CHAPTER 2: LITERATURE REVIEW

The literature review addresses three fundamental questions:

1.) Are there theories or models which explain how to analyze a student’s verbal behavior in writing samples?

2.) How is data mining being used in education, and is it currently being used to find patterned relationships between writing samples and technology use?

3.) What current model or theory best explains students’ LMS usage and efforts?

In reviewing the theory around LMS usage in higher education courses, there seem to be as many ways to analyze the data and to build research around them as there are individual studies in the field. LMS usage analysis in unsupervised studies basically equated to Web log studies, since most LMSs record page views and other typical Web log data. Therefore, the literature was surveyed broadly to become acquainted with the full range of research methods and approaches used therein.

In reviewing the literature around the process of data mining, and the special case of literature that focuses on educational data mining (EDM), a book published by the International Educational Data Mining Society (Romero, 2010) was invaluable in communicating the current EDM research activities and specifically focusing resources on this relatively new field of study.

Philosophical and Conceptual Approach to the Study

This study was built and planned around a data classification design, which promotes two distinct datasets and separate processing of those data until they are compared in the final step. Though this approach shall be discussed at length in the
methodology chapter that follows, it is important here to review the literature to inform
the theoretical underpinnings in this process. Thus, a diagram, Figure 2.1, is included
below to display the relationship among the parts of the study and to point out the
significant fields which will eventually affect the choices made in the methodological
stage of the study.
The classification procedures occur separately on both branches using initial classification models provided in the literature, and the results of these decision tree models are (were) compared after separate analyses have occurred on both branches of inquiry. This is important, because the decision was made early on not to analyze the data on either branch in such a manner as to anticipate the needs of the other branch. Instead, each branch required an analytical approach befitting the theory and practice of the fields that enveloped each one.
In the case of the branch that incorporated the student writing samples (WS), or the left branch in Figure 2.1, verbal behavior in the form of student class compositions were analyzed. This textual analysis falls within the verbal behavior field, and especially its ways and means of the analysis of verbal behavior.

In the case of the branch that incorporated the student page views data (PV), or the right branch in Figure 2.1, computer usage behavior in the form of student clicks inside an LMS were analyzed, so this behavior analysis falls clearly within the behavior analysis field, and especially its ways and means of the analysis of computer usage behavior while inside a learning management system.

Only after these separate analyses occurred did the comparative process come into play, and at that point the field of data mining itself provided the ways and means of analyzing data in a manner validated within the field itself. Though educational data mining has its own applications of these theories, the knowledge discovery theory that underpins it is that same theory that underpins all of data mining, no matter what the discipline. This is as it should be, for educational data mining has much to learn from the historical developments of data mining, especially those rooted in commercial purposes, and to divorce the procedures from those tried and true ones that have been successfully used in other fields would leave the educational pursuit virtually unarmed in this endeavor.

**Verbal Behavior and Textual Analysis**

It is common practice in higher education for instructors to require students to submit works of compositional writing (i.e., papers) in partial (or total) completion of
course requirements. The practice is so ubiquitous that it seems second-nature to everyone in the academic community. Particularly in the humanities, the assignment of term papers and other forms of compositional writing is considered the rule, and the course where no written work is required the rare exception.

However, in its familiarity lies a tendency to ignore what is really happening when instructors employ these instructional methods. We are so accustomed to the process that we are often blinded to the dynamics that underlie it, dynamics that include power structures and punishment and a student’s sometimes futile quest for approval. Skinner (1957) provided us with a method of theoretical observation for these structures that largely eluded his forebears. Chomsky (2000) almost immediately refuted Skinner’s externally-motivated theory of verbal behavior with one that focused on the cognitive theory of language attainment, and the two poles were quickly delineated.

Textual Analysis as an Analysis of Human Verbal Behavior

Textual analysis has a long history, which began well before computers helped to automate the process. Holmes (1994) provides a fine historical account of the ways and means of authorship attribution, which he summarizes in the pursuit of discovering the authorship of Biblical texts.

The field of text analysis has a vast number of studies, perhaps characterized by qualitative and quantitative methods applied to text analysis (Bernard & Ryan, 1998), including a critical view of misunderstandings of the usage of the term “textual analysis” (Carrera-Fernández, Guàrdia-Olmos, & Peró-Cebollero, 2011).
A more recent lens through with textual analysis is viewed is in Discourse Analysis, both critically (Bucholtz, 2001; Huckin, 2002), but also in the areas of social linguistics (Gee, 2007; Gee, 2010) and social semiotics (Hodge & Kress, 1988; Kress, 2006; Lemke, 1998). Discourse analysis is a relatively new field, and though it is theoretically related to textual analysis as a study of verbal behavior, in its dedication towards the critical aspects and social significance of verbal behavior, it is not the most informative theoretical framework to address this study’s research questions, which are based more on exploring the relationships between individual student behavior in compositional writing and LMS usage.

Finally, there is a theoretical discussion about the digitization of what were formerly printed texts, and the implications of textual analysis on digital texts, which have been critically edited by human beings well past the lifespan of the texts’ original authors (Bayne, 2006). Since the corpus of student texts to be analyzed in the current study consists of those that were submitted in digital format, by authors who are still very much alive, and no attempt at third-party critical editing on the texts have taken place, it can be inferred that this discussion, though interesting in its own right, has little bearing on the study at hand.

Some of the varied ways and means of textual analysis can be found in a multitude of studies, that include analyzing error patterns (Doolan & Miller, 2012), a comparison of essays submitted by sociology and English students (Bruce, 2010), genre differences between technical and argumentative writing (Wang & Cho, 2010), and sensing affect from text messages (Shaikh, Helmut, & Ishizuka, 2006), among others. For a full view of how text analysis has occurred and is occurring, the reader should refer
to one of the many papers and books on the subject, such as Lemke’s (1998) fine overall
view and analysis of the pursuit. Those specific studies that directly inform the
methodology for this project will be discussed below.

**Textual Analysis Using Lexical Bundles**

Theoretically, recent textual analysis methods have derived and analyzed lexical
bundles, which are “multi-word sequences that recur frequently and are distributed
widely across different texts” (Adel & Erman, 2012, p. 82). The seminal text in this field
(Pawley & Syder, 1983) provides a solid foundation for their use.

This lexical bundles approach also acknowledges the importance of idiomaticity,
the presence of lexical bundles that arise within certain knowledge domains or academic
fields over others. Thus, if a student finds herself in an English course, she could expect
different lexical bundles to arise than if she should find herself in a computer science
course, for example. Of root importance to the framework of lexical bundles is the
“idiom principle” (Sinclair, 1987, “The Nature of Evidence” cited in Adel & Erman,
2012, p. 82), which is the principle that words are co-selected instead of selected one at a
time.

Significant time was spent in the literature deliberating between using 4-word
bundles, versus 3-word bundles and 5-word bundles (Adel & Erman, 2012, p. 84). The
general consensus is that 4-word bundles present around 100 bundles for most texts. 3-
word bundles produce too many bundles to be manageable (and are usually encapsulated
in the 4-word bundles anyway), and 5-word bundles produce too few. Various thresholds
for determining which word sequences can be considered bundles have been discussed,
but the general consensus in the literature tends to place the threshold between 25 instances and 40 instances per million words processed (Adel & Erman, 2012; Cortes, 2004).

The categorization of lexical bundles was also of special consideration. The literature reviewed provided three categories (referential, stance, and “text organising” [sic]) (Adel & Erman, 2012, p. 89), and four categories (basically the former three plus interactional) (Cortes, 2004, p. 401). The application of textual analysis using lexical bundles has provided significant results in studying differences in compositional writing between native and non-native speakers (as in Adel, cited above) and comparing professional and student writers in the fields of history and biology (as in Cortes, also cited above).

One area of recent inquiry is the categorization of clausal bundles (not specifically called “lexical bundles” in this study) into areas defined within a “discourse taxonomy” (de Waard & Maat, 2012, p. 359) that includes discourse “segment types” of “Fact, Problem, Hypothesis, Implication, Goal, Method, and Result.” This further classification of lexical bundles appears to provide a better base to correlate verbal behavior with psychological correlates, as the de Waard study attempted in its specific focus on verb tense.

The bundling of phrases appeared to be a significant branch in the linguistic research of compositional writing but was too idiosyncratic to be undertaken in this study. Instead, the choice was made to undertake linguistic inquiry into the word choices made by students as they composed their required course writing assignments, especially
using available validated instruments.

**Tools to Aid in Textual Analysis**

There have been a few word count tools that have been developed in past years. Colin Martindale (1990) used computer analysis to explore novel expressions among artists and other creative people, and Philip Stone, et al. (1968), as far back as 1966, pioneered computerized textual analysis. However, Pennebaker, Francis & Booth’s (2001) Linguistic Inquiry and Word Count (LIWC) has been utilized in many textual analysis projects since its inception in 2001 and appears to be the “industry standard” among researchers who desire to do particle analysis of text (including Faliagka, Kozanidis, Stamou, Tsakalidis, & Tzimas, 2011; Groom & Pennebaker, 2002; Mohtasseb & Ahmed, 2010; Nguyen, Phung, Adams, & Venkatesh, 2011; Vercellone-Smith, Jablokow, & Friedel, 2012). A more complete view could be gained by typing LIWC into any research database query application.

Carla Groom (2002) eloquently sets out the theoretical underpinnings of the LIWC software and effectively ties it into Skinner’s Verbal Behavior theory through the discussion of “Freudian slips” (2002, p. 616). As she asserts that “everyday utterances” could “reveal details of the unconscious,” Skinner (1957, p. 372), too believed that our verbal behavior reveals details of the unconscious and readily expresses these details, especially when errors are made in verbal.

Groom and others (cited above) who have used LIWC for textual analysis claim its ease of use, its non-obtrusiveness, and its happy avoidance of the need for subjects to self-report.
LIWC is simple to install on a personal computer and use for any number of files, either in plain text format or either of the Microsoft Word formats (.doc or .docx). Whereas textual analysis software in past years was created for mainframe computer analysis, LIWC conveniently and effectively leverages the computing power of personal computers.

One of the other great strengths of LIWC is that one can collect data after the fact, download it to her computer, and process those data in an unobtrusive manner. One does not have to corral subjects and bring them to the lab to acquire experimental data. The subjects produce the data in their natural surroundings, and the researchers collect those data after the subjects have departed the course environment, be it a physical one or a virtual one. This makes the use of LIWC unobtrusive within the study environment. Furthermore, since the data consist of natural verbal behavior produced by the subjects, there is no need for those same subjects to report their behaviors after the fact. The assessment is more direct -- the text is the text. Although some researchers have chosen to collect interview or survey data from the participants to validate their experiences, for many studies it is simply not necessary, as LIWC has been validated as a standalone textual analysis tool in its own right (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007).

One of the main complaints about LIWC among the academic community is that it is set up by default to consider single words (referred to as “particles”) (Groom & Pennebaker, 2002) versus phrases, as illustrated by the studies that use semantic webs and/or lexical bundles as mentioned above. Groom (2002, p. 618) states that “particles form the linguistic ‘glue’ that hold content words together.” By this description, the
particles hold an autoclitic role in the text, albeit an important one that they provide the “personal style” by which the speaker expresses the content. As such these particles can be considered style markers, and therefore are useful in comparative studies among subjects emitting verbal behavior while performing the same task within the same verbal environment. LIWC has been created with the help of human judges (Groom & Pennebaker, 2002, p. 617) and extensively validated by a separate group of judges (Pennebaker, Mehl, & Niederhoffer, 2003). Furthermore, the successful use of the software in a vast number of projects provides evidence that textual analysis of particles in the “bottom-up” manner of LIWC can be an effective approach to psycholinguistic analysis.

Groom and Pennebaker (2002, p. 620) complete their paper with a suggestion that “It will also prove helpful to combine word-count approaches with more inductive strategies” and PsychoNet (Mohtasseb & Ahmed, 2010) may be that hoped-for combination. The authors combine the psycholinguistic taxonomy of LIWC together with the common sense knowledgebase called ConceptNet to create a psycholinguistic commonsense ontology called PsychoNet. The authors use cosine difference to measure the similarity degree among semantic graphs (Mohtasseb & Ahmed, 2010, p. 162), providing mood classification as an example application.

Perhaps the most informative study that uses LIWC in analyzing textual behavior is one that predicts age, sentiment, and connectivity from text collected in the realm of social media (Nguyen et al., 2011). This study uses probabilistic text modeling to predict age (old and young), sentiment (happy and sad), and connectivity (social versus solo) among 10,000 bloggers in the LiveJournal blog site. Posted age and bloggers’
“followers” and “following” provide the discriminators for age and connectivity, respectively, and the “mood tag” feature within each blog post provided the discriminator for sentiment. Since all of the features in all of the polar categories were found to be not of normal distribution, the differences were evaluated using the Mann-Whitney U non-parametric test (Nguyen et al., 2011, p. 6). All of the elements of this proposed study are present in their study; only the source of the data is different (student composition data versus social media data).

The Nguyen, et al., study appears to have a few weaknesses (e.g., age and the selection of the mood tag were self-reported, and thus are based on the whims of the bloggers in an environment where users are not held accountable for these kinds of assertions), its setup and use of LIWC and selection of LIWC categories for further analysis and reporting were invaluable in the consideration of like tasks for the current study. As such, the Nguyen study provided some structure and methodology for the current study, though as we discovered, the approaches between the two studies were different in that they had somewhat different purposes.

**Other Considerations in Textual Analysis**

Three additional considerations in the literature appeared to relate to this study. The first had to do with “legitimate textual borrowing” (Petric, 2012), which studied the behavior of quoting by L2 (second language) students in classroom compositions. In Skinnerian terms, the behavior of quoting others’ work could be considered echoic behavior, and also falls under the realm of transcription. Petric (2012, p. 111) provides four “reasons” for quoting, which includes reasons related to source, writer’s goals,
external pressure, and student’s beliefs and fears. In the current study, the amount of echoic behavior used by a student in composing a class paper may belie one of the above-mentioned reasons, which may have an effect, perhaps, on motivation and engagement.

The second consideration had to do with “expressing emotion” in textual verbal behavior, and especially in class compositions (Hancock, Landrigan, & Silver, 2007). Since the reader is not normally in the same room as the writer, the reader cannot determine the emotion with which the writer is engaged while composing the text. Therefore, the normal non-verbal cues are not present, and the reader, if the determination of emotional state is important, must take all cues from the text itself. This study used LIWC to analyze a portion of the text, providing an example of a regression model (Hancock et al., 2007, p. 93) that was not of direct use in the current study, but could perhaps be of use in a future study.

The third area of interest was the determination of sub-technical vocabulary (words that are used readily in all contexts but which have more specific meanings in technical contexts, though widely across different fields). Several studies have pursued this line (Cowan, 1974; Huizhong, 1986; King, 1989), and it initially appeared to be a promising approach for the current study, since the collected textual data are student compositions that are of a somewhat technical nature.

Once initial analysis of student textual behavior had occurred, the results were a rich array of dimensions that had a relationship with student LMS usage, and therefore were “mined” for relevance. There were a number of approaches to this mining process within the domain of Educational Data Mining from which to choose, some of which were discussed immediately below.
Educational Data Mining

There is a relatively recent but robust research program in Educational Data Mining (EDM), set forth by an international organization (the International Educational Data Mining Society), which has held a conference since 2008, and produced a journal (the Journal of Educational Data Mining).

Though members of this organization do not produce all of the research occurring in this area, they are aware of the bulk of it, and provide some excellent review materials that characterize much of the research occurring in this area and an excellent synthesis of this work. One paper in particular, “Educational data mining: A review of the state-of-the-art” (Romero & Ventura, 2010), provides an excellent summary of advancements in the field and a good assessment of future work to be done. It also includes a comprehensive list of over 300 references, which makes the literature review in this area much easier to accomplish.

EDM Basics

First, however, it was helpful to review literature that pertained to Educational Data Mining as a subfield of Data Mining (DM) itself, and the special goals and challenges that researchers face when they are researching in this area.

In the paper “Applying Data Mining Techniques to e-Learning problems, Castro et. al. (2007) give an excellent overview of data mining and its various approaches in general as well as within the context of e-Learning. Data Mining is the undertaking of special analysis methods to discover knowledge in what is normally a large amount of data (Pahl & Donnellan, 2002). Data miners try to determine patterns in the data and
build models from these patterns. They may use clustering tools to attempt to classify people or things into groups, and regression tools and decision paths and association rules to predict trends and behaviors. There is a wide variety of approaches and applications of data mining, especially in the areas of business and commerce, which have built and provided many of the tools and much of the theory supporting the validity and efficacy of the data mining process.

Educational Data Mining (EDM) takes these tools and theoretical framework and applies them specifically to the practice of data mining using educational data focused on educational research questions. The educational data are of varied origin and detail, but often include those data that come from traditional classroom practices, e-learning and web-based learning tools, learning management systems, adaptive educational systems, psychometric tools such as tests and surveys, physical and online content, among others. A table of these items, complete with references, is available in Romero’s (2010) paper, which is publicly available on the Web.

Because it can easily be said that educational data mining has been done in every area where students, instructors, and administrators are motivated to obtain knowledge about one another, the research questions are much more difficult to characterize. Romero and Ventura (2010) organize these questions into a number of groupings, which include:

- the analysis and visualization of data,
- providing feedback for supporting instructors,
- recommendations for students,
- predicting student performance,
• student modeling,
• detecting undesirable student behaviors,
• grouping students,
• social network analysis,
• developing concept maps,
• constructing courseware, and,
• planning and scheduling.

As an interesting side note, no references to research undertaken by students to learn more about their teachers and the administrators at their institutions were uncovered. These data mining tools are freely available publicly to anyone, and the data about instructors, though much of it may be biased and unreliable, are available on Websites (such as RateMyProfessors.com), so it is not unreasonable to think that some analysis of these data has been done by students, with or without the awareness of their teachers and administrators and, perhaps, researchers in the field. As such, these studies would also be considered educational data mining, though perhaps not supported and endorsed by any educational institution.

**Discovering Models in Educational Data**

It has been said, in the educational sector, the “abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation” (Hanna, 2004). The current study was performed to reduce that reality as it processed existing data to provide additional information that will hopefully (upon publication of this document) be helpful to the educational community.
Studies that attempt to build models in educational data, especially those that use the LMS to collect the data, are reviewed below.

**LMS Usage Analysis and Modeling**

LMS usage analysis consists of attempting to determine how, when, and where (and sometimes why) students use learning management systems that are provided by the institution and configured and populated with course content and activities by instructional designers and/or instructors to support student learning within a course. These materials that are delivered on learning management systems (often in the guise of a “course”) may support instruction by providing all or a majority of the course content (a “distance” or online course), a significant amount of course content and activities to support face to face courses (a hybrid course), or a smaller amount of content and activities, perhaps limited to the single function of serving as a drop box for assignment submission, a download site for articles and handouts, or an easy and secure way to communicate grades to students, for example.

No matter what level of intended use, the ultimate objective in providing an LMS for student use is to provide 24-7 access to course content and activities for students and teachers, and to provide an organized and scalable way to support student learning. These objectives provide a rich source of data for researchers, as student behavior within an LMS normally consists of mouse clicks and keyboard typing, both of which have been the staples of computer usage studies for decades (Kelly & Teevan, 2003). However, recent trends in LMS design, especially in the example of Instructure’s Canvas™ LMS, provide for student interaction with the LMS in the use of images, audio recording, and
video recording. These online behaviors, which may begin and end with simple mouse clicks, are content-dependent, as a large variation of student usage behavior happens between those mouse clicks, most of which does not immediately manifest itself for analysis by researchers. In these cases, qualitative analysis of images, audio, and video data in terms of discourse analysis must be undertaken, which is a time-consuming and resource-intensive undertaking.

A significant number of studies have focused upon LMS usage data in the forms of mouse clicks and keyboard typing. These data have been mined for significant patterns and have models of LMS usage based thereupon including grammars of LMS usage. Srivastava, et al. (2000), summarized the state of usage mining research as it was heading into the 21st century. Their study outlined a structural approach to LMS usage analysis, including the focus on preprocessing, calling it “arguably the most difficult task in the Web Usage Mining process due to the incompleteness of the available data” (Srivastava et al., 2000, p. 14). Fortunately, with Instructure’s Canvas™ LMS, the preprocessing stage is made easier through a vendor design choice to capture all of the page views data for every student in every course. These data are saved in the cloud on a series of database servers, and are available through the API as well as through a downloadable csv file available to administrators.

Srivastava, et al. (2000, pp. 16-17) also briefly discussed the primary methods of pattern discovery, including statistical analysis, association rules, clustering, classification, sequential patterns, and dependency modeling, all of which are viable options for usage mining analysts, and each with specific goals for specific applications. Romero and Ventura, in providing their list of approaches to pattern discovery (2010, p.
Cocea and Weibelzahl (2007a) used “classification via regression” to predict levels of engagement, resulting in the classification of two disengaged groups, students who click quickly through content pages and those students who take an extraordinary amount of time (a longer amount of time than needed to read the page) on each page. In both cases, these groups appeared to be clearly-delineated groups with clearly-delineated LMS usage patterns, and these methods were therefore considered significant as a possible starting point for pattern exploration in the current research project.

The most relevant study dealt with prosodic data in a linguistical context (Xydas, Spiliotopoulos, & Kouroupetroglou, 2004), using an XML annotation scheme that may be quite helpful within this study. To extend this approach, broad support for ensemble voting in regard to multiclass support vectors (Chen, et al., 2008) and Perfect Random Tree Ensembles (Cutler & Zhao, 2001) showed a higher success rate at classification. Dzeroski and Zenko (2004) extended the ensemble approach to classification by using the technique of “stacking” (combining methods in addition to combining voters). Both of these methods (ensemble voting and combining methods) were used during the data analysis phase and will be described and then utilized in the remaining chapters.

Data clustering has also been successfully applied (Talavera & Gaudioso, 2004), especially in the case of fuzzy clustering (Hogo, 2010), and classification by association rules, mimicking an e-commerce approach, found success among two representative
studies (C. B. Baruque, Amaral, Barcellos, da Silva Freitas, & Longo, 2007; Chanchary, Haque, & Khalid, 2008).

Finally, Srivastava, et al. (2000) briefly discussed pattern analysis, especially the pursuit of filtering out “uninteresting” rules or patterns, and the call for data visualization to assist in this process. Both of these considerations merited further consideration in this study and were used when applicable.

Classification and Exploration of LMS Usage Patterns from Web Log Data

Of particular interest to the research questions that drove this study is the concept of exploring student usage patterns within the LMS. The classification models and implications of the studies found in the literature were important to note, as they set up the classification environment which provided the pre-processing step which then helped generate the exploratory models for this study.

The intent of this study was to explore LMS usage patterns in a non-judgmental way. Therefore, constructs such as engagement, disengagement, and motivation have connotations that, if used to drive the study, might steer the exploration. This study was conducted in a way that focused more on pattern discovery and its possible use in classification models, over the discovery of patterns matching one or more constructs.

Within the Unified Learning Model, motivation is defined as “the impetus for directing working memory to a task” (Shell et al., 2009). Cocea and Weibelzahl (2007b) agree that motivation is a cognitive process, and though they assert that there is “no specific definition for engagement as a psychological concept” (2009, p. 2), they cite two theories which refer to it: Flow Theory (Csikszentmihalyi, 1998) and Theory of
Engagement (Shneiderman, Alavi, Norman, & Borkowski, 1995). Flow Theory is described by Csikszentmihalyi as “when a person faces a clear set of goals that require appropriate responses” (1998, p. 1). Though the article does not explicitly tie Flow Theory to engagement, it seems by definition that flow is the epitome of engagement; that is, when one is fully engaged in a task, one is in a perfect state of “flow.” In the Theory of Engagement, espoused by Schneiderman, et al. (1995, p. 7), the theory is built upon experiences with novelty in technology-rich classrooms and the heightened level of engagement that students and faculty members report from those experiences. Since “smart” classrooms are no longer new, the honeymoon may be over, but the Theory of Engagement still refers to the appropriate use of instructional technology, novel or otherwise, as a way to engage students.

As introduced briefly above, one research group of prime interest in this study was that of Cocea and Weibelzahl, who have worked together to produce a number of studies directly relevant to this one. In one study in particular (2006), they formulated decision trees to estimate the learner’s level of motivation. They discovered that motivation inside the LMS environment could be classified as “engaged” or “disengaged” based primarily upon the time spent on a content page and the number of timeouts that occurred while the students were logged in. Disengaged students tended to either click through pages very quickly (too quickly to actually read any of the content with any kind of comprehension), or click through very slowly (thus multi-tasking during time that should have been spent on reading the page content) with a number of timeouts due to lack of activity in the LMS. As shall be seen in the analysis phases, this was the primary point upon which this study began, though the value-laden constructs of
“disengaged” was quickly abandoned for the less loaded terms of “Fast Clicker” and “High Times Out” students. Although Cocea and Weibelzahl concluded that fast clicking and a high amount of user times out constitute “disengagement” in the context of tutorial software, here they are viewed as convenient points of departure in the exploration of student usage patterns in an LMS, especially as those patterns relate to similar findings in student writing assignment text.

Other ways that researchers have classified students in terms of engagement have included using “time on task percentage” and “average session duration” (Beck, 2004; Hershkovitz & Nachmias, 2009b), pace of students’ navigation among the content page areas (especially relative to the class average) (Hershkovitz & Nachmias, 2009a), and two espousing Keller’s (1987) ARCS model (attention, relevance, confidence, and satisfaction), with an emphasis on the attention and confidence components on the model (Hurley, 2006; Zhang, Cheng, He, & Huang, 2003).

In addition to studies focusing specifically on classification methods, others focused generally on various aspects of student motivation, including producing a number of rules (inputs leading to outputs) that aid in motivation diagnosis (De Vicente & Pain, 2002), and the dynamic mixture model, that combines Item Response Theory and Hidden Markov Modeling to predict both student motivation and proficiency (Johns & Woolf, 2006). One study emphasized the importance of the design of the e-learning environment and its effect on student motivation (Ramaha & Ismail, 2012). Others used special approaches to analysis, such as a Bayesian network (Arroyo & Woolf, 2005), and “text replays,” which consists of a special coding procedure based on the textual aspect of the log files, in contrast to replaying video screen capture data (Baker & de Carvalho,
Four additional studies focused specifically on “gaming the system” (Baker et al., 2008; Baker, Corbett, & Koedinger, 2004; Rowe, McQuiggan, Robison, & Lester, 2009; Walonoski & Heffernan, 2006). Though this phenomenon is certainly related to student engagement with the LMS, it is difficult to determine with log files alone, and is therefore not within the scope of this study.

**Tools and Practices for Analyzing Web Logs**

Analyzing Web logs is a somewhat different undertaking compared to LMS usage analysis, though it is related to it in that Web logs are normally built around user navigation events, which consist almost entirely of mouse clicks among various navigational links provided within the LMS (Ingram, 2000). Web logs normally contain information such as the URL of the page visited by the user, the time and date of the visit, the IP address of the computer that the user was using to visit the Web page, and the computer platform and browser environment that the user was using at the time.

In the field of education, there is also a great interest in which pages students visit when engaged with online educational content and activities, when they visit them, in which order they visit them, and ultimately why they visit them. Research in this area (Hanna, 2004) is often centered upon gaining knowledge about actual student navigational behavior as compared to how teachers and designers think or hope students will navigate. The ultimate objective in most of these research projects is to discover knowledge about how students learn online, so that instructors and designers can make design decisions that facilitate greater student learning.
A significant number of educational data mining studies have centered on Web log analysis, with exemplars including ranking topic relevance in the context of the individual user (Wang, Chen, Tao, Ma, & Wenyin, 2002), analyzing learner’s behavior (Psaromiligkos, Orfanidou, Kytagias, & Zafiri, 2011), and supporting e-learning decision-making (Monk, 2005). For a full description of sources related to Web log analysis in an educational environment, please refer to Chapter 1 of the text (Romero, 2010).

**Instructure’s Canvas™ as a Log Data Collection Tool**

The LMS has been sometimes cited as a factor in the outcome of studies focused on other subjects, such as the teacher’s involvement in the Moodle LMS (Clark, Beer, & Jones, 2010) and data mining as it is done in a Moodle environment (Baruque, Amaral, Barcellos, Freitas, & Baruque, 2007; Baruque et al., 2007). Instructure was created in 2008 and released the Canvas™ LMS in 2011 in part to provide a viable LMS alternative to those offered by BlackBoard™. Since that time, Canvas™ has been adopted by over 300 educational institutions at all levels (Canvas -- about us), and is growing in its user base daily. Thus, at the time of proposal for this project there had been little research done on the suitability of the Canvas™ LMS for data mining projects (MacNeill, et al., 2012, Voss 2013). The current study has discovered and utilized features of the Canvas™ LMS that support and promote data mining.

The Canvas™ LMS automatically captures page view log data on all students for all courses in its entire network. Thus, in this design choice alone, Canvas™ is unequalled in its support of educational data mining. It makes these data available to
users and instructors through its newly-released Analytics feature, with a visualization scheme that makes it simple for educators to see trends in student activity in the areas of conversation, assignment submissions, and grade attainment.

At a higher level, administrators can supply instructors with page views data inside the Canvas™ LMS, which are similar to the Web usage log data that have been described at length earlier in this chapter. The last 300 page views data are available in handy .csv file format, or all of the page views data stored in the database are available via screen-scrape within the administrator’s console (though this method requires much manual manipulation to gather all of the available data).

At a still higher level, instructors and administrators may collect data using the Canvas™ API, which provide a still richer array of dimensions. However, there is a difficult learning curve to using the API, and most users will most likely find this avenue of data collection challenging.

For the purposes of this study, the log data provided in Canvas™’ page views functionality were sufficient to classify students by their usage patterns within the LMS.

**Predicting Student LMS Use**

Data mining models are built to help predict future behaviors. Hanna (2004) described the model as being “built once” and then subjected to visualization methods based on the analyst’s needs. This approach is differentiated from that used by analysts when employing statistical packages, which Hanna asserts has to be selected, executed, and reported each time the analyst desires to discover or predict any relevant behavior.
Gaudioso, et al. (2012), and Iksal and Choquet (2005) also provided frameworks for eliciting predictive models.

Gasevic, et al. (2011), discuss a “group-on”-style approach (referred to as “folksonomy”) to elicit suggestions for generating an ontology-enhanced learning environment based on Semantic Web Mining. Tanimoto (2007) discussed ways to improve the prospects for educational data mining, including a good discussion of the concept of “unobtrusive assessment” (Tanimoto, 2007, p. 2).

A few researchers investigated motivation and engagement. Cocea and Weibelzahl (2007b) used a classification via regression approach to investigate motivation diagnosis in adaptive systems, engagement prediction (2007a), and disengagement detection (2009). Beyond being the recognized authorities on studying and predicting engagement, they provide a good framework of classes of engagement including “engaged”, “disengaged”, and “neutral”. Hershkovitz and Nachmias (2008) also provide dimensions of engagement, using Learnograms to produce clusters fitting their three dimensions of motivation. A few other approaches have been taken to extract significant items from e-learning data, including Semantic Web (Jovanovic et al., 2007; Velásquez, Dujovne, & L’Huillier, 2011) and genetic programming (Zafra & Ventura, 2010).

Finally, a philosophical paper (Knox, 2010) warns those who conduct educational data mining to tread carefully, for although one might think that the research is “unobtrusive,” the very fact that the students are aware that their clicks are being tracked may already be affecting the educational process in a negative manner.