imperative that the algorithm which numerically describes the message sufficiently recognize words. This means that misspellings may threaten the

ANN's ability to correctly classify messages. To determine the effect spelling has on the ANN model, a sample of 100 messages was selected from the message set, corrected for spelling errors, coded by the automatic content analysis tool, and compared to the original set of 100 messages. The hypothesis is that correcting spelling errors will not change how the messages are categorized.

A systematic, random sample of 100 messages was chosen from the 8 courses (four history and four political science) used in the second experiment of this study. The model from the second experiment was chosen because it is more robust than the first experiment's model and would therefore be more sensitive to spelling errors. To ensure consistency in detecting spelling errors, the text of each message body was placed in Microsoft Word©. Word automatically identifies the spelling errors and those errors are corrected until Word identifies no further spelling errors. This step is repeated for all 100 messages. The spell-corrected set of messages is then placed into the database and the method used for numerically describing messages is applied. Once the messages are numerically described, the ANN algorithm is applied to each message in order to categorize that message into cognitive presence categories. Again, the hypothesis is that the ANN model will place correctly-spelled messages into the same category as their misspelled counterparts.

For this analysis, an error is defined as any word misspelled so that it is unrecognizable by the General Inquirer and self-defined dictionaries. Grammatical errors or errors in word choice do not count. Grammatical errors are unaffected by the dictionary. For example, the dictionary does not care that "i" is lower-case; it is still recognized as a first-person pronoun. Errors in word choice do not count because it is not

possible to discern the intent of the author. Though it may be a word choice error, changing the word for this study may bring about a meaning that the author did not intend thereby creating a greater error than the word choice error. Some error will be introduced into the model in the form of homonyms (e.g. “to,” “too,” and “two”).

Comparing Human and ANN Coding Decisions

To ultimately answer the first part of the research question which hypothesizes, “An artificial neural network (ANN) analyzes messages as well as a human,” human coding decisions must be compared to ANN coding decisions. Six human coders coded the same set of 100 messages and reliability scores for the human group were calculated. However, we must now determine the reliability between the set of human coders and the artificial neural network. To do this, an aggregate categorization of all six human coders was first derived. Table 5 illustrates the process for gleaning the aggregate categorization.

Table 5

Decision Logic for Aggregating Human-Coded Messages

Example number	Coder Decision A	B	C	D	E	F	Aggregate Decision	Action
1.	2	2	2	2	2	2	2	Use mode
2.	0	0	1	1	1	1	1	Use mode
3.	0	1	2	3	4	4	4	Use mode
4.	0	0	1	1	2	2	1	Use mode first, then mean.
5.	0	0	3	3	4	4	3	Use mode first, then mean, and select the choice closest to the mean.
6.	1	1	1	2	2	2	1	Score of least reliable coder is thrown out and the mode is used to make the decision

First, the messages have been coded into one of five categories in which 0 is noncognitive, 1 is a triggering event, 2 is an exploration message, 3 is an integration message, and 4 is a resolution message. Each of the six coders is represented by a letter where A is the first coder, B is the second, and so on. The following decision logic was used. The mode of all raters is used first in order to give the greatest weight to the decisions made by each individual. The intent is for the artificial neural network to

model the decision-making of all coders. In the first three examples above, the mode is sufficient for making the aggregate decision. However, if there is a tie among the coders, as in examples four, five, and six, the mean should be used to break the tie. In example four, there is a tie among three possible modes; therefore, the mean, one, is employed to break the tie. Garrison, Anderson, and Archer (2001) lead us to believe that the cognitive presence categories are situated along a continuum. They refer to them as “phases” (p. 10) and suggest that online discourse progresses through each phase beginning with a triggering event and culminating in resolution. For this reason, the mean score of the six coders is used as a tie-breaker if the mode fails. Looking again to example four, some feel the message is higher along the continuum than others. The mean is chosen because it incorporates each coder’s decision into the reliability score and centers that group decision. However, in those situations like example five in which there is a three-way tie and the mean is a category number that no coder has chosen, then the mode nearest the mean should be selected. Again, this tactic more fully incorporates each coder’s decision into the reliability score. Finally, in the rare event that neither the mode nor the mean is sufficient to identify an aggregate coding, then one coder’s ranking should be discarded in order to break the tie. Selecting the one coder to remove is based on each coder’s average pairwise reliability score. The coder with the lowest average pairwise reliability should be removed from the decision on that single item. Example six above assumes that coder F’s average pairwise reliability is lower than all the others; therefore, F’s score is removed and the mode is used.

Sample ANN Analyses

Once an acceptable artificial neural network model has been constructed, that model may be run against the entire body of messages returning a cognitive presence value for every message in every analyzed course. A number of analyses immediately surface from the data, and the following is a sample of the analyses we should expect once each message has been assigned a cognitive presence weight.

Descriptive Analyses

Mean Cognitive Presence Weights

For a given body of messages such as those from one course or from a number of courses by topic, a mean cognitive presence weight can be derived. This weight shows the overall cognitive presence, or intellectual effort, exerted by the course participants. The mean cognitive presence weight is an average of messages whose cognitive presence value falls along a continuum between zero and four as follows:

0. Non-cognitive
1. Triggering Event
2. Exploration
3. Integration
4. Resolution

Figures 8 and 9 exemplify mean cognitive presence weights for a body of messages. Figure 8 shows the cognitive presence weight by course allowing for the comparison of instructors and Figure 9 shows the cognitive presence weight by course topic allowing, in this case, the overall cognitive presence in history to be compared to the overall cognitive presence in political science.

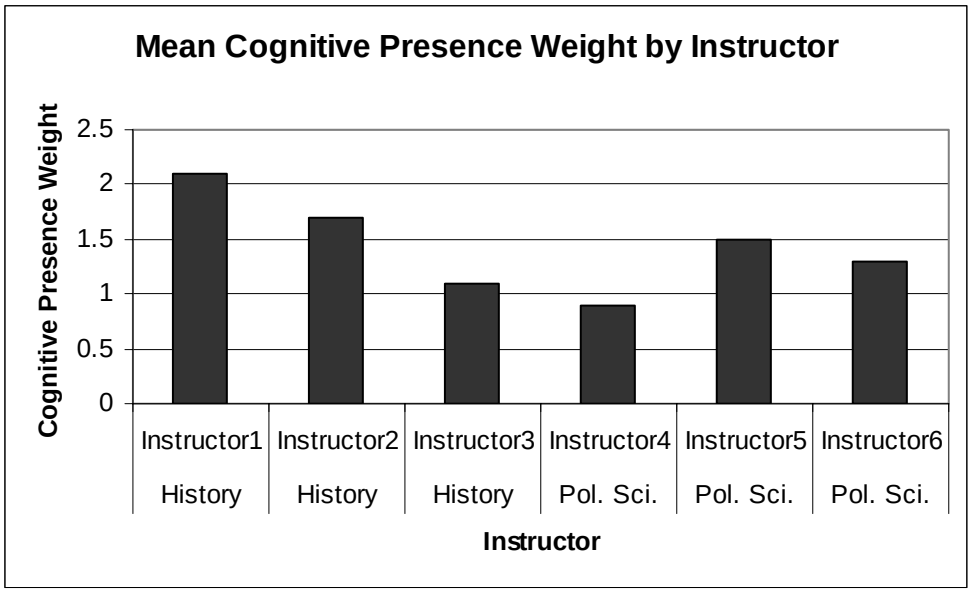


Figure 8. Mean cognitive presence weight by instructor.

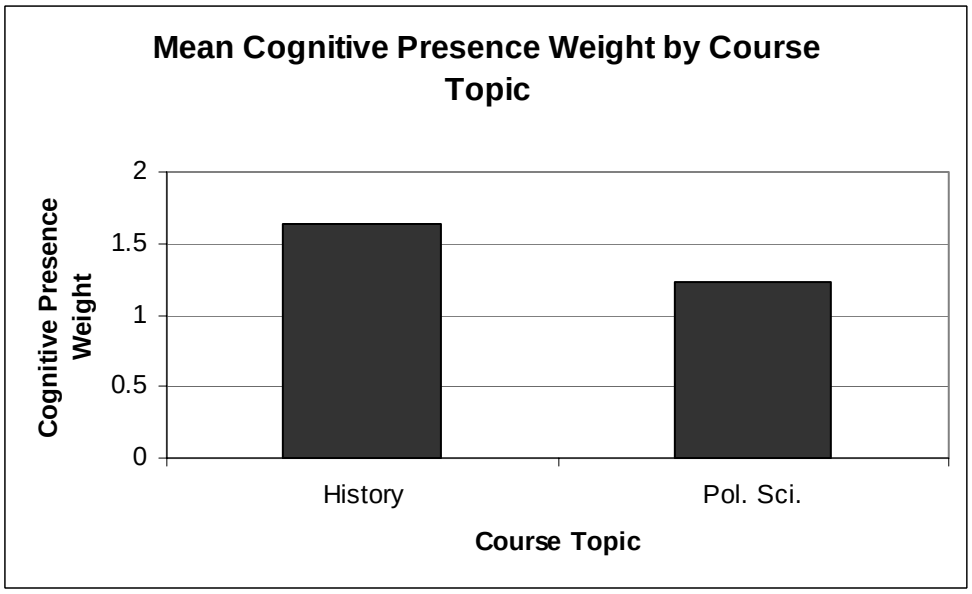


Figure 9. Mean Cognitive Presence Weight by Course Topic

Occurrences of Each Cognitive Presence Category

The above analysis provides a broad-level view of each student's cognitive output, but does not indicate how cognitive presence is distributed. That is, the above analysis does not answer the question, "how many messages were non-cognitive?" For

that analysis, a percentage of the occurrence of each cognitive presence category may be generated. Figures 10 and 11 provide the percentage of messages falling into each cognitive presence category by course section and topic respectively. From this example, we see the reason why Instructor 1 had the largest overall cognitive presence; this instructor's class generated far more integration messages than the others. Further, we can also compare course topics (history and political science) along each cognitive presence category allowing us to see, for example, which topic generates more triggering events and exploration (see Figure 11).

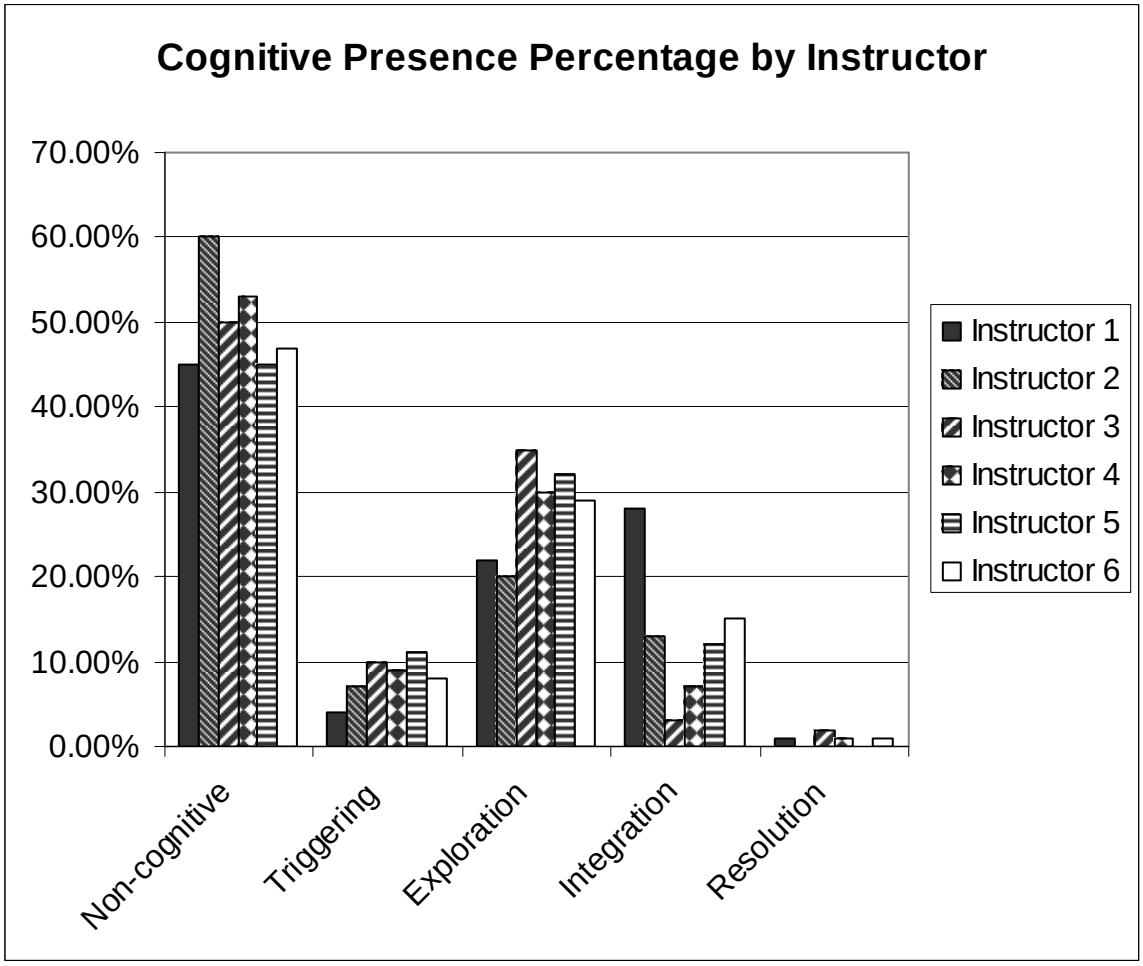


Figure 10. Cognitive Presence Percentage by Instructor

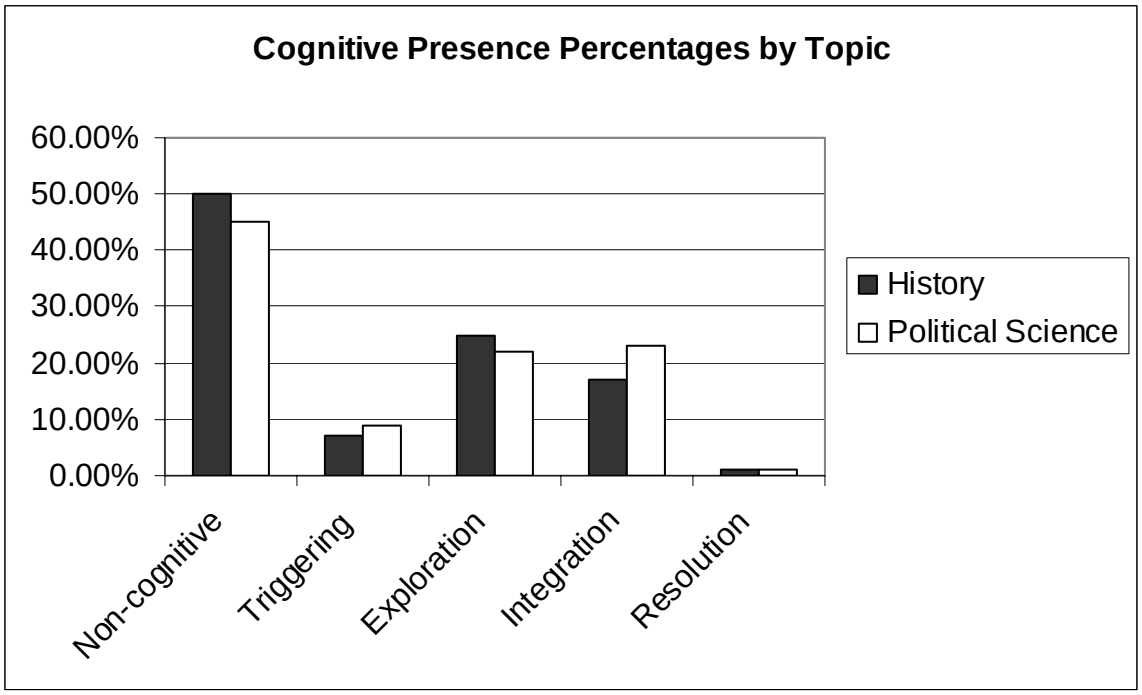


Figure 11. Cognitive Presence Percentages by Topic

Analyses Based on Course Section Variables

Aside from discerning global information about a single course section or a topic, fine-grained analyses related to a single course are also revealed as a result of a cognitive presence value associated with each message. Two examples are shown below; the first is an analysis of each student's performance, and the second views the cognitive presence from each week of a course. Table 6 shows the mean cognitive presence weight of each student over the course of one term along with the total number of messages each student contributed and the number of messages falling into each cognitive presence category. The following equation is used to assign mean cognitive presence weights to students

$$\frac{(MC_0 \text{ x } W_0) + (MC_1 \text{ x } W_1) + (MC_2 \text{ x } W_2) + (MC_3 \text{ x } W_3) + (MC_4 \text{ x } W_4)}{MT} \qquad (1)$$

where MC is the number of messages from each cognitive presence category, W is the weight of each category (in which noncognitive = 0, triggering event = 1, exploration = 2, integration = 3, and resolution = 4), and MT is the total number of messages the student generated.

Table 6

Sample Cognitive Presence Values by Student

Student	Cog weight	0	1	2	3	4	Total
FirstName1 LastName1	4.00	0	0	0	0	1	1
FirstName2 LastName2	0.67	2	0	1	0	0	3
FirstName3 LastName3	0.75	78	7	18	15	0	118
FirstName4 LastName4	1.56	25	4	28	20	0	77
FirstName5 LastName5	1.73	7	1	10	8	0	26

Note: This shows sample cognitive presence values by student in which 0 represents noncognitive messages, 1 represents triggering events, 2 represents exploration messages, 3 represents integration messages, and 4 represents resolution messages.

The cognitive presence weight in Table 6 is derived using Equation 1. The range of cognitive presence values a student may receive is the same as the range of cognitive presence values, zero to four. In the above example, it becomes clear that a student may receive a high overall cognitive presence value even though that student submitted only one message. This indicates that a true measure of course-long cognitive effort must incorporate the number of messages a student contributed.

The second course section variable is topic. In this case, we may assume that a topic is given each week in the course, and we may see how the level of cognitive presence is distributed as the course progresses. We may assume that the first and last weeks will contain relatively low cognitive presence since little content is discussed during those times. Figure 12 depicts the type of analysis an instructor would receive showing the cognitive presence values for each week of the semester.

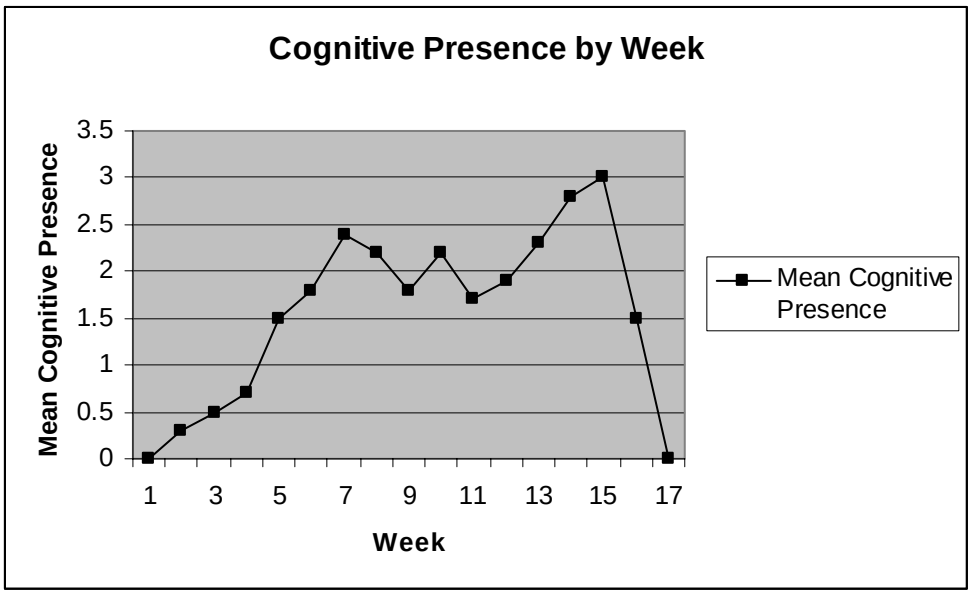


Figure 12. Cognitive presence by week.

Beyond Descriptive Analyses

The above descriptive analyses immediately surface as a result of associating cognitive presence values with each message; however, the researcher or instructor with statistical experience may perform a number of more rigorous statistical analyses in which the cognitive presence weight serves as the dependent variable. Specifically, the instructor may perform analyses of variance (ANOVA) or regression analyses holding the cognitive presence weight as the dependent variable and the following as independent variables:

1. Instructor: The researcher may compare multiple instructors across multiple topics.
2. Student: The researcher may compare various demographic groups or treatment groups.
3. Message length: The researcher may ascertain whether message length correlates with cognitive presence value.
4. Course Topic: The researcher may determine whether the topic explains a portion of the variability in cognitive presence values.
5. Number of messages: The researcher may question whether the amount of student participation as displayed by the number of messages a student posts contributes to increases or decreases in cognitive presence.
6. Instructor participation: The researcher may also determine whether the amount of participation by an instructor across various course topics makes a difference in the overall cognitive presence elicited by the students.

Overall, associating a cognitive presence weight to each message in an online educational discussion provides not only a wealth of descriptive statistics but may also provide a dependent variable enabling even richer analyses.

Limitations and Bias

Figure 4 at the beginning of this chapter outlines the artificial neural network and human content analysis procedures. This graphical outline also shows the points within the model where limitations may occur. I'll begin the description of the limitations by following the model from the message input into the artificial neural network content analysis, through the human content analysis, and finally to the five output categories.

First, the message itself is limited in that we are using it as a unit of analysis as opposed to a theme-based unit of analysis. The limitation is that the entire message is being categorized into one cognitive presence category. It is true that a message may contain aspects of multiple categories, it may serve as a triggering event and an exploration message within the posting. This is addressed at the coder level in that each coder is asked to code the message based on how that coder would respond to the message. Here we are relying on the latent projections (see "Cognitive Structures" above for a discussion of Potter & Levine-Donnerstein's (1999) latent projections) each coder brings to make the appropriate coding decisions.

Another set of limitations appears in converting message text to a numeric description. The most obvious limitation is that this stage does not feed the message into the decision-making artificial neural network; instead, it feeds an array of numbers describing that message. In defense of the artificial neural network, it may be making decisions based on more identifiable information about each message than human coders

do. Even if the ANN is categorizing messages based on only the best 40 discriminators, it would be difficult for a set of human coders to identify 40 factors contributing to the decisions they make when categorizing messages. Also, within the numeric description of the model is a dictionary analysis. One limitation to the dictionary analysis is that the numeric description of each message becomes an array of weighted themes. For example, the message as described as having a large amount of "positiv" or a low amount of "strong." This means that the meaning underlying the message is replaced by the prevalence or absence of a series of themes. One response to this limitation is the argument that if the ANN reliably makes the same decisions as humans, then it does not matter that humans and the ANN have arrived at those decisions through separate means.

Another limitation to the dictionary analysis is that the array of weighted themes describing each message is based on correctly spelled words. This means the misspellings may inadvertently skew the numeric description of each message. A spelling analysis was conducted in which a sample of messages were corrected and re-analyzed using the full ANN model. Presumably, if spelling threatens the accuracy of the model, then the model would code correctly spelled messages differently than their misspelled counterparts. A fuller, but much more time-consuming study could be done in which a model is constructed using only correctly-spelled messages and then compared to a model constructed from original, misspelled messages. Further, neural networks are designed to work well with fuzzy, incomplete, and noisy data.

Further, hierarchical structure is a limitation only in that this information may not help discriminating one category of messages (e.g. triggering event) from another (e.g. exploration). This further becomes problematic in generalizing from one course to

another. Some instructors may pay closer attention to ensuring that each discussion topic has its own message thread whereas others may not. This means that a model constructed from messages in which the instructors adhere to strict message structures may perform less well with a new set of messages from courses lacking strict structure. The best response to this limitation occurs in the implementation of the ANN content analysis tool. The best implementation of such a tool would be to allow the instructor to modify the ANN model much like a user of speech recognition technology "trains" the speech recognition software to recognize the user's voice. This implementation would allow each instructor to train the ANN to more accurately categorize his/her own set of messages.

Another set of limitations surrounds the trained and tested artificial neural network (ANN) model. Garson (1998) explains the largest limitation of ANN modeling:

> It can be difficult to understand how neural nets arrive at their results. Systems designed thus far do not include the capacity of alternative techniques like expert systems to provide an audit trail fully explaining how the system arrived at its conclusions…. While there are approaches to causal analysis using neural models, it is still fair to state that the social scientist's core concern with explication, not simple prediction, has been the primary reason why neural models have not spread more than they have. (p. 16)

Garson (1998) responds to this limitation by saying that neural networks not only outperform statistical approaches but are robust under conditions in which the input data are "noisy, nonlinear, and with missing measurements" (p. 162).

In the model outlined in Figure 4, the artificial neural network (ANN) models the decision-making of six human coders and each coder brings his/her own bias to the coding decision. Three measures have been employed to reduce the bias inherent in human decision-making. First, all six coders have either taught, taken, or administered

distance learning courses meaning that they have some understanding of online learning. Second, each coder has been trained through numerous coder training sessions to set their biases aside and to categorize messages reliably. Finally, the coders rely on a rubric to guide their decision-making and they are trained to consult the rubric in all message categorizing decisions. Just as bias exists within the coders, bias may also exist within the rubric itself. Though the rubric is derived from previous online discussion list coding rubrics, the rubric may still contain culture bias. It treats higher-order thinking, as exemplified in integration and resolution categories, as those messages displaying justified knowledge claims and lower-order thinking as brainstorming and personal narrative.

Certainly, the model is also limited by the number of output categories comprising cognitive presence which is understood as the amount of intellectual effort exemplified by a single posting. Intellectual effort is far more complex than the five output categories imply. The output categories, therefore, should be viewed as broad categories of cognitive presence under which lies deeper complexity. This limitation is brought about in part by the limitations of content analysis. In order to perform a reliable content analysis, the number of coding categories must be limited to a bare minimum. The presence of more categories increases the complexity of the coders' tasks and reduces reliability.

Further, at least one limitation has been revealed by projecting the types of analyses an automatic content analysis tool would create. That is, a student may receive a high cognitive presence value even if that student only submitted one resolution message the entire term. Therefore, when this tool is implemented, cognitive presence

weights assigned to each student must be reported alongside the number of messages each student generated. Further, such a report should also detail each student's number of messages assigned to each cognitive presence category (see Table 6 for an example).

Summary

Overall, the research question, "how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans," has two parts. The first part hypothesizes, that an artificial neural network (ANN) analyzes messages as well as a human. The second part describes the information expected from an ANN content analysis tool.

Two methods are used to answer the first question. First, a human content analysis is performed and reliability statistics are calculated among coders. Second, an artificial neural network (ANN) is built from the set of coded messages, and the reliability between the set of human coders and the ANN is calculated to determine how well the ANN model performs. The ANN is applied to all messages, and a series of sample analyses answers the second research question. Those analyses include descriptive analyses that compare cognitive presence values by cognitive presence category and by course section variables and statistical analyses comparing means by variables such as instructor, student, and course topic.

CHAPTER 4

RESULTS

This section describes the results to the methods used to answer the research question, “how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans?” This question has two parts. The first part hypothesizes, that an artificial neural network (ANN) analyzes messages as well as a human. The second part describes the information expected from an ANN content analysis tool. This chapter is divided into two sections, one addressing the first part of the research question and one addressing the second part. The first section is further divided into two subsections, one for each iteration of the method, experiment one and experiment two. The following graphic overview (Figure 13) of the research method guides the presentation of the results.

Comparing Artificial Neural Networks to Humans

The method used to answer the hypothesis that an ANN analyzes messages with the same accuracy as a human was performed twice with two separate groups of coders, and the results from both experiments are presented in the sections “First Experiment” and “Second Experiment” below. Lessons learned from the first experiment are outlined in the section of this chapter entitled, “Modifications to the Human Content Analysis.” Those lessons are applied in the second experiment with the intent of improving the inter-

rater reliability among the group of human coders and thereby improving the ANN model.

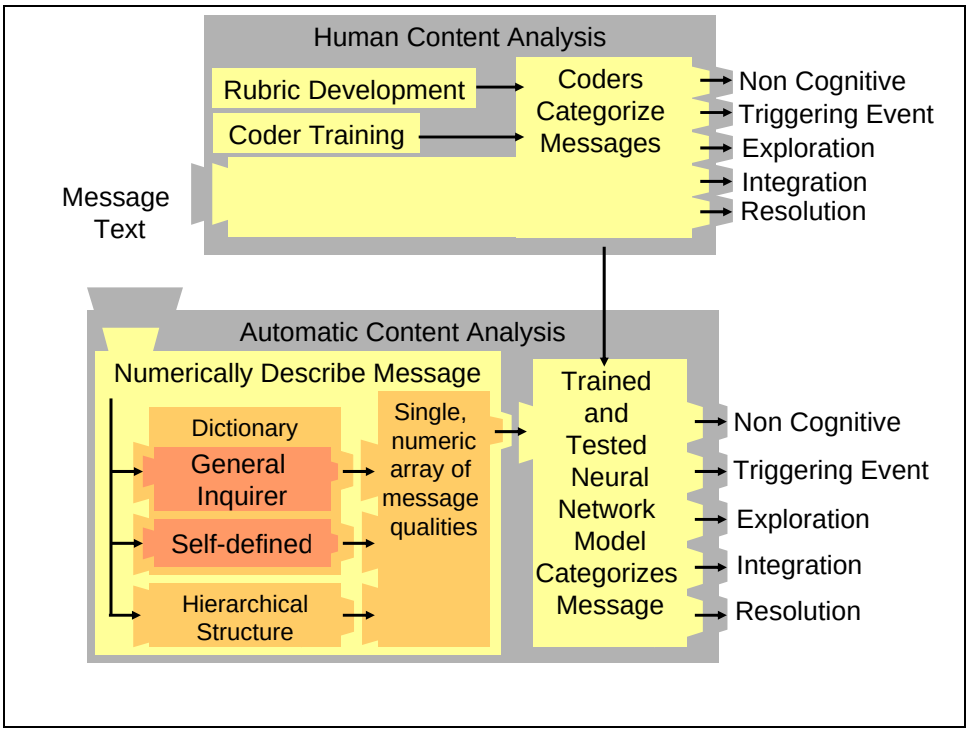


Figure 13. Graphic overview of the research methods.

First Experiment

The first part of the research question asks whether an automatic content analysis tool can categorize messages as well as a set of human content analysts. Stated in measurable terms, it asks whether an automatic content analysis tool codes messages with a Cohen's kappa value comparable to the mean pairwise kappa values of the human coders. This question is answered in two parts. The first part determines how reliably a group of humans categorize messages, and the second part compares a trained artificial neural network (ANN) model to the humans. To do this, Cohen's kappa is used as a common reliability measure among the human coders and between the human coders and

the ANN model. Recall from Chapter 3 that the human coders have been trained to use the cognitive presence coding rubric outlined by Garrison, Anderson, and Archer (2000). During that analysis, each of the six human coders was asked to code the same 100 messages chosen using a systematic random sampling technique from six eCore™ courses. As for the second part, the trained ANN coded the same 100 messages the humans coded, and the reliability scores among the humans is compared to the reliability score between the group of humans and the ANN. Looking at Figure 13, this means that the reliability score is derived by comparing how similarly the two large blocks, human content analysis and automatic content analysis, categorize messages. Overall, to answer the first research question, we must know how reliably each human coder rated messages compared to the other human coders and then how reliably the artificial neural network coded messages compared to the group of human coders. The next section, "Human Content Analysis," describes how reliably each human coder rated messages compared to the other human coders.

Human Content Analysis

This section describes how reliably each coder rated messages compared to the other human coders. This section describes the message sample used in this comparison and then compares the coders' performance. A systematic, random sample of messages was chosen from discussions held in six courses: three sections of history and three sections of political science. Table 7 shows the total number of messages in each course by section and topic. Interestingly, the history sections consistently generated over twice as many messages as the political science sections.

Table 7

Number of Messages by Course Section and Topic

Section	History	Political Science
Section 1	1,564	688
Section 2	1,060	429
Section 3	1,619	677
Total	4,243	1,794

Note: Total number of messages by topic.

After three training sessions, the coders reached pairwise Cohen's kappa reliability scores over 0.70, indicating that training could cease and that they were prepared to code messages. The six coders were then asked to code 300 messages. Of those 300 messages, 100 messages rated by each coder were used for a reliability comparison, and 200 from each coder (1,200 in all) were used to train the ANN. Table 8 shows the pairwise reliability of the six human coders and ANN using Cohen's kappa above the diagonal and percentage agreement below the diagonal. The mean Cohen's kappa among the human coders is 0.6, the average percent agreement is 74%, and coder A is the lowest performing coder (kappa=0.566). A human-to-ANN comparison is provided in the next section. The reliability scores for this set are lower than the scores reached during the last training session, and possible reasons for that are outlined in the section headed "Modifications to the Human Content Analysis" below. Rife, Lacy, and Fico (1998), accept kappa values below 0.80 for content analyses breaking new ground, a category into which this research clearly fits. Landis and Koch (1997) offer a less

conservative interpretation of kappa values by considering those between 0.41 and 0.60 as moderate and those between 0.61 and 0.80 as substantial agreement.

Table 8

Agreement and Kappa Scores for Each Coder

	A	B	C	D	E	F	ANN
A	-	.535	.559	.524	.532	.680	.509
B	70%	-	.742	.682	.617	.531	.504
C	72%	84%	-	.714	.600	.586	.446
D	69%	80%	82%	-	.645	.563	.490
E	69%	75%	74%	77%	-	.494	.426
F	79%	69%	73%	71%	66%	-	.541
ANN	70%	70%	67%	69%	64%	71%	-

Note: Percentage agreement scores are below the diagonal and kappa values are above the diagonal comparing each coder to the others.

Automatic Content Analysis

To address the hypothesis that an ANN categorizes messages as accurately as a set of humans, we must compare ANN reliability scores to human reliability scores. The group of human coders generated reliability scores between 0.494 and 0.742 with a mean score of 0.6. Therefore, a model coding within that range would be acceptable, and an ANN reliability value close to 0.6 would be desirable.

To briefly recap, the ANN model was constructed from 1,200 messages coded by the group of six human coders. Each coder rated 300 messages; 200 of

those are used to train the ANN whereas 100 are used to calculate reliability. Three coders rated messages from history courses and three coders rated messages from political science courses.

During the creation of the ANN, 1,100 messages are used to train the model and 100 messages are randomly reserved for testing. Numerous models were built (see the section of Chapter 3 entitled “Comparison of Models”), and the best of these yielded a percent agreement of 71% and a kappa value of 0.519, within the range of human pairwise kappa values extending from 0.494 and 0.742. A full description of the ANN settings is provided in Appendix C.

At first glance, the ANN-to-human reliability value (kappa = 0.519) appears comparable to the range of human kappa values (kappa range = 0.494 to 0.742), but this is tempered by the fact that the human reliability scores are low and that the ANN coded messages unlike human coders would have. Table 9 compares the messages coded by humans to the messages coded by the ANN. This table shows human coding decisions from top to bottom and ANN coding decisions from left to right. Numbers along the diagonal represent the number of messages in which both the humans and the ANN agree. Numbers off the diagonal show disagreement and indicate how the model may be flawed. For example, the human coders rated 51 (45 + 5 + 1) messages as non-cognitive, and the ANN agreed with the humans on 45 of those 51 ratings. Table 9 shows that the ANN model over-generalizes on non-cognitive, exploration, and integration categories and under-generalizes on triggering events and resolution messages. That is, the ANN failed to code messages into those categories for which there were few human-coded

messages. This results in a model that only codes into three of the five cognitive presence categories.

Table 9

Comparison of Coding Decisions for the First Experiment

		Desired (Aggregate of Human Coders)				
		Non-Cognitive	Triggering Event	Exploration	Integration	Resolution
Predicted (ANN)	Non-Cognitive	**45**	4	7	1	0
	Triggering Event	0	**0**	0	0	0
	Exploration	5	4	**19**	2	0
	Integration	1	0	5	**7**	0
	Resolution	0	0	0	0	**0**

Note. This comparison of coding decisions for the full artificial neural network model yields a kappa of 0.519.

These findings point to a set of modifications to make to the ANN. The most fundamental modification is that the ANN training set must be improved. The results above show that low reliability among the coders may result in conflicting decision logic within the ANN. Therefore, to improve the ANN's training set, we must first improve the reliability among the human coders. These modifications are outlined in the section below entitled "Modifications to the Human Content Analysis."

Modifications to the Human Content Analysis

The research method described in Chapter 3 was conducted, and based on the results from that experiment, the research method was modified, and a second experiment was conducted. This section outlines the modifications made to the research method between the first and second experiment. Specifically, this section describes modifications to the human content analysis in an effort to improve the reliability among the human coders thereby providing a better training set for the ANN.

Results from the first experiment reveal a relatively low agreement among the coders. Since the ANN model is built from the logic of the human coders, there is reason to believe that the accuracy of the ANN model would be increased by improving the reliability among the human coders. This was done by improving the coding rubric and the coder training. That is, improvements in the human content analysis are necessary before improvement in the automatic content analysis can be realized, and those modifications are outlined below.

First, the number of coding categories was reduced from five to four. Garrison, Anderson, and Archer's (2000) resolution category contains very few messages in their own research, in the pilot study, and in the first experiment of this study. In general, less than four percent of all messages were coded as resolution messages. Further, the Garrison, Anderson, and Archer (2001) study mentions that the course topic and facilitation medium do not lend themselves to the type of real-world hypothesis-testing required of this category (p. 6). Empirical, concrete hypothesis testing is difficult, perhaps impossible, to achieve in electronic discussion forums on history and political science. Finally, Garrison, Anderson, and Archer (2001) refer to the categories within

the cognitive presence domain as sequential, “the idealized, logical sequence of the process of critical inquiry” (p. 2). Taken as a sequence, the resolution category is the closest neighbor to the integration category. For these reasons, the resolution category was removed, and any messages which would have fallen into this category were placed into the integration category, the sequentially closest category to the resolution category.

Second, a systematic process for modifying and clarifying the rubric was used. Two coders went through three rounds of coding messages from the first experiment set and made modifications to the rubric based on their disagreements. The most significant clarification of the rubric occurred between the exploration and integration category. In Garrison, Anderson, and Archer’s (2001) rubric, disagreement or divergence fell within the exploration category while agreement or convergence fell within the integration category (p. 10-11). Going back to Figure 7, message 307 disagrees with the message to which it replies, but it does so by offering plenty of justification. Under the rubric from experiment one, coders would be confused. They would be unsure whether it should be an exploration message because it shows disagreement or whether it should be an integration message because it presents a justified claim. The rubric was changed so that agreement or disagreement without justification is placed in the exploration category while agreement or disagreement with justification is placed in the integration category. Besides collapsing integration and resolution messages into one category, exploration was made distinct from integration messages by defining exploration messages as falling along the lower order thinking skills in Bloom’s taxonomy (Bloom and Krathwohl, 1956), and integration messages were tied to Bloom’s higher order thinking skills (analysis, synthesis, and evaluation). This change was primarily made during training

because each of the coders was familiar with Bloom's taxonomy and understood dividing messages along these lines.

Third, another way to improve coding accuracy is to construct the ANN from good examples. Chi (1997) recommends removing difficult-to-code items from the final results. In the second experiment, coders were asked to code all messages but indicate which messages were difficult to code. This modification allows the ANN model to be constructed from the best examples of each category thereby removing some of the noise within the model.

Fourth, not only was each coder asked to code more messages in the second experiment, but each coder was also asked to code messages from both history and political science. In the first experiment, each coder was asked to code messages from one section of one topic. Overall, each coder rated 500 messages, 200 from political science, 200 from history, and the same 100 messages from both history and political science used for reliability testing.

Fifth, in the first experiment, messages were selected using a systematic, random sampling technique. This meant that the coders had little or no context for each message making it difficult to accurately categorize messages especially when a coder was struggling to determine whether a message should be coded as an exploration or integration message. The message set for the second experiment was modified so that coders were given sets of 20 or 25 contiguous messages. For example, a single coder would rate ten sets of 20 messages from one history course, ten sets of 20 messages from one political science course, and for the reliability set, four sets of 25 messages from two history courses and two political science courses. A systematic, random sampling

technique was used to determine the first message in each set of 20 or 25 contiguous messages.

Sixth, the coders were required to meet for face-to-face training sessions, and a different set of coders was used for the second experiment. In the first experiment, coders were scattered across Georgia and training was done primarily by telephone with one or two coders on the line at a time. The coders never gathered in the same physical space and never knew how each other coded. For the second experiment, the coders were asked to not only meet in the same physical space but to scrutinize each others' coding decisions. Further, when a coder rated a message differently from the others, that coder was asked to justify his/her ranking. I hypothesize that this process of justifying rankings solidified the coders' ratings and ultimately led each coder to rely more heavily on the rubric for making final coding decisions. Together, the coders solidified their understanding of each coding category, but more importantly, shared their process of coding. Sharing the process of coding ensured that each coder was using similar decision logic when making coding decisions. Further, fewer coders were used in the second experiment (n=4) than in the first experiment (n=6). Fewer coders provides a greater likelihood for increased inter-rater reliability.

Finally, to determine whether these modifications were valid and could be communicated easily during training, one outside coder unfamiliar with the project was asked to code a set of 200 messages using the new rubric. This coder was trained, and this coder's categorization decisions were analyzed for accuracy. This exercise revealed complex aspects of training allowing the trainers to anticipate questions the coders would ask. Primarily, this training revealed that extra effort is required to train coders on the

differences between exploration and non-cognitive categories and between exploration and integration categories.

Overall, this body of changes was applied to the rubric, to the set of messages coded, to coder training, and to the selection of coders. The method outlined in Chapter 3 was repeated with these modifications, and the results are detailed in the following section.

Second Experiment

Human Content Analysis

Like the first experiment, this section describes how reliably each coder rated messages compared to the other human coders. This experiment repeats the method outlined in Chapter 3 with the modifications outlined in the previous section. This section first describes the message sample then compares the coders. A systematic, random sample of messages was chosen from history and political science eCore™ courses. The sampling technique for this experiment was modified so that ten sets of 20 messages were chosen from each course section beginning from a random starting point and at a randomly chosen interval. Messages were chosen from eight eCore™ courses in history and political science held during the 2002 calendar year. Table 10 shows the total number of messages in each course by section and topic. Unlike the first experiment, there is approximately the same number of messages in both history and political science with the exception of the fourth section of political science.

Table 10

Number of Messages by Course Section and Topic

	History	Political Science
Section 1	1,100	1,053
Section 2	1,423	1,117
Section 3	1,357	782
Section 4	1,096	2,716
Total	4,976	5,668

Note. Total number of messages by topic for the second experiment.

After four training sessions, the coders reached pairwise Cohen's kappa reliability scores over 0.80 indicating that they were prepared to code messages. The four coders were then asked to code 500 messages: 200 from one section of history, 200 from one section of political science, and 100 from both topics used for reliability comparison. Table 11 shows the pairwise reliability of the four human coders and the ANN. Cohen's kappa values are reported above the diagonal, and percentage agreement is below the diagonal. The mean pairwise Cohen's kappa among the humans is 0.848 while the mean percentage agreement is 90%. ANN-to-human comparisons are provided in the next section. Coder B has the lowest mean kappa value (kappa = 0.816). On average, the kappa values exceed Rife, Lacy, and Fico's (1998) minimum of 0.8. Moreover, these reliability values exceed those from the first experiment indicating that the modifications outlined in the previous section may have had a positive impact.

Table 11

Agreement and Kappa Values for Each Coder

	A	B	C	D	ANN
A	-	0.798	0.866	0.864	0.697
B	87%	-	0.789	0.862	0.704
C	91%	86%	-	0.911	0.664
D	91%	91%	94%	-	0.687
ANN	81%	82%	78%	80%	-

Note. Percentage agreement scores are below the diagonal, and kappa values are above the diagonal.

Automatic Content Analysis

Just as in the first experiment, the messages coded in the human content analysis are used to develop the automatic content analysis tool, also referred to as the ANN model. The trained ANN model codes the same set of 100 messages used to determine human inter-rater reliability, and the ANN results are compared to the aggregate of human coders. The ANN model was constructed from 1,600 human-coded messages. Each of the four coders rated 500 messages; 200 from history, 200 from political science, and 100 reserved for reliability comparison.

To build the ANN model, the 1,600 human-coded messages were used to train and test the neural network. During training, the coders were asked to identify those messages which were difficult to code. The coders identified 125 difficult-to-code messages which were removed from the neural network training/testing set leaving 1,475 messages. From that set of messages, 1180 messages were used to train the model while

295 messages (25%) were reserved for testing. By default, the ANN software, Ward Systems'® Neuroshell 2, reserves 25% of the cases for testing. Numerous models were built from this set of messages, and the best of these was saved. These models were then used to code the set of 100 messages used for human inter-rater reliability, and the single model with the best overall reliability value, measured using Cohen's kappa, was kept. An aggregate of all four coders was used to generate one set of human-coded messages against which to compare the machine-coded messages. Where there was a tie among human coders, the algorithm described in Chapter 3, "Comparing Human and ANN Coding Decisions" was employed. The highest-performing ANN model yielded a percentage agreement of 81% and a Cohen's kappa value of 0.704. This kappa value shows an increase from the model built during the first experiment but does not equal the higher reliability achieved during the second experiment human content analysis.

Recall that 13 full models were constructed and the one model that most closely matches the human coders' decisions was retained. This model was built using the most discriminating 40 inputs. Appendix D provides a fuller description of the ANN settings and Appendix E provides a list of the 40 inputs and their descriptions. Interestingly, this model contains none of the hierarchy structure categories (thread number, message width, message depth, number of children, and number of grandchildren). The most discriminating input is word count, the number of words in a message. Presumably, longer messages are more indicative of integration messages which make substantiated claims. The next most discriminating input is whether or not the message contains a question. This input most likely separates triggering events from other messages. The third most discriminating input is whether or not a message contains the name of a person

from the class. We can hypothesize that this input separates exploration and integration messages from the others because a reply to a triggering event may contain the name of the person writing the triggering message.

Table 12 compares the coding decisions between the ANN model and the aggregate of human coders. This comparison reveals not only higher accuracy than the model from the first experiment but also that every category contains a message. This indicates that the second experiment model performs better than the model from the first experiment which undergeneralized on some categories while overgeneralizing on others.

Table 12

Comparison of Coding Decisions for the Second Experiment

		Desired (Aggregate of Human Coders)			
		NT	TE	EX	IN
	NT	**37**	1	10	0
Predicted	TE	1	**1**	0	0
(ANN)	EX	0	0	**7**	0
	IN	1	1	5	**36**

Note. This confusion matrix for the full artificial neural network model yields a kappa of 0.704. Here, NT is non-topical, TE is triggering event, EX is exploration, and IN is integration.

Table 13 summarizes the research results from the Garrison, Anderson, and Archer (2001) study, the first experiment, and the second experiment of this study. Overall, the second experiment shows improvement over the first and shows that the

automatic content analysis tool from the second experiment is approaching the human content analysis reliability of the seminal Garrison, Anderson, and Archer (2001) study.

Table 13

Synopsis of Reliability

Study	Content Analysis Type	Agreement	Kappa
Garrison, Anderson, & Archer (2001)	Manual	84%	0.74
First Experiment	Manual	74%	0.608
	ANN	71%	0.519
Second Experiment	Manual	90%	0.848
	ANN	81%	0.704

Before addressing the second part of research question, results of two other analyses should be addressed: the comparison of models and the effect of spelling errors.

Comparison of Models

This research has assumed that a single ANN model could be used to categorize both history and political science messages. The purpose of this section is to determine whether this assumption is valid. This is done by comparing three types of ANN models: the full model comprised of both history and political science messages, topic models, and section models. The full model is an ANN constructed from all of the human-coded messages from all sections of history and political science; there is one full model for experiment one and another for experiment two. The topic models are two models constructed from messages in each topic, history and political science. The section

models are built from messages in each of the course sections, three history sections and three political science sections from experiment one and four history sections and four political science sections from experiment two. The purpose of comparing these models is to determine whether one model, the full model, may be built to code both history and political science messages. If the full model is either more accurate than or just as accurate as the topic and section models, then this work may proceed along the more efficient path of constructing one model for all messages. The results from both experiments are presented below.

Experiment One

To reiterate, the full model from experiment one is built from 1,200 messages, and 100 of those messages are reserved to test the model. The kappa reliability statistic for the full model is 0.519 with a percentage agreement of 71%. In comparison, there are two topic models, one comprised of all coded history messages and one for political science messages. Each topic model is built from 600 messages, 500 to train the model and 100 to test the model. Reliability statistics for the topic models are outlined in Table 14.

Table 14

Reliability Statistics for Topic Models

Model	**Agreement**	**Kappa**
History	82%	0.7
Pol. Science	74%	0.56

Note. Both history and political science topic models exceed the full model (kappa=0.519; Percent Agreement=71%).

The reliability statistics for both the history and political science topic models exceed the reliability statistics of the full model.

There are also six section models, three for history and three for political science. Each section model is built from 200 messages, 150 to train the model and 50 to test the model. Reliability statistics for the section models are outlined in Table 15. Four of the six section kappa values exceed the full model (kappa = 0.519). However, each section model is lower than the best topic model (history kappa = 0.7) and comparable to the lowest topic model (political science kappa = 0.56).

Table 15

Reliability Statistics for Section Models

Model	Agreement	Kappa
History Section 1	78%	0.61
History Section 2	72%	0.54
History Section 3	80%	0.66
Pol. Sci. Section 1	70%	0.43
Pol. Sci. Section 2	70%	0.50
Pol. Sci. Section 3	74%	0.56

Note. Four of the six section models exceed the full model (kappa = 0.519).

This finding suggests that, in experiment one, individual topic models may be more accurate.

The above comparisons for the first experiment reveal that the topic models and section models may outperform the full model. The reliability values for the first experiment are low which calls the overall comparison of models for the first experiment

into question. If the topic and section models continue to outperform the full model on the more robust second experiment, then there may be reason to question whether one, generic model should be constructed. The analysis below compares the full model to the topic and section models for the second experiment.

Experiment Two

For the second experiment, the full model is comprised of 1,475 messages, and 25% (295 messages) of those were reserved for testing. The kappa reliability statistic for the full model is 0.704 with a percentage agreement of 81%. Like the first experiment, there are two topic models, one for history and one for political science. With the difficult-to-code messages removed from each topic model, the history model was built from 726 messages. Recall that each coder categorized the same set of 100 messages for a reliability comparison. Fifty of those were history messages, and 50 were political science messages. Therefore, the history topic model is comprised of 726 human-coded messages with 50 messages reserved for testing, and the political science topic model is built from 749 human-coded messages with 50 messages reserved for testing. Reliability statistics for the topic models are outlined in Table 16.

Table 16

Reliability Statistics for Topic Models

Model	Agreement	Kappa
History	80%	0.621
Pol. Science	78%	0.639

Note. Neither the history nor political science topic models are more accurate than the full model (kappa=0.704; Percent Agreement=81%).

Unlike the first experiment, the full model is more accurate than both topic models.

The performance of the full model compared to the topic models for the second experiment indicates that one, generic, full model is not only more efficient to construct, but more accurate than individual topic models. The following comparison studies the performance of the full model against individual section models. For the second experiment, there are four sections of history and four sections of political science. Numerous section models were constructed and the models with the highest kappa values are reported. Like the first experiment, there are 200 messages from each section: 150 are used to train the model and 50 are reserved for testing. Reliability statistics for the section models are outlined in Table 17.

Table 17

Reliability Statistics for Section Models for Experiment Two

Model	Agreement	Kappa
History Section 1	66%	0.451
History Section 2	84%	0.731
History Section 3	76%	0.605
History Section 4	82%	0.729
Pol. Sci. Section 1	82%	0.722
Pol. Sci. Section 2	66%	0.488
Pol. Sci. Section 3	78%	0.66
Pol. Sci. Section 4	94%	0.802

Note. Four of the eight section models exceed the full model (kappa = 0.704).

Overall, the full model for the second experiment is more accurate than four of the eight section models and more accurate than both topic models.

For the first experiment, most topic and section models outperformed the full model. However, this finding did not persist into the second experiment in which the full model outperformed all topic models and half of the section models. These initial findings indicate that the full model is a viable candidate against the more pinpointed topic and section models.

Spelling Analysis

Recall from Chapter 3 that the ANN model is built from the presence or absence of words in a discussion list message. Therefore, misspellings may threaten the ANN's ability to correctly classify messages. A systematic, random sample of 100 messages was chosen from the 8 courses (four history and four political science) used in the second experiment of this study. The text of each message body was placed in Microsoft Word which automatically identified the spelling errors. Those errors were corrected until Word© identified no further spelling errors, and this process was repeated for all 100 sampled messages. The spell-corrected set of messages was then placed into the database and the method used to numerically describe messages was applied. The ANN algorithm was then applied to each of the 100 messages in order to categorize that message into cognitive presence categories. Again, the hypothesis is that the ANN model will place correctly-spelled messages into the same category as their misspelled counterparts.

Of the 100 messages sampled, each misspelling that Microsoft Word© identified was counted. Based on that count, the average number of spelling errors per message which could contribute to the message being falsely categorized is 0.92 (SD = 2.34). Had

the entire population been analyzed, the mean number of misspelling errors would fall between a lower confidence boundary of 0.46 and an upper confidence boundary of 1.39 assuming 95% confidence. Though this low number does not make it seem that spelling affects the outcome of the model, greater certainty may be reached by sending the corrected messages back through the ANN model to see if the existing model categorizes the corrected messages differently than the messages with misspellings.

Both the original set of uncorrected messages and the set of messages corrected for misspellings were sent back through the ANN model from the second experiment. Both sets of messages were coded exactly the same by the model. This indicates that the spelling errors were not significant enough contribute to errors in coding. It may still be the case that spelling errors in some messages cause them to be miscoded by the ANN, but the current analysis suggests that the number would be less than one percent of all messages.

Overall, there is an average of approximately one spelling error per message. There may appear to be more spelling errors because messages contain a number of errors other than spelling errors including grammatical and word choice errors. Many of these errors do not affect the correct message classification. For example, the message below appears riddled with errors making it at first appear to be a reasonable candidate for false classification.

> i [*sic*] thought the same thing. the basic behavior of humanity really hasen't [*sic*] changed and i'm [*sic*] sorry to say that i [*sic*] doubt it ever will. just think, if osama had had as strong an army as cortes did (as opposed to the aztecs), the ones of us left would be wearing burka's [*sic*] and growing beards.

This message contains poor capitalization, an emotional topic, an arguably exaggerated point, the incorrect use of possessive case, but just one misspelling. Recall that the average misspellings per message is 0.92. This message contains many errors, but just one is an error that would possibly contribute to its false classification.

Sample ANN Analyses

Recall that the research question, ""how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans" has two parts. The first part hypothesizes that an ANN analyzes messages as accurately as a group of humans, and that part was addressed in the section above. This section addresses the second part of the research question, a description of the information we should expect from an ANN content analysis tool.

After deriving the ANN model that categorizes messages closest to the set of human coders, the model was run against every message in every course. The most outstanding benefit of an ANN model is that it tirelessly categorizes every message. At the very minimum, such a tool should offer descriptive statistics in the form of mean cognitive presence values for a body of messages, the distribution of cognitive presence categories among a body of messages. However, a manual content analysis will provide that information. An ANN content analysis, on the other hand, goes beyond manual content analyses by offering fine-grained analyses of course variables such as cognitive presence by student, weekly topic, or major thread. Since the results from the previous section reveal that the model from experiment two possesses a higher reliability than the model from experiment one, the examples used come from messages analyzed using the second experiment's model.

Comparison of Means

First, an instructor or administrator may conduct broad-level comparisons based on mean cognitive presence values by both topic and section. As mentioned in Chapter 3, the mean cognitive presence weight is an average of messages whose cognitive presence value falls along a continuum between zero and four as follows:

0: Non-cognitive

1: Triggering Event

2: Exploration

3: Integration

4: Resolution

Given the modifications to the model from experiment one to experiment two, the first category, non-cognitive, is more accurately named "non-topical" and the final category, resolution, is removed leaving the following values:

0: Non-topical

1: Triggering Event

2: Exploration

3: Integration

For example, Figure 14 compares mean cognitive presence values by topic, history and political science. Here, political science displays a slightly higher mean cognitive presence than history. Upon further inspection, however, Figure 15 shows one political science section is responsible for skewing the results of all reported political science courses. In fact, on removing that section from the analysis, history sections slightly outperform the political science sections. A closer inspection of Figure 15 also prompts

the instructor or administrator to ask what factors in political science 4 are responsible for its improved performance.

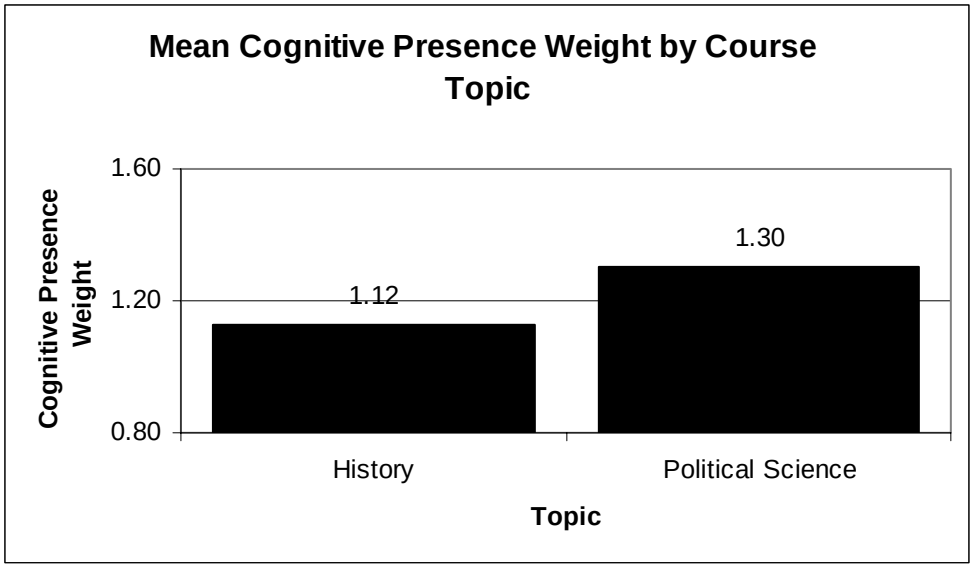


Figure 14. Mean cognitive presence value by course topic.

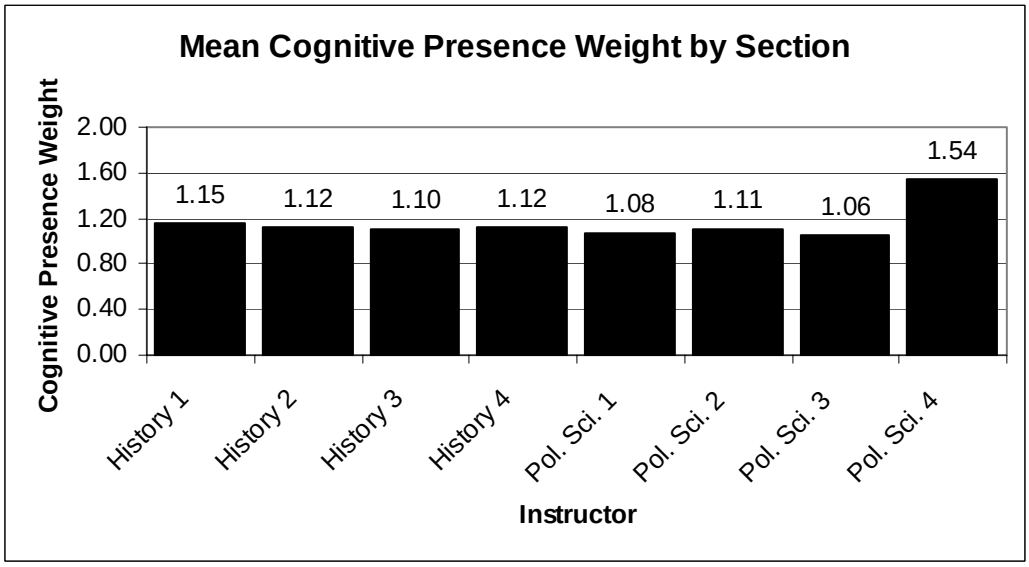
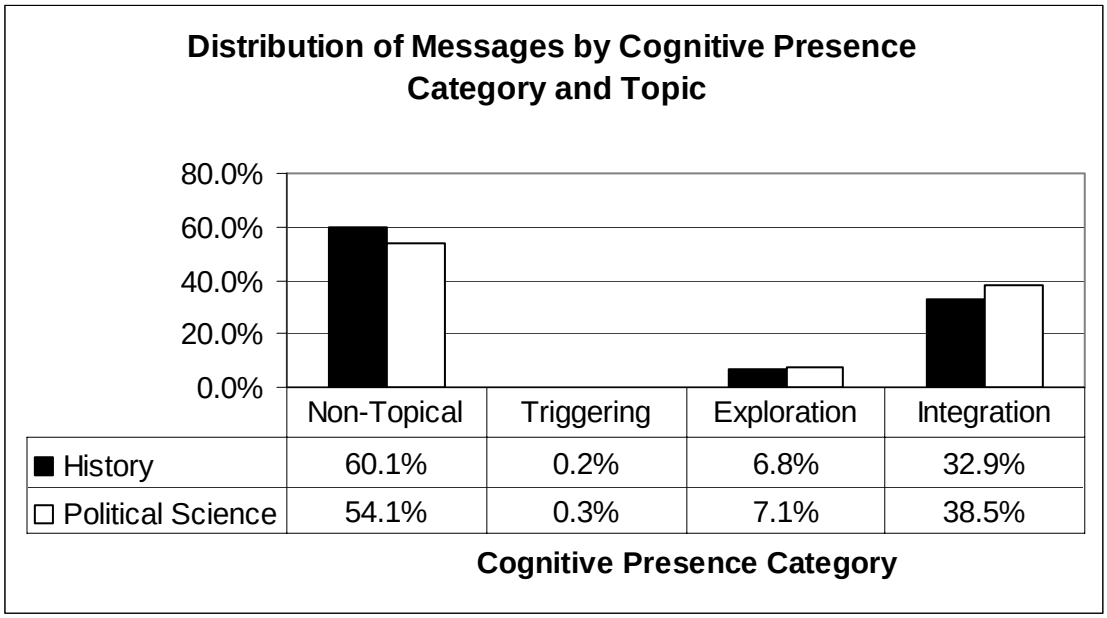


Figure 15. Mean cognitive presence weight by section.

Distribution of Messages by Cognitive Presence Category

This section exemplifies the automatic content analysis tool's ability to provide gradually deeper levels of information about a body of messages. This section shows the distribution of messages by cognitive presence category (non-topical messages, triggering events, exploration messages, and integration messages) and across topics and sections.

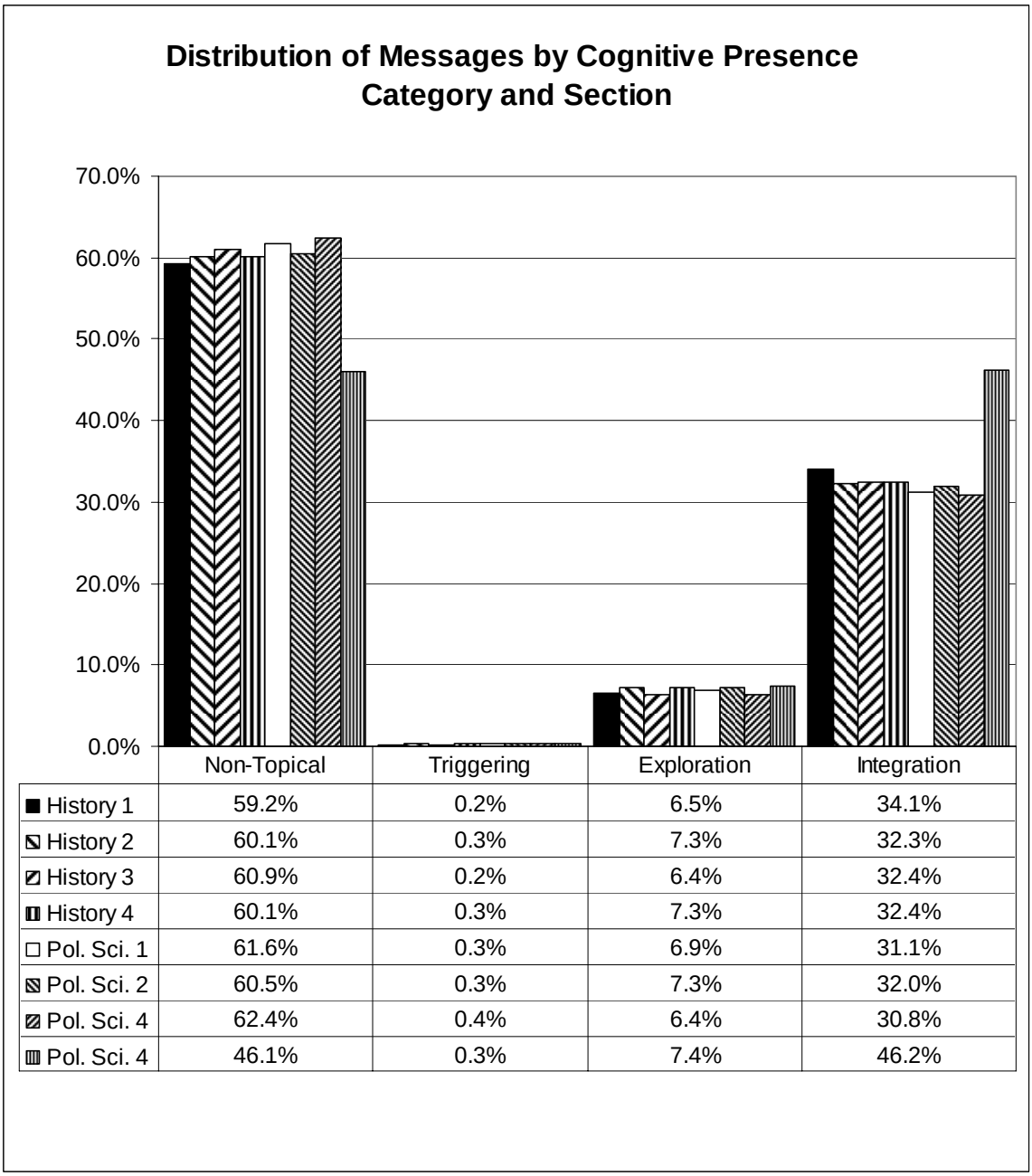
Figure 16 shows the distribution of messages by cognitive presence category over each topic, history and political science. Notice that over half of all messages in both topics are non-topical messages devoted most likely to technical support and social exchanges such as greetings. This analysis also shows little fluctuation between topics on any of the cognitive presence categories. Interestingly, there are far more integration messages than exploration messages meaning that students may be justifying their claims more than they are engaging in exploratory activities such as brainstorming. This certainly prompts instructors and administrators to ask whether this is desirable.



	Non-Topical	Triggering	Exploration	Integration
History	60.1%	0.2%	6.8%	32.9%
Political Science	54.1%	0.3%	7.1%	38.5%

Figure 16. Distribution of messages by cognitive presence category and topic.

Figure 17 shows how each cognitive presence category is distributed over each section, three sections of history and three sections of political science. As expected, a majority of messages are coded as non-cognitive. Also notice that in Figure 15, political science section 4 showed the highest mean cognitive presence value. Figure 17 offers an explanation: this section had fewer non-topical and more integration messages than any other section. Looking back at Table 10, political science section 4 generated 2,716 messages, two to three times more messages than any other course possibly because this course had 48 students while most of the other courses had between 28 and 32 students. Certainly, an array of questions emanates from this combination of variables: “How was this instructor able to maintain a high number of integration messages?” “What contributed to the relatively low percentage of non-topical messages?”



	Non-Topical	Triggering	Exploration	Integration
History 1	59.2%	0.2%	6.5%	34.1%
History 2	60.1%	0.3%	7.3%	32.3%
History 3	60.9%	0.2%	6.4%	32.4%
History 4	60.1%	0.3%	7.3%	32.4%
Pol. Sci. 1	61.6%	0.3%	6.9%	31.1%
Pol. Sci. 2	60.5%	0.3%	7.3%	32.0%
Pol. Sci. 4	62.4%	0.4%	6.4%	30.8%
Pol. Sci. 4	46.1%	0.3%	7.4%	46.2%

Figure 17. Distribution of messages by cognitive presence category and section.

Analyses Based on Course Section Variables

The value of an automatic content analysis tool is that it categorizes every message instead of a sample of messages. This capability allows fine-grained analyses related to a single course. Three examples, all related to one section of one course, are

shown below; the first is an analysis of each student's performance, and the second two describe a specific discussion thread. Table 18 shows the mean cognitive presence weight of each student along with the total number of messages each student contributed and the number of messages falling into each cognitive presence category. This reveals that the number of messages is no indicator for a high cognitive presence value. In fact, the student contributing the largest number of messages (FirstName47 LastName47 who posted 221 messages) possesses a mid-range cognitive presence value of 1.48. The range goes from a low of 0.00 to a high of 2.21; the highest value a student could possibly achieve is a three. Although a high message count does not predict a high mean cognitive presence value, a high mean cognitive presence value can still be achieved by making far too few contributions to the class. In fact, the student with the highest cognitive presence value, FirstName 33 LastName 33 with a mean cognitive presence value of 2.21, made only 28 contributions to the discussion. At this early stage, it is too soon to tell what should be considered an acceptable number of messages and an acceptable cognitive presence weight.

Table 18

Mean Cognitive Presence Values by Student

Student	Mean Cognitive Presence Value	NT	TE	EX	IN	Total
FirstName1 LastName1	1.21	24	1	1	16	42
FirstName2 LastName2	1.21	41	0	4	26	71
FirstName4 LastName4	1.93	48	0	16	86	150
FirstName6 LastName6	1.41	62	0	10	51	123
FirstName7 LastName7	0.00	1	0	0	0	1
FirstName8 LastName8	1.38	57	0	2	48	107
FirstName12 LastName12	0.67	7	0	0	2	9
FirstName13 LastName13	1.65	21	1	1	26	49
FirstName14 LastName14	1.00	3	0	1	1	5
FirstName16 LastName16	1.81	54	0	12	80	146
FirstName17 LastName17	1.68	45	1	11	55	112
FirstName19 LastName19	1.09	7	0	0	4	11
FirstName20 LastName20	1.41	65	0	10	54	129
FirstName22 LastName22	0.83	4	0	1	1	6
FirstName23 LastName23	1.29	7	1	1	5	14
FirstName24 LastName24	1.70	14	0	1	18	33
FirstName26 LastName26	1.00	6	0	0	3	9

Table 18 (continued)

Mean Cognitive Presence Values by Student

Student	Mean Cognitive Presence Value	NT	TE	EX	IN	Total
FirstName27 LastName27	1.74	56	0	17	74	147
FirstName28 LastName28	1.31	29	0	4	21	54
FirstName29 LastName29	0.64	10	1	1	2	14
FirstName31 LastName31	0.75	9	0	0	3	12
FirstName32 LastName32	1.61	17	0	2	19	38
FirstName33 LastName33	2.21	7	0	1	20	28
FirstName34 LastName34	1.88	37	1	8	62	108
FirstName35 LastName35	1.28	60	0	8	41	109
FirstName36 LastName36	1.77	12	0	2	17	31
FirstName39 LastName39	1.91	11	0	3	19	33
FirstName40 LastName40	0.00	1	0	0	0	1
FirstName41 LastName41	1.81	28	2	5	43	78
FirstName42 LastName42	1.33	55	0	7	41	103
FirstName44 LastName44	2.00	11	0	6	22	39
FirstName46 LastName46	1.83	53	0	14	81	148
FirstName47 LastName47	1.48	108	0	13	100	221
FirstName48 LastName48	2.00	12	0	0	24	36

Aside from analyzing each student's individual cognitive contributions, we may also discern information related to course topics. Here, course topics may be teased from the data in two ways, by time interval, such as a week, or by message thread number. First, an instructor may ask students to contribute to a specific topic each week in which case each week is associated with a unique topic. Figure 18 shows the cognitive presence values for each topic, assuming that a new topic is introduced each week. This shows that the first week is for general introductions which tend to be non-cognitive and that the final week of messages is usually comprised of well-wishing and thanks which would also be considered non-cognitive. In between, however, an instructor may analyze each topic for the level of intellectual effort exemplified in student messages. Figure 18 shows that the topic from weeks six and nine carried the greatest effort while that from week four carried the least effort. Further, Figure 19 shows the number of messages contributed each week for the same course. Over 500 messages were generated in the first week, yet the mean cognitive presence value was low, further indicating that most messages during this time were greetings and introductions. After week one, the weekly message count tapered off to between 200 and 300 messages per week until the final two weeks of this summer course.

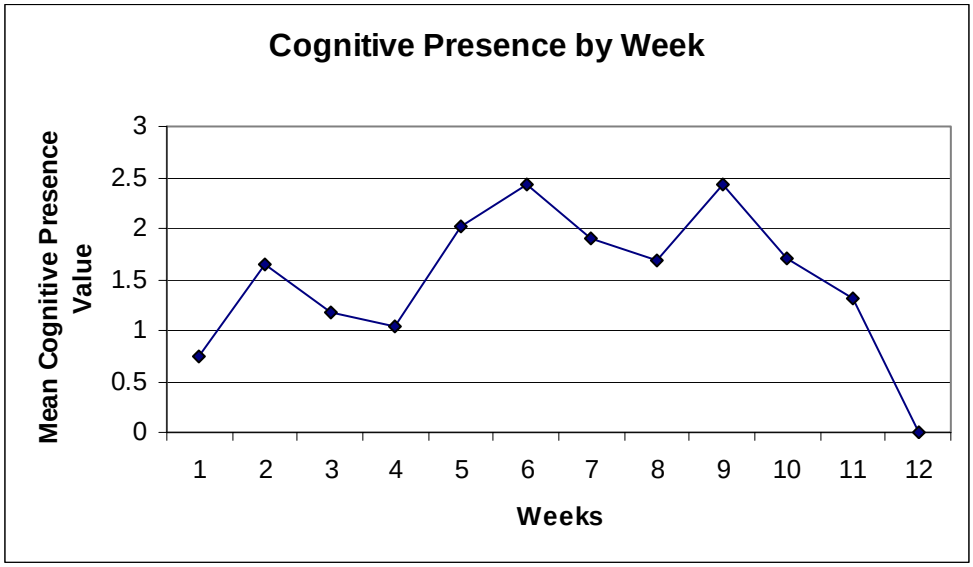


Figure 18. Mean cognitive presence values by week.

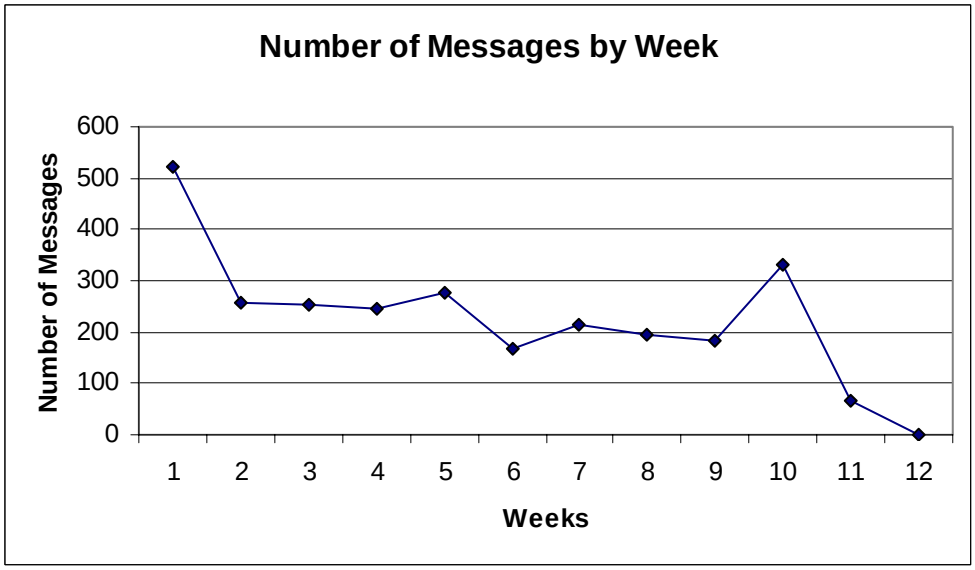


Figure 19. Number of messages by week

An instructor may also break the discussion of a particular topic into individual threads in which case one thread represents one topic. In this case, the cognitive presence of a single discussion is measured allowing the instructor to see which threads generated discussion that goes beyond non-cognitive or exploratory comments. Figure 20 shows

that, of the more than 1,000 discussion threads in History 4, thread 619 generated the highest mean cognitive presence value, but this does not indicate how many messages are in this thread nor the individual cognitive presence weights. For that, Figure 21 shows the cognitive presence distribution for a single message thread, thread 619. From this, we see that there are 17 messages in this thread and that 13 of these messages fell into the integration category meaning that the majority of contributions to this topic may be substantiated claims.

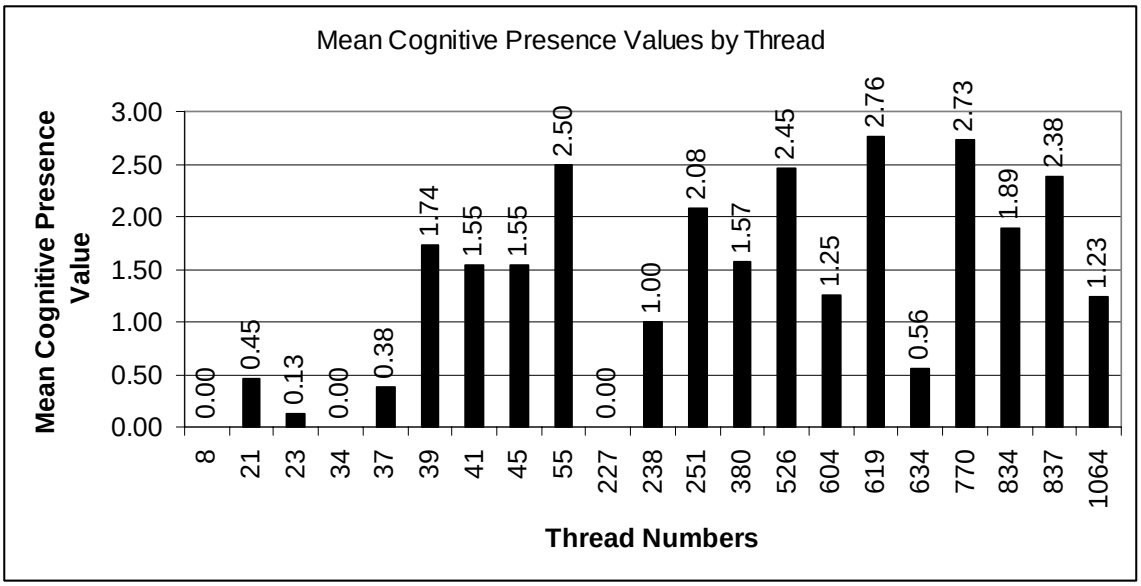


Figure 20. Mean Cognitive Presence Values by Thread for History 4

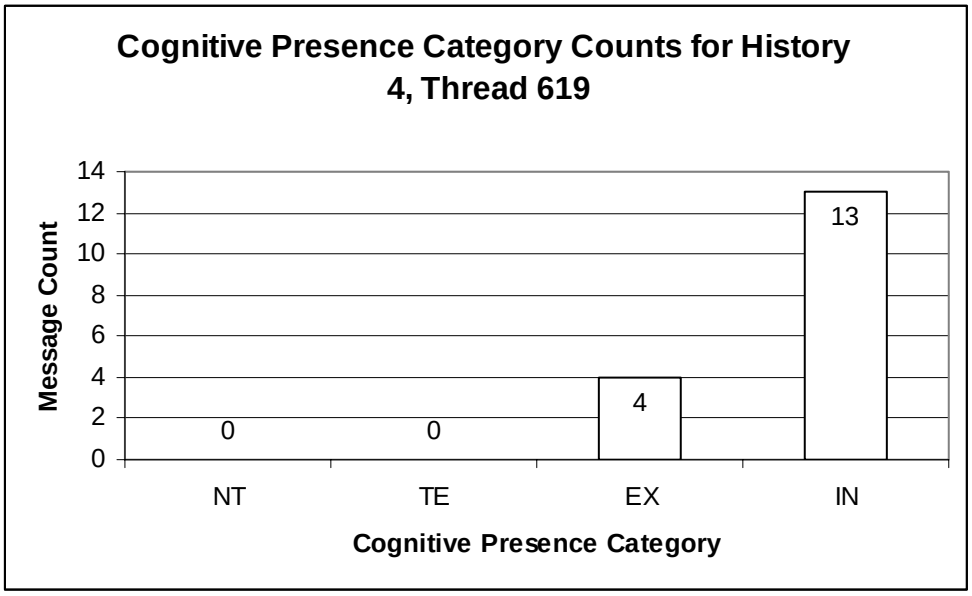


Figure 21. Cognitive Presence Category Counts for History 4, Thread 619

Secondary analyses

The above analyses immediately surface as a result of associating cognitive presence values with each message; however, the researcher or instructor with statistical experience may perform a number of more rigorous statistical analyses in which the cognitive presence weight serves as the dependent variable. Specifically, the instructor may perform analyses of variance (ANOVA) or regression analyses among various factors. Some of those factors are (a) instructor, (b) student, (c) message length, (d) thread number, (e) week in course, (f) topic (assuming that an instructor presents a topic for discussion over a specific period of time), (g) student participation either in length of messages or number of messages, and (h) instructor participation.

Summary

Overall, this chapter presents the results of the research question: “how well does an artificial neural network (ANN) analyze and describe the cognitive effort students

exhibit in online educational discussions as compared to humans?" This question has two parts. The first part hypothesizes, that an artificial neural network (ANN) analyzes messages as well as a human. The second part describes the information expected from an ANN content analysis tool. Compared to humans, the ANN models from both the first and second experiment possess lower reliability statistics measured using Cohen's kappa. The ANN model from the second experiment, however, possesses a greater reliability value than the mean pairwise kappa values from the human content analysis of the first experiment, suggesting that the modifications made during the second experiment improved the model's accuracy. This chapter answers the second research question by demonstrating the analyses an instructor should expect from a automatic content analysis tool. In brief, an instructor should expect a cognitive presence variable showing results over time, by topic, section, instructor, student, and thread. This variable should also enable action research through statistical procedures more sophisticated than descriptive analyses.

CHAPTER 5

DISCUSSION AND RECOMMENDATIONS

This chapter explains the results of the research question, "how well does an artificial neural network (ANN) analyze and describe the cognitive effort students exhibit in online educational discussions as compared to humans?" This question has two parts. The first part hypothesizes, that an artificial neural network (ANN) analyzes messages as well as a human. The second part describes the information expected from an ANN content analysis tool. Broadly, Chapter 3 describes the methods used to answer the research question, Chapter 4 outlines the results, and this chapter offers possible explanations for those results. This set of explanations is followed by the limitations and bias of the study beyond those explained in Chapter 3, the study's major contributions, and a road map for future research. To guide the discussion, the graphic overview of the research method is presented again in Figure 22.

Comparing Artificial Neural Networks to Humans

The first part of the research question which hypothesizes that an ANN analyzes messages as well as a human is addressed by the two iterations of the research method. Explanations of the results from the two experiments are presented in the following two sections.

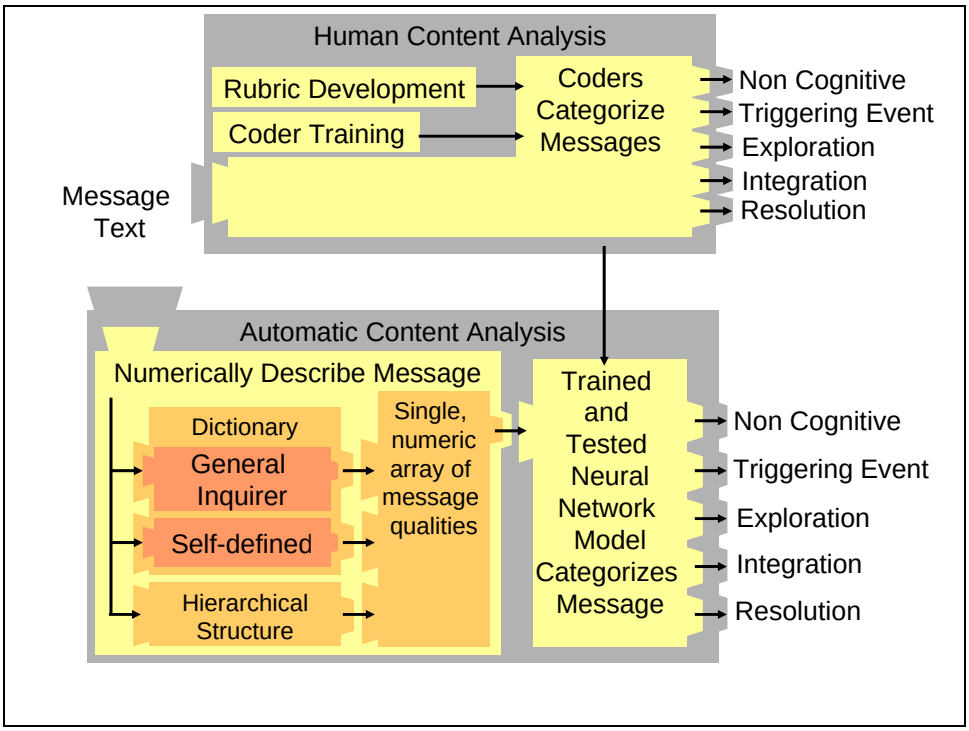


Figure 22. Graphic overview of the research methods.

First Experiment

In the first experiment, the ANN model and the aggregate of human coders describe the cognitive landscape of online discussions approximately as well as the lowest-performing coders. That is, based on reliability statistics, the range of pairwise Cohen's kappa values among human coders (0.504 to 0.747) encompasses the pairwise kappa value between the ANN model and the aggregate of human coders (kappa=0.519). In this example, the ANN model could be introduced as a seventh coder and emerge with pairwise reliability statistics that do not distinguish it as non-human. However, upon looking at the way the ANN model and the group of humans code messages, differences emerge making the ANN model look clearly non-human. The following sections outline those differences and offer possible explanations.

The Differences Between ANN- and Human-Coded Messages

Three distinct differences emerge between the way the ANN model codes messages and the way the humans code the same messages. The ANN model is not as sensitive to categories with few messages, the ANN model confuses non-cognitive with exploration messages, and the ANN model differs in the way it categorizes exploration and integration messages. The first difference between the ANN model and the human coders is that the group of human coders may be more sensitive to cognitive presence categories with fewer messages. Table 9 shows that the group of human coders placed messages into four of the five cognitive presence categories, but the ANN only placed messages into three of the five categories. Though one coder identified a message as a resolution message, the aggregate of human coders placed no messages into the resolution category. The aggregate of human coders identified eight triggering events whereas the ANN model placed no messages into the triggering event category. Instead, the ANN model distributed all 100 messages into either the non-cognitive, exploration, or integration categories. This finding indicates that, at this stage, the ANN model may not be as sensitive as human coders are to rare messages.

The second difference between the ANN and the group of human coders is that the ANN model confuses non-cognitive and exploration messages. Table 9 shows that the ANN model agrees with the humans on 45 non-cognitive messages, but the ANN non-cognitive category held seven human-coded exploration messages. Further, the ANN exploration category held five human-coded non-cognitive messages. Though the ANN and humans agree on the vast majority of non-cognitive messages, the ANN non-cognitive category remains partially entangled in the exploration decision space.

The third major difference between the way humans code messages and the way the ANN model codes is found in the way the two code exploration and integration messages. As for exploration messages, Table 9 shows that the ANN model places human-coded exploration messages into both the non-cognitive and integration categories, ultimately disagreeing with the human coders on 12 of 31 human-coded exploration messages. Further, the ANN exploration category contains 11 messages that humans do not consider being exploration messages. This exemplifies the most common coder complaint during training that the line separating exploration and integration is not clear. Further, the ANN model and the humans disagree on some integration messages. The ANN model missed three of the 10 human-coded integration messages and placed them, instead, into the non-cognitive and exploration categories. Inversely, the ANN integration category contains one human-coded non-cognitive message and five human-coded exploration messages.

Overall, this reveals differences between the ANN and the aggregate of human coders, but this does not show that the ANN errors are different than the errors among humans. The next section, “Renegade Coding,” analyzes those types of errors.

Renegade Coding

To more deeply understand the differences between the human and ANN coding decisions, renegade coding patterns may be analyzed. Renegade coding occurs when one coder (human or ANN) makes a coding decision unlike any other human coder.

Among the 100 reliability messages comparing the ANN model to the human coders, there are twelve instances in which the ANN model disagrees with every human coder. For example, in the message below, four coders rated this message an integration

message, two coders rated it as an exploration message, yet the ANN rated it as non-cognitive.

> What a sad state of affairs you must bear witness to my brother. You must continue in your efforts to circumspent [*sic*] these horrific acts against the Natives and thus against our Holy Father's way. In your efforts you must always remember that you do the most divine of works and try to take comfort in the fact that there are brothers trying to do the same. Hopefully the Spaniards of which you speek [*sic*] will come to the Lord and see the evilness they have wrought and repent. I feverently [*sic*] pray this will come to pass in the name of all that is good. Humbly, FirstName25 LastName25

This message exemplifies the role playing many of the instructors in this study asked of their students. Students adopted the persona of a historical figure and posted messages in character. One by-product of this is that the students also adopted the manner of speaking of their character. This by-product may throw off an algorithm designed to code messages based on lexical cues. In this example, that may have been the case. The coders all rated it as either an exploration or integration message while the ANN model placed it in the non-cognitive category.

Interestingly, though there are 12 instances in which the ANN model coded unlike any single coder, there are 20 instances total in which one coder rated a message unlike any other coder. Table 19 outlines the instances of renegade coding. It shows that the ANN coded more messages unlike any human coder, but that human coders were also fallible.

Table 19

Renegade Coding for Experiment One

Coder	Instances of Renegade Coding	Comparison
ANN	12	Compared to the six human coders
A	6	Compared to the other five human coders
B	4	Compared to the other five human coders
C	3	Compared to the other five human coders
D	2	Compared to the other five human coders
E	1	Compared to the other five human coders
F	4	Compared to the other five human coders

Note. The ANN codes unlike any human coder more often than the human coders, but the human coders also exhibit independence.

Though 12 instances of renegade coding distinguish the ANN model from the set of human coders, a few instances of renegade coding are certainly within the realm of human expectation.

Further, there are three instances in which there is uniform agreement among all human coders yet disagreement between the ANN model and the human coders. In the message below, every human coder rated this message as a triggering event, yet the ANN model placed it in the non-cognitive category: "You have noted how government affects your lives. How do you account for the 50% of Americans over the age of 18 who don't vote? Why do people not vote?" This shows that the model is not picking up on the very distinct triggering event cues. This message is clearly a question designed to evoke a

response. At this stage, there may be too few training messages for the ANN model to distinguish it from other messages.

The following message is the second example of uniform agreement among the human coders but disagreement between the aggregate of human coders and the ANN model. Each human coder called it an integration message, yet the ANN model coded it as an exploration message.

> In his statement "our detached and distant situation enables us to pursue a different course", Washington was saying since we are not apart [*sic*] of and we are so far from the other nations we should be able to keep our affairs separate. Being separated by to [*sic*] vast bodies of water the Atlantic and the Pacific oceans made is [*sic*] a lot easier to [*sic*] we are a world unto ourselves, what goes on here stays here and what goes on there stays there.

One distinguishing feature of an integration message is that it refers to another source, and this message clearly does. Three linguistic cues show that the author of this message is invoking an outside source, yet the ANN model is not sensitive enough to pick up on the combination of quotation marks and reference phrases like “In his statement” and “was saying.” Again, this may be explained by the fact that there simply were not enough examples or that competing noise in the message made it appear exploratory.

Explanation of the Differences Between ANN- and Human-Coded Messages

Broadly, the first experiment shows that the ANN model codes messages within the range of human pairwise kappa values. However, three errors make it appear non-human. It identifies no triggering events, it confuses non-cognitive and exploration messages, and it confuses exploration and integration messages. By looking at the instances of human renegade coding, some error is to be expected from a model built from human decisions.

Numerous potential reasons may explain why the ANN model describes the cognitive landscape of discussion list messages differently than a group of humans. The most obvious of these is that the ANN model is not human. It may be too much to expect that software trained on a few hundred messages will make decisions identical to a group of humans with a biological neural networks constructed from a lifetime of experience. Most other reasons for the differences emanate from the single reason that the information used to train the ANN model forces it to make decisions distinctly different than humans. The discussion below explains why the information used to train the ANN model created a model unlike the lineup of human coders. The graphic overview of the research methods shown in Figure 22 provides the causal structure leading up to an ANN model reliable at kappa = 0.519 to a group of human coders. That causal structure is as follows: The ANN model is built from a set of training messages with a less-than-desirable inter-rater reliability; the less-than-desirable inter-rater reliability of the training set was produced by human coding decisions; the human coding decisions are a product of coder training; the quality of coder training is affected by the coding rubric. If error is introduced to any link in this causal chain, the entire ANN model is potentially weakened. The discussion below follows the causal chain from the rubric to coder training to the human coding decisions to the less-than-ideal set of messages used to train the ANN model.

To begin, the rubric may have failed to adequately guide coder decisions in two obvious ways. First, many coders complained that they had trouble understanding the difference between exploration and integration messages indicating that the guidelines and examples were not adequate. It may also indicate that the constructs themselves,

exploration and integration, may not be clearly distinct from one another. Garrison, Anderson, and Archer's (2001) rubric mentions that disagreement or divergence falls within the exploration category while agreement or convergence falls within the integration category (p. 10-11). Message 307 in Figure 7 disagrees with the message to which it replies but offers justification:

> I don't feel as though he were exaggerating at all. The Aztec civilization was trly [*sic*] flourishing in the 16th century. Just because they did not have a religion like that of the European explorers, did not mean that they were a backward people. Prior to the Aztecs, the Mayan people had calendars much more accurate than those in use by Spain, England, Portugal, and others. Their architecture was a sight to behold and it still exists to this day on the Yucatan peninsular [*sic*].

The coders were unsure whether a message like this should be an exploration message because it shows disagreement or whether it should be an integration message because it presents a justified claim.

Second, coders had difficulty correctly coding triggering events that were intentionally initiated by the instructor. The following example is from one of the history courses:

> I would like you focus on the following ideas. How accurate do you think Cortes' account is? How much of it do you think was inflated, or reflected ideas that he might not have actually had first hand experience with? Why do you think Cortes wrote this letter? Do you find any elements in it that could be construed as self-serving? What about this account surprised you? Remember that for evaluation purposes, you need to make at least two substantive posts to this discussion. That means that messages such as "I agree with so and so.." are not sufficient. You should provide ideas that contribute to the discussion, and respond to others' ideas by addressing those ideas directly, and providing insight, analysis, etc.

In some instances, a message like this was coded under the non-cognitive subcategory called "Unrelated/Course Management." This may be due to the lack of a triggering event subcategory that is consistent with an instructor-initiated triggering event. Current

subcategories under triggering event are “Sense of Puzzlement” and “Recognizes Problem,” subcategories that are inconsistent with an intentional, instructor initiated, triggering event. Clearly, the instructor was not puzzled, nor was the instructor recognizing a problem. This issue is easily corrected by adding a new triggering event subcategory for instructor-initiated triggering events. These two issues with the rubric could cause the coders to define certain messages for themselves without using the rubric as a guide, a process that might bleed into other coding decisions. That coder uncertainty is translated into ANN uncertainty.

During coder training, a number of issues may partially explain the difference between the human and ANN performance. That is, coder training may not have prepared coders to code messages exactly alike. First, the coders in experiment one may not have coded enough messages during training. Most coders were trained on 90 messages. During training, coders were asked to code 30 messages during each training session and reliability scores were calculated after each session. In hindsight, this low number of messages inflated the reliability scores during the third training session falsely indicating that training could cease. The mean pairwise reliability of the third and final training session was 0.7, much higher than the reliability achieved during coding, 0.608. Ending training early may have meant that the coders were not unified in their decision-making. Further, the threshold for ending coder training was a mean pairwise Cohen’s kappa reliability score among coders of 0.70 which may have been too low. Aside from ending training too early, coder training for experiment one may not have been rigorous enough. Coders were trained at a distance usually via teleconference and were not asked to justify their responses before a group of peer coders. This strategy did not ensure

similar coding and may have created multiple, conflicting decision strategies among coders.

Finally, the sample of messages provided to coders, both during training and during coding, provided no context for messages. The systematic, random sampling technique plucked single messages from their context and prevented coders from discerning the conversation topic by knowing the body of messages surrounding the ones they were reading. This meant that some exploration messages may have been coded as non-cognitive because the coder could not discern the topic. Again, this creates competing, noisy decision logic for the ANN.

Another explanation for the discrepancies between human and ANN coding is that the message set used to train the ANN included too many conflicting examples. Though Garson (1998) states that artificial neural networks are robust under conditions in which the input data are "noisy, nonlinear, and with missing measurements," (p. 162) there is a threshold at which the ANN cannot compensate for poor training cases. Specifically, difficult-to-code messages were not eliminated, two conflicting coding strategies were fed into the ANN, and the ANN model may need more examples of rare messages than were provided. First, Chi (1997) describes a coding discrepancy in which coders are unsure which code to assign to a coding unit, the message in this case. Chi's recommendation is that those difficult-to-code items be removed from the final results. Removing the messages that the coders found difficult ensures that the ANN is trained on a more consistent and less noisy set of messages and may also correct some of the error of the rubric. Second, inspecting the pairwise kappa values for each coder shows two potentially conflicting coding strategies. Table 8 shows the kappa values among the six

human coders. From this, two high-reliability groups of coders emerge. One group, comprised of coders B, C, and D show reliability greater than the group mean (kappa = 0.6). The mean kappa value of these three coders (B and C = 0.742; B and D = 0.682; and C and D = 0.714) is 0.713. The second group, coders A and F, also shows pairwise reliability (A and F = 0.68) above the group mean. Further, the mean pairwise reliability between these two groups of coders (kappa = 0.55) is below the mean pairwise reliability for the entire group of coders. These kappa reliability values indicate two distinct groups of coders. Since the ANN model is constructed from the coding logic of the human coders, the ANN model may have been constructed from two competing decision strategies. Finally, there simply may not be enough triggering event and resolution messages to adequately train the ANN. To adequately categorize messages, the ANN may need a larger set of rare messages than it was provided. Ultimately, the set of messages used to train the ANN may have possessed competing decision logic for some cognitive presence categories and too little decision logic for other categories.

The causal structure described above linking the rubric to the creation of a less-than-ideal training set of messages describes where error may occur within the human content analysis. Figure 22 shows that yet another explanation for the difference between ANN and human message categorization may lie within the automatic content analysis. Specifically, the strategy used to describe each message before it is sent to the ANN may introduce error. Recall from Chapter 3 that messages are translated into an array of numbers, and each number describes a specific quality of that message. Those qualities are defined by two dictionaries, the General Inquirer and a self-defined dictionary of qualities developed for this study to help discriminate one cognitive presence category

from another. That is, each quality description should provide enough numeric information to the ANN that it can distinguish, for example, a non-cognitive message from a triggering event. It may be that the numeric description of each message does not contain the full set of message qualities needed to adequately categorize messages.

Overall, Figure 22 shows four areas where error may be introduced ultimately creating a less-than-ideal training set used for constructing the ANN model. The three areas related to the human content analysis (the rubric, coder training, and message categorization) are addressed during the second experiment, and those changes are described in the section of Chapter 4 entitled “Modifications to the Human Content Analysis.” An explanation of the impact of those modifications is described in the section below.

Second Experiment

For the second experiment, changes were made to reduce the error in the human content analysis section of the model. Figure 22 shows three areas in the human content analysis portion of the diagram where error may occur, and a description of the changes made to the rubric, coder training, and message categorization is presented in the section of Chapter 4 entitled “Modifications to the Human Content Analysis.”

In the second experiment, the ANN model performs less well than all the coders but better than the model from the first experiment. The range of pairwise kappa values among the four human coders extends from coder B's low of 0.816 to coder D's high of 0.879. This displays a narrower range than the first experiment. For the first experiment, the range in pairwise kappa values is 0.243, and the range for the second experiment is 0.063. The reliability between the aggregate of human coders and the ANN model is

0.704. This reliability value is below the range of human coders for the second experiment but is higher than the value from the first experiment (kappa = 0.519). If the ANN model were introduced as the fifth coder, a person could most likely distinguish it from the lineup of human coders based on kappa values alone. Further comparison of the errors the ANN model makes versus the errors the human coders make reveals deeper distinctions between the two and an outline of areas to address in future research.

The Differences Between ANN- and Human-Coded Messages

Two major differences emerge between the way the ANN model codes messages and the way the humans code the same messages. The ANN model confuses non-topical messages with exploration messages and does not perfectly discriminate between exploration and integration messages.

Like the model from the first experiment, the second ANN model also confuses non-topical and exploration messages. Table 12 shows that the ANN model agrees with the aggregate of human coders on 37 non-topical messages, but the ANN model consumed ten human-coded exploration messages. However, the ANN exploration category did not falsely code any human-coded messages. This suggests that the ANN non-topical category overgeneralizes while the ANN exploration category undergeneralizes.

Also like the model from the first experiment, the second ANN model confuses exploration and integration messages. Table 12 shows that the ANN model agrees with the human coders on 36 integration messages, but the ANN model's integration category also picked up five human-coded exploration messages. Like the errors mentioned in the

previous paragraph, the ANN integration category overgeneralizes while the ANN exploration category undergeneralizes.

Overall, the model from the second experiment shows improvements over the first experiment's model in that it possesses higher reliability and it codes messages into all coding categories. However, the second experiment ANN continues to have non-topical/exploration and exploration/integration errors.

Renegade Coding

As in the first experiment, an analysis of renegade coding patterns offers a clearer picture of the differences between the human and ANN coding decisions. Recall that renegade coding occurs when one coder (human or ANN) makes a coding decision unlike any other human coder. A closer look at the 100 reliability messages coded by all four coders and the ANN model reveals a potential source for the confusion between the non-topical and exploration messages. The ANN codes unlike any human coder more often than the human coders, but the human coders also exhibit independence.

Table 20

Renegade Coding for Experiment Two

Coder	Instances of Renegade Coding	Comparison
ANN	11	Compared to the four human coders
A	3	Compared to the other three human coders
B	6	Compared to the other three human coders
C	3	Compared to the other three human coders
D	0	Compared to the other three human coders

Note. The ANN model codes unlike any single human coder more often than any single human coder codes unlike any other coder.

Table 20 shows that the ANN model missed 11 messages in which there was uniform agreement among the human coders. Seven of those errors are instances in which all human coders rated the message as an exploration message, but the ANN model coded it as non-topical. Similarly, looking at the instances of renegade coding among the human coders, Coder B had six errors in which that coder differed from the uniform coding of all the other coders. Of those six disagreements, four were exploration/non-topical errors. All other human coders rated the messages as exploration, but Coder B rated them as non-topical. Coder B differed from the other coders in the same way that the ANN model differed from the human coders. To further confuse the ANN model, one of Coder A's errors was identical to Coder B's described above, and two of coder C's errors were exactly opposite. While the other coders uniformly rated a message as non-topical, Coder C rated it as exploration. Overall, seven

of the 12 human renegade coding errors were non-topical/exploration errors. Since the ANN model is constructed from messages the humans coded, there is reason to believe that error in the ANN model emanates from disagreements among the human coders.

Interestingly, analyzing accounts of human renegade coding shows the other primary error found in the ANN model, exploration/integration errors. While seven of the 12 human renegade coding errors were non-topical/exploration errors, the remaining five were exploration/integration errors. Three of those are instances in which one human coder rated a message as exploration while the remaining coders rated it as integration. The remaining two errors are the opposite; one coder rated the message as integration that the others considered to be exploration. Again, confusion in the ANN model's training set will most likely manifest itself as confusion in the ANN model.

Explanation of the Differences Between ANN- and Human-Coded Messages

Looking at Figure 22, there are three areas where the human content analysis may introduce error: the rubric, coder training, and message coding. Since the 12 human renegade coding errors are the same type as the ANN renegade coding errors, it would appear that coder error has indeed been translated into ANN error. The similarity in coding error also suggests that there is error in those areas each coder shares. Of the variables this study controls, each coder shares the same rubric and in the second experiment the same training experience. First, the rubric may cause coders not to code uniformly. This divergence from uniformity exists in Coder B's non-topical/exploration errors. Since Coder B made errors unlike other coders in 6 of 100 cases, it would appear that coders A, C, and D applied the rubric more systematically than coder B. The error then lies in either the coder training or the message coding. It could be that during coder

training, Coder B required more coaching on non-topical/exploration errors. It may also be that the Coder B did not apply the rubric as stringently as the others during coding. During training, the coders often mentioned that upon looking at the message a second time, they can see that they made a mistake, but could not explain it. The coders would say of these errors, "Yes, I just made a mistake in coding. It's clear that that message is not what I coded it as." Indeed, the human content analysis introduces error into the ANN model indicating that further improvements can be made to the rubric and coder training; however, little can be done to correct obvious human error.

As in the first experiment, another explanation for the difference between ANN and human message categorization may lie within the automatic content analysis. The explanation of errors for the first experiment suggests that translating messages into an array of numbers may introduce error. This is one area that was not modified between the two experiments which means that any error this caused in the first experiment would also appear in the second experiment. Certainly, one area that the numeric description of messages must address is identifying parts of speech. The current, numeric description does not identify parts of speech which means that messages with words like “account” contain all meanings of the word including a record of events, a list of financial transactions, and to allow for as in “take into account.” This is an important, technically possible, yet labor-intensive task requiring the technical capability to accurately identify the part of speech of each word in every message.

Overall, though the ANN model from the second experiment performs better than the one from the first experiment, the gap between the kappa values of the human coders and ANN model is larger for the second experiment than the first. By looking at the

errors in the second experiment, the ANN model makes roughly the same errors that the set of human coders makes, it just makes more of them. This suggests that future improvements should be made to the automatic content analysis portion of the model, to improving the numeric description of messages and improving the self-defined dictionary so that it better discriminates among cognitive presence categories.

Explaining the Shift in Exploration and Integration Decision Space Between Experiments

Looking at the Garrison, Anderson, and Archer (2001) study, the pilot study mentioned in Chapter Three, and the first experiment, messages are distributed approximately the same. About half of the messages are non-Cognitive or non-topical, a little under 10% are triggering events, about 25% to 30% of the messages are exploration messages, about 10% are integration and between 0% and 4% of the messages are resolution messages. The second experiment, however, distributes exploration and integration messages differently. Looking at the 100 messages coded to measure reliability among coders (see Table 12), the aggregate of human coders codes 22% as exploration messages and 36% as integration. This shows a shift in the exploration and integration decision space in which exploration has shrunk while integration has grown. This section seeks to explain that shift in decision-making.

Since the shift is noticed in the human coding, studying the human content analysis will reveal the most likely causes of the shift. Looking again at the human content analysis portion of Figure 22, there are three areas within the human content analysis to study: the rubric, coder training, and message coding. Intentional changes in the rubric most likely account for the greatest shift in decision space. The rubric used from the first experiment (see Appendix A) was offered from Garrison, Anderson, and

Archer (2001) and modified to fit this context. A closer look at the subcategories used to describe exploration and integration from that rubric offer a source of coder confusion. Specifically, coders expressed confusion over the “divergence within” and “divergence among” subcategories of the exploration category and over the “convergence within” and “convergence among” subcategories of the integration category. Coders were told that a major difference between exploration and integration was justification. In general, a message is an exploration message if it offers no substantiation and an integration message if it offers some substantiation. The coders became confused over poor or illogical substantiation claiming that the rubric was unclear on messages in which the student was clearly justifying a claim but doing so poorly. The “divergence within” and “divergence among” subcategories forced them away from looking at substantiation to looking at arguments that showed disagreement or multiple conflicting ideas. This may have brought about miscoding true integration messages into the exploration category because a well-justified argument disagreed with a previous message and was therefore coded as an exploration message. Further the “convergence among” and “convergence within” subcategories within the integration category may compete with the justification this category requires. A message showing agreement or merging of ideas may not display justification of a claim. Figure 23 show the competing decision logic within the exploration and integration categories. Either a coder makes decisions along the x or y axis, but not both.

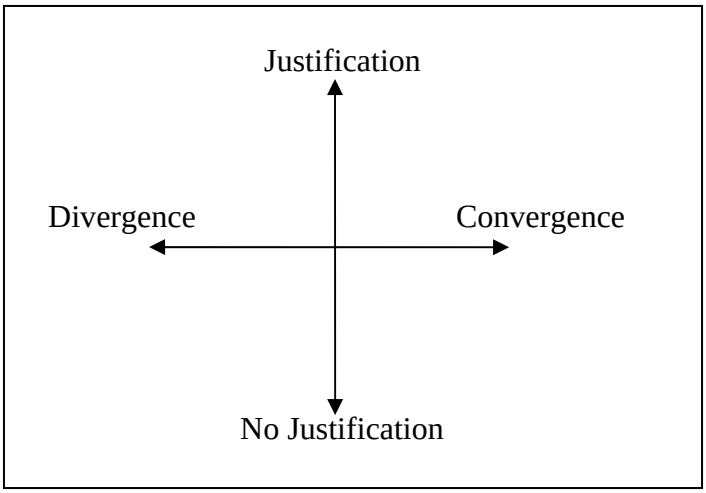


Figure 23. Decision logic within exploration and integration categories.

For the second experiment, intentional modifications were made to the exploration and integration categories focusing on justification over divergence or convergence. This modification has possibly reduced the confusion between the categories, has generated a greater reliability among human coders, and has produced an ANN model with a reliability statistic which agrees with human coders 0.108 greater than the model from the first experiment. However, this modification may mean that cognitive presence cannot be measured accurately with one rubric. It may be that cognitive presence is best measured with one rubric focusing on justification and another focusing convergence/divergence. This study does not seek to validate the constructs underlying cognitive presence, but to clarify the existing constructs in order to improve the reliability between a group of human coders and an artificial neural network. That clarification could not happen without choosing either substantiation over convergence/divergence or vice versa. Perhaps it is better to say that the cognitive presence model used in the second experiment is the substantiation variant of the cognitive presence model. As such, this variant may reveal more integration messages.

Another explanation could be the message set itself. Many instructors from the second experiment required their students to engage in role-play activities. These activities require students to adopt the persona of a historical figure and to post messages in character. Such messages fit most cleanly into the “interpretation” and “synthesis” subcategories of the integration category of the second experiment rubric. Clearly, adopting the persona of a historical figure requires students to interpret a character’s experience, an act often requiring the integration of ideas. Here, the reason for more integration messages may simply be that the instructional strategy used in these courses may lend itself to a greater number of integration messages than those from the previous studies.

Question Two: Sample ANN Analyses

This work lays the groundwork for an inexpensive, rapid, frequent, and objective measure of expressed critical thinking. The second part of the research question describes the information expected from an ANN content analysis tool. Chapter 4 displays the type of reports instructors and administrators may expect from an automatic content analysis tool applied to online discussions. The broad implication of such reports is that it offers objective evidence that an instructor may use to modify instruction and that it may be used as an action research tool promoting more specific questions about the quality and use of online discussions.

Moreover, with a nationwide focus on accountability among K-12 public schools, school leaders are demanding this type of measure, one that allows them to gauge the progress of their students between state tests. State tests, often offered only once a year, are often not reported until many months and in some cases almost a year after they are

administered, yet these tests are used to determine whether some schools receive federal funding. Where such accountability measures have been in place, especially in Texas, district-level school leaders say that frequent monitoring is critical (Skrla, Scheurich, & Johnson, 2000; Massell, 2000). Though the automatic content analysis tool has not been analyzed to predict success on state tests nor has it been used with secondary students, researchers and education leaders may adapt the method applied in this research. That is, this work outlines how to use an artificial neural network to build a tool that objectively measures one aspect of student learning. Researchers may verify the reliability of such tools by comparing cognitive presence values to standardized test performance to determine how well such tools predict success on state measures. Though much work would have to be done, it is certainly feasible to develop a battery of predictive and reliable tools that allow teachers to monitor their students' performance on what has previously gone unmeasured and ultimately to adjust their own approach. Skrla, Scheurich, & Johnson (2000) refer to this latter aspect, using data to inform and alter what happens in the classroom, as a critical step to improving student achievement.

Limitations and Bias

At the end of Chapter 3, the limitations and bias of the methods is presented. The same methodological limitations and biases apply, and this section outlines additional limitations and biases revealed during the study.

One area of concern emerged during the comparison of ANN models from the first and second experiments. That is, the categories which operationalize cognitive presence may not perfectly define cognitive presence. Most notably, Figure 23 reveals that the rubric forces coders to decide along two competing axes, justification and

convergence/divergence. Certainly, more work is needed to validate the cognitive presence coding categories.

Other limitations and biases emerge in the comparison of models. One model is not equal to another of different parameters, so any comparison is forced. The most notable differences among the full, topic, and section models are that they are constructed from different message sizes. The full model is built from the most messages while the other models are built with considerably fewer messages. As a rule, models built from more cases should outperform those constructed from fewer cases provided the decision logic is similar. Therefore, the full model should outperform the other models. However, the decision logic among the cases that built each model may not be similar. Topic and section models most likely contain more homogenous messages, while the full model most likely contains more heterogenous messages. In both experiments, section models are rated by one coder which means that the accuracy of the model depends upon the consistency of the coder. Consistency of language used within the course may also be a factor. Without the ability to hold these variables, message number and homogeneity, constant, a better comparison is not possible.

For the spelling analysis, bias is controlled by using an external mechanism, Microsoft Word©, to define a spelling error. This reduces human bias, but introduces the bias of the software in making misspelling decisions. Second, a more thorough analysis may be performed by employing the time-consuming, though accurate, task of correcting every message from an entire course, constructing an ANN model from the completely corrected set, and then analyzing the coding decisions of each model. Given the resources required, this effort is not feasible. The spelling analysis does, however, point

to one broad bias. Messages filled with errors may not be coded as highly as those without errors. Coders may carry the false assumption that an integration message is inherently better than an exploration message and may code error-filled messages into what they perceive to be a lower category. Appropriate training can correct for this bias, but it must be done deliberately.

Additionally, student cognitive presence within on-line messages is dependent upon a number of factors such as verbal ability, comfort with technology, comfort with communicating to a group of unknown course participants, and undistracted time to devote to the course. In private correspondence, Terry Anderson (2002) mentions that in the original Garrison, Anderson, and Archer work (2000, 2001) little attention is devoted to validity. It may be that both a human and an automatic content analysis of cognitive presence does not measure pure cognitive presence. Factors, such as those mentioned above, may be inextricably tied to cognitive presence meaning that future research may seek to understand the degree to which other factors confound the cognitive presence measure. Also, the scope of this study is to analyze the cognitive effort displayed through on-line messages within on-line courses. Other aspects of the course such as quizzes, exams, telephone conferences, face-to-face meetings and written assignments, contain indicators of cognitive effort which lie outside that scope.

Potential users of automatic content analysis tools to measure cognitive presence are cautioned against using the tool as either a measure of individual student performance or cognitive ability. Instead, this tool should be used to evaluate the structure and delivery of the course as well as the instructor's teaching approach.

Major Contributions of this Study

This study makes three major contributions to the research literature. It demonstrates a method for transferring a human decision-making process to a computational decision-making process; it demonstrates that a computer model may categorize student-generated text with near-human accuracy; and it sets expectations instructors should demand from a content analysis of online discussion list messages.

This study adds to the body of knowledge by showing that an artificial neural network may be used to perform a task traditionally reserved for humans alone. This study is an example of transferring some of the cognitive load of educational evaluations to a computer. This work provides a road map for developing an artificial content analysis tool to measure the other dimensions within Garrison, Anderson, and Archer's (2000) model, namely social presence and teaching presence. Looking beyond Garrison, Anderson, and Archer's (2000) model, this approach may be applied to other areas of educational decision-making. Wherever human evaluations are made, we may also ask whether the process of making those decisions may be captured and performed using a computational model. The major contribution, in this case, is that this work expands what we typically expect computers to do.

This study also adds to the body of knowledge in the research literature by demonstrating a method of computationally evaluating student-generated text. Scant education literature is devoted to using computers to analyze text, and to date, even less of that literature is devoted to analyzing text with the goal of improving instruction. Most computational text-analysis literature is devoted to grading students (see Burstein, Marcu, Andreyev, & Chodorow, 2001; Hearst, 2000).

Finally, this study adds to the body of knowledge by demonstrating the information instructors should reasonably expect from a content analysis whether that content analysis is conducted by humans or by using artificial intelligence. Specifically, it sets the quantitative expectation that all units of analysis, in this case all messages, can be measured instead of a sample. This expectation allows for a much deeper analysis of the subgroups within the message set. For example, it allows for analyses by instructor, by course topic, by weekly topics within a course, and by student. This work sets the expectation that performing a content analysis should not be an end in itself but should provide a variable enabling further, deeper analysis.

Suggestions for Future Research

Figure 24 shows an outline of the research efforts that would take this work from its current state to widespread adoption among educators. From its current state, there are three research options, tracks one to three.

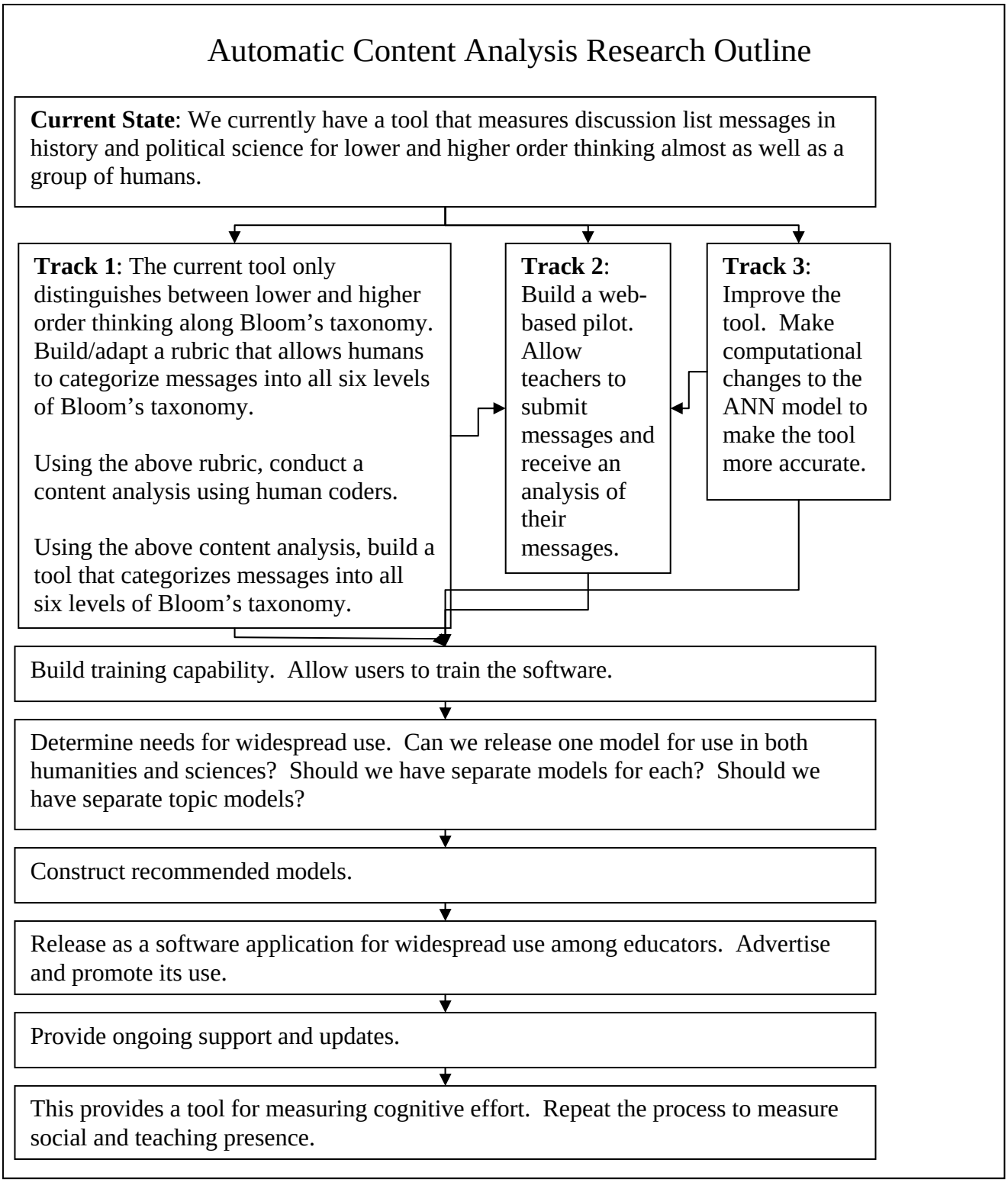


Figure 24. Automatic content analysis research outline.

The first track requires expanding the coding categories of the current tool. Currently, messages are coded into four categories (non-topical, triggering event, exploration, and integration). The last two, exploration and integration are broadly considered to encompass lower- and higher-order thinking respectively along Bloom's taxonomy (Bloom & Krathwohl, 1956). There are six levels to Blooms taxonomy: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Exploration messages, therefore, include behaviors such as recognizing recalling information, whereas integration messages require behaviors such as inferring, planning, and appraising (see Domin, 1999, p. 109). Caution should be taken when expanding the number of coding categories for content analysis research. More categories will require improving the coder training to ensure that coders place messages into the correct category, and more coding categories may also result in poorer reliability. One possible direction would be for coders to code messages in two passes. In the first pass, the coders place messages into the four categories described in this research study (non-topical, triggering event, exploration, and integration). In the second pass, the coders look specifically at messages in the exploration and integration categories. All exploration messages are coded into Bloom's lower order categories, and all integration messages are coded into Bloom's higher order categories. The same two-pass strategy may also be used when building the artificial neural network (ANN) models. One model may code broadly into the current four categories while two subsequent models code only exploration and integration messages.

The second track requires releasing the automatic content analysis tool for instructor use. Studying how instructors use it and structuring feedback from instructors

on the usefulness of this tool provides a list of needs for improving the usability of the tool. This answers such questions as: “How easy is it for instructors to submit messages for analysis?” “How long does it take to provide feedback?” “Is this lag-time acceptable?” “Do instructors use the feedback?” “How do instructors use the feedback?” “What changes have instructors made to their delivery of instruction?” “Have these changes resulted in improvement in student learning?”

The third track involves improving the artificial neural network (ANN) model to make the tool more accurate. Results from the second experiment show that human coders have a reliability of kappa = 0.848 whereas the artificial content analysis tool is about has a reliability of kappa = 0.704.. This indicates a reliability gap of approximately 0.15 meaning that humans will categorize messages with 15% more accuracy than the current ANN. Improving the model would mean reducing this 15% gap between the human coders and the automatic content analysis tool. This effort requires expertise in artificial neural networks and computational linguistics.

The three research tracks all feed into a final set of research and development areas including building training capability, determining needs for widespread use, developing recommended models, releasing and supporting the software, and reiterating all the steps to build social and teaching presence models. Of these, the first two and the last one require research efforts.

To disseminate an automatic content analysis tool among instructors, we must first determine whether users should be given the ability to train the tool to accurately categorize messages unique to each instructor’s setting. We may look to other tools using artificial neural network (ANN) technology for guidance. Most notably, ANNs are

used in speech and handwriting recognition software. Most speech and handwriting recognition software may be used both with and without training. The mode not requiring training allows nearly any person to use the software, but the recognition is not as accurate as models trained to a specific user. Just as a person would train speech and handwriting recognition to recognize their own personal voice or handwriting style, so instructors may also train ANNs to categorize messages unique to their own instructional context. We do not know for certain that one model is generalizable across multiple topics, though one model has been constructed from both history and political science courses. We do not know, for example, that a model built from messages in which an instructor predominantly uses one strategy (e.g. role play) will most correctly categorize messages from a course in which an instructor uses another strategy. Therefore, future iterations of discussion list analysis tools may consider allowing instructors to train the model to their own specific context. A base model may be provided and the instructor may simply submit sample messages from his/her course which exemplify cognitive presence categories. The instructor tweaks the model to his/her own course.

With networked computers as the norm in educational institutions, trained models may be created not only at the instructor level but at the department or university level. This means that a group of instructors may train a single model that most closely matches departmental instructional strategies. This means that more instructors may submit more messages to train the model which improves the model's accuracy. We may have already reached the threshold in the accuracy of non-trained models. Adaptive models allowing for user-training may be the only way to improve the accuracy of categorizing messages.

Introducing an artificial neural network content analysis tool into instruction could change instructional strategy decisions. Such a tool would provide a measurement of cognitive effort at any time during a course providing immediate feedback which may confound or support other objective measures of the effectiveness of an instructional strategy such as assessments and feedback surveys.

Second, the research efforts involved in determining needs for widespread use are similar to track 2 above. The same research questions focusing on meeting instructor needs are applicable to determining widespread use; however, these research efforts should be expanded to all stakeholders including administrators and students as well as instructors. At this point, research may also be conducted to determine whether an automatic content analysis tool focusing on cognitive presence predicts student achievement.

Finally, Garrison, Anderson, and Archer's (2000) Community of Inquiry model is comprised of three broad domains: cognitive presence, social presence and teaching presence. This work focuses solely on developing an automatic content analysis tool for the cognitive presence domain. This work may be replicated to build tools which assign social presence and teaching presence values to messages as well. Once such tools are built, researchers may then analyze the rich correlation among the domains. This effort would begin answering questions such as: "What percentage of the variability in cognitive presence is explained by social presence or teaching presence?" "Do students exhibiting high social presence also exhibit high cognitive presence?" "Is social presence an inhibitor or a catalyst for cognitive presence?" "How well to the three domains predict student achievement?"

Summary

Overall, an ANN can be constructed to categorize discussion list messages with near-human accuracy. Improvements made to the second experiment indicate that the human content analysis can be improved; however, the two models did not keep pace with each other. A 24.8% improvement in the human content analysis from experiment one (kappa = 0.6) to experiment two (kappa = 0.848) resulted in only an 18.5% improvement in the ANN model from experiment one (0.519) to experiment two (0.704). Since most of the improvement efforts between the two experiments were made to the human content analysis, future work should focus on improving the elements within the automatic content analysis, the second major area of the model (see Figure 22). The results also indicate that spelling errors have little effect and that a full model constructed from both history and political science courses is not only a simpler solution but also no worse than a combination of topic and section models. Finally, users of automatic content analysis tools should expect to perform analyses on the entire population of cases and should expect to describe fine-grained detail of their courses.

In the first chapter, I outlined a brief history of technology use in education. This history provides repeated examples of initial exuberance over the promise of a technology to improve education only to be followed by a retrospective look which finds little, if any, impact. Kozma (1994) suggests that if we are to break out of this cycle, we must use our technology do to what cannot be otherwise be done. If the combination of computers and the internet are to have any lasting impact, we must use them to create learning opportunities, strategies, and environments that otherwise would not exist. This work gives researchers the groundwork for a tool that allows them to study online

discussions in a way that we have never before experienced. It enables the analysis of every message instead of a sample; it does so with near-human accuracy; with further development, it can be used in action research projects with very little resources; it can be deployed at departmental levels to understand the quality of messages among multiple courses, and it serves as one of potentially many similar tools upon which to confirm theories of online learning. This work provides one of many necessary examples of using our technology to encourage fundamentally different learning opportunities which is critical if we are to break out of Reiser's (2002) hundred-year cycle.

References

Alston, W. (1994). Belief-forming practices and the social. In F. F. Schmitt (Ed.), *Socializing epistemology: The social dimensions of knowledge*. Lanham, MD: Rowman and Littlefield.

Alston, W. (1996). *A realist conception of truth*. Ithaca, NY: Cornell University Press.

ALT distance education student profile survey. (2001). Retrieved October 23, 2003, from http://alt.usg.edu/research/studentprofile_2001_crossterm.pdf

ALT distance education student profile survey. (2002). Retrieved October 23, 2003, from http://www.alt.usg.edu/research/studentprofile_2002_crossterm.pdf

Angeli, C., Bonk, C., & Hara, N. (1998, November). Content analysis of online discussion in an applied educational psychology course. Retrieved February 20, 2002, from http://crlt.indiana.edu/publications/crlt98-2.pdf

Bloom B. S., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, book 1: Cognitive domain.* New York: Longmans, Green.

BonJour, L. (1978). Can empirical knowledge have a foundation? *American Philosophical Quarterly, 15*, 1–13.

Bullen, M. (1998). Participation and critical thinking in online university distance education. *Journal of Distance Education/Revue de l'enseignement à distance: 13*, 2. Retrieved January 15, 2002, from http://www.icaap.org/iuicode?151.13.2.1

Burstein, J., Marcu, D., Andreyev, S., & Chodorow, M. (2001). Towards automatic classification of discourse elements in essays. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 90-97.

Boyles, D. (2000). Students as knowers: An argument for justificatory social epistemology by way of blind realism. *Social Epistemology, 14*(1).

Center for Social Organization of Schools. (1983). School uses of microcomputers: Reports from a national survey (Issue no. 1). Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools.

Chiang, W. C., Urban, T. L., & Baldridge, G. W. (1996). A neural-network approach to mutual fund net asset value forecasting. *Omega-International Journal of Management Science, 24*(2), 205-215.

Code, L. (1999). Is the sex of the knower epistemologically significant? In L. P. Pojman (Ed.), *The theory of knowledge: classical and contemporary readings*. London: Wadsworth Publishing Co.

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

Daft, R., & Lengel, R. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Danielson, W. A., & Lasorsa, D. L. (1997). Perceptions of social change: 100 years of front-page content in *The New York Times* and *The Los Angeles Times*. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing inferences from texts and transcripts* (pp. *103-115).* Mahwah, NJ: Lawrence Erlbaum.

Doerfel, M. L., & Barnett, G. A. (1999). A semantic network analysis of the International Communication Association. *Human Communication Research, 25*, 589-603.

Domin, D. S. (1999). A content analysis of general chemistry laboratory manuals for evidence of higher-order cognitive tasks. *Journal of Chemical Education 76*(1), 109-112.

Dybowski, R., Weller, P., Chang, R. & Grant, V. (1996). Prediction of outcome in critically ill patients using artificial neural network synthesized by genetic algorithm. *The Lancet, 347*(9009), 1146-51.

Ennis, R. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: theory and practice.* New York: W.H. Freeman.

Evans, W. (2000). Teaching computers to watch television: Content-based image retrieval for content analysis. *Social Science Computer Review, 18*, 246-257.

Evans, W. (2001). Computer environments for content analysis: Reconceptualizing the roles of humans and computers. In O. V. Burton (Ed.), *Computing in the social sciences and humanities*. Champaign, IL: University of Illinois Press.

Fahy, P. J., Crawford, G. & Ally, M. (2001, July). Patterns of interaction in a computer conference transcript. *International Review of Research in Open and Distance Learning, 2*(1). Retrieved January 17, 2002, from http://www.icaap.org/iuicode?149.2.1.4

Feldman, R. (1994). Good arguments. In F. F. Schmitt (Ed.), *Socializing epistemology: The social dimensions of knowledge*. Lanham, MD: Rowman and Littlefield.

Franzosi, R. (1995). Computer-assisted content analysis of newspapers: Can we make an expensive research tool more effective? *Quality and Quantity, 29*, 157-172.

Fausett, L. (1994). *Fundamentals of neural networks: Architectures, algorithms, and applications.* New Jersey: Prentice Hall.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists.* Thousand Oaks: Sage.

Garrison, D. R. (1992). Critical thinking and self-directed learning in adult education: An analysis of responsibility and control issues. *Adult Education Quarterly, 42*(3), 136-148.

Garrison, D. R., Anderson, T., Archer, W. (2000). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education, 2*(2-3), 87-105.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7-23.

Gay, G., Pena-Shaff, J., & Martin, W. (2001). An epistemological framework for analyzing student interactions in computer-mediated communication environments. *Journal of Interactive Learning Research 12*(1), 41-68. Retrieved July 10, 2003, from http://dl.aace.org/6402

Georgia distance learning numbers increase dramatically; more statewide have internet access, according to Georgia GLOBE research. *Yahoo!Finance*. Retrieved December, 17, 2001, from http://biz.yahoo.com/prnews/011211/attu016_1.html

Gilbert, M. (1994). Remarks on collective belief. In F. F. Schmitt (Ed.), *Socializing epistemology: The social dimensions of knowledge*. Lanham, MD: Rowman and Littlefield.

Goldman, A. (1999). *Knowledge in a social world.* Oxford, England: Oxford University Press.

Griffin, E. (1997). Information theory of Claude Shannon and Warren Weaver [Electronic version]. In E. Griffin, *A first look at communication theory* (5th ed. Chapter 4). New York: McGraw-Hill. Retrieved September 3, 2004, from http://www.afirstlook.com/archive/information.cfm?source=archther

Hara, N. (2000, April). *Visualizing tools to analyze online conferences.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved January 17, 2002, from http://www.ils.unc.edu/~haran/paper/fca/fca_aera.html

Hara, N., Bonk, C., & Angeli, C. (1998, March). *Content analysis of online discussion in educational psychology courses.* Paper presented at the meeting of the Society for Information Technology and Teacher Education 98, Washington, DC. Retrieved January 17, 2002, from http://www.coe.uh.edu/insite/elec_pub/HTML1998/re_hara.htm

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22-37.

Henri, F. (1992). Computer conferencing and content analysis. In A. R. Kaye (Ed.), *Collaborative learning through computer conferencing: The Najaden papers* (pp. 115-136). New York: Springer.

Howell-Richardson, C., & Mellar, H. (1996). A methodology for the analysis of patterns of participation within computer mediated communication courses. *Instructional Science, 24*, 47-69.

Kanuka, H., & Anderson, T. (1998). On-line social interchange, discord, and knowledge construction. *Journal of Distance Education, 13*(1), 57-74.

Kitcher, P. (1994). Contrasting conceptions of social epistemology. In F. F. Schmitt (Ed.), *Socializing epistemology: The social dimensions of knowledge*. Lanham, MD: Rowman and Littlefield.

Knight, J. E. (1990). Coding journal entries. *Journal of Reading, 34*(1), 42-46.

Kuehn, S. A. (1994). Computer-mediated communication in instructional settings: A research agenda. In *Communication Education 43*(2), 171-184.

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Landis, J. R., & Koch, G. G., (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

MacKinnon, G. R. (2000). The dilemma of evaluating electronic discussion groups. *Journal of Research on Computing in Education, 33*(2), 125-131.

MacKinnon, G., & Aylward, L. (2000). Coding electronic discussion groups. *International Journal of Educational Television 6*(1), 53-61.

McDonald, J. (1998). Interpersonal group dynamics and development in computer conferencing. *American Journal of Distance Education, 12*(1), 7-25.

McCluskey, F. D. (1981). DVI, DAVI, AECT: A long view. In J.W. Brown & S. N. Brown (eds.), *Educational media yearbook: 1981*. Littleton, CO: Libraries Unlimited.

McKlin, T., Harmon, S., Jones, M., & Evans, W. (2001). Cognitive presence in web-based learning: A content analysis of students' online discussions. *Proceedings of the 2001 Association for Educational Communications and Technology International Convention.*

McQuail, D. and Windahl, S. (1981). *Communication models for the study of mass communications.* New York: Longman.

Misulis, K. (1997). Content analysis: A useful tool for instructional planning. *Contemporary Education, 69*(1), 45-50.

Moore, C. (2001, December 11). E-learning leaps into the limelight. *CNN.Com/Sci-Tech*. Retrieved December 17, 2001, from http://www.cnn.com/2001/TECH/internet/12/11/elearning.leaps.idg/index.html

Newman, D. R., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology: An Electronic Journal for the 21st Century, 3*(2), 56-77.

Pattern Recognition Workbench (Version 2.1.253) [Computer software]. (1992-1997). Waltham, MA: Unica Technologies, Inc.

Pattie, D. C., & Haas, G. (1996). Forecasting wilderness recreation use – neural network versus regression. *AI Applications 10*(1): 67-74.

Picciano, A. G. (1998). Developing an asynchronous course model at a large, urban university. *Journal of Asynchronous Learning Networks, 2*(1). Retrieved July 13, 2002, from http://www.aln.org/publications/jaln/v2n1/pdf/v2n1_picciano.pdf

Pojman, L. P. (2001). *What can we know: An introduction to the theory of knowledge, second edition.* U.S.: Wadsworth.

Pomeroy, D. (1993). Implication of teachers' beliefs about the nature of science: comparison of the beliefs of scientists, secondary science teachers, and elementary teachers. *Science Education, 77*, 261-278.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. .*Journal of Applied Communication Research*, *27*, 258-284.

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing media messages: Using quantitative content analysis in research.* Mawah, New Jersey: Lawrence Erlbaum.

Reiser, R. A. (2002). A history of instructional design and technology. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and Issues in Instructional Design and Technology* (pp. 26-53). Upper Saddle River, New Jersey: Merrill Prentice Hall.

Romiszowski, A. J., & Mason, R. (1996). Computer-mediated communication. In D. Jonassen (Ed.), *Handbook of research for educational communciations and technology, first edition* (pp 438-456). New York: MacMillan.

Romiszowski, A. J., & Mason, R. (2003). Computer-mediated communication. In D. Jonassen (Ed.), *Handbook of research for educational communciations and technology, second edition* (pp. 397-431). New York: MacMillan.

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (1999). Assessing social presence in asynchronous text-based computer conferencing. *The American Journal of Distance Education, 14*(2). Retrieved March 7, 2001, from http://cade.athabascau.ca/vol14.2/rourke_et_al.html

Saettler, P. (1968). *A history of instructional technology*. New York: McGraw-Hill.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions.* San Francisco, CA: Jossey Bass.

Seaver, J. D., Smith, T., & Leflore, D. (2000). Constructivism: A path to critical thinking in early childhood. *International Journal of Scholarly Academic Intellectual Diversity, 4*(1).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379-423.

Shera, J. (1970). *Sociological foundations of librarianship*. New York: Asia Publishing House.

Sherry, L., Tavalin, F., & Billig, S. H. (2000). Good online conversation: Building on research to inform practice. *Journal of Interactive Learning Research, 11*(1), 85-127.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* Toronto, ON: Wiley.

Shrout, P. E., & Fleiss, J. L., (1979), Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sproull, L., & Keisler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*, 1492-1513.

Tsai, C. (1999). Content analysis of Taiwanese 14 year olds' information processing operations shown in cognitive structures following physics instruction. *Research in Science and Technological Education, 17*(2), 125-138.

Weiss, R., & Morrison, G. (1998). Evaluation of a graduate seminar conducted by listserv. *Proceedings of Selected Research and Development Presentations at the National Convention of the Association for Educational Communication and Technology (AECT).*

Zhu, E. (1998). Learning and mentoring: Electronic discussion in a distance-learning course. In C. J. Bonk & K. S. King (Eds.), *Electronic collaborators: Learner-centered technologies for literacy, apprenticeship, and discourse* (pp. 233-259). New Jersey: Lawrence Erlbaum Associates, Inc.

Appendices

Appendix A

Coder Training Rubric for Experiment One

The following rubric is adapted from that used by Garrison, Anderson, and Archer (2002) with examples and instructions provided with assistance from Patricia Oliver.

The task will involve the following:

1. Read a student message
2. Select the most appropriate Cognitive Presence Subcategory for this message
3. Indicate the Cognitive Presence Subcategory by writing the corresponding Cognitive Presence Subcategory Code in the space provided for message's Code.

General Guidelines:

If a message contains multiple categories, determine the essence of the message, and code the message accordingly. If you are unable to determine the essence of the message, then select the category that is highest within the cognitive hierarchy. For example:

> "Greetings fellow students. Several of us are planning to meet at Rocky Mountain Pizza on Friday night. Please feel free to join us. On another note, did anyone experience problems logging in to take the quiz? I had some problems last night. Name_1's discussion about the negative aspects of colonialism reminded me of some disturbing elements within Conrad's Heart of Darkness. I find it hard to comprehend Cortes and some of the other explorers. Cortes must have been somewhat like Kurtz. It would be interesting to study the different motivational forces at work among different explorers and colonists. What do you all think was the primary motivational factor for most explorers or for

> colonists? While the text, Historical narratives and his letters all mention gold as a motivating factor for Cortes, his letters also indicate religious motivations. From his actions during the incident described in the text we might infer some additional motivating forces such as power. I agree with Name_3's and Name_4's characterization that Cortes' primary motive was personal gain (Power and Money). I've researched several websites re: Cortes. The Catholic Encyclopaedia's Website mentions his ambition and desire for power and that he had 'no excess of scruples in morals'. That doesn't appear to be a very ringing religious endorsement. Additionally, the article characterized his use of the Church as primarily utilitarian. The U. Michigan website describes him as 'the perfect Machiavellian blend of will power and good luck'. These sources, along with the text, his letters and Historical narratives all lend credence to the view that Cortes was motivated by ambition, power and wealth".

An analysis of the above message yields the following:

1. The greeting and invitation to Rocky Mountain Pizza fall within the Unrelated (UR) category.
2. The statement and question regarding problems logging in to take the quiz fall within the Technical Support (TS) category.
3. The reference to Name_1's discussion and "disturbing elements within" <u>Heart of Darkness</u> contain elements of Personal Narrative (PN).
4. The statement "Cortez must have been somewhat like Kurtz" is a Leap to Conclusion (LC).
5. The statement and question about explorers' and colonists' motivations show a Sense of Puzzlement (SP).
6. The statements referring to conflicting information related to Cortes' motivations from the Text and Cortes' letters show Divergence Within (DW).
7. The support of prior messages augmented with the information from the websites, text, Cortes' letters and historical narrative shows Convergence Among (CA) the messages and these sources.

Although the above message contains elements that fall into a significant number of Cognitive Presence Subcategories, the majority of the message is in support of Convergence Among (CA). The first reference to explorers' and colonists' motivations (SP) are a lead in to a discussion that culminates in Convergence Among. The example message would be coded as CA because that is the essence of the message and, of the cognitive subcategories present within the message, it is the highest within the hierarchy.

Table A1

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
Not Cognitive	Unrelated	UR	Statements that are not related to the course concepts, nor are they related to technical issues regarding the eCore course; Social Pleasantries.	"Do you have plans for Friday night? Several of us are planning to meet at Ruby's."

Course Management	CM	Statements regarding logistics or management of the course (materials, schedules, assignments)	“When do we meet next?” “The bookstore has finally obtained some additional course texts.” “When is the exam?”

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Technical Support	TS	Providing technical support or assisting others with technical issues related to the course.	"When I logged on last night, the server was unavailable. Did anyone else have similar trouble?" "First download v 5.5, and execute the install".
	External Reference	ER	Reference to an external source for additional information.	"The following link has some information which you might find useful: http://www.newinfo.org"

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
Triggering Event (Evocative)	Recognizes Problem	RP	Presenting background information that culminates in a question.	"In an earlier post, Name_2 reminded us that their diet was very similar to ours. I wonder if the frequency of diet related diseases in their culture was similar that in our culture".

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Sense of Puzzlement	SP	Asking questions, Messages that take discussion in a new direction, Sense of Puzzlement	"In H's earlier post, he mentioned that it was difficult to fit half day kindergarten into the 5 year olds' busy schedules. If this is true, how busy will the children be by 4th grade? What does all of this fast-paced living do to their health?"
Exploration (Inquisitive)	Leap to Conclusion	LC	Offers unsupported conclusions	"They must have been very angry about the intrusion into their culture."

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Personal Narrative	PN	Personal narratives containing descriptions or facts that are not used as evidence to support a conclusion.	"I found their actions to be quite disturbing." "I have a friend who grew up on an Indian Reservation and he said..."
	Brainstorming	BS	Adds to the established points but does not systematically defend, justify or develop this addition.	"I'm beginning to wonder if this might just be the key…" "What if…" "Here's an idea..."

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Information Exchange	IE	Descriptions or facts that are not used as evidence to support a conclusion, but that are course or topic related.	"The author states…" "One of the narratives was completed shortly after the events occurred. The second narrative was completed many years later".
	Suggestion	SU	Suggestion(s) for consideration. Author explicitly characterizes the message as exploration.	"I'm beginning to wonder if this might just be the key… What do you think? Am I on target?" "Does that seem about right?" or "Am I way off base?" "What do you think?"

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Divergence Among	DA	Divergence or disagreement within the online community; unsubstantiated contradiction of prior ideas, divergence among messages	"I disagree with that statement…" "I'm not so sure about that..."

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Divergence Within	DW	Presenting many different ideas or themes within one message; divergence within a message.	"Narrative one states ….. while narrative 2 states the opposite." "While one author suggests that the settlers were quite austere, other evidence suggests that they held, and enjoyed parties, games and other social events."

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
Integration (Tentative substantiation, combining ideas)	Creating Solutions	CS	The writer explicitly characterizes the message as a solution.	"I believe this is the key because…" "The following hypothesis ties it all together…"

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Synthesis	SY	Connecting ideas, integrating information from various sources – textbook, articles, personal experience	"There is definitely social interaction in WebCT. As mentioned in Khan, p. 363, A free exchange of ideas, opinions, and feelings is the lifeblood of collaborative learning. In evaluating the success of this class, we can't overlook the value of the open forum we have enjoyed on the bulletin board."

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Convergence Within (one message)	CW	Justified, developed, defensible, yet tentative hypothesis.	"The settlers must have been less austere than the author proposes. The archeological evidence taken together with the social events described within the diaries and the town records all point towards the settlers enjoying an active social life".

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
	Convergence Among (one or more messages or sources)	CA	Reference to one or more prior message(s) followed by substantiated agreement, e.g., "I agree because…"; building on and adding to others' ideas	"I agree with Name12. I think the settlers had an active social life. As well as the town records, personal diaries and archeological evidence, a settler, John E. described a very gay party in letters to England. He discusses the food, drink, music, dancing and games at the party as follows: ……..".

Table A1 (continued)

Coder Training Rubric for Experiment One

Cognitive Presence Category	Cognitive Presence Subcategory	Code	Description	Example
Resolution (Committed)	Resolution	RE	Committed, characterized by applications to the real world, testing solutions and defending solutions.	"Based on the overwhelming evidence, it is apparent that the author's account of the settler's austerity is incorrect. The settlers definitely had an active social life. This is supported by the following: the remains of several musical instruments have been found at the site. Equipment for making, storing and serving wine and ale have also been found at the site. Letters

	exist which describe social occasions in significant detail. Town records and diaries also include accounts of parties and social occasions. The evidence of an active social life in the settlement is overwhelming ……..”.

Appendix B

Coder Training Rubric for Experiment Two

The following rubric is a modified version of the coder training rubric from experiment one which was adapted from that used by Garrison, Anderson, and Archer (2002) with examples and descriptions provided with assistance from Patricia Oliver.

Table B1

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
Non-Topical: 1	Unrelated	Statements that are not related to the course concepts, nor are they related to technical issues regarding the eCore course; Social Pleasantries.	"Do you have plans for Friday night? Several of us are planning to meet at Ruby's."

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Course Management	Statements regarding logistics or management of the course (materials, schedules, assignments)	"When do we meet next?" "The bookstore has finally obtained some additional course texts." "When is the exam?"
	Technical Support	Providing technical support or assisting others with technical issues related to the course.	"When I logged on last night, the server was unavailable. Did anyone else have similar trouble?" "First download v 5.5, and execute the install".
	External Reference	Reference to an external source for additional information with no reference to the course topic.	"The following link has some information which you might find useful: http://www.newinfo.org"

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Simple Restatement or Question	Restatement or simple clarification, or a simple question.	"I meant to say that Montezuma was …"
	Simple Agreement or Disagreement	Statement of agreement or disagreement related to non-course content or unknown content	"I absolutely agree with you." "I don't think I can agree with you on that."
	Topic Undetermined	If you are unable to determine if the message is related to course content.	
	Topical Compliment	Statements such as 'good job' or 'great budget' which do not tell why.	Statements complimenting a team member on a joint project on a piece of submitted work.

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Topical Course Management	Statements or questions which are course related or topical but are course or project management.	"Have you finished your piece on the Revolutionary War?" "I plan to report on the Cuban Missile Crisis."
Triggering Event (2): Problem-posing events, evocative in terms of conceptualizing a problem or issue.	Recognizes Problem	Presenting background information that culminates in a question.	"In an earlier post, Name_2 reminded us that their diet was very similar to ours. I wonder if the frequency of diet related diseases in their culture was similar that in our culture".

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Sense of Puzzlement	Asking questions, Messages that take discussion in a new direction	"In H's earlier post, he mentioned that it was difficult to fit half day kindergarten into the 5 year olds' busy schedules. If this is true, how busy will the children be by 4th grade? What does all of this fast-paced living do to their health?"

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Challenge	One student challenges another student's position or opinion, but does not substantiate, develop or justify.	"Tell me one good thing this person has done with his presidency".

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
Exploration (3): Inquisitive, search for relevant information, inquisitive or divergent process in search for ideas to make sense of a problem	Opinions or Information Exchange	Offers unsupported conclusions. Leaps to conclusions.	"They must have been very angry about the intrusion into their culture."
		Suggestion(s) for consideration. Author explicitly characterizes the message as exploration.	"I'm beginning to wonder if this might just be the key… What do you think? Am I on target?"
		Information, descriptions or facts that are not used as evidence to support a conclusion, but are course or topic related. Facts or descriptions can be divergent.	"The author states…" "One of the narratives was completed shortly after the events occurred. The second narrative was completed many years later".

	Personal narratives containing descriptions or facts that are not used as evidence to support a conclusion.	"I found Montezuma's treatment of the disabled or exceptional to be quite disturbing."
	Brainstorming.	"My friend grew up on an Indian Reservation.."
	Drawing Parallels without offering explanation of relationships.	"I'm beginning to wonder" "What if…"
	Adding to the established points with no justification.	"I think this is similar to the Cuban Missile Crisis."
		"In addition, to your arguments, avoiding war would have caused additional unrest."

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Topical Agreement or Disagreement	Agreement or disagreement within the online community with respect to the subject matter; unsubstantiated contradiction or support of prior ideas, divergence among messages	"I disagree with your assessment of the Columbian Exchange" "I agree with your statement about the British not acting in accordance with their stated principles." "I agree with your position on states rights."

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
Integration (4): Constructi on of a possible solution, or a tentative conversion, or connecting relevant ideas capable of providing insight	Drawing Conclusions	The writer may explicitly characterize the message as a solution. He or she may connect ideas, integrate information from various sources – textbook, articles, personal experience which lead to a conclusion. Evidence of a tentative, yet developed, defensible, hypothesis with some justification. Reasoned comparison and	"I believe this is the key because…" "The following hypothesis ties it all together…" "There is definitely social interaction in WebCT. As mentioned in Khan, p. 363, A free exchange of ideas, opinions, and feelings is the lifeblood of collaborative learning." "The settlers must have been less austere than the author proposes. The archeological evidence

	contrast	taken together with the social events described in diaries all point towards…”.
Substantiated Agreement or Disagreement	Reference to one or more prior message(s) followed by substantiated agreement or disagreement, building on and adding to others’ ideas.	“I agree with Name12. I think the settlers had an active social life. The town records show social events such as…”

Table B1 (continued)

Coder Training Rubric for Experiment Two

Cognitive Presence Category	Cognitive Presence Subcategory	Description	Example
	Interpretation and Synthesis	Logical progression showing cause and effect among numerous ideas or events.	The economic difficulties led then to social unrest. The social unrest began first in large cities, then spread to smaller cities. This social unrest progressed to talk of revolution, plans for revolution and, eventually…
		Showing internalization through summarization.	
		Drawing parallels with descriptions of how events or concepts are related.	Stating ideas from external sources in their own words, not simply quoting or parroting a text or source.
		Narrative justification: the effective use of narrative.	

	Parallels that more than just relate or link items, but describe or explain relationships.
	Narrative offering significant insight, meaning, richness to the topic; may support a conclusion.

Appendix C

Neural Network Settings for Experiment One

This model was built using the neural network package, Pattern Recognition Workbench, using the best 102 predictors (see the section of Chapter 3 entitled “Predictor Order”). Specifically, a backpropagation/MLP model using two hidden layers, a learning rate of 0.05, momentum of 0.1, 102 inputs (in order of importance -- see "Predictor Order" in Chapter 3), five outputs (one for each cognitive presence category), two hidden neurons in the first layer, five hidden neurons in the second layer, an automatically generated random selection pattern with a random seed of 1563029628 and training saved on the best test set. Pattern Recognition Workbench describes this type of model as follows:

> Milti-layer perceptron (MLP), also known as a “backpropagation neural network," is a neural network algorithm which generates input-to-output mappings based on computations of interconnected nodes. Nodes are arranged in layers. Each node's output is a nonlinear function of the weighted sum of inputs from the nodes in the preceding layer. (Unica Technologies, Inc, 1992 – 1997)

Appendix D

Second Experiment Neural Network Settings for Experiment Two

This model was built using Ward Systems' Neuroshell 2 Release 4.0. Specifically, a three-layer backpropagation network was employed with a learning rate of 0.05, momentum of 0.5, 40 inputs (in order of importance – see "Predictor Order" in Chapter 3), 4 outputs (one for each cognitive presence category), 56 hidden neurons, a rotational (as opposed to random) pattern selection with training saved on the best test set.

Appendix E

Descriptions of the 40 Categories Used in Experiment Two

The following list names and describes each of the top 40 categories used by the artificial neural network to categorize discussion list messages in experiment two.

WordCount: number of words in the message. Words are separated by spaces.

Question: number of questions (identified as question marks) in the message.

PersonNames: number of names of fellow classmates in the message.

Increas*: 111 words indicating change connoting increase

Know*: 348 words indicating awareness or unawareness, certainty or uncertainty, similarity or difference, generality or specificity, importance or unimportance, presence or absence, as well as components of mental classes, concepts or ideas.

Self*: 7 pronouns referring to the singular self

Region*: 61 words referring to regions and routes between them.

POLIT*: 507 words having a clear political character, including political roles, collectivities, acts, ideas, ideologies, and symbols.

MALE*: 56 words referring to men and social roles associated with men.

WltTot*: 378 words in wealth domain.

EVAL*: 314 words which imply judgment and evaluation, whether positive or negative, including means-ends judgments.

Polit@*: 263 words having a clear political character, including political roles, collectivities, acts, ideas, ideologies, and symbols.

Race*: 15 words (with important use of words senses) referring to racial or ethnic characteristics.

TrnLoss*: Transaction loss, 113 general words of not accomplishing, but having setbacks instead.

Web: self defined category referring to the WWW.

Social*: 111 words for created locations that typically provide for social interaction and occupy limited space

ThirdPersonPronouns: Self defined category of third person pronouns.

TimeSpc*: a general space-time category" with 428 words.

Pleasur*: 168 words indicating the enjoyment of a feeling, including words indicating confidence, interest and commitment.

WltOth*: 271 wealth-related words not in the above, including economic domains and commodities.

Object*: category with 661 words subdivided into *Tool*, (318 words), *Food* (80 words), *Vehicle* (39 words), *BldgPt* (46 words for buildings, rooms in buildings, and other building parts), *ComnObj* (104 words for the tools of communication) and *NatObj* (61 words for natural objects including plants, minerals and other objects occurring in nature other than people or animals). Last, a list of 80 parts of the body (*BodyPt*)

PowAuth*: Authoritative power, 79 words concerned with a tools or forms of invoking formal power.

Strong*: 1902 words implying strength.

PtLw*: A list of 68 actors not otherwise defined by the dictionary.

Exprsv*: 205 words associated with the arts, sports, and self-expression.

Thread: structure category identifying the thread to which the message belongs.

Width: structure category identifying the depth of the message (1 is a top level message in a thread; 3 is a grandchild message to a top level message).

ArenaLw*: 34 words for settings, other than power related arenas in *PowAren*.

RcRelig*: Religion, 83 words that invoke transcendental, mystical or supernatural grounds for rectitude.

Time@*: 273 words indicating a time consciousness, including when events take place and time taken in an action. Includes velocity words as well.

TrnGain*: Transaction gain, 129 general words of accomplishment

Quality*: 344 words indicating qualities or degrees of qualities which can be detected or measured by the human senses. Virtues and vices are separate.

Work*: 261 words for socially defined ways for doing work.

TranLw*: 334 words of transaction or exchange in a broad sense, but not necessarily of gain or loss.

TransitionsContrast: self-defined category of transition terms dealing with contrast.

PowAuPt*: Power authoritative participants, 134 words for individual and collective actors in power process

Decreas*: 82 words indicating change connoting decrease

Exch*: 60 words concerned with buying, selling and trading.

Exert*: 194 movement terms connoting exertion.

EnlPt*: Enlightenment participant, 61 words referring to roles in the secular enlightenment sphere.

Reply: self-defined category indicating that a message is a reply to another message.

*categories belonging to the General Inquirer (http://www.wjh.harvard.edu/~inquirer/homecat.htm)