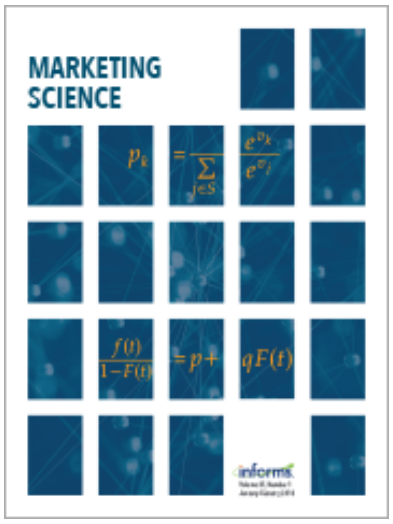

Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Efficient Methods for Sampling Responses from Large-Scale Qualitative Data

Surendra N. Singh, Steve Hillmer, Ze Wang,

To cite this article:
Surendra N. Singh, Steve Hillmer, Ze Wang, (2011) Efficient Methods for Sampling Responses from Large-Scale Qualitative Data. Marketing Science 30(3):532-549. https://doi.org/10.1287/mksc.1100.0632

MARKETING SCIENCE
CELEBRATING 30 YEARS
Vol. 30, No. 3, May–June 2011, pp. 532–549
ISSN 0732-2399 | EISSN 1526-548X | 11 | 3003 | 0532

informs

DOI 10.1287/mksc.1100.0632

Efficient Methods for Sampling Responses from Large-Scale Qualitative Data

Surendra N. Singh, Steve Hillmer
School of Business, University of Kansas, Lawrence, Kansas 66045
{ssingh@ku.edu, hillmer@ku.edu}

Ze Wang
College of Business Administration, University of Central Florida, Orlando, Florida 32816,
zwang@bus.ucf.edu

The World Wide Web contains a vast corpus of consumer-generated content that holds invaluable insights for improving the product and service offerings of firms. Yet the typical method for extracting diagnostic information from online content—text mining—has limitations. As a starting point, we propose analyzing a sample of comments before initiating text mining. Using a combination of real data and simulations, we demonstrate that a sampling procedure that selects respondents whose comments contain a large amount of information is superior to the two most popular sampling methods—simple random sampling and stratified random sampling—in gaining insights from the data. In addition, we derive a method that determines the probability of observing diagnostic information repeated a specific number of times in the population, which will enable managers to base sample size decisions on the trade-off between obtaining additional diagnostic information and the added expense of a larger sample. We provide an illustration of one of the methods using a real data set from a website containing qualitative comments about staying at a hotel and demonstrate how sampling qualitative comments can be a useful first step in text mining.

Key words: consumer-generated media; consumer-generated content; customer feedback on the Web; text mining; qualitative comments; large-scale qualitative data sets; sampling open-ended questions
History: Received: October 10, 2008; accepted: December 8, 2010; Eric Bradlow served as the editor-in-chief for this article. Published online in *Articles in Advance* March 15, 2011.

1. Introduction

The World Wide Web has fundamentally altered how marketers receive and manage customer feedback. Businesses still conduct cross-sectional and longitudinal surveys to gauge customer sentiment in an orderly and focused manner, but there is a vast quantity of user-generated posts on the Web. This content is commonly referred to as consumer-generated media and appears in multiple fora, such as resellers' sites, online opinion review sites, message boards, entertainment sites, blogs, and public discussion boards. The vastness and exponential growth of the consumer-generated media has caught marketers by surprise. In response, the Marketing Science Institute and Wharton Interactive Media Initiative recently issued a joint call for research proposals to stimulate, promote, and facilitate research in this emerging field.

Consumer-generated media have three crucial characteristics. First, the content is not controlled by marketers. A viewer-posted video on YouTube that compared Ferrari to Lamborghini generated 19,395,225 views and 16,518 comments, none of which came from either car company. Second, the sheer volume of information is staggering. There are well over 100,000 Usenet news groups, of which 20,000 or so are active (see Wikipedia). According to eMarketer (Verna 2008), there were 77 million user-generated content creators in 2007, and the number is projected to reach 108 million by 2012. The number of active users of Facebook has extended beyond 200 million (The Future Buzz 2009), and the number of tweets to date is over 29.7 billion (Giga Tweet 2011). Third, consumer-generated media influence buyer behavior (Zhu and Zhang 2010). In a 2005 survey of consumers ($n = 3{,}331$) by Deloitte & Touche, nearly two-thirds (62%) of the consumers reported reading consumer-generated reviews online. Of these, 82% said that reviews had a direct impact on their purchase decision. Furthermore, 69% of those who read the reviews shared them with people in their social circle, magnifying the impact of reviews (Deloitte 2007).

The question therefore is not whether but rather how to harness consumer-generated media for the benefit of the company. The vast corpus of consumer-generated media provides a treasure trove of potentially meaningful information that no business can

ignore, yet the sheer volume and unstructured, qualitative nature of this feedback makes extracting meaningful information from it a daunting task. Text mining seems to be the primary method for handling the mountains of qualitative comments encountered on websites (Ross 2005). Text mining programs scan unstructured text for key words or user-specified phrases. Programs also identify frequently appearing words in the data set that the user can examine more closely in various (positive or negative) contexts in which the words appear. These search programs can help determine the frequency with which different comments occur in unstructured text, but even methods that mimic natural language processing to extract fuller meaning from unstructured text (Atkinson 2007) are limited in their ability to extract important diagnostic information that offers insights into the phenomena in question (Feldman and Lynch 1988). Atkinson (2007, p. 146) notes that "the most sophisticated approaches to text mining or knowledge discovery for texts are characterized by an intensive use of external electronic resources including ontologies, thesauri, etc., which highly restricts the application of the unseen patterns to be discovered." Thus, though mining unstructured text can extract quite a bit of useful information, it is unlikely to find the most unusual and novel information without some prior knowledge of the key diagnostic information in the text.

The problem in theory is not unlike that encountered in dealing with the qualitative responses to open-ended questions in large-scale surveys, some of which—if conducted on the Web—can generate tens of thousands of responses (Daukantas 2000, Kaplowitz et al. 2004). In such cases, researchers suggest selecting a sample of responses and thoroughly examining it (McDaniel and Gates 2005). The information extracted from the sample then can be used either as a representative summary of the total information contained in the data set or as input to develop codes that the text mining programs employ to analyze the remaining responses. It therefore seems logical that a sampling approach also could handle large volumes of qualitative data in the context of the Web. In this sense, the results of sampling responses might even serve as input to the text mining to enable more informed searches for what Atkinson (2007, p. 146) describes as the "unseen patterns to be discovered."

One major problem hinders this recommendation, however: literature in marketing or any other discipline—including qualitative research—provides no guidance about how to select the sample or what sample size will be adequate, as if the type of sampling scheme and the size of the sample were immaterial. Common qualitative research techniques, such as adding cases until the study achieves saturation (Eisenhardt 1991) or selecting a maximum of 15 cases (Carson et al. 2001), are inapplicable when the research does not include in-depth interviews or case research. The problem we attempt to address is a specific, well-defined one: given a data set of unstructured, open-ended, qualitative responses, how (sampling method) and how many (sample size) respondents should be sampled to gain maximum insights (diagnostic information to facilitate meaningful change) in return for a given cost?

We argue that the choice of sampling method and the size of the sample matter. To the extent that different sampling methods lead to the selection of different subsets of qualitative responses, the choice of sampling method alters the nature and extent of insights gained from the data set. Furthermore, in contrast with the common misconception that sample size is irrelevant in qualitative research (Patton 1990, Sandelowski 1995), size is critical in a population in which a significant proportion of participants offer unique diagnostic information. Even qualitative researchers admit that sample size represents an important consideration, if only for time and cost reasons (Carson et al. 2001). The dual goals of this research therefore are to (1) compare a limited number of sampling schemes—including two novel methods that we propose—that might be used to sample qualitative responses from consumer-generated media on the Web and select those that result in the most diagnostic information, and (2) provide guidance regarding adequate sample size.

We treat each respondent's comment as a collection of information units (IUs), where each IU contains one distinct idea. Furthermore, IUs vary in their diagnosticity. For example, an IU about a website's home page design might read, "Disliked the ugly, brown color of the text on the home page." This IU is diagnostic, because it provides information the company could act on, such as changing the ugly brown text. Another IU might instead indicate, "I think it could be a much better home page." This statement reveals a generally negative tenor but fails to "diagnose" anything useful about the home page, and it is therefore nonactionable.

With these data, we consider four sampling plans, including two of the most commonly used probability-based plans: simple random sampling (SRS) and stratified random sampling (STRS) (Cochran 1977, Thompson 2002). A probability sample employs probability theory to develop inferences about the population from the sample with a known degree of certainty. When an analyst can divide a population into groups on the basis of its known characteristics, STRS may be the best choice. Furthermore, if the

groups are more homogeneous than the entire population, STRS can reduce sampling variance (Cochran 1977). A third plan assumes that comments that contain the most IUs should be more informative when included in the sample; we call this approach the largest number of IUs (LNIU) plan. Finally, the fourth plan randomly selects comments sequentially until a predetermined number of IUs, rather than a fixed number of comments, appear in the sample; we call this approach sequential random sampling (SeqRS). In a data set containing qualitative comments, the four sampling plans would work as follows.

(1) SRS: Select a simple random sample of size n from the set of comments. Each comment has an equal chance of being selected. This plan serves as a benchmark of information that can be obtained from the simplest type of sampling.

(2) STRS: Select a stratified random sample of size n, with strata defining groups of comments according to the number of IUs contained in each comment. The stratified sample selects items such that they are proportional to the total number of IUs within each stratum. After determining the sample size for each stratum, the analyst randomly selects the sample from various strata. In this plan, strata that contain more IUs provide a larger proportion of the sample.

(3) LNIU: Predicated on the belief that sampling comments that have the most IUs should be most informative, this approach takes a random sample of the desired size (e.g., 25) from the stratum that has the most IUs. If the size of this group is less than the desired size (e.g., 10), the entire group gets selected, and the remaining sample comes from the stratum that has the second most IUs. If this group contains fewer than 15 respondents, all members get selected, and the remaining portion comes from the next group, and so on. For this plan, a sample of comments of size n will lead to a known number of N IUs being selected.

(4) SeqRS: The desired number of IUs N, rather than a target number of comments n, is chosen. The target number of IUs is specified to be the same as that in the LNIU plan (N) with n comments; however, the process differs. This method randomly selects customer comments until the number of IUs in these comments is as close as possible to but does not exceed N (the number of IUs in the LNIU plan).

In the following, we demonstrate the superiority of the LNIU and SeqRS sampling plans according to theoretical (i.e., deductive and analytical) considerations. We then apply each plan to eight hypothetical sets of comments and a real data set that contains qualitative comments (feedback on Web page design) to determine which plan is most efficient. In eight simulations, we determine whether the findings hold in varying conditions, including extreme violations of our baseline assumption. In a ninth simulation, we generalize the findings to a relatively large data set ($n = 10{,}064$ simulated comments) and determine an adequate sample size for selecting a sample of comments from among thousands. We also illustrate the use of our proposed sampling methods with a real-world data set and show how valuable insights can be gained from the sample itself when used in combination with text mining. In the concluding section, we discuss the limitations of our research, its implications, and some directions for further research.

2. Analytical Comparison of the Four Plans

To help clarify our exposition, we employ two real-world data sets, one of which involves customer feedback on a popular website for reviewing and booking hotels. It lists 1,801 customer reviews (each by a different customer) about a particular hotel in a popular vacation spot. Ten typical customer reviews are listed in Appendix A. We refer to each customer review as a comment, and each comment may contain one or more IUs. For example, the second comment in Appendix A offers two IUs: "The room we booked was not available" and "We were switched to a smaller room." Shorter reviews tend to contain fewer IUs than longer reviews. Furthermore, not all IUs are diagnostic, and even the diagnostic IUs have differing values to hotel management. For example, the IU commenting, "The location was great" is not as valuable as one that noted, "The room was not well vacuumed," because management can institute changes to address the latter, actionable comment.

The 1,801 customer reviews on the website contain a significant amount of managerially relevant information, if management could extract most unique IUs and determine the frequency with which they appear. However, to read, interpret, and categorize all the customer reviews would be labor intensive, costly, and time consuming. Therefore, the hotel could sample some of the comments, develop a list of unique diagnostic IUs in the sample, and then use text mining software to search the entire set of reviews and count the number of times each unique diagnostic IU or its equivalent occurs. However, for this approach to succeed, the hotel needs a sampling plan with a high probability of identifying a large percentage of unique diagnostic IUs at a reasonable cost.

2.1. Problem and Assumptions

Consider comments from 10 customers (Appendix A) about their stay in a hotel. Each comment contains one or more IUs, and in theory, we could classify them into one of K groups such that each group has the

same diagnostic value. (Depending on the nature and purpose of the analysis, managers also could define a diagnostic value of an IU. For this illustration, we consider an IU diagnostic if it provides actionable insights.) Of the $K+1$ groups of IUs, those in group G_0 would have no diagnostic value, those in group G_K provide the greatest diagnostic value, and IUs in group G_k have more diagnostic value than those in G_{k-1}, but less than those in G_{k+1}, where $k = 1$ to $K-1$. To operationalize our results, it is not necessary to define how these $K+1$ groups are determined because we use this conceptual grouping only to derive our theoretical results.

The fourth comment in Appendix A contains seven IUs, one of which, "I wouldn't ever stay there again," has no diagnostic value, because it conveys dissatisfaction but not any specific actionable information. The IU that the "room looked old and worn out" joins the group G_1, because it includes specific information about why the customer is dissatisfied. The remaining IUs then comprise G_2, because they point to particular reasons that are more specific than G_1 and that the hotel can address to reduce customer dissatisfaction. Of course, as in any interpretation of qualitative information, reasonable people can disagree about the classification scheme, but the method a particular company uses to assign diagnostic values has no bearing on the methodology we propose.

To operationalize three of our potential sampling plans, we assume it is possible to determine the number of IUs contained in each customer comment. We illustrate one way to estimate this information for a large set of comments in §6. After estimating the number of IUs in each comment, we can construct Table 1 and thereby execute our sampling plans and compute the information relevant for determining an appropriate sample size (see §5).

Assume there are T customer comments, and each comment C_i contains a known number m_i of IUs with varying diagnostic value, where $i = 1, \ldots, T$. We let IU_{ij} denote the jth IU in comment C_i such that $C_i = \bigcup_{j=1}^{m_i} IU_{ij}$. For example, in the statements about the hotel, the second customer comment (C_2) contains two IUs: IU_{21} ("The room that we booked was not available") and IU_{22} ("We were switched to a smaller room"). Let $M = \sum_{i=1}^{T} m_i$ denote the total number of IUs in all the customer comments, and let $IU_{\text{all}} = \bigcup_{i=1}^{T} \bigcup_{j=1}^{m_i} IU_{ij}$ denote the set of IUs contained in them. We further assume that all IUs in the set IU_{all} can be grouped into one of $K+1$ mutually exclusive groups $(G_0, G_1, \ldots, G_K)$, so the diagnostic value of all the IUs within group G_k is the same. These groups are ordered such that the values of IUs in a higher indexed group are greater than those in a lower indexed group.

Table 1 All Comments in the General Case

Group: No. of IUs in each comment	No. of comments in group	Total IUs in group
1	I_1	(1) I_1
2	I_2	(2) I_2
⋮	⋮	⋮
J	I_J	(J) I_J

To develop the theoretical (analytical) criterion that we will use to evaluate various sampling plans, we also must make an assumption about the rate at which people provide diagnostic IUs of the same value. The *quality of IUs* may depend on factors such as the type of experience with the product or service, the person's knowledge of the target phenomenon, and her or his motivation to respond (McDaniel and Gates 2005). The *quantity of IUs* relies on such factors as the person's disposition (taciturn or verbose) and ability to articulate (Dillon et al. 1987). Therefore, as a reasonable baseline assumption, we assert that IUs are randomly distributed across all comments such that if the IUs in group G_k constitute a proportion p_k of the total number of IUs in the set IU_{all} for $k = 1, \ldots, K$, the probability that any IU in any customer comment is from group G_k equals p_k. This assumption allows for different probabilities for IUs with different diagnostic value. In practice, even if the rate of diagnostic IUs of a given value varies, as long as it does not deviate too much from the baseline assumption, the theoretical criterion we develop should hold.

2.2. Determining the Best Sampling Plan

The best sampling plan captures the most diagnostic IUs, which makes it most useful for effecting meaningful changes. In general, the comments will consist of J strata, based on the number of IUs in each comment. Let I_j be the number of comments in the jth stratum for $j = 1 - J$. The breakdown of the number of IUs is in Table 1.

THEOREM 1. *Assume a sample of n respondent comments selected using SRS, STRS, LNIU, or SeqRS sampling. The expected total number of IUs selected by SRS is N_1, by STRS is N_2, by LNIU is N_3, and by SeqRS is N_4. Then, $N_1 < N_2 < N_3 = N_4$.*

The proof of Theorem 1 is in Appendix B.

We also consider segmenting IUs in all the comments in IU_{all} into homogeneous groups, so that the value of the IUs within each group is the same. In practice, this grouping is possible only after we learn the content of every IU; however, we can conceptualize it for our theoretical investigation. We denote these groups $G_0, G_1, \ldots, G_K$. Consider an arbitrary group G_k, with $k = 1 - K$, in which the value of all IUs are the same but different from the value

of those in any other group. If W^k is the number of unique diagnostic IUs from group k in a sample, we can use the expected value of W^k, $E(W^k)$, to compare the proposed sampling plans. All diagnostic IUs have the same utility, so a larger $E(W^k)$ value indicates more diagnostic information. In this group, some IUs appear once, and others will be repeated. We let U denote the number of unique diagnostic IUs in this group and index the unique diagnostic IUs from 1 to U. The random variable W^k can be represented as $W^k = W_1 + \cdots + W_U$, where W_i is an indicator variable that takes a value 1 if at least one unique diagnostic IU_i appears in the sample and 0 otherwise. Thus, W_i indicates whether the ith unique diagnostic IU appears in the sample. The random variable W^k counts the number of unique diagnostic IUs from this group.

It follows from elementary properties of expected value that $E(W^k) = E(W_1) + \cdots + E(W_U)$. Furthermore, because each W_i is an indicator variable, $E(W_i) = P(W_i = 1)$ is the probability of observing the ith unique diagnostic IU at least once in the sample. To compute $E(W^k)$, it is sufficient to compute $P(W_i = 1)$ for $i = 1 - U$.

Lemma 1. *Using W_i, assume two samples such that the number of IUs for sample 1 is N_1, and the number of IUs for sample 2 is N_2, where $N_1 < N_2$. Then $P_1(W_i = 1)$ denotes the probability that $W_i = 1$ for sample 1, and $P_2(W_i = 1)$ denotes the probability that $W_i = 1$ for sample 2. If the baseline assumption is true, then*

$$P_j(W_i = 1) = 1 - P_j(W_i = 0) = 1 - \frac{\binom{R}{0}\binom{M-R}{N_j}}{\binom{M}{N_j}}, \quad j = 1, 2, \tag{1}$$

where R is the number of times the ith unique diagnostic IU repeats, and $P_2(W_i = 1) > P_1(W_i = 1)$.

The proof of Lemma 1 is in Appendix B. Thus, we can prove the following theorem.

Theorem 2. *For a given set of verbal comments, if the IUs are grouped into a set of K groups such that the value of each IU within the group is the same, W^k is the number of unique diagnostic IUs in group k, and $W = W^1 + \cdots + W^K$ is the number of unique diagnostic IUs in all groups, then if the baseline assumption is true, $E(W^k)$ is largest for LNIU and SeqRS for $l = 1 - K$, and $E(W)$ is the largest for LNIU and SeqRS.*

Proof. From Theorem 1, the number of IUs for the LNIU and SeqRS are greater than the number of IUs for both SRS and STRS. Note that $E(W^k) = E(W_1) + \cdots + E(W_U) = P(W_1 = 1) + \cdots + P(W_U = 1)$. If the baseline assumption is true, from the lemma it follows that $P(W_i = 1)$ for every i is greater for LNIU and SeqRS than for SRS or STRS, because the number of IUs obtained from LNIU and SeqRS are the largest. In turn, $E(W^k)$ is the largest for LNIU and SeqRS for all of the k groups. Because $W = W^1 + \cdots + W^K$, it also follows that $E(W) = E(W^1) + \cdots + E(W^K)$; thus, $E(W)$ is the largest for LNIU and SeqRS because each term in the sum is the greatest for both these sampling plans. This proves the theorem. Q.E.D.

Our preliminary investigation of the performance of the four proposed sampling plans based on analytical considerations thus favors LNIU and SeqRS. To validate these results, we investigate their performance using simulations with empirical data. For simplicity, the simulations always feature only two groups: one in which the IUs have no value and another in which all IUs are diagnostic with the same value. It should be clear from our preceding discussion that this simplification causes no loss of generality, because the results we derive apply to each of the k groups $G_1, \ldots, G_K$ if that group were treated separately.

3. Empirical Validation

To evaluate the performance of the four sampling plans, we use a qualitative data set that contains feedback about the Web page design, as well as two additional data sets constructed from the original data. Because a simulation must be based on some characteristics of a population, our use of real data should lend more credence to the findings. For empirical validation, we also deliberately chose a small data set to ensure that we fully understood the characteristics of the qualitative IUs, such as whether they are diagnostic, repeated, or unique, and the frequency with which they appear in the data.

3.1. Data Source

In a study designed to understand Web page perceptions, 323 students browsed the home page of an existing commercial website and provided written feedback by responding to several Likert and semantic differential items related to their involvement, perception, attitudes, and behavioral intentions. In addition, the participants provided qualitative feedback in their responses to a few open-ended questions about the Web page design (e.g., "What features of the Web page did you like most? Like least?"). Twenty-seven participants did not provide any qualitative responses, which reduced the effective population to 296 people. The target home page describes the website of a small manufacturer of handcrafted, earth-friendly greeting card gifts, which feature flower seeds embedded in handmade paper that can be

planted. The site sells directly to consumers; these products are relevant to the student participants.

In several 30-minute sessions, each involving 10–25 people, participants arrived at a lab and read instructions on assigned computer screens, which indicated that they would be viewing and recording their impressions of a Web home page. To minimize potential viewing differences, the computers had identical configurations, the same browser version, and 17″ monitors with identical display resolution settings (800×600 pixels), color palette, font size, and refresh frequencies. The participants browsed the home page from the perspective of someone interested in buying a greeting card; after viewing the page, they completed a questionnaire that they accessed by clicking on a link on the computer screen.

3.2. Original Population

In this actual population, 296 respondents each made one comment (containing 1–7 IUs) about the Web page, for a total of 932 IUs, of which 188 are diagnostic. Among these 188 diagnostic IUs, some appear only once, which means they are unique (e.g., "Too many links on the left side of the page made it difficult to read"; "The brief description of what exactly the product was instantly grabbed me"). Others appear repeatedly in comments by different people (e.g., "Liked categories on the side"; "Liked list of items on the side"). In all, we identify 51 distinct IUs; in the first two columns of Table 2, we break down the number of times the respondents repeated diagnostic IUs. In the population of 51 distinct IUs, 69% were mentioned only once.

We also can characterize this population by stratifying it on the basis of the number of IUs contained in each comment. The numbers of IUs vary from one to seven, so we include seven strata in Table 3. As this breakdown shows, 17 respondents produced one IU, 59 produced two IUs, and so on. These counts include both diagnostic and nondiagnostic IUs.

Table 2 **Frequency of IU Repetitions in Original and Constructed Populations**

Population 1 (original)		Population 2 (constructed)		Population 3 (constructed)	
Times IUs repeated	Frequency of occurrence	Times IU repeated	Frequency of occurrence	Times IU repeated	Frequency of occurrence
1	35	1	17	1	6
2	3	2	13	2	24
4	2	3	3	3	3
5	1	4	1	4	1
6	2	7	10	7	15
8	1	8	3	8	2
9	1	9	2		
11	1	10	2		
14	1				
15	1				
20	1				
22	1				
23	1				

Table 3 **Breakdown of All IUs for the Original Population**

Stratum	No. in stratum	No. of IUs in stratum	Frequency of total IUs in stratum
1	17	17	0.0182
2	59	118	0.1266
3	130	390	0.4185
4	57	228	0.2446
5	22	110	0.1186
6	8	48	0.0515
7	3	21	0.0225
Total	296	932	1.0000

3.3. Construction of Two Hypothetical Populations

The sample size required to obtain a given amount of diagnostic information is a function of the distribution of the diagnostic IUs in the population. A population with a larger (smaller) percentage of diagnostic IUs that are repeated infrequently requires a larger (smaller) sample. We test the efficacy of the four sampling plans in populations that vary on this dimension. The original population contains a large percentage of infrequently repeated diagnostic IUs (69%), so in constructing our hypothetical populations, we deliberately reduce the percentage of diagnostic IUs that occur exactly once.

To facilitate comparisons, we retain all characteristics of the actual population except for the frequency of diagnostic IUs. Thus, for the actual and two constructed populations, we consider 296 respondents who generate a total of 932 IUs, of which 188 are diagnostic, with 51 distinct ideas. We list the changed frequency of repeated IUs for the constructed populations in the last four columns of Table 2. Intuitively, for any given sample size, it should be easier to capture a larger proportion of 51 distinct IUs when sampling from Population 3 than from Population 2. Likewise, for any fixed sample size, it should be easier to obtain a larger proportion of diagnostic information from Population 2 than from the original population.

3.4. Specifics of the Sampling Schemes

Because the particular allocation of samples for the four sampling plans depends only on the characteristics of the population (see Table 3), the allocation of the samples remains the same for all three populations. We vary the total sample size from 25 to 200 in increments of 25 and evaluate each sampling plan for each increment. For SeqRS, the desired number of IUs (N) is the same as chosen by LNIU sampling for

Table 4 Sample Sizes for the Different Strata: LNIU Plan

Sample size (n)	IUs (N)	Stratum 1	2	3	4	5	6	7
25	139	0	0	0	0	14	8	3
50	247	0	0	0	17	22	8	3
75	347	0	0	0	42	22	8	3
100	437	0	0	10	57	22	8	3
125	512	0	0	35	57	22	8	3
150	587	0	0	60	57	22	8	3
175	622	0	0	85	57	22	8	3
200	737	0	0	110	57	22	8	3

a given sample size. For example, if $n = 25$ in LNIU sampling, we set $N = 139$ IUs for SeqRS (see Table 4). The SRS takes a simple random sample of desired size from the entire set of participants who commented. For STRS, the sample size for each stratum is proportional to the frequency of the total number of IUs in the stratum (Table 3). In Table 5, we provide the allocation of the STRS plan for different sample sizes. (For details, see Appendix C.) Therefore, for a given sample size n, simple random samples of the determined size come from each of seven strata, which then combine to form the stratified random sample. The LNIU plan selects the samples first from the stratum with seven IUs, then from the stratum with six IUs, and so forth, until it achieves the desired sample. In Table 4, we display the allocation of the samples for each n. Although the desired number of IUs (N) rather than the sample size (n) is fixed in SeqRS, we use the sample size associated with the LNIU to refer to the situation in which SeqRS is comparable to LNIU. The target number of IUs for the SeqRS plan appears in the second column of Table 4 for the corresponding sample size.

3.5. Evaluating the Relative Performance of the Four Sampling Plans

We evaluate the performance of the four sampling plans for each of the three populations with a theoretical criterion, based on the baseline assumption. We then evaluate their performance in simulations that do not rely on the baseline assumption, which is violated slightly in all three populations (see the goodness-of-fit test in Appendix D). That is, we draw a sample of size n $(25, 50, 75, \ldots, 200)$ for all plans except SeqRS, for which the comments are randomly selected until we reach the number of IUs in the second column of Table 4. We then compute the value of W, or the number of distinct diagnostic IUs in each sample, and repeat the process 1,000 times for each combination of n, sampling plan, and population. The result is an empirical distribution of W, or the number of distinct diagnostic IUs for each combination. The $E(W)$ statistic provides a measure of the center of the empirical distribution for W and summarizes the amount of information expected from each sampling plan. We report the simulation-generated $E(W)$ values in columns 2–5 of Table 6 (panels (a)–(c) refer to Populations 1–3, respectively). The theoretical $E(W)$ values reflect the information about the three populations using the method outlined in the proof of Lemma 1 and Theorem 2. For these results, we refer to columns 6–8 of Table 6.

As our comparison of $E(W)$ values shows, in Population 1, with the greatest number of unrepeated IUs,

Table 5 Sample Sizes for the Different Strata: STRS Plan

Sample size (n)	IUs (N)	Stratum 1	2	3	4	5	6	7
25	94	0	3	10	6	3	2	1
50	142	1	7	20	12	6	3	1
75	270	1	10	30	19	9	4	2
100	359	2	13	40	25	12	5	3
125	437	2	16	51	31	15	7	3
150	533	3	20	61	37	18	8	3
175	621	3	23	71	45	22	8	3
200	697	4	27	84	52	22	8	3

Table 6 Expected Number of Unique Diagnostic IUs, $E(W)$

	Based on simulation				Based on theory		
n	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$	SRS $E(W)$	STRS $E(W)$	LNIU and SeqRS $E(W)$
				(a) Population 1			
25	10.92	12.92	18.72	15.75	10.43	11.85	15.23
50	17.25	19.68	24.29	22.04	16.41	17.70	20.40
75	22.01	24.41	29.14	27.42	20.93	22.66	25.36
100	25.82	28.69	33.03	31.32	24.80	26.88	29.42
125	29.44	32.73	35.33	34.53	28.41	30.85	32.63
150	32.66	36.34	37.67	37.50	31.83	34.38	35.73
175	36.04	39.94	40.00	40.34	35.12	37.96	38.76
200	39.29	42.52	42.32	43.56	38.28	40.98	41.72
				(b) Population 2			
25	12.88	14.91	20.70	19.84	12.86	14.90	19.86
50	21.62	23.60	26.21	28.47	21.57	23.39	27.01
75	27.69	29.08	31.08	33.73	27.68	29.79	32.81
100	32.24	33.51	35.13	37.69	32.20	34.37	36.81
125	35.87	36.83	37.91	40.41	35.86	38.09	39.60
150	39.03	39.80	40.45	42.63	38.94	41.05	42.10
175	41.59	42.49	42.92	44.86	41.63	43.75	44.32
200	44.03	44.49	45.26	46.78	43.98	45.84	46.32
				(c) Population 3			
25	13.29	15.16	21.64	20.63	13.23	15.41	20.78
50	22.85	24.81	28.89	30.27	22.66	24.68	28.71
75	29.61	30.92	34.62	36.07	29.46	31.82	35.17
100	34.60	35.78	39.23	40.17	34.50	36.88	39.47
125	38.58	39.62	41.77	43.01	38.48	40.81	42.34
150	41.61	42.61	43.62	45.41	41.67	43.73	44.72
175	44.30	45.22	45.09	47.12	44.28	46.20	46.69
200	46.48	47.01	46.28	48.73	46.40	47.91	48.28

the theory-based $E(W)$ values are lower than those based on the simulations for all four sampling plans. In addition, the difference between the LNIU $E(W)$ values and those of SRS and STRS is substantially greater in the simulation results than in the analytical results for sample sizes of less than or equal to 125. The $E(W)$ values for LNIU based on the simulations are lower than the values for SeqRS when the sample sizes are less than 125. For Population 1, LNIU is the best sampling plan, followed by SeqRS. When the sample is smaller than 125, the analytical results substantially understate the advantages of the LNIU and SeqRS plans compared with SRS and STRS.

In Population 2, the simulation- and theory-based $E(W)$ values are nearly the same for SRS; theory-based $E(W)$ values are slightly higher than the simulation values for STRS and LNIU; and theory-based $E(W)$ values are slightly lower for SeqRS. The simulation and analytical results together suggest that LNIU leads to more diagnostic IUs than SRS and STRS, and SeqRS results in more diagnostic IUs.

Finally, in Population 3, which has the fewest unrepeated comments, $E(W)$ values based on the simulation are roughly consistent with $E(W)$ values based on theory for SRS, STRS, and LNIU. Only for SeqRS are the simulated values of $E(W)$ slightly larger than suggested by theory. Furthermore, SeqRS performs slightly better than LNIU in this population, and both perform substantially better than SRS and STRS.

3.6. Additional Simulations: Further Relaxing the Baseline Assumption

In the first three simulations, we evaluate the sampling plans when the proportion of nonrepeated diagnostic IUs varies and violations of the baseline assumption are not extreme. In the next five simulations, we hold the proportion of nonrepeated IUs constant but vary the probability of a diagnostic IU across the strata in a more extreme manner, which enables us to evaluate the performance of the four sampling plans when the baseline assumption is severely violated. For these additional simulations, all populations have the same general structure as in Table 3 and include the frequency of repeated IUs for Population 2 from Table 2. The varying probabilities of a diagnostic IU within each of the seven strata appear in Table 7, along with the probabilities for Population 2 previously used.

For each population, the 188 diagnostic IUs (out of a total of 932) indicate 20% diagnosticity. For Populations from 2-1 to 2-3, the probabilities of diagnostic IUs are greater than 20% for comments that contain fewer IUs and less than 20% for those that contain more IUs. Because it selects the comments with the most IUs, LNIU should perform worse than predicted if the baseline assumption holds such that Population 2-1 should provide the greatest challenge. However, in Populations 2-4 and 2-5, the probability of a diagnostic IU increases for comments containing the most IUs, so the LNIU plan should perform better than predicted by the baseline assumption, because it selects from the strata with a greater chance of producing diagnostic IUs.

Table 7 Probability of Diagnostic Comments in Simulated Populations

Stratum	Pop 2	Pop 2-1	Pop 2-2	Pop 2-3	Pop 2-4	Pop 2-5
1	0.18	0.29	0.24	0.30	0.12	0.12
2	0.32	0.30	0.25	0.30	0.12	0.15
3	0.17	0.21	0.20	0.19	0.13	0.15
4	0.15	0.21	0.20	0.18	0.30	0.21
5	0.26	0.10	0.19	0.18	0.30	0.34
6	0.27	0.10	0.15	0.19	0.31	0.35
7	0.19	0.10	0.14	0.19	0.29	0.38

We conduct all simulations as described previously and compute the expected number of unique diagnostic IUs, $E(W)$, from the empirical distribution. For each population and sample size n, we determine an empirical distribution for the four sampling plans from 1,000 repetitions. We provide the resulting means of these empirical distributions in Table 8.

These simulation results offer the following conclusions: for Population 2-1, the $E(W)$ values for the LNIU plan are substantially lower than those predicted if the baseline assumption were to hold; those for the STRS are slightly lower, and the $E(W)$ values for the SRS and SeqRS are nearly the same as predicted with the baseline assumption. Except for $n = 25$, the $E(W)$ values for LNIU are nearly the same or larger than those for SRS and STRS. The values of $E(W)$ for SeqRS are substantially larger than the other three plans. For Population 2-2, the $E(W)$ values are roughly the same as predicted by theory for all four plans. For Population 2-3, the values for the STRS and LNIU plans are slightly smaller than predicted by theory at sample sizes of 75 or more; the values for the LNIU plan are greater than those for SRS and STRS for every sample size. In addition, SeqRS outperforms LNIU for every sample size except $n = 25$. That is, even when the probability of a diagnostic IU is greater for comments that contain fewer IUs, LNIU tends to outperform SRS and STRS, and SeqRS performs the best.

In both Populations 2-4 and 2-5, the $E(W)$ values for STRS and LNIU from the simulations are substantially greater than predicted by the baseline assumption; those for SRS and SeqRS are close to the theoretical values. The increase in $E(W)$ for the LNIU plan tends to be greater than the increase for the STRS. Although SeqRS performs better than SRS and STRS for these populations, LNIU substantially outperforms the other three plans.

Table 8 $E(W)$ **Values Based on Simulation and Theory**

	Population 2-1				Population 2-2			
n	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$
25	12.89	13.30	12.57	19.68	12.93	14.16	18.12	19.70
50	21.55	21.92	21.77	28.11	21.56	23.02	27.06	28.26
75	27.71	28.11	29.40	33.81	27.68	29.54	32.98	33.72
100	32.23	32.63	34.58	37.59	32.36	34.56	37.34	37.65
125	36.03	36.55	38.13	40.38	35.90	38.46	40.28	40.26
150	38.85	39.69	41.18	42.70	38.99	41.33	42.93	42.65
175	41.66	43.63	44.00	44.89	41.52	44.09	45.32	44.88
200	44.00	44.96	46.33	46.89	44.00	46.08	47.43	46.86
	Population 2-3				Population 2-4			
n	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$
25	13.01	14.49	20.41	19.86	12.94	16.59	25.75	19.63
50	21.64	23.01	27.62	28.22	21.61	25.19	33.71	28.19
75	27.67	29.11	30.75	33.77	27.66	31.74	39.24	33.66
100	32.40	33.60	33.37	37.67	32.32	36.33	42.50	37.58
125	35.93	37.51	36.62	40.33	35.91	39.71	43.87	40.29
150	39.00	40.26	39.72	42.69	38.90	42.22	45.20	42.75
175	41.63	42.53	42.56	44.99	41.67	44.79	46.58	44.84
200	44.06	44.52	45.06	46.76	44.00	46.60	47.88	46.92
	Population 2-5				Based on theory			
n	SRS $E(W)$	STRS $E(W)$	LNIU $E(W)$	SeqRS $E(W)$	SRS $E(W)$	STRS $E(W)$	LNIU and SeqRS $E(W)$	
25	12.86	17.03	28.02	19.55	12.86	14.90	19.86	
50	21.37	25.65	35.53	28.33	21.57	23.39	27.01	
75	27.63	32.18	41.75	33.88	27.68	29.79	32.81	
100	32.08	36.93	45.51	37.62	32.20	34.37	36.81	
125	35.88	41.07	46.86	40.33	35.86	38.09	39.60	
150	38.91	44.10	48.08	42.65	38.94	41.05	42.10	
175	41.63	46.93	49.12	44.85	41.63	43.75	44.32	
200	44.04	48.57	50.22	46.80	43.98	45.84	46.32	

3.7. Discussion

Our evaluations of the proposed sampling plans rely on simulations that require the construction of a realistic set of assumed populations. We begin with actual comments about a real website and make modifications to test the performance of the sampling plans when the violations of the baseline assumption are extreme. Thus, although the results of these simulations are limited to the populations we used, we have carefully chosen populations grounded in real data that pose real challenges to the baseline assumption. If the baseline assumption was true, the LINU and SeqRS methods would uncover more distinct diagnostic IUs than the SRS and STRS methods.

The simulation results verify this advantage for LINU and SeqRS for a variety of populations that violate the baseline assumption, which suggests more useful information is available from LINU or SeqRS than from SRS or STRS. However, SRS and STRS might cost less as methods to analyze the selected comments because they select fewer IUs, and the cost of analyzing comments is proportional to the number of IUs they contain. One way to think comparatively about SRS and STRS versus LNIU and SeqRS is that the former sample a predetermined number of comments, whereas the latter sample a predetermined number of IUs. The superior strategy seems to be to select a predetermined number of IUs rather than a predetermined number of comments. Yet in practice, this information must be balanced against the budget for analysis. If a target number of IUs is given, decision makers can use the information we outline in §5 to evaluate this balance; the probability of detecting IUs that repeat infrequently and the cost to analyze the target number of IUs can help decision makers select the appropriate number of IUs.

The cost associated with the LNIU and SeqRS methods is nearly the same because both select the same number of IUs. It may be slightly less time consuming to sort comments by the number of IUs and select those with the most than to generate a set of random numbers and use them to identify the specific comments to be selected; however, the difference in cost is minimal. Furthermore, if the baseline assumption is true, both methods give the same result on average. Thus, because it is slightly easier and somewhat less costly to select the sample for LNIU, it might be preferable if the decision maker is confident the baseline assumption is true. Overall, the $E(W)$ values from the simulations for SeqRS were closer to the $E(W)$ based on theory than were the simulated values of $E(W)$ for LINU. This is most evident in Populations from 2-1 to 2-5. Intuitively, this makes sense because choosing comments randomly mimics the conditions in the baseline assumption.

Another consideration is the standard deviations of the simulated results for LNIU and SeqRS, which can be estimated from the empirical distributions derived from the 1,000 repeated simulations for each method, each population, and each sample size. The standard deviation for SeqRS is roughly twice that for LNIU in the same conditions; therefore, there is more risk of getting smaller or larger than average results for SeqRS than for LNIU.

These observations lead to some suggestions for how to choose between them. Without information about how the probability of a diagnostic IU varies with the number of IUs in a comment, a conservative choice would be SeqRS, which will perform about the same, regardless of whether the baseline assumption is correct. If there is reason to believe the probability of a diagnostic IU is significantly greater for comments that contain fewer IUs, SeqRS again is the better choice. If the probability of a diagnostic IU is significantly larger for comments with more IUs, however, LNIU is a better choice. When the main concern

is the average performance, as measured by $E(W)$, SeqRS provides more conservative results than LNIU, because it protects against violations of the baseline assumption that disadvantages LNIU but performs as well as LNIU when the baseline assumption holds. However, if decision makers are relatively certain that the baseline assumption is not violated, LNIU is more conservative, because its average results will be more consistent (less variable) than the SeqRS results. Ultimately, the choice depends on prior beliefs about the structure of the population being sampled.

4. Generalizing to Large-Scale Surveys

Sampling is most useful when there are many comments; therefore, we investigate the performance of the four sampling schemes in a relatively large population. We construct a hypothetical population of more than 10,000 comments, beginning with the characteristics of Population 1 described in Tables 2 and 3 and replicating it 34 times. The resulting population contains 10,064 comments, each with one to seven IUs. The diagnostic IUs for this population also reflect the same proportion as in the first two columns in Table 2. Thus, many diagnostic IUs occur a very small percentage of the time; 35 of the 51 unique diagnostic IUs occur only in 0.34% of the cases.

We again use simulations to evaluate the performance of each sampling plan and report these results in Table 9. The expected number of unique diagnostic IUs, $E(W)$, is nearly equal to 51—the total number of unique diagnostic IUs for relatively small sample sizes for LNIU and SeqRS and moderate sample sizes for STRS and SRS sampling plans. In addition, the probability of observing all 51 unique diagnostic IUs (denoted in Table 9 as "Prob.") is nearly 1.00 for sample sizes as small as 800 with the LNIU and SeqRS plans. In this example, we can thus obtain all important diagnostic information using either LNIU or SeqRS and sampling a relatively small proportion (less than 10%) of the population. As theoretically expected, the performance of these sampling plans is best for this large population, and the STRS performance is better than that of SRS. The simulation also reveals that LNIU and SeqRS perform substantially better than the other two plans for small sample sizes.

Table 9 **Expected Number of Unique Diagnostic IUs and Probability of Selecting All Unique Diagnostic IUs from a Large Population**

	SRS		STRS		LNIU		SeqRS	
n	$E(W)$	Prob	$E(W)$	Prob	$E(W)$	Prob	$E(W)$	Prob
200	32.29	0.000	33.87	0.000	43.25	0.000	42.51	0.000
400	41.88	0.000	43.35	0.000	47.69	0.002	48.80	0.099
600	46.57	0.010	47.58	0.023	50.37	0.491	50.31	0.498
800	48.85	0.110	49.50	0.229	50.94	0.940	50.80	0.819
1,000	50.02	0.367	50.36	0.524	51.00	1.000	50.94	0.941

5. Determining an Adequate Sample Size

Because the LNIU and SeqRS plans are superior to the other two plans, we base our discussion of appropriate sample size on them; a similar analysis is possible for the other two plans. When we determine a sample size n for LNIU, we also determine the target number of IUs N for SeqRS, which we illustrate with the hypothetical population of 10,064 described in the previous section. We know the number of IUs contained in the comments for this population, so we can assume the information in Table 10 is known. We also illustrate how to estimate this information for any population using an actual example in the next section.

To determine an appropriate sample size, we compute the probability of selecting a diagnostic IU that repeats R times in the population. Let the random variable W_J equal 1 if an IU repeated R times in the population is selected and 0 otherwise. Let M be the total number of IUs in the population; for the example from Table 10, $M = 31{,}688$. Also from Table 10, we calculate the total number of IUs N observed for any sample of n respondents. For example, if $n = 100$ with LNIU, we select 100 individuals who produced seven IUs, so $N = 700$ IUs. If $n = 200$, we select 102 individuals who generated seven IUs and 98 individuals who generated six IUs to obtain $N = (7(102) + 6(98)) = 1{,}302$ IUs. The total number of IUs for the different sample sizes appears in the second column of Table 11.

Table 10 **Frequency of IUs in the Hypothetical Population**

No. of IUs per individual	1	2	3	4	5	6	7	Total
Frequency	578	2,006	4,420	1,938	748	272	102	10,064
Total	578	4,012	13,260	7,752	3,740	1,632	714	31,688

Table 11 **Probability of Observing One or More Diagnostic IUs Repeated R Times**

LNIU	SeqRS	$R = 10$	$R = 20$	$R = 30$	$R = 40$
n	N	$P(W_J = 1)$,	$P(W_J = 1)$,	$P(W_J = 1)$,	$P(W_J = 1)$,
100	700	0.2003	0.3606	0.4887	0.5912
200	1,302	0.3429	0.5683	0.7164	0.8137
300	1,902	0.4618	0.7103	0.8442	0.9162
400	2,476	0.5570	0.8038	0.9131	0.9616
500	2,976	0.6273	0.8611	0.9482	0.9807
600	3,476	0.6874	0.9023	0.9695	0.9904
700	3,976	0.7386	0.9317	0.9822	0.9953
800	4,476	0.7821	0.9525	0.9896	0.9977

With the baseline assumption, we determine the probability of observing a diagnostic IU repeated R times using Equation (1). From Table 10, it is easy to determine the values of M and N for any sample size n in the LNIU and SeqRS sampling. The value of R is given, so we easily compute $P(W_J = 1)$ from Equation (1) and list the associated probabilities for the population in Table 11.

In an actual application, researchers cannot know the number of times an IU is repeated, but the values in Table 11 can be helpful for selecting a sample size. For example, in a population of more than 10,000, it would be reasonable to specify a range for the percentage of individuals who produce the most infrequent IU and compute the probability of observing that IU at least once for the given range of values. If the most infrequent IU were provided by between 0.1% and 0.4% of the population, the range for the associated R values are those in Table 11. In this case, a sample of size 800 (4,476 IUs) would offer a high probability of observing even the most infrequent IU and might be a reasonable choice. In practice, a manager would need to integrate the information about the probability of observing an IU repeated R times with the cost of obtaining a sample of the required size. This involves balancing the benefit of an increased probability of observing a rare IU with the increased cost of a larger sample.

In practice, information equivalent to that in Table 10 can be obtained from the focal population with commercially available software, and researchers can make some assumptions about the frequency of the most infrequent IU. Equation (1) then can help compute $P(W_J = 1)$ for different sample sizes or different target IU values, which then suggests the appropriate sample size as a function of the probability of capturing an IU that occurs with certain frequency.

6. Using SeqRS to Sample Comments: An Illustration with a Large Real-World Data Set

To illustrate SeqRS sampling in practice, we consider a real-world qualitative data set of 1,801 customer reviews about a hotel. To perform the sampling, we first construct a table similar to Table 1 that groups the comments by number of IUs. We randomly sampled 25 comments from the 1,801 customer reviews and, using text processing software, counted the number of words in each comment (after removing the 25 words in Appendix A). We then manually coded the 25 comments to determine the number of IUs. Two coders performed all coding, both for these 25 identified and several used later, by dividing each comment into idea units, each of which contained one distinct idea. As an example, the comment, "The hotel is rundown and needs improvement," contains two distinct idea units: "The hotel is rundown," and "The hotel needs improvement." The coders worked independently and resolved any disagreements through mutual discussion. The intercoder agreement reached 0.92, and Krippendorff's (1980) α is 0.84.

We fit a linear regression equation for the number of IUs regressed on the number of edited words for the sample of 25 comments and obtained $\text{IU} = 3.07 + 0.122 \times EW$. The R^2 value for the regression equation was 87.8%. To predict the number of IUs in the remaining 1,776 comments, we used text processing software to determine the total edited words in the 1,776 comments, which provided the input into the regression equation. Finally, we combined the actual IUs in the 25 coded comments with the estimated IUs for the remaining 1,776 comments.

The results that we summarize in Table 12 are an approximation of the breakdown of the actual number of IUs in the set of comments, based on information easily obtained from the set of comments through the use of text processing software. (The coded 25 comments are part of the SeqRS sampling plan, so processing these comments to determine the number of IUs does not constitute additional effort.)

Table 12 indicates a total of 14,516 IUs in the hotel comments. Given this total number, $M = 14{,}516$, along with the target number of IUs (N) and an assumed value of R (the number of times an IU is repeated in the entire set of comments), we use Equation (1) to compute the probability of selecting a comment (repeated R times) from the entire set, as we show in Table 13.

Table 12 Breakdown of the Estimated Number of IUs in the Comments on the Hotel Stay

IUs in group	2	3	4	5	6	7	8	9	10	11	12	13
Comments in group	1	9	199	275	320	233	175	151	92	79	70	45
Total IUs	2	27	796	1,375	1,970	1,631	1,400	1,359	920	869	840	585
IUs in group	14	15	16	17	18	19	20	21	22	23	24	25
Comments in group	26	26	16	16	17	8	8	6	3	6	3	3
Total IUs	364	390	256	272	306	152	160	126	66	138	72	75
IUs in group	26	27	30	31	37	38	Total					
Comments in group	5	3	1	2	2	1	1,801					
Total IUs	130	81	30	62	74	38	14,516					

Table 13 Probability of Observing One or More Diagnostic IUs Repeated R Times in the Hotel Comments

Target no. of IUs	$R=5$	$R=10$	$R=15$	$R=20$	$R=25$	$R=30$	$R=35$	$R=40$
250	0.0832	0.1595	0.2295	0.2936	0.3526	0.4065	0.4560	0.5013
500	0.1608	0.2958	0.4091	0.5052	0.5840	0.6510	0.7072	0.7544
750	0.2330	0.4118	0.5489	0.6541	0.7348	0.7967	0.8442	0.8806
1,000	0.3002	0.5103	0.6574	0.7603	0.8324	0.8828	0.9180	0.9427
1,250	0.3626	0.5937	0.7411	0.8351	0.8949	0.9331	0.9574	0.9729
1,500	0.4204	0.6641	0.8054	0.8873	0.9347	0.9622	0.9781	0.9873

These probabilities provide guidance for determining the target number of IUs to sample. The primary determinant is the analyst's or manager's judgment of the desirability of sampling unique comments. With a target of 1,000 IUs, the probability of sampling a comment that occurred only 10 times is 0.5103, the probability of one that occurred 20 times is 0.7603, and the probability of a 40-repeat comment is 0.9427. We consider a sequential sample targeted at 1,000 IUs as reasonable. Even though it contains only 7% of the 14,516 IUs, the chance of selecting an IU that is repeated 20 times is slightly greater than 0.5, and the chance of getting an IU that is repeated 40 times is close to certain (0.9427). Because the number of IUs for the remaining 1,776 comments has been estimated, it is easy to select comments randomly until the total estimated number of IUs equals 1,000; we sequentially and randomly sampled an additional 87 comments.

After coding these 87 randomly selected customer comments, we combined them with the original 25 coded comments. This manual coding of 112 comments revealed a total of 1,015 IUs, of which 370 were nondiagnostic and 645 contained diagnostic information. The sampled diagnostic IUs contain valuable customer feedback, both positive and negative, that can be used in a couple of ways. They might offer a preliminary indication of how customers perceive the hotel and the changes needed to improve it. Just as important, these sample IUs also could provide a starting point for text mining software that searches the entire set of 1,801 customer comments to draw inferences about the hotel. Using information from the sample, we mined the entire data set. The 645 diagnostic IUs from the sample are grouped according to common themes. For example, one prevalent theme in the sampled IUs is that the hotel is a "good value"; 33 IUs in the sample belong to this category. Some representative phrases in this category include "Great room for the money"; "I chose this hotel because of the price"; "The price of the room is very good; for the price, the hotel is worth it"; "This is the best deal in the area"; and so on. Next, for each category identified from the sampled IUs, using the phrases for the IUs in that category from the sample and synonyms for these phrases, the entire set of comments are searched (via the concordance program MonoConc Pro) to determine the counts of the IUs in the category. For instance, to obtain the number of IUs in the "good value" category, all the phrases for this category, along with their synonyms, are used as the search criteria. Thus, the phrase "good room for the money" is used as one of the search terms because it is a synonym for "great room for the money." The total counts of the IUs for each category in the entire set of 1,801 customer comments are reported in Appendix E.

We also determined joint counts of the IUs in all combinations of pairs of the categories from the sample using the MonoConc Pro program. As an illustration of how this is done, the phrases associated with "good value" are combined with phrases in IUs that are associated with "good location," and the entire set of 1,801 comments is searched for the number of times the two phrases both occurred in a customer's comment. We found 233 joint occurrences of "good value" with "good location"; in contrast, similar analyses reveal only four occurrences of "good value" with "good bathroom" and with "good restaurant."

7. Illustrative Results from Mining an Entire Data Set Using the Sampled IUs

A total of 2,954 diagnostic IUs are obtained as a result of data mining (see Appendix E for details). Of these, 1,586 are positive and 1,368 are negative. These IUs are further grouped into two sets: noncommensurate ones (a positive perception not countered by a negative one, and vice versa) and the commensurate ones. Overall customer evaluations are positive, in that the positive IUs (1,586) exceed the negative ones (1,368).

7.1. General Impressions from the Mined Data

An in-depth examination of IUs (excluding items with fewer than 10 IUs) is quite informative. Positive noncommensurate IUs indicate that the hotel is perceived as having a good location (349), the view from the rooms is good (101), rooms are of good size (69), and mirrors on the walls/ceiling are appreciated (17), as is the restaurant (13). Looking at the noncommensurate negative IUs, the biggest gripe is that the hotel

is old (140), followed by hidden fees (107), escalator problems (87), problems with the room TV (65), outdated rooms (51), bad-smelling rooms (51), smelly carpets (38), and the hotel itself smelling bad (35). There are also complaints about room temperature control (32), the lack of a coffee machine in the room (25), problems in finding the room (22), no refrigerator in the room (19), bad carpets (12), and noisy rooms (12).

Examining the commensurate IUs, the number of IUs indicating the hotel as a good value (446) far exceeds those who characterize it as a poor value (74); ditto for pool (129 good, 34 bad), staff service (129 good, 84 bad), and casino (46 good, 6 bad). However, there are more negative than positive IUs about check-in (114 bad, 91 good), the room being clean (96 not clean, 78 clean), the bathroom (146 bad, 33 good), the beds (63 bad, 1 good), the hotel being clean (23 not clean, 21 clean), and hotel security (10 bad, 5 good). Finally, an equal number (20) of good and bad IUs are obtained vis-à-vis the cleaning service.

Even the foregoing simple summary reveals those areas that may need immediate attention such as bathrooms (the largest number of complaints, 146), followed by poor check-in experience (114) and escalator issues (87). The fact that the hotel is considered old (140) may suggest a need for renovation, as long as it does not conflict with the hotel's historic image.

7.2. Joint Counts of IUs for Gaining Further Insights

Additional insights are obtained by analyzing IUs jointly, where appropriate. IUs concerning such problems as the escalators, room TV, room temperature control, noisy room, and lack of a coffee machine and refrigerator in the room are easy to interpret. But the cases where (a) there are related issues, e.g., the hotel is old, has outdated rooms, and has historic value, or (b) where there is little information available prima facie (e.g., hidden fees (107)), using joint frequency counts of IUs might prove valuable.

To ensure that the managerial recommendations resulting from the IUs are compatible with one another, we computed joint counts of "old hotel" and "historic value"; of "old hotel" and "outdated rooms"; and of "historic value" and "outdated rooms." The resulting joint counts are as follows (values in parentheses): "old hotel" and "historic value" (2), "old hotel" and "outdated rooms" (7), and "historic value" and "outdated rooms" (0), suggesting that there is little association between the hotel being old and outdated and the hotel being historic. Thus, the hotel might be remodeled without jeopardizing its historic value.

In the case of hidden fees, the IUs in the data set reveal that some customers are unhappy with what they perceive as hidden charges related to a resort fee for using the pool, spa, and fitness center; an Internet access fee; a fee for parking; and a charge for local or 800 phone calls. To understand customer perceptions of these fees, the hotel needs to know how many customers are unhappy about each type, as well as the number of customers dissatisfied with two or more fees. We further text mined to search all customer comments for joint phrases associated with two or more fees and found that of the 1,801 customers, 74 are displeased with the resort fee, 15 dislike the Internet fee, 5 complain about the parking fee, and 13 hate the phone charges. In addition, two customers complained about both the resort fee and the Internet fee; two, about the resort and parking fees; and five, about the resort fee and phone charges. No customer complaints combined any other set of two or more fees. Thus, it appears that the most annoying extra fee is the resort fee; customers who are charged the resort fee may be more likely to be dissatisfied with the other added fees.

7.3. Developing Customer Segments with Similar (Compatible) Desires

We considered all items with sizable IUs (>100) as bases for segmentation. Comments in Appendix E reveal only one natural segment that is based on "value." The highest number of IUs (520) belongs to the value dimension—446 IUs describe the hotel as a good value, whereas 74 consider it a poor value. The second-highest number of IUs concerns good location (349), but we do not treat it as a segmentation variable, for a joint count shows that 67% of those who consider the hotel as a good value do so, in part, because of its good location. The same goes for the pool (129 IUs): 53% of those who consider the hotel as a good value include the pool in their choice. The fourth-largest set of IUs (129) concerns staff service, but we decided against using it as a desirable segmentation variable because although 60% of the IUs on this dimension are positive, 40% are negative. Thus, staff service should be treated as a desirable attribute of good value. We refrain from using "good view from room" (101 IUs) also as a segmentation basis. Rather, it too could be incorporated as an attribute contributing to good value as 36 of the 101 IUs are common with good value.

Using the number of IUs as a proxy for the segment size, we estimate that 520 out of 2,954 or about 18% of the customers find "value" as an explicitly relevant dimension. Of these, 15% (446/2,954) find it a good value and 2.5% (74/2,954) find it a poor value.

Those who consider the hotel a good value seem to particularly emphasize its good location, good pool, clean rooms, good staff service, good view from the room, and good-sized rooms (for details, see Appendix F). On the other hand, some in this segment

also find the hotel to be old, the staff service poor, bad bathrooms, unclean rooms, and rooms outdated and smelly (for details, see Appendix G).

Of those who consider the hotel a poor value, most complain about poor staff service (11), check-in problems (7), rooms not being clean (6), bad bathrooms (5), and the hotel being old (5) (see details in Appendix H). However, the same guests also praise the hotel for its cleaning service (11), good location (9), its historic value (9), and a good pool (7) (see details in Appendix I). (Note that there are 1,326 sets of pairwise joint counts and we examined them all; no additional insights were gained, however.)

Overall, our analyses suggest that the management of the hotel needs to address not only the general problems identified in §7.1, but also tackle issues that the "good value" segment finds particularly undesirable (e.g., the perception that the hotel is old, staff service is poor, and so on). Equally important, many of these same shortcomings are emphasized by those who find the hotel a poor value. Addressing these will increase the size of the desirable "good value" segment while shrinking the size of the undesirable "poor value" segment.

As a validity check of our qualitative analyses, we use a quantitative piece of information found on the hotel's feedback website: 972 customers had indicated that they will recommend the hotel to others versus 829 who indicated that they will not—a ratio of 54% to 46%. We mined the qualitative data set looking for phrases such as "will return," "stay again," "come back," "will recommend" (and their respective negatives). We found 70 customers in the "recommend" category and 62 in the "nonrecommend" category: a ratio of 53% to 47%, which is virtually identical to the ratios obtained using the quantitative input, bolstering confidence in our qualitative analyses.

8. Implications, Limitations, and Further Research Directions

Papa John's national pizza chain places an Internet feedback form on its website and requests open-ended feedback about site features, products, the organization, or "anything else that comes to mind." Microsoft Inc. receives a flood of e-mails from readers of its various websites, asking questions, providing suggestions for new features, and so on (Ross 2005). These are just two examples of the vast array of qualitative data being generated on the Web. The data contain invaluable insights for improving product and service offerings by "identifying customer pain points, issues, and trends" (Ross 2005). Furthermore, because this feedback is in the customers' own words and unbiased by framing—as occurs when consumers respond to structured questions in surveys—it can be a rich source of wording for promotional campaigns (McDaniel and Gates 2005).

Harnessing insights contained in these large consumer-generated data is a challenging task, because coding and interpreting qualitative statements is a tedious, time-consuming, and expensive endeavor (Aaker et al. 2004). Computer programs such as NUD $*$ IST have automated the coding, but the interpretation is still left to the analyst (Churchill and Iacobucci 2005). Sampling a fraction of comments seems to provide a logical solution for extracting meaningful diagnostic information, regardless of whether the information supports inference making (i.e., getting a feel for the tenor of respondents' comments) or helps develop additional codes for further text mining (Nisbet and McQueen 1992). Yet extant literature is mute about which sampling methods to use and which sample sizes to choose. We attempt to address this long-standing problem by developing an analytical framework that we use to compare two established sampling schemes—SRS and STRS—with two proposed plans: LNIU and SeqRS. For a wide variety of populations, the LNIU and SeqRS plans offer substantial advantages, especially if the rate of diagnostic IUs is uniform across the frequency of IUs. However, even with a nonuniform rate of IUs, evidence from our simulations suggests that the LNIU and SeqRS plans still perform as well or better than the other two sampling plans.

The value of the proposed framework is also derived from its ability to compute the probability of sampling specific diagnostic IUs, repeated R times in the population, from readily obtainable information about the number and frequency of IUs per comment. Such information should enable managers to make informed choices that recognize the trade-off between obtaining additional diagnostic information and the added cost of coding and analyzing that information.

One apparent limitation of the proposed method is that it relies on data sets in which all comments appear together as a block. Can the method also be applied to blogs or other cyber communications, in which one person's comments may be spread throughout the thread? One way to deal with this issue would be to group each person's comments into a block and then proceed as we have illustrated; even a simple word processing program or standard statistical package could accomplish such grouping. Another limitation is that the proposed method does not do automatic data scraping, as is the case with some other methods (e.g., Feldman et al. 2010, Lee and Bradlow 2007). Unlike Lee and Bradlow (2007), our method does require some intervention.

The potential for further research is significant. Researchers could investigate how the most diagnostic information identified by LNIU or SeqRS compares with, say, the "themes" identified by online influence measures that attempt to ascertain how information is diffused though social networks and who the key influencers are (Dwyer 2009). Are the people who make the most comments the same as those identified as the most respected in new measures of online influence (Dwyer 2009)?

The Web has changed the very nature of marketing; marketers are no longer in control of their message. Consumer-generated content dwarfs market-generated information, and for many products, this content significantly influences consumer behavior (Zhu and Zhang 2010). Consumer-generated content thus offers a degree of transparency that marketers must address (Mayzlin 2006). To that end, many businesses track social media outlets (Needleman 2009). Our proposed methods for sampling qualitative comments—by enabling extraction of meaningful information from the consumer-generated content—suggest valuable tools that offer a measure of control and can facilitate the effective use of these qualitative comments by business and public policy makers.

Appendix A. Verbatim Customer Comments About a Hotel

ID	Comment
1	Great location for the price.
2	[T]he room that we booked was not available and we were switched to a smaller room.
3	If you are just needing a place to hang you hat, this is a good value for the money. The casino is old school, the hotel needs a little updating, but all in all it's a good place for the money.
4	The hotel room looked old and worn out. It smelled and was dirty. My wife was afraid to sit down in the bath tub because the bathroom was so dirty and worn. The carpets in the hotel had numerous holes covered by duct tape. I wouldn't ever stay there again.
5	I stayed at the back of the hotel. I believe the older part of the hotel. It looked run down but for what I paid for it, I think it was alright. The room was big enough with a patio. And the location was great. I would definitely stay there again.
6	Even after confirming on telephone before checking in I was charged and $8/day as some hotel fees which was also not mentioned any where on the site. I asked for [a] nonsmoking room, I got [a] smokers room which was stinking with cigarette smoke. Did not get proper information from local experts of the hotel regarding the city and attractions near by.
7	The rooms were clean and big. The lighting was a little poor. The location was great. Many of the higher star resorts were within easy walking distance as is access to the monorail. The casino itself is great because they have lower limit blackjack and craps, even on Saturdays. The brunch buffet was terrible however, very few selections for the same price as you'd pay at one of the higher end casinos.
8	The hotel was conveniently located and reasonable prices but it is showing signs of being well past its prime. The carpets in the hallways were very dirty and frayed, the elevators were extremely slow, and our room was not well vacuumed. However, the room was comfortable enough and the location was a definite plus. We definitely do NOT recommend the buffet, which was far below expectations with mediocre food, very limited selection, and high prices. This hotel is OK for those on a limited budget, but it certainly was nothing to write home about.
9	We had a terrible experience, very unfriendly staff, messed up room reservations, toilet that would not flush, television that we had to pay for that only got like seven channels and came in fuzzy—when we asked them they said there was no technician to fix it. Our room was not super clean—it was a cheap hotel—[I] have stayed in other hotels in this chain and I thought they all had to uphold the same standards but this one obviously slipped—at least for us! [Y]ou get what you pay for and it's a cheap hotel—location is at end of strip so—that is not bad at all—within walking distance to everything or monorail . . . .
10	Overall, I was very pleased with the hotel. It is an older hotel, but, still in good shape. We asked if any available upgrades were available and was shown a book of rooms. Although the upgrade prices were decent, we opted to stay with the standard room. We were a little too early to check in so they held our bags while we looked around. When we came back to get our keys, we had an upgrade!! At no charge! The bed was a little hard and we didn't have the widescreen TV, but, we had a great view and big room. The casino was pretty good but needed more variety. Now the not so good, the elevators!! Only 2 out of 8 ran the entire time. We had to go up to go down and vice versa. Sometimes we had to wait 15 minutes. The buffet was overpriced and small. All in all, the location made this hotel worth the money we paid.

Note. We edited the following words out of customer comments (in order of frequency): the, and, to, a, I, for, in, of, is, we, my, it, but, you, they, our, if, so, an, as, it's, or, me, this, your.

Appendix B. Proof of Theorem 1 and Lemma 1

Proof of Theorem 1. In LNIU, the sample selection involved choosing the comments with the most IUs; thus, it follows that N_3 is as large or larger than the number of IUs in SRS and STRS. The number of IUs in SeqRS equals the number of IUs in LNIU. Both SRS and STRS should select at least one comment that has fewer IUs than LNIU sampling; it follows that N_3 and N_4 are equal and larger than either N_1 or N_2.

For SRS, the expected number of IUs in each stratum is proportional to the number of IUs in that stratum, multiplied by the probability of choosing a member of the stratum. Because in SRS, all respondents have the same chance of being selected, P (selecting a comment from stratum j) $= I_j/(I_1+\cdots+I_J)$. Thus, the expected number of IUs from SRS is

$$N_1 = n\left[\frac{(1)I_1+\cdots+(J)I_J}{I_1+\cdots+I_J}\right].$$

For STRS, the number of comments selected from the jth stratum is proportional to the number of IUs in the stratum divided by the total number of IUs. Thus, for STRS, the number of comments from stratum j is $n[jI_j/(I_1+2I_2+\cdots+JI_J)]$, and the number of IUs from STRS is

$$N_2 = n\left[\frac{(1)I_1+(2)^2I_2+\cdots+(J)^2I_J}{I_1+2I_2+\cdots+JI_J}\right].$$

We wish to show that $N_1 < N_2$ or that

$$n\left[\frac{(1)I_1+\cdots+(J)I_J}{I_1+\cdots+I_J}\right] < n\left[\frac{(1)I_1+\cdots+(J)^2I_J}{I_1+2I_2+\cdots+JI_J}\right].$$

This inequality is true if and only if

$$[I_1+(2)I_2+\cdots+(J)I_J]^2 < [I_1+I_2+\cdots+I_J][I_1+(2)^2I_2+\cdots+(J)^2I_J]. \quad \text{(B1)}$$

The left-hand side of inequality (B1) equals

$$I_1^2+(2)^2I_2^2+\cdots+(J)^2I_J^2+\sum_{i\neq j}2(i)(j)I_iI_j, \quad \text{(B2)}$$

and the right-hand side of inequality (B1) is equal to

$$I_1^2+(2)^2I_2^2+\cdots+(J)^2I_J^2+\sum_{i\neq j}[i^2+j^2]I_iI_j. \quad \text{(B3)}$$

The inequality in Equation (B1) is true if and only if the value of Equation (B2) is smaller than or equal to Equation (B3). The first J terms in Equations (B2) and (B3) are identical. The term involving I_iI_j in Equation (B3), minus the term involving I_iI_j in Equation (B2), is $[i^2-2ij+j^2]I_iI_j = [i-j]^2I_iI_j > 0$, because $i \neq j$. Thus, the value of Equation (B3) is greater than that of Equation (B2), and $N_1 < N_2$. We have shown that $N_3 = N_4$ and both are larger than either N_1 or N_2, so the theorem is proved. Q.E.D.

Proof of Lemma 1. Let R be the number of times that the ith unique diagnostic IU is repeated, so $1 \leq R$. Note that $P(W_i=1) = 1-P(W_i=0)$. Next, to compute $P(W_i=0)$, we sample without replacement from a population of size M, which contains one group of size R and a second group of size $M-R$. Because, according to the baseline assumption, all IUs of the same value contained in group k are randomly distributed throughout the population, the hypergeometric distribution applies, and Equation (1) follows. To prove the lemma, it is sufficient to show that $P_1(W_i=0) > P_2(W_i=0)$. From Equation (1), this is true if and only if $\binom{M-R}{N_1}/\binom{M}{N_1} > \binom{M-R}{N_2}/\binom{M}{N_2}$, which is true if and only if $(M-N_1)!/(M-R-N_1)! > (M-N_2)!/(M-R-N_2)!$, which is equivalent to $(M-N_1)\cdots(M-N_1-R+1) > (M-N_2)\cdots(M-N_2-R+1)$. But because $N_1 < N_2$, each term on the left-hand side of the inequality is greater than the corresponding term on the right-hand side, the inequality is true, and the lemma is proved. Q.E.D.

Appendix C. Allocation to Strata for STRS

In STRS, the sample size selected for each stratum is proportional to the fourth column of Table 5. Thus the plan would select 1.82% of the sample from stratum 1, 12.66% from stratum 2, and so forth. Although the frequency of the sample cannot match the frequency in the last column of Table 5 exactly, we attempt to obtain an allocation of the samples that is as close as possible. Therefore the sample sizes for the different strata in Table 4 for $n=25$ result from the frequencies in the fourth column of Table 5, multiplied by the total sample size (25), and rounded to the nearest whole number. For example, the sample size for stratum 1 is $(0.0186)\times(25) = 0.4650$, which we round to 0.

We modify this method of allocation when the desired stratum sample is greater than or equal to the total number in the stratum, which we can illustrate with an example when the total sample size is $n=175$. In this case, the desired sample for stratum 7 is $(0.0230)\times(175) = 4.025$, which we round to 4. Because only three persons appear in stratum 7, we select all three of them, and the remaining 172 items for the sample should be distributed among the remaining strata in the same proportion as the frequencies in Table 5, recomputed after removing stratum 7.

Appendix D. Goodness-of-Fit Test When the Baseline Assumption Holds, Populations 1–3

In Table D.1, we list the observed (same) number of diagnostic comments for Populations 1–3, as well as the expected number of diagnostic comments if the baseline assumption holds. The p-value for testing the null hypothesis that the baseline assumption holds (standard goodness-of-fit chi-squared test) is 0.006. We reject the null hypothesis because there are more diagnostic comments than we expected in stratas 2, 5, and 6 and fewer than expected in stratas 3 and 4.

Table D.1 Observed and Expected Diagnostic Comments for the Three Populations

No. of comments in the stratum	1	2	3	4	5	6	7
Expected diagnostic comments	3.43	23.80	78.67	45.99	22.19	9.68	4.24
Actual diagnostic comments	3	38	66	35	29	13	4

Appendix E. Noncommensurate and Commensurate IUs Summary After Mining the Entire Hotel Data Set

(a) Noncommensurate IUs

Positive IUs	*n*	Negative IUs	*n*	Negative IUs	*n*
Good location	349	Hotel is old	140	No refrigerator in room	19
Good view from room	101	Hidden fees	107	Hotel carpet bad	12
Good room size	69	Escalator problems	87	Noisy rooms	12
Mirrors on walls and ceilings	29	Room TV problems	65	Undesirable guests	9
Hotel's historic value	17	Outdated room	51	Annoying vendors	9
Good restaurant	13	Room smells badly	48	Hotel not maintained	5
		Room carpets badly	38	Bad room furniture	5
		Hotel smells badly	35	Room window problems	4
		Room temp. control bad	32	No microwave in room	4
		No coffee in room	25	Poor room lighting	3
		Problems getting to room	22	Misleading ads	2

(b) Commensurate IUs

Positive IUs	*n*	Negative IUs	*n*
Good value	446	Poor value	74
Good pool	129	Bad pool	34
Good staff service	129	Bad staff service	84
Check-in good	91	Check-in bad	114[a]
Clean room	78	Room not clean	96[a]
Good bathroom	33	Bad bathroom	146[a]
Good beds	1	Bad beds	63[a]
Good casino	46	Bad casino	6
Clean hotel	21	Hotel not clean	23[a]
Good cleaning service	20	Bad cleaning service	20
Good room service	9	Bad room service	9
Good hotel security	5	Poor hotel security	10[a]

[a]Negative IUs exceed positive IUs.

Appendix F. Joint Counts of Good Value and Positive IUs

Good value and ——

Second IU	Joint count
Good location	233
Good pool	69
Clean room	40
Good staff service	36
Good view from room	36
Good room size	34
Check-in good	33
Clean hotel	17
Good casino	16
Mirrors on walls and ceiling	13
Good cleaning service	11
Hotel's historic value	9
Good bathroom	4
Good restaurant	4

Note. "0" joint counts with good beds, good room service, and good hotel security.

Appendix G. Joint Counts of Good Value and Negative IUs

Good value and ——

Second IU	Joint count
Old hotel	42
Bad staff service	23
Bad bathrooms	18
Room TV problems	15
Rooms not clean	13
Outdated room	13
Escalator problems	12
Hidden fees	12
Hotel smells badly	9
Parking problems	8
Room temperature control bad	8
Bad beds	7
Bad room carpets	7
Room smells badly	7

Note. "0" joint counts with no coffee in room, problems getting to the room, no refrigerator in the room, bad hotel carpet, noisy rooms, undesirable guests, annoying vendors, hotel unmaintained, bad room furniture, room window problems, no microwave in the room, poor room lighting, misleading ads, bad pool, bad staff service, bad check-in, bad casino, hotel not clean, bad cleaning service, bad room service, and poor hotel security.

Appendix H. Joint Counts of Poor Value and Negative IUs

Poor value and ——	
Second IU	Joint count
Bad staff service	11
Check-in bad	7
Room not clean	6
Bad bathrooms	5
Old hotel	5
Hotel smells badly	4
Room smells badly	3
Bad room carpets	3
Bad pool	3
Escalator problems	2
Room TV problems	2
Room temperature control bad	2
Outdated room	2
Bad cleaning service	2
Poor hotel security	2
Hotel not clean	1
Hidden fees	1
No refrigerator in room	1
No coffee in room	1
Bad hotel carpet	1
Problems getting to rooms	1

Note. "0" joint counts with bad beds, noisy rooms, undesirable guests, annoying vendors, hotel unmaintained, bad room furniture, room window problems, no microwave in the room, poor room lighting, misleading ads, bad casino, and bad room service.

Appendix I. Joint Counts of Poor Value and Positive IUs

Poor value and ——	
Second IU	Joint count
Good cleaning service	11
Hotel's historic value	9
Good location	9
Good pool	7
Good restaurant	4
Clean room	3
Check-in good	3
Good view from room	1
Good room size	1
Clean hotel	1
Good casino	1
Mirrors on walls and ceiling	1

Note. "0" joint count with good staff service, good bathroom, good beds, good room service, and good hotel security.

References

Aaker, D. A., V. Kumar, G. S. Day. 2004. *Marketing Research.* John Wiley & Sons, New York.

Atkinson, J. 2007. Evolving explanatory novel patterns for semantically-based text mining. A. Kao, S. R. Poteet, eds. *Natural Language Processing and Text Mining.* Springer-Verlag, London, 145–170.

Carson, D., A. Gilmore, C. Perry, K. Gronhaug. 2001. *Qualitative Marketing Research.* Sage, London.

Churchill, G. A., Jr., D. Iacobucci. 2005. *Marketing Research: Methodological Foundations*. South-Western, Mason, OH.

Cochran, W. G. 1977. *Sampling Techniques*. John Wiley & Sons, New York.

Daukantas, P. 2000. Web helps ease AF workload. *Government Comput. News* (January 20) 33–34.

Deloitte. 2007. New Deloitte study shows inflection point for consumer product industry: Companies must learn to compete in a more transparent age. Press release (October 2). Retrieved February 10, 2011, http://goo.gl/YERQh.

Dillon, W., T. J. Madden, N. Firtle. 1987. *Marketing Research in a Marketing Environment*. Times Mirror/Mosby College Publishing, St. Louis.

Dwyer, P. 2009. Measuring interpersonal influence in online conversations. MSI Report 09-108, Marketing Science Institute, Cambridge, MA.

Eisenhardt, K. M. 1991. Better stories and better constructs: The case for rigor and comparative logic. *Acad. Management Rev.* **16**(3) 620–627.

Feldman, J. M., J. G. Lynch Jr. 1988. Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psych.* **73**(3) 421–435.

Feldman, R., J. Goldenberg, O. Netzer. 2010. Mine your own business: Market structure surveillance through text mining. Working paper, Wharton Interactive Media Initiative, http://www.whartoninteractive.com.

Future Buzz. 2009. Future marketing trends—By the numbers. Retrieved July 10, 2010, http://thefuturebuzz.com/2009/07/10/future-marketing-trends.

Giga Tweet. 2011. Retrieved February 10, 2011, http://gigatweeter.com/counter.

Kaplowitz, M. D., T. D. Hadlock, R. Levine. 2004. A comparison of Web and mail survey response rates. *Public Opinion Quart.* **68**(1) 94–101.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage, Newbury Park, CA.

Lee, T. Y., E. T. Bradlow. 2007. Automatic construction of conjoint attributes and levels from online customer reviews. The Wharton School Working Paper OPIM WP 06-08-01, University of Pennsylvania, Philadelphia.

Mayzlin, D. 2006. Promotional chat on the Internet. *Marketing Sci.* **25**(2) 155–163.

McDaniel, C., R. Gates. 2005. *Marketing Research.* John Wiley & Sons, Hoboken, NJ.

Needleman, S. E. 2009. For companies, a tweet in time can avert PR mess. *Wall Street Journal* (August 3) B6.

Nisbet, L. A., D. V. McQueen. 1992. Assessing the qualitative data component in large-scale quantitative surveys: Computer-aided qualitative survey data management, retrieval and analysis. *Health Ed. Res.: Theory Practice* **7**(4) 547–553.

Patton, M. Q. 1990. *Qualitative Evaluation and Research Methods*, 2nd ed. Sage, Newbury Park, CA.

Ross, S. 2005. A view from the glass house: Competing in a transparent marketplace. Retrieved November 28, 2007, http://research.microsoft.com/displayarticle.aspx?id=1067.

Sandelowski, M. 1995. Sample size in qualitative research. *Res. Nursing Health* **18**(2) 179–183.

Thompson, S. K. 2002. *Sampling*. John Wiley & Sons, New York.

Verna, P. 2008. User-generated content: In pursuit of ad dollars. Report, eMarketer, New York. http://www.emarketer.com/Reports/All/Emarketer_2000400.aspx.

Wikipedia. Usenet newsgroup. Retrieved May 27, 2010, http://en.wikipedia.org/wiki/Usenet_newsgroup.

Zhu, F., X. Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* **74**(2) 133–148.