Measuring Critical Thinking within Discussion Forums using a Computerised Content Analysis Tool

Stephen Corich, Kinshuk, Lynn M. Hunt

Eastern Institute of Technology, New Zealand, Massey University, New Zealand, Massey University, New Zealand

scorich@eit.ac.nz, kinshuk@ieee.org, l.m.hunt@massey.ac.nz

ABSTRACT

The use of discussion forums as a means of promoting collaboration and interaction between distance education students is increasing as a result of the growing popularity of online learning. The transcripts of these discussion forums have provided researchers an opportunity to analyse the interactions between participants and investigate evidence of cognitive and metacognitive activity.

This paper outlines some of the methodologies adopted by researchers who have attempted to measure evidence of cognitive processes and critical thinking among discussion forum participants. It also investigates the use of computers in the content analysis process and describes the use of automated tools to analyse discussion forum transcripts.

The paper then introduces a computerised tool, which has been designed with the aim of allowing different content analysis methodologies to be compared. The paper describes how the tool was used to analyse the discussion forum transcripts from a first year undergraduate degree course and discusses how the results obtained using the tool compared to a manually coded set of results.

Keywords

Quantitative content analysis, discussion forums, critical thinking, automatic measurement

INTRODUCTION

The increased use of online discussions in recent years within courses that are exclusively online or use online technologies to enhance on-campus courses is clearly identified (Meyer, 2004). One of the benefits associated with the use of online discussions is that a written record of activity is created that can be referred to by students for reflection. These written records can also be studied by academics who may wish to investigate the types of interactions between participants.

Researchers investigating the use of asynchronous discussion forums have attempted to identify evidence of critical thinking, which, as Bloom et al. (1956) suggested, should be a prime objective of any form of education, including online learning. A number of models, the majority of them based loosely on Bloom's taxonomy, have been documented and tested. These models have attempted to measure the extent to which knowledge is constructed through the collaborative discourse among discussion forum participants. The more commonly cited researchers include Henri (1991), Gunawardena, Lowe & Anderson (1997), Newman, Webb & Cochrane (1995), Garrison, Anderson & Archer (2000 & 2001) and Hara, Bonk & Angeli (2000).

While there is an abundance of evidence of researchers manually coding models to measure discussion forum activity, there is little evidence of computers being used to automate the coding process. McKlin et al. (2002) document one such occurrence, describing the successful use of an automated tool that used neural network software to categorize messages from a discussion forum transcript.

This paper describes a computerized automated content analysis tool (ACAT) which builds on the positive findings of McKlin et al. (2002), automating the coding of discussion forum transcripts against a variety of models. The tool can also be used to assist with the manual coding process and allows manually coded results to be compared with the automatically generated coding results.

The paper describes how the tool was used to code the transcripts obtained from a first year undergraduate degree course involving students who were enrolled in a computing systems degree and as such were familiar with using information technology. The model used for coding was a four stage cognitive-processing model

developed by Garrison et al. (2001). The automatically generated results were compared with manually coded results to determine the validity of the ACAT system.

The paper concludes by examining the potential for a computerized tool such as ACAT to analyse discussion forum transcripts and considers areas for possible future developments.

BACKGROUND

A review of literature suggests that quantitative content analysis (QCA) is one of the most popular approaches used by researchers to evaluate evidence of critical thinking in discussion forum postings. QCA is described as "a research technique for the objective, systematic, quantitative description of the manifest content of communication" (Berelson, 1952, p. 519). In its simplest form, QCA involves breaking transcripts into units, assigning the units to a category and counting the number of units in each category. QCA is described by many, who have used it, as “difficult, frustrating, and time-consuming” (Rourke et al., 2001, p12). Agreement between coders varies considerably and very few researchers duplicate their original models to validate their findings.

The more commonly cited researchers who have used QCA techniques to analyse discussion forum transcripts include Henri (1992), Gunawardena et al. (1998), Newmanet al. (1995), Garrison et al. (2000) and Hara et al. (2000). Henri (1992) developed an analytical model that highlights five dimensions of the learning process that can be found in messages. Gunawardena et al. (1997) introduced a model of analysis to assess the social construction of knowledge and collaborative learning. Newman et al. (1995) developed an analytical method for the study of critical thinking, which presented a list of indicators of critical thinking. Hara et al. (2002) used a content analysis approach, based largely on Henri’s (1992) cognitive and metacognitive dimensions, to support the investigation of quality online discussions. Garrison et al. (2000) assessed inquiry capabilities as well as critical thinking through a three dimensional model which measured cognitive presence, teaching presence, and social presence.

The methodologies adopted by Henri (1991) and modified by Hara et al. (2000), and Garrison et al. (2000) are two of the most popular content analysis approaches. These two methodologies have been either duplicated or incorporated into models developed by other researchers. Corich, Kinshuk & Hunt (2004) describe the application of these two methodologies to a first year undergraduate degree course.

AUTOMATED CONTENT ANALYSIS

The written transcripts produced as a result of activity between participants in a discussion forum can usually be exported to a text file that can then be subjected to quantitative content analysis. A review of literature suggests that even though the output is produced in a machine readable format, there is little evidence of using computers to assist with the analysis (McKlin et al., 2002). There are a number of software tools that can be used to assist in the task of analyzing text. These text-analysis tools include Wordnet, WordStat, NUD*IST, HyperQual and General Inquirer (Rourke et al., 2001). The tools are primarily text-processing systems which identify words as units and except for problems arising from the use of special alphabets, tend to be language-independent. The more powerful tools allow researchers to break a transcript into units and assign the units to a number of different coding categories. Once the transcripts have been coded, the results can be imported into statistical programs for more detailed quantitative analysis.

The majority of automated text-analysis tools are generic and can be applied to a number of text analysis situations. Since the tools are generic, they do not come with built-in pre-existing word categories that could be applied to categorize cognitive activities and critical thinking; they rely on the researchers to create word categories.

McKlin et al. (2002) described the use of an automated tool that uses neural network software to categorize messages from a discussion forum transcript. They suggested that the tool may ultimately be used to gauge, guide, direct and manipulate the learning environment. The analysis was based on Garrison et al. (2000, 2001) community of enquiry model. The study reported coefficient of reliability figures of 84% and 76% when compared to results of human coders, suggesting that a neural network has the potential to successfully code transcripts to identify cognitive presence. The study also suggested that the tool would be refined to produce a system that could be used to reliably classify messages into cognitive presence categories. Despite a review of current literature and an attempt to contact the research team, no evidence could be found of further publications relating to the tool.

AUTOMATED CONTENT ANALYSIS TOOL (ACAT)

Considering the positive findings of McKlin et al. (2002), a decision was made to design and develop a web based automated content analysis tool (ACAT) that could be applied to discussion forum transcripts to identify evidence of critical thinking among discussion forum participants. The aim of the development was to produce a system that can be used to apply a number of recognized QCA models to categorize messages for differing levels of cognitive activity and critical thinking.

Figure 1 shows the concept model for the ACAT system. Users can import transcripts and manually code them against one of a number of recognized QCA models. Users can also import transcripts and allow the system to automatically code them against the QCA models and then compare the manual and automatically coded results.

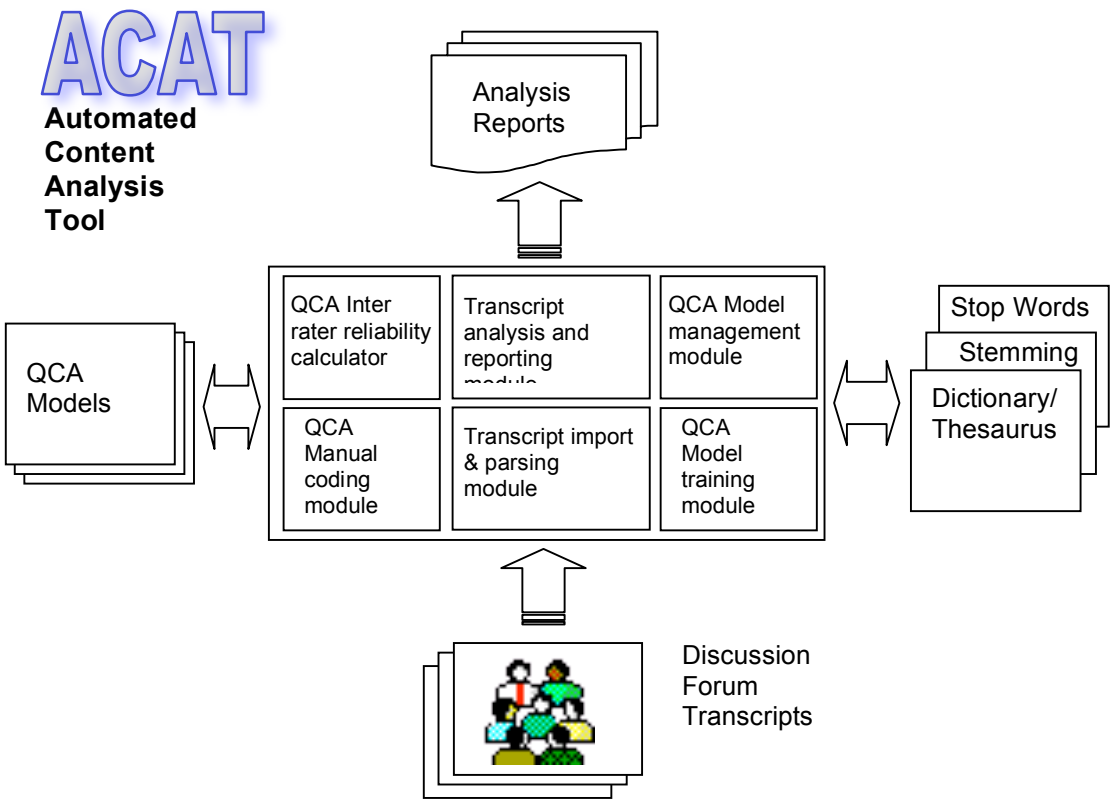

Figure 1: The concept model for the ACAT system

The transcript importing and parsing module allows any discussion forum transcript saved as a raw text file to be imported into the system. The system then parses the text, breaking the content into individual sentences which are stored in a table to allow classification to occur. While researchers have experimented with different units of analysis with varying measures of success (Rourke et al., 2000), Campos (2004) suggests using the sentence as the human cognitive unit of analysis. The ACAT system has been built to analyse transcripts on the basis of individual sentences.

In addition to saving the transcripts as raw text, the transcript importing and parsing module also applies a stop word algorithm to each sentence saving each sentence with stopwords removed. Stopword removal involves the removal of words which are very frequent and do not carry meaning. Stopword removal is said to improve the reliability and effectiveness of text analysis systems (Ginsparg et al., 2004). The same module also applies a word stemming algorithm and saves each sentence as a stemmed word sentence. Word stemming is the process of removing suffixes from words to get the common origin of the word, and is said to help when comparing texts to identify words with a common meaning and form as being identical (Hull & Grefenstette , 1996).

The manual coding module allows a user to manually code sentences from any imported transcript against a selected QCA model and stores the result so that it can be used by the transcript analysis and reporting module to produce a manual coding report.

The QCA training module allows a user to train any recognized QCA model that has been created using the QCA model management module. The training module allows a user to take a transcript that has been manually coded and add it to the QCA model dictionary, which then forms the basis for the transcript analysis and reporting module to produce coding reports. The model dictionary contains a set of phrases for each category of a QCA model which are used by the transcript analysis and reporting module when it performs an analysis of an imported transcript against the individual model categories.

The QCA model management module allows a user to create a QCA model with a number of categories and code acceptable phrases which are used in the analysis process for each category.

The transcript analysis and reporting module produces a manual coding report which lists the occurrences of sentences against each category of a QCA model. The transcript analysis and reporting module also performs the automated analysis of any imported transcript, producing a report that is similar to the manual coding report, listing occurrences of sentences against each category of a QCA model. While there is little evidence of using computers to classify the transcripts of discussion forums, several studies have reported favorably on the computer grading of essays (Page, 1966 & 1994; Landauer, Foltz & Latham, 1998; Chung & O'Neil, 1997). Rudner & Liang (2002) reported that there is promising literature in the information science field regarding the use of Bayes Theorem as the underlying model behind text classification. Bayesian networks have become widely accepted and are being used in essay grading systems, help desk applications, medical diagnosis, data mining, intelligent learning systems and risk assessment tools (Rudner & Liang, 2002). Bayesian Networks use probability theory to assign items to various categories. McCallum and Nigam (1998) provide an excellent overview of the use of Baysian Networks. The apparent success of using Bayesian networks to classify text suggests that a similar process could be applied to classify transcripts. The analysis technique adopted in ACAT development is based on an essay scoring system described by Rudner and Liang (2002) who used a four point scale to categorize the features of essays.

The QCA inter rater reliability calculator module allows the results of a manually coded transcript to be compared with the results that have been automatically produced by the ACAT system and a coefficient of reliability to be calculated. The most commonly used method of reporting reliability between coders is the percent agreement statistic, which reflects the number of agreements per total of coding decisions. Hosti's coefficient of reliability (Hosti, 1969) and Cohen's kappa statistic are two of the popular methods of reporting coding reliability (Rourke et al., 2000). Acceptable levels of agreement have yet to be established, with some researchers stating that anything less than 80% is unacceptable (Riffe, Lacey & Fico, 1998), while others report levels as low as 35% (Garrisonet al., 2000). The ACAT system uses Hosti's coefficient of reliability.

USING ACAT TO CODE TRANSCRIPTS

Corich et al. (2004) describe how the transcripts obtained from a first year data communications course were manually coded using coding schemes developed and tested by Garrison et al. (2001) and Hara et al. (2000). The same set of transcripts was used to test the ACAT system.

The research in the Corich et al. (2004) study was ethnographic due to its small sample size and lack of statistical testing. It was designed as a preliminary exercise to a larger research project that will use larger samples across a variety of institutes, utilizing intelligent software to perform the content analysis coding. The research was conducted to allow the researchers to become familiar with two of the most popular quantitative content analysis models and to attempt to identify if the models could be applied to determine the level of critical thinking for individual students.

The research was conducted during the second semester of a first year undergraduate degree course. All the students were enrolled in a computing systems degree and as such were familiar with using information technology. The course was an introductory data communications and networking class that was delivered using a blended learning environment, combining traditional face-to-face activities with web publishing, on-line review and discussion forum activities. On-line activities, which included publishing the results of a research project, evaluating the work of peers and participation in a discussion forum formed a significant part of the course. The use of the discussion forum was seen as a way to encourage participation as well as to provide a tool to promote discussion over a period of time to a topic that was a key component of the course curriculum. Previous offerings of the course did cover the same topic, the future of data communications, in a normal classroom setting, using face-to-face discussion over a period of at most two hours. Using the discussion forum approach, students were allowed three weeks to participate in on-line discussion.

The software used to support the discussion forum was an integral part of the Blackboard learning management system. All students had previously used Blackboard to retrieve course materials and to participate in on-line

tests in their earlier courses; however none of the students had participated in discussion forums during their previous academic study.

The class consisted of fifteen students, three females and twelve males, aged between 18 and 38, and of varying academic abilities. Students were given the topic for the discussion early in the course and instructions were provided to the students as to what was expected in the discussion forum. The instructions were given as a guide to encourage higher level critical thinking. The student postings were monitored by an instructor who provided encouragement, added pedagogical comments and provided reinforcement and expert advice.

During the three weeks that the discussion forum was operational a total of 104 posts were made, 30 of which were made by the course instructor. Once the instructor postings were removed, the remaining 74 posts generated 484 sentences for coding.

The resulting transcripts were coded by two individual coders and the results of the transcript analysis for the two instructors were evaluated to establish the level of agreement that existed, using the coefficient of reliability developed by Holsti (1969). The coefficient of agreement was 87% using the Garrison et al. (2001) model and 81% using the Hara et al. (2000) model.

For the purposes of testing the ACAT system, the QCA model with the higher coefficient of reliability was used and the two coders used the QCA model management module to create a model based on the Garrison et al. (2001) model which has four categories of cognitive activity. An initial dictionary was built based on the triggers described by Garrison et al. (2001) and the experiences gained from the earlier coding exercise. The coders then used the ACAT transcript importing and parsing module to import the transcript and the manual coding module to manually code the transcript.

RESULTS AND FINDINGS

The ACAT transcript analysis and reporting module was used to produce a manual coding report and an automatic coding report. The two reports were then compared. Table 1 shows the output produced.

Category	Manual Coding	Automatic Coding (Raw Text)	Automatic Coding (Stopwords removed)	Automatic Coding (Stemmed text)
1. Triggering	73 (15%)	78 (16.1%)	80 (16.5%)	75 (15.5%)
2. Exploration	124 (25.6%)	128 (26.5%)	129 (26.7%)	119 (24.6%)
3. Integration	209 (43.2%)	218 (45%)	217 (44.8%)	225 (46.5%)
4. Solution	58 (12%)	60 (12.4%)	58 (12%)	65 (13.4%)
Not categorised	20 (4.1%)	0	0	0
Total number of units	484 (100%)	484 (100%)	484 (100%)	484 (100%)

Table 1: Number of sentences in each of Garrison et al. (2001) categories

It was interesting to note that the automated system, did not leave any items uncategorized where as human coders were unable to categorize 4% of the total. The algorithm which categorized the sentences matched on the basis of the greatest probability of fit in a specific category, a more in-depth analysis identified that none of the 484 categories had a match less than the probability of a random occurrence in an individual category. The in-depth analysis also identified a problem with the algorithm, which categorized sentences with equal match probabilities. When the algorithm identified a sentence having identical probabilities in more than one category the algorithm automatically placed the sentence in the first of the categories in the processing loop.

While the percentages of sentences were similar across all categories for manual coding, raw text automatic coding, stopwords removed text automatic coding and stemmed text automatic coding, the values for Hosti's coefficient of reliability did not indicate such a high level of coding correlation. The coefficient of reliability between the manually coded transcript and the automatically coded raw text was 0.64. This is significantly less than the 80% figure indicated as being acceptable by Riffe, Lacey & Fico (1998), but is still an improvement on the figures quoted by Garrison et al. (2000). The figure is also lower than the 87% correlation reported by Corich et al. (2004) using the same transcript for two coders using manual coding methods. When compared to McKlin et al. (2002) the correlation coefficient is better than the neural network system before in-depth system training.

The coefficient of reliability between the manually coded transcript and the automatically coded transcript text with stopwords removed was 0.65, indicating a slight improvement in the coding process. The coefficient of

reliability between the manually coded transcript and the automatically coded transcript text using stemwords was 0.71, indicating a more significant improvement in the coding process.

Both coders, involved in creating the QCA model and model training, suggested that the reliability of the system would be improved if more time had been spent training the system and by increasing the size of the model dictionary. It was also suggested that transferring the manually coded units to the model dictionary would increase the effectiveness of the system for subsequent validation trials. These suggestions support the findings of McKlin et al. (2002).

CONCLUSIONS AND FUTURE DEVELOPMENTS

Fahey (2002) suggests that detection of critical thinking in discussion forum activity is a difficult and very time consuming task. This task is said to be inherently subjective, inductive and prone to errors (Rourke & Anderson, 2002). The subjectivity arises from the interpretation of coders as they attempt to assign topics to categories. To reduce the likelihood of subjectivity during coding, researchers employ multiple coders and compare coding results to ensure that they come to the same coding decisions (Rourke et al., 2000). Providing the tool can be shown to be reliable, the use of an automated analysis tool such as the ACAT system has the potential to reduce the time involved in coding and eliminate much of the subjectivity.

While the coefficient of reliability figures were not as high as those reported for manual coding by Corich et al. (2004) the initial investigations of using the ACAT system as an alternative to manual coding would suggest that the system merits further investigation. The findings reinforce the conclusions of McKlin et al. (2002) who suggest that automated tools could be used to categorize discussion forum activities into cognitive categories such as those proposed by Garrison et al. (2000), Henri (1991) and modified by Hara et al. (2000).

The ACAT system described in this paper is a work in progress. The current system, while having the potential to automatically code transcripts against any recognized QCA model, has only been used with the Garrison et al. (2000) model. Creating multiple models and training the models would give the system the potential to compare different models. The current system only calculates Hosti's coefficient of reliability. Adding the ability to calculate Cohen's kappa statistic would enable the system to compare the two reliability calculation measures.

The current ACAT system codes complete transcripts, paying no attention to the identity of the contributor, providing levels of cognitive activity for all participants. The system could be modified to identify individuals, allowing the system to be used to provide individual feedback to participants on the level of cognitive activity.

While the existing ACAT system performs its coding algorithm using Bayesian Network probability theory to assign items to various categories, the system could be extended to include a number of different algorithms such as Nearest Neighbor, Centroid-Based Document Classifier, Latent Semantic Indexing, Log-Entropy weighting, or Term Frequency & Inverse Document Frequency. The effectiveness of each of these techniques could then be compared.

Current research has focused on the identification levels of cognitive activity, the system has the potential to be extended to measure levels of activity against a knowledge domain area. Such a system may have the potential to automatically report on participant activity against a prescribed domain.

As McKlin et al. (2002) note, there is potential for a well-trained automated system to classify messages as reliably as human coders. Such tools will allow instructors to make adjustments to their approach in order to bring about desired displays of cognitive effort.

REFERENCES

Berelson, B. (1952). Content analysis in communication research. Illinois: Free Press.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W.H. & Krathwohi, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. New York: David McKay Co. Inc.

Campos, M. (2004, April). A Constructivist method for the analysis of networked cognitive communication and the assessment of collaborative learning and knowledge building. Journal of American Learning Networks, 8(2),1-29.

Chung, G, K, W. K., & O'Niel, H. F., Jr. (1997). Methodological approaches to online scoring of essays. ERIC Document Reproduction Service, No. ED 418 101, 39pp.

Corich, S.P., Kinshuk. & Hunt, L. M., (2004). Assessing Discussion Forum Participation: In Search of Quality. International Journal of Instructional Technology and Distance Learning, 1(12), 1 -12.

Fahy, P. (2002). Use of linguistic qualifiers and intensifiers in a computer conference. American Journal of Distance Education. 16(1).

Garrison, D.R., Anderson, T., & Archer, W. (2000). Critical thinking in a text-based environment. Computer Conferencing in higher education. Internet in Higher Education, 2(2), 87-105.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical Thinking, Cognitive Presence, and Computer Conferencing in Distance Education. The American Journal of Distance Education 15(1), 7–23.

Green, J. (2000). The online education bubble. The American Prospect, 11(22), 32-35.

Ginsparg. P., Houle, P., Joachims, T., & Sul, J. (2004). Mapping subsets of scholarly information. Proceedings of National Academy of Sciences of the United States of America 101 (1) , 5236-5240.

Gunawardena, C., Lowe, C. & Anderson, T. (1997). Analysis of a global on-line debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. Journal of Educational Computing Research, 17(4), 395-429.

Hara, N., Bonk, C., & Angeli, C., (2000). Content analyses of on-line discussion in an applied educational psychology course. Instructional Science. 28(2), 115-152.

Henri, F. (1991). Computer conferencing and content analysis. In A. R. Kaye (Ed.), Collaborative learning through computer conferencing: The Najaden Papers (pp. 116-136). Berlin: Springer-Verlag.

Hosti, O. (1996). Content analysis for social sciences and humanities. Don Mills, ON: Addison Wesley.

Hull,D.A., & Grefenstette, G.(1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science. 47 (1), 70-84.

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

McCallum, A. & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization. Available online http://citeseer.ist.psu.edu/mccallum98comparison.html

McKlin, T.,Harmon, S. W.,Evans, W., & Jones, M, J. (2002). Cognitive Presence in Web-Based Learning: A Content Analysis of Student's Online Discussions.

Meyer, K.A. (2004, April). Evaluating online discussions: four different frames of analysis. Journal of American Learning Networks, 8(2),101-114.

Newman, G., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face computer supported group learning. Interpersonal Computing and Technology, 3(2), 56-77.

Page, E. B. (1996). Grading essays by computer: Progress report. Notes from the 1996 Invitational Conference on Testing Problems, 87-100.

Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. Journal of Experimental Education, 62(2), 127-142.

Riffe, D., S. Lacey, & F. Fico (1998). Analyzing media messages: Using.quantitative content analysis in research. Mahweh: New Jersey.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. International Journal of Artificial Intelligence in Education, 12(1), 8-22.

Rudner, L. M., & Liang, T. (2002). Automated Essay Scoring Using Bayes' Theorem. National Council on Measurement In Education. New Orleans, April.

Spatariu, A., Hartley, K. & Bendixen, L.D. (2004, Spring) Defining and Measuring Quality in On-line Discussion. Journal of Interactive Online Learning. Vol 2(4).