Validation of Database Search Terms for Content Analysis: The Case of Cancer News Coverage

By Jo Ellen Stryker, Ricardo J. Wray, Robert C. Hornik, and Itzik Yanovitzky

While databases are increasingly used for content analyses in mass communication and journalism research, concerns about sampling error have been largely ignored. We introduce a method to measure the quality of a search phrase according to two criteria: recall (its ability to accurately call up items of interest) and precision (ability to avoid extraneous ones). We present a detailed description of the evaluation procedure, offer an example of its use assessing an online search for news reports about cancer, and discuss limitations of the procedure and further potential uses.

Mass communication research has become increasingly reliant on content analytic methods.[1] The growth of content analysis has triggered increased recognition of the need to develop standards of methodological rigor.[2] Yet systematic reviews of published content analyses reveal the relative lack of attention to methodological detail. For example, a review of content analyses published in *Journalism & Mass Communication Quarterly* between 1975 and 1995 by Riffe and Freitag[3] found that only 56% reported some measure of intercoder reliability. Similarly, a review of 200 content analyses published in communication journals between 1994 and 1998 found that 69% mentioned intercoder reliability, but only a small fraction of these reported any procedural or statistical details.[4]

While intercoder reliability is vital to the internal validity of content analytic studies, an equally important methodological concern pertains to sampling, which can strengthen the external validity of content analyses. In probability sampling, required for statistical inferences,[5] the central principle is that each sampling unit has an equal likelihood of being selected. Historically, mass media content analyses have relied more on convenience or purposive than probability sampling: Riffe and

Jo Ellen Stryker is assistant professor in the Department of Behavioral Sciences and Health Education, Rollins School of Public Health at Emory University; Ricardo J. Wray is assistant professor in the Health Communication Research Laboratory, School of Public Health at St. Louis University; Robert C. Hornik is professor in the Annenberg School for Communication, University of Pennsylvania; and Itzik Yanovitzky is assistant professor in the School of Communication, Information and Library Studies at Rutgers, The State University of New Jersey. Previous versions of this paper were presented at the 1998 and 2005 International Communication Association annual conferences in Jerusalem and New York.

J&MC Quarterly
Vol. 83, No. 2
Summer 2006
413-430

Frietag's[6] review found that 77.8% of published content analyses relied upon non-probability sampling. Presumably, logistical difficulties precluded random sampling.

However, researchers are increasingly relying upon electronic databases, such as Lexis-Nexis, for content analysis purposes. Our own review of 198 content analyses published in *Journalism & Mass Communication Quarterly, Journal of Communication, Communication Research, Journal of Broadcasting & Electronic Media, Political Communication,* and the *Journal of Health Communication* between 2000 and 2005 revealed that 42% (eighty-three) sampled from databases. One appeal of databases is the capability to retrieve a large quantity of relevant items with a single search term, thereby providing easier application of random sampling methods.

However, the results and the inferences drawn depend on whether the texts are representative. Reliance on electronic databases expands the reach and power of the content analyst, but also heightens concern about reliability and validity of the selection process. Of the eighty-three content analyses utilizing electronic databases in our review, only 39% provided the search term used, and only 6% discussed the validity of the search term. Content analysis with samples from electronic databases will be strengthened if researchers can make claims about the reliability and validity of methods for selection of texts. Yet, with the exception of our own research,[7] a systematic evaluation of search phrases, with estimates of their ability to return relevant stories and reject extraneous ones, is rarely provided in the literature[8] or in research methods texts.[9] In this manuscript, we present a methodological study of approaches to validating search terms when sampling electronic databases.

Background

The procedure to carry out the selection of relevant content from a universe of texts requires the development of a query to the database, in the form of a search phrase. Different databases use different search interfaces with varying capabilities, but the underlying logic is universal. A search phrase is the combination of concepts designed to return records relevant to the researchers' interest.[10] "Practically, the search phrase is a strainer that allows the researcher to extract from the digital soup only those data that bear directly and specifically on the problem at hand. The trick is to assure the appropriate spaces in the strainer's mesh."[11]

How to evaluate efficiency and effectiveness of search strategies is a central concern of the information sciences.[12] The capacity of the search phrase to select relevant records, and eliminate irrelevant ones, is evaluated by measures termed *recall* and *precision*.[13] These are used to judge the performance of a search[14] and to measure incremental improvements gained from refinements to the search phrase.[15]

Recall and precision are both proportions. Recall is an estimate of the conditional probability that a particular text will be retrieved, given that it is relevant, calculated by dividing the number of relevant items returned by a search phrase by the total relevant records in the database. Precision is an estimate of the conditional probability that a particular text

TABLE 1

Calculations for Recall and Precision

Citations	Not Relevant	Relevant	Total
Not Retrieved	a	b	a + b
Retrieved	c	d	c + d
Total	a + c	b + d	a + b + c + d
Recall = (d/ b + d)	Precision = (d/ c + d)		

is relevant, given that it is retrieved, calculated by dividing the number of relevant items by the total number of items returned in a specific search. Like Type I and Type II errors, precision and recall represent two different kinds of errors—those of omission and commission, respectively.[16] An inverse relationship is often found between recall and precision: as one improves, the other gets worse.[17] However, skilled researchers score higher on both measures than novices.[18] These relationships are depicted in Table 1.

Cell "a" represents items neither relevant nor retrieved by the search phrase. Cell "b" represents relevant items that are not retrieved. Cell "c" represents retrieved items that are not relevant. Cell "d" represents relevant items that were captured by the phrase. Any search phrase easily divulges values for cells "c" and "d," making calculation of precision a matter of dividing the relevant number of items found by the total number of items. Establishing a value for cell "b," required to calculate recall, takes more effort, as we will show. For online searches, the cell "a" quantity is in practice most often unknown. The text databases are for all practical purposes so large that "a" is not calculable. Fortunately, "a" is not needed to calculate recall or precision.

Method

The method to evaluate how well a search phrase selects texts from an online database for the purposes of content analysis involves three stages.

Stage I. In the first and formative stage, the researcher establishes three things, defined by the substance of the research question: (1) the relevant universe of texts (e.g., all Associated Press files from 1970 to 2001; *New York Times* editorials from 1994 to 1995; transcripts of CNN broadcasts during the one month before and after the 2000 presidential election); (2) the definition of story relevance (e.g., stories that question the usefulness of mammograms; stories discussing immigration policy; stories that quote unnamed sources); and (3) a specification of what will be considered a satisfactory level for the recall and precision for this study.

First, the researcher must decide how initial texts will be selected for review: manually or electronically. This discussion is limited to texts that can be retrieved electronically. Our own formative research demon-

strated that a hard copy search produced lower recall values than a broad-based electronic search.[19] While recognizing the fallibility of both procedures, the analyst must choose one (or if the resources are available, both) with as much care as possible.

Second, story relevancy must be determined. Like all constructs in content analysis, story relevancy is threatened by inter-subjectivity of assessment, so a traditional measure of reliability, such as Krippendorff's alpha,[20] should be applied to the decision criteria for relevancy. Although this internal consistency criterion has been recommended for computerized content analysis, it has rarely been applied.[21] An advantage of establishing intercoder reliability is that the task of assigning relevance to texts retrieved by the open search term can be distributed among coders.

Third, the researcher must establish desired levels of recall and precision, which will vary according to research objectives.[22] For example, a researcher seeking a maximum number of relevant stories for human content analysis (and not concerned with the resources required to manually cull out irrelevant stories) needs high recall. A researcher interested in a reliable measure of the number of relevant stories in different periods over time, and intending to rely on computer-aided coding, may want to be sure to eliminate the irrelevant proportion and thus seeks high precision. With equivalent levels set for both recall and precision (i.e., the proportion of extraneous stories is the same as the proportion of missed relevant stories) the resulting numbers of stories can be used with confidence as an estimate of the true number of stories on the topic. Alternatively, in the absence of equivalency, taking the proportion of precision to recall produces the approximate rate of over- or under-sampling that can then be used to adjust the media measures for sampling error.

Stage II. During the second stage, the search phrase is developed and refined with a random sub-sample drawn from the universe of texts. As different iterations of the phrase are tested, recall and precision vary. The stage ends when no significant improvements in recall and precision are observed.

First, then, an "open" search term is developed to capture any and all relevant stories relevant to the topic of interest. The search term uses every conceivable word that might refer to the topic. To generate an open search term, researchers should review a small sample of stories retrieved by key word searches. Any other word encountered during this review process that might also be associated with the topic of interest should also be included. For example, when looking for articles that discuss domestic violence, perhaps the first phrase tried is "domestic violence." However, some texts that are relevant to the topic do not use those terms; they use the term spousal abuse, or mention that a man hit his wife, or a wife killed her spouse. These words should also be included. The open search term is designed to elicit all possible texts, but will also include many irrelevant ones; a good open search term should yield perfect recall (it captures every relevant story), regardless of precision.

Second, an effective "closed" search phrase is designed and refined, using a sample from the universe of texts retrieved by the expansive or "open" search phrase. The analyst reads a sample of texts, and makes a

first effort at defining a search term, which should attempt to capture relevant stories. Continuing with the earlier example, a search term which includes "man or husband" and "spouse or woman or wife" and "kill or hurt" may produce many irrelevant stories (e.g., a husband and wife were killed in an airplane crash), and the phrase will have to be refined further. This process will involve judgment by the analyst, advice by other colleagues, and constant informal reading of texts. In an iterative process, the analyst refines and tests a set of vocabulary terms to include in the online search phrase. Incremental additions and subtractions of words, manipulation of proximity terms, and the like, can be compared to optimize recall and precision.

Numerous books propose strategies for developing search phrases, and, specifically, for improving their recall or precision.[23] For example, to achieve better recall the searcher can use as many synonyms as possible, use the Boolean term "OR" rather than "AND," avoid proximity terms, and include the entire text (rather than limited fields) for the search. To achieve high precision, the searcher can use specific rather than general index terms, limit the search to specific fields, and link as many terms as possible with "AND" and proximity rules.[24]

Once the informal development of a search phrase is complete, the phrase is tested with a systematic sample of texts. Sample size will vary with the likelihood of finding relevant articles, but a reasonable goal would be to have thirty relevant texts for this first test. Although not statistically derived, our experience suggests that thirty relevant texts will provide sufficient variation to adequately test the validity of the search phrase in this formative phase. If a relevant text is likely to be found once for every five stories retrieved by the open search term, the total sample would have to include about 150 stories (30 X 5).

These 150 stories should be chosen randomly from the universe of stories. However, generating a truly random sample would be time-consuming, requiring an enumeration of the census of texts retrieved by the open search term in order to randomly select texts for review. We propose strategies below for stratified sampling that minimize sampling error.

The selected texts should be reviewed for relevance (assuming that intercoder reliability has already been established). Assume the review process revealed thirty relevant stories. The closed search term would then be applied to the 150 stories retrieved by the open search term. The result would then be evaluated against the criteria of recall and precision. If the closed search term retrieved thirty-five articles, and twenty-five were relevant and ten not relevant, the precision would be defined as 25/35=.71 (or 71% precision).

The next task will be to make a recall estimate. That requires an estimate of how many of the unretrieved texts were relevant. If the closed search term had produced twenty-five relevant texts, and the expansive search term turned up five more relevant texts, then the estimated recall for the candidate search term would be 25/30 or 83.3%.

The two estimates (here 71% and 83.3%) would be assessed against the *a priori* criteria for recall and precision that had been speci-

fied. If the analyst had set as a goal recall and precision at 80%, then the criteria were not met in these examples . The next task would be to refine the search term so that for the sample of 150 stories the criteria would be met, by reviewing the texts of the irrelevant retrieved articles and irrelevant unretrieved articles to see what modifications of the search term would produce the requisite recall and precision. The search would then be rerun on the 150-story sample and the analysis repeated to establish that the modified search terms produced the needed recall and precision.

This process will produce a search term that works for the sample. This only establishes that it is possible to create a good search term for that sample. Establishing that the term is relevant to the population of texts requires an additional test.

Stage III. Third, in a formal test, the best search phrase is confirmed using a new random sample. New values of recall and precision are calculated, and confidence intervals can be generated for the recall and precision statistics. We can estimate how many texts will be required to provide an estimate within a specified confidence interval (see Table 2). These are essentially conventionally calculated power estimates assuming 80% power and an alpha level of .05. We present sample sizes required for recall or precision between .70 and .99, and 95% and for confidence intervals from1% to 10%.

We used NQuery software to generate these estimates, using a test of a single expected proportion against a null value. These calculations assume that the sample of texts was selected through simple random sampling. The only way to generate a random sample would be to apply the open search term to the entire study period for every text source, and randomly select texts for review. This process would be inefficient, particularly when trying to apply the closed search term to the same sample of texts. Stratified or cluster sampling is more efficient, and, to the extent that the intra-class correlations associated with this type of sampling can be minimized, then the estimated sample sizes presented in Table 2 are adequate approximations for stratified or cluster sampling.

Using the example from above, assume we wanted to set our recall at 90%, with a 5% confidence interval. Looking at Table 2, then we would need to sample 283 relevant articles to be 95% sure that the observed recall was no less than 0.85. If relevant articles were one-fifth of the total, then one would need 1,415 (283 X 5) articles in the full sample. If the open search term retrieved thirty relevant stories out of 150, and we need 283 relevant stories in our sample to generate our estimates, then we need to increase the sample of our open search 283/30 or by a ratio of 9.433. Hence, we would need to retrieve 150 X 9.433 or 1,415 stories using our open search term to generate our desired estimate.

The newly selected sample would be used to provide the estimates of recall and precision. There is some risk, of course, that the desired levels of recall and precision will not be reached. The analyst then needs to decide whether to do a new modification of the search term and a new test, or simply report the results.

It is possible to develop a search phrase that has high recall and high precision but inadequate validity given the topic of the search. This may

TABLE 2

Required Sample Sizes for Recall or Precision, Assuming Specified Confidence Interval

Desired Precision or Recall	Desired Confidence Interval									
	-1%	-2%	-3%	-4%	-5%	-6%	-7%	-8%	-9%	-10%
.99	987	332	184	123	90	71	58	48	42	36
.98	1588	485	253	162	116	89	71	59	50	43
.97	2171	631	318	199	140	105	83	68	57	49
.96	2740	772	380	234	162	121	95	77	64	55
.95	3296	910	441	268	184	136	106	85	71	60
.94	3839	1044	500	301	205	150	116	93	77	65
.93	4369	1175	558	333	225	164	126	101	83	70
.92	4887	1303	614	364	245	178	136	108	89	75
.91	5393	1428	668	395	264	191	146	116	95	79
.90	5886	1549	722	424	283	204	155	123	100	83
.89	6366	1667	773	453	301	216	164	129	105	88
.88	6834	1782	824	481	318	228	173	136	110	92
.87	7290	1894	873	508	335	240	181	142	115	96
.86	7733	2003	920	534	352	251	189	149	120	99
.85	8164	2109	967	560	368	262	187	154	125	103
.84	8583	2212	1011	584	383	273	205	160	129	107
.83	8989	2311	1055	608	399	283	212	166	133	110
.82	9383	2408	1097	631	413	293	219	171	138	113
.81	9764	2501	1137	654	654	302	226	176	142	116
.80	10133	2591	1177	675	441	312	233	181	145	119
.79	10490	2679	1215	696	454	320	239	186	149	122
.78	10834	2763	1251	716	466	329	245	190	152	125
.77	11166	2844	1286	735	478	337	251	195	156	128
.76	11486	2922	1320	754	490	345	257	199	159	130
.75	11793	2996	1352	772	501	352	262	203	162	133
.74	12088	3068	1383	789	511	359	267	207	165	135
.73	12370	3137	1413	805	521	366	272	210	168	137
.72	12640	3202	1441	820	531	372	276	213	170	139
.71	12898	3265	1468	835	540	378	281	217	172	141
.70	13143	3324	1493	849	548	384	285	220	175	142

* Estimates generated from NQUERY program, test of a single expected proportion against a null value, assuming 80% power and a one-sided test

happen if the definition of what was relevant or not turns out to be inappropriate. Then the numbers generated by the search would be irrelevant. One method for confirming the validity of a search term is through comparison with expected change based on experience and history. For example, the analyst might test the results of the search phrase to see whether known periods of high or low public attention to the issue of concern correspond with high or low levels of the number of articles located with the online search phrase.

The Example of Press Coverage of Cancer. In this example, researchers were conducting a content analysis of cancer news coverage in top major U.S. and ethnic/minority papers for 2003 (the example that follows reports only the results for coverage in the major daily newspa-

pers). For the universe of texts, researchers were interested in the fifty U.S. newspapers with the highest circulation; however, the Lexis-Nexis database provided full-text continuous coverage for only forty-four of the top fifty papers (see Appendices). Although the study time period was 2003, we decided to validate the search term on a sample of stories appearing in 2002. Once the term was validated, all relevant stories in 2003 were entered into a database for random selection, to be reviewed for a variety of different constructs by four coders.

During the formative process, a working definition of a relevant story was established. We started with a broad conceptual definition of relevance (a story containing a minimal amount of cancer information), and ultimately moved to a more rigidly defined definition of relevance to increase our intercoder reliability (see Appendices for decision criteria for relevance). In this study, stories would ultimately be reviewed by human coders; thus, we aimed to find most of the relevant stories, with less consideration with filtering out irrelevant ones. Hence, our goal was to set recall at a minimum of 93%, meaning 93% of relevant stories would be picked up by the search term, with no set criteria for precision estimates.

We utilized a two-stage stratified sampling process to try to approximate a random selection process. Ideally, we would want every story to have an equal chance of being selected. But every iteration of the closed search term must be tested using the same sample of stories. The most efficient way to re-run searches in Lexis-Nexis is to include every story retrieved by the search term for a given time period, such as one day. However, with forty-four newspapers, the chances of identical or similar stories appearing in the sample is quite large, thereby increasing the intra-class correlations and increasing sampling error. To minimize the chance that the stories in our sample were related to each other, but to maximize the efficiency of re-running our searches, we randomly selected a newspaper, then randomly selected ten days of coverage, until we had a sample greater than N=250 articles.

This sample was gathered using an open search term designed to capture any and all cancer news stories. Thus, every conceivable word that might be used to refer to cancer was gathered by reviewing a small sample of stories about cancer and from results of our earlier work. New words were added if discovered during the initial phases of searching.

The open search included the following words (an exclamation point is used for truncated searches in Lexis-Nexis):

> Body[25] (cancer! or leukemia! or lymphoma! or melanoma! or hodgkin! or tumor! or sarcoma! or carcino! or retinoblastoma! or adenoma! or astrocytoma! or blastoma! or glioma! or macroglobulinemia! or meningioma! or mesothelioma! or mycosis! or myelo! or neoplas! or neuroblastoma! or osteosarcoma! or pheochromocytoma! or rhabdomyosarcoma! or anticancer! or oncol!)

Decision criteria were developed for what constituted a relevant cancer story, and the first random sample of stories was reviewed by four

coders. Decision criteria for relevancy were discussed and revised through three successive random samples of stories containing between 250-325 stories retrieved by the open search term. Reliability among the four coders was established in the fourth random sample of 314 stories, with Krippendorff's alpha = .81. The fourth sample was set aside for validation purposes.

The three earlier reviews allowed us to estimate the sample size we would need for our final validation test. In our samples, approximately one out of every 3.4 stories was relevant. Having set our recall at .93, and the decision to set a 95% confidence interval at plus or minus 5%, we needed to include 225 relevant stories in our sample (Table 2). This meant that we would need to retrieve 765 (3.4 X 225) stories for the final validation process. Hence, after reliability was established, each of the four coders coded additional stories, yielding a final sample of 768 stories that could be used to validate the search term.[26]

The process of developing the final search term (the closed term) began with formative work to develop a candidate search term. It was further developed by reviewing the stories from one of the earlier samples (after re-reviewing the stories for relevancy once we had finalized our decision criteria), looking for key words that could reduce the number of irrelevant stories retrieved by the search term without eliminating the valid stories (for the search terms attempted during this formative period, see the Appendices).

Ultimately, our closed search term was the following:

> "OPEN TERM" and BODY(atleast 2 (cancer! or leukemia! or lymphoma! or melanoma! or hodgkin! or tumor! or sarcoma! or carcino! or retinoblastoma! or adenoma! or astrocytoma! or blastoma! or glioma! or macroglobulinemia! or meningioma! or mesothelioma! or mycosis! or myelo! or neoplas! or neuroblastoma! or osteosarcoma! or pheochromocytoma! or rhabdomyosarcoma! or anticancer! Or oncol!)) and not body((feline pre/1 leukemia) or (capricorn))

While the elements of this term are specific to this search, they may be analogous to what other searches will require. This search term retrieves all stories that mention cancer or cancer synonyms at least twice, which made it much more likely the story was really about cancer while thus filtering out some of the irrelevant stories . Additionally, the word "cancer" often came up in the context of horoscopes. To filter out references to horoscopes, any story which mentioned "capricorn" was not to be included. Similarly, stories about feline leukemia were filtered out by instruction to disregard text where the only occurrences of cancer synonyms was the word "leukemia" appearing immediately after the word "feline."

Using a sample of 303 stories, of which fifty-four were relevant, this search term produced a recall score of .93, and a precision score of .45. This was considered satisfactory for the purposes of this study, and

TABLE 3

Recall and Precision Estimates for Cancer Search Term

Citation	Not Relevant	Relevant	Total
Not Retrieved	411	14	435
Retrieved	122	221	333
Total	533	235	768

Precision = 221/333 = .67

Recall = 221/235 = .94

the search term was then applied to the 768 stories we had reserved for final validation. The results of this analysis can be seen in Table 3.

Our search term retrieves 94% (±5) of all cancer stories deemed valid by our decision criterion, and 67% of retrieved stories are relevant.

Although precision and recall estimates are not comparable, we can adjust the sample to get a reliable estimate of the number of cancer news stories appearing in 2003. Using the closed search term retrieved 26,874 stories. Because we filtered out fewer irrelevant stories than relevant ones, this adjustment corrects for our over-sampling. Using the proportion of recall to precision, we need to adjust our sample by .713. Hence, the true number of relevant stories is approximately 19,161. This adjustment factor can be applied to the number of articles published in any shorter period, if it can be shown that the recall and precision estimates are more or less the same over the entire time period studied. However, since the recall/precision ratio carries a sampling error with it, estimates that are based on brief time periods, while unbiased, may be subject to substantial error.

Discussion

We have outlined a procedure for providing estimates of sampling error associated with online searches. In the cancer example provided here, we utilized the validation procedure to gather the sample, but relied on human coders to code the measures. However, this procedure can also function as a reliable method for coding data. Content analysis is increasingly carried out using specially designed computer software, that typically counts or relates the occurrence of phrases selected by the analyst.[27] If the researcher is interested in counting stories on a certain topic within a broader population of stories, she may be justified in adding a specific word, phrase, or set of words to the original search phrase, to establish a count of stories on the smaller topic of interest. For example, Stryker[28] used one general term to gather the census of stories on marijuana use, and then two additional terms to examine the positive and negative consequences of marijuana use within those stories.

Research on topics ranging from marijuana use to drunk-driving,[29] mammography utilization,[30] and seat-belt use[31] suggests that search terms can be generated to capture thematic media frames, including pos-

itive or negative consequences of a specific behavior, attributions of responsibility, and advocated solutions to public problems. Generating these more specific frames requires additional validation testing (in the form of precision and recall) for each sub-category. Examples of non-health related issues that researchers have studied utilizing similar methods to analyze frames within a sample[32] include positive and negative news coverage of presidential campaigns,[33] different media frames used to describe the federal budget deficit,[34] and elite cues and media biases in presidential campaign coverage.[35] It must be recognized, however, that the use of any computer-assisted content analytic tools is subject to particular biases: while large quantities of data can be analyzed rapidly, it may come at the cost of having less refined measures, and may not necessarily produce results similar to human coders.[36]

We have presented a technique to assess systematically and quantitatively the quality of a search term in terms of relevant stories retrieved and extraneous ones excluded. We contend that these measures can strengthen the toolkit of content analysts, by establishing a way to estimate the sampling error around an online search term. At the same time, a number of cautions and caveats are worth noting.

An important component of developing an open search term is to begin with the known synonyms, metaphors, and word combinations, but also to review retrieved texts to locate additional key words. The appropriate search phrase must be sensitive to differences in the terminology used to describe a certain object, either between sources selected for inclusion in the study or within each source over time. For example, depending on political affiliation or leaning, some media outlets may use the term "conservative" to denote a "Republican" while others may not. Similarly, reference to "drunk-driving" within a single media outlet may have been replaced over time by acronyms such as "DUI" and "DWI." In both cases, without adjusting accordingly the search phrase used, estimates of actual media attention to these issues, as well as the corresponding estimates of recall and precision, are likely to be significantly biased.

While electronic databases offer a wealth of informational resources to the scholar, the results of ad hoc searches may be misleading for a number of reasons. As the neophyte researcher soon learns, the structure of database services and their associated search engines varies considerably. Even in a service as familiar as Lexis-Nexis or Dialog, the researcher needs to be conscious of date of entry of specific publications, particularly if using files that include combinations of publications. In other words, over time increasing numbers of publications have been added, changing the shape of the universe of concern.

The easy entry into electronic databases is seductive. Their digital form suggests static or error-free content that may not be altogether faithful to the original. It is interesting to note, for example, that the paper version of the New Jersey or West Coast editions of the *New York Times* that someone might read does not necessarily match the electronic version. The online version is taken from the late edition of the New York City version, which is different in certain respects from other edi-

tions.[37] Thus, it is important to be leery of the ease of using online services.

Recently, many print publications have been forced to choose between removing electronic traces of articles contributed by freelancers, or to pay royalties to those freelancers. Many have chosen to remove the freelance articles. In addition, online search companies may try to ease client searches by providing automated synonyms for search terms and retrieving texts with the synonyms as well. This can be helpful in locating extra relevant texts, but if the rules about what synonyms are included vary over time, searches done at different times may produce different results as more synonyms are included. All the more reason, we suggest, that search phrases be tested rigorously, and the searches be done at one time.

We restrict the proposed use of this technique to cases where a finite text-driven database is used, such as those available through Lexis-Nexis, Dialog, Ethnic NewsWatch, or the Vanderbilt Television Archives. We do not claim that the proposed method may be used in evaluating Internet searches, as the Web is that much more mutable, and not finite or defined in the way databases are. We argue that the proposed evaluation method will be useful in the specific context of online databases representing defined universes of written texts.

For researchers looking at mass media content, and the association of the media with individual and social phenomena, electronic databases are an invaluable resource. They enhance the power of content analysis by putting entire archives at the fingertips of researchers, making it possible to search large bodies of text efficiently and economically. But the ease of their use can be deceptive. Even as they simplify the work of content analysis, they can reduce the level of attention the researcher brings to the selection and analysis of items.

This study offers a technique to enhance the rigor of one part of this process: evaluating whether an online search phrase effectively locates texts that are relevant to the topic of concern. The proposed method takes an explicit measure of how well a search phrase selects relevant items and rejects extraneous ones. The resulting estimates of recall and precision together attest to the reliability and validity of a search phrase. Recall and precision show that the search phrase is valid, by revealing how the items that it selects are those the analyst is seeking. In the formal test, the reliability of the search phrase is examined.

Such transparency in selecting items for analysis can only benefit journalism and mass communication research. Coupled with the power of access that online services offer, the proposed technique strengthens the study of texts in a communication context. We hope that this contribution will enhance the rigor of content analysis, and thereby add to its value as a research method in social science inquiry.

APPENDIX A

List of Newspapers Used in Content Analysis

1. Atlanta Journal and Constitution
2. Baltimore Sun
3. Boston Globe
4. Boston Herald
5. Buffalo News
6. Charlotte Observer
7. Chicago Sun-Times
8. Chicago Tribune
9. Columbus Dispatch
10. Daily News (New York)
11. Dallas Morning News
12. Denver Post
13. Detroit Free Press
14. Fort Worth Star-Telegram
15. Houston Chronicle
16. Indianapolis Star
17. Investor's Business Daily
18. Kansas City Star
19. Los Angeles Times
20. Miami Herald
21. Milwaukee Journal Sentinel
22. New York Post
23. New York Times
24. Newsday
25. Orange County Register
26. Oregonian
27. Orlando Sentinel
28. Philadelphia Inquirer
29. Pittsburgh Post-Gazette
30. Plain Dealer
31. Rocky Mountain News
32. San Antonio Express-News
33. San Diego Union-Tribune
34. San Francisco Chronicle
35. San Jose Mercury News
36. Seattle Times
37. St. Louis Post-Dispatch
38. St. Petersburg Times
39. Star Tribune (Minneapolis MN)
40. Sun-Sentinel (Fort Lauderdale)
41. Tampa Tribune
42. Times-Picayune
43. USA Today
44. Washington Post

APPENDIX B

Decision Criteria for Cancer Relevance

CANCER "SYNONYMS": leukemia, lymphoma, melanoma, hodgkin's, sarcoma, carcinogen, retinoblastoma, adenoma, astrocytoma, meningioma, mesothelioma, mycosis, myeloma, neuroblastoma, osteosarcoma, pheochromocytoma, rhabdomyosarcoma, anticancer, oncolog(-y, -ist), precancerous lesion, actinic keratoses
NOTE: references to feline leukemia, astrology, cancer metaphorically DON'T count.

CANCER "PSEUDO-SYNONYMS"
WORDS THAT DO NOT NECESSARILY MEAN CANCER: tumor, adenoma, blastoma, glioma macroglobulinemia, neoplasm
Rules to follow:

a. If "malignant" precedes any of these words, then it IS cancer.
b. If these words are explicitly associated with CANCER once in the story, then all occurrences of the word count as references to cancer UNLESS explicitly discussed as NOT being cancer.

Decision rules for relevance

a. At least 2 paragraphs that EXPLICITLY mention CANCER, CANCER SYNONYMS, or follows rules for CANCER PSEUDO-SYNONYMS, or the explicit referent (e.g., "It," "her disease" (previously mentioned as cancer)) is to CANCER (or synonyms or follows rules for cancer pseudo-synonyms)?

If a is true, then:

i. Are the mentions strictly logistical information about an event?
YES, STORY IS NOT RELEVANT
ii. Is it a laundry list of events (community calendar)?
IF YES, STORY IS NOT RELEVANT
iii. Is it a laundry list of death notices?
IF YES, STORY IS NOT RELEVANT
iv. Is at least one of these paragraphs more than a sentence (If more than 2 paragraphs, rule does not apply)?
IF NO, STORY IS NOT RELEVANT

If a is NOT true, then:

i. Is a formatting error combining separate ideas into one paragraph, and 2 of those ideas make explicit mention of cancer, cancer synonyms, follows rules for cancer pseudo-synonyms, or explicitly reference any of the former?
IF YES, STORY IS RELEVANT

Decision rules for relevancy of news briefs

a. Two separate briefs cannot be combined to form 2 explicit mentions of cancer; one brief must have two explicit mentions.
b. If more than one brief is valid, the story should be coded twice; and call them "a" and "b" (i.e., 12a and 12b)
c. Is it a 1 paragraph story, OR is it a health brief that is one paragraph, and more than one sentence long?
IF YES, does it contain at least 2 explicit references IN SEPARATE SENTENCES to cancer, cancer synonyms or follow rules for cancer pseudo-synonyms (including headline)?
Does the paragraph contain more than strictly logistical information about an event?
IF YES, STORY IS RELEVANT

NOTES

1. Daniel Riffe, Stephen Lacy, and Frederick G. Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (Mahwah, NJ: Lawrence Erlbaum, Associates, 1998)

2. See, for example, Kimberly A. Neuendorff, *The Content Analysis Guidebook* (Thousand Oaks, CA: Sage, 2002); Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2d ed. (Thousand Oaks, CA: Sage, 2004); Riffe, Lacy, and Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research.*

3. Daniel Riffe and Alan Freitag, "A Content Analysis of Content Analyses: Twenty-Five Years of Journalism Quarterly," *Journalism & Mass Communication Quarterly* 74 (winter 1997): 873-82.

4. Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken, "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research* 28 (October 2002): 587-604.

5. Riffe, Lacy, and Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research.*

6. Riffe and Freitag, "A Content Analysis of Content Analyses: Twenty-Five Years of *Journalism Quarterly.*"

7. See, for example, Ricardo Wray, Kimberly Maxwell, and Robert Hornik, "Validation of On-Line Searches of Media Coverage: An Approach to Evaluation with an Example of Reporting Domestic Violence" (paper presented at the International Communication Association, Jerusalem, July 1998); Itzhak Yanovitsky and Jo Ellen Stryker, "Mass Media, Social Norms, and Health Promotion Efforts: A Longitudinal Study of Media Effects on Youth Binge Drinking," *Communication Research* 28 (2001): 208-39; Jo Ellen Stryker, "A Longitudinal Analysis of the Effects of News Media Messages on Health Behaviors" (Ph.D. diss., University of Pennsylvania, 2001); Jo Ellen Stryker, "Media and Marijuana: A Longitudinal Analysis of News Media Effects on Adolescents' Marijuana Use and Related Outcomes, 1977-1999," *Journal of Health Communication* 8 (2003): 305-28; Itzhak Yanovitsky, "Effect of News Coverage on the Prevalence of Drunk-Driving Behavior: Evidence from a Longitudinal Study," *Journal of Studies on Alcohol* 63 (2002): 342-51.

8. In our review of content analyses published in the last five years, no other authors reported quantitative estimates of the ability of a search term to return relevant stories and reject extraneous ones.

9. See, for example, Paula M. Poindexter and Maxwell E. McCombs, *Research in Mass Communication: A Practical Guide* (Boston, MA: St. Martin's, 2000); Carl W. Roberts, ed., *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences fromTexts and Transcripts* (Mahwah, NJ: Lawrence Erlbaum Associates, 1997).

10. Tom Koch, *Journalism for the 21st Century: Online Information, Electronic Databases and the News* (New York: Praeger, 1991); Charles T. Meadow, *Text Information Retrieval Systems* (San Diego, CA: Academic Press, Inc., 1992).

11. Koch, *Journalism for the 21st Century: Online Information, Electronic*

Databases and the News, 201.

12. Gerald Salton and Michael J. McGill, *Introduction to Modern Information Retrieval* (New York, NY: McGraw-Hill, 1983).

13. Meadow, *Text Information Retrieval Systems.*

14. Alice Y. Charmis, *Vocabulary Control and Search Strategies in Online Searching* (Westport, CT: Greenwood, 1991); Meadow, *Text Information Retrieval Systems*; Salton and McGill, *Introduction to Modern Information Retrieval.*

15. Charles R. Hildreth, *Intelligent Interfaces and Retrieval Methods for Subject Searching in Bibliographic Retrieval Systems* (Washington, DC: Library of Congress, 1989).

16. Meadow, *Text Information Retrieval Systems.*

17. Chris J. Armstrong and Andrew J. Large, *Manual of Online Search Strategies* (Boston, MA: G.K. Hall and Company, 1988).

18. Meadow, *Text Information Retrieval Systems.*

19. Wray, Maxwell, and Hornik, "Validation of On-Line Searches of Media Coverage: An Approach to Evaluation with an Example of Reporting Domestic Violence."

20. Krippendorf, *Content Analysis: An Introduction to Its Methodology.*

21. For a good discussion about the internal consistency criterion, and its lack of application in computerized content analysis, see Mike Conway, "The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis," *Journalism & Mass Communication Quarterly* 83 (spring 2006): 186-200.

22. Salton and McGill, *Introduction to Modern Information Retrieval.*

23. See, for example, Armstrong and Large, *Manual of Online Search Strategies*; Koch, *Journalism for the 21st Century: Online Information, Electronic Databases and the News;* Meadow, *Text Information Retrieval Systems;* Roberts, ed., *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts.*

24. Armstrong and Large, *Manual of Online Search Strategies.*

25. The term "Body" specifies that only the text body be searched. Lexis-Nexis catalogs stories through a system of "intelligent searching." This means that staff librarians read through stories and link related concepts to the search term. For example, "tumor" is linked to "cancer," such that a search term containing the word cancer will retrieve stories about tumors even if tumor is not part of the search phrase. Librarians are constantly updating their linked files, so a search term could retrieve a different number of stories if applied on different days (assuming that both searches are for the same time period). The solution to this problem is to search the body of the news stories only. Ideally, a search would include both the headline and the body, but this is not an available option.

26. Although our calculated estimate suggested that we only needed to retrieve 765 articles, an even distribution of stories among four coders resulted in three additional stories.

27. Roberts, ed., *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts.*

28. Stryker, "Media and Marijuana: A Longitudinal Analysis of

News Media Effects on Adolescents' Marijuana Use and Related Outcomes, 1977-1999."

29. Yanovitsky and Stryker, "Mass Media, Social Norms, and Health Promotion Efforts: A Longitudinal Study of Media Effects on Youth Binge Drinking"; Yanovitsky, "Effect of News Coverage on the Prevalence of Drunk-Driving Behavior: Evidence from a Longitudinal Study."

30. Itzhak Yanovitsky and Cynthia Blitz, "Effect of Media Coverage and Physician Advice on Utilization of Breast Cancer Screening by Women 40 Years and Older," *Journal of Health Communication* 5 (2000): 117-34.

31. Stryker, "A Longitudinal Analysis of the Effects of News Media Messages on Health Behaviors."

32. The similar methods that we are referring to rely on electronic search phrases to identify relevant themes in texts, and make use of David Fan's Infotrend Program. However, there are several key distinguishing features. First, this method can only be used with access to the specialized software, whereas our proposed method requires no such access. Second, while the research requires a filtering stage akin to developing precision, there is no standardized method for reporting this statistic.

33. David Domke, David P. Fan, Michael Fibison, Dhavan V. Shah, Steven S. Smith, and Mark D. Watts, "News Media, Candidates and Issues, and Public Opinion in the 1996 Presidential Campaign, *Journalism & Mass Communication Quarterly* 74 (winter 1997): 718-37.

34. Amy E. Jasperson, Dhavan V. Shah, Mark Watts, Ronald J. Faber, and David P. Fan, "Framing and the Public Agenda: Media Effects on the Importance of the Federal Budget Deficit," *Political Communciation* 15 (1998): 205-24.

35. Mark D. Watts, David Domke, Dhavan V. Shah, and David P. Fan, "Elite Cues and Media Bias in Presidential Campaigns: Explaining Public Perceptions of a Liberal Press," *Communication Research* 26 (April 1999): 144-75.

36. Conway, "The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis."

37. Kathleen A. Hansen, "Online Inaccuracies: The Use and Misuse of Electronic Information Sources" (paper presented at the annual meeting of AEJMC, Washington, DC, August 1995).