RESEARCH ARTICLE

Six Americas short survey

Breanne Chryst[a,b], Jennifer Marlon[b], Sander van der Linden[c], Anthony Leiserowitz[b], Edward Maibach[d], Connie Roser-Renouf[d]

[a]Department of Statistics, Yale University, New Haven, CT 06511, USA; [b]Yale Program on Climate Change Communication, Yale University, New Haven, CT 06511, USA; [c] Department of Psychology, Downing Site, University of Cambridge CB2 3EB, Cambridge, UK; [d] Center for Climate Change Communication, George Mason University, Fairfax, VA 22030, USA

ARTICLE HISTORY
Compiled July 11, 2018

ABSTRACT
Audience segmentation has long been used in marketing, public health, and communication, and is now becoming an important tool in the environmental domain as well. Global Warming's Six Americas is a well-established segmentation of Americans based on their climate change beliefs, attitudes, and behaviors. The original Six Americas model requires a 36 question-screener and although there is increasing interest in using these segments to guide education and outreach efforts, the number of survey items required is a deterrent. Using 14 national samples and machine learning algorithms, we identify a subset of four questions from the original 36, the Six Americas Short SurveY (SASSY), that accurately segment survey respondents into the Six Americas categories. The four items cover respondents' global warming risk perceptions, worry, expected harm to future generations, and personal importance of the issue. The true positive accuracy rate for the model ranges between 70% and 87% across the six segments on a 20% hold-out set. Similar results were achieved with four out-of-sample validation data sets. In addition, the screener showed test-retest reliability on an independent, two-wave sample. To facilitate further research and outreach, we provide a web-based application of the new short-screener.

KEYWORDS
Six Americas; Global Warming; Segmentation

1. Introduction

The majority of climate scientists have concluded that human-caused climate change is happening and poses serious risks to society (Cook et al., 2016; Pachauri et al., 2014). Addressing global climate change will therefore require urgent and substantial changes in human behavior, decision-making, and policy-support (van der Linden, Maibach, & Leiserowitz, 2015). In order to enable more effective mitigation and adaptation, it is important to understand how the public thinks, feels, and acts on the issue of global warming. The efficacy of public engagement programs typically improves with the ability to tailor specific messages to a well-defined target audience (Hine et al., 2014). Tailored information is generally perceived as more credible, and it is more

CONTACT Breanne Chryst. Email: breanne.chryst@yale.edu

likely to be read and recalled. Communication campaigns that employ audience segmentation have proven successful not only in changing perceptions, but in changing behavior, in domains from politics to public health (Harris, Lock, Phillips, Reynolds, & Reynolds, 2010; Maibach, Weber, Massett, Hancock, & Price, 2006; Noar, Benac, & Harris, 2007). The process of segmentation involves identifying (within a target population) relatively homogeneous subgroups that share similar psychographic profiles. Understanding the unique beliefs, attitudes, and behaviors of each subgroup then allows for the development of tailored frames and communications. For example, prior research has shown that framing climate change as a health, national security, or environmental issue can have diverging effects on different audiences (Myers, Nisbet, Maibach, & Leiserowitz, 2012).

A comprehensive and well-known inventory of American views on global warming was conducted jointly by the Yale Program on Climate Change Communication and the George Mason University Center for Climate Change Communication (Maibach, Leiserowitz, Roser-Renouf, & Mertz, 2011). Since 2008, 14 nationally-representative surveys of American adults that include a large number of identical questions have been carried out (see Leiserowitz, Maibach, Roser-Renouf, Feinberg, and Rosenthal (2016)). An audience segmentation analysis of the first survey (n=2,164) that was applied to each subsequent survey (n>18,000) identified six unique "interpretive communities" within society who each respond to the issue of global warming in their own distinct ways; the "Six Americas" (Leiserowitz, 2005; Maibach et al., 2011; Roser-Renouf, Stenhouse, Rolfe-Redding, Maibach, & Leiserowitz, 2014). The original Six Americas model used 36 variables to classify respondents into segments based on Latent Class Analysis (LCA) using the LatentGold 4.5 software (Magidson & Vermunt, 2002; Vermunt & Magidson, 2002) (Maibach et al., 2011). The survey items cover a wide range of global warming beliefs, risk perceptions, policy support, and behaviors. The six segments range from those very concerned about global warming to those who are strongly opposed to taking action against global warming (Fig. 1). The six segments are the Alarmed, the Concerned, the Cautious, the Disengaged, the Doubtful, and the Dismissive. Each segment differs meaningfully in their beliefs, attitudes, issue involvement, behaviors, and policy-preferences about climate change (Maibach et al., 2011; Roser-Renouf et al., 2014). For example, the Alarmed segment includes those individuals who are most convinced that human-caused climate change is happening, are highly engaged with the issue, and ready to take action. In contrast, the Dismissive are on the other end of the spectrum, and strongly believe that global warming is not happening or human-caused and actively oppose any action on climate change (Maibach et al., 2011). Individuals in the middle four groups vary in their sense of urgency or certainty about the problem, and tend to display lower personal and political engagement than those in the extreme categories (Leiserowitz, 2005).

FIGURE 1 HERE

The Six Americas instrument was originally developed to profile the American population, and has been used for that purpose to support an array of research, education, and communication efforts by scholars, practitioners, and decision makers (Akerlof, Bruff, & Witte, 2011; Costello, 2014; Fisher, 2014; Leiserowitz et al., 2016; Maibach et al., 2011). The method has also been used to profile specific sub-populations, such as American zoo, national park, and aquarium visitors (Kelly et al., 2014; Schweizer, Davis, & Thompson, 2013) and agricultural agents (Bowers, Monroe, & Adams, 2016), it's been used to study specific audience segments, such as the the Alarmed (Doherty

& Webler, 2016), and used changes in audience composition as an outcome measure to assess the impact of targeted education interventions (Flora et al., 2014).

Recently, a growing interest in the value of tailoring communications about climate change to specific audiences has led to the development of segmentation studies around the world (Hine et al., 2014). For example, one study in Australia identified four segments in the Australian population based on knowledge of and concern about climate change (Ashworth, Jeanneret, Gardner, & Shaw, 2011). A more recent study identified "Six Australias", analogous to the Six Americas (Morrison, Duncan, Sherley, & Parton, 2013). In Germany, five segments - notably missing a dismissive group - were identified, based on beliefs, attitudes, and media consumption (Metag, Füchslin, & Schäfer, 2015). Six segments were also identified in India, using clustering analysis, ranging from the Informed to the Disengaged (Leiserowitz, Thaker, Feinberg, & Cooper, 2013). A recent study in Singapore identified three segments, the concerned, the disengaged, and the passive (Detenber, Rosenthal, Liao, & Ho, 2016) and the British Broadcasting Company conducted a segmentation analysis with over 33,000 residents from six countries in Asia (BBC Media Action., 2013) with the explicit aim of developing more effective communication strategies.[1] Segmentation analyses have also been applied to various sub-populations, such as US corn belt farmers' views on climate change (Arbuckle et al., 2014). In short, regardless of the population or exact number of segments, scholars around the world have found audience segmentation to be a valuable tool, whether for assessing current issue understanding, developing communication strategies, or developing new messages to advance dialogue and action.

The effectiveness of a segmentation tool depends on the development of a measure that is concise, reliable, and valid in describing individual differences in public opinion, including cognitive and affective issue engagement, and behavior (Slater, 1996). This need, and the growing interest in climate change audience segmentation in particular, motivates our current work. In particular, more empirical research on shorter versions of the original 36-item screener have been called for (Hine et al., 2014), especially since the length of the full 36-item survey may be prohibitive in many research studies. Moreover, in light of "Big Data" opportunities, there is an increasing demand for short measures of psychological constructs and scales that are not cognitively taxing or time-consuming. A shorter Six Americas survey instrument would make segmentation feasible for diverse researchers, and would allow them to quickly gauge the range and distribution of climate opinions held by audiences of interest. Previous work by Maibach, et al. identified a reduced set of 15 items, the Reduced Discriminant Model Tool (RDM), for identifying the Six Americas (Maibach et al., 2011). The principal aim of this research is to identify the smallest subset of the 15-item RDM Tool capable of identifying each of the six segments with sufficient accuracy, defined as roughly 70% accurately categorized respondents in each of the six segments on out-of-sample validation data (Fawcett, 2006; Kleinbaum & Klein, 2010). Accordingly, we advance a new screener tool that achieves this goal with only four survey items.

[1] It should be noted that attitudes toward climate change in Western nations are structured differently than in non-Western nations.

2. Methods

2.1. Data and Sample

To develop the Six Americas short survey we use data from 14 nationally representative opinion surveys of American adults conducted between 2008 and 2016 (Table 1). Thirteen of the fourteen surveys were probability-based and conducted online (by GfK Knowledge Networks). One survey was conducted by telephone using the same wording as the online panel surveys (by Abt SRBI). All observations that were missing an assigned Six Americas segmentation were excluded from the analysis (423 observations across 9 surveys). We focus our analysis on the fifteen questions from the Reduced Discriminant Model (Maibach et al., 2011). In total, there are 41 item non-responses, roughly 0.3% of the data. These missing observations were filled using hot deck imputation (Myers, 2011), which fills the missing responses with those from respondents that are otherwise similar, based on their responses to the other survey items (Cranmer, Gill, Jackson, Murr, & Armstrong, 2016; Cranmer & Gill, 2013).

We evaluate the test-retest reliability of the model using observations collected by the George Mason University Center for Climate Change Communication in February and March of 2017. This survey was administered online with a 14-21 day follow-up. We focus our test-retest analysis on the control group from this experiment (n=241). This research received ethical approval from the Yale Institutional Review Board.

Table 1 HERE

2.2. Empirical strategy

2.2.1. Variable Selection using Supervised Machine Learning

The construction of a classification procedure from a set of data for which the correct classes are known is often referred to as supervised machine learning (Michie, Spiegelhalter, & Taylor, 1994). We implement a version of this approach using the existing Six Americas segments as the known classes. To validate our model we rely on cross-validation, a two-stage process including a training stage and a testing stage. In the training stage, a classifier algorithm is developed using the responses (data points) from individual participants and the correct categories associated with them to learn a specific pattern for how the data points map onto the categories. Once the classifier is trained, it then acts as a function to take in additional data points and produce the predicted classifications. We use cross-validation to perform the analysis on one subset of the data (the training set) and subsequently evaluate the performance of the model on the other subset (the testing set). Finally, to avoid biasing our ultimate model choice, we validated the chosen model on held-out samples. In contrast with standard selection methods (e.g., correlation, stepwise regression), a (supervised) machine learning approach can be evaluated by its effectiveness in making accurate predictions using new, independent samples.

Our model was trained and tested on the first 11 of the 14 surveys. The October 2015 survey, which included just the 15 items from the Reduced Discriminant Model, was withheld from the training of the model and used for validation purposes. The most recent surveys from March and November of 2016 were also withheld for validation to test the accuracy of the final model. The remaining 11 surveys were separated into two non-overlapping sets, 80% of the observations in a training set and 20% in a testing set (Pentreath, 2015), using the R package *caret*, such that the distribution of the Six Americas segments was approximately constant between the two sets.

The analysis was conducted in R and focused on the 15 questions included in the Reduced Discriminant Model. A generalized boosted regression modeling (GBM) algorithm (200 trees, 10-fold cross validation, multinomial distribution) was employed to identify the key variables for predicting segment membership (Kuhn, 2008). GBM is a broad method that uses classification and regression trees (boosting refers to an ensemble method where final predictions are the result of aggregated predictions from individual models). For a user-friendly introduction to machine learning, regression trees, and classification please see Strobl, Malley, and Tutz (2009). Whereas traditional ensemble techniques such as random forests rely on simple averaging of the models, gradient boosted machines consecutively fit new models to provide a more accurate estimate of the response variable (Natekin & Knoll, 2013). Specifically, we implemented the GBM package in R (Friedman, 2001; Ridgeway, 2007).

2.2.2. Six Americas Categorization using Multinomial Logistic Regression

In the second part of the analysis, multinomial-logistic regression models were iteratively fit to the data using Six Americas segmentation as the dependent variable and the top variables from the previous GBM procedure as independent variables. That is, the first model was fit using just the top variable; the second with the top two, etc., the next variable on the list was added to the model in order until satisfactory accuracy, roughly 70%, was achieved (Bekkar, Djemaa, & Alitouche, 2013; Fawcett, 2006; Kleinbaum & Klein, 2010). Multinomial-logistic regression identifies group membership for multi-class dependent variables (Venables & Ripley, 2013). The model achieved our pre-specified level of accuracy using the top four variables. The top four variables were used in a multinomial-logistic regression model that achieved the desired accuracy for both within- and out-of-sample test data. In short, we use a multinomial logistic regression to generate coefficients (odds-ratios) for each of the four questions. The (multi-class) dependent variable for the multinomial regression is the Six Americas (six categories). Based on the responses to the 4 items, the odds-ratios predict the likelihood of membership into each of the six categories (e.g. the odds of being in the Disengaged vs. the Alarmed). The highest odds-ratios across the four questions are used to classify respondents into one of the Six Americas segments.

3. Results

Results indicate that four variables are sufficient to identify the Six Americas. The four questions include respondents' worry about global warming, risk perceptions of the impact that global warming will have on them personally and on future generations, and personal importance of the issue. Our model achieves a minimum of 70% true positive rate, i.e., the proportion of accurately labeled respondents, in each of the Six Americas segments on a test set and very similar results on four out of sample validation sets. The model coefficients, listed as odds ratios with the Alarmed segment as the base, are shown in Table 2. The coefficients can be interpreted as follows: If a respondent answered "Not at all" versus "Don't know" to the question "How much do you think global warming will harm you personally?" the odds of being classified in the Dismissive segment (versus the Alarmed) increase by a factor of 22.56 to 1.

Table 2 HERE

To evaluate model performance we fit five separate testing/validation sets. The first set is the 20% holdout from surveys 1-11, called the "test set." The second, third, and fourth were all nationally representative surveys, two that segmented respondents based on the 15 item RDM screener from October 2015 and November 2016 (Appendix, Tables A1 and A3) and one using the full 36 item survey from March 2016 (Appendix, Table A2). A fourth set was a representative state-wide survey of Colorado conducted in 2013 (Leiserowitz, Feinberg, Howe, & Rosenthal, 2013) (Appendix, Table A4).

Several model performance measures were calculated for the test set data, including confusion matrices, precision, recall, F_1-scores, and average accuracy. The average accuracy, or the average per-category effectiveness of the classifier, for the four questions screener is 0.91 (Sokolova & Lapalme, 2009). The confusion (or "error") matrix describes the performance of a classification model on a set of test data by listing the true positive, false positive, true negative, and false negative values of the model fit. Here the values in the diagonal represent the true positives, the off-diagonal column entries are false negatives, and the off-diagonal row entries represent false positives (Table 3). The precision metric measures the class agreement with the labels, while recall measures the effectiveness of the classifier in identifying true positives, and the F_1-score measures the relationship between the classifier and the data's labels. The macro-level versions (i.e., measures of the model fit over all six segments) are each adjusted for multi-class classifiers using the sums of the per-class decisions (Sokolova & Lapalme, 2009). For reference, we list the expected value for each of the macro-level measures if the observations were classified at random in Table 4 (Rickert, 2016). From Table 5 it is clear that the model performs best in the Dismissive category, with precision, recall, and F_1-score all near 0.9, and least well in the Doubtful segment, with all three measures around 0.7.

The model showed a high level of accuracy[2] classifying respondents in a) the test set (Table 3), b) the 15 item screener survey from October 2015 (Appendix, Table A1), and c) on an out-of-sample surveys from March 2016 (Appendix, Table A2). The model has a true positive rate of at least 69% in each of the six segments for all three nationally representative out of sample validation data sets. The state-wide telephone survey from Colorado faired slightly worse, with a 62% true positive rate for the Disengaged segment but higher (68%-85%) for all other segments. The lower accuracy on this set may be due to the survey method, phone versus online or the limited geography of the sample (Colorado versus national). [3]

Table 3 HERE

Table 4 HERE

Table 5 HERE

We also evaluated the test-retest reliability of the model with an online study (n=241). The observations were taken 14-21 days apart, each using our short screener for segment identification. We evaluated performance using pearson's product-moment correlation coefficient. The correlation between the two segment fits is (r=0.66); the full confusion matrix is listed in the appendix.

[2]We also tested a six-item screener, incorporating introductory items: "Do you believe that global warming is happening" and "assuming global warming is happening, do you believe it is human-caused?" The accuracy of this model is similar and the relevant statistics can be found in the Appendix Tables B1 through B3.

[3]Phone surveys are less likely to elicit "Don't know" responses, which is an important identifier for the Disengaged segment.

4. Discussion

In this study we have developed and validated a new four-item screener to accurately segment the American public into one of Global Warming's Six Americas, originally identified in Maibach et al. (Maibach et al., 2011) with 36 items. Using these four questions, a true positive rate was achieved of around 70 percent across all six segments in the hold-out data, with the highest accuracy (> 0.85) found among the Dismissive. Moreover, although it is possible to increase overall classification accuracy by adding more relevant questions to the screener, the performance of the 4-item tool compares favorably to the much longer 15-item "Reduced Discriminant" screener, which reported an average classification accuracy of 83.8 percent for the overall sample, ranging from 0.60 to 0.97 across the six segments (Maibach et al., 2011). Across all samples, the average accuracy of the 4-item screener ranges from 0.77 (Colorado phone survey) to 0.94 (October 2015 RDM). The lowest true positive accuracy amongst online surveys occurs for the Doubtful segment, with the lowest hit rate of 0.69 in the March 2016 data (which is consistent with the 15-item tool (Maibach et al., 2011)). The lowest hit rate for the test-retest and telephone surveys occurs in the Disengaged category, having a true positive rate of 0.62 in the Colorado survey and 0.5 in the test-retest survey, though it is important to consider that the Disengaged segment (as fit by the first survey and 36 question screener respectively) represented a very small proportion of both samples, 1% and 2% respectively.

The current study also complements other recent research. For example, Swim and Geiger (2017) highlight positive correlations between the full 36-item screener and self-categorization into the Six Americas using a single-item (Swim & Geiger, 2017). Although promising, internal reliability cannot be estimated when researchers use single items. Nonetheless, it is worth noting that although the Spearman's correlations between our four-item and the 36-item screener are substantially higher (0.89-0.92 *vs.* 0.67-0.82), the test-retest reliability between Swim and Geiger's (2017) single-item measure and the current 4-item screener are similar (0.67 vs. 0.66).

The short segmentation tool provides a new, cost-effective survey instrument for understanding diverse public perceptions about climate change that is consistent with the literature. For example, risk perception, worry, and personal importance have long been identified as important predictors of climate change engagement and policy support (Ding, Maibach, Zhao, Roser-Renouf, & Leiserowitz, 2011; Malka, Krosnick, & Langer, 2009; Roser-Renouf et al., 2014; Smith & Leiserowitz, 2014; van der Linden, 2017). While social marketing tools like audience segmentation are sometimes characterized as forms of "data mining" that lack a theoretical basis, variations in beliefs, attitudes, and behaviors identified through the Six Americas segmentation have already been shown to reflect meaningful differences in risk perceptions and decision making across a variety of applications (Myers et al., 2012; Roser-Renouf et al., 2014). Indeed, two well-established theoretical dimensions underpin the Six Americas (Roser-Renouf et al., 2014), namely; attitudinal valence (the inclination to accept or reject climate science) and issue involvement (cognitive and affective issue engagement). The 4-item screener covers both attitudinal valence as well as issue involvement, which we define by the extent to which people think about and have firm beliefs (i.e. attitudinal certainty) about the issue of climate change (Roser-Renouf et al., 2014). For example, the SASSY screener accurately captures the proportion of Disengaged as identified by the full screener (Fig. 1). The disengaged are particularly characterized by a frequent "don't know" response, which in our view reflects low belief certainty and low issue involvement. It is of course possible to maintain a different definition of issue engagement. For example, some scholars have argued that many Americans might consider themselves disengaged for alternative reasons, such as prioritizing other issues, not knowing what to do about climate change, or lacking a general sense of self-efficacy (Swim & Geiger, 2017). In other words, our definition of the disengaged, in both the short and full screener, may underestimate the true proportion of disengaged in the population if these broader dimensions are considered, given that people's self-categorization into the Six Americas may differ from the screeners' (Swim & Geiger, 2017). However, it is worth noting that in the current screener, those individuals who do have opinions but aren't engaged, are more likely to end up in the Cautious category. We therefore encourage further work to assess whether these definitional issues are consequential in accurately representing how people respond to information about climate change. Further research on the Six Americas is likely to strengthen and add to the theoretical development of the literature on risk and science communication, opinion leadership, information-processing and social influence and persuasion.

Aside from the conceptual contribution, an interesting methodological question is how the machine learning method adopted in the current paper compares to more traditional approaches used in model selection, such as stepwise regression (where predictors are included or deleted one at a time in successive order). Forward or backwards stepwise regression is often used to select the "best" model with a set of q predictors. Although there are parallels between tree building and stepwise regression, scholars increasingly warn against the use of stepwise regression methods because they capitalize on sampling error, which leads to overfitting and poor out-of-sample prediction accuracy (Strobl, Malley, & Tutz, 2009; Thompson, 1995). Unlike stepwise regression, the gradient boosted machine learning approach employed in the current paper has the advantage of averaging over a large number of (decision tree) models.[4] Furthermore, in comparison, a stepwise regression approach selected a model that retained nearly all of the 15-item screener variables. This is not entirely surprising, as the inclusion of more variables leads to a relatively greater amount of model fit (e.g., as measured by R^2). Yet, the goal of the SASSY screener is to accurately classify respondents in *new* samples with the least number of questions rather than optimizing the amount of variance the variables can explain in *current* samples. In addition, stepwise regression methods also suffer from order effects whereas the advantage of ensemble methods (which employ parallel tree models) is that the order effects counterbalance, so that the overall importance ranking of the variables is much more reliable across samples (Strobl et al., 2009). In short, we note that there are a number of important advantages to a machine learning approach, including the use of training and (unseen) holdout datasets to evaluate predictive accuracy, a reduced risk of overfitting, greater model stability, and the fact that it requires fewer distributional assumptions.

To further support the current research, we have developed a web application that gives users the ability to find their Six Americas segment by taking this short screener. It also allows researchers to input their own survey data (with proper formatting) and receive a dataset with the respondents' segments appended as output. (The application can be accessed here `http://climatecommunication.yale.edu/visualizations-data/sassy/`.) We look forward to future research and practice exploring the value of Global Warming's Six Americas in predicting and explaining how the public responds to and engages with the issue of climate change.

[4] An all-possible-subset regression using Dominance analysis (Budescu, 1993) yielded a similar 4-item ranking, but this approach suffers from some of the same shortcomings (Matsuki, Kuperman, & Van Dyke, 2016).

References

Akerlof, K., Bruff, G., & Witte, J. (2011). Audience segmentation as a tool for communicating climate change. *Park Sci*, *28*(1), 56–64.

Arbuckle, J., Hobbs, J., Loy, A., Morton, L. W., Prokopy, L., & Tyndall, J. (2014). Understanding corn belt farmer perspectives on climate change to inform engagement strategies for adaptation and mitigation. *Journal of Soil and Water conservation*, *69*(6), 505–516.

Ashworth, P., Jeanneret, T., Gardner, J., & Shaw, H. (2011). Communication and climate change: What the australian public thinks. *Report No. EP112769. Canberra: CSIRO Publishing.*

BBC Media Action. (2013). *Climate asia methodology series. regional segment profiles.* BBC Media Action, London, UK.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced datasets. *J Inf Eng Appl*, *3*(10).

Bowers, A. W., Monroe, M. C., & Adams, D. C. (2016). Climate change communication insights from cooperative extension professionals in the us southern states: finding common ground. *Environmental Communication*, *10*(5), 656–670.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, *114*(3), 542.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R., Verheggen, B., Maibach, E. W., ... others (2016). Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), 048002.

Costello, C. (2014). Why are we still debating climate change. *CNN. com. Online at http://www. cnn. com/2014/02/24/opinion/costello-debate-climate-change/, accessed February*, *24*, 2014.

Cranmer, S., Gill, J., Jackson, N., Murr, A., & Armstrong, D. (2016). hot.deck: Multiple hot-deck imputation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=hot.deck` (R package version 1.1)

Cranmer, S. J., & Gill, J. (2013). We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, *43*(02), 425–449.

Detenber, B., Rosenthal, S., Liao, Y., & Ho, S. S. (2016). Climate and sustainability— audience segmentation for campaign design: Addressing climate change in singapore. *International Journal of Communication*, *10*, 23.

Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*, *1*(9), 462–466.

Doherty, K. L., & Webler, T. N. (2016). Social norms and efficacy beliefs drive the alarmed segment [rsquor] s public-sphere climate actions. *Nature Climate Change*, *6*(9), 879–884.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fisher, D. (2014). *Global warming theories + political reality = tesla subsidies.* Forbes. com. Online at `https://www.forbes.com/sites/danielfisher/2014/04/05/global-warming-theories-political-reality-tesla-subsidies/#77f365976be3`, accessed January.

Flora, J. A., Saphir, M., Lappé, M., Roser-Renouf, C., Maibach, E. W., & Leiserowitz, A. A. (2014). Evaluation of a national high school entertainment education program: The alliance for climate education. *Climatic change*, *127*(3-4), 419–434.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Harris, P., Lock, A., Phillips, J. M., Reynolds, T. J., & Reynolds, K. (2010). Decision-based voter segmentation: an application for campaign message development. *European Journal of Marketing*, *44*(3/4), 310–330.

Hine, D. W., Reser, J. P., Morrison, M., Phillips, W. J., Nunn, P., & Cooksey, R. (2014). Audience segmentation and climate change communication: conceptual and methodological considerations. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(4), 441–459.

Kelly, L.-A. D., Luebke, J. F., Clayton, S., Saunders, C. D., Matiasek, J., & Grajal, A. (2014). Climate change attitudes of zoo and aquarium visitors: implications for climate literacy education. *Journal of Geoscience Education*, *62*(3), 502–510.

Kleinbaum, D. G., & Klein, M. (2010). Introduction to logistic regression. In *Logistic regression* (pp. 1–39). Springer.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Leiserowitz, A., Feinberg, G., Howe, P., & Rosenthal, S. (2013). *Climate change in the coloradan mind: July 2013.*

Leiserowitz, A., Maibach, E., Roser-Renouf, C., Feinberg, G., & Rosenthal, S. (2016). Climate change in the american mind: March, 2016. *Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication*.

Leiserowitz, A., Thaker, J., Feinberg, G., & Cooper, D. (2013). Global warmings six indias. *Yale University, New Haven, CT*.

Leiserowitz, A. A. (2005). American risk perceptions: Is climate change dangerous? *Risk analysis*, *25*(6), 1433–1442.

Magidson, J., & Vermunt, J. K. (2002). Latent class modeling as a probabilistic extension of k-means clustering. *Quirks Marketing Research Review*, *20*, 77–80.

Maibach, E. W., Leiserowitz, A., Roser-Renouf, C., & Mertz, C. (2011). Identifying like-minded audiences for global warming public engagement campaigns: An audience segmentation analysis and tool development. *PloS one*, *6*(3), e17571.

Maibach, E. W., Weber, D., Massett, H., Hancock, G. R., & Price, S. (2006). Understanding consumers' health information preferences development and validation of a brief screening instrument. *Journal of health communication*, *11*(8), 717–736.

Malka, A., Krosnick, J. A., & Langer, G. (2009). The association of knowledge with concern about global warming: Trusted information sources shape public thinking. *Risk Analysis*, *29*(5), 633–647.

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33.

Metag, J., Füchslin, T., & Schäfer, M. S. (2015). Global warmings five germanys: A typology of germans views on climate change and patterns of media use and information. *Public understanding of science*, 0963662515592558.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification.

Morrison, M., Duncan, R., Sherley, C., & Parton, K. (2013). A comparison between attitudes to climate change in australia and the united states. *Australasian Journal of Environmental Management*, *20*(2), 87–100.

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, *5*(4), 297–310.

Myers, T. A., Nisbet, M. C., Maibach, E. W., & Leiserowitz, A. A. (2012). A public health frame arouses hopeful emotions about climate change. *Climatic Change*, *113*(3-4), 1105–1112.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? meta-analytic review of tailored print health behavior change interventions. *Psychological bulletin*, *133*(4), 673.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., ... others (2014). *Climate change 2014: synthesis report. contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change.* IPCC.

Pentreath, N. (2015). *Machine learning with spark.* Packt Publishing Ltd.

Rickert, J. (2016). *Computing classification evaluation metrics in r* (Blog No. March 11). `http://blog.revolutionanalytics.com/2016/03/com_class_eval_metrics_r.html`.

Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package. *Update*, *1*(1), 2007.

Roser-Renouf, C., Stenhouse, N., Rolfe-Redding, J., Maibach, E. W., & Leiserowitz, A. (2014). Engaging diverse audiences with climate change: message strategies for global warming's six americas.

Schweizer, S., Davis, S., & Thompson, J. L. (2013). Changing the conversation about climate change: A theoretical framework for place-based climate change engagement. *Environmental Communication: A Journal of Nature and Culture*, *7*(1), 42–62.

Slater, M. D. (1996). Theory and method in health audience segmentation. *Journal of health communication*, *1*(3), 267–284.

Smith, N., & Leiserowitz, A. (2014). The role of emotion in global warming policy support and opposition. *Risk Analysis*, *34*(5), 937–948.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Swim, J. K., & Geiger, N. (2017). From alarmed to dismissive of climate change: A single item assessment of individual differences in concern and issue involvement. *Environmental Communication*, *11*(4), 568–586.

Thompson, B. (1995). *Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial.* Sage Publications Sage CA: Thousand Oaks, CA.

van der Linden, S. (2017). Determinants and measurement of climate change risk perception, worry, and concern. *Oxford Encyclopedia of Climate Change Communication*.

van der Linden, S., Maibach, E., & Leiserowitz, A. (2015). Improving public engagement with climate change: Five best practice insights from psychological science. *Perspectives on Psychological Science*, *10*(6), 758–763.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with s-plus.* Springer Science & Business Media.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*, 89–106.

Survey Data

ID	Time (and mode)	Domain	Type	Sample	Purpose
1	October 2008 (online)	US	Full	2139	Test/Train
2	January 2010 (online)	US	Full	1001	Test/Train
3	June 2010 (online)	US	Full	1024	Test/Train
4	May 2011 (online)	US	Full	486	Test/Train
5	November 2011 (online)	US	Full	981	Test/Train
6	April 2012 (online)	US	Full	996	Test/Train
7	September 2012 (online)	US	Full	1058	Test/Train
8	April 2013 (online)	US	Full	1035	Test/Train
9	December 2013 (online)	US	Full	823	Test/Train
10	October 2014 (online)	US	Full	1272	Test/Train
11	March 2015 (online)	US	Full	1263	Test/Train
12	October 2015 (online)	US	RDM	1329	Validation
13	March 2016 (online)	US	Full	1203	Validation
14	November 2016 (online)	US	RDM	1226	Validation
15	June 2013 (phone)	Colorado	Full	780	Validation
16	February 2017 (online)	US	SASSY	241	Test-Retest

Table 1. Dataset time and mode, domain, sample sizes, and purpose. RDM refers to the 15-item Reduced Discriminant Model from Maibach, et al. 2011. SASSY refers to the 4-item Six Americas Short Survey described in this work.

Multinomial-logistic Model

		Six Americas Segment:				
		Conc.	Caut.	Diseng.	Doubt.	Dismis.
How much do you think global warming will harm future generations of people?	Not at all	1.20 (1.20)	34.71*** (1.23)	0.02*** (6.82)	17.61*** (1.25)	747.34*** (1.27)
	Only a little	0.29*** (.66)	13.27*** (.65)	0.04*** (.93)	8.21*** (.66)	5.55*** (.71)
	A mod. amt.	1.38 (.31)	13.51*** (.37)	0.00*** (1.02)	0.71 (.39)	0.27** (.54)
	A great deal	0.38*** (0.24)	0.19*** (0.32)	0.00*** (0.59)	0.00*** (0.37)	0.01*** (0.53)
How important is the issue of global warming to you personally?	Not too imp.	19.06*** (1.07)	8.54*** (1.06)	8.98*** (1.08)	3.73** (1.06)	2.14 (1.07)
	Somewhat	3.13*** (.80)	0.28*** (.79)	0.49 (.83)	0.04*** (.80)	0.02*** (.84)
	Very	0.37*** (.80)	0.01*** (.80)	0.02*** (.86)	0.00*** (.85)	0.00*** (.96)
	Extremely	0.08*** (.81)	0.00*** (.87)	0.00*** (.99)	0.00*** (.99)	0.00*** (1.14)
How worried are you about global warming?	Not very	0.87 (.76)	0.40*** (.77)	0.42* (.80)	0.09*** (.78)	0.02*** (.79)
	Somewhat	0.41*** (.71)	0.03*** (.72)	0.03*** (.77)	0.00*** (.74)	0.00*** (.93)
	Very	0.12*** (.71)	0.01*** (.75)	0.00*** (.85)	0.00*** (1.29)	0.00*** (4.95)
How much do you think global warming will harm you personally?	Not at all	1.12 (.32)	7.23*** (.39)	0.10*** (.51)	5.79*** (.41)	22.56*** (.53)
	Only a little	0.61** (.23)	2.14** (.30)	0.05*** (.40)	0.64 (.34)	1.58 (.54)
	A mod. amt.	0.58* (.22)	1.29 (.29)	0.05*** (.40)	0.58 (.36)	2.09 (.67)
	A great deal	0.28*** (.23)	1.31 (.33)	0.06*** (.58)	0.86 (.53)	1.57 (.99)
Akaike Inf. Crit.	11,869.920					
Note:				*$p<0.05$; **$p<0.01$; ***$p<0.001$		

Table 2. Coefficients are listed as odds-ratios. Conc. is Concerned, Caut. is Cautious, Diseng. is Disengaged, Doubt. is Doubtful, and Dismis. is Dismissive. Reference Category: Alarmed. Reference response by question: Don't know; Not at all important; Not at all worried; Don't know. Standard errors are provided in parenthesis.

Confusion Matrix Test-set

	Observed					
Model	Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.73**	0.09	0.01	0.00	0.01	0.00
Concerned	0.26	**0.73**	0.16	0.05	0.01	0.00
Cautious	0.01	0.16	**0.71**	0.00	0.14	0.02
Disengaged	0.00	0.03	0.00	**0.82**	0.06	0.00
Doubtful	0.00	0.00	0.11	0.013	**0.70**	0.11
Dismissive	0.00	0.00	0.01	0.00	0.08	**0.87**

Table 3. Confusion matrix of the 36 item segment fit and the 4 item multinomial-logistic fit for the testing set (20% hold out from surveys 1-11).

Macro-Performance Measures

	Macro-Precision	Macro-Recall	Macro-F_1-score
Model	0.76	0.76	0.76
Expected if Random	0.17	0.17	0.17

Table 4. Macro-performance measures for test set (20% holdout) data for multinomial logistic regression model.

Performance Measures by Segment validation data set

	Precision	**Recall**	**F_1-score**
Alarmed	0.73	0.79	0.76
Concerned	0.73	0.72	0.73
Cautious	0.71	0.70	0.71
Disengaged	0.82	0.76	0.79
Doubtful	0.70	0.67	0.68
Dismissive	0.87	0.89	0.88

Table 5. Performance measures for test set (20% holdout data) for each segment.

Appendix A. Tables

Confusion Matrix October 2015 (RDM)

	Observed					
Model	Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.81**	0.10	0.00	0.00	0.00	0.00
Concerned	0.19	**0.76**	0.09	0.04	0.01	0.00
Cautious	0.00	0.13	**0.82**	0.02	0.14	0.00
Disengaged	0.00	0.01	0.00	**0.76**	0.04	0.00
Doubtful	0.00	0.00	0.08	0.18	**0.78**	0.00
Dismissive	0.00	0.00	0.00	0.00	0.04	**1.00**

Table A1. Confusion matrix of the 15-item Reduced Discriminant Model (RDM) segment fit and the 4 item multinomial-logistic SASSY fit for the October 2015 survey.

Confusion Matrix March 2016

Model	Observed Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.70**	0.11	0.01	0.02	0.00	0.00
Concerned	0.30	**0.76**	0.14	0.08	0.01	0.00
Cautious	0.00	0.12	**0.74**	0.02	0.20	0.01
Disengaged	0.00	0.00	0.00	**0.76**	0.04	0.00
Doubtful	0.00	0.00	0.11	0.13	**0.69**	0.08
Dismissive	0.00	0.00	0.00	0.00	0.07	**0.91**

Table A2. Confusion matrix of the 36 item segment fit and the 4 item SASSY fit for the March 2016 survey.

Confusion Matrix November 2016

Model	Observed Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.80**	0.07	0.01	0.02	0.00	0.00
Concerned	0.20	**0.78**	0.10	0.05	0.00	0.00
Cautious	0.00	0.13	**0.78**	0.02	0.08	0.01
Disengaged	0.00	0.01	0.00	**0.73**	0.07	0.00
Doubtful	0.00	0.00	0.12	0.20	**0.80**	0.00
Dismissive	0.00	0.00	0.00	0.00	0.05	**0.99**

Table A3. Confusion matrix of the 15 item RDM segment fit and the 4 item SASSY fit for the November 2016 survey.

Confusion Matrix Colorado

Model	Observed Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.73**	0.12	0.00	0.00	0.00	0.00
Concerned	0.26	**0.71**	0.20	0.12	0.00	0.01
Cautious	0.01	0.14	**0.68**	0.12	0.25	0.02
Disengaged	0.00	0.01	0.01	**0.62**	0.03	0.00
Doubtful	0.00	0.00	0.11	0.12	**0.69**	0.12
Dismissive	0.00	0.01	0.01	0.00	0.03	**0.85**

Table A4. Confusion matrix of the 36 item segment fit and the 4 item SASSY fit for a survey completed in Colorado only.

Confusion Matrix Test-Retest

Second Fit	First Fit Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.76**	0.06	0.12	0.00	0.00	0.00
Concerned	0.04	**0.59**	0.15	0.00	0.25	0.25
Cautious	0.19	0.16	**0.69**	0.00	0.00	0.00
Disengaged	0.01	0.04	0.00	**0.50**	0.00	0.00
Doubtful	0.00	0.04	0.02	0.00	**0.67**	0.30
Dismissive	0.00	0.10	0.02	0.50	0.08	**0.45**

Table A5. Confusion matrix of the first 4 item segment SASSY fit and the second 4 item segment SASSY fit for the test-retest data.

Appendix B. Six Question Model

Confusion Matrix Test Set; Six Question Model

	Observed					
Model	Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.72**	0.08	0.00	0.01	0.00	0.00
Concerned	0.27	**0.80**	0.14	0.05	0.00	0.00
Cautious	0.00	0.11	**0.77**	0.00	0.10	0.01
Disengaged	0.00	0.02	0.01	**0.83**	0.05	0.00
Doubtful	0.00	0.00	0.08	0.11	**0.76**	0.11
Dismissive	0.00	0.00	0.00	0.00	0.09	**0.89**

Table B1. Confusion matrix of the 36 item segment fit and the 4 item SASSY fit for the 20% hold out data set.

Confusion Matrix March 2016 Survey; Six Question Model

	Observed					
Model	Alarmed	Concerned	Cautious	Disengaged	Doubtful	Dismissive
Alarmed	**0.69**	0.10	0.01	0.00	0.00	0.00
Concerned	0.30	**0.81**	0.14	0.08	0.00	0.00
Cautious	0.00	0.09	**0.78**	0.02	0.14	0.01
Disengaged	0.00	0.00	0.00	**0.81**	0.03	0.00
Doubtful	0.00	0.00	0.06	0.09	**0.75**	0.07
Dismissive	0.00	0.00	0.00	0.00	0.08	**0.92**

Table B2. Confusion matrix of the 36 item segment fit and the 4 item SASSY fit for the March 2016 survey.

Coefficients Multinomial-logistic 6 Question Model

		Conc.	Caut.	Diseng.	Doubt.	Dismis.
(Intercept)		152***	4410***	469864***	815591***	18108***
How much do you think global warming will harm future generations of people?	A great deal	0.42***	0.20***	0.00***	0.01***	0.01***
	A mod. amt.	1.62	17.48***	0.00***	0.71	0.36
	Not at all	3.37	98.35**	0.11	62.47**	2926.8***
	Only a little	0.39	16.59***	0.05**	9.41**	6.36*
How important is the issue of global warming to you personally?	Extremely	0.06***	0.00***	0.00***	0.00***	0.00***
	Not too imp.	23.20**	15.27*	13.46*	6.62	4.46
	Somewhat	2.70	0.30	0.39	0.04***	0.02***
	Very	0.30	0.01***	0.01***	0.00***	0.00***
How worried are you about global warming?	Not very	1.13	0.61	0.58	0.12*	0.04***
	Somewhat	0.57	0.06***	0.04***	0.00***	0.00***
	Very	0.17*	0.01***	0.01***	0.00***	0.00***
How much do you think global warming will harm you personally?	A great deal	0.24***	0.88	0.04***	0.66	1.60
	A mod. amt	0.60***	1.30	0.05***	0.49	1.44
	Not at all	1.29	12.54***	0.14***	12.81***	54.05***
	Only a little	0.66	3.30***	0.06***	1.01	2.46
Assuming global warming is happening, do you think it is caused by...	human activities and natural changes	1.50	0.96	0.34	0.13	0.01**
	mostly human activities	1.68*	1.22***	0.92**	0.14**	0.08**
	mostly natural changes in the environment	8.61*	39.62***	26.97**	47.74**	69.84**
	global warming isn't happening	0.87	2.61	3.09	4.00	16.11
Do you think global warming is happening?	no	61.63	103.84	82.20	701.02	2947.35
	yes	0.12***	0.01***	0.01***	0.01***	0.00***

Table B3. Coefficients are listed as odds-ratios. Reference Category: Alarmed. Reference response by question: Don't know; Not at all important; Not at all worried; Don't know. * Significant at the 0.05 level. ** Significant at the 0.01 level. *** Significant at the 0.001 level.