EXTRACTING QUANTITATIVE INFORMATION FROM NONNUMERIC MARKETING DATA: AN AUGMENTED LATENT SEMANTIC ANALYSIS APPROACH

By

INIGO ARRONIZ
B.A. Universidad de Navarra, 1997

A dissertation proposal submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Marketing
in the College of Business Administration
at the University of Central Florida
Orlando, Florida

Spring Term
2007

Major Professors:
Ronald E. Michaels
Steven M. Shugan

ABSTRACT

Despite the widespread availability and importance of nonnumeric data, marketers do not have the tools to extract information from large amounts of nonnumeric data. This dissertation attempts to fill this void: I developed a scalable methodology that is capable of extracting information from extremely large volumes of nonnumeric data.

The proposed methodology integrates concepts from information retrieval and content analysis to analyze textual information. This approach avoids a pervasive difficulty of traditional content analysis, namely the classification of terms into predetermined categories, by creating a linear composite of all terms in the document and, then, weighting the terms according to their inferred meaning. In the proposed approach, meaning is inferred by the collocation of the term across all the texts in the corpus. It is assumed that there is a lower dimensional space of concepts that underlies word usage. The semantics of each word are inferred by identifying its various contexts in a document and across documents (i.e., in the corpus). After the semantic similarity space is inferred from the corpus, the words in each document are weighted to obtain their representation on the lower dimensional semantic similarity space, effectively mapping the terms to the concept space and ultimately creating a score that measures the concept of interest.

I propose an empirical application of the outlined methodology. For this empirical illustration, I revisit an important marketing problem, the effect of movie critics on the performance of the movies. In the extant literature, researchers have used an overall numerical rating of the review to capture the content of the movie reviews. I contend that valuable information present in the textual materials remains uncovered. I use the proposed methodology to extract this information from the nonnumeric text contained in a movie review. The proposed setting is particularly

attractive to validate the methodology because the setting allows for a simple test of the text-derived metrics by comparing them to the numeric ratings provided by the reviewers.

I empirically show the application of this methodology and traditional computer-aided content analytic methods to study an important marketing topic, the effect of movie critics on movie performance. In the empirical application of the proposed methodology, I use two datasets that combined contain more than 9,000 movie reviews nested in more than 250 movies. I am restudying this marketing problem in the light of directly obtaining information from the reviews instead of following the usual practice of using an overall rating or a classification of the review as either positive or negative.

I find that the addition of direct content and structure of the review adds a significant amount of exploratory power as a determinant of movie performance, even in the presence of actual reviewer overall ratings (stars) and other controls. This effect is robust across distinct opertaionalizations of both the review content and the movie performance metrics. In fact, my findings suggest that as we move from sales to profitability to financial return measures, the role of the content of the review, and therefore the critic's role, becomes increasingly important.

Dedicated with love to my wife Wendy,
our daughter Natalia,
and all my family and friends.

ACKNOWLEDGMENTS

This dissertation will not have been possible without the help and encouragement of many. First to my wife for all the patience, countless pep talks and always being ready to listen to my difficulties. Not only did you read the manuscript several times, edit it, checked it again and again but you gave me the inspiration and motivation to keep going even when all odds were against me.

While my wife has been the rock that offers me a place to anchor my ship even in the most turbulent times my committee members have been the compass that ensured that the ship arrived to safe port after all its trips and adventures. Among the committee member I want to express my special thanks to my co advisors Steve Shugan and Ron Michaels. From Steve not only I learned what constitutes a worthwhile research problem but Steve's enduring optimism and amazing conceptual clarity carried the day in many instances when seemingly insurmountable obstacles appeared in the horizon. I want to thank Steve again for actually agreeing to chair the dissertation to begin with. This was way beyond the line of duty. From Ron Michael I have learned how to manage a project that not only is incredibly complex and challenging but at times I thought there will be no end to. From Raj Echambadi I have learned how to persevere under all circumstances, even when the storm is such that staying afloat seems an impossible task. From Jai Ganesh I have learned how to make sure that every journey is such that as it has a beginning it needs to have an end in sight. And from Steve Sivo that while research is one of the most important things in academia, there are other things that need to take precedence at times and that family and friends are more important that a brilliant professional career.

I also want to thank Mohan Sawhney for believing in me even at the time when my ship had not arrived to port yet. Mohan took me under his wing and taught me all that I have learned so far

about "real" business and expanded my thoughts regarding marketing as true value discipline. You my friend are a true gentleman and a scholar.

I want to give thanks for all the help and support that I got from all the PhD students at UCF and UF. I truly think that going through a PhD program is a life altering experience. The experience is not just comprised of books and lectures but of shared experiences with other individuals with whom you go through the journey together. A special thanks to my cohort marketing students is in order. To Nacef Mouri for all those long nights that we spent together in the cubicles looking at papers and textbooks when we could have just gone home and enjoyed the many pleasures that Florida has to offer. To Mike McCardle for those amazing Super Bowl parties that you threw for the rest of the students in the program. To both of you for all the encouragement and help during the program.

To my parents Jose Antonio Arroniz and Mentxu Aquerreta for always emphasizing the importance of an education despite not having the opportunity to attend college themselves. From my father I have learned that while luck is important, with hard work the odds are always in your favor. From my mother I have learned how to put all aside to help others when they need someone to help them.

I also want to thank my daughter Natalia Arroniz, the source of my inspiration in the later stages of the dissertation. From you I have learned many things about life and about myself in the short year that you have been with us; but if I have to mention just one thing that you have taught me is how hard you can fight to achieve something when you really want it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

"The investigation of the meaning of words is the beginning of education"

–Antisthenes

Marketers face an interesting conundrum. The amount of machine-readable textual or nonnumeric data available has grown exponentially in recent years with the widespread proliferation of computer databases, e.g., Lexis Nexis and the advent of the Internet (Urban & Hauser, 2004). Despite the explosion of nonnumeric data, however, there is no tool currently available to extract information from these vast arrays of unstructured data, prompting calls for tools that can help researchers reliably obtain valid information from nonnumeric data (Shugan, 2002, p. 376). This dissertation attempts to fill this void by developing a methodological approach to extract quantitative information from large sets of textual data.

Why is extracting information from nonnumeric data important? Textual data that may be of interest to marketers, for example, include consumers' descriptions of their experiences in chat rooms, user participation in blogs, consumer communications in brand communities, professional reviews of products, and analyst reports about companies. Fundamentally, textual data, as compared to quantitative data, may possess nuanced information and, hence, may be very useful to marketers. Moreover, information from nonnumeric data has the potential to supplement, or in some cases, supplant numeric data. While some of the readily available nonnumeric data may be collected in numeric form by the application of primary research techniques, such exercises oftentimes require large amounts of resources and time. Therefore, traditional quantitative data collection may not be

feasible in some instances, and, in such cases, information from nonnumeric sources may be used to supplant numeric data.

Moreover, the level of obtrusiveness of traditional measurement and other serious measurement limitations (i.e., recall effects in survey research) may tip the scale toward the use of available nonnumeric data. In other cases, in which the timing of information is of critical importance, managers can use the information from nonnumeric data to obtain a quick feel of the problem studied. Sometimes, extracting information from existing nonnumeric data is the optimal solution. For example, the difficulty in collecting numeric information makes measurement of word-of-mouth effects a vexing research issue (e.g., Rust, Lemon, & Zeithaml 2004), even though there is abundant nonnumeric data available on the Internet (www.epinions.com is one example).

The use and analysis of textual data are by no means new to the marketing literature (c.f., Abernathy and Franke (1996) for examples on the use of content analysis in advertising) and the social sciences in a more general fashion (Pooping, 2000). Historically, most content analysis has been conducted using human coders. This technique has proven very valuable in its own right; however, the extant form of content analysis has several important limitations that hamper its use in large collections of documents: a) it is extremely taxing and consumes large amounts of resources, including the use of expert time training coders and/or coding the texts; b) the coding is necessarily subjective and, therefore, different coders will code the same text differently; and c) this coder subjectivity necessitates the use of multiple coders to test the degree of uniformity in coding, i.e., intercoder reliability, which increases resource requirements exponentially. Considering the increase in computing power and the large amounts of textual information commonly available, the perfect window of opportunity now exists for the development and use of new approaches, methods that will access and quantify the wealth of textual information currently residing on computer networks

and databases with fewer requirements of human resources. That is precisely the objective of this dissertation research.

Specifically, my research goals in this dissertation are threefold. I intend to show the following: a) that rich information resides in textual data, b) that there are systematic and scalable ways of extracting and analyzing information obtained from texts, and c) that this information can be used by marketers to better understand interesting phenomena and hence make informed decisions.

In order to accomplish the goal of extracting information and quantifying textual data, I propose and develop an approach called Augmented Latent Semantic Analysis (ALSA). In the proposed method, I move away from the traditional content analysis literature, which characteristically groups words or expressions into discrete content categories. Traditional approaches necessitate the creation of a set of formal rules or sometimes general guidelines that allow assigning each word or set of words to a given content category by the coder. Instead, in the proposed methodology, I attempt to convert a major weakness of analyzing a large set of documents (i.e., resource requirements because of the size) into an advantage by learning from a large set of documents how rules of assignment should be created.

I borrow from developments in the information retrieval discipline. In doing so, I propose creating linear composites by which each word in the document is weighted according to its inferred meaning to measure concepts of interest for the researcher. I use collocation, or the relationship between two words or groups of words that often go together and form a common expression, to infer the meaning that words have in the text. I do so by analyzing the placement of the particular term, in the context of its proximate terms, and across each of the texts in the available group of documents (i.e., corpus).

While we as researchers merely observe words and expressions in each of the documents in our dataset, I argue that, according to communication theory, there is an underlying lower dimensional space of concepts that drives word usage when composing the message. This underlying space of content implies that the observed word occurrence is by no means random both within and across documents. I suggest that the (scaled) frequency of appearance of a given term in a document is linked to those underlying latent concepts that the composer of the message intends to communicate to the reader. I attempt to retrieve this lower dimensional latent space of concepts from the observed word usage using a Singular Value Decomposition of the term document matrix. The semantics of each word are inferred by identifying its various contexts across documents. This is accomplished by creating a similarity measure that scores the degree of proximity between any two terms in the concept space. Then, I derive similarity weights to create linear composites that capture the essence of the concepts that I intend to measure across the documents. To ensure robustness of the method, I propose a two-stage procedure. In the first stage, the weights are computed based on a calibration dataset, and in the second stage, these computed weights are used to obtain the measures that capture the intended construct of interest.

An important step to make this methodology useful in any empirical application is the selection of seed words. A small set (typically one or two) seed words will be used to anchor the construct of interest in the measurement process. The search for seed words should be most appropriately guided by the theoretical understanding of the constructs of interest. The proposed methodology maps terms observed in documents to a set of latent constructs that are in principle difficult to label; these difficulties limit their usefulness in applied contexts. Note, however, that in cases in which the research area is in a nascent stage the search for seed words may entail the use of

either traditional content analysis in a subsample of the text and/or leveraging the large size of the dataset to uncover important concepts within the text.

I intend to demonstrate the usefulness of the proposed methodology by examining the potential exploratory power that quantified information extracted from nonnumeric data has compared to traditional measures of content. To accomplish this objective, I will examine an important marketing problem—the impact of critiques/ratings of professional reviewers on movie performance.

I chose this particular area for three primary reasons. First, the movie industry is particularly interesting from the product development and launch process perspective since product life cycles are extremely short, e.g., most movies do not stay in the theaters for more than eight weeks. This brief window of opportunity provides marketers with little room to maneuver in case of an initial poor response from the movie-going public. Under these circumstances, a priori forecasting of the success or failure of the motion picture becomes critical. If information from movie critics can be used prior to the launch, then better product introduction decisions can be made based on this information. The prevalent use of prescreenings and audience showings in the film industry facilitates the a priori forecasting of movie success based on prescreening reviews that take place prior to the launch of the movie. Second, extracting information from movie reviews is a particularly challenging endeavor as critics in this product category often use sarcastic language and connotation. If extraction of information from movie reviews is possible, this is evidence that the task will be simpler in other cases in which denotative use of the language is predominant. Third, this setting allows a strong and objective external validity test of the developed text-based metrics to be conducted. By using the readily available metric of review content, i.e., star ratings, I will also demonstrate the predictive validity of the text-based metrics that I derive.

There is a large amount of empirical evidence that suggests professional movie critics' reviews are related to box office revenues (Basuroy, Chatterjee, & Ravid, 2003; Eliashberg & Shugan, 1997; Jedidi, Krider, & Weinberg, 1998; Litman, 1983; Litman & Ahn, 1998). Past studies in this area have used indirect measures of the actual content of the movie review. In particular, many critics provide an overall rating of the movie, often on a 0 to 4- or 1- to 5-star scale, and it is this numerical rating that is most frequently used in the extant literature to assess the effect of reviewers on movie-going behavior, and ultimately on box office revenues.

In this dissertation, I suggest that measuring the impact critics could have on moviegoers' experiences is undermined by this oversimplification of the actual process. If the overall judgment or the rating were ultimately the only valuable source of information in the critics' reviews, we would rarely find long and intricate movie reviews in the marketplace, as moviegoers would not use them. Also, movie-going experiences, similar to many other hedonic product consumption experiences, is dependent on customer preferences, and those preferences vary greatly. If this is the case, then a holistic evaluation may not suffice, and the actual content delivered in the review may be indeed an important factor. I intend to show that the way the content is delivered in the review and the content of the review itself can be quantified directly, and thus used to assess the effect of critics on movie performance.

In addition to the proposed ALSA method, I also apply two other conventional computer-assisted quantitative content analysis approaches that have been used in marketing to accomplish the same goal. This exercise will allow me to evaluate the relative efficacy of the proposed ALSA approach in quantifying textual content when compared with existing methods.

The proposed empirical setting requires the collection of a large set of professional reviews for a set of movies. To do so, I use a crawler or spider that allows for the controlled collection of

web-based information given a set of prespecified parameters. In this process, I will tap into existing databases such as Internet Movie Database to access the records. I combine these movie-specific reviews with box office and marketing effort information to investigate the impact that reviewers have on how a movie performs.

The remainder of the dissertation proposal is organized as follows. Chapter Two presents a review of the pertinent literature dealing with the analysis of nonnumeric information. Chapter Three draws upon traditional content analytic techniques and extant models of search and document retrieval involving textual information to develop and explain the theoretical and mathematical underpinnings of the proposed ALSA approach. In Chapter Four, the design of an empirical demonstration of the usefulness of the proposed approach is presented. Moreover, a comprehensive comparison is made between the proposed ALSA approach and two traditional computer-aided content analytic approaches used in the extant marketing literature.

CHAPTER TWO: LITERATURE REVIEW

Measuring Content in Textual Data: Content Analysis

Content analysis, which involves obtaining quantifiable information from textual and other nonnumeric data, is by no means a novel idea (see Webber, 1990, or Pooping, 2000, for reviews on content analysis history and applications). Techniques that relate to the essence of content analysis have been described as early as the Middle Ages when scholars studied the Bible and tried to uncover premonitions inscribed in the text using the information in dates and references made in the Holy Scriptures.

Late in the 19th century, mass communication researchers started to develop content analytic methodologies that were the precursors of the techniques traditionally used in marketing. The basic theoretical underpinnings that drove research in mass communication can be summarized in the realization that the message transmitted through the media has an effect on the receiver(s) of the message. If this is the case, it is of particular interest to examine the nature of the content in the message in order for researchers to understand the potential effects that mass communication has on the audience.

In one of the first accounts of this technique, Speed (1893) studied the type of coverage that different subjects received in four New York newspapers. Speed argued that there was evidence that over the period studied (1881 to 1893), newspapers changed to include more gossip and scandal stories and devoted less coverage to cultural stories. Other authors (Wilcox, 1900; Street, 1909) continued with this line of research, finding support for the fact that newspapers were becoming more sensational in content at the time. The methods used to measure content were usually based

on measuring the length of the articles that were published containing the different types of news (as determined subjectively by the author). To measure the length, the author used a ruler as opposed to measuring length by some other measure such as counting the number of words, sentences, or paragraphs.

Other research questions such as whether the coverage of news was slanted or biased were studied using content analytic techniques in the beginning of the 20th century by Lippmann and Merz (1920). In their study, they compared the news published in the *New York Times* to factual information available post facto regarding activities on the Russian front during World War I. Looking at the content in the news, they determined that the reporting of the events was ". . . almost always misleading" (p. 42).

Later in the century, a group of political scientists headed by Harold Lasswell undertook the study of how governments use mass media outlets as a military weapon to diffuse propaganda. The initial methods developed during this period were mainly based on the count of words and using these counts to conduct analysis (Lasswell, 1927, 1941) that set the stage for the more advanced techniques that would evolve later with the use of the computer. These researchers were faced with the usual caveats of labor-intensive methods: constraints on time and resources made the analysis difficult and tedious. De Sola Pool (1980) writes regarding the tediousness of the job involved: "I stopped doing content analysis before Phil Stone had developed the General Inquirer, because it was too hard. The amount of work involved for the product was enormous" (p. 245).

However, despite the impracticality in large datasets, content analysis was already an often-used research method by the 1940s (see Diefenbach, 2001, and references therein). In addition to researchers in mass communication and political science, disciplines such as psychology were using content analytic techniques as a means to capture important information that could be used to infer

the state of mind of subjects. The basic theoretical grounding for the use of content analysis in psychology is that personality traits or psychopathologies are manifested in verbal or written communication involving the patients. If this is the case, then studying the content of the messages should be valuable as a diagnostic tool for assessing the mental state of the patient. In this line of work some of the initial work involved verbal behavior.

In a dedicated issue of *Psychological Monographs,* Johnson (1944) proposed the use of the type-token ratio (TTR) as a tool to diagnose some mental disorders. The TTR measures the ratio of the number of distinct words in a text to the total number of tokens (words) in the text given a figure of the vocabulary diversity. The author proposes the use of a standardized measure that tackles the problem of lexical diversity increasing with message length. Other theoretical work appeared using different measures of vocabulary diversity to assess language behavior. For example, Boder (1940) proposed the use of the ratio of adjectives to a set number of verbs used in verbal or written expression. In the same volume of *Psychological Monographs,* Fairbanks (1944) and Mann (1944) reported on empirical findings that validated the use of these types of measures of lexical diversity in looking at standardized TTR and adjective-to-verb ratios for schizophrenic and nonschizophrenic (college undergraduate freshmen) subjects. Finally, Chotlos (1944) commented on the length needed for adequate reliability when pursuing standardized TTR measures used to assess physiological pathologies.

New tests and techniques that allow clinical psychologists to assess the state of patients using their verbal behavior were developed later (cf., Gottschalk, 1995). Overall, results in this literature can be summarized by saying that there is relationship between patterns of speech and psychological state and that "the hypothesis that speech variability increases with successful therapy has generally been supported" (Holsti, 1969, p. 75).

Although initially limited to studies that examined texts for the frequency of the occurrence of identified terms (word counts), by the mid-1950s researchers were already considering the need for more sophisticated methods of analysis, focusing on concepts rather than simply words, and on semantic relationships rather than just presence or absence of terms or tokens (de Sola Pool, 1959). The generalization of the use of the computer in content analysis methods freed the researcher from the arduous task of manually repeating mechanical tasks that are very efficiently programmed and executed by machines (Pooping, 2000). While the use of the computer is becoming ubiquitous in the different types of content analytic methods, the level of automatization of tasks varies greatly from method to method. However, it is important to note that more automatization usually comes at a cost of more shallow analysis since, as many researchers have argued, computers help in the process but they are still "dumb clerks" (Stevenson, 2001).

Methods integrating the use of computers originated in the 1960s with the introduction of the General Inquirer by Stone and colleagues (Stone et al., 1966; Kelly & Stone, 1975). This content analytic approach mapped words to categories (dictionary) that were considered meaningful for different research projects. The core of the program was an engine of rules that allowed homograph words (that is those that have the same spelling but different meaning) to be disambiguated to be later classified in different categories.

Contemporary computer-based methods build on past applications that allow researchers to assess, among others, mental maps (Palmquist, Carley, & Dale, 1997; Carley, 1990). New techniques are flourishing in other research disciplines such as Natural Language Processing (NLP) and Information Retrieval (IR) that hold the promise of making possible more automatization of complex tasks that had been performed by human raters. One example of such promising technology is recent work in automatic summarization of text that allows one to obtain a

compilation of text that has only been processed by the computer (cf., Mani & Maybury, 1999, and references therein).

Marketing and the Analysis of Textual Data

Content analytic techniques were praised as promising by the marketing discipline as early as the 1970s by Kassarjian (1977) and Hoolbrook (1977). Both noted the potential in empirical applications that content analytic techniques offered to the marketing discipline. Those authors also noted that with the more common use of mainframe computers the use of computer-assisted content analytic techniques offered great potential.

Marketing has a long history of using human coded content analysis to obtain information from textual and nontextual messages (e.g., visual ads). There is a rich tradition involving the use of content analyses of advertising messages to assess the impact that ad content or type messages have on the effectiveness of particular advertising (see Abernethy and Franke 1996 for an excellent review of empirical results in this area). But there are few examples of the use of computer-assisted content analysis in the marketing literature.

While analyzing textual data, computers have been used as a tool that allows the researcher to organize categories and query them after the process of traditional coding has already taken place. Kennedy, Goolsby, and Arnould (2003) provide a case in point in which this type of qualitative content analysis is aided by computers. The authors conducted a study that examined the dynamics involved in the implementation of a customer orientation program in a public school district. The authors used a paired-comparison ethnographic design in which two institutions are studied in depth, one in which a customer orientation program is successfully underway and one in which the implementation of a similar program was struggling. The study provides insights as to how an

organization adopts a customer orientation philosophy by modifying the roles of leadership, interfunctional coordination, and the collection and dissemination of market data. The authors used NUDIST (nonnumerical unstructured data indexing, searching and theorizing), a qualitative content analysis program, to process 99 transcripts of meetings and interviews that the authors conducted with different sets of stakeholders. The authors are not explicit about how they used NUDIST while conducting their empirical analysis. Given the available information in the article, it seems likely that the authors used the program as a tool to formalize the usual human coding process, and to store and organize the categories and original transcripts. This is a form of qualitative content analysis in which the software is used in a way that facilitates the work of the coder but the human coder still plays a central role. Other studies in marketing use similar methods for content coding (cf., Wheeler, Jones, & Young 1996; Baines, Scheucher, & Plasser, 2001; Craig-Lees & Hill, 2002). Although qualitative content analysis is an important methodology in its own right, this study focuses on quantitative methods.

In other instances, researchers create fixed rules that assign words or expressions to categories, and the computer is used to carry out the classification. Along similar lines, Rosa et al. (1999) studied the origins and evolution of product markets from a sociocognitive perspective. The authors describe product markets as socially constructed knowledge structures that arise from the interaction of producers' and consumers' conceptual systems. The authors examined the evolution that occurs in emerging product markets. To study the dynamic nature of the product market from inception to the establishment of a stable category, the authors used the stories that consumers and producers tell each other in several publications. These stories were interpreted by two individuals who manually coded a portion (about 10%) of the available textual materials. From this coding system, they derived a set of rules and then used a program to code the remainder of the text. Using

this semiautomated system, the authors coded general references to the minivan, car, station wagon, and van categories. The authors also coded for the use of these categories as points of reference and for comments on the acceptability of existing minivan models on a set of predetermined attributes. Using the counts in these categories, the authors found that category stabilization creates significant differences in how consumers and producers use product category labels for emerging and preexisting product categories. This article is one of the few instances in marketing in which semiautomated content analytic methods are used for quantitative hypothesis testing. Unfortunately, the authors do not explain in detail the coding process used in the study.

A similar although more elaborate rule set is used in other instances. Not only are inclusion rules used, but after the first coding of inclusions is conducted, a second classification (count) of exclusions based on another set of rules is conducted. Rosa's work (2001) is an example of this coding methodology. The author studied the use of embodied concepts to understand and solve ill-defined problems presented to marketing managers. The author conducted two studies: the first assessed the level of use of embodied concepts among marketing managers; the second study focused on the potential impact environmental cues and dispositional factors have on the use of embodied concepts by managers. Study one consisted of two separate data collections. In the first collection, 33 managers responded to questions regarding the future of their businesses (i.e., an ill-defined problem). In the second part, excerpts of the answers previously obtained were used as stimuli to elicit embodied concepts from 80 managers. Study two was a field experiment conducted with 68 marketing managers playing a brand-management simulation during which verbal protocols were collected. The authors counted embodied concepts in several categories. To collect embodied concepts and given their referential nature, instances in which expressions were used for alternative meanings were not counted. To resolve this issue without using human coders, a list of words and

expressions that refer to the embodied concepts was first developed. The computer program, VBPro in this case, codes the texts for these categories. After the first count takes place, another list of exceptions for each initial rule is given with the intention of disambiguating the embodied concepts from other uses of the expression or word. The difference between the initial count and the second count is the number of references to the embodied concept. Results show that embodied-concept use is common among marketing managers, and that it is influenced by dispositional factors and environmental factors.

While traditional content analysis and semiautomated content analysis methods are useful approaches in terms of analyzing nonnumeric data, there are limitations. First, the theory construction and the technique used are inherently correlated to create a reverse demand effect, i.e., researchers know what they are looking for and then look for it using a priori classifications. This could inject bias into the process. Second, these approaches require a great deal of time and effort in terms of developing exhaustive, subjective coding schemes. Third, because of the effort-intensive nature of the coding, these approaches are typically used for smaller sets of documents, and any such quantification may be plagued by small number problems. Fourth, all words or terms used in a document are equally weighted, which may be an untenable assumption since some expressions have greater degree of emphasis than others.

Outside of marketing, there is an array of promising methods being developed to examine textual content in other disciplines. For example, resonance theory, a communication-based theory, dictates that a word is relevant in the communication process if the word plays a central role in its relationship to other words in the message (Corman, Kuhn, McPhee, & Dooley, 2002). With the help of a computer, textual material can be converted into nets or trees of interrelated nodes of words or expressions. Once a complete network of the text(s) is created, network centrality

measures can be used to determine word importance within the text. An alternative method is the concept mapping approach that draws from the information theory paradigm (Miller & Riechert, 1994). If a document has a word that has a dramatically higher probability of occurrence, chances are that the word is related to an important aspect of the document and these probabilities of occurrences are used to examine the importance of content.

In summary, while the aforementioned studies are an indication of the interest that text-based methods elicit and of their potential, they are not appropriate for the research goals outlined in this study. Newer perspectives are needed to develop a scalable, quantitative approach that allows for the measurement of concepts across multiple documents.

CHAPTER THREE: THEORETICAL UNDERPINNINGS AND PROPOSED METHODOLOGY

The basic methodology involved in the transformation of textual content into numerical data for data analysis can be traced back to set theory (Franzosi, 1994). The basic idea is simple; we can assign different terms, words, or expressions to sets. After the assignment rules have been determined and every token is classified into each set, the cardinal numbers allow us to convert the once nonnumeric data (the text) into variables that are essentially counts (i.e., frequencies) of the elements in each set.

The essential issue for the researcher is how the symbols, words, etc. are assigned to the categories or sets in order to form variables of interest. Based on the extant marketing literature that uses textual data, we identify two commonly used methods to create the underlying rules of term assignment. On one hand, a researcher, guided by theory and a clear definition of the content category, makes a judgment call as to where each piece of textual data should be classified. This rule of assignment leads to what we term traditional content analysis conducted by human coders. On the other hand, a researcher can build a more rigid set of rules that map each term to a content category. A set of these rules is usually known as a dictionary in the computer-enabled content analysis literature. In a broad sense, there are two types of dictionaries, those that are context specific (hereafter referred to as ad hoc) and those that are more general in nature (for examples of general purpose dictionaries, see the Harvard IV or the Lasswell dictionaries).

While the aforementioned rules of assignment are useful in their own right, each has strengths and weaknesses. As I have already noted, the subjective assignment rule is the one that makes the most out of the context and hence is able to extract nuances from the text that a simple

dictionary may be hard-pressed to obtain. However, the amount of resources required as the number and length of pieces of text grows is so large that it becomes impractical even with reasonably small document sets. It is my contention that while maximum information may be extracted through the use of human coders, the coding process itself limits the amount of text that researchers can realistically use; hence, this method is impractical in many applications. In the tradeoff between the scope and breath of research and depth, the former is favored given the expansion of available text datasets. Implicitly, some validity is traded for better reliability of the coding of textual data since rigid rules such as dictionaries are easier to replicate and, hence, more reliable. It is interesting to note, however, that while for some tasks human coders perform significantly better than the more shallow techniques covered here, studies have shown that in some applications performance differences are negligible (cf., Simon & Xenos, 2004).

In juxtaposition to the subjective rules of coding, once the dictionary is available, a computer can quickly score any text. This makes the study of large datasets feasible and practical. On the positive side, there is also the high reliability of this method as the coding can be easily replicated. The difficulty in the case of the dictionary is its construction and effectiveness. As I have already mentioned, the same concept can be expressed using many words and expressions so the number of rules is inherently large for most constructs. General dictionaries are appealing since this rule-generating process has already been completed. Although the use of general dictionaries is an option when the construct of interest has already been studied using this methodology, often new constructs are of interest to researchers and the need for a more context-specific set of rules makes dictionaries a more effort intensive option.

To overcome these limitations, I propose using theory to anchor the meaning of the construct and the available text itself to derive rules that underline the measurement of the content

categories. This leverages the fact that there are large amounts of text available and shifts what was a practical limitation into a strength. To do so, I build on a rich set of methods dealing with text and information existing in other disciplines.

Information Retrieval and Search Models

I examined the information retrieval and document indexing literature to devise an automatic content extraction method that leverages the large number of documents in the dataset. The basic precept of these disciplines is to create systems that allow for an efficient search of all documents in a database by examining textual content and then, based on scoring algorithms, select the documents in the database that best match the users' queries (Baeza-Yates & Ribeiro-Neta, 1999). At a fundamental level, these models have much to offer toward information extraction of nonnumeric data. While their goal is not the measurement of constructs in the text, the models' performance in document retrieval is inherently affected by the content in the documents. As a result, I draw the theoretical and mathematical underpinnings for my approach from vector and latent semantic models commonly used in information retrieval. I explain these models in detail in the following sections.

The vector model

Vector models rely on the premise that the meaning of a document can be derived from its constituent terms. Each document is represented as a vector in a high-dimensional term space, and each unique term in the document corresponds to a dimension in the space. Let k_i be an index term (a word), d_j be a document (say, a movie review), and w_{ij} be a weight associated with each index and

document (k_i,d_j). The weight w_{ij} quantifies the importance of the index term for describing the document contents. Finally, let q be a user generated query (string of words).

Define: $w_{ij} > 0$ *whenever* $k_i \in d_j$

$w_{iq} >= 0$ associated with the pair (k_i,q)

$vec(dj) = (w_{1j}, w_{2j}, ..., w_{tj})$

$vec(q) = (w_{1q}, w_{2q}, ..., w_{tq})$

Each indexing term k_i is associated a unitary vector *vec(i)*. The unitary vectors *vec(i)* and *vec(j)* are assumed to be orthonormal (i.e., index terms are assumed to occur independently within the documents). The *t* unitary vectors *vec(i)* form an orthonormal array on a *t*-dimensional space. In this space, queries and documents are represented as weighted vectors. The similarity between the query, *q*, and the document, d_j, denoted by Sim(q,d_j) is given by the cosine formed by the two vectors in the *t* dimensional space (see Figure 1). Note that the similarity can then be measured by the cosine of the two vectors, that is:

$$\text{Sim}(q,d_j) = \cos(\Theta) = [vec(d_j) \cdot vec(q)] \;/\; |dj| * |q| = [\Sigma(w_{ij} * w_{iq})] \;/\; |d_j| * |q| \qquad (1)$$

Since $w_{ij} > 0$ and $w_{iq} > 0$, $0 <= \text{sim}(q,dj) <= 1$

The set of documents that best matches (lowest distance or highest similarity) the query is presented to the user. The key question is then: how do we compute the weights w_{ij} and w_{iq} ? A representative weight must take into account two properties that are desirable when the goal is document extraction: a) Quantification of intradocument contents, that is, the importance of the term inside the document. This is called the *tf* factor, or the term frequency within a document. b) Quantification of interdocuments separation or how discriminating is that

term considering the other documents inside the collection or corpus (dissimilarity). This term is called *idf* factor, or the inverse document frequency.

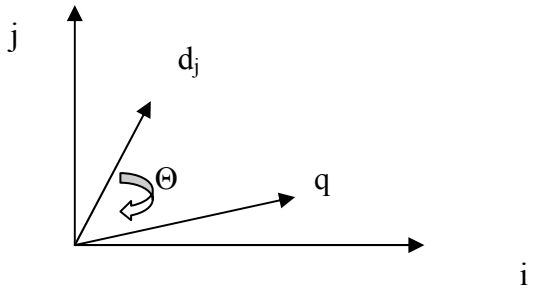

Figure 1 Graphical Representation of The vector Model.

Similarity between two vectors is given by the cosine of their angle. Θ

The final weight is computed as the product of the two parts $w_{ij} = tf(i,j) * idf(i)$. Let N be the total number of documents in the collection, n_i be the number of documents that contain the indexing term k_i, and freq(*i,j*) be the raw frequency of k_i within d_j. A normalized *tf* factor is given by $tf(i,j)$ = freq(*i,j*) / max(freq(*l,j*)) where the maximum is computed over all terms that occur within the document d_j. The *idf* factor is computed as *idf*(i) = log (N/n_i). The best term-weighting schemes (called *tf-idf* weighting scheme) use weights that are given by $w_{ij} = tf$(i,j) * log(N/n_i).

Although the vector model is intuitively appealing, there is one major limitation—the terms are assumed to be independent. This may lead to the problem of synonomy, wherein many terms can be used to express the same thing. For example, car and automobile mean the same thing, but the similarity between some relevant documents may be low because they do not share the same words. Latent Semantic Indexing techniques can be used to avoid this problem of synonymy.

Latent semantic indexing (LSI)

Latent semantic analysis (LSA) is a generalization of factor analytic techniques for document term matrices. LSI is an empirical application of LSA to the document indexing problem. As in vector models, LSI relies on the constituent terms of a document to learn about the document's meaning. However, the LSI model assumes that the variability of word choice partially obscures the document's meaning. In other words, the terms in a document are somewhat weaker indicators of the concepts contained in the document. Therefore, LSI attempts to find the latent structure in term usage. This model assumes that words are chosen based on an underlying latent structure and that words are correlated mainly because of this underlying latency. These correlations between words are driven by the inferred meaning. LSI accomplishes the goal of finding the latent structure by reducing the dimensionality of the term-document space, thereby revealing the underlying, semantic relationships between documents.

Let t be the total number of index terms (tokens, words, or sets of words) and let N be the number of documents. Finally, let M_{ij} be a term-document matrix with t rows and N columns. Originally, the frequencies observed in each document are the components used in M. A transformation of the frequencies (e.g., logarithmic) is used to reduce distributional concerns that arise with the use of frequency data (i.e., skewness). In other instances, weight w_{ij} associated with the pair $[k_i,d_j]$ has been proposed. The weight w_{ij} can be based, for example, on a *tf-idf* weighting scheme. The matrix M_{ij} can be decomposed into three matrices (singular value decomposition) as follows:

$$M_{ij} = K\ S\ D^t \qquad (2)$$

where K is the matrix of eigenvectors derived from M M^t, D^t is the matrix of eigenvectors derived from M^t M, S is an r x r diagonal matrix of singular values where, and r = min(t, N), that is, the rank of M_{ij}.

The space reduction takes place after the decomposition per equation (2), and the reduction assumes that only the first *s* concepts are relevant. To reduce the dimensionality from min(*t, N*) to *s* dimensions, only the *s* largest singular values in S are selected, and the rest are substituted by zeros. This matrix is called S_s. The corresponding columns in K and D^t are kept. The resultant matrix called M_s is then given by

$$M_s = K_s S_s D^t \qquad (3)$$

where *s, s* < *r* is the dimensionality of the concept space.

The number of dimensions retained should be large enough to allow fitting the characteristics of the data but small enough to filter out the nonrelevant representational details. That is, if we choose too small a value for *s*, it will not fit the actual relationships well, as we are downplaying the semantic complexity of the text; a large value of *s* will capture too many nuances that are not descriptive of the concepts, hence adding unnecessary noise to the data.

The user query can be modeled as a pseudo-document in the original M matrix. The matrix $M^t M_s$ quantifies the relationship between any two documents in the reduced concept space. If the query was placed as the first document in the matrix M, the first row of $M^t M_s$ provides the rank of all the documents with regard to the user query.

I build on the concepts outlined above in LSA to create a new way of measuring content that is present in a document. I call this the 'Augmented Latent Semantic Analysis' approach.

Proposed Methodology: The Augmented Latent Semantic Analysis (ALSA) Approach

Unlike extant literature that uses ad hoc or general purpose dictionaries, I propose a new way of conducting content analysis that combines content analytic foundations and information retrieval methods to efficiently extract content from text. The proposed method builds on LSA and is capable of inferring semantic similarity from contextual information. The ability to obtain a measure of how similar elements of texts are to each other provides an opportunity for generating dictionaries quickly without the need to use an expert(s) to code some or all the text to create the dictionary. Not only can we assign words to dictionary categories, but we may be able to use the information in the complete text instead of focusing on a limited amount of information that is obtained from the terms used in a dictionary.

The process in the proposed methodology begins by obtaining a list of words that are informative about the research question that we are trying to answer. Frequently, this list of words and expressions will be determined by the theory guiding the research. After these words (seed words) are identified (denoted as set I containing the seed words), then vector-based similarities are used to create a weighted average encompassing all relevant information across all words that appear in each document with regards to the seed words. The steps in the process are detailed as follows:

1) Obtain an LSA representation of the raw data, the term document matrix, based on Singular Value Decomposition (SVD) of the original document term matrix. That is, compute Mp_s following equations (2) and (3).
2) Obtain the matrix of cosine measures of proximity, $\cos(\Theta)$, for each of the words with respect to the rest of the words in the corpus following equation (1)(as the number of

documents grows in the corpus, this matrix becomes very large as the dimension of this matrix is number of tokens by number of tokens). $\cos(\Theta)_{ij} = Mp_s \, Mp_s^t / | Mp_{s\,i} | * | Mp_{s\,j} |$

3) Select the set of words that are in the list, call this set I. Set $\cos(I) = \cos(\Theta)_{ij}$ with $i \in I$. Although, selection can be done based on multiple methods, there are two main philosophical avenues to follow:
    a. Theory-based approaches. Identify constructs that are predictive of the phenomenon of interest. In our empirical illustration, the valence of review can influence the perception of the moviegoers if the review is read prior to the movie attendance. Given this rationale, the overall valence of the review may be informative; thus, words that are expressive of the valence can be used to anchor the content measure of valence.
    b. Empirical-based methods use metrics derived from the actual text to determine the relevance of words. Several distinct approaches can be used to identify potential candidates if there is no a priori theory that dictates what constructs and words are of interest in the documents:
        i. One potential selection criteria could be to use the terms that have high scores in a *tf-idf* model, that is, these are words that are prevalent inside the document but relatively less common across all other documents in the corpus.
        ii. An alternative procedure is based on the idea of the amount of information (Shannon & Weaver, 1949). The information is quantified by the level of surprise of the marginal probability of the word in the document compared

with the marginal probability of the word across the entire corpus or a given set of documents; see Miller and Riechert (1994) for a similar strategy.

iii. In the case of lengthy individual texts, one can look at the distributions of the words within the document. In particular, the distance, measured as the number of words from a word to the next time the same word appears in the text, has properties that can be exploited to distinguish relevant or content-bearing words from words that function as links and structure (e.g., auxiliary verbs). Link words are randomly distributed across the text and will have an exponential distribution of those distance measures. Strong departures from this distribution imply that the words have potentially relevant meaning (Ortuño et al., 2002).

iv. Finally, words can be chosen such that if we construct a network composed of the textual elements in the document, the chosen word has a high degree of centrality in the network of words (see Corman, Kuhn, McPhee, & Dooley 2002 for a similar approach).

4) Use the matrix of selected word similarities to compute weighted linear composites of the initial metrics. That is, compute P such that P= Cos(I) Mp. P contains a set of linear composites that contain all the information in the text per the similarity exhibited by each of the words with the elements inside set I.

This proposed method is conceptually distinct from traditional content analysis and dictionary-based methods since this method shifts the underlying rationale from uniquely assigning words to word categories to creating linear combinations of word frequencies or marginal

probabilities. Therefore, the methodology moves from using part of the information in the term document matrix to using all the information available in the text. Also, the same word will have different weights in each of the latent vectors or weighted linear combinations that will be created to capture the concepts intended to be measured. The weights obtained based on the proximity of the tokens in the lower dimensional space of concepts could range from -1 to +1. A weight of -1 implies that the appearance of the particular term in the text is perfectly but inversely related to the concept that we want to measure using the linear composite. A weight of +1 implies that the word has literally the same meaning (pattern of co-occurrence across documents) in the lower dimensional space as the concept of interest. If a word has a weight close to zero, the word's presence or absence in the document does not provide meaningful information regarding the concept that is being measured via the linear composite.

CHAPTER FOUR: EMPIRICAL ILLUSTRATION—INFLUENCE OF MOVIE CRITICS ON MOVIE PERFORMANCE

I propose the use of an empirical application of the ALSA approach to show its potential use in marketing. In doing so, I re-examine the role that professional critics have on movie performance. I chose this area as an illustrative example for three reasons. First, the movie industry is particularly interesting from the product development and launch process perspective since product life cycles are extremely short (most movies do not stay in the theaters for more than eight weeks), making a decision to introduce a movie very important. This makes a priori forecasting of the success or failure of the motion picture critical. If we can use the information from the critical reviews prior to the launch, then better product introduction decisions can be made based on this information. Second, extracting information from movie reviews is a particularly challenging endeavor as critics in this product category oftentimes use sarcastic language and connotation. If we can show that the extraction of information from movie reviews is possible, this is evidence that this task will be simpler in other cases in which denotative use of the language is predominant. Third, this setting allows a test of external validity of the developed metrics, providing objective evidence of their soundness. By using the readily available metric of movie review content, i.e., star ratings, I can perform a predictive validity test to authenticate the text-based metrics that we have derived.

There is a large amount of empirical evidence that suggests that professional movie critics' reviews are related to box office (Jedidi, Krider, & Weinberg 1998; Litman, 1983; Litman & Ahn, 1998). Though there is evidence of a positive relationship, these studies did not analyze the mechanism by which critics' assessments correlate to box office figures. Eliashberg and Shugan (1997) were the first to propose two potential roles under which the reviews of critics and box office

are correlated. On one hand, critics may very well play an important role in moviegoers' decision making and hence be influencers. On the other hand, it is also plausible that critics are simply representative of their audiences and, thus, act as mere predictors without significantly shifting moviegoers' decision making. To disentangle these two competing explanations, the authors looked at relationships through the course of the movie's lifecycle, that is, they looked at the longitudinal and cross-sectional variations in box office. The authors found that movie critics mainly play the role of predictors within their respective markets.

In a recent study Basuroy, Chatterjee, and Ravid (2003) conducted a similar empirical study to assess the role of movie critics using time series cross-sectional regression of movie box office revenues on the number and ratio of positive and negative reviews. The authors studied the potential moderating effects of star power and budget on the relationship between movie critic reviews and box office revenues. This study shows results that are somewhat at odds with Eliashberg and Shugan (1997) because Basuroy and colleagues found some evidence that movie critics can be influencers, though their results are mixed on this issue.

However, these studies and others have used indirect measures of actual review content. In particular, many critics provide an overall rating of the movie, oftentimes in a 0 to 4- or 1- to 5-star scale. This rating or a reader's overall subjective judgment is used to classify the review as positive, negative, or mixed. I suggest that measuring the impact that critics could have on the moviegoers' experience is undermined by this oversimplification of the actual process. If the overall judgment or the rating were ultimately the only valuable source of information in the critics' reviews, we would be hard-pressed to find long and intricate movie reviews in the marketplace. Also, movie-going experiences, similar to many other hedonic product consumption experiences, are dependent on customer preferences, and those preferences vary greatly. If this is the case, then an overall

evaluation may not suffice, and the actual content delivered in the review may indeed be an important factor. I intend to show that content and structure of reviews can be directly quantified and used to assess the effect of critics on movie performance.

Measures of Movie Review Content

Based on the proposed Augmented Latest Semantic Analysis (ALSA) framework, it can be argued that movie critics possess a complex set of latent attitudes and affects toward the movie. This set of attitudes is reflected in the content (i.e., what they say) and structure (i.e., how they say it) used by the film critics in their reviews. Specifically, I am interested in assessing the overall attitude toward the movie as measured using the number of positive and negative comments in the reviews.

To obtain the weighted scoring of the distribution of words in each review to obtain valence scores for the reviews, I first created an ALSA-based reproduction of the document matrix, Mp_s, per (2) and (3). As a second step, I computed cosine measures of similarity for all pairs of words following (1) $Cos(\Theta)$[1]. Then I selected a set of words that captured the trait of interest. Given that the underlying measurement approach is similar to a semantic differential, I chose two extreme adjectives—one at each end of the construct. In this case, for the overall attitude toward the movie I selected the words "good" and "bad" to anchor this construct in the semantic space. I also selected another construct of interest to illustrate that several traits can be measured using this method. Movie enjoyment is an important correlation of overall attitude that may be able to explain success in movies that may not score high on overall quality. A movie may be enjoyable while having

[1] Because of constraints in computing power, I limit the words that are analyzed to those that appear at least in four (0.3%) of all the reviews in the first dataset. The number of words used for this analysis is in excess of 10,000. Note this procedure involves inverting and multiplying large matrices.

[2] Throughout the empirical analyses, I use word occurrence marginal probabilities instead of word occurrence frequencies. This choice is motivated by the need to disentangle some of the structural elements (such as length of message) from content-related effects.

average acting, directing, and special effects, and hence, may succeed. To capture this construct, I selected the adjectives "enjoyable" and "dull" as anchoring scales for the construct in latent semantic space. From the similarity matrix, Cos(Θ), only the rows containing these four words comprise set I. This portion of the matrix is referred to as Cos(I). After selecting the words I created scores for each of the words by multiplying the distances by the Mp matrix, $P= Cos(I)\ Mp_s^2$. P contains the scores for each of the anchors of the two constructs of interest, overall movie attitude, and movie enjoyment.

Comparison with Alternative Textual Data Approaches

I used two traditional computer-assisted quantitative content analysis approaches to the aforementioned empirical issue in order to compare the efficacy of the ALSA approach in quantifying textual content.

The first approach is labeled ad hoc dictionary because I created a new dictionary for each concept based on the context studied. The creation of the ad hoc dictionary can be explained with an example for coding review content. Words in the movie reviews can be classified into words that have a positive connotation, a negative connotation, or a neutral connotation. To create an ad hoc dictionary, I started with some simple adjectives, such as good and bad, that are commonly used to make evaluative judgments about a movie. I used the Microsoft synonym feature in Microsoft Word to find words that are similar to these two seed words. I repeated this step using the newly found synonyms as the new seed words. Once the set of synonyms was exhausted, the process was stopped. Table 1 contains the words that possess positive and negative connotations that were obtained using this strategy.

Table 1
Valance Based Ad Hoc Dictionary

Positive valence	Negative valence
Amusing	Absurd
Best	Annoying
Brilliant	Awful
Convincing	Bad
Enjoy	Badly
Enjoyable	Dire
Enjoyed	Dreadful
Excellent	Hate
Fantastic	Hideous
Favorite	Hopeless
Fine	Horrible
Fun	Horrific
Funny	Inadequate
Great	Outrageous
Greatest	Painful
Hilarious	Pitiable
Humorous	Poor
Interesting	Ridiculous
Like	Silly
Love	Stupid
Memorable	Terrible
Outstanding	Unfortunate
Perfect	Unpleasant
Pretty	Useless
	Worst
	Worthless
	Wrong

To determine which words should be assigned to each set, the following rule of assignment was used. If the word was listed as having positive valence in Table 1, then the word was assigned to the positive set. Alternatively, if the word was listed as negative valence in Table 1, then the word was assigned to the negative set. The remaining words were assigned to the neutral set. Based on these rules, I created frequency counts for each review. The values obtained are the marginal

distribution of frequencies f(wi) of each word within the text. Based on this marginal distribution of the frequency of each word in the texts, one can use the single words as distinct entities (sets of one element) or group the tokens or words in sets that have common meaning and or behavior (e.g., positive words ={good, fantastic, amusing…}).

The second approach for comparison is called the general purpose dictionary. The Harvard-IV and Lasswell are two general purpose dictionaries that are used by the General Inquirer for the analysis of texts (see http://www.wjh.harvard.edu/~inquirer). These dictionaries have more than 10,000 different words that are classified into multiple categories. The rule assignment is made based on the similar rationale to the ad hoc dictionary approach. For example, to code review content, I used the Positive and Negative classifications provided by the General Inquirer. These dictionaries have advantages and disadvantages over ad hoc or empirically derived dictionaries. These general purpose dictionaries typically have been validated in other empirical settings (e.g., Holsti 1964) and hence may make the researcher's task simpler in terms of generating the sets of rules. However, these general purpose dictionaries may not be well suited as measures of the concepts of interest in the research study.

Given that multiple sets of variables[2] are involved, three different operationalizations of the content-related variables for the ad hoc and general purpose dictionary were considered. The constructs were initially operationalized by entering the marginal probabilities associated with each word in the dictionary as an independent variable and hence estimating individual beta weights for each term. I call this individual term formulation. Because there were numerous words, the list of independent variables grew quickly, potentially making the estimation of the model infeasible (not enough data points).

[2] In both dictionaries, we observe marginal probabilities and counts of different terms for each of the constructs.

The second operationalization tackled this limitation by creating a summated scale formulation in which all scale frequencies related to one construct were summed. This approach has been used by marketers who have employed automated content analysis (e.g., Rosa et al. 1999). The summated scale approach, while successfully addressing the overparameterization alluded to earlier, has its own limitations. First, all words are forced to have the same weight and contribution in measuring the construct of interest. In other words, according to this operationalization, the words "good" and "great" have the same effect in determining the valence of the review. A secondary effect of this operationalization is that the reduction of variables comes at the expense of variance explained when compared to the individual term approach.

The third approach considered is a latent variable approach, which is a significant departure from the extant research in this area. My theory suggests that critics develop a complex set of attitudes about the movie and some of its components (Was the casting adequate? What is the level of acting? Are the special effects realistic?) and that the choice of words in the reviews best reflects the attitudes that the critics have formed during and after the movie experience. This implies that scaled frequencies of words are mere indicators of latent attitudes that have developed in the critic's mind.

The latent variable model requires the specification of a measurement model that links the observed scaled frequencies to the latent constructs of interest. There are two distinct types of measurement models that have very different natures and implications: reflective and formative (Fornell & Bookstein, 1982; Bollen & Lennox, 1991). The most widely used measurement model, reflective, stems from classical true score theory that postulates that items (measures) are created as a composite of true score and error, which is later decomposed into systematic and random components. Under this formulation, a variation on the true score or the trait that we are interested

in measuring will translate into a change in the item score. That is, if multiple items or measures of the same construct are used, they will covary to the extent that the true score of the construct changes; this is the typical common factor model.

However, some measures do not exhibit this behavior and therefore should not be modeled as reflective. In particular, formative measures are those that compose or create the construct as a weighted linear combination of the items plus some error component. Note that the causal chain is reversed, and that, while in reflective measures the construct is what causes the change in the item, in the case of formative measures, it is the measure that causes the construct to change (see Jarvis, Mackenzie, and Podsakoff (2003) for an excellent discussion on the differences between the two types of measures). It is important to realize that word frequencies (or scaled word frequencies) are formative indicators of the attitude toward the movie. Formative measurement models are appropriate when the different variables compose or create the construct.

There are two reasons why a formative model seems more adequate in this circumstance. As previously mentioned, the choice of words indicates the state of the attitude of the reviewer; however, space and time for communication are limited, and the number of potential words and expressions that a reviewer can use to express a particular attitude is large. Therefore, chances are that that once one of the words or expressions is used, the mere usage will preclude the critic from using many words and expressions available in his or her vocabulary. A second reason to use formative models is that ultimately we are interested in the effects that the content of the review has (if any) on the movie-going behavior of consumers. If this is not the case, then it is not necessarily the attitude that the critic intends to convey with the review but the attitude that is inferred from the review by the consumers (readers) that is of interest. If this is the case, and noting that the attitude is composed by processing the words and phrases in the review, a formative measurement model is

most appropriate, that is, words shape the consumer's attitude and not the other way around. This explains that, especially at the individual review level,[3] a formative measurement model is more appropriate.[4] Following the guidelines provided by Jarvis, Mackenzie, and Podsakoff (2003), I modeled the measures as formative.

Table 2
Operationalizations of Review Content

Approach	'Content' operationalization
Ad hoc dictionary	1. Individual term 2. Summated scale 3. Latent Variable
General purpose dictionary	1. Summated scale 2. Latent variable
ASLA	1. Continuous weights

Recall that I intend to test three textual data approaches—ad hoc, general purpose dictionary, and ALSA. The ALSA approach uses continuous weights. The ad hoc dictionary uses three independent variable operationalizations, as described previously, and the general purpose dictionary approach uses two, as the individual term approach is not feasible given the large number of terms in the general purpose categories (Table 2).

[3] Note that as reviews are aggregated to the movie level the originally formative nature is diluted as vocabulary choice becomes less important since more and more reviews are averaged out. So as the number of reviewers that is aggregated grows, a reflective model will fit the observed data better.

[4] Another potential way of explaining this is from the communication side. There are two potential effects of reviews: prediction of success and influence. If we are interested in the second effect, then it is not per se the attitude that the reviewer has that we are concerned with but the attitude that is communicated though the review. The communicated attitude is composed of the words used in the review, and, therefore, words are clearly formative in nature.

Overall, I posit that, regardless of the method used, the metrics measuring positive comments should exhibit a positive effect on the a priori movie evaluations and thus positively affect movie performance by encouraging movie attendance. Similarly, the metrics measuring negative comments and judgments by reviewers should have the opposite effect on movie performance. Finally, I also suggest that valence comments will have diminishing marginal effects on movie performance as hearing the same (negative) message 10 times probably will not detract twice as many people from going to the theater as hearing it just 5 times.

The Effect of Structure: Measures of Length and Complexity

In many circumstances, both as sources and as receivers in the communication process, human beings express their opinions, attitudes, and emotions using both stated messages (content) and more subtle nonexplicit cues to transmit these basic attitudes. Similar to nonverbal cues in nonwritten communications, the structure of the message can convey much about the attitude of the writer. This duality, content versus structure, provides the possibility of extracting relevant information from the explicit concepts that are transmitted in the text (the content) as well as from the way the text is written (the structure of the text). I consider two such measures of structure—length and complexity of the message.

Measure of review length

Past research has shown that length of the communication is affected by the attitude of the composer toward the object that is being described. This effect is clearly visible when the sender of the message finds the content that he or she can use in the message limited by explicit rules or societal norms. For example, research conducted in evaluation of letters of recommendation finds

that length is a good predictor of attitude. Given that it is usually not acceptable to write a negative letter of recommendation, the recommender is relatively limited in his or her capacity to convey his or her judgment or attitude regarding the recomendee. Mehrabian (1965) found that recommendation letter writers wrote longer letters for subjects for whom the letter writers had a more positive attitude. Wiens, Jackson, Manaugh, and Matarazzo (1969) replicated this basic result in a similar setting.

Interestingly, receivers also use length as a cue. Past research has shown that evaluators who are given longer letters of recommendation containing similar factual content tend to evaluate candidates more positively than candidates with shorter recommendation letters. In an experimental setting, Kleinke (1978) showed that longer letters were deemed more favorable than shorter letters. He also found that length played a more salient role in evaluations when less information was available to the receiver of the message. This effect is consistent with the argument that cues are used more heavily when other more direct information, say content, is scarce. This finding is consistent with the signaling literature (see Kirmani & Rao, 2000, for a review of the signaling literature), which suggests that cues are used when direct knowledge or information is lacking. In summary, I expect that the length of the message will have a positive effect on the attitudes of the receiver of the message. This effect would be stronger for complex messages. I operationalize the length of the message as the number of tokens in the message.

Measure of review complexity

Complexity of the message is potentially an important factor in the communications process. Research in human information processing has suggested that humans are limited in the amount of information that they can process (cf., Miller, 1956). This limit in processing ability will entail that

there is an optimal level of information in stimulus and that levels below and above it will yield lower levels of affect. In other words, intermediate levels of complexity are optimal with high levels of effect and both low and high levels of complexity showing lower effect and preference levels. Based on this literature, I posit that complexity will have a nonlinear effect on positive effect or liking such that intermediate levels of complexity are optimal.

The complexity of a textual message can be divided into two distinct components: a) lexical complexity of the message and b) syntactical complexity of the message. While both affect the overall complexity of the message, they are distinct in their nature. Lexical complexity is a measure of the level of vocabulary that is used by the writer whereas syntactical complexity relates to how the words are interlinked.

I calculated lexical complexity using the Type-Token Ratio (TTR) (Johnson, 1944) which is computed by dividing the number of different words (types) by the total number of words (tokens) found in the text. Because this measure is not independent from the length of the message, I computed standardized TTR (STTRx) for each document. The value of TTR was computed for x number of words within the document, where x is constant across all messages that are studied. After the TTRs were computed for the given window (100 words in this case), an average across occasions became STTRx. This measure used all possible information in the message and corrected for the relationship of TTR to message length.

I computed the entropy of word distribution as another measure of message complexity. This measure was computed as $H=-k \sum p_i \log p_i$ for all i. Note that this measure of message complexity is solely based on the structure of the words, such as STTR, and not based on the function that words fulfilled within the communication process.

Finally, I used lexical density to measure the ratio of content-bearing words to total number of words. Psycholinguistic studies have shown (see for example, Perfetti, 1969) that there is a correlation between lexical density and sentence comprehension. Sentences with a high lexical density are more difficult to absorb and so controlling the lexical density of a text is one way of helping less able readers (Bradac et al., 1977). To compute lexical density, I used the tags existing in the general purpose dictionary built in by General Inquirer and classified words as content bearing or not and created a simple ratio.

Data

For this study, two main movie review datasets were used. The first dataset, the calibration set from here on, was collected from the Internet Movie Data Base (IMDB) and was composed of approximately 1,400 reviews. These reviews represented a wide spectrum of movies evaluated mainly from the years 1997 to 2001. Detailed reviewer level data were collected for this database. The data include overall rating, scale used for the overall rating, movie that was reviewed, reviewer, and a complete breakdown of the documents into a document term matrix, M, as described in the previous sections. This dataset was used for validity checks as well as a calibration for the ALSA-based content analytic procedure.

The second main dataset, the validation set, is composed of 242 movies that were aired in the mid-90s and all reviews available in the Movie Review Query Engine for each of those movies. This set totals more than 8,000 reviews nested within the 242 movies. For this dataset information on the type of movie genre, the NPAA rating, whether the movie was a sequel, the amount spent in advertising, the maximum number of screens, number of screens at opening, and the budget of the

movie were also collected. This dataset was used to assess the effects of critics on movie performance.

The Internet has played a critical role in facilitating the efficient collection of information. While other specific software such as a crawler can be used to collect the data, mainstream software such as Adobe Acrobat, with its open website function, allows users to collect information from Internet websites efficiently. Users may also save the collected information in Rich Text Format, a format that some content analysis programs can read. Acrobat allows the user to specify a website and then determine how many levels down on the tree of the website he or she wants the content to be collected. Adobe collects the information from each site and every link into it. The user may also constrain the content to the same path or server in order to avoid the collection of content that is not relevant to the study. The validation dataset was collected using this procedure. The Movie Review Query Engine was used for two reasons: it contains a broad database of professional reviews, and its format is particularly appealing to the collection of reviews using Adobe Acrobat.

To compute the document term matrix, M, software that breaks down the documents into word lists was needed. There are a number of choices that could efficiently complete this task. Wordsmith 4.0 was used in this study, but other software packages such as VBPro could be used. Most of this software is easy to use and allows the user to conduct analyses and organize the documents in the database. After the frequencies of words in each document were collected, they were assembled into a term document matrix using IML in the SAS v. 8.02 environment. Two matrices, M, containing the term frequency data $f(w_{ij})$, and Mp, containing empirical marginal probabilities $p(w_{ij})$ for each term in each document, were assembled. To compute Mp, the values in M were divided by the total number of tokens in each document.

Overall Analysis Strategy

I used a two-stage approach in testing the effect of reviews on movie performance. In the first step, I tested the predictive validity of the content-derived metrics using the proposed operationalizations. After the validity test was satisfied, I tested whether there was additional information to predict movie performance on the reviews above and beyond the effect of movie ratings. To accomplish this task, I used two different datasets.

First, I analyzed the calibration dataset. This dataset contains individual-level reviews with information regarding the movies and reviewers who wrote the reviews. The cross-sectional nested panel nature of this dataset provided the required information to account for unobservable factors pertaining to both reviewers and movies that may bias the relationships.

To test the direct effect of direct content and structure on reviewer ratings, I used Linear Mixed Models (Bryk & Raudenbush, 1992; Goldstein, 1987) to account for the clustered nature of the data. Reviews are not independent of one another since the same reviewer reviews multiple movies, and similarly, the same movie has multiple reviews. I modeled the unobserved heterogeneity via a random effects formulation. This enabled controlling for reviewer-specific idiosyncrasies (e.g., reviewer style) and movie unique characteristics that potentially could contaminate the testing of the effects of interests. In this case, the general mixed model has the following form:

$$y_{ij} = X_{ij}\beta + Z_i b_i + W_j c_j + \varepsilon_{ij} \quad (4)$$

$$bi \sim N_q(0,\Psi) \quad (5)$$

$$c_j \sim N_r(0,\Xi) \quad (6)$$

$$\varepsilon_{ij} \sim N_N(0,\sigma^2\Lambda_{ij}) \quad (7)$$

where y_{ij} is the dependent variable with the observation of the i^{th} reviewer for the j^{th} movie. X_{ij} is the $N \times p$ model matrix corresponding to the fixed effects. β is the $p \times 1$ vector of fixed-effect

coefficients. Z_i is the N × q model matrix for the random effects for observations in group i. b_i is the q × 1 vector of random-effect coefficients for group i. W_j is the N × r model matrix for the random effects for observations in group j. c_j is the q × 1 vector of random-effect coefficients for group j. ε_{ij} is the N × 1 vector of errors for observations in group ij. Ψ is the q × q covariance matrix for the reviewer random effects. Ξ is the r × r covariance matrix for the movie random effects, and finally, $\sigma^2\Lambda_{ij}$ is the N × N covariance matrix for the errors.

For both the ad hoc and general purpose dictionaries, I specified content, as shown in Table 2. While the individual term and summated scale operationalization of the content in the reviews was straightforward, the latent variable approach needed clarification. To obtain latent scores from the formative terms, Partial Least Squares (PLS) was employed. In PLS, the dependent variable is used as part of the optimization procedure to determine the word weights that form the underlying latent constructs. I determined two latent constructs, positive attitude and negative attitude, assigning the words for each of these categories, as shown in Table 2. I then obtained two sets of PLS weight-based scores for each method, one for positive comments and one for negative comments. I used these two sets of latent scores as independent variables to model critic ratings in a PLS framework.

Second, after testing the validity of the content metrics was completed, the analysis shifted to the effect of the content of the reviews on movie performance. Given that a primary interest was in analyzing the effects of reviews on movie performance, the unit of analysis changed from the review to the movie level. Since the dependent variable was at the movie level, I aggregated the frequencies and scaled frequencies (probabilities) for each set of movies.

Movie performance is operationalized in three different ways: box office revenues, gross profits, and return on budget. I used a set of controls for spurious variance on the movie

performance measures. Specifically, I included advertising/media spending (Media), number of screens (Screens), dummy variables to account for the movie being a sequel (sequel), genre (dummies for family, action, drama, comedy, and thriller), and a dummy variable for whether the movie is classified by the MPAA as R or not. Since this involved an estimation of aggregate-level models, control of heterogeneity was not possible. Thus, the models can be written as the following:

$$P = \beta_w g(W) + \beta_x X + e \quad (8)$$

where P is a vector of a movie performance measure (e.g., box office revenues, gross profits…), β_w is a vector of parameters that correspond to the word scaled frequencies, g(W) is the matrix of the scaled frequencies of the relevant words, β_x is a vector of parameters for the covariates, X is a matrix of covariates, and e is a vector of errors. Similarly, R is vector a movie ratings, β_w' is a vector of parameters that correspond to the word scaled frequencies, $g(W^*)$ is the matrix of the frequencies of the relevant words, β_x' is a vector of parameters for the covariates, X^* is a matrix of covariates, and u is a vector of errors. The parameters in (9) are estimated using Ordinary Least Squares (OLS).

Validity Test : Predicting Rating with Content and Structure

Since the methods employed in this research are novel, the starting point was providing evidence of the predictive validity of the content measures proposed in the previous sections. The basic idea for this test is simple: if the metrics created from the raw text can predict the actual rating provided by the reviewers, then we have a valid measure of content. This test was conducted using both datasets.

I began by analyzing the calibration dataset. This dataset provided a particularly good testing scenario since there are individual reviews with information regarding the movie and reviewer who wrote it. The cross-sectional nested panel nature of this dataset provided the required information to

account for both reviewer and movie unobservable factors that may bias the relationships between the metrics and the actual ratings of the reviewers.

The essential testing strategy was relatively straightforward. Given that we have information regarding the overall attitude of the reviewer about the movie and the movie rating (Srating[5]), I assessed whether the content- and structure-based measures predicted attitude. To do so, I used linear mixed models to test the potential effect of direct content and structure on reviewer ratings (Bryk & Raudenbush, 1992; Goldstein, 1987). The rationale for using mixed models is that they allow modeling of the clustered nature of the data. The reviews were not independent of each other as the same reviewer reviewed multiple movies and the same movie had multiple reviews. The lack of independence can be understood as a consequence of idiosyncrasies that are unique to each of the grouping elements that are not explicitly measured. These idiosyncrasies are usually referred to as unobserved heterogeneity, as oftentimes they cannot be measured even if the researcher attempts to do so. In the literature, authors have warned against the potential biases that occur when these effects are present unless the heterogeneity is modeled explicitly (e.g., Hutchinson, Kamakura, & Lynch, 2000). We model these idiosyncrasies via random and fixed effects.

Ad hoc dictionary

In the particular implementation of this dataset, the general model presented in equations (4) to (7) takes a simpler form. In this case, I specified random effects for both movies and reviewers.[6] The initial model for the ad hoc dictionary is given by the following:

[5] Note that the ratings are rescaled so that they are all expressed in a 0 to 4 scale.
[6] The fit of a series of combination of fixed, random, and mixed models was tested. The all random effects model was found to be superior to the rest using fit and information criteria measures.

$$Sratings_{ij} = X_{ij}\beta + u_i + v_j + \varepsilon_{ij} \quad (9)$$

$$u_i \sim N(0,\Psi) \quad (10)$$

$$b_j \sim N(0,\Xi) \quad (11)$$

$$\varepsilon_{ij} \sim N_N(0,\sigma^2\Lambda_{ij}) \quad (12)$$

Vector β contains the fixed effects that assessed the effect that content and structure have on ratings. Only random intercepts were estimated in the models.[7] Note that in this type of model the heterogeneity in each of the two dimensions is assumed to be independent of the other random variables (this includes also the "usual" individual specific error term ε_{ij}).

One important question that we need to address is how to best operationalize the two sets of words or variables: positive and negative. As a first step, I specified a model in which the frequencies for all the words were entered into the equation as an individual variable, and a separate beta weight was estimated for each of the words. This operationalization represents the belief that each word is distinctly important and will determine the overall attitude capture by the rating differently. In this first model, X_{ij} contains the marginal probability of observing each of the words in Table 1. The maximum likelihood (ML) estimates for the parameters in this model are presented in Table 3. Table 3 also reports a similar model (same independent variables) in which the clustering of the variables is ignored and simple ordinary least squares (OLS) regression is used to estimate the parameters in β.

As can be seen from the results, there is evidence that the content in the reviews as measured by this set of variables is significantly related to the reviewer rating. This is an initial check that provides evidence of predictive validity for the ad hoc dictionary as a measure of content of the

[7] I tried different specifications including random slopes and found, in general, no evidence that the effects change either across movies or reviewers, and overall, the substantive results of the models were invariant to changes in the random effects structure.

review. Note that the clustering effect is particularly strong for the unobservables at the movie level, and that it is also statistically significant although substantively weaker at the reviewer level.[8] This indicates that it will be no surprise that the ratings are more similar within movie than they are across movies. This also indicates that the same reviewer tends to give similar ratings across movies above and beyond what would be expected from the movie itself. This latter effect, however, is relatively small when compared to the movie clustering. Nevertheless, it is still significant and important to model it to obtain consistent fixed effects in the model, β, and correct standard errors.

[8] The intraclass correlation measured as the ratio between the within group variance and the total variance is .05 for reviewer and .40 for movie.

Table 3
Mixed Model and OLS Regression for Ad hoc Dictionary (Predicting Ratings Using Individual Word's Marginal Probability)

	MIXED		OLS	
Effect	Estimate	Standard error	Estimate	Standard error
Intercept	2.190 ***	0.054	2.15323	0.05111
Tokens	-	-	-	-
Entropy	-	-	-	-
Lexicaldensity	-	-	-	-
STTR100	-	-	-	-
Absurd	-24.303	103.580	-59.8003	113.37126
Aggravating	-115.870	331.890	-239.90338	386.05483
Annoying	4.675	40.903	-12.15396	43.84463
Awful	-255.050 ***	51.778	-288.49661 ***	55.81479
Badly	-126.960 **	53.033	-152.21626 ***	56.16244
Dire	-73.274	104.810	-151.68261	116.55742
Disgusting	-23.809	76.788	-1.40872	87.53276
Dislike	72.343	154.260	171.80808	169.7438
Dismal	-56.910	109.790	-32.07071	125.95933
Dreadful	-230.270 **	99.766	-153.1447	109.81946
Dull	-178.480 ***	36.136	-225.05311 ***	39.78827
Exasperating	-474.840	323.880	-590.08569 *	356.54507
Frustrating	-119.010	115.890	-172.38051	130.71158
Grim	37.618	128.530	105.68955	138.73018
Hate	-24.554	45.894	-18.32995	49.18769
Hideous	31.179	43.903	36.66829	45.9124
Hopeless	-167.190	109.930	-208.53497 *	122.92935
Horrendous	56.744	101.250	5.40147	113.99696
Horrible	-166.320 ***	53.970	-177.05434 ***	59.32718
Horrific	192.330	134.630	261.73884 *	153.66929
Inadequate	-202.330	316.380	-543.11221	343.66363
Irritating	-149.870 **	69.966	-147.49007 *	76.39484
Meaningless	-252.140 **	125.430	-423.25007 ***	144.85692
Ominous	201.990 *	120.500	193.43979	133.46063
Outrageous	-130.000 *	67.350	-126.76344 *	72.4424
Painful	-88.118	70.626	-86.22727	76.34591
Pathetic	-140.800 ***	47.314	-206.72216 ***	53.04774
Poor	-10.226	41.313	-10.21229	45.16942
Ridiculous	-180.630 ***	49.391	-213.36345 ***	55.39306
Silly	-48.468	41.567	-48.71125	46.7913
Stupid	-160.670 ***	30.149	-159.13712 ***	32.63242
Sucks	-16.530	107.860	-55.13634	119.21097
Terrible	-135.750 ***	41.913	-153.83725 ***	45.74059
Unfortunate	-30.375	80.739	19.18847	85.80919
Unpleasant	-0.803	81.187	-43.43951	87.10259
Useless	52.497	86.216	-66.71386	96.26541
Worst	-121.320 ***	25.084	-145.90884 ***	26.83031
Worthless	-127.480	103.820	-90.05639	116.54929
Wrong	23.308	27.579	34.25246	30.83258
Amazing	61.389	39.903	69.7798	42.99592
Amusing	8.594	39.822	-7.50814	43.14683

	MIXED				OLS		
Effect	Estimate		Standard error		Estimate		Standard error
Best	51.853	***	15.219		56.6904	***	16.57251
Brilliant	194.090	***	50.019		232.98505	***	56.07504
Convincing	52.790		47.145		42.44113		52.62533
Dazzling	323.410	*	175.890		411.94389	**	196.94166
Enjoy	-2.533		37.555		14.06901		40.65541
Enjoyable	90.747	**	39.548		82.44474	*	43.37951
Enjoyed	95.069	*	51.051		82.95801		57.15141
Excellent	77.624	**	31.922		87.54073	**	34.12853
Exceptional	182.080		121.860		189.25055		127.92366
Extraordinary	187.050	**	73.704		248.0064	***	85.10409
Fantastic	135.210	**	63.347		162.62103	**	70.99815
Favorite	-104.170	**	51.480		-89.36579		56.25079
Finest	45.007		76.843		138.60423	*	83.39319
Fun	47.896	**	20.115		58.77334	***	22.33942
Gorgeous	215.730	**	93.763		240.18112	**	102.69075
Great	79.344	***	14.387		87.04196	***	15.72927
Greatest	70.552		42.828		121.12013	**	47.10902
Incredible	2.651		62.483		24.21015		70.31472
Interesting	-38.959	*	20.667		-45.92987	**	22.33748
Joyful	396.710		459.680		274.12095		535.64441
Like	-8.223		8.613		-13.4863		9.18952
Love	24.330	*	13.601		17.50191		14.11259
Marvelous	81.027		119.010		122.25306		132.36336
Memorable	137.450	***	46.479		125.51022	**	50.53833
Outstanding	207.350	***	71.965		271.68382	***	77.52489
Perfect	72.811	***	26.355		102.57671	***	29.12531
Pretty	-32.509	*	19.710		-38.82888	*	21.3619
Remarkable	143.960	**	68.930		146.46541	**	74.3507
Splendid	125.520		230.570		200.10486		251.59172
Superb	39.902		52.590		88.46689		57.38334
Terrific	127.200	**	59.578		158.21343	**	66.25563
Tremendous	-5.238		110.170		-17.39831		120.70076
Wonderful	75.126	**	38.135		98.16894	**	42.09934
- 2Log Likelihood	3222.9			R^2	0.3118		
LR null model	119.08	***		$AdjR^2$	0.2724		
Var(Ui)	0.031	***	0.011	F	7.91	***	
Var(Vi)	0.272	***	0.034				
Var(Ei)	0.3791	***					
AIC	3378.9						

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

I also estimated a model that adds to the previous model the variables related to the structure of the message (how the message is communicated).The results for this second set of estimates are given in Table 4.

Given that the random structure is common across both mixed models, the fixed parameters are nested, and ML was used for the estimation, a likelihood ratio (LR) test[9] is appropriate to determine whether the inclusion of this set of variables significantly improves fit. In this case, the test statistic LR=50.3 is significant at 1% when compared to a chi square with four degrees of freedom (df). This implies that the model fit improves beyond chance when we include structure, and, thus, there is also an effect of how the message is delivered on ratings. In particular, the estimates suggest that the longer and the less complex the review, the more positive its rating.

A careful examination of the coefficients and their respective standard errors in the model shows several important issues. First, while most signs are in the expected direction, there are some that are opposite to what is expected a priori (e.g., FAVORITE) by the classification provided in Table 1 (i.e., negative coefficient for negative words and positive coefficient for positive words). Another pattern that arises is that many of the coefficients in the model are not significant, i.e., the ratio of the coefficient to the standard error is not large. Both of these issues could be explained by multicollinearity and overlap in variance among the variables (all words within a category are after all measuring the same thing!).

[9] The test statistic is constructed as $LR=2LL_{unconstrained}-2LL_{constrained}$ in which LL is the likelihood of each of the models at the optimum value for the set of parameters estimated under each model.

Table 4
Mixed Model and OLS Regression for Ad Hoc Dictionary (Predicting Ratings Using Individual Words and Structure)

	MIXED			OLS		
Effect	Estimate		Standard error	Estimate		Standard error
Intercept	5.063	***	0.808	5.354	***	0.817
Tokens	0.026		0.016	0.042	**	0.017
Entropy	-0.087		0.108	-0.001		0.117
Lexicaldensity	-1.151		0.831	-2.063	**	0.862
STTR100	-0.041	***	0.009	-0.032	***	0.010
Absurd	-45.127		101.920	-84.716		111.255
Aggravating	-286.360		328.450	-402.848		379.191
Annoying	20.690		40.412	2.413		43.133
Awful	-252.330	***	50.940	-276.018	***	54.783
Badly	-131.460	**	52.097	-156.090	***	55.157
Dire	-62.418		103.230	-127.859		114.475
Disgusting	-2.712		75.991	26.698		86.153
Dislike	30.885		152.020	126.229		166.585
Dismal	-41.561		108.170	-29.553		123.494
Dreadful	-197.930	**	98.191	-127.161		107.813
Dull	-164.350	***	35.656	-207.359	***	39.116
Exasperating	-518.680		318.770	-602.362	*	349.886
Frustrating	-111.190		114.150	-155.356		128.186
Grim	47.289		126.360	128.264		136.045
Hate	-28.138		45.148	-21.382		48.240
Hideous	31.925		43.068	41.180		45.031
Hopeless	-127.370		108.460	-164.740		120.752
Horrendous	68.276		99.758	2.426		111.817
Horrible	-172.470	***	53.133	-178.133	***	58.183
Horrific	178.930		132.790	219.673		151.042
Inadequate	-224.100		310.950	-537.905		336.942
Irritating	-154.520	**	68.802	-137.461	*	74.926
Meaningless	-246.080	**	123.800	-412.028	***	142.118
Ominous	192.190		118.550	170.137		130.905
Outrageous	-122.370	*	66.198	-111.984		71.066
Painful	-93.522		69.487	-84.344		74.895
Pathetic	-127.160	***	46.710	-182.543	***	52.163
Poor	-23.681		40.747	-30.837		44.419
Ridiculous	-167.500	***	48.739	-203.015	***	54.376
Silly	-43.728		40.923	-35.917		45.927
Stupid	-154.920	***	29.707	-153.162	***	32.044
Sucks	-25.635		106.440	-72.845		117.412
Terrible	-130.830	***	41.264	-139.974	***	44.942
Unfortunate	-27.647		79.270	13.585		84.122
Unpleasant	12.371		79.843	-31.804		85.519
Useless	58.665		84.850	-43.981		94.515
Worst	-114.080	***	24.685	-133.640	***	26.389
Worthless	-107.720		102.400	-92.649		114.443
Wrong	23.040		27.163	36.284		30.257
Amazing	62.330		39.283	75.754	*	42.221
Amusing	7.846		39.240	-11.025		42.439

	MIXED				OLS		
Effect	Estimate		Standard error	Effect	Estimate		Standard error
Best	49.254	***	14.973		55.004	***	16.253
Brilliant	177.450	***	49.303		206.488	***	55.097
Convincing	53.930		46.431		44.824		51.622
Dazzling	305.850	*	173.190		406.212	**	193.379
Enjoy	-4.126		36.941		14.660		39.891
Enjoyable	94.643	**	38.930		93.611	**	42.614
Enjoyed	97.436	*	50.283		79.112		56.044
Excellent	78.204	**	31.422		89.713	***	33.607
Exceptional	177.570		119.570		182.517		125.501
Extraordinary	205.130	***	72.662		260.174	***	83.474
Fantastic	140.660	**	62.408		169.534	**	69.724
Favorite	-116.650	**	50.679		-100.198	*	55.175
Finest	58.004		75.672		144.248	*	81.868
Fun	50.069	**	19.866		60.564	***	21.989
Gorgeous	197.750	**	92.401		221.648	**	100.869
Great	71.671	***	14.196		75.463	***	15.510
Greatest	68.126		42.197		117.571	**	46.255
Incredible	-16.249		61.610		1.841		69.150
Interesting	-42.251	**	20.338		-54.870	**	21.956
Joyful	362.970		453.300		182.780		525.544
Like	-9.807		8.478		-16.808	*	9.046
Love	27.163	**	13.413		18.575		13.922
Marvelous	30.046		117.410		59.271		130.135
Memorable	135.250	***	45.712		125.302	**	49.594
Outstanding	219.630	***	71.058		282.757	***	76.206
Perfect	78.685	***	25.985		110.035	***	28.608
Pretty	-32.993	*	19.388		-39.001	*	20.966
Remarkable	168.940	**	68.154		150.299	**	73.450
Splendid	138.660		226.840		212.816		247.062
Superb	44.600		51.717		95.726	*	56.293
Terrific	119.990	**	58.742		144.916	**	65.033
Tremendous	22.941		108.550		-14.789		118.449
Wonderful	74.981	**	37.572		97.068	**	41.337
- 2Log Likelihood	3172.6			R^2	0.3408		
LR null model	3336.6	***		$AdjR^2$	0.3009		
Var(Ui)	0.027	***	0.01031	F	8.54	***	
Var(Vi)	0.255	***	0.03212				
Var(Ei)	0.3714	***	0.02463				
AIC	3378.9						

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

There are different ways of addressing this issue. A simple approach is to obtain frequencies or probabilities for each of the sets previously proposed, that is, positive and negative. Following this rationale, the second operationalization consists of an index created by adding all the positive

variables in one case and all the negative variables in the other. I labeled these two variables PositiveAdhocSum and NegativeAdhocSum such that,

$$\text{PositiveAdhocSum} = \sum_{j \in S_1} \text{prob}(w_j) \tag{13}$$

$$\text{NegativeAdhocSum} = \sum_{j \in S_2} \text{prob}(w_j) \tag{14}$$

where S_1 and S_2 correspond to the Positive and Negative sets of words given in Table 1.

Following this operationalization of the ad hoc dictionary, the model was re-estimated. The results of this model are given in Table 5. Note how the problem regarding the signs and significance of the individual word probabilities has been successfully resolved. In this case, all variables measuring content are significant at 1%, and their parameters have the theoretically expected sign. Note also that, as before, there is a large amount of unobserved heterogeneity, particularly across movies, but also across reviewers.

It is interesting to note that the most important variable predicting review ratings is the one that accounts for negative comments. This seems to indicate that while positive words are also related to ratings, the negative words are most reflected in the rating generated by the critic.

Table 5 also shows the additional effect that structure has on ratings. The results in this case are similar to the previous case. The addition of this variable is also significant, LR(4)=46,8 p<1%. The effect of length of the message is positive and the effect of complexity is negative.

Table 5
Mixed Model and OLS Regression for Ad Hoc Dictionary (Predicting Ratings Using Summated Scales and Structure)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	2.1633	***	0.05695		2.11864	***	0.05273
Tokens	-		-		-		-
Entropy	-		-		-		-
Lexicaldensity	-		-		-		-
STTR100	-		-		-		-
Positiveadhocsum	35.3608	***	4.4162		40.92842	***	4.63482
Negativeadhocsum	-90.6266	***	8.2296		-112.68763	***	8.71878
- 2Log Likelihood	3441.9			R^2	0.186		
LR null model	185.53	***		AdjR^2	0.183		
Var(Ui)	0.037	***	0.01166	F	51.91	***	
Var(Vi)	0.380	***	0.03679				
Var(Ei)	0.408	***	0.02558				
AIC	3453.9						

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	4.907	***	0.852		5.313	***	0.863
Tokens	0.034	***	0.017		0.052	***	0.018
Entropy	-0.035	***	0.112		0.057		0.123
Lexicaldensity	-0.185	***	0.864		-0.589		0.892
STTR100	-0.041	***	0.010		-0.036	***	0.010
Positiveadhocsum	34.096	***	4.365		37.842	***	4.594
Negativeadhocsum	-86.739	***	8.154		-106.525	***	8.625
- 2Log Likelihood	3395.1			R^2	0.186		
LR null model	175.8	***		AdjR^2	0.183		
Var(Ui)	0.03377	***	0.0117	F	51.91	***	
Var(Vi)	0.3574	***	0.0368				
Var(Ei)	0.4004	***	0.0256				
AIC	3415.1						

Dependent variable is Sratings. *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test.

While the summated scale operationalization overcame some of the limitations of the first approach, the operationalization has drawbacks. The summated scale forces all terms within a

category of the dictionary to have the same weight and contribution when the summated variables[10] are created. That is, according to this operationalization, the word "good" and the word "great" have the same effect in determining the valance of the review. A secondary issue that stems from the equal weighting of the summated index is that the reduction of variables comes at the cost of variance explained. This finding implies that the weights used for creating the index variables are not optimal, at least using the amount of explained variance as a criterion of optimality.

While there are other technically valid approaches to deal with these aforementioned drawbacks (see footnote 10 for example), a more elegant solution is to account for the variation of the observed variables (in this case, the word marginal probabilities) using latent variables. In particular, it is contended that in the case of the movie critics, after the movie is watched the critics develop a complex set of attitudes and effects about the movie and some of its components (Was the casting adequate?, What is the level of acting?, Are the special effects adequate?, etc.). After this natural evaluation occurs, the reviewer chooses words that best match the attitudes that he or she has formed during and after the movie experience.

Following this line of thought, the word frequencies, and by extension their marginal probabilities, are mere indicators of latent attitudes that have been developed in the critic's mind. That being the case, we can use the scaled frequency data to use either exploratory (exploratory factor analysis [EFA], principal component analysis [PCA]) or confirmatory techniques (structural

[10] Note that the creation of the variables Positive and Negative can be generalized to a linear convex combination of the original words such that $\text{Positive} = \sum_{j \in S_1} v_j f(w_j)$ with the sum of the weights of each word, vj, adding to one. The same can be done for Negative. The issue is how to obtain estimates of the weights. One potential simple approach is to use least squares to minimize the difference between Positive and Negative and the dependent variable ratings (which is similar to what we did in the first analysis where all words were entered into an OLS regression. The OLS estimates are the weights). Other simple alternatives require the collection of additional data. For example, a questionnaire can be used to determine weights based on responses to adequacy of each word to describe a movie for different degrees of effect.

equation models [SEM] and partial least squares [PLS]) to model the impact of the content of the review in movie performance.

I first used exploratory factor analysis (EFA) and principal components analysis (PCA) to analyze the data obtained with the first criteria to choose words. Given our previous discussion, a principal component extraction is most appropriate to mach as closely as possible with the formative nature of the data. Given that this research has been guided by theory, a two-stage approach to principal component regression was employed. A single component PCA was fitted to each of the two sets of words, positive and negative, separately. After the correlation matrix was decomposed, regression-based factor scores were obtained for each of the two latent constructs. These factor scores were used to estimate a similar model to those estimated thus far. Table 6 provides the estimates of the model. As shown in the table, the PCA-based model fits the data better than the summated scales model discussed earlier, that is, equal weights are not supported in this particular application.

Table 6
Mixed Model and OLS Regression for Ad Hoc Dictionary (Predicting Ratings Using PCA- and PLS-based Scores and Structure)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	4.976	***	0.848		5.337	***	0.828
Tokens	0.034	**	0.017		0.050		0.017
Entropy	-0.078		0.113		0.011		0.119
Lexicaldensity	-1.022	**	0.861		-1.870		0.862
STTR100	-0.039	***	0.010		-0.032		0.010
PositiveadhocPCA	0.248	***	0.023		0.305	***	0.024
NegativeadhocPCA	-0.231	***	0.023		-0.287	***	0.024
- 2Log Likelihood	3360.6			R2	0.239		
LR null model	118.450	***		AdjR2	0.236		
Var(Ui)	0.034	***	0.011	F	71.28	***	
Var(Vi)	0.277	***	0.035				
Var(Ei)	0.435	***	0.028				
AIC	3380.6				LR Structure	52.6	***

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	4.761	***	0.793		5.153	***	0.771
Tokens	0.026	*	0.016		0.040	***	0.016
Entropy	-0.094		0.107		0.002		0.111
Lexicaldensity	-0.939		0.812		-1.625		0.807
STTR100	-0.038	***	0.009		-0.031	***	0.009
PositiveadhocPLS	0.226	***	0.017		0.267	***	0.018
NegativeadhocPLS	-0.259	***	0.017		-0.289	***	0.018
- 2Log Likelihood	3195.800			R^2	0.336		
LR null model	96.900	***		AdjR^2	0.333		
Var(Ui)	0.025	***	0.009	F	114.8	***	
Var(Vi)	0.222	***	0.030				
Var(Ei)	0.404	***	0.026				
AIC	3215.800				LR Structure	48.8	***

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

It is interesting to note that the PCA weightings for the words are based solely on their own properties and that no external influences (i.e., the dependent variable that is movie ratings) are used to "optimize" the weightings. Even under these circumstances, there is an improvement in fit that emanates mainly from the positive set of comments. In particular, it seems as if some of the words are inversely related to the dependent variable, and, when averaged, they cancel each other, reducing

(biasing) the estimated effect. Again, based on the results presented in Table 6, it should be noted that there is significant heterogeneity and that structure as a set is relevant with a substantive pattern similar to that found when operationalizing content using the summated scale.

I also modeled the words using PLS. PLS allows for the word weights that determine the underlying latents to be determined using the dependent variable as part of the optimization procedure. Given that we concluded earlier that the words are formative measures of attitude, and therefore a formative model is a more suitable data-generating process, we use partial least squares to estimate the covariance analysis model (Fornell & Bookstein, 1982). For application and prediction, a PLS approach is often more suitable, especially if formative models are to be estimated. Since the approach estimates the latent variables as exact linear combinations of the observed measures, PLS avoids the indeterminacy problem suffered by common factor models and provides an exact definition of component scores. Other advantages of PLS are that the distributional assumptions are not as restrictive as in the covariance modeling approach (also known as LISREL because of the software package) and that the sample size requirements for stability are not as demanding (Wold, 1985). PLS also is better suited than multivariate regression because PLS accounts for measurement error and avoids possible multicollinearity problems (Ryan, Rayner, & Morrison, 1999).

In relating measures to constructs and permitting the construction of a system of equations, PLS attempts to maximize both the variance explained by the measures (indicators) and simultaneously create variates[11] that maximize the variance explained among the endogenous constructs. The estimation of PLS uses a series of OLS regressions that are optimally weighted to

[11] In multivariate techniques a variate, V, is defined as a linear combination of the form $V = x\omega$ where ω is a $t \times 1$ vector of weighs and $\boldsymbol{x}$ is a $n \times t$ matrix that contains the original variables. The resulting variate is a vector with $n \times 1$ elements.

create latent scores that subsequently will be run using OLS to determine the structural or path estimates. This iterative process is repeated until an optimal value for the path coefficients and weights is reached.

The modeling approach used in PLS was similar to the previous approaches discussed. I determined two latent constructs, positive attitude and negative attitude, assigning the same words already mentioned in Table 1 to each latent according to the two-list classification. I obtained two sets of PLS weight-based scores, one for positive comments and one for negative comments. These scores were then used to run OLS and mixed models. Table 6 also reports the estimation of this model. As can be seen from the table, this was the best-fitting model thus far. It should come as no surprise that the two latents have stronger effects on critics' ratings since the weights that PLS obtained are computed to maximize the variance explained. While negative comments have an effect in attitude that is about 20% higher than positive comments, the difference is not as marked as before. The rest of the results are consistent with the previous analysis and indicates that this is the best approach tested thus far since it explains the largest amount of variance using theoretically sound coefficients.

To summarize the testing so far, I found that we can explain significant amounts of variance in overall ratings by observing the likelihood of the words in the two categories of the ad hoc dictionary, i.e., length of the review and the complexity of the review. This result is robust to the operationalization of the content metrics, attesting to the validity of the dictionary as a means to capture content. I found, however, that not all operationalizations of the ad hoc dictionary were equally efficient in extracting content and that the summated scale traditionally used in marketing is inferior to both principal components and PLS, which is the most effective of all the operationalizations tested for this dictionary.

General purpose dictionary

The next step in the proposed analysis was to use the information in the general purpose dictionary to similarly validate it using the calibration dataset. Remember that we use the categories Positive and Negative, used by the General Inquirer, and which ultimately reflect the Harvard-IV and Lasswell dictionaries. While the dictionary registers in excess of 2,000 negative terms and about 2,000 positive terms, 1,430 distinct negative words and 1,023 positive words were present in our dataset.

In testing the validity of this dictionary, there is an important difference in the operationalizations that are feasible in the case of the general purpose dictionary (GPD). Individual term operationalizations are not included in this case since the number of independent variables would be larger than the number of observations in the dataset, and hence the model cannot be estimated (i.e., has negative degrees of freedom).

Our first model then included the summated scale of these two sets, called NegativeGISum and PositiveGISum. Table 7 reports the results from fitting such a model. As can be seen, the model has similar substantive results compared to the summated index in the ad hoc dictionary. In both cases, the variables capturing content are significant and possess the appropriate sign. There is also a substantive amount of unobserved heterogeneity, mostly across movies.

Table 7
Mixed Model and OLS Regression for General Purpose Dictionary (Predicting Ratings Using Summated and PCA-based Scores and Structure)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	5.612	***	0.870		6.445	***	0.8774
Tokens	0.036		0.017		0.064	***	0.0182
Entropy	-0.021		0.114		0.123		0.1266
Lexicaldensity	-1.444		0.910		-2.205	**	0.9467
STTR100	-0.045	***	0.010		-0.040	***	0.0102
PositiveGISum	20.803	***	2.223		23.409	***	2.3284
NegativeadhocGISum	-12.462	***	2.326		-13.888	***	2.4449
- 2Log Likelihood	3528.600			R^2	0.1596		
LR null model	209.400			AdjR^2	0.1559		
Var(Ui)	0.051	***	0.015	F	43.04	***	
Var(Vi)	0.419	***	0.041				
Var(Ei)	0.420	***	0.027				
AIC	3550.6				LR Structure	55.4	***

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	6.452	***	0.8975		7.403	***	0.907
Tokens	0.045	**	0.01759		0.060	***	0.019
Entropy	0.000		0.1181		0.131		0.132
Lexicaldensity	0.031		0.9092		-1.488		0.946
STTR100	-0.788	***	0.01004		-0.050	***	0.011
PositiveGIPCA	-0.050	***	0.02568		0.153	***	0.026
NegativeGIPCA	0.139		0.02279		0.030		0.026
- 2Log Likelihood	3528.700		0.0228	R^2	0.080		
LR null model	209.790			AdjR^2	0.076		
Var(Ui)	0.050	***	0.015	F	19.78	***	
Var(Vi)	0.419	***	0.041				
Var(Ei)	0.420	***	0.027				
AIC	3548.700				LR Structure	48.8	***

Dependent variable is Sratings. *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test.

The fit of this and the same operationalization in the ad hoc dictionary are now compared. Given that the variables included in the two models are different and hence the models are not nested (they have the same degrees of freedom), I used AIC, an information criteria measure, to determine the best model in this circumstance (Akaike, 1973; Bozdogan, 1987). AIC is defined as $AIC=n\ ln(RSS)+2m$, where *ln* is the neperian logarithm, RSS is the residual sum of squares for the

model, *m* is the total number of parameters in the model, and n is the sample size. As can be seen from the fit statistics,[12] this model fits worse than the one based on the ad hoc dictionary that uses a summated scale. The overall results show the same significance pattern, but the variable that captures the negative comments effect exhibits a much weaker effect. Given this pattern, the most likely explanation for these two combined results is that many of the words are negative in a general sense, implying that their actual valence varies greatly introducing a large amount of noise into the model. This potential explanation was also confirmed by the relative size of the coefficients within the model that uses the summated index in the GPD. Note that the negative construct is the one that has the largest number of terms, more than 1,400, while the strongest effect is for the positive categories having just over 1,000 terms. Further empirical evidence of this hypothesis was found for the next two models.

The second operationalization of the GPD was the PCA-based regression factor score model. The same procedure conducted for the ad hoc dictionary was used. The two factor score variables are PositiveGIPCA and NegativeGIPCA. The results of this model may be surprising to some since the model fits the data poorly. This poor fit arises from the demands imposed on the methods. PCA requires the computation and manipulation of covariance matrices. In the situation for the positive case, the covariance matrix can be computed (with little accuracy due to the large number of variables); in the case of the negative comments, the number of variables is larger than the number of data points. This explains the nonsignificant result of the negative variable in this model. This is further evidence that the number of variables used is the source of the results we observe. Given these shortcomings, this model is highly questionable, and hence no further substantive implications will be drawn from it.

[12] Note that, as AIC is computed by SAS, the lower its value, the better fitting the model is.

The third model that uses the general purpose dictionary tries to leverage the fact that the information set is large. That is, if we can choose only the words that are real correlates of the dependent variable, and perhaps weight them according to the strength of the relationship, then we may be able to achieve the best of both worlds. To do this, I used PLS to obtain factor scores for the positive and negative variables. PLS is particularly effective at reducing a large number of variables for predictive purposes. The results of this exercise are presented in Table 8. First, note that the fit of the model is excellent, corroborating that when a large amount of potential information is available, PLS does an excellent job in extracting it. It does such a good job that the structural variables as a set become nonsignificant in this model.

Table 8
Mixed Model and OLS Regression for General Purpose Dictionary (Predicting Ratings Using PLS-based Scores and Structure)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	3.103	***	0.572		3.62106	***	0.526
Tokens	-0.009		0.011		-0.0027		0.011
Entropy	-0.059		0.077		-0.0089		0.076
Lexicaldensity	-0.131		0.573		-0.44407		0.547
STTR100	-0.016	**	0.006		-0.01607	***	0.006
PositiveGIPLS	0.233	***	0.010		0.2362	***	0.010
NegativeadhocGIPLS	-0.262	***	0.010		-0.26413	***	0.010
- 2Log Likelihood	2199.000			R^2	0.694		
LR null model	36.360	***		$AdjR^2$	0.692		
Var(Ui)	0.019	***	0.007	F	513.37	***	
Var(Vi)	0.041	***	0.013				
Var(Ei)	0.242	***	0.014				
AIC	2219.000				LR Structure	5.9	

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

This last observation is worrisome since throughout all the analyses we have observed that the structural variables are consistently significant under all operationalizations. The large number of variables used to fit the PLS model has one potential drawback. PLS is prone to overfitting when the

number of independent variables is extremely large. To verify this potential problem, a cross-validation check of the original PLS models was conducted. I used more than eighth dataset splits to conduct the cross-validation. When the cross-validation was conducted, the computer returned a warning, indicating that the cross-validation observation results were far from the training set and that the results may be numerically sensitive. Using the press statistics and cross-validation to determine the number of underlying PLS dimensions in the data, I selected zero dimensions. This is an indication that the model is overfitting the data; thus, our best effort at modeling this data is the summated scale, which as previously noted, exhibits poorer fit than our ad hoc dictionary.

ALSA

Finally, I looked to the ALSA-based methodology to construct valid measures based on a linear combination of all the content in the data. As described earlier, two pairs of scores, one for the good/bad pair and one for the enjoyable/dull pair, will be generated. Note that any other expression can be treated in a similar manner. As a first step, I generated an ALSA-based representation of our observed data matrix Mp. To do this, I needed to first decompose the original matrix according to equation (2).

After the decomposition was accomplished, I selected the number of dimensions that would be used to reproduce the data. I followed a simple and conservative test philosophy; I selected 100 underlying dimensions or constructs in the reviews following general literature directions in LSA/LSI research (e.g., Deerwester at al., 1990; Foltz & Dumais, 1992). Note that a grid search method can be used to find the optimal number of latent dimensions that are used by ALSA. However, to avoid overfitting, I stayed on the low side of the range suggested by the methodology,

100 to 300 dimensions, and I did not use the grid search to find optimal number of dimensions.[13] After the number of dimensions was decided, I used (3) to reproduce the original data based on the underlying 100 dimensions. After this was accomplished, I computed the distance, or cosine matrix per (1), which I used to obtain the scores by multiplying the Mp matrix by the four-word columns of the distance matrix that contains the words good, bad, enjoyed, and dull to obtain the latent scores.

Note that in this process I have not used information other than that contained in the documents. This is important because I avoided overfitting problems. The latent score is a reflection of the position of the words in the documents. The underlying assumption is that if words are more likely to appear together in the same document (other units of analysis such as phrases can be used also), they are more likely to be related, especially if the occurrence is repeated across multiple occasions. The results of this effort are provided in Table 9. We see that the model fits the data best of all the models that I have tested and validated so far. The AIC for this model is lower than that of any other model estimated. We also see again that there is a large amount of heterogeneity across the clustering variables and that the same results that arise with the other analyses are replicated in this case. All signs of significant variables are as expected. Both negative comment scores show stronger effects than their positive counterparts.

[13] This is, therefore, a conservative test as the results I found are a lower bound of what we could find were the number of dimensions be determined using the grid search.

Table 9
Mixed Model and OLS Regression for ALSA-based Content (Predicting Ratings Using ALSA-based Scores and Structure)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	4.873	***	0.858		5.33911	***	0.8427
Tokens	0.277	***	0.068		0.24685	***	0.0641
Entropy	-0.292	***	0.111		-0.21161	*	0.1134
Lexicaldensity	-0.959		0.833		-1.72151	**	0.8150
STTR100	-0.055	***	0.010		-0.04975	***	0.0098
AALSA_BAD	-0.066	***	0.004		-0.06682	***	0.0034
AALSA_DULL	-0.025	***	0.003		-0.02802	***	0.0029
AALSA_ENJOYED	0.014	***	0.003		0.02022	***	0.0030
AALSA_GOOD	0.062	***	0.004		0.06065	***	0.0037
- 2Log Likelihood	3190.7			R^2	0.3290		
LR null model	116.630			$AdjR^2$	0.3251		
Var(Ui)	0.068	***	0.018	F	83.24	***	
Var(Vi)	0.186	***	0.030				
Var(Ei)	0.409	***	0.026				
AIC	3214.7				LR Structure	42.4	***

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

Given that this is the best-fitting linear model, I investigated the possibility that, as I have theorized, the effects of structure are nonlinear and that the positive and negative content in the words may interact. To do this, I followed a hierarchical modeling approach. I have already estimated the model using simple linear effects. In the second stage, I introduced square terms for the structure variables and the content variables to test the hypothesized nonlinearities. The results for the addition of quadratic effects are provided in Table 10. Note that modeling the effects using a quadratic function is supported overall by the data, LR (8)=101.5 and highly significant, suggesting that the effects are indeed nonlinear. A similar conclusion is reached if we use AIC for model selection.

Table 10
Mixed Model and OLS Regression for ALSA-based Content (Predicting Ratings Using ALSA-based Scores and Structure with Quadratic Effects)

	MIXED				*OLS*		
Effect	Estimate		Standard error		Estimate		Standard error
Intercept	-16.3955		10.52700		-17.34027		10.79651
Tokens	0.5773	***	0.17960		0.56545	***	0.17509
tokens2	0.0004		0.00876		-0.00431		0.00873
Entropy	5.5601	***	1.20500		4.86889	***	1.21129
Entropy2	-0.4138	***	0.09035		-0.36667	***	0.09028
Lexicaldensity	12.6457		15.69630		14.90832		16.56629
lexicaldensity2	-14.0698		16.25430		-17.25827		17.17876
STTR100	-0.0675		0.24010		0.00331		0.24661
STTR1002	0.0003		0.00154		-0.00008		0.00158
ALSA_BAD	-0.1255	***	0.00957		-0.12952	***	0.00975
ALSA_BAD2	0.0001	***	0.00001		0.00008	***	0.00001
ALSA_DULL	-0.0577	***	0.00815		-0.06825	***	0.00798
ALSA_DULL2	0.0000	***	0.00001		0.00005	***	0.00001
ALSA_ENJOYED	0.0277	***	0.00854		0.03720	***	0.00813
ALSA_ENJOYED2	-0.00002	**	0.00001		-0.00003	**	0.00001
ALSA_GOOD	0.1271	***	0.01042		0.13169	***	0.01053
ALSA_GOOD2	-0.0001	***	0.00001		-0.00008	***	0.00001
- 2Log Likelihood	3089.2			R^2	0.3816		
LR null model	106.67			$AdjR^2$	0.3743		
Var(Ui)	0.063	***	0.0174	F	52.06	***	
Var(Vi)	0.166	***	0.0266				
Var(Ei)	0.384	***	0.0242				
AIC	3129.2			LR quadratic	101.5	***	

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

For all content words, I found that the quadratic effects are significant and have the opposite sign to the simple effect, that is, there are diminishing marginal returns to the increase in the probability of observing a word in a given set. This is not surprising since once someone says that a movie is horrible and really bad, adding additional negative comments may not alter judgments about the movie that much. It is also interesting to note that there is a change in the role of complexity. Entropy played no role in the determination of critic ratings, but once the square term is used to model its effect, it becomes strongly significant. This is evidence of a strong nonlinear effect. I found that while initially the effect on increased entropy is positive, after entropy increases enough,

its effect could become negative. Note, however, that within the range of values of entropy that I observed in the data, the effect of entropy was always positive, making this a purely theoretical possibility. The length of the message retains its positive linear effect on ratings. There is considerable heterogeneity, as I have found previously.[14]

Finally, Table 11 shows the estimates for a model that, in addition to the quadratic effects, adds interaction terms among the content variables to investigate whether there are synergistic effects. As can readily be seen, this model does not improve fit significantly, LR=5.4 with df =6 p=0.49. Therefore, there is no evidence that the four dimensions of content included in the model interact.

[14] I tried several different specifications for the random component part. In particular, I tried random slopes for the coefficients in the model but failed to identify any model that was stable and improved fit.

Table 11
Mixed Model and OLS Regression for ALSA-based Content (Predicting Ratings Using ALSA-based Scores and Structure with Quadratic Effects and Interactions of Content Variables)

	MIXED			*OLS*		
Effect	Estimate		Standard error	Estimate		Standard error
Intercept	-16.19820		10.5122	-17.1825		10.8073
Tokens	0.53580	***	0.1824	0.5133	***	0.1787
tokens2	0.00319		0.0090	-0.0012		0.0090
Entropy	5.49060	***	1.2048	4.7497	***	1.2138
Entropy2	-0.40840	***	0.0903	-0.3576	***	0.0905
Lexicaldensity	12.07690		15.6868	13.9497		16.5963
lexicaldensity2	-13.49230		16.2425	-16.3234		17.2067
STTR100	-0.06221		0.2400	0.0160		0.2468
STTR1002	0.00029		0.0015	-0.0002		0.0016
ALSA_BAD	-0.12630	***	0.0103	-0.1311	***	0.0105
ALSA_BAD2	-0.00001		0.0002	0.0001		0.0003
ALSA_DULL	-0.06257	***	0.0090	-0.0708	***	0.0090
ALSA_DULL2	-0.00003		0.0002	0.0000		0.0002
ALSA_ENJOYED	0.03105	***	0.0094	0.0402	***	0.0092
ALSA_ENJOYED2	0.00006		0.0002	0.0000		0.0002
ALSA_GOOD	0.12960	***	0.0110	0.1337	***	0.0112
ALSA_GOOD2	-0.00044	**	0.0002	-0.0005	**	0.0002
ALSA_BADxALSA_DULL	0.00009		0.0003	-0.0001		0.0003
ALSA_BADxALSA_ENJOYED	-0.00032		0.0003	-0.0004		0.0003
ALSA_BADxALSA_GOOD	0.00034		0.0004	0.0004		0.0004
ALSA_DULLxALSA_ENJOYED	-0.00023		0.0002	-0.0001		0.0002
ALSA_DULLxALSA_GOOD	0.00022		0.000292	0.0003		0.0003
ALSA_GOODxALSA_ENJOYED	0.00034		0.000282	0.0004		0.0003
- 2Log Likelihood	3083.8			R^2	0.384	
LR null model	105.95			AdjR^2	0.374	
Var(Ui)	0.064	***	0.018	F	38.13	***
Var(Vi)	0.165	***	0.027			
Var(Ei)	0.383	***	0.024			
AIC	3135.8			LR Interactions	5.4	

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

In this section, I have analyzed the predictive validity of the ALSA measure. I found that the general purpose dictionary (GPD) performs poorly when compared both to the ad hoc dictionary and the ALSA-based method. The general purpose dictionary is "generic," inducing too much noise to capture the content accurately. While the ad hoc dictionary performed better than the GPD, I

found that the ALSA-based method performs best, offering a flexible way of including content without the need to create exhaustive list of words. In the following section, I analyze the effects that content has on movie performance.

Validation Dataset

I turn now to the validation dataset. Note that while the coding of the information takes place at the review level, given that we are interested in the effect that content and its delivery has at the movie level, the unit of analysis requires aggregation of the review information to the movie level. This is necessary because the effect of an individual review is not distinguishable given the information available, i.e., overall movie performance (what would the box office have been if one of the reviews had not been written). Given that the dependent variable is at a movie level, I aggregated the frequencies and scaled frequencies (i.e., marginal probabilities) for each set. Following the same testing procedure, I first used the observed scaled frequencies directly as predictors of the performance of the movies, for which I analyzed three different operationalizations: box office, gross profit, and return on budget.

To provide evidence of the validity of the approach, I briefly look at the predictive validity of the measures using review ratings as the dependent variable. As in the previous section, I checked the predictive validity of content and structure. Note, however, that because I am estimating models at an aggregated movie level, the previous control of heterogeneity is not possible. If the words are treated as independent entities, then a simple regression can be used to see how the frequency of each word contained in the reviews affects the performance of the movie. The models used can be written as

$$P = \beta_w\, g(W) + \beta_x\, X + e \qquad (15)$$

$$R = \beta_w' g(W^*) + \beta_x' X^* + u \quad (16)$$

where P is a vector of a movie performance measure (e.g., box office, gross profit…), β_w is a vector of parameters that correspond to the word scaled frequencies, g(W) is the matrix of the scaled frequencies of the relevant words, β_x is a vector of parameters for the covariates, X is a matrix of covariates, and e is a vector of errors. Similarly, R is a vector of movie ratings, β_w' is a vector of parameters that correspond to the word scaled frequencies, $g(W^*)$ is the matrix of the frequencies of the relevant words, β_x' is a vector of parameters for the covariates, X^* is a matrix of covariates, and u is a vector of errors.

For the model predicting ratings, the results of the OLS regression are given in Table 12. Note that the model explains more than 63% of the variance in the movie ratings. This increase in the predictive ability over a similar model that predicts individual ratings should not be surprising. Predicting individual behavior is, in general, much more difficult than doing so for the average critic. If we consider this as a criterion validity test, we can see that the words selected correlate highly with movie reviews.[15] We can interpret this result as lending support for the idea that the probability of observing these words in the review is an indication of the overall valance of the review.[16]

[15] If we construct a linear combination of the words creating a composite with weights equal to the regression weight, this variable and ratings correlate at approximately 0.8.

[16] In the case in which ratings for a particular movie review are not available, this or other similar procedures could be used to obtain approximate ratings (forecasting).

Table 12
OLS Regression (Predicting Ratings Using Individual Words in the Ad Hoc Dictionary)

Variable	Parameter	Std. error	T statistic
Intercept	5.9019***	0.468	12.611
Tokens	0.0001**	0.0001	2.2324
Absurd	0.0216	0.1727	0.125
Amazing	0.1486	0.1038	1.4319
Amusing	-0.0736	0.0815	-0.903
Annoying	-0.1061	0.0936	-1.134
Bad	-0.0166*	0.0085	-1.9532
Best	0.0159*	0.0092	1.7194
Brilliant	0.0932	0.09	1.0356
Convincing	-0.0882	0.0955	-0.9234
Dislike	-0.1344	0.2545	-0.5282
Dull	-0.2559**	0.1132	-2.2604
Enjoy	0.0881	0.0821	1.0738
Enjoyable	0.1089	0.0979	1.1124
Enjoyed	-0.192	0.1283	-1.4969
Excellent	0.138**	0.0623	2.2141
Extraordinary	0.1277	0.2261	0.5646
Fantastic	-0.0746	0.1633	-0.4567
Favorite	0.1894***	0.0707	2.6797
Finest	0.0327	0.2342	0.1398
Fun	-0.0349	0.0285	-1.2234
Good	0.0047	0.0081	0.5818
Great	-0.0043	0.0074	-0.579
Greatest	-0.0239	0.0715	-0.3343
Hate	-0.0216	0.0175	-1.2309
Horrible	-0.2128	0.14	-1.5202
Horrific	0.148	0.2537	0.5834
Incredible	0.0922	0.0633	1.4557
Interesting	0.0721	0.0497	1.4505
Like	-0.0344**	0.0153	-2.249
Love	0.0037	0.009	0.4142
Memorable	0.1391	0.1223	1.1373
Outrageous	-0.086	0.1623	-0.5299
Painful	-0.0355	0.1647	-0.2156
Perfect	0.0924*	0.0486	1.9027
Poor	-0.1942**	0.0831	-2.3379
Pretty	-0.0374	0.0385	-0.9736
Remarkable	-0.0015	0.1885	-0.0079
Ridiculous	-0.1901	0.1167	-1.6284
Silly	-0.0826	0.0669	-1.2336
Stupid	0.009	0.0578	0.1563
Superb	0.3064**	0.1399	2.1894
Terrible	-0.2949**	0.1352	-2.1819
Terrific	0.1626	0.1035	1.5715
Tremendous	0.0866	0.3269	0.265

Variable	Parameter	Std. error	T statistic
Unfortunate	0.0258	0.2292	0.1125
Wonderful	0.0381	0.0805	0.4732
Worst	-0.191 ***	0.067	-2.8493
Wrong	0.0895	0.0543	1.6472
Dependent Variable: SRATING			
R Square	0.632	Adjusted R Square	0.541
Overall Model significant at 1%			

Note that again I am using marginal probabilities (scaled frequencies) instead of word frequencies as input for this and other models. This was done to break down two effects that are otherwise mixed in the frequency data, especially at the movie level. On one hand, more words are observed if there are more reviews for a particular movie and/or if the reviews for the movie happen to be longer. Common sense dictates that movies that are widely released will have more reviews and therefore will have more words when the reviews are aggregated to the movie level. On the other hand, some movies, regardless of the number or length of the reviews, have greater frequency of a particular set of words. This is the effect that I am interested in capturing. To tackle this issue, I again used scaled frequencies, dividing each frequency by the total number of words for that particular movie. After this transformation, the frequency became an estimate of the probability of observing a particular word when a review is chosen at random. For the reminder of the analysis, and unless otherwise noted, I will use scaled frequencies for all analyses.

Procedures similar to those reported in the previous section (with minor required modifications) were followed. First, I did not estimate any PLS-based general purpose models since they already showed clear evidence of model overfit. Second, I created the ALSA-based scores using an extra conservative approach. Instead of estimating a new matrix of distances before the scaled frequency matrix is weighted to create the scores, I used the distance matrix already calculated from

the first dataset. This procedure ensures that the original results did not occur because of overfitting. Following this analysis strategy, I found that the ALSA-based procedure performs well, but a new difficulty was encountered. The level of multicollinearity could have been considered high in the first dataset, with maximum variance inflation factors[17] in the 30s range for the model, including quadratic effects. However, in this dataset, multicollinearity is approximately similar in magnitude, but the smaller sample size at the unit of analysis level (242 movies versus more than 1,400 reviews) limits the stability of the results. Tables 13 and 14 show the quadratic and linear models for the ALSA-based scores. Note that while the fit of the model that includes quadratic terms is good with more than 50% of the variance explained,[18] a linear model does significantly no worse, $F_{8,224}=1.16$ $p>0.1$. Hence, the more simple linear specification is favored at the aggregate level. As can be seen from the tables, the overall results are highly consistent with those obtained in the first dataset.

[17]Variance Inflation Factors (VIF) are computed as $(1 - R_j^2)^{-1}$ for each regressor b_j where R_j^2 is obtained by regressing the j's predictor in X on the remaining predictors. These terms are usually used to diagnose potential collinearity problems. Maximum VIF values of over 10 are considered potentially harmful however as noted by other (Mason & Perreault, 1991) if the sample size is high enough multicollinearity may not have negative effects in estimation and testing.

[18] Note that 50% shared variance implies that is I created a linear composite using these weights, and I used this and ratings as two measures of reviewers' overall attitude their reliability will be in excess of 0.7, providing further evidence of measurement validity.

Table 13
OLS Regression for ALSA-based Content (Predicting Ratings Using ALSA-based Scores and Structure with Quadratic Effects)

	OLS			
Effect	Estimate		Standard error	VIF
Intercept	5.7811		0.09766	0
MCTokens	4.94E-06		3.98E-06	9.79
MCTokens2	-2.47E-11		3.09E-11	4.11
MCentropy	1.3791	**	0.65140	5.58
Mcentropy2	-0.0505		1.22970	1.55
MCLexicaldensity	-0.5379		4.99607	1.83
Mclexicaldensity2	12.6303		224.58695	1.81
MCSTTR	-0.2521	***	0.08617	1.86
MCSTTR2	0.0541		0.05088	1.13
MCALSA_BAD	-169.1297	***	17.57324	15.20
MCALSA_BAD2	-1569.2074	*	810.93477	9.21
MCALSA_DULL	-27.9906		22.36052	17.23
MCALSA_DULL2	-774.8542		1361.29050	12.85
MCALSA_ENJOYED	62.0283	***	20.01254	15.57
MCALSA_ENJOYED2	-859.07767		1030.31645	9.84
MCALSA_GOOD	128.0538	***	23.19516	31.23
MCALSA_GOOD2	2171.3220	**	937.97294	16.92
R^2	0.5128			
AdjR^2	0.478			
F Null model	14.740	***		
F All quadratic terms=0	1.160			

Dependent variable is Sratings. *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test. MC stands for Mean Centered.

I also estimated PLS scores based on the ad hoc dictionary. Results for this analysis are shown in Table 15. Given that the quadratic terms do not add significant amount of explanatory power, $F_{6,226}$=0.24 p>0.1, I report only results from the linear model. As can be seen in the table, this modeling approach fits the data better than the ALSA model. Specifically, almost 65% of the variance in ratings can be explained by this set of variables. Note again that the results are consistent in sign and relative effect importance with those found in the previous model, thus demonstrating the robustness of the effects. Given that this approach has been shown to be superior in the aggregated dataset, I focused on this approach for the remainder of the analysis. Also, while this model uses a two-step approach, first estimating the PLS-based scores and then using the scores to

run an OLS regression, I can estimate these steps simultaneously, and improve efficiency. All the following models are based on a single-step PLS modeling approach with the content coming from the ad hoc dictionary.

Table 14
OLS Regression for ALSA-based Content (Predicting Ratings Using ALSA-based Scores and Structure Linear Terms Only)

	OLS			
Effect	Estimate		Standard error	VIF
Intercept	5.74916	***	0.05027	0
MCTokens	0.000003		0.000002	2.77
MCTokens2	-		-	
Mcentropy	1.62552	***	0.46101	2.79
Mcentropy2	-		-	
MCLexicaldensity	1.49616		4.76940	1.66
Mclexicaldensity2	-		-	
MCSTTR	-0.27024	***	0.08042	1.61
MCSTTR2	-		-	
MCALSA_BAD	-175.40730	***	17.07661	14.30
MCALSA_BAD2	-		-	
MCALSA_DULL	-29.58562		21.62541	16.06
MCALSA_DULL2	-		-	
MCALSA_ENJOYED	54.07716	***	18.95542	13.92
MCALSA_ENJOYED2	-		-	
MCALSA_GOOD	141.64295	***	21.81557	27.54
MCALSA_GOOD2				
R^2	0.4938			
AdjR^2	0.4763			
F Null model	28.29	***		
F structure	11.31	***		

Dependent variable is Sratings. *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test.

Effect of Content and Structure of the Reviews on Movie Performance

Following most past studies, I measured performance by looking at box office revenues and box office revenues per screen (Eliashberg & Shugan, 1997; Basuroy, Chatterjee, & Ravid, 2003). In addition, I also considered gross profit (net contribution) and ROI. For model selection, I estimated models with different transformations of the dependent and independent variables and found that

the substantive results were robust to the specification form of the different models. To avoid clutter, I report the results of the untransformed variables and/or those that are most conservative.

Table 15
OLS Regression for PLS-based Content Using Ad Hoc Dictionary (Predicting Ratings Using PLS-based Scores and Structure Linear Terms Only)

	OLS			
Effect	Estimate		Standard error	VIF
Intercept	7.4676	**	3.74662	0
Tokens	1.45E-06		1.68E-06	2.50
Entropy	9.88E-01	***	3.75E-01	2.68
Lexicaldensity	-2.9355		3.36257	1.20
STTR	-0.2126	***	0.06518	1.54
Positiveadhocpls	0.1855	***	0.02665	1.61
Negativeadhocpls	-0.2772	***	0.02782	1.34
R^2	0.6476			
AdjR^2	0.6386			
F Null model	71.680	***		
F All quadratic terms=0	0.240			

Dependent variable is Sratings. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test.

To analyze the effect of critics on movie performance, I used a set of variables that have already been shown to influence box office revenues in the past. In particular, I included as independent variables advertising/media spending (media), number of screens (screens), dummy variable to account for the movie having been a sequel (sequel), genre (dummies for family, action, drama, comedy, and thriller), and dummy for whether the movie is classified by the MPAA as R or not.

Tables 16 and 17 present the measurement and structural parts of the models that estimate the effect of critics on performance. I reproduce only the measurement model for the model that uses box office revenues as the dependent variable to avoid clutter. The models for the other movie performance variables and specifications are substantially similar. Remember that in the case of formative models, the usual measures of evaluation (e.g., reliability, average variance extracted

(AVE)) do not apply because the items need not necessarily correlate (Diamantopoulos & Winkhofer, 2001; Jarvis, Mackenzie, & Podsakoff, 2003). The only reflective measure in the model, length of the message, exhibits strong internal consistency and reliability. I fitted models in which the dependent variable was in levels or in log form. I comment on the results for the model that fits the data better for each of the three conceptual performance metrics: sales (box office revenues), profit, and return on budget. Table 18 provides the fit statistics for each of the estimated models.

Table 16
PLS Model with Movie Box Office as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Measurement Model for Multiple-item Constructs)

Construct	Variable	Loading		S.E.†	T-statistic††	Weight		S.E.	T-statistic
Positive	Amazing	0.6163	***	0.0885	6.9609	0.5366	***	0.1094	4.9057
	Amusing	-0.137	*	0.082	1.6715	-0.0848		0.0845	1.0036
	Best	0.3553	***	0.0581	6.1139	0.2502	***	0.0899	2.784
	Brilliant	0.2931	***	0.0839	3.494	0.1353		0.0873	1.5499
	Convincing	0.0598		0.0721	0.8289	0.0178		0.0869	0.2049
	Dazzling	0.3104	***	0.1062	2.9215	0.0124		0.1082	0.1146
	Enjoy	0.0409		0.0755	0.5418	0.0477		0.0926	0.5154
	Enjoyable	0.1761	**	0.0874	2.0139	0.1126		0.0982	1.1471
	Enjoyed	0.2369	**	0.0998	2.3741	0.1228		0.0993	1.2364
	Excellent	0.3842	***	0.0871	4.41	0.1168		0.1192	0.98
	Exceptional	0.1115		0.1268	0.8791	0.0885		0.1717	0.5156
	Extraordinary	0.1086		0.0936	1.1607	0.0079		0.087	0.0908
	Fantastic	0.393	***	0.0961	4.0916	0.2568	**	0.1048	2.4514
	Favorite	0.2818	**	0.1289	2.1868	0.1589		0.1117	1.4223
	Finest	0.2907	***	0.105	2.7676	0.1042		0.1018	1.0239
	Fun	0.1462		0.1049	1.3933	0.0543		0.103	0.5274
	Good	0.2765	***	0.1039	2.6612	0.1497	*	0.0887	1.6871
	Gorgeous	-0.0714		0.0777	0.9187	-0.1717	*	0.0896	1.917
	Great	0.0356		0.1121	0.3176	0.0442		0.0851	0.5196
	Greatest	0.0058		0.0816	0.0711	-0.1945	**	0.0789	2.4636
	Incredible	0.1564		0.141	1.1089	0.0376		0.0835	0.4506
	Interesting	0.2425	***	0.0832	2.9154	-0.0587		0.1	0.587
	Joyful	-0.1279	**	0.0605	2.1138	-0.1837	*	0.1061	1.7318
	Like	0.0252		0.0724	0.3479	-0.0928		0.1003	0.9254
	Love	0.0374		0.0787	0.4755	0.2824	***	0.089	3.1737
	Marvelous	-0.0482		0.0677	0.7121	-0.2321	**	0.1032	2.2484
	Memorable	0.1264		0.0772	1.6369	-0.1321		0.0803	1.6443
	Perfect	0.2427	**	0.1049	2.3144	-0.0863		0.1133	0.762
	Pretty	0.004		0.085	0.0471	-0.0423		0.0935	0.4525
	Remarkable	0.0515		0.0822	0.6268	-0.1464		0.1113	1.3151
	Splendid	0.0324		0.1023	0.3168	-0.1037		0.1025	1.0116
	Superb	0.2215	**	0.0903	2.4525	0.0511		0.1049	0.4872
	Terrific	0.4036	***	0.0963	4.191	0.2848	***	0.1069	2.6639
	Tremendous	0.2169	**	0.1065	2.0359	-0.053		0.0995	0.5328
	Wonderful	0.2703	***	0.0885	3.0525	0.0094		0.1028	0.0915
	Outstanding	0.4219	***	0.0928	4.5455	0.1815		0.1355	1.3399
Complexity	Entropy	0.9643	***	0.1109	8.6939	0.9102	***	0.0971	9.3756
	Lexicaldensity	-0.5105	***	0.1506	3.3887	-0.2479		0.1884	1.3159
	STTR	0.0423		0.1943	0.2177	-0.1023		0.254	0.4028
Negative:	Absurd	-0.3014	***	0.0916	3.291	-0.2091		0.1281	1.6327
	Aggravating	0.1587	**	0.0624	2.5442	0.0771		0.1263	0.6103
	Annoying	0.1083		0.0953	1.1366	0.0507		0.1449	0.35
	Awful	0.1962	**	0.0797	2.4632	0.0275		0.1194	0.2303
	Bad	0.1211		0.0902	1.3424	0.0747		0.1282	0.5825
	Badly	0.2607	***	0.085	3.066	0.2257	*	0.1242	1.8179

Construct	Variable	Loading		S.E.†	T-statistic††	Weight		S.E.	T-statistic
Negative:	Dire	0.0514		0.0774	0.664	0.0971		0.1088	0.8922
	Disgusting	-0.1323		0.1031	1.2828	-0.1086		0.1459	0.7441
	Dislike	0.045		0.0892	0.5042	0.0808		0.1233	0.6551
	Dismal	0.142	*	0.0829	1.7135	0.1768		0.1072	1.6496
	Dreadful	0.0084		0.1644	0.0511	-0.0191		0.1668	0.1145
	Dull	0.213	**	0.1055	2.0199	-0.0587		0.1398	0.4199
	Exasperating	0.1185	*	0.0701	1.6905	0.1075		0.0896	1.1999
	Frustrating	-0.1103		0.1089	1.0125	-0.0736		0.1265	0.5819
	Grim	0.1633	**	0.0645	2.5322	0.1859	*	0.1051	1.7694
	Hate	0.1013	*	0.0577	1.7552	-0.016		0.1115	0.1435
	Hideous	-0.1569		0.1261	1.2441	-0.0539		0.1278	0.4218
	Hopeless	0.0611		0.0838	0.7288	0.1038		0.1127	0.9211
	Horrendous	0.0435		0.0941	0.4624	0.1306		0.1475	0.8854
	Horrible	-0.1451		0.1242	1.1683	-0.1128		0.1585	0.7115
	Horrific	0.0029		0.1079	0.0269	0.0734		0.1477	0.4969
	Inadequate	-0.2028		0.2018	1.0052	-0.1273		0.1828	0.6962
	Irritating	0.041		0.0884	0.4637	0.0019		0.1081	0.0176
	Meaningless	-0.1173		0.107	1.0961	-0.1204		0.1348	0.893
	Ominous	-0.1436		0.1682	0.8537	-0.1498		0.1516	0.9881
	Outrageous	-0.2142	*	0.1223	1.7515	-0.2463	**	0.1236	1.9926
	Painful	0.2416	**	0.1068	2.2613	0.19		0.1158	1.6409
	Pathetic	0.1112		0.0955	1.1649	0.1513		0.1382	1.0949
	Pitiable	-0.4689	***	0.175	2.6789	-0.4608	***	0.1757	2.6221
	Poor	0.3666	***	0.0896	4.0906	0.281	**	0.1222	2.299
	Ridiculous	0.1213		0.0992	1.2227	-0.0189		0.1248	0.1515
	Silly	0.2656	***	0.0905	2.9344	0.3409	**	0.1382	2.4661
	Stupid	-0.0302		0.0955	0.3161	-0.0837		0.1382	0.6055
	Sucks	0.0247		0.1057	0.2336	-0.0408		0.1339	0.3048
	Terrible	0.2409	**	0.0956	2.5189	0.2181	*	0.128	1.7042
	Unfortunate	0.0758		0.1071	0.7078	0.1351		0.1342	1.0065
	Unpleasant	0.1331		0.0884	1.5061	0.1428		0.1136	1.2566
	Useless	-0.0806		0.0938	0.8594	-0.1911		0.1207	1.5835
	Worst	-0.0746		0.0986	0.7567	-0.3062	**	0.1371	2.2338
	Worthless	-0.0377		0.104	0.3626	-0.231	*	0.1334	1.7315
	Wrong	-0.0718		0.0946	0.7593	-0.1283		0.1361	0.9429
Length	Reliability	0.97		AVE	0.945				
	Tokens	0.9865	***	0.0232	42.4947	2.2511	**	1.0832	2.0781
	Sentences	0.9571	***	0.0388	24.6411	-1.2752		1.1096	1.1493

†Standard errors are computed empirically based on 500 bootstrap resamples. †† T is computed as the absolute value of the ratio between the parameter estimate and its standard error. Reliability is computed for reflective scales only as Reliability = ((Σλyi)2 / ((Σλyi)2 + Σ var(εi)) where var(εi) = 1 - λyi2 where λyi is the loading for construct y and item I and εi is the item error term. Average Variance Extracted (AVE) is only competed for reflective scales as AVE = $\Sigma\lambda_{yi}^2$ / $\Sigma\lambda_{yi}^2$ + Σ var(ε_i) with var(ε_i) = 1 - λ_{yi}^2 *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test.

I began by evaluating the effects of critics' reviews on box office revenues (in this case, the log of box office revenues, since it fits the data better for both the linear and the quadratic specifications). Assessing the parameters requires explanation. Since PLS makes no distributional

assumptions, traditional parametric methods of significance testing (e.g., confidence intervals, chi-square, etc.) are inappropriate. Therefore, bootstrapping was used to assess the statistical significance of the parameter estimates (Efron & Gong, 1983). Bootstrapping is sampling with replacement from observed data to estimate the variability in a statistic of interest. Instead of assuming that the variables have certain distributional properties (i.e., normality), I approximated the empirical sampling distribution of the statistic that I wanted to test by drawing from the actual sample with replacement (i.e., using my sample as a micro-population). To obtain an approximation of the density function, several samples are obtained from the original sample, resampling with replacement. For each of these new samples, the parameters of interest were calculated, that is, the model was estimated and the frequency distribution of the values is an approximation of the empirical distribution of the statistic as the number of resamples increases. This also allows the calculation of the standard errors of the statistic and confidence intervals. Standard errors were computed on the basis of 500 bootstrapping runs.

Table 17
PLS Model with Movie Performance as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Structural Effects with Linear Effects Only)

Dependent Variable	Box Office Parameter		SE	T†	Dependent Variable	Log of Box Office Parameter		SE	T
Positive	0.245	***	0.0617	3.9698	Positive	0.244	***	0.0568	4.2937
Negative	-0.133	**	0.0622	2.1385	Negative	-0.125	***	0.0413	3.0291
Complexity	-0.035		0.0644	0.5434	Complexity	0.02		0.0592	0.3377
Length	0.142		0.1009	1.4073	Length	0.066		0.0583	1.1311
Ratings	-0.014		0.0498	0.2811	Ratings	0.027		0.0475	0.5686
Budget	-0.012		0.0516	0.2326	Budget	-0.018		0.0487	0.3695
Media	0.382	***	0.0492	7.7601	Media	0.437	***	0.0484	9.0364
Screens	0.179	***	0.0532	3.3646	Screens	0.206	***	0.0543	3.7911
Sequel	0.026		0.0444	0.5861	Sequel	0.044		0.0354	1.2413
R	-0.12	***	0.0455	2.6392	R	-0.078	*	0.0415	1.88
Action	0.053		0.0715	0.741	Action	0.052		0.0563	0.923
Comedy	0.021		0.0631	0.3327	Comedy	0.033		0.0591	0.5583
Drama	-0.021		0.0551	0.3812	Drama	-0.035		0.0496	0.706
Family	-0.02		0.0498	0.4018	Family	0.005		0.0462	0.1083

Dependent Variable	Profit Parameter		SE	T	Dependent Variable	Log of Profit Parameter		SE	T
Positive	0.284	***	0.0595	4.7699	Positive	0.232	***	0.082	2.8304
Negative	-0.254	***	0.0523	4.8589	Negative	-0.254	***	0.0499	5.091
Complexity	0.041		0.0599	0.6843	Complexity	-0.043		0.0625	0.6881
Length	0.116		0.079	1.4681	Length	-0.068		0.0596	1.1405
Ratings	-0.005		0.0593	0.0843	Ratings	0.101	*	0.0607	1.6648
Budget	-		-	-	Budget	-		-	-
Media	0.209	***	0.056	3.7329	Media	0.061		0.0626	0.9752
Screens	0.015		0.0605	0.248	Screens	0.053		0.0654	0.8104
Sequel	0.037		0.055	0.6724	Sequel	-0.032		0.0934	0.3428
R	-0.079		0.0587	1.3448	R	0.033		0.0622	0.5302
Action	-0.123		0.0904	1.3609	Action	-0.122		0.0839	1.454
Comedy	-0.052		0.0796	0.6534	Comedy	-0.004		0.0874	0.0458
Drama	-0.128	*	0.0693	1.8465	Drama	-0.091		0.0697	1.305
Family	-0.082		0.0627	1.3078	Family	-0.025		0.0656	0.3811

Dependent Variable	Return Parameter		SE	T	Dependent Variable	Log of Return Parameter		SE	T
Positive	0.275	***	0.0646	4.2575	Positive	0.321	***	0.0608	5.2796
Negative	-0.36	***	0.0555	6.492	Negative	-0.351	***	0.0559	6.2824
Complexity	0.017		0.0554	0.3067	Complexity	-0.008		0.0559	0.1432
Length	0.046		0.0651	0.7061	Length	0.037		0.0583	0.6342
Ratings	0.003		0.0591	0.0507	Ratings	0.021		0.0597	0.3517
Budget	-		-	-	Budget	-		-	-
Media	0.158	*	0.0881	1.7927	Media	0.157	**	0.0674	2.3277
Screens	-0.183	**	0.0734	2.4941	Screens	-0.135	**	0.064	2.1083
Sequel	0.028		0.0418	0.67	Sequel	0.044		0.0432	1.0175
R	0.062		0.057	1.0872	R	0.02		0.0534	0.3748

Dependent Variable	Return Parameter		SE	T	Dependent Variable	Log of Return Parameter		SE	T
Action	-0.189	**	0.0867	2.1799	Action	-0.15	*	0.0829	1.8103
Comedy	-0.117		0.0876	1.335	Comedy	-0.1		0.0806	1.2407
Drama	-0.24	***	0.079	3.0397	Drama	-0.203	***	0.0728	2.7872
Family	-0.121	*	0.0655	1.8466	Family	-0.093		0.0621	1.4974

† T is computed as the absolute value of the ratio between the parameter estimate and its standard error. .*** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test. Budget is not used with profits and return as budget is algebraically related to them.

In reviewing the structural effects, we see that there are two conceptually distinct sets of variables that affect box office revenues. On one hand, marketing efforts have a large effect on movie box office as measured by media spending (media) and distribution coverage (screens). Both variables have positive impacts on ticket sales, and as a set, it has the largest impact on the dependent variable. On the other hand, information about the content of the movie has an important effect on the sales of the movie. In my model, information about the movie (product) is transmitted through at least two different channels: critics, with the variable ratings, positive comments in the average review (positive) and negative comments in the average review (negative), and MPAA ratings (R for rated R). It is of particular interest that when the content and structure of reviews are included the effect of movie ratings disappears from the model, $T_{ratings}=0.437$ $p>0.1$. It may be that this cancellation is simply a shift in variance explained from one variable to the other as one ceases to be significant when others are entered.[19] However, the inclusion of the content variables and text structure, $F_{4, 145}=7.25$ $p<0.01$[20], increases the variance explained significantly beyond that captured by ratings. This fact points to the possibility that the simple rating measure or

[19] Note, however, that this effect is somewhat present as in the model that does not include content or structure variables ratings is a significant positive predictor of movie box office sales.

[20] Note that I compute the degrees of freedom for PLS models in a conservative way. PLS is a two-stage model: the first part is the measurement I which weights are created to compute scores, and the second part is a structure in which the scores regress on each other. I then compute the degrees of freedom as $df=n-k_1-k_2$, where n is the sample size and k_1 and k_2 are the parameters used to fit the first and second stages of the model.

a categorization (into positive, mixed, and negative reviews) does not capture all the potentially relevant information to assess the influence of critics on movie box office revenues.

Table 18
PLS Models with Movie Performance as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Testing for Quadratic Effects in Content and Structure)

Dependent variable	R^2 quadratic	R^2 linear	AIC_c quadratic	AIC_c linear	Delta R^2	Delta df	df quadratic	F statistic	P value
Box office	0.635	0.625	238.798	218.675	0.010	4	141	0.966	0.428
Log of box office	0.704	0.701	216.867	194.970	0.003	4	141	0.357	0.839
Profit	0.427	0.403	280.141	261.803	0.024	4	142	1.487	0.209
Log of profit	0.271	0.257	305.343	284.701	0.014	4	142	0.682	0.606
Return	0.431	0.379	279.408	265.928	0.052	4	142	3.244	0.014
Log of return	0.456	0.429	274.705	257.142	0.027	4	142	1.762	0.140

As in the previous section, I also fitted models in which the independent variables have quadratic effects on the sales figures. The results for this and other models, including quadratic terms, are shown in Table 18. The rationale behind the quadratic effects is that the probability of observing positive or negative comments in a review may have a nonlinear effect on the reader's decision to patronize the movie. To see whether there is evidence of improvement in fit when effects are allowed to be nonlinear, and given that the linear effects model is nested in the quadratic effects model, I computed F tests to see if the improvement in fit is indeed significant. Results for these tests are provided in Table 18. Note that in the case of box office sales there is no need to include the quadratic terms, and, hence, I did not interpret these models.

Table 19
PLS Model with Movie Performance as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Structural Effects for Different Performance Metrics with Quadratic Effects for Content and Structure)

Dependent Variable	Box office Parameter		SE	T†	Dependent Variable	Log of box office Parameter		SE	T
Positive	0.217	***	0.0658	3.2956	Positive	0.219	***	0.0633	3.46
Positive2	0.038		0.0762	0.4989	Positive2	-0.026		0.0516	0.504
Negative	-0.158	***	0.0548	2.8818	Negative	-0.112	**	0.0474	2.3618
Negative2	-0.001		0.0641	0.0156	Negative2	-0.056		0.047	1.1922
Complexity	0.047		0.1011	0.4647	Complexity	0.018		0.0722	0.2491
Complexity2	-0.006		0.0531	0.1129	Complexity2	-0.013		0.0404	0.3219
Length	-0.102		0.1577	0.6467	Length	0.025		0.0997	0.2506
Length2	0.176		0.1337	1.3168	Length2	0.046		0.0637	0.7217
Ratings	0.026		0.0569	0.457	Ratings	0.022		0.0504	0.4369
Budget	0.002		0.0527	0.0379	Budget	-0.022		0.0489	0.4495
Media	0.386	***	0.0496	7.7865	Media	0.437	***	0.0459	9.5177
Screens	0.19	***	0.0575	3.3042	Screens	0.207	***	0.0518	3.9958
Sequel	0.039		0.0437	0.8923	Sequel	0.04		0.0334	1.1969
R	-0.117	**	0.0494	2.3687	R	-0.08	**	0.0397	2.0153
Action	0.027		0.0696	0.3877	Action	0.044		0.0553	0.7957
Comedy	0.008		0.0673	0.1188	Comedy	0.029		0.0586	0.4952
Drama	-0.043		0.0565	0.7615	Drama	-0.044		0.0498	0.8831
Family	-0.035		0.0574	0.6099	Family	-0.004		0.0494	0.081

Dependent Variable	Profit Parameter		SE	T	Dependent Variable	Log of Profit Parameter		SE	T
Positive	0.218	***	0.0772	2.8229	Positive	0.197	**	0.0772	2.5517
Positive2	0.117	*	0.0685	1.7078	Positive2	0.097		0.0839	1.1557
Negative	-0.254	***	0.0602	4.219	Negative	-0.233	***	0.0604	3.8583
Negative2	0.005		0.0624	0.0802	Negative2	-0.002		0.0659	0.0303
Complexity	0.034		0.0672	0.5058	Complexity	-0.048		0.0759	0.6324
Complexity2	-0.079		0.0501	1.5764	Complexity2	-0.068		0.0502	1.3534
Length	-0.049		0.1096	0.4469	Length	-0.032		0.0833	0.3842
Length2	0.177		0.1296	1.3653	Length2	0.102		0.1087	0.9381
Ratings	0.027		0.0641	0.421	Ratings	0.061		0.0573	1.0642
Budget	-		-	-	Budget	-		-	-
Media	0.207	***	0.0581	3.5636	Media	0.049		0.0593	0.8258
Screens	0.026		0.0647	0.402	Screens	0.026		0.0581	0.4472
Sequel	0.053		0.053	0.9995	Sequel	-0.041		0.0965	0.4248
R	-0.056		0.0589	0.9511	R	0.034		0.0614	0.5537
Action	-0.135		0.0908	1.4869	Action	-0.111		0.0834	1.3314
Comedy	-0.05		0.0864	0.5788	Comedy	0.004		0.09	0.0444
Drama	-0.119		0.0729	1.633	Drama	-0.081		0.0758	1.0684
Family	-0.092		0.0641	1.4362	Family	-0.01		0.0731	0.1368
Positive	0.218	***	0.0772	2.8229	Positive	0.197	**	0.0772	2.5517
Positive2	0.117	*	0.0685	1.7078	Positive2	0.097		0.0839	1.1557
Negative	-0.254	***	0.0602	4.219	Negative	-0.233	***	0.0604	3.8583

Dependent Variable	Return Parameter		SE	T	Dependent Variable	Log of Return Parameter		SE	T
Positive	0.178	**	0.0686	2.5957	Positive	0.232	***	0.0672	3.4534
Positive2	0.252	***	0.0637	3.9572	Positive2	0.186	***	0.0628	2.9633
Negative	-0.3	***	0.061	4.9209	Negative	-0.312	***	0.0599	5.2087
Negative2	-0.069		0.0612	1.1283	Negative2	-0.049		0.0603	0.8129
Complexity	0.099		0.0917	1.079	Complexity	0.032		0.0756	0.4232
Complexity2	-0.079		0.0689	1.1468	Complexity2	-0.046		0.0684	0.6722
Length	-0.038		0.0763	0.4981	Length	-0.033		0.0823	0.401
Length2	0.092		0.0725	1.2693	Length2	0.075		0.0757	0.9904
Ratings	0.003		0.0558	0.0537	Ratings	0.029		0.0578	0.5014
Budget	-		-	-	Budget	-		-	-
Media	0.149	*	0.0893	1.6691	Media	0.15	**	0.0666	2.2517
Screens	-0.193	**	0.0744	2.5948	Screens	-0.128	**	0.0632	2.0261
Sequel	0.037		0.042	0.8818	Sequel	0.052		0.0421	1.2352
R	0.076		0.0573	1.3264	R	0.032		0.055	0.5821
Action	-0.213	**	0.0902	2.3602	Action	-0.172	**	0.0851	2.0205
Comedy	-0.106		0.0923	1.1479	Comedy	-0.096		0.085	1.1295
Drama	-0.253	***	0.0775	3.264	Drama	-0.214	***	0.0745	2.8706
Family	-0.172	**	0.0684	2.5131	Family	-0.127	**	0.0636	1.9956

† T is computed as the absolute value of the ratio between the parameter estimate and its standard error. *** *p* value<.01, ** *p* value<.05, * *p* value<0.1 for a two-tail test. Budget is not used with profits and return since budget is algebraically related to them.

Next, I considered the effect that critics have on movie gross profits (measured as box office revenues minus budget). Arguably, movie gross profit is a more managerially relevant metric than theatrical sales. The results for the main effects (linear) models using both profit and its logarithm[21] are provided in Table 17. Given the fit information, I focus my comments on the raw profit model. In checking the structural effects, we see, again, that the same two conceptually distinct sets of effects are present among the significant predictors of profitability of the movie. It is interesting to note that the balance of importance shifts from marketing variables to information regarding the content of the movie. This implies, not surprisingly, that if the movie is not liked by reviewers, either because of their influencing role (Basuroy, Chatterjee, & Ravid, 2003) or because of their representativeness as potential patrons (Eliashberg & Shugan, 1997), the movie will not be

[21] Given that some of the movies do indeed lose money and the log of nonpositive numbers is not defined, I took the log of the gross profits-Min(profit)+1, where the minimum is computed across all movies and is hence constant after it is determined for this dataset.

profitable. I still found evidence that media spending positively influences profitability. I also found that the number of screens does not seem to have any influence in the profitability of the movie although it did help in increasing the movie's sales. I found in this case that the content in the critics' reviews is particularly important for profitability and that the dummy for rated R is not significant anymore. This last fact is interesting because it shows that while a restricted MPAA rating limits the overall market potential (some people who may be interested in attending the theater are not allowed to do so), the number of moviegoers in the potential market actually persuaded appears to be independent of the restriction, and hence profitability is not affected by it.

I replicated a similar test to see if again the introduction of content and structure added significant amount of explained variance over ratings alone. I found again that the addition of the direct information from the reviews, content, and structure significantly added to the variance explained by the model, $F_{4,147}=7.79$ $p<0.01$. I also found that this variance comes from two sources. First, a variable that was significant before the introduction of content and structure becomes nonsignificant, $T_{ratings}=0.437$ $p>0.1$, with variables in the model and $T_{ratings}=4.093$ $p<0.01$, with no review information in it. Second, there is a significant gain in R^2 as I have already showed, and hence there is additional insight in the content since ratings are still in the model.

I also analyzed regression models in which the independent variables allow for nonlinear effects through higher order polynomials of content and structure variables. The results for these models are shown in Table 19. To evaluate whether there is evidence of improvement in fit when effects are allowed to be nonlinear, I used an F test. As can be seen in Table 18, there is no evidence that quadratic effects are present.

Finally, I looked at the third movie performance metric-return on budget. The dependent variable is defined as gross profit divided by the movie budget. The first step is to decide whether

the logarithm or the variable in levels will be used for analysis. Reviewing Table 18, we notice that the decision cannot be made as before by comparing R^2, or performing a simple F test to see whether the quadratic or the linear model should be interpreted. This is because the F test yields different results for the variable in levels (where the quadratic model is preferred) and with the logarithm of the variable (where the model in levels is favored). This implies that we should compare the fit of the model with raw returns and quadratic terms to that of the log of returns and linear terms. While it was acceptable to compare fit for models having log and level variables before, when the F test yields the same conclusion and hence the independent variables in the model are the same, it is not so. Therefore, I used a variation of AIC to determine the best model in this circumstance. Given that AIC's calculation is based on asymptotic approximations, it is only valid for large sample sizes. Given the sample size in the calibration dataset, this is not an issue. However, the sample size for this dataset is far from what will be reasonable to invoke asymptotic properties. To this end, I used a finite sample correction for AIC. The finite sample correction of AIC, denoted as AIC_c, is computed as $AIC_c = n \ln(RSS) + 2m + {2m(m+1)}/{(n-m-1)}$ (cf., Sugiura, 1978) and is recommended when $n/m<40$ (Burnham & Anderson, 2002, p. 445). In this case, I report values for AIC_c for all the estimated models in Table 18. Note that using this criterion is straightforward; the model with the lowest value of AIC_c is selected. In this case, I selected the log of the return model with linear specification.

The structural parameters in the case of financial return demonstrate that the trend initiated in the profit model accentuates in the case of return, in which content plays even a more important role as measured by the absolute size of the coefficients. I also note that while both marketing efforts are significant, the effect of screens is negative, implying that an increase in the number of

screens when the remaining factors are kept constant will lead to decreasing financial returns. As I did with the other two metrics, I tested whether the content and structure variables, when added to the model, added significant amounts of explained variance. In this case also, I found that the set improves fit significantly with $F_{4,146}=15.53$, $p<0.01$. Similar to the other case, ratings are also not significant with content in the model, $T_{ratings}=0.352$, $p>0.1$. However, in the model in which content is left out, the rating is significant, $T_{ratings} = 3.698$, $p<0.01$. Thus, there is clear evidence that the content in the reviews overlaps in variance with ratings, but there is evidence of additional information that can be used to predict movie performance.

As a last step in the modeling program, I investigated the possibility that there may be interaction effects among the constructs studied. I have argued that it is plausible that the two content variables may interact. This may happen because when positive comments are present in the absence of negative ones, the positive comments are likely to have a larger effect than when they appear alongside negative comments. This suggests an interaction between the positive and the negative latent constructs. I also looked for potential interaction between the content of the review and the structure. Longer reviews may create stronger changes in attitude because of the increase in exposure to the content. More complex reviews may cause weaker effects as they become more difficult to be processed. Finally, I hypothesized a potential interactive effect of media spending and content. The rationale for this effect is that as the amount of dollars spent on a movie (for example, in advertising) increases, it will have at least two potential effects. One is to convince customers that the movie is good enough to induce purchase. The other potential positive effect of advertising is increased interest and search for additional information to make a decision. If this second effect is important, we may see that the effect of content is dependent on the media effort and hence there is a potential interaction between content and marketing effort.

To test these interaction effects, I followed a two-step procedure. I first computed PLS-based scores from the respective linear models.[22] In the second step, I created product terms (variables are mean centered) to capture the interaction, and these terms were used to estimate regression models that test the effect.[23]

Table 20 shows the results from testing several models against simple effects to capture the interaction effects. First, I tested for each of the three dependent variables that were chosen as best fitting in the linear model in which all the hypothesized interactions are entered simultaneously (ALL interactions). As shown in the Table, the model is not supported in any of the three, that is, according to the F test, there are no significant amounts of variance explained in movie performance. I also looked at all content and structure interactions as a set (including product terms for both positive and negative and both length and complexity (named P Value contentXStructure in the table). I did not find support for these models either. Next, I looked at the possibility that positive content and negative content in the review interact. While there is no evidence of this interaction effect in the case of profit and box office revenue, I found that when return is considered there is a significant interaction effect as measured by both the F test and the T test (computed from bootstrapping, $F_{1,146}$=2.848, $p<0.1$ and $T_{PosXNeg}$ = 2.593 $p<0.05$). The negative interaction effect indicates that as the number of negative comments in the review increases, the positive effect that the positive comments found in the same review have on financial return is mitigated or weakens.

[22] Note given that the dependent variable changes the PLS weights that are used for the creation of the scores will also change, requiring computation of scores for each model.
[23] We still use bootstrap to obtain standard errors to ensure that the results are robust and comparable to those obtained using a full-fledged PLS model.

Table 20
PLS Models with Movie Performance as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Testing Interaction Effects among Content, Structure, and Media Effort)

R Square

Dependent variable	R^2 Linear Model	R^2 ALL interactions	R^2 contentXStructure	R^2 PosXNeg	R^2 ComplexityXContent	R^2 LengthXContent	R^2 MediaXContent
LBO	0.701	0.705	0.704	0.704	0.701	0.703	0.705
Df	145	138	141	144	143	143	143
Profit	0.403	0.438	0.418	0.412	0.409	0.418	0.434
LROI	0.429	0.449	0.434	0.440	0.429	0.434	0.434
Df Model	146	139	142	145	144	144	144
Delta df		7	4	1	2	2	2

F Statistics

Dependent variable	F Linear	F ALL interactions	F contentXStructure	F PosXNeg	F ComplexityXContent	F LengthXContent	F MediaXContent
LBO	-	0.267	0.357	1.459	0.000	0.481	0.969
Profit	-	1.237	0.915	2.219	0.731	1.856	3.810
LROI	-	0.721	0.314	2.848	0.000	0.636	0.636

Pvalues

Dependent variable	P Value Linear	P Value ALL interactions	P Value contentXStructure	P Value PosXNeg	P Value ComplexityXContent	P Value LengthXContent	P Value MediaXContent
LBO	-	0.966	0.839	0.229	1.000	0.619	0.382
Profit	-	0.287	0.457	0.138	0.483	0.160	0.024
LROI	-	0.655	0.869	0.094	1.000	0.531	0.531

Next, I looked at the potential content-structure interaction, but in separate sets according to structure. I first investigated a potential interaction of content measures and complexity. I found no evidence for this interaction. Next, I looked at length, and again I failed to identify any interaction effects with content. Finally, I looked at the interaction of content and media spending. I found that there are some interaction effects between these two variables when predicting gross profits, $F_{2,146}=3.810$, $p<0.05$. I found that the effect that negative review comments have on return is accentuated by increases in media spending, $T_{NegXMedia} =2.298$, $P<0.05$.

Previously, I observed the negative simple effect of screens on performance. I now estimated a new interaction effect between media spending and number of screens. This effect is

substantiated by the fact that after advertising heavily, the positive effect of advertising will not translate to profitability unless the movie is widely distributed. After testing this potential effect with all the three performance metrics, I found that there is some evidence of the interactive nature of the two marketing mix variables. I found that in all cases the product term of media expenses and distribution is positive and significant at least at 10% using a bootstrap-based t statistic for the product term. This implies that while the simple effect of screens may be zero or even negative, when I account for media spending higher distribution efforts pay off.

Table 21
PLS Model with Movie Performance as Dependent Variable and Content, Structure, Marketing Effort, and Controls (Structural Effects for Different Performance Metrics with Interaction Effects among Content, Structure, and Media Effort)

Dependent Variable	Profit Parameter		SE	T†	Dependent Variable	Log of Return Parameter	Variable	SE	T
Positive	0.263	***	0.060	4.424	Positive	0.312	***	0.064	4.890
Negative	-0.238	***	0.059	4.050	Negative	-0.363	***	0.055	6.626
Complexity	0.039		0.049	0.791	Complexity	0.018		0.049	0.368
Length	0.119		0.099	1.200	Length	0.041		0.062	0.665
Ratings	0.001		0.063	0.016	Ratings	0.042		0.065	0.642
Budget	-		-	-	Budget	-		-	-
Media	0.158	**	0.065	2.446	Media	0.151	**	0.072	2.108
Screens	0.018		0.069	0.262	Screens	-0.128	*	0.074	1.737
Sequel	0.037		0.061	0.604	Sequel	0.047		0.048	0.989
R	-0.084		0.065	1.296	R	0.003		0.065	0.046
Action	-0.122		0.101	1.207	Action	-0.164		0.101	1.620
Comedy	-0.041		0.091	0.451	Comedy	-0.113		0.101	1.116
Drama	-0.124		0.081	1.536	Drama	-0.231	**	0.092	2.513
Family	-0.122		0.101	1.207	Family	-0.164		0.101	1.620
PosXNeg	-		-	-	PosXNeg	-0.111	**	0.043	2.593
PosXComplexity	-		-	-	PosXComplexity	-		-	-
NegXComplexity	-		-	-	NegXComplexity	-		-	-
PosXLength	-		-	-	PosXLength	-		-	-
NegXLength	-		-	-	NegXLength	-		-	-
PosXMedia	0.021		0.092	0.229	PosXMedia	-		-	-
NegXMedia	-0.169	**	0.074	2.298	NegXMedia	-		-	-

† T is computed as the absolute value of the ratio between the parameter estimate and its standard error. *** p value<.01, ** p value<.05, * p value<0.1 for a two-tail test. Budget is not used with profits and return since budget is algebraically related to them.

It is clear from the results, even while maintaining the rating variable in the equation, that the text-based constructs are significant and add to the explained variance in the entire set of movie performance metrics. It is notable that my results suggest that the role of content in the reviews becomes increasingly important as we move from sales to financial gain metrics (gross profit, ROI). That is, advertising and other marketing efforts are usually successful in increasing sales. However, this influence comes at an obvious cost in the case of distribution. When the bottom line is also considered, the effect of the content of the reviews becomes prevalent (and it is also harder to predict movie success as it can be seen by the drop in variance explained). This quote from a review penned by a moviegoer (not used in the analysis) summarizes the idea, "Publicity got me to the theatre. Advice will take you away from this waste of time. Very bad everything." ("Consumer review").

This section shows unequivocally that the information content of reviews is a) relevant to the prediction of new product success and b) relatively untapped, in that there are resources in the reviews that have not been fully used because of the lack of methods that pertain to textual information analysis. In the following section, I summarize the primary contributions of this research study and provide guidelines for future research in this arena.

CHAPTER FIVE: DISCUSSION AND FUTURE RESEARCH

Discussion

In any organization, the most powerful competitive weapons managers have access to are information acquisition, dissemination, and use. This is particularly true in competitive environments where innovation is likely a key driving competitive force (Baumol, 2001). In such cases, obtaining accurate and relevant information in a timely manner is critical to the success of the firm.

Moreover, this is true not only for managers and businesses but also for researchers and academicians who are ultimately in the business of knowledge creation and dissemination. New sources of information are critical for improving and expanding research. It is my contention that while there are exciting methodologies in the making that allow researchers to extract more information from numeric data (e.g., neural networks, support vector machines), the real breakthroughs will come from analyzing nonnumerical data.

The amount of digital, nonnumerical information to which researchers have access has grown exponentially with the advent of the computer (e.g., databases such as Lexis Nexis) and the Internet. In addition, with the associated advances in computing power and speed, new methodologies are now possible that were difficult, if not impossible, to implement a few decades ago. The interaction of these two factors has created a window of opportunity that I consider to be critical for development of the technology and methods that use information available in nonnumerical (text) form.

Advances in technology facilitate the use of nonnumerical data by marketing researchers during all stages of the research process. The process begins with the capture, collection, and

organization of raw text pertinent to the particular marketing problem and then moves to the subsequent processing and analyzing stages. The collection is enabled by new technologies such as OCR, digital voice recording, and transcription software (helped by advances in Natural Language Processing), digital databases, the Internet, and other distributed networks. These and other advancements allow users to have ready access to large amounts of information in textual form with low processing time. This is important because the cost of making text available in machine readable form was one of the deterrents to the use of nonnumeric information in the past (de Sola Pool, 1959; Miller, 1995). In this study, I used the Internet and general purpose software (Adobe Acrobat) to collect thousands of documents from a remote database. This was accomplished with minimal investment of time involved in the data collection process. The process proceeded with the organization and cleaning of the data in which software and computers eased the painstaking job of sorting and storing thousands of records in relational databases where queries can readily be answered. Finally, after the data were preprocessed, the bulk of the analysis was facilitated by advances in computing power, software, and techniques developed in other fields (e.g., statistics and artificial intelligence).

In this study, I proposed a methodology that departs from the classical philosophy associated with content analysis in which words phrases and expressions are categorized in sets to be counted later (Pooping, 2000). I integrated concepts from the information literature in general, and specifically, latent semantic indexing and content analysis to propose a different approach to analyzing textual information. I moved from the aforementioned classification approach to an approach in which terms in the text are weighted according to their inferred meaning. Meaning is inferred in this method by the collocation of the documents across texts. It is assumed that there is an underlying lower dimensional space of concepts that underlies word usage. Information is

obtained about word semantic similarity by observing the documents in which a given word appears and those in which it does not. After the semantic similarity space was inferred from the data, the words in each document were weighted to obtain its representation in the lower dimensional concept space. This simplified the need to create ad hoc dictionaries to classify words in the dictionary categories. Using a vector-based method, I began with a seed word and computed a variate that incorporated all the information in the text that was semantically similar or dissimilar to that particular word.

I demonstrated the application of this methodology and traditional computer-aided content analytic methods to the study of an important marketing topic, the effect of movie critic reviews on film performance. In my empirical application, I used two datasets that, combined, contain more than 9,000 movie reviews. It is noteworthy that the amount of work involved in manually hand-coding this volume of text is prohibitive, even using software that facilitates the hand-coding process (e.g., NUDIST). I studied this marketing problem in light of directly obtaining information from the reviews instead of using an overall rating or a classification of the review as either positive or negative.

It is my contention that this particular research topic and others can benefit significantly from a more thorough analysis of nonnumeric data. It was demonstrated in this case that the numeric measures most frequently used by researchers in this arena do not capture all the information in the critics' reviews, and hence may underestimate the effect that reviewers have on movie performance. To do this, I first tested the validity of the three proposed methods for extracting information from text: a) the creation of an specific dictionary to categorize words and expressions into meaningful categories, b) the use of a general purpose dictionary to categorize words, and c) the use of a vector space method that uses ALSA to facilitate the creation of a score(s)

that captures the essence of each word. Within these three broad categories, I used different operationalizations to measure the effect of the content on the overall rating provided by the reviewer. The operationalizations refer to the way the observed frequencies (marginal probabilities) within each class in the dictionary are combined to create the measure of the construct of interest (attitude of the reviewer in the movie example). In particular, I analyzed individual words treated as independent entities, the classic summated scale, a principal component-based factor scores approach, PLS-based scores, and, finally, for the ALSA-based model, I computed the scores for weighted probabilities of observing all the words in the document.

I found that all the operationalizations were not equally efficient in capturing the content in the reviews. Table 22 summarizes the efficacy of the different textual approaches and their different operationalizations. We see that if the ad hoc dictionary is used, the summated scale provides the least efficient means to capture the information since it assumes that all words in each set are equally important. The PCA factor scores are more efficient than the summated scales because the words receive different weightings according to the covariation. However, it does not use information about ratings to determine the weights. PLS-based weights have better predictive properties since they explicitly use the dependent variable in determining the optimality of the weights used to combine the original variables. Finally, the ALSA-based method performs similarly to the PLS-based weighting, although no information about the dependent variable was used. Its advantage emanates from two fronts: a) it uses the information in all words and b) it taps into the similarity in semantic meaning and enhances the information only using words that are closely related to the seed, hence reducing the amount of error. This last advantage is made clear when I compared the results obtained with the ad hoc dictionary (containing only 77 terms) with those obtained when the general-purpose-dictionary-based categories Positive and Negative are used (containing more than

4,000 terms combined). When using the general purpose dictionary, I observed that the more advanced weights (PCA and PLS) perform poorly for different reasons. On one hand, PCA cannot extract a meaningful component from a large set of variables that has little communality. On the other hand, PLS has too much information to predict and overfits the data. ALSA overcomes this problem because it successfully weights 5 times more information without falling into either one of these extremes.

Table 22
Summary of Results on the Different Methods to Extract Quantitative Information from Text

Data type	Method	Operationalization	Advantage(s)	Disadvantage(s)	Performance
Disaggregated (review level)	Ad hoc	Individual terms	Max variance explained within method	Wrong parameter signs High collinearity	Good forecasting Poor hypotheses testing
		Summated scale	Simple	Assumes equal weights among terms	Moderate forecasting Moderate hypotheses testing
		PLS-based latents	Treats concepts as underlying latents	Potential to overfit	Good forecasting Good hypotheses testing
	General purpose	Individual terms	Max Variance explained	May not be identified unless number of documents is really high; Wrong signs; May overfit	Uncertain forecasting Very poor hypothesis testing
		Summated scale	Simple	Assumes equal weights among terms; Introduces excess noise	Poor forecasting Poor hypotheses testing
		PLS-based latents	Treats concepts as underlying latents	Potential to overfit	Good forecasting Uncertain hypotheses testing
	ALSA		No need to create dictionary Uses all information in text	Needs training of data; potential multicollinearity if used for multiple constructs	Good forecasting Good hypotheses testing

Finally, in this study, I investigated the effect that content of the critics' reviews had on movie performance. I found that the best predictive validity of the data was obtained from the PLS based ad hoc dictionary, and therefore I used this operationalization of content in the evaluation of my research question. I found that the addition of content and structure of the review added significant amounts of explanatory power, even in the presence of controls and the ratings. This effect is robust across operationalizations of the performance metrics. In fact, I found that as we move from sales to financial return measures the role of the content of the review, and therefore the critic's role, becomes increasingly important.

I have shown in this study that the use of direct content to evaluate the effect of critics can shed new light on the important role that reviews play in the movie marketplace. We see that the content-based measures have more explanatory power and hence may uncover insights that were hidden with other indirect measures of the review content.

Limitations of the Study

As is true in virtually all research, this research has limitations. I found serious multicollinearity problems when using the proposed ALSA-based method with multiple vectors included in a single model. This problem is surmountable if the sample size in the study is large enough, as in the case of our validity test; however, with a sample size of approximately 240 observations, multicollinearity makes my estimates imprecise and prevents me from exploring the methodology further.

Another concern stems from the fact that I did not have access to weekly advertising figures. This prevented me from studying the critic review—movie performance relationship in a causal

manner. Of course, this precludes me from providing some additional insight about the role of critics.

What Lies Ahead: Direction for Future Research

On the substantive side, there are multiple applications where these metrics can help researchers better understand the world. Important marketing problems such as the effect of word-of-mouth communication can also be investigated using these methods. Other less obvious areas such as managerial decision making can benefit from the possibility of mapping mental representations using textual information in protocols and descriptions (Palmquist, Carley, & Dale, 1997).

While i have shown that the outlined methodologies do a good job at capturing movie review content, there are several additional questions that could be answered within the same area of study. The main strength of this methodology is its ability to handle textual information quickly when facing large amounts of text. In the case of movies, and given the short span of the life cycle of any given movie, having a quick uptake on the reaction of reviewers to the movie, these insights can be used to adjust the marketing effort accordingly over the few weeks that the movie is distributed on theaters.

Another example of an application that may arise using this and other text-related methods is the discovery of optimal product creation attributes based on insights from reviews. In this example, I could extract a set of content attributes that describe optimal movie features and themes so that new movies can be created that use this information in the process.

On the methodological side, there are many exciting developments and refinements that can be adapted to the methodology proposed in this paper. Gains can be made by increasing the number

of contextual units used to estimate the ALSA conceptual space. The gain in contextual units can come from two fronts. One way is to increase the number of reviews in the sample. The less obvious way is by using other textual units instead of the document. We could code the text at the paragraph or sentence level to try to improve our accuracy of the ALSA space formation. There is an important tradeoff in this adaptation; as we zoom in, we obtain more contextual units (more data points), but the matrix of observations becomes more and more sparse (there are large numbers of zeros). The investigation of the optimal level of analysis for this type of methodology can increase the effectiveness of the technique in future applications.

Other refinements to the methodologies included here could be the use of stemming words to aggregate same root words instead of treating them as separate entities. Doing this may increase the precision of the probability (frequency) estimates in the raw data (cf., Hull, 1996). Stemming is another methodology that emerged from the information retrieval to improve precision as researchers noted that the particular form of the word used was usually not critical, but that its stem or root was the content-bearing part. There are a number of stemming algorithms that have been proposed to conduct this task, and their efficiency varies. The use of some of these methods to improve the quality of the data prior to analysis may increase the quality of the obtained measures.

I have also suggested the application of techniques that require the use of seed words to either construct a dictionary or create a variate that weights the words in the documents according to the similarity with the seed term. Methods that allow the identification of good seed words should be of great value. There are some developments in different fields that promise exciting advancements in this aspect of the methodology. Work conducted by Corman, Kuhn, McPhee, and Dooley (2002) using neural nets allowed the researcher to select words that have relevance in the message. Another promising area is the study of DNA, in which genes are arrays of tokens for

which we do not know the syntax or the meaning. Methods are being created to identify relevant sequences within this unknown language. The potential use of these techniques for finding relevant words is promising. Appendix A provides additional detailed information on this topic.

Other refinements of the method are possible. Word sense disambiguation algorithms, similar to those built into the General Inquirer, can be used to separate words that have the same spelling but conceptually different meanings. We could also use syntactical information to fine-tune the information available in the words. While this last possibility seems promising, there are not fully automatic reliable solutions for syntax parsing, though advancements are taking place in this field. In this study, I have proposed methodologies that can be used to move from text to numbers in a consistent and scalable way, allowing for the processing of information in thousands of documents. While I am aware that my study only begins to scratch the surface, I hope that this and other efforts will be the initiators of a stream of work in marketing and other disciplines to allow researchers the efficient use of textual and other nonnumeric data to gain insights and to conduct hypothesis testing.

APPENDIX : EMPIRICAL-BASED MODELS OF DICTIONARY BUILDING

I report here on attempts to construct a dictionary from the information contained in the reviews themselves and information regarding their overall valence and whether the movie reviewed succeeds. The first way of assigning words to categories is by using either the frequency for each word or the marginal probabilities as described previously, and information related to whether the movie could be considered a success or a failure. We then need a definition for failure and success for this particular case. A movie was considered successful if it stayed in the theaters eight weeks or longer (Eliashberg & Shugan, 1997), and the per-screen box office revenues were greater than $30,000.[24] A movies was considered a failure if the per screen box office was lower than $7,000. Note that these figures were chosen because they represent the top and bottom 20% in a database composed of more than 3,000 movies. The following sets were created:

Unique words are those that appear rarely or that are specific to a movie. To operationalize this set, I used words that appear only in five movies or less (that is about 2% or less of the movies in the dataset). This category encompasses most of the distinct tokens in the sample (about 70% of the distinct words in the sample). A list is not provided as it is used to "eliminate" words from other sets.

24 This figure was computed as the ratio of the total gross over the number of screens at the widest distribution of the movie. The threshold was obtained looking at the distribution of cumulative box office for more than 3,000 movies and corresponds roughly with the top 20%.

Table 23
Most Frequent Common Words Between Good and Bad Words

Word	Sumfreq*	Word	Sumfreq	Word	Sumfreq
the	741955	An	51062	No	26422
of	366171	Who	49634	If	25428
and	332970	Reviews	49547	Can	25352
to	292712	All	48531	some	25288
in	202470	Be	47095	their	24711
is	195834	One	46645	We	24676
it	129162	Pm	44522	which	24252
that	124152	not	42379	character	23559
for	109331	has	41519	good	23328
with	98022	was	41439	Just	22799
as	97530	have	39199	Time	22708
by	92141	her	38105	Into	22592
this	88699	www	37740	dvd	22496
movie	87149	they	36972	html	21299
on	78413	out	36971	him	21261
his	75471	about	34055	new	21009
film	75242	or	33112	than	20648
he	72591	movies	32796	only	19932
but	70469	there	32610	get	19135
are	61557	up	32238	other	19054
review	60146	more	32036	will	18806
you	58896	like	31359	its	18047
com	56937	so	30328	even	17997
at	56333	when	27839	story	17710
from	53267	what	26820	most	17653
http	51533	she	26709	first	17503

*Sumfreq is the sum of the observed frequency counts for each word across the 243 calibration reviews.

Table 24
Most Frequent Successful Words Based on Frequencies

Word	Sumfreq*	Word	Sumfreq	Word	Sumfreq
scream	4627	low	1853	whom	1663
batman	4205	dvds	1842	talent	1648
George	3597	herself	1830	presented	1645
robin	3278	oscar	1826	ways	1643
mission	2845	supporting	1825	stand	1640
seven	2736	Gary	1815	collection	1638
truth	2548	scary	1808	wild	1632
Disney	2469	widescreen	1803	change	1629
romantic	2388	jokes	1802	reference	1628
Washington	2382	romance	1800	certain	1615
gay	2377	minor	1777	Nicholson	1599
Missouri	2362	brings	1775	asp	1584
ahicks	2340	toy	1753	trailers	1569
independent	2302	offers	1746	liked	1564
hunt	2287	brilliant	1743	plus	1563
voice	2265	third	1732	various	1557
alone	2245	talking	1729	introduced	1549
wonderful	2227	truly	1728	liar	1535
Carrey	2224	hilarious	1726	Eddie	1534
agent	2217	incredible	1719	break	1529
crime	2179	successful	1708	giving	1520
Tim	2051	manages	1701	success	1516
roles	2042	within	1700	laugh	1513
living	2023	theme	1699	party	1511
sexual	2001	physical	1694	writing	1511
impossible	1975	mix	1683	win	1508
secret	1922	girlfriend	1677		
Apollo	1896	emotional	1675		
Howard	1871	premise	1675		
Jurassic	1856	wedding	1674		
laughs	1855	staff	1670		

*Sumfreq is the sum of the observed frequency counts for each word across the 243 calibration reviews.

Good words are those that appear frequently in the set of movies that were previously classified as successful. To operationalize this set, I included words that are in the top 1% of the frequency distribution of all words in the complete sample for so-called successful movies. This amounts to more than 1,200 words.

Bad words are those that appear frequently in the set of movies that were previously classified as unsuccessful. To operationalize this set, I included words that are in the top 1% of the frequency distribution of all words in the complete sample for the unsuccessful movies. This amounts to more than 1,200 words.

Table 25
Most Frequent Unsuccessful Words, Based on Frequencies

Word	sumfreq	Word	sumfreq	Word	sumfreq
Scott	2877	sfgate	1848	explosion	1590
con	2839	festival	1847	giant	1587
boys	2631	political	1819	opens	1585
novel	2562	below	1803	begin	1582
gun	2561	Runs	1799	manager	1580
stone	2425	battle	1788	Stallone	1577
future	2412	query	1777	disaster	1573
Johnny	2337	Jobs	1767	mad	1566
los	2271	investing	1750	familiar	1565
Francisco	2227	middle	1747	crap	1558
starship	2226	Guns	1740	forum	1556
devil	2187	Rich	1724	Nixon	1551
husband	2120	Leads	1715	poor	1544
child	2106	register	1715	ii	1540
fox	2098	Land	1697	opinion	1539
engine	2079	Patrick	1697	critics	1537
troopers	2043	machine	1690	board	1509
escape	2028	Latest	1688	judge	1494
Willis	2013	public	1683	columnist	1478
gore	1973	State	1681	century	1472
water	1961	ratings	1665	shoot	1470
Angeles	1958	Ca	1657	killing	1467
former	1951	Tense	1638		
bay	1924	Sam	1623		
jam	1919	villain	1618		
law	1905	articles	1612		
friendly	1887	Hot	1608		
nudity	1877	Near	1607		
dtl	1851	Cars	1602		
newspaper	1850	obviously	1599		

*Sumfreq is the sum of the observed frequency counts for each word across the 243 calibration reviews.

Common words are those that appear in both the good and the bad words list. These are usually prepositions pronouns and other commonly words needed for basic construction of sentences (see Table 23).

Successful words are good words but are not unique nor are they common. See Table 24 for results.

Unsuccessful words are bad words but are not unique nor are they common. See Table 25 for results.

Discriminat words are those that are either successful or unsuccessful, but not both. Note that this is successful plus unsuccessful, given that they are disjointed.

Table 26
Most Frequent Successful Words, Based on Frequencies

Word	sumfreq	Word	sumfreq
Jack	4677	enjoy	2004
Scream	4627	easy	1985
Batman	4205	al	1979
George	3597	impossible	1975
Robin	3278	uses	1967
Disc	3255	serious	1931
commentary	2853	secret	1922
Mission	2845	Apollo	1896
Aliens	2809	document	1876
Seven	2736	definite	1871
Jim	2701	Howard	1871
Sequel	2555	addition	1858
Truth	2548	Jurassic	1856
Disney	2469	laughs	1855
Chris	2397	dvds	1842
Gay	2377	herself	1830
Missouri	2362	oscar	1826
Ahicks	2340	supporting	1825
Brother	2305	scary	1808
independence	2302	widescreen	1803
Hunt	2287	jokes	1802
Alone	2245	romance	1800
Wonderful	2227	minor	1777
Carrey	2224	brings	1775
Agent	2217	toy	1753
Cage	2204	offers	1746
Crime	2179	simple	1746
Tim	2051	brilliant	1743
Living	2023		

*Sumfreq is the sum of the observed frequency counts for each word across the 243 calibration reviews.

Similarly, instead of using frequencies, standardized frequencies or marginal probabilities are used, that is, accounting for the fact that the number of reviews and length are not constant across movies and reviews. The information in Tables 26 and 27 summarizes the results for successful and unsuccessful words.

Table 27
Most Frequent Unsuccessful Words, Based on Probabilities

Word	sumfreq	Word	sumfreq
cop	3219	Angeles	1958
con	2839	Sci-fi	1954
ship	2793	former	1951
fiction	2677	bay	1924
Moore	2664	jam	1919
police	2653	law	1905
boys	2631	nudity	1877
novel	2562	newspaper	1850
gun	2561	festival	1847
creature	2468	political	1819
van	2426	below	1803
stone	2425	battle	1788
future	2412	query	1777
los	2271	lots	1753
starship	2226	investing	1750
devil	2187	middle	1747
husband	2120	guns	1740
child	2106	rich	1724
troopers	2043	leads	1715
tale	2038	register	1715
escape	2028	female	1705
Willis	2013	land	1697
media	2010	Patrick	1697
camera	1989	UK	1695
gore	1973	latest	1688
water	1961	executive	1675
Sean	1959		

*Sumfreq is the sum of the observed frequency counts for each word across the 243 calibration reviews.

As can be seen in Tables 24 to 27, this approach to obtaining sets of words is not very promising. The main disadvantage is that because of the nature of the classification many of the words are somewhat unique to particular movies.

To obtain better sets or rules to assign words to groups or composites, we need some insight as to which words to select. There are different approaches to obtain the words. I started by collecting a calibration sample of reviews that are not part of the original dataset. Given the

availability of data, I collected a separate sample of 240 reviews that are selected such that they correspond to a balanced 2 by 2 design with factors successful unsuccessful and positive and negative.

Note that we are interested in obtaining words that are useful in discriminating between a) reviews that have a positive and a negative valance (also rating) and b) words that discriminate between successful and unsuccessful movies. Selecting the calibration sample in the proposed way instead of randomly provides the researcher with information that is valuable in the selection on the words and assignment to each of the groups. In a way, this selection allows the researcher to obtain data that have a structure similar to experimental data in that the conditions are known to the researcher; therefore, this information can guide the word selection patterns.

The criteria for selection were simple. The review was considered positive if the site www.Rottentomatoes.com rated the review as a red tomato and negative otherwise. This website has a set of reviewers who get their reviews disseminated through Rottentomatoes.com. For classification, the site uses experts who read the review and classify it as suggesting a good movie, red tomato, or a bad movie, green tomato. The movie was considered successful if it followed the aforementioned pattern.

One way of selecting words is looking at the differences in the marginal distributions of words across the cells given in the table. That is, words are selected such that they have different characteristics: 1) words that have high discriminat power between good and bad reviews, 2) words that have high discriminat power between successful movies, and 3) a combination of the above. There are several ways of conducting this analysis. In the first instance, I used chi-square tests and *t*-tests to determine which words are the ones to be selected.

We can assume that if words are to provide meaningful information they should appear in different proportions across cells. Thus, we can calculate X^2 for each word across the four cells as:

$$\chi^2 = \sum_i \frac{(x_i - E_i)^2}{E_i} \tag{17}$$

I also computed *t*-tests for the difference of proportion for the words across two groups Positive-Negative and Successful-Not Successful movies. When testing the null hypothesis that H_0: $(p_1 - p_2) = 0$ or, equivalently, H_0: $p_1 = p_2$, that is, that the rate of occurrence of a particular token is the same across groups, the best estimate of $p_1 = p_2 = p$ is found by dividing the total number of successes in the combined samples by the total number of observations in the two samples. That is, if x_1 is the number of successes in group 1 which has n_1 observations (e.g., tokens, words…) and x_2 is the number of successes in group 2 out of n_1 observations, then the overall number of successes under the null of no differences is given by:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{18}$$

In this case, the best estimate of the standard deviation of the sampling distribution of the difference between the rates, $D=(p_1 - p_2)$, is found by substituting $\hat{p}$ for both the sample estimates of p_1 and p_2:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{19}$$

We define the statistic for the difference among the two rates as:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{\hat{p}_1 - \hat{p}_2}} \tag{20}$$

Where D_0 is the hypothesized value of the difference and $q=(1-p)$ in our case as we are testing for the null of equal rates D_0 is zero. Z follows a standard normal distribution.

To conduct this procedure, I calculated z for each word using the rates computed for each word according to the categories shown in the table. The resulting words using this statistic to order the words that behave more dissimilarly across groups is given by rank ordering and removing noncontent words. Table 28 shows the most frequent terms. Given their nature, these terms are removed in many cases as they are mainly auxiliary verbs and pronouns. Tables 22, 23, and 24 show the words that exhibit the most differences across the cells as measured by the chi-square statistic. Note how using more information than before and even using a statistical test to rank the most relevant words, we still get many words that are unique to one movie, either in the title or part of the cast and/or characters of the movie.

Table 28
Most Frequently Used Words in the Calibration Sample of Reviews

Word	Frequency	Word	Frequency
the	8173	All	428
of	3628	they	425
and	3548	At	421
to	3160	like	421
in	2178	was	398
is	2159	there	387
it	1587	more	366
that	1551	Up	359
as	1155	when	354
with	1089	Or	345
for	993	So	340
his	941	out	336
this	853	which	323
but	847	about	320
by	806	her	315
he	793	can	303
on	734	time	300
film	666	If	297
you	652	into	297
are	635	their	291
movie	623	than	279
who	618	We	272
be	610	what	272
an	581	some	270
not	514	even	269
from	509	little	266
one	509	just	262
has	468	will	252
have	460		

One potential explanation comes from the nature of the data. Note that for most of the reviews the counts for many potentially important words will be low. For example, how many times does a reviewer actually repeat the word "awful" in a 1,000-word review? Arguably once or twice, at most, should suffice to communicate the overall impression regarding the movie. So if we use chi-square or Z tests when testing these less frequently used words, they will not be picked up by the

test unless huge amounts of text are available. Why is this? Because for the chi-square to have adequate testing power (detecting effects that are truly there), at least five counts per cell are needed for asymptotics to apply. If, similar to most individual reviews, the size of the text is fixed, more accurate tests that do not require the approximation of normality may perform better.

Table 29
Words with Largest Chi-Square Statistic (No Words Removed)

Word	Chi-square	Z_PN	Z_SU	Word	Chi-square	Z_PN	Z_SU
The	2123.78	4.60	-3.32	Genie	132.04	3.75	0.00
And	870.22	2.11	-1.75	Kung	129.65	-2.97	5.18
Of	846.70	1.01	-1.14	Be	128.76	-0.47	0.02
To	644.51	-1.23	0.51	You	126.16	0.32	0.31
Is	605.07	3.15	-1.57	Not	122.75	1.42	-0.07
In	583.87	3.34	-2.02	Holy	112.87	-3.99	5.26
It	384.05	1.17	-0.90	Has	112.45	0.54	-0.10
That	327.14	-0.28	-0.60	Time	112.08	1.43	-1.82
As	277.46	1.51	-0.59	Titanic	108.90	5.35	-3.91
His	263.74	2.17	-1.65	Little	103.53	-0.71	-1.60
He	241.01	2.34	-1.50	Leopold	102.42	2.31	-3.22
with	223.44	-0.15	0.05	From	101.59	-1.31	0.60
For	200.20	0.27	0.29	Pow	100.54	-2.14	4.32
Are	194.43	2.49	-1.57	Patriot	97.79	3.43	-3.42
Stuart	189.27	1.64	-3.95	Alien	97.04	-5.59	5.17
terminator	181.43	8.67	-5.08	Grinch	96.42	2.33	-3.15
Film	179.19	2.35	-0.53	We	96.26	1.12	-1.80
djinn	176.05	4.34	0.00	An	95.55	-2.25	1.83
But	163.17	-0.14	0.45	Will	94.18	1.40	-1.40
who	161.19	1.59	-1.28	One	93.63	-0.49	0.32
species	160.73	-5.88	5.92	More	91.57	-0.07	-0.65
oedekerk	160.50	-4.09	6.07	Ship	90.31	4.45	-3.15
This	159.85	-0.17	0.70	Craven	89.82	2.66	0.27
By	153.39	-0.51	0.34	About	89.67	0.38	-0.74
Cameron	150.48	6.47	-4.62	Toy	88.91	6.02	-3.50
debney	146.71	3.96	0.00	Atkins	88.02	3.07	0.00
wishmaster	146.71	3.96	0.00	Divoff	88.02	3.07	0.00
Murphy	146.31	-5.09	6.17	First	87.78	1.95	-0.99
mummy	139.20	-3.53	-1.52	Movie	86.75	-2.62	1.71
On	138.57	-1.06	1.04				
action	134.30	2.79	-3.19				

Is this in the reference list, and what is the name? Following Dunning (1993), I used a binomial-based LR test to test the same differences. The idea is that every word included in the text (review) is considered a Bernoulli trial, that is, the word that we are looking at is either the target word (and therefore a success) or not (every other word). Note that implicit in the Bernoulli is the

assumption of independence of the probability of a word appearing, given that another word has appeared in the last "trial." This, of course, is not true; one reason for this is semantic and syntactic rules that govern language, and this assumption works well as word dependency (correlation) becomes small rapidly as we move farther and farther from the target word. Under this assumption, we can model each word's marginal frequencies as the results of *n* Bernoulli experiments and therefore each review as a T dimensional Binomial[25] variate, where T is the number of distinct tokens or words in the review (text).

[25] Note that a binomial distribution is equivalent to repeating N Bernoulli independent experiments each with parameter p and counting the total number of successes across the N experiments.

Table 30
Words with Largest Chi-Square Statistic (most frequent (60) words removed)

Word	Chi-square	Z_PN	Z_SU	Word	Chi-square	Z_PN	Z_SU
Stuart	189.27	1.64	-3.95	Lintz	79.63	1.06	1.81
terminator	181.43	8.67	-5.08	Kate	78.95	2.50	-2.47
Djinn	176.05	4.34	0.00	Ii	77.18	-3.99	4.05
species	160.73	-5.88	5.92	War	74.35	1.70	-2.63
oedekerk	160.50	-4.09	6.07	Fang	73.35	2.80	0.00
Cameron	150.48	6.47	-4.62	Fist	73.05	-1.91	3.73
debney	146.71	3.96	0.00	Toys	72.53	5.73	-3.16
wishmaster	146.71	3.96	0.00	Horror	72.38	3.16	-0.30
murphy	146.31	-5.09	6.17	Man	72.16	-1.33	3.04
mummy	139.20	-3.53	-1.52	Ricky	71.41	-2.23	3.85
action	134.30	2.79	-3.19	Eddie	70.73	-2.55	3.97
Genie	132.04	3.75	0.00	Sarah	69.83	5.58	-3.05
Kung	129.65	-2.97	5.18	Special	69.17	3.06	-2.19
Holy	112.87	-3.99	5.26	Martin	69.03	1.65	-2.19
titanic	108.90	5.35	-3.91	Eve	68.30	-4.45	4.16
leopold	102.42	2.31	-3.22	Most	67.78	0.92	-0.64
Pow	100.54	-2.14	4.32	Buzz	67.19	5.60	-3.13
patriot	97.79	3.43	-3.42	Gibson	67.07	0.56	-2.22
Alien	97.04	-5.59	5.17	Mouse	66.85	2.09	-2.66
grinch	96.42	2.33	-3.15	Silverstone	66.53	-3.43	4.11
Ship	90.31	4.45	-3.15	John	66.36	3.33	-2.31
craven	89.82	2.66	0.27	Shopping	65.92	-1.81	3.54
Toy	88.91	6.02	-3.50	How	65.19	1.39	-1.08
atkins	88.02	3.07	0.00	Effects	64.96	1.66	-1.61
divoff	88.02	3.07	0.00	Henstridge	63.40	-4.07	3.93
First	87.78	1.95	-0.99	schwarzenegger	63.02	2.28	-2.64
goldblum	86.34	-3.04	4.46	ryan	62.56	1.51	-2.40
buddy	86.21	-1.49	3.75	sommers	62.47	-2.71	-0.86
t2	80.63	6.13	-3.43	preston	62.10	-2.40	3.72
Story	79.64	2.37	-1.24	matrix	61.02	-0.29	-1.63

The binomial distribution has the following probability function:

$$b(x;n,p) = P(X = x) = p^{x}(1-p)^{n-x}\frac{n!}{x!(n-x)} \qquad (21)$$

P(X=x) is the probability of *x* occurrences of an outcome out of a total of *n* trials where *p* is the probability of the outcome (in our case, *p* is the rate at which a given token (word) occurs in the text out of n distinct words or tokens). Given this information, we can create a likelihood ratio test that compares the value of the parameter *p* across two groups (e.g., texts, extracts of text, groups of documents). To do so, we need to derive the likelihood that we will observe *x* successes if the words in the text were generated by a series of *n* Bernoulli experiments with rate of success *p* is given. In this case, given that we are referring to the likelihood of observing an event (x success out of n trial), the likelihood is equal to the probability of that observation[26]:

$$L(n, p; x) = p^{x}(1-p)^{n-x}\frac{n!}{x!(n-x)} \qquad (22)$$

We could compute the likelihood for any two strings of words (e.g., pieces of texts) for any target word, and therefore it is possible to compare whether it is likely that both texts where generated from a binomial distribution with the same *p* parameter. The ratio of the likelihood that the data are generated by a single binomial with the same rate of success *p* to the likelihood that two distinct binominals with rates p_1 and p_2 is given by:

$$R = \frac{p^{x_1}(1-p)^{n_1-x_1}\frac{n_1!}{x_1!(n_1-x_1)}p^{x_2}(1-p)^{n_2-x_2}\frac{n_2!}{x_2!(n_2-x_2)}}{{p_1}^{x_1}(1-p_1)^{n_1-x_1}\frac{n_1!}{x_1!(n_1-x_1)}{p_2}^{x_2}(1-p_2)^{n_2-x_2}\frac{n_2!}{x_2!(n_2-x_2)}} = \frac{p^{x_1}(1-p)^{n_1-x_1}p^{x_2}(1-p)^{n_2-x_2}}{{p_1}^{x_1}(1-p_1)^{n_1-x_1}{p_2}^{x_2}(1-p_2)^{n_2-x_2}} \qquad (23)$$

[26] In general, however, the likelihood of observing a series of events *m* events ($x_1, x_2, \ldots, x_m$) generated from binomials with parameters ($n_1, p_1; n_2, p_2; \ldots n_m, p_m$) is given by the product of their probabilities: $L\left[(n_1, p_1; n_2, p_2 \ldots n_m, p_m); x_1, x_2 \ldots x_m\right] = \prod_{i=1}^{m} {p_i}^{x_i}(1-p_i)^{n_i-x_i}\frac{n_i!}{x_i!(n_i-x_i)}$

Taking the log and multiplying the expression by -2 to ensure that the statistic has the desired distributional properties (i.e., is chi-squared distributed with one degree of freedom), we obtain the LR test statistics:

$$LR\{(n_1, p_1; n_2, p_2); x_1, x_2\} = 2\{p_1^{x_1}(1-p_1)^{n_1-x_1} + p_2^{x_2}(1-p_2)^{n_2-x_2} - p^{x_1}(1-p)^{n_1-x_1} - p^{x_2}(1-p)^{n_2-x_2}\}$$

(24)

Where $\hat{p}_1 = \frac{x_1}{n_1}$, $\hat{p}_2 = \frac{x_2}{n_2}$ and $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ are maximum likelihood estimates of the rate of the binomial in each of the cases. Following this method, we compute LR statistics for each of the words in the text across all the words in the 240 documents using positive-negative and successful-unsuccessful as groups for the computation.

Table 31
Words with largest Z Statistic Positive vs. Negative (Most frequent [60] Words Removed)

Word	Chi-square	Z_PN	Z_SU	Word	Chi-square	Z_PN	Z_SU
terminator	181.43	8.67	-5.08	epic	41.99	3.46	-2.35
Cameron	150.48	6.47	-4.62	perfect	45.89	3.43	-2.03
T2	80.63	6.13	-3.43	patriot	97.79	3.43	-3.42
Toy	88.91	6.02	-3.50	leo	33.92	3.38	-2.08
Toys	72.53	5.73	-3.16	john	66.36	3.33	-2.31
Buzz	67.19	5.60	-3.13	conner	23.52	3.31	-1.85
Sarah	69.83	5.58	-3.05	robot	23.52	3.31	-1.85
titanic	108.90	5.35	-3.91	meyer	26.46	3.31	-1.97
Andy	59.53	5.00	-2.96	horror	72.38	3.16	-0.30
machine	59.22	4.88	-2.82	cal	26.84	3.15	-1.98
hamilton	50.39	4.85	-2.71	neo	46.70	3.08	-2.50
furlong	47.03	4.68	-2.62	eyes	36.33	3.08	-1.78
Ship	90.31	4.45	-3.15	light	32.43	3.08	-1.66
djinn	176.05	4.34	0.00	gladiator	23.17	3.07	-1.84
nuclear	43.06	4.32	-2.51	spectacular	23.17	3.07	-1.84
woody	43.47	4.14	-1.67	stunts	23.17	3.07	-1.84
connor	39.79	4.14	-2.23	atkins	88.02	3.07	0.00
unlike	30.30	3.98	-1.89	divoff	88.02	3.07	0.00
debney	146.71	3.96	0.00	cowboy	20.16	3.07	-1.72
wishmaster	146.71	3.96	0.00	lightyear	20.16	3.07	-1.72
Rose	58.76	3.91	-2.87	tiny	20.16	3.07	-1.72
amazing	40.26	3.86	-2.43	voice	38.32	3.06	-1.41
genie	132.04	3.75	0.00	special	69.17	3.06	-2.19
metal	33.15	3.74	-2.00	leader	27.67	3.04	-2.00
dicaprio	43.18	3.62	-2.49	lethal	27.08	3.04	-1.63
Boat	41.39	3.62	-2.20	meet	21.84	3.04	-1.63
wishes	51.56	3.59	-0.71	winslet	34.75	2.98	-1.99
cyborg	26.88	3.54	-1.98	effect	24.28	2.98	-1.57
liquid	26.88	3.54	-1.98	created	26.65	2.95	-1.09
serious	28.98	3.50	-1.15	may	45.22	2.93	-1.15

Table 32
Words with Largest Z Statistic Successful vs. Not Successful Movies (Most Frequent [60] Words)

Word	Chi-square	Z_PN	Z_SU	Word	Chi-square	Z_PN	Z_SU
murphy	146.31	-5.09	6.17	helgenberger	43.89	-3.39	3.27
oedekerk	160.50	-4.09	6.07	del	41.29	-2.55	3.21
species	160.73	-5.88	5.92	toro	41.29	-2.55	3.21
holy	112.87	-3.99	5.26	murray	39.17	-3.48	3.20
kung	129.65	-2.97	5.18	siam	39.95	-2.77	3.19
alien	97.04	-5.59	5.17	arts	57.43	-2.65	3.17
goldblum	86.34	-3.04	4.46	gorilla	47.11	-1.78	3.11
pow	100.54	-2.14	4.32	lazard	39.02	-3.20	3.08
king	47.31	-4.93	4.23	man	72.16	-1.33	3.04
Eve	68.30	-4.45	4.16	flipper	37.97	-2.00	2.95
silverstone	66.53	-3.43	4.11	tiger	36.44	-1.98	2.91
Ii	77.18	-3.99	4.05	kelly	40.00	-1.69	2.89
eddie	70.73	-2.55	3.97	justin	34.14	-2.99	2.88
henstridge	63.40	-4.07	3.93	williamson	34.14	-2.99	2.88
ricky	71.41	-2.23	3.85	crane	31.08	-2.30	2.80
baggage	56.08	-3.45	3.78	walken	31.08	-2.30	2.80
buddy	86.21	-1.49	3.75	betty	38.54	-1.56	2.79
Fist	73.05	-1.91	3.73	marg	31.70	-2.88	2.78
preston	62.10	-2.40	3.72	fu	30.04	-2.02	2.77
anna	60.12	-3.25	3.72	enter	44.61	-1.16	2.74
Sex	58.80	-3.77	3.65	ventura	25.67	-3.20	2.73
steve	56.42	-2.91	3.60	baby	21.70	-2.98	2.73
martial	58.18	-3.14	3.56	chosen	31.87	-1.85	2.71
network	53.13	-2.71	3.55	ross	33.55	-2.68	2.71
shopping	65.92	-1.81	3.54	dna	29.26	-2.77	2.67
excess	49.08	-3.25	3.54	sil	29.26	-2.77	2.67
emily	47.37	-2.88	3.53	jeff	36.56	-1.26	2.59
elephant	46.33	-3.48	3.36	alicia	26.83	-2.04	2.58
madsen	46.33	-3.48	3.36	hammerstein	26.83	-2.04	2.58
mars	46.33	-3.48	3.36	cromwell	24.71	-2.88	2.58

REFERENCES

Abernathy, A. M., & Franke, G. R. (1996). The information content of advertising: A meta-analysis. *Journal of Advertising, 25*(2), 1–18.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceeding of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado.

Baeza-Yates, R., & Ribeiro-Neta, B. (1999). *Modern information retrieval*. New York: ACM Press.

Bagozzi, R. P. (1984). A prospectus for theory construction in marketing. *Journal of Marketing 48*(Winter), 11–29.

Baumol, W. J. (2002). *The free-market innovation machine: Analyzing the growth miracle of capitalism*. Princeton: Princeton University Press.

Barclay, D., Higgins, C., & Thompson, R. (1995). The partial least squares (PLS) approach to causal modeling: Personal computer adoption and use as an illustration. *Technology Studies, 2*(2), 285–309.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing, 67*(4), 103–117.

Baines, P. R., Scheucher, C., & Plasser, F. (2001). The "Americanisation" myth in European political markets: A focus on the United Kingdom. *European Journal of Marketing, 35*(9/10), 1099–1117.

Berlyne, D. E. (1960). *Conflict, arousal and curiosity*. New York: McGraw-Hill.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Berlyne, D. E. (1973). Interrelations of verbal and nonverbal measures used in experimental aesthetics. *Scandinavian Journal of Psychology, 14,* 177–184.

Berlyne, D. E. (1974a). The new experimental aesthetics. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics* (pp. 1–25). Washington, DC: Hemisphere.

Berlyne, D. E. (Ed.). (1974b). *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation*. New York: Wiley.

Berlyne, D. E. (1974c). Verbal and exploratory responses to visual patterns varying in uncertainty and in redundancy. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 121–158). New York: Wiley.

Berlyne, D. E., & Madsen, K. B. (Eds.). (1973e). *Pleasure, reward, preference*. New York: Academic Press.

Berlyne, D. E. & Madsen, K. B. (1974) Information and motivation. In A. Silverstein (Ed.), *Human communication: Theoretical explanations* (pp. 19-45). Hillsdale, NJ: Lawrence Erlbaum.

Bielecki, A. (1994). *Aesthetic measure for fractals consist of intervals*. Cracow: Jagellonian University, Institute of Computer Science.

Birkhoff, G. (1968). A mathematical approach to aesthetics. In *Collected mathematical papers* (Vol. 3, pp. 320--333). New York: Dover. (Reprinted from *Scientia, 50*(September), 133–146.)

Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality and Quantity, 18,* 377–385.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin 110*(2), 305–314.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Bozdogan, H. (1931). Polygonal forms, Sixth Yearbook of Nat. Council of Teachers of Math, 1931, 165-195. Reprinted in: Birkhoff, G., Collected mathematical papers, Vol. 3, New York: Dover, 1968.

Bradac, J. J., Konsky, C. W., & Davies, R.A. (1976a). Two studies of the effects of linguistic diversity upon judgments of communicator attributes and message effectiveness. *Communication Monographs, 43,* 70–79.

Bradac, J. J., Courtright, J. A., Schmidt, G., & Davies, R. A. (1976b). The effects of perceived status and linguistic diversity upon judgments of speaker attributes and message effectiveness. *The Journal of Psychology, 93,* 213–220.

Bradac, J. J., Desmond, R. J., & Murdock, J. I. (1977). Diversity and density: Lexically determined evaluative and informational consequences of linguistic complexity. *Communication Monographs, 44,* 273–283.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Buckley, C., Singhal, A., Mitra, M., & Salton, G. (1996). New retrieval approaches using SMART: TREC 4. P*roceedings the Fourth Text Retrieval Conference (TREC-4)*, 25-48.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Corman, S., Kuhn, T., McPhee, R., & Dooley, K. (2002), Studying complex discursive systems: Centering resonance analysis of organizational communication. *Human Communication Research, 28*(2), 157–206.

Carley, K. (1990). Content analysis. In R.E. Asher (Ed.), *The encyclopedia of language and linguistics*. Edinburgh: Pergamon Press.

Carley, K. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology, 23,* 75–126.

Carley, K. & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces, 70*(3), 601–636.

Craig-Lees, M., & Hill, C. (2002). Understanding voluntary simplifiers. *Psychology and Marketing, 19*(2), 187.

Crazier, J. B. (1974). Verbal and exploratory responses to sound sequences varying in uncertainty level. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 27–90). New York: Wiley.

de Sola Pool, I. (1959). *Trends in content analysis*. Urbana, IL: University of Illinois Press.

Diamantopoulos, A., & Winklhofer, H. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277.

Diefenach, D. L. (2001). Historical foundations of computer-assisted content analysis. In M. D. West (Ed.), *Theory, method, and practice in computer content analysis* (pp. 13–41). London: Ablex.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science, 41*(6), 391–407.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*(1), 36–48.

Eliashberg, J., & Shugan, S. M. (1997). Film critics: Influencers or predictors? *Journal of Marketing, 61*(2), 68–78.

Foltz, P. W., & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM, 35*(12), 51–60.

Foltz, P. W., & Dumais, S. T. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197–202.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research, 19*(November), 440–452.

Franzosi, R. (1994). From words to numbers: A set theory framework for the collection, organization, and analysis of narrative data. *Sociological Methodology, 24,* 105–136.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika,* 74, 430–431.

Guadagni, P. M., & Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science, 2*(3), 203–238.

Hagerhall, C. M., Purcell, T., & Taylor, R. (2004). Fractal dimension of landscape silhouette outlines as a predictor of landscape preference. *Journal of Environmental Psychology, 24*(2), 247–255.

Hare, F. G. (1974). Verbal responses varying in distributional redundancy and in variety. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 169–173). New York: Wiley.

Heyduk, R. G. (1975). Rated preference for musical compositions as it relates to complexity and exposure frequency. *Perception and Psychophysics, 17,* 84–91.

Holbrook, M. B. (1977). More on content analysis in consumer research. *Journal of Consumer Research, 4*(June), 176–177.

Holsti, R. (1964). An adaptation of the "General Inquirer" for the systematic analysis of political documents. *Behavioral Science, 9*(4), 382–388.

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

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic Management Journal, 20*(2), 195–204.

Jarvis, C., Scott, M., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(Sept), 199–218.

Jedidi, K., Krider, R. E., & Weinberg, C. B. (1998). Clustering at the movies. *Marketing Letters, 9*(4), 393–405.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception and Psychophysics, 12*(4), 354–356.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*(June), 8–18.

Kelle, U. (Ed). (1995). *Computer-aided qualitative data analysis: Theory, methods and practice*. London: Sage.

Leong, E. K. F., Ewing, M. T., & Pitt, L. F. (2004). Analysing competitors' online persuasive themes with text mining. *Marketing Intelligence and Planning, 22*(2), 187–200.

Lynch, J., & Gimblett, H. R. (1992). Perceptual values in the cultural landscape: A computer model for assessing and mapping perceived mystery in rural environments. *Journal of Computers, Environment and Urban Systems, 16,* 453–471.

McMullen, P. T. (1974a). The influence of complexity in pitch sequences on preference responses of college-age subjects. *Journal of Music Therapy, 11,* 226–233.

McMullen, P. T. (1974b). The influence of number of different pitches and melodic redundancy on preference responses. *Journal of Research in Music Education, 22,* 198–204.

McPhee, R., Corman, S., & Dooley, K. (2002). Organizational knowledge expression and management: Centering resonance analysis of organizational discourse. *Management Communication Quarterly, 16*(2), 130–136.

Miller, M. M., & Riechert, B. P. (1994, August). *Identifying themes via concept mapping: A new method of content analysis*. Paper presented at the Communication Theory and Methodology Division of the Association for Education in Journalism and Mass Communication Annual Meeting, Atlanta, GA.

Moles, A. (1958). *Information theory and esthetic perception*. Urbana, IL: University of Illinois Press.

Normore, L. F. (1974). Verbal responses to visual sequences varying in uncertainty level. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 109–119). New York: Wiley.

Perfetti, C. (1969). Lexical density and phrase structure depth as variables in sentence retention. *Journal of Verbal Learning and Verbal Behavior, 8,* 719–724.

Rosa, J. A., Porac, J. F., Runser-Spanjol, J., & Saxon, M. S. (1999). Sociocognitive dynamics in a product market. *Journal of Marketing, 63*(June), 64–77.

Rosa, J. A., Porac, J. F., Runser-Spanjol, & Saxon, M. S. (2001). Embodied concept use in sensemaking by marketing managers. *Psychology and Marketing, 18*(5), 454–474.

Rosa, J. A., & Porac, J. F. (2002). Categorization bases and their influence on product category knowledge structures. *Psychology and Marketing, 19*(6), 503–531.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68*(1), 109–127.

Ryan, M. J., Rayner, R., & Morrison, A. (1999). Diagnosing customer loyalty drivers. *Marketing Research, 11*(2), 18–26.

Saklofske, D. H. (1975). Visual aesthetic complexity, attractiveness and diversive exploration. *Perceptual and Motor Skills, 41*(3), 813–814.

Spehar, B., Clifford, C. W. G., Newell, B. R., & Taylor, R. P. (2003). Universal aesthetic of fractals. *Computers and Graphics, 27*, 813–820.

Simon, C. R., & Wohlwill, J. F. (1968). An experimental study of the role of expectation and variation in music. *Journal of Research in Music Education,* 16, 227–238.

Stamps, A. E., III. (1994). A study in scale and character: Contextual effects on environmental preferences. *Journal of Environmental Management, 42*(3), 223–245.

Stamps, A. E., III. (2002a). Entropy, visual diversity, and preference. *The Journal of General Psychology,* July, 300-320.

Stamps, A. E., III. (2002b). Fractals, skylines, nature and beauty. *Landscape and Urban Planning, 60*, 163–184.

Steck, L., & Machotka, M. (1975). Preference for musical complexity: Effects of context. *Journal of Experimental Psychology: Human Perception and Performance, 104*(2), 170–174.

Stiny, G., & Gips, J. (1978). *Algorithmic aesthetics: Computer models for criticism and design of artwork.* Berkeley: University of California Press.

Sprott, J. C. (1996). *The computer artist and art critic, fractal horizons: The future use of fractals*. New York: St. Martin's Press.

Stevenson, R. L., (2001). In praise of dumb clerks: Computer-assisted content analysis. In M. D. West (Ed.), *Theory, method, and practice in computer content analysis* (pp. 3–12). London: Ablex.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods, A7*, 13–26.

Taylor, R. P. (2001). Architects reach for the clouds. How fractals may figure in our appreciation of a proposed new building. *Nature, 410*, 18.

Taylor, R. P., Spehar, B., Wise, J. A., Clifford, C. W. G., Newell, B. R., & Martin, T. P. (in press). Perceptual and physiological responses to the visual complexity of Pollock's dripped fractal patterns. *Journal of Non-linear Dynamics, Psychology and Life Sciences*.

Taylor, R. P., Spehar, B., Wise, J. A., Clifford, C. W. G., Newell, B. R., & Martin, T. P. (1964). Preferences for rates of information presented by sequences of tones. *Journal of Experimental Psychology, 68*(2), 176–183.

Taylor, R. P., Spehar, B., Wise, J. A., Clifford, C. W. G., Newell, B. R., & Martin, T. P. (1968). Information, run structure and binary pattern complexity. *Perception and Psychophysics, 3,* 275–280.

Taylor, R. P., Spehar, B., Wise, J. A., Clifford, C. W. G., Newell, B. R., Martin, T. P., & Todd, R. C. (1969). A coded element model of the perceptual processing of sequential stimuli. *Psychological Review, 76,* 433–449.

Urban, G. L., & Hauser, J. (2004). "Listening in" to find and explore new combinations of customer needs. *Journal of Marketing, 68*(April), 72–87.

Wheeler, C., Jones, M., & Young, S. (1996). Market entry modes and channels of distribution in the UK machine tool industry. *European Journal of Marketing, 30*(4), 40.

Wold, H. (1985). Partial least squares. In S. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences*, Vol. 6, 581–591.