Published in final edited form as:

Obesity (Silver Spring). 2015 February ; 23(2): 296–300. doi:10.1002/oby.20955.

Large-Scale Automated Analysis of News Media: A Novel Computational Method for Obesity Policy Research

Rita Hamad, MD, MPH, MS[a], **Jennifer L. Pomeranz, JD, MPH**[b], **Arjumand Siddiqi, ScD**[c,d], and **Sanjay Basu, MD, PhD**[e,f]

[b]Temple University, Center for Obesity Research and Education, Department of Public Health

[c]University of Toronto, Dalla Lana School of Public Health

[d]University of North Carolina, Gillings School of Global Public Health

[e]Stanford University, Prevention Research Center

[f]London School of Hygiene and Tropical Medicine, Department of Public Health and Policy

Abstract

Objective—Analyzing news media allows obesity policy researchers to understand popular conceptions about obesity, which is important for targeting health education and policies. A persistent dilemma is that investigators have to read and manually classify thousands of individual news articles to identify how obesity and obesity-related policy proposals may be described to the public in the media. We demonstrate a novel method called “automated content analysis” that permits researchers to train computers to “read” and classify massive volumes of documents.

Methods—We identified 14,302 newspaper articles that mentioned the word “obesity” during 2011–2012. We examined four states that vary in obesity prevalence and policy (Alabama, California, New Jersey, and North Carolina). We tested the reliability of an automated program to categorize the media’s “framing” of obesity as an individual-level problem (e.g., diet) and/or an environmental-level problem (e.g., obesogenic environment).

Results—The automated program performed similarly to human coders. The proportion of articles with individual-level framing (27.7–31.0%) was higher than the proportion with neutral (18.0–22.1%) or environmental-level framing (16.0–16.4%) across all states and over the entire study period ($p<0.05$).

Conclusion—We demonstrate a novel approach to the study of how obesity concepts are communicated and propagated in news media.

[a]Corresponding author: Stanford University, Division of General Medical Disciplines, 1070 Arastradero Road, Palo Alto, CA 94304, USA; rhamad@stanford.edu.

RH contributed to the literature review; study design; data collection, analysis, and interpretation, and manuscript preparation. SB, JLP, and AS contributed to study design, data interpretation, and manuscript preparation. All authors approved of the submitted manuscript.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

Keywords

Health policy; research methods; research design; longitudinal; trends

INTRODUCTION

The news media influences public perceptions about obesity and obesity-related policy proposals (1). Media studies have observed that the manner in which newspapers frame obesity as either an individual problem or a consequence of an "obesogenic" environment influences readers to support or oppose policy proposals, such as soda taxes (2, 3, 4). Analyzing how the media frames obesity policies is arduous, however. Commonly, investigators review and classify ("code") individual articles manually, limiting most studies to single locations, brief time periods, and non-representative subsamples of media outlets (5, 6, 7).

Here, we demonstrate the use of a novel analytic method, "automated content analysis," developed and validated in the field of computational linguistics (8, 9, 10). The method allows large-scale media analysis using supervised machine learning, a process enabling a computer to "learn" from human coders to identify how different word clusters express an opinion or frame a problem. The computer applies this knowledge to classify an unlimited number of text documents. We demonstrate the reliability and scalability of this method by applying it to newspaper articles published over a two-year period in four states that differ in obesity prevalence rates and that vary in their passage of Commonsense Consumption Acts – laws that emphasize the role of "personal responsibility" and disallow lawsuits against food establishments (11, 12).

METHODS

We gathered all newspaper articles (N= 14,302) that included the word "obesity" and were published in 2011–2012 from Access World News, an online news database available at http://infoweb.newsbank.com. Four states were included: Alabama and North Carolina (in which CCAs had been passed), and California and New Jersey (in which CCAs had not been passed).

We established four mutually exclusive categories to code the articles: (1) articles in which the attribution of or solution to obesity are at the individual level; (2) articles focusing exclusively on environmental or systemic causes; (3) articles mentioning both of these frames (i.e., "neutral" articles); or (4) irrelevant articles. The latter category ensured that the list was exhaustive and included articles that mentioned obesity but did not discuss framing or attributions. An inter-class correlation coefficient of 0.80 was achieved among four coders implementing the categorization scheme on a sample of 200 articles, indicating an acceptable level of inter-rater reliability. Hand-coding was then performed on a subset of 354 articles, a sufficient number to minimize the root mean square error (13). These were used to "train" the program.

We applied an automated content analysis algorithm that first strips all words in the documents to their stems (e.g., “writing” and “written” become “writ”). The algorithm then creates a word stem profile, **S**, for each document, representing the “bag of words” contained in the document. The framing of each document (i.e., individual-, environmental-, or neutral characterization of obesity) is represented as **D**. The frequency distribution, P(**S**), can be expressed as:

$$\mathrm{P}(\mathbf{S})=\mathrm{P}(\mathbf{S}|\mathbf{D})*\mathrm{P}(\mathbf{D})$$

The algorithm estimates P(**D**), the distribution of articles among predetermined categories, by first calculating P(**S**), the frequency distribution of the words in the documents, then dividing by P(**S**|**D**), which is obtained from the “training set” of hand-coded documents provided to the program by the researchers. Importantly, by using a Bayesian approach, this method does not require that the training set be representative of the larger data set, only that it be drawn from the same population of texts. A formal derivation and proof have been published previously (13).

Comparisons were performed at the state level, the level at which the CCA policies were enacted. The algorithm tabulated the proportion of articles in each category in each state over time, expressed in six-month blocks to ensure a large enough sample for standard error calculations. Bootstrapped standard errors were estimated through 300 replications. The analysis was performed using the content analysis algorithm included in the package ReadMe in the statistical program *R* (v. 3.1.0, R Foundation for Statistical Computing, Vienna).

RESULTS

Of the 14,302 articles mentioning “obesity” during the study period, 9,598 were deemed relevant by the algorithm: 822 in Alabama, 5,554 in California, 1,481 in New Jersey, and 1,741 in North Carolina (Supplemental Figure S1).

In each state, the proportion of articles with individual-level framing (27.7–31.0%) was significantly higher than those with neutral framing (18.0–22.1%) or environmental-level framing (16.0–16.4%) ($p<0.05$). The distribution of articles into categories as tabulated by the automated algorithm matched the distribution by human hand-coders for the training set (Figure 1). There were surprisingly no significant differences across states despite differing policy climates, contrary to our *a priori* hypothesis (Figure 1). In all but the last time period, there was a significantly higher proportion of articles with individual-level framing relative to environmental-framing in Alabama ($p<0.05$) (Figure 2). In California, New Jersey, and North Carolina, articles with individual-level framing significantly outnumbered articles with environmental-level framing and neutral framing at the majority of time points ($p<0.05$). During each of the four time periods, there were no significant differences in each framing category across states (Figure 3).

DISCUSSION

In this study, we demonstrate the use of a novel method for large-scale media analysis. This overcomes the challenge of hand-coding large volumes of documents, which has limited previous research to single locations, brief time periods, and non-representative subsamples of media outlets. This method "learns" from researchers' classifications of documents, then "reads" large volumes of text to apply the coding scheme.

Using a publicly available automated content analysis program, we demonstrate that this approach reliably "learns" from and matches the findings of hand-coders, consistent with prior literature that has validated this method in political science and sociology studies (8, 9, 10). When applied to media articles on obesity, we found that newspaper articles from states with differing policy climates consistently attributed obesity to individual-level responsibility rather than environmental factors or both. Testing the hypothesis that these states differed in their media framing would typically require months or years for hand-coders, but took just days on a university server.

In addition to processing large numbers of articles of any length, there are several advantages to this novel method. The hand-coded articles do not have to be representative of the larger corpora of documents to provide an accurate estimate of the distribution of document classifications, because the method employs a Bayesian approach that does not assume representativeness of the training set from which it "learns." The estimation procedure also allows the calculation of standard errors to more confidently make statistical inferences across time and space. Moreover, methods that code small samples of individual articles and then infer proportions at the population level likely result in biased estimates, while the algorithm we employ has been shown to give unbiased and statistically consistent estimates of document category proportions (13). Unlike unsupervised machine learning, this supervised technique allows researchers to define the categories of interest rather than having a computer define these categories.

Those interested in implementing the method should be aware of three limitations: (a) the algorithm is computationally intensive, requiring several days on a university server to calculate bootstrapped standard errors; (b) the method is limited by its investigator-designed coding scheme, which in our case included only four categories that may have limited our ability to detect variation (i.e., combining vehement editorials with mildly worded articles); and (c) the algorithm estimates frequencies of categories across the entire body of articles rather than labeling each article, which is not useful for researchers trying to find specific text in a volume of documents.

Despite these limitations, our application of the method has important implications for obesity policy researchers. In addition to offering a new method to extend prior analyses of obesity policy research limited to single locations (6), our unexpected finding that obesity media coverage is uniformly focused on individual-level framing across states with differing policy climates reveals that the public health community's emphasis on environmental contributors to obesity (14) is not conveyed in the media. Previous studies have been contradictory with regard to the predominance of individual-level or environmental-level

frames in the U.S. Some have found more frequent individual-level framing in social media accounts (15), polling of the general public (16), and the media (17, 18). Others suggest a predominance of environmental framing in mass media based on small non-representative samples of documents (136 and 83, respectively) (19, 20). Our larger-scale study using automated content analysis allows us to clarify debates in the literature by characterizing a volume of articles that would be practically impossible through hand-coding.

CONCLUSION

We demonstrate a novel method to conduct large-scale analysis of news media. Our research sets the stage for future comparative studies that may serve to improve health education and the targeting of evidence-based obesity policies, overcoming the gap between scientific evidence and the translation or dissemination of science into the public sphere.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Ashley Geo and Ashley Overbeek for their assistance in hand-coding articles for this research.

References

1. McCombs ME, Shaw DL. The agenda-setting function of mass media. Public Opin Q. 1972; 36:176–187.
2. Barry CL, Brescoll VL, Brownell KD, Schlesinger M. Obesity metaphors: how beliefs about the causes of obesity affect support for public policy. Milbank Q. 2009; 87:7–47. [PubMed: 19298414]
3. Dorfman L, Wallack L, Woodruff K. More than a message: Framing public health advocacy to change corporate practices. Health Educ Behav. 2005; 32:320–336. [PubMed: 15851542]
4. Gollust SE, Niederdeppe J, Barry CL. Framing the consequences of childhood obesity to increase public support for obesity prevention policy. Am J Public Health. 2013; 103:e96–e102. [PubMed: 24028237]
5. Gollust SE, Eboh I, Barry CL. Picturing obesity: Analyzing the social epidemiology of obesity conveyed through US news media images. Soc Sci Med. 2012; 74:1544–1551. [PubMed: 22445762]
6. Mejia, P.; Nixon, L.; Cheyne, A.; Dorfman, L.; Quintero, F. Two communities, two debates: news coverage of soda tax proposals in Richmond and El Monte. Berkeley Media Studies Group; Berkeley, California: 2014.
7. Roberto, C.; Camp, C.; Gagnon, G.; Werth, P. A Content Analysis of Public Discourse About the New York City Sugar-Sweetened Beverage Portion Limit Policy. Annual Scientific Meeting of The Obesity Society; 2013;
8. Ceron A, Curini L, Iacus SM, Porro G. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society. 2014; 16:340–358.
9. Grimmer J, Messing S, Westwood SJ. How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation. American Political Science Review. 2012; 106:703–719.
10. King G, Pan J, Roberts ME. How censorship in China allows government criticism but silences collective expression. American Political Science Review. 2013; 107:326–343.

11. Adams R. Fast food, obesity, and tort reform: An examination of industry responsibility for public health. Business and Society Review. 2005; 110:297–320.
12. Wilking CL, Daynard RA. Beyond Cheeseburgers: The Impact of Commonsense Consumption Acts on Future Obesity-Related Lawsuits. Food & Drug LJ. 2013; 68:229.
13. Hopkins DJ, King G. A method of automated nonparametric content analysis for social science. American Journal of Political Science. 2010; 54:229–247.
14. Johnston LM, Matteson CL, Finegood DT. Systems Science and Obesity Policy: A Novel Framework for Analyzing and Rethinking Population-Level Planning. Am J Public Health. 2014; 104:1270–1278. [PubMed: 24832406]
15. Harris JK, Moreland-Russell S, Tabak RG, Ruhr LR, Maier RC. Communication About Childhood Obesity on Twitter. Am J Public Health. 2014; 104:e62–e69. [PubMed: 24832138]
16. Saad L. Public balks at obesity lawsuits. Gallup. 2003
17. Saguy, AC.; Almeling, R. Fat in the Fire? Science, the News Media, and the “Obesity Epidemic” 2. Wiley Online Library; 2008. p. 53-83.
18. Barry CL, Jarlenski M, Grob R, Schlesinger M, Gollust SE. News media framing of childhood obesity in the United States from 2000 to 2009. Pediatrics. 2011; 128:132–145. [PubMed: 21690111]
19. Lawrence RG. Framing Obesity The Evolution of News Discourse on a Public Health Issue. The Harvard International Journal of Press/Politics. 2004; 9:56–75.
20. Ries NM, Rachul C, Caulfield T. Newspaper reporting on legislative and policy interventions to address obesity: United States, Canada, and the United Kingdom. J Public Health Policy. 2011; 32:73–90. [PubMed: 21109764]

WHAT IS ALREADY KNOWN ABOUT THIS SUBJECT

- The language and "framing" of news media stories about obesity are thought to critically influence individual behavior and public policy.
- Comprehensively studying how the media discusses obesity as a problem and describes potential policy solutions is an arduous task limited by the absence of methods to analyze large volumes of text.
- Understanding differences in media discussions about obesity across different locations and times may facilitate a targeted approach to health education and enhance our understanding of how obesity research is disseminated to the public.

WHAT THIS STUDY ADDS

- We demonstrate the use of a novel, free computational method called automated content analysis, which allows a computer to "learn" from researchers' classification of media stories into categories, "reading" large volumes of media documents and categorizing them according to the human-selected coding scheme.
- We apply the method to analyze all newspaper articles related to obesity in four states over a two-year period, demonstrating the reliability of the method and describing how it can be used to characterize different popular conceptions about obesity across place and time.

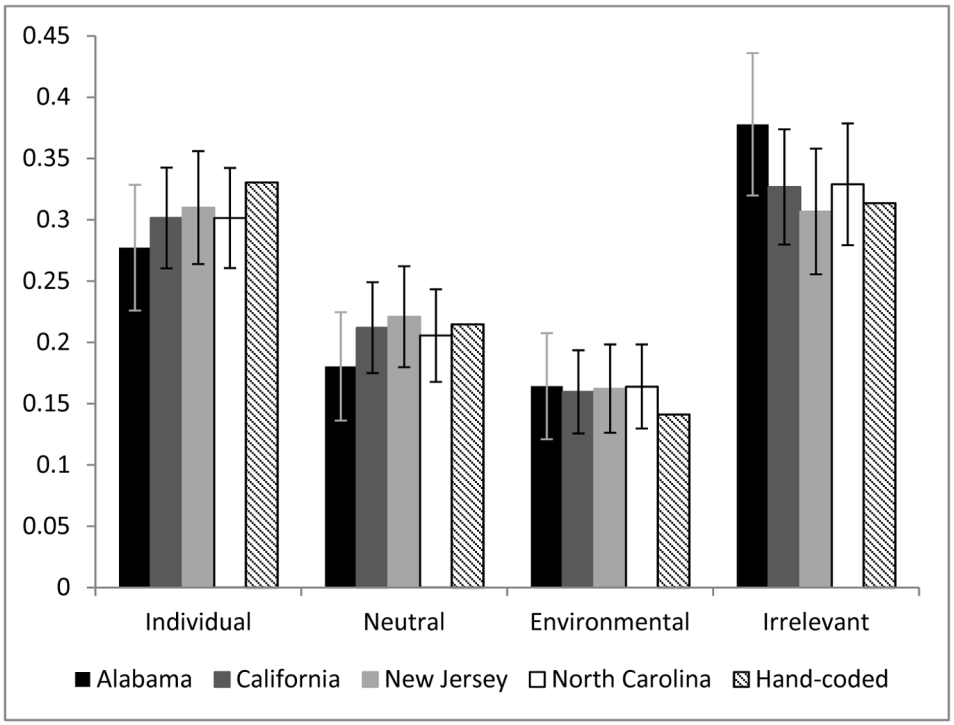


Figure 1. Overall proportion of articles in media framing categories by state (2011–12)

Note: Error bars represent 95% confidence intervals calculated using bootstrapping, which is not possible for hand-coded articles.

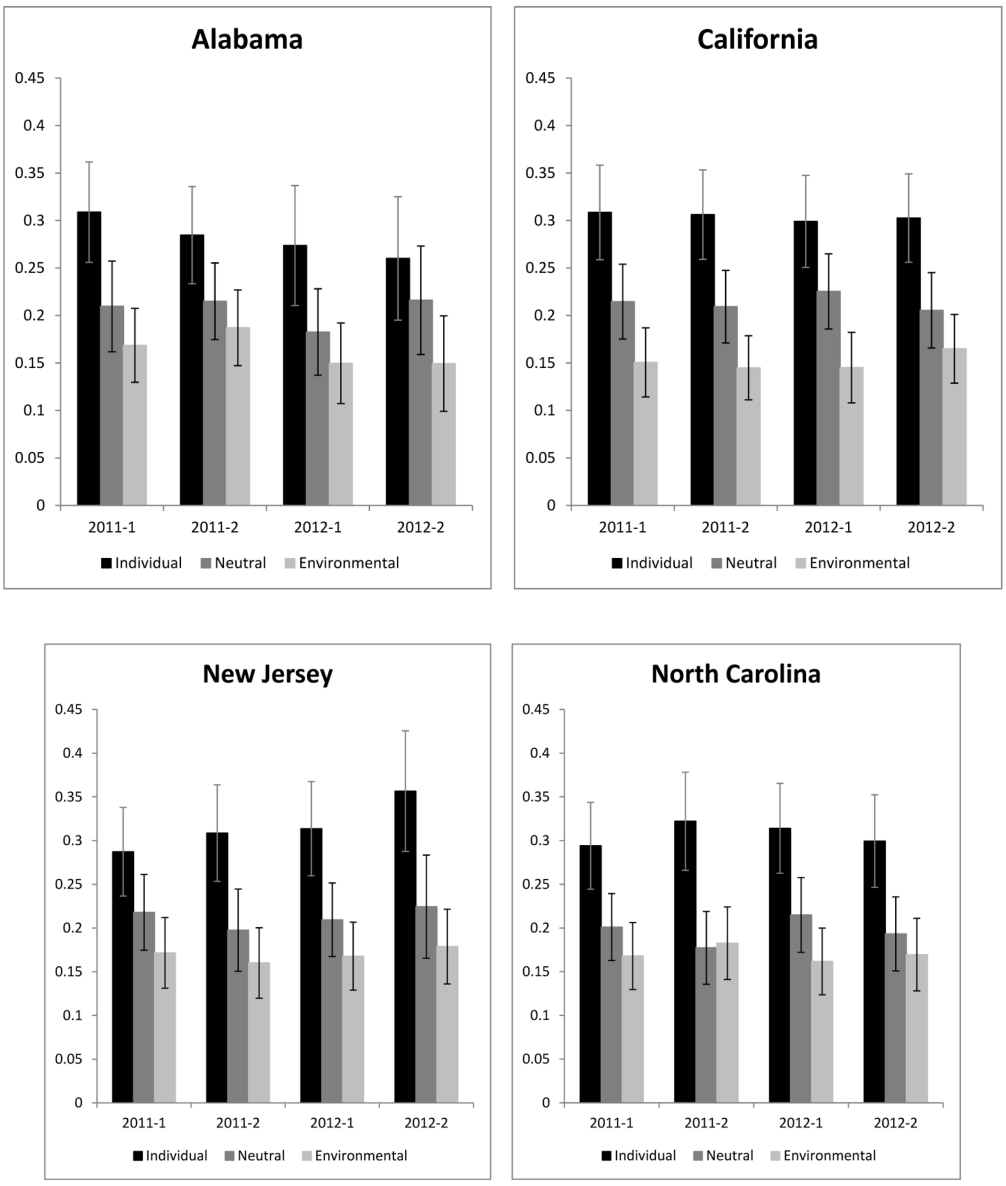
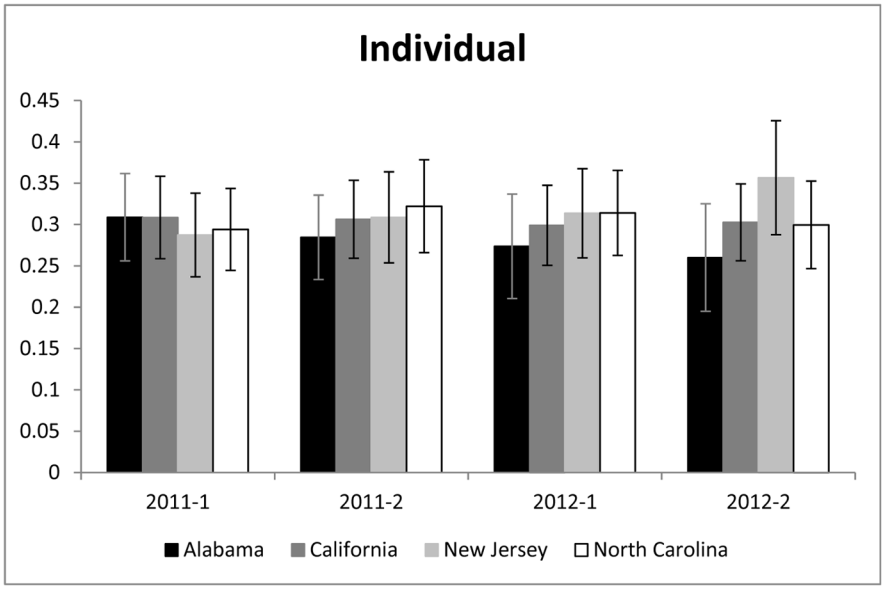
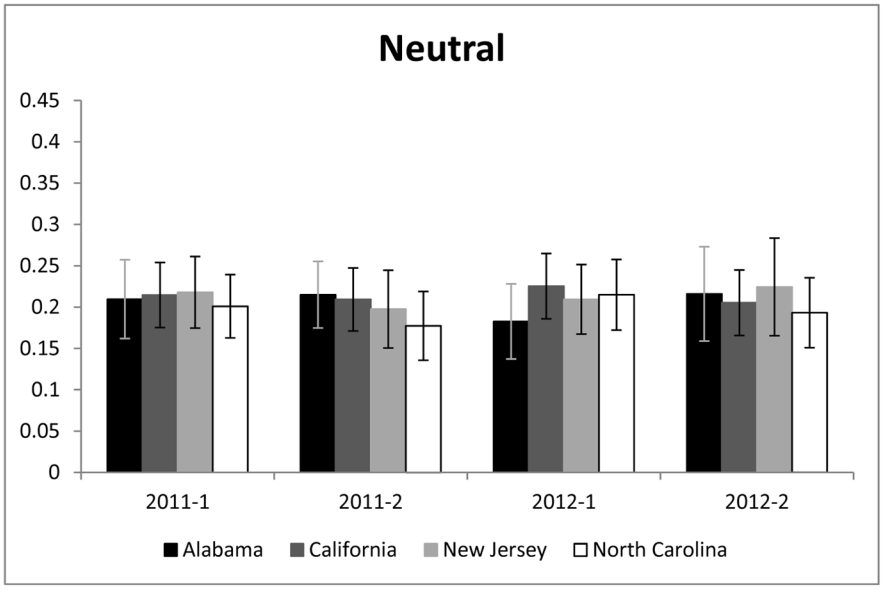


Figure 2. Longitudinal distribution of proportion of articles in media framing categories by state
Note: Error bars represent 95% confidence intervals, calculated using a bootstrapping procedure. “Irrelevant” category not displayed.





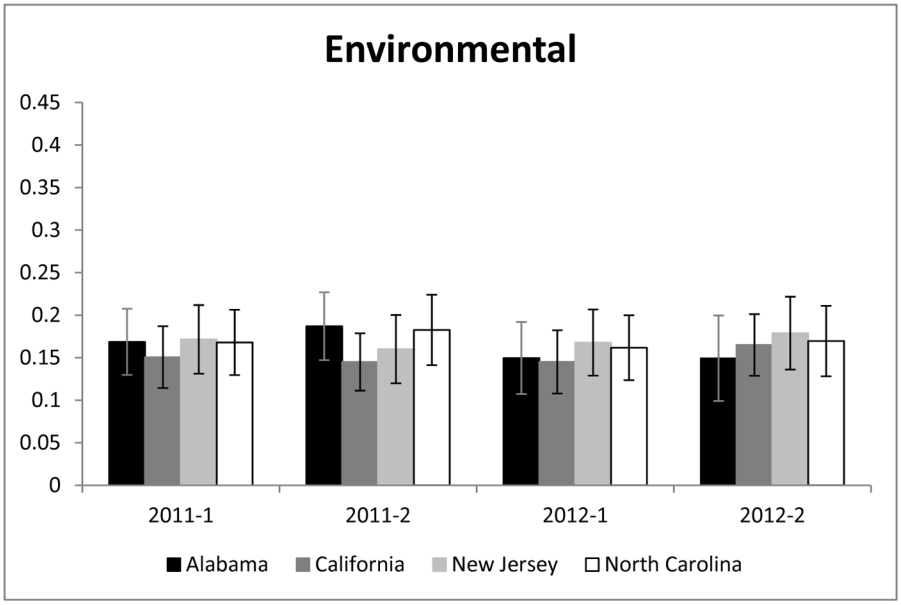


Figure 3. Longitudinal distribution of proportion of articles, by media framing category and state

Note: Error bars represent 95% confidence intervals, calculated using a bootstrapping procedure. “Irrelevant” category not displayed.