An Experimental Study of Recommendation Algorithms for Tailored Health Communication

Hyun Suk Kim, Sijia Yang, Minji Kim, Brett Hemenway, Lyle Ungar, Joseph N. Cappella

CCR 1 (1): 103–129
DOI: 10.5117/CCR2019.1.005.SUKK

Abstract

Recommendation algorithms are widely used in online cultural markets to provide personalized suggestions for products like books and movies. At the heart of the commercial success of recommendation algorithms is their ability to make an accurate prediction of a target person's preferences for previously unseen items. Can these algorithms also be used to predict which health messages an individual will evaluate favorably, and thereby provide effective tailored communication to the person? Although there is evidence that message tailoring enhances persuasion, little research has examined the effectiveness of recommendation algorithms for tailored health interventions aimed at promoting behavior change. We developed a message tailoring algorithm to select smoking-related public service announcements (PSAs) for smokers, and experimentally test its effectiveness in predicting a target smoker's evaluations of PSAs and encouraging smoking cessation. The tailoring algorithm was constructed using multiple levels of data on smokers' PSA rating history, individual differences, content features of the PSAs, and other smokers' PSA ratings. We conducted a longitudinal online experiment to examine its efficacy in comparison to two non-tailored methods: "best in show" (choosing messages most preferred by other smokers) and "off the shelf" (random selection from eligible ads). The results showed that the tailoring algorithm produced more accurate predictions of smokers' message evaluations than the simple-average method used for the "best in show" approach. Smokers who viewed PSAs recommended by the tailoring algorithm were more likely than those receiving a random set to evaluate the PSAs favorably and quit smoking. There was no significant difference between the

"best in show" and "off the shelf" methods in message assessment and quitting behavior.

Keywords: Recommendation algorithms, message tailoring, health communication

The design of communication campaigns is a complicated and multidimensional undertaking whether for public health goals (Slater, 2006) or for public relations or strategic communication goals (Buhmann, Likely, & Geddes, 2018; Macnamara, 2018). One typical component of communication campaigns is message selection (Kim & Cappella, in press). The process of selecting messages to employ in a media campaign is crucially important to the campaign's success. Three broad approaches to message selection can be identified: "off the shelf" (OTS hereafter); pretesting in targeted groups; and tailoring to individuals in targeted groups.

1. Literature Review

1.1 Three Approaches to Campaign Messages

OTS approaches are directed less by scientific principles and well-defined data and more by intuitions of experienced campaign designers, data from interviews or focus groups (Carey, 1994), or convenience based on the availability and pertinence of messages. The approach is certainly not random but is an alternative given limited resources and/or the need to be immediately responsive to a targeted problem.

Pre-testing of messages takes a variety of forms including designing campaign messages from scratch starting with psychographic profiles and belief structures in the targeted audience for the behavior under scrutiny (Parvanta et al., 2013; Zhao et al., 2016) to the selection of a subset of messages from those pretested for their effectiveness by members of the targeted population (Kelder, Pechmann, Slater, Worden, & Levitt, 2002; Nonnemaker, Farrelly, Kamyab, Busey, & Mann, 2010). The latter of these two approaches assumes that a set of relevant messages already exists and testing allows the section of the best performing messages. We label these selections "best in show" (BIS hereafter) messages because they exhibit the highest scores among competing messages on various criteria of effectiveness.

Tailoring involves selecting message which are geared to individuals in the targeted group based on their individual characteristics (Petty, Barden, & Wheeler, 2009) and is one of the most studied communication strategies

(Kreuter, Farrell, Olevitch, & Brennan, 2000; Noar & Harrington, 2016). One form of tailoring typically begins with a survey about demographic and other individual characteristics that are relevant to the targeted behaviors. These individual-level data (e.g., a person's age), in combination with relevant message-level features (e.g., age of a person in a campaign ad), are used as tailoring variables to craft various versions of intervention messages. Messages deemed most effective for a target person are selected from a "message library" of options and directed to the person (Noar & Harrington, 2016; Rimer & Kreuter, 2006). There is meta-analytic evidence that message tailoring enhances persuasion (Lustria et al., 2013; Krebs, Prochaska, & Rossi, 2010). However, there are an infinite number of individual- and message-level features on which tailoring can proceed. That is, typical tailoring approaches, while effective, are inefficient as well as costly to create and test using conventional approaches (Cappella, Yang, & Lee, 2015).

Tailoring via machine learning-based recommendation algorithms is a viable alternative way to tailor (Cappella et al., 2015; Sadasivam et al., 2016). Computer algorithms are widely used by commercial vendors such as Amazon and Netflix to make personalized suggestions for which books to read and which movies to watch (Resnick & Varian, 1997; Jannach, Zanker, Felfernig, & Friedrich, 2011). These recommender systems have the same goal as tailored health communication interventions, namely to predict a target individual's ratings of items that have not been seen by the individual and provide suggestions for the individual accordingly. Developments in recommendation algorithms such as collaborative filtering have significantly increased the accuracy of the prediction of, for example, an individual's movie preferences (Koren, Bell, & Volinsky, 2009; Amatriain & Basilico, 2015). Unlike the conventional message-tailoring approach, recommender systems (a) automate the tailoring process using computer algorithms, (b) do not require extensive pretests to determine which psychographic features are predictive in the specific context (Cappella et al., 2015), and (c) tailor on actual choices not on psychographic predictors of choices. Despite the successful implementation of recommender systems in commercial arenas and their potential application to message tailoring, little research, especially experimental research, has tested the predictive performance of recommendation algorithms for health messages like public service announcements (PSAs) that are presumably less heterogeneous in content and format than books and movies. Also, little is known empirically about whether and how algorithm-selected, tailored health messages shape attitudes and behaviors.

1.2 Comparing Three Approaches to Message Selection

In our study, we compare three approaches to message selection in the context of anti-smoking messages directed at adult smokers. The three approaches compare messages selected via the OTS, BIS, and tailored recommendation algorithm techniques. The goal is to determine whether a recommendation algorithm for anti-smoking PSAs can perform as well or better than two simple alternatives. Those alternatives are ones often used in message selection for health and other kinds of campaigns. To our knowledge, no research has tested whether a recommendation algorithm for complex health messages is effective in behavior change at least in contrast to simple but common alternatives. Although not the focus of the current research, tailored recommendation algorithms also have the ability to improve predictive accuracy over time as more evaluation data is covertly gathered from users (Koren & Bell, 2015), whereas the other two conventional approaches lack such dynamic "learning" ability. The test that we propose and carry out involves the development of a computational procedure for message selection and a behavioral test of its efficacy in comparison to the two other approaches. The comparison that we will offer is not just a prediction of what is selected but a step beyond involving prediction of behavioral change in response to messaging driven by these three different selection procedures.

We build on methodological advances in recommender systems (Aggarwal, 2016; Adomavicius & Tuzhilin, 2005; Koren & Bell, 2015; Hastie, Mazumder, Lee, & Zadeh, 2015) to develop a message tailoring algorithm for anti-smoking PSAs. The specific form of the algorithm we developed is described in detail in the methods section below. In brief, the algorithm incorporates collaborative filtering (item and user based as well as matrix factorization), certain message features, and individual characteristics into a multilevel framework to achieve predictive success (Hastie et al., 2015: Raudenbush & Bryk, 2002). These components represent kinds of social influence (choices by similar others) and content preferences (content choices of messages similar to those previously preferred), while the multilevel model with random intercepts and slopes allows key preference parameters to vary by individual. Most importantly, the tailoring is not psycho-social, it is more behavioral and content oriented. The algorithm offers predictively useful selections based on what others of like mind would choose and messages that are congruent with prior selections.

The algorithm's performance is compared behaviorally to BIS (choosing messages most preferred by other smokers) and OTS (choosing anti-smoking messages using random sampling from a set of established anti-smoking

PSAs). We hypothesize that our algorithmic selections for each person in the sample will outperform messages selected using the other two methods in predicting subsequent quit attempts for adult smokers.

2.0 Materials and Methods

2.1 Overview

Wave 1 was a baseline survey (N = 1,057) to collect data to be used as a basis for designing message interventions at Wave 2. During the two-week gap between Waves 1 and 2, Wave 1 data were used to produce a tailoring algorithm-predicted persuasiveness rating for all 72 unseen PSAs respectively for each participant. Wave 2 was a randomized experiment (N = 675). Participants were randomly assigned to one of three experimental conditions: (a) the *tailoring* condition; (b) the *OTS* condition where four randomly selected PSAs were shown to participants; and (c) the *BIS* condition where the top-four PSAs based on overall means of persuasiveness evaluations by other participants at Wave 1 were selected. Wave 3 (N = 525), a follow-up survey, was conducted after about another two weeks to assess behavioral consequences of the message interventions at Wave 2. The retention rate was 63.9% for Wave 2 and 49.7% for Wave 3, both compared with the Wave 1 sample. The flow diagram of the longitudinal experiment is presented in Figure 1.

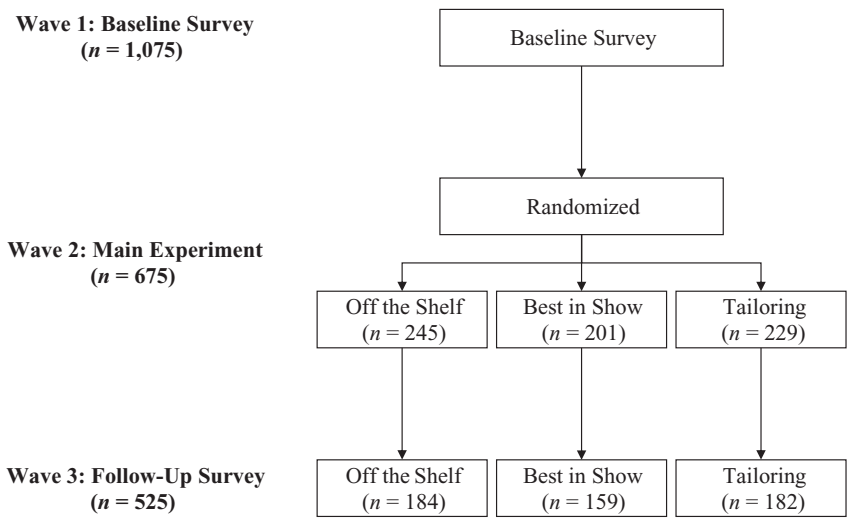

Figure 1. Flow diagram of the study.

Eighty anti-smoking PSAs (all 30-seconds) were used as message stimuli. The PSAs varied in perceived effectiveness (PE), an efficient measure of a message's actual persuasive effectiveness especially useful when a large number of messages need to be ranked (Dillard, Shen, & Vail, 2007; Cappella, 2018), representing eight levels of PE (ten PSAs per level). The PE data were obtained from independent samples of smokers not a part of this three-wave study.

2.2 Wave 1: Baseline Survey

2.2.1 Sample and procedure.

Participants were 1,057 smokers who smoked at least 100 cigarettes in their lifetime and currently smoked daily or on at least some days (age ranged from 18 to 65, M = 42.1, SD = 12.2). Participants were recruited from an online research panel hosted by Survey Sampling International. Upon consenting, participants answered a *pre-exposure survey* about smoking-related individual characteristics (e.g., stage of change toward quitting). They were then taken to a *main survey* phase where they watched and evaluated eight PSAs. For each PSA, they reported PE ratings of the PSA and emotional responses to the PSA. After viewing and rating eight PSAs, they completed a *post-exposure survey* about quitting intentions and demographic characteristics.

The PSA sample used in this study came from a larger research project on 200 anti-smoking PSAs. The 200 PSAs were obtained from professional health agencies such as CDC and American Legacy. Many of them were aired on TV as part of state or national campaigns between 1998 and 2007. The research project created an archive of 200 PSAs that were (a) coded on various content and format features (e.g., message sensation value) by independent coders and (b) rated by nationally representative samples of 2,354 U.S. smokers across four studies (e.g., PE).

Based on these data, we classified the 200 PSAs into eight levels of PE using an octile-split. We then obtained a representative sample of 80 PSAs for the current study by stratified random sampling (ten PSAs per each of the eight levels of PE). For the *main survey*, we sampled eight out of the 80 PSAs for each participant using stratified random sampling (one PSA per each of the eight PE levels) to ensure that the eight PSAs assigned to each participant were representative of the varying levels of PE among the population of 200 PSAs. The PSA sampling and assignment procedure resulted in the average number of participants per PSA being 105.7 (SD = 11.6, Min = 71, Max = 135).

As with the PSA sample, the Wave 1 survey items were also adopted from the larger research project mentioned above. We analyzed the 200-PSA evaluation data obtained from nationally representative samples of 2,354 U.S. smokers, and identified individual characteristics of smokers that were reliably associated with the PE ratings of the PSAs. We included survey questions about those individual characteristics in the current study.

2.2.2. Survey measures.

2.2.2.1. Pre-exposure measures. Participants reported how many times they had previously quit smoking on purpose for more than one complete day. The distribution of the number of *previous quitting attempts* was positively skewed, and thus it was log-transformed (M = 1.14, SD = .78). To measure *nicotine dependence*, participants were asked to complete the Fagerström test (Heatherton, Kozlowski, Frecker & Fagerström, 1991; M = 4.46, SD = 2.41). *Perceived effects of smoking on health* was measured using a single item with response options ranging from "not at all" (= 1) to "very much" (= 5): "to what extent do you feel your overall health has been affected by smoking?" (M = 2.99, SD = 1.06). Using the same response options, participants reported *perceived health benefits of quitting* in response to a single item: "how much do you think that quitting smoking could help your health?" (M = 3.84, SD = 1.12). *Need for cognition* (Cacioppo, Petty, & Kao, 1984) was measured using four items from the original scale (α = .54, M = 3.54, SD = .77). Participants reported their *stage of change toward quitting* (M = 6.33, SD = 2.66) using the Contemplation Ladder (Biener & Abrams, 1991) which ranged from "I have no thoughts about quitting smoking" (= 0) to "I am taking action to quit smoking" (= 10).

2.2.2.2. Main survey measures. After exposure to each PSA, participants evaluated its PE and reported emotional responses to it. PE was measured using four five-point items (Bigsby, Cappella, & Seitz, 2013). The items showed high internal consistency (α = .72). PE score was created by averaging the three items (M = 3.39, SD = .89). Participants indicated their emotional responses to each PSA. They reported how strongly they agreed with the following statements about their negative emotional responses (afraid, guilty, disgusted) (α = .87, M = 3.04, SD = 1.16) and positive emotional responses (hopeful, proud) (α = .72, M = 2.62, SD = 1.08).

2.2.2.3. Post-exposure measures. Intention to quit smoking was measured using the five items used successfully in prior research (Bigsby et al., 2013). Items were standardized and then averaged (α = .90, M = 0, SD = .84). This variable was not used in developing a tailoring algorithm-based prediction model of PE. Instead, we included it as a covariate in the analysis

of the Wave 2 and Wave 3 data. Lastly, participants answered demographic questions. About 61.4% of them were female. 79.9% were non-Hispanic White. 65.6% were currently employed. 62.0% were married or living with partner. 85.9% were the head of household. 51.8% had children under the age of 18 in their household. The distribution of the highest level of education was as follows: less than high school (2.9%), high school (19.2%), some college (36.0%), college or higher (41.8%). The household income was distributed as follows: less than $10,000 (5.5%), $10,000 to $14,999 (5.1%), $15,000 to $19,999 (4.6%), $20,000 to $34,999 (15.8%), $35,000 to $49,999 (16.4%), $50,000 to $74,999 (23.1%), $75,000 to $99,999 (16.5%), $100,000 to $199,999 (11.4%), $200,000 or more (1.6%).

2.2.3. Message characteristics of PSAs.

We incorporated message features of the 80 PSAs as part of the tailoring algorithm described in the next section. Message features were included because we wanted to be sure that any effects from the recommendation algorithm were over and above features common across the PSAs (e.g. strong arguments). The message features were either (a) coded by trained research assistants (*presence of narrative*, *presence of smoking cue*, *presence of efficacy information*, *message sensation value*, *video- and audio-level information introduced*), or (b) obtained by aggregating evaluations provided by independent samples of smokers (*argument strength*) or the current sample (*positive and negative emotions*). Data on all message features except positive and negative emotions were collected through the samples provided by the larger research project mentioned earlier and, therefore, completely independent of the ratings provided by the current sample. The features coded and descriptive statistics included the following.

2.2.3.1. Presence of narrative, smoking cue, and efficacy information. Out of the 80 PSAs, 26 PSAs (32.5%) contained a narrative, defined as the presence of a person's story typically with a "point" or moral. A PSA was considered to have a smoking cue if it presented one of the following visual scenes (Kang, Cappella, Strasser, & Lerman, 2009): (a) objects associated with smoking (e.g. ashtrays), (b) holding or handling cigarettes, and (c) actual smoking behaviors. Forty-three PSAs (53.8%) contained smoking cues. Efficacy information (Bandura, 2004) was defined as verbal or visual information directing audience to a quit line or nicotine replacement therapies (e.g. nicotine patches), and was presented in 28 PSAs (35.0%).

2.2.3.2. Message sensation value (MSV). Eighteen MSV features were coded and grouped into three categories (Morgan, Palmgreen, Stephenson, Hoyle, & Lorch, 2003): (a) MSV-video ($M = 1.05$, $SD = .49$; animation, number

of cuts, number of edits, number of faces, special visual effects, slow motion, fast motion, unusual color, intense moments), (b) MSV-audio (M = 1.13, SD = .66; sound saturation, music, sound effects, slow voice, and fast voice), and (c) MSV-content (M = 1.20, SD = 1.06; act out vs. talking head, unexpected format, surprising/twisted end). Number of cuts, number of edits, and number of faces were normalized to vary between 0 and 2; others were binary-coded (present vs. absent).

2.2.3.3. Video- and audio-information introduced (I^2). I^2 represents a set of message executional characteristics that can affect message elaboration by activating or in some circumstances reducing resource allocation and elaboration (Lang, 2006). We coded eight visual features (camera change [CC], emotion change, new object, unrelated scenes, object change, distance change, perspective change, and form change) and six audio features (orienting eliciting structural features [OESF] measuring any voice or sound change, new audio, unrelated audio, audio form change, emotional audio, and emotion change). CC and OESF produce orienting responses, but the information introduced following them can alter cognitive demands and resources allocated. Each I^2 category was coded for its presence at each CC/OESF, summed up, and then divided by the length of PSAs (seconds) to create final scores (Video-I^2: M = 32.73, SD = 21.06; Audio-I^2: M = 9.26, SD = 4.92).

2.2.3.4. Argument strength. Coders extracted central arguments from PSAs (one per each PSA). The arguments' persuasive strength (Zhao, Strasser, Cappella, Lerman, & Fishbein, 2011) was then assessed by independent samples of current smokers. Each argument was evaluated by 40 smokers on average (ranging 25 to 69). The argument-strength scores were aggregated by PSA (M = 3.45, SD = .39).

2.2.3.5. Positive and negative emotions. The emotional responses measured at the Wave 1 survey were used. Negative emotional responses were averaged across three items ("I felt *afraid*/*guilty*/*disgusted*") and positive emotional responses were averaged between two items ("I felt *hopeful*/*proud*"). Both emotion scores for a given PSA were obtained by aggregating (averaging) emotional responses from all participants who viewed that PSA (M = 2.61, SD = .19 for positive emotions; M = 3.05, SD = .31 for negative emotions).

2.2.4. Developing a Tailoring Algorithm-Based Prediction Model

Using the Wave 1 data, we developed a random intercepts and slopes multilevel model of participant i's PE evaluations of PSA j (PE_{ij}). The prediction model had the following three groups of components: (a) predicted scores

from user-based, item-based, and matrix factorization collaborative filtering algorithms; (b) individual-level (e.g., stage of change toward quitting smoking), message-level (e.g., argument strength), and cross-level (e.g., argument strength × stage of change toward quitting smoking) characteristics; and (c) random, participant-specific, intercepts and slopes.

We began with the best-performing model that was selected based on the analysis of the aforementioned 200-PSA evaluation data. The model was then fitted to the Wave 1 data to obtain optimal weights of its components and thereby maximize its predictive power. We used 87.5% of the data to train our prediction model and 12.5% of the data to test its predictive accuracy. This was done for each of the eight PSA evaluations. That is, to predict a participant's PSA rating in the *n*th evaluation, we used the data from the other seven evaluation occasions for training the model.

The tailoring algorithm-based prediction model of PE_{ij} had 69 degrees of freedom and included 13 random-effects parameters: one random-intercept, six random-slope, and six covariance parameters.[1] Key elements of the model are summarized below.

2.2.4.1. User-based collaborative filtering (UBCF) Regarding participant *i*'s PE evaluation of PSA *j*, UBCF calculates the similarity-weighted average PE ratings of PSA *j* made by all of the "neighbors" of participant *i*. Neighbors were defined as (a) those who co-rated one or more PSAs other than PSA *j* with participant *i* or (b) those who co-rated only PSA *j* with participant *i*. Since the data was sparse and the number of co-rated PSAs for user-user pairs was low (42.9% of all possible pairs had zero co-rated PSA, and 56.6% only had one, two, or three co-rated PSAs), we chose the inverted Euclidean distance, normalized by the number of co-rated PSAs for each participant-participant pair, as the similarity function. The commonly used cosine similarity performed worse. Participants' PE evaluation scores were mean-centered at the participant level before calculation to remove participant biases in the PE ratings of PSAs.

2.2.4.2. Item-based collaborative filtering (IBCF). IBCF calculated the PSA-to-PSA similarity-weighted and participant bias-removed average of participant *i*'s Wave 1 PE evaluations. That is, while UBCF weighted neighbors' PE ratings by the corresponding participant-participant similarity vector, IBCF computed prediction by examining participant *i*'s own history of PE ratings, and weighted the ratings for PSAs by the PSA-PSA similarity. IBCF operates under the premise that for a given smoker, her preference structure for PSAs should be relatively stable during the time period when the recommendation system is operating and hence should

give comparable PE ratings to pairs of PSAs that are similar—similar in the sense of receiving close PE ratings repeatedly by different smokers in the system. The PSA-PSA similarity was assessed using Sarwar's adjusted cosine similarity (Sarwar, Karypis, Konstan, & Riedl, 2001) after mean-centering by each participant's average PE rating. The parameter k – the number of PSAs from one's past evaluated set – in IBCF was set to the maximum (= 7).

2.2.4.3. Matrix factorization (MF). Hastie et al.'s MF algorithm (Hastie et al., 2015) combining alternative least squares with the singular value decomposition was employed. Unlike memory-based UBCF and IBCF, the MF approach to collaborative filtering assumes a generative model underlying the rating matrix, and its goal is to recover the generative model based on observed data (i.e. actual PE ratings) and then predict missing values in the same matrix using the recovered model. MF projects the observed matrix onto a lower-rank space where labels are characterized by latent content features and smokers by preferences for the same set of content features. Notably, such content features are latent, in the sense of being indirectly inferred from PE ratings rather than specified a priori by researchers. The value for the regularization parameter lambda was empirically calibrated on the training sample.

2.2.4.4. Individual characteristics of participant i. These included stage of change toward quitting, nicotine dependence, previous quitting attempts, perceived effects of smoking on health (single item), perceived health benefits of quitting, need for cognition, age, gender, race/ethnicity, education, household income, marital status, employment status, household head, and presence of children under the age of 18 in household.

2.2.4.5. Message characteristics of PSA j. these included argument strength, presence of narrative, negative emotions, positive emotions, presence of smoking cue, presence of efficacy information, video-, audio-, and content-level MSV, and video- and audio-level I^2.

2.2.4.6. Interactions. Interactions of individual characteristics of participant *i* with UBCF, IBCF, and message characteristics of PSA *j* were included as follows: (a) UBCF × perceived health benefits of quitting, (b) UBCF × age, (c) UBCF × presence of children under the age of 18 in household, (d) IBCF × gender, (e) IBCF × age, (f) IBCF × education, (g) negative emotions × nicotine dependence, (h) presence of efficacy information × nicotine dependence, (i) presence of efficacy information × need for cognition, (j) MSV audio × presence of children under the age of 18 in household, (k) MSV content × age.

2.2.4.7. Random intercept. Participant-specific intercepts.

2.2.4.8. Random slopes. Participant-specific random slopes for UBCF, IBCF, MF, negative emotions, positive emotions, and I^2 video.[2]

As indicated above, we used seven-evaluation data (87.5%) for training and one-evaluation data (12.5%) for testing. The predictive accuracy of the tailoring algorithm model was assessed in terms of the discrepancy between the model-predicted and actual PE ratings. The discrepancy was calculated using the root mean square error (RMSE). As a baseline comparison, we used a simple average of other participants' PE ratings for a given PSA (i.e., the information we used to provide BIS recommendations) because this is the standard metric typically employed to select messages for large-scale public health communication interventions. As each participant evaluated eight PSAs at Wave 1, we obtained eight RMSE values, respectively for the tailoring algorithm-predicted PE scores and the overall means of others' PE scores. The average of the eight RMSE values was .54 for the tailoring algorithm and .88 for the simple-average method. The tailoring model we developed outperformed the method typically used in practice in terms of predictive accuracy across all eight PSA evaluations (see Figure 2).

We fitted the tailoring algorithm-based prediction model to all eight evaluations from the Wave 1 data, and used parameter values from this model to obtain the predicted PE ratings for the 72 PSAs that participants had not watched at Wave 1 (i.e., predicted PE scores of the tailoring algorithm). Likewise, we calculated global means of the PE ratings provided by

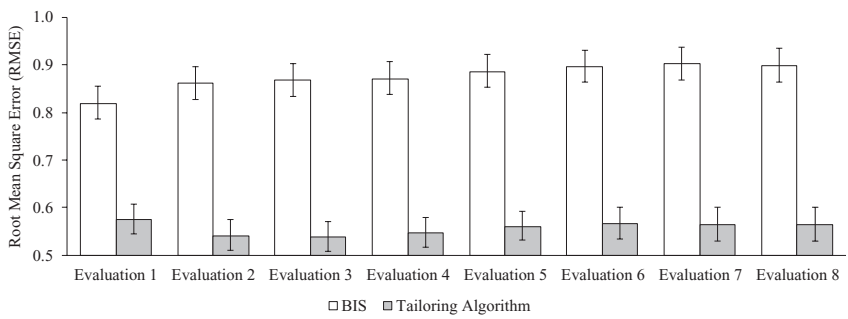

Figure 2. Predictive accuracy of the tailoring algorithm and BIS method (Wave 1 data). Simple averaging (global mean of other participants' ratings of a given PSA) was used for the BIS method. Values represent means and 95% confidence intervals of root mean square error of the predicted PE scores produced by the two methods. Due to uncertainty about the underlying distribution of the RMSE-difference estimate, bootstrapping was used to estimate means and bias-corrected confidence intervals (5,000 resamples).

other participants at Wave 1 for the 72 unexposed PSAs respectively for each participant (i.e., predicted PE scores of the BIS method). We produced both versions of prediction for each participant because random assignment of participants to experimental groups was designed to be made at the outset of Wave 2 and therefore it was necessary for us to be prepared for all possible results of the random assignment.

In sum, Wave 1 data were used to develop a prediction model tested at Wave 1 but applied at Wave 2 to PSAs as yet unseen by participants. The bases for the prediction model were algorithmic surrogates of social influence and item similarity over and above individual demographics and message features.

2.3. Wave 2: Randomized Experiment

A total of 675 participants from Wave 1 enrolled in the Wave 2 study, a web-based randomized experiment. Upon consenting, participants were randomly assigned to experimental conditions. They first answered a *pre-exposure survey* about their stage of change toward quitting smoking using the same measure as that of the Wave 1 survey (M = 6.76, SD = 2.85). They then proceeded to a *main experiment* phase in which they watched and rated PE of four PSAs. After evaluating four PSAs, they answered a *post-exposure survey* about their smoking cessation intention. The intention items were identical to those of the Wave 1 survey (α = .88, M = 3.36, SD = 1.19).

Participants were randomly assigned to one of three experimental conditions: (a) the *OTS* condition where four randomly selected PSAs were presented to participants (n = 245); (b) the *BIS* condition where the top-four PSAs based on average PE ratings of other participants at Wave 1 were selected for participants to watch (n = 201); (c) the *tailoring* condition where participants were exposed to the top-four from the list of 72 unseen PSAs based on predictions generated by the tailoring algorithm developed using the Wave 1 data (n = 229). After exposure to each of the four PSAs, participants evaluated PE using the same items as those of Wave 1 (α = .69, M = 3.56, SD = .85).

2.4. Wave 3: Follow-Up Survey

A total of 525 participants (about 49.7% of the Wave 1 sample; 77.8% of the Wave 2 sample) joined the Wave 3 study which was a follow-up survey. Of the 675 Wave 2 participants, there was no significant difference in attrition across experimental conditions between those who participated in the Wave 3 survey (n = 525) and those who did not (n = 150) at Wave 2, $\chi^2(2) = 1.60$, p = .45. Upon consenting, participants reported whether they

had changed or thought about changing their smoking behavior using a six-point scale ranging from "I have not made any changes to my smoking behavior" (= 1) to "I quit, and I'm still quit" (= 6) The mean of the quitting-behavior scale was 2.93 (*SD* = 1.54).

2.5. Statistical Analyses

2.5.1. Predictive accuracy. We tested the predictive accuracy of the tailoring algorithm and the simple-average approach using the data obtained from the participants in the OTS condition. The tailoring algorithm-predicted PSA ratings and the simple averages (global means) of other participants' PSA ratings were compared to the actual PE scores of the OTS condition participants. As participants were randomly assigned to the OTS condition at Wave 2, and the four PSAs they watched were randomly sampled from the pool of 72 PSAs that they had not watched at Wave 1, the comparison between the actual and predicted ratings of these participants allowed us to test the prediction precision of the tailoring and the BIS methods for the full range of predicted PE ratings. Predictive accuracy was measured by the discrepancy between the predicted and actual PE ratings using RMSE. Each participant evaluated four PSAs, so we calculated four RMSE scores for each method.

2.5.2. Message-intervention effects on PSA evaluations. To examine message-intervention effects on PE ratings, we fitted a cross-classified multilevel model that included random intercepts both at the participant and PSA levels. Specifically, as shown in *Equation 1*, the model regressed participant *i*'s PE rating of PSA *j* (PE_{ij}) on experimental conditions (i.e., *OTS* vs. *BIS* vs. *Tailoring*), individual characteristics of participant *i* (*I*), and the PSA-evaluation order indicators (*S*). The model had three disturbance terms: participant-level random intercept (α_i), PSA-level random intercept (α_j), and an idiosyncratic error term that varied across participant *i* and PSA *j*. The participant-level unobserved heterogeneity (α_i) was included because each participant rated four PSAs and thus the observations were not independent. Similarly, the PSA-level unobserved heterogeneity (α_j) was included because the set of four PSAs evaluated by participants differed not only across the experimental conditions but also within the same condition. The cross-classified multilevel model was estimated using the maximum likelihood method. The results are presented in Table 1.

$$PE_{ij} = \beta_0 + \beta_1 BIS_i + \beta_2 Tailoring_i + \delta_1 I1_i + \ldots + \delta_k Ik_i + \theta_1 S2_j + \ldots + \theta_3 S3_j + \alpha_i + \alpha_j = \varepsilon_{ij} \quad (1)$$

Table 1. Linear mixed effects regression of PE with maximum likelihood estimation

	Perceived Effectiveness (PE)
BIS (vs. OTS)	.104 (.064)
Tailoring (vs. OTS)	.118* (.056)
Stage of Change (Wave 2)	.037*** (.010)
Quitting Intention (Wave 1)	.397*** (.036)
Nicotine Dependence	.021* (.010)
Previous Quitting Attempts	−.051+ (.031)
Smoking Effects on Health	.036 (.025)
Health Benefits of Quitting	.022 (.024)
Need for Cognition	−.024 (.030)
Female	−.004 (.049)
Age ($\times 10^{-1}$)	.072*** (.021)
African American (vs. White)	.168* (.085)
Other (vs. White)	.028 (.067)
High School (vs. < HS)	.249+ (.129)
Some College (vs. < HS)	.234+ (.126)
BA or Higher (vs. < HS)	.235+ (.130)
Employed	.058 (.057)
Income	.011 (.015)
Married or Living w/ Partner	−.032 (.053)
Household Head	−.162* (.074)
Children < 18 Years Present	.108* (.052)
Evaluation Order: 2nd vs. 1st	−.003 (.027)
Evaluation Order: 3rd vs. 1st	−.030 (.027)
Evaluation Order: 4th vs. 1st	−.020 (.027)
Residual Variance	
Between-Participant	.268 (.018)
Between-PSA	.027 (.008)
Within	.247 (.008)
df	28
Log likelihood	−2551.944
AIC	5159.888
BIC	5325.116

Note. N = 2,700 (675 participants × 4 evaluations). Cell entries are unstandardized regression coefficients with standard errors in parentheses.
+ $p < .10$, * $p < .05$, *** $p < .001$.

2.5.3. Indirect effects of message interventions on behavioral intention and behavior change. We examined the behavioral consequences of the PSA recommendation methods using mediation analyses. Based on the integrative model of behavioral prediction (Fishbein & Ajzen, 2010), we focused on intention to quit smoking measured at Wave 2 and quitting behaviors reported at Wave 3 as key outcome variables. Our full mediation model,

from message interventions (Wave 2) to PE (Wave 2) to quitting intention (Wave 2) to quitting behavior (Wave 3), included three levels of variance: (a) between-participant (e.g., PE, quitting intention, quitting behavior, and individual-difference covariates), (b) between-PSA (PE), and (c) within (PSA-rating order). We tested the cross-classified multilevel mediation model using Bayesian estimation (Luo, 2017).[3]

3.0. Results

Our results show that the tailoring algorithm outperforms the simple-average approach in predictive accuracy. We compared the "observed" PE scores of the participants in the OTS condition at Wave 2 (i.e., those who watched four PSAs that were randomly sampled from the 72 previously unexposed PSAs) with the tailoring and BIS versions of the "predicted" PE scores for them. The tailoring algorithm generated more accurate predictions, showing significantly lower RMSE than the BIS method across all four PSA evaluations (see Figure 3).

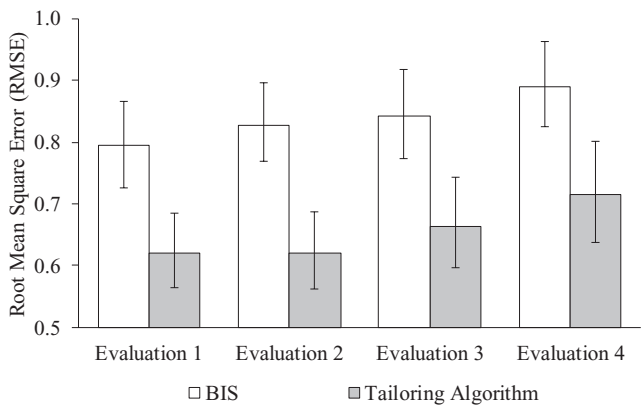

Figure 3. Predictive accuracy of the tailoring algorithm and BIS method (Wave 2 data). Simple averaging (an average of other participants' ratings of a given PSA) was used for the BIS selection. Values are predicted means and 95% confidence intervals. Due to uncertainty about the underlying distribution of the RMSE-difference estimate, bootstrapping was used to estimate the means and bias-corrected confidence intervals (5,000 resamples).

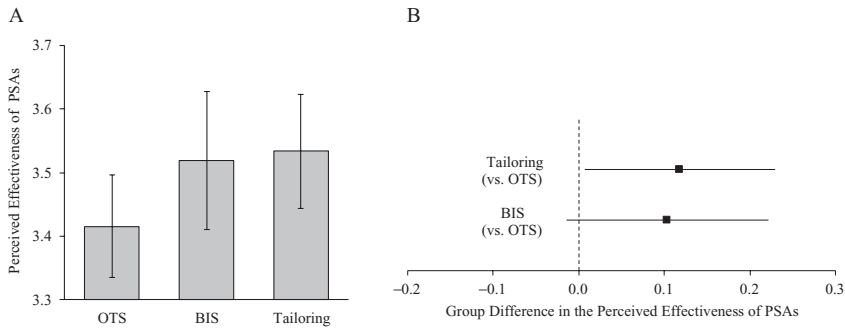

Figure 4. Effects of recommendation methods on PE of PSAs. Values represent predicted PE scores (A) and group mean differences (B) with 95% confidence intervals.

The results also reveal that message recommendation methods exert attitudinal and behavioral effects on smokers. First, as shown in Figure 4, participants who watched PSAs recommended by the tailoring algorithm, evaluated the PSAs as significantly more persuasive than those who viewed OTS PSAs. The difference between the BIS and OTS methods was not statistically significant, and the tailoring algorithm did not outperform the BIS method at a statistically significant level.

Why, then, did the tailoring algorithm produce higher PE ratings than OTS selection, whereas the BIS method failed to do so? An ancillary analysis was conducted to address this question, focusing on the role of the differential predictive accuracy of the tailoring and BIS methods. We focused on the possibility that differential within-participant variability in PE ratings (i.e., across the four PSA evaluations) might explain the result. The coefficient of variation (CV) was used to quantify the within-participant variability for each participant (N = 675): the ratio of the standard deviation of a participant's four PE ratings to the mean of the four ratings. The distribution of CV was positively skewed (M = .13, SD = .10, $Median$ = .10, Min = 0, Max = .60), and the square-root transformation normalized it (M = .32, SD = .16, Min = 0, Max = .77). We thus conducted an analysis of variance using the square root of CV: $F(2, 672) = 5.33$, $p = .005$. We found that the BIS (M = .32, SD = .15) and OTS (M = .34, SD = .17) conditions did not significantly differ in within-participant variability across the PE ratings of four PSAs (see Figure 5), although the BIS method selected four PSAs that received highest PE scores from other participants and thus was expected to induce lower within-participant variability. In contrast, the tailoring algorithm (M = .29, SD = .16) produced significantly lower within-participant variability

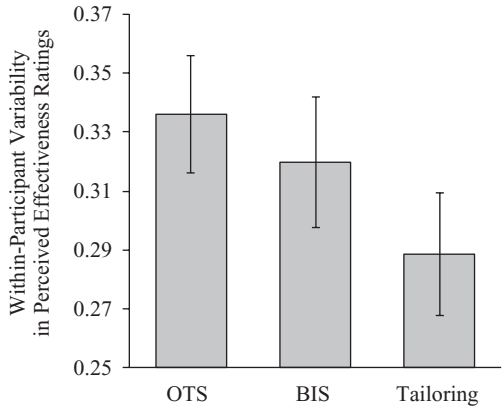

Figure 5. Group differences in within-participant variability in the PE ratings of four PSAs. As a measure of variability, we used the coefficient of variation. Values are predicted means and 95% confidence intervals.

than the two non-tailored methods; that is, it better approximated participants' latent preferences for antismoking PSAs. In sum, the results suggest that the relatively low predictive accuracy of the BIS method might have resulted in the non-significant difference with OTS selection in PE ratings.

Next, we examined how the perceived persuasiveness of the PSAs – shaped by the PSA recommendation methods – further influences quitting intention and behavior (see Table 2). Figure 6 shows that PE was positively associated with intention to quit smoking, identifying a significant indirect effect of the tailoring algorithm (vs. OTS selection) on intention by enhancing the PE ratings of the PSAs, *b* (Bayesian posterior median estimate) = .044, 95% credibility intervals [.003, .100]. Quitting intention in turn predicted quitting behavior. The PSAs recommended via algorithm were more likely than those dictated by the OTS method to promote smoking cessation behavior by increasing PE of the PSAs and quitting intention, *b* = .022, 95% credibility intervals [.002, .052]. The corresponding indirect effect of the BIS procedure (vs. OTS selection) was not statistically significant on either intention (*b* = 032, 95% credibility intervals [–.006, .070]) or behavior (*b* = .016, 95% credibility intervals [–.003, .039]).

4.0. Discussion

The results offer some evidence that an algorithmic selection of tobacco cessation PSAs tailored to the specific person is predictive of smokers'

Table 2. Linear mixed effects regression of PE with maximum likelihood estimation

	PE (Wave 2)	Intention (Wave 2)	Behavior (Wave 3)
Between-Participant			
BIS (vs. OTS)	.092 [−.017, .218]	−.022 [−.094, .053]	.185 [−.090, .441]
Tailoring (vs. OTS)	.124* [.009, .267]	−.040 [−.127, .027]	.234 [−.013, .448]
PE (Wave 2)	n/a	.367* [.286, .425]	
Quitting Intention (Wave 2)	n/a	n/a	.517* [.323, .706]
Stage of Change (Wave 2)	.035* [.015, .058]	.078* [.062, .090]	.146* [.098, .202]
Quitting Intention (Wave 1)	.398* [.323, .465]	.452* [.400, .508]	
Nicotine Dependence	.019* [.001, .038]	.000 [−.015, .015]	−.057* [−.104, −.004]
Previous Quitting Attempts	−.050 [−.101, .008]	.029 [−.014, .076]	.093 [−.043, .272]
Smoking Effects on Health	.037* [.000, .079]	.001 [−.039, .033]	.026 [−.105, .125]
Health Benefits of Quitting	.021 [−.025, .074]	−.045* [−.079, -.018]	.002 [−.113, .116]
Need for Cognition	−.021 [−.079, .043]	.042 [−.014, .076]	.121 [−.002, .242]
Female	−.006 [−.120, .093]	−.049 [−.106, .038]	-.206 [−.425, .085]
Age ($\times 10^{-1}$)	.075* [.037, .115]	−.029* [−.059, −.002]	.041 [−.058, .150]
African American (vs. White)	.154* [.005, .363]	.032 [−.062, .144]	.588* [.200, .954]
Other (vs. White)	.020 [−.131, .166]	.064 [−.049, .144]	.208 [−.126, .465]
High School (vs. < HS)	.237 [−.044, .481]	.169 [−.032, .371]	−.598* [−1.15, −.064]
Some College (vs. < HS)	.216 [−.019, .510]	.074 [−.108, .251]	−.531 [−1.245, .056]
BA or Higher (vs. < HS)	.203 [−.101, .447]	.053 [−.123, .219]	−.262 [−.900, .299]
Employed	.056 [−.057, .177]	.059 [−.015, .135]	−.175 [−.414, .111]
Income	.014 [−.023, .040]	−.015 [−.039, .010]	−.006 [−.079, .065]
Married or Living w/ Partner	−.024 [−.139, .079]	.103* [.018, .163]	.307* [.117, .590]
Household Head	−.166*[−.323,−.039]	.093 [−.020, .176]	.141 [−.103, .458]
Children < 18 Years Present	.108* [.003, .203]	−.049 [−.126, .013]	−.024 [−.287, .237]
Within			
Evaluation Order: 2nd vs. 1st	−.003 [−.046, .055]		
Evaluation Order: 3rd vs. 1st	−.029 [−.073, .022]		
Evaluation Order: 4th vs. 1st	−.019 [−.057, .038]		
Residual Variance			
Between-Participant	.278* [.246, .325]	.176* [.161, .194]	1.578* [1.425, 1.837]
Between-PSA	.027* [.015, .042]	n/a	n/a
Within	.248* [.235, .265]	n/a	n/a

Note. $N = 2{,}700$ (675 participants × 4 evaluations). Cell entries are Bayesian estimates of unstandardized coefficients (*medians* of posterior distributions) with 95% credibility intervals in brackets. Point estimates with asterisk are those whose 95% credibility intervals do *not* include zero. Markov chain Monte Carlo (MCMC) algorithms were used to iteratively obtain an approximation of posterior parameter distributions (iterations carried out in two parallel and independent chains). The convergence criterion was that a Proportional Scale Reduction (PSR) factor was close enough to 1 for each parameter, more specifically, less than 1.1. Non-informative priors were used: N (0, ∞) for regression slopes and IG (−1, 0) for variances. Model fit was assessed by the Bayesian posterior predictive checking using the likelihood-ratio chi-square: 95% confidence interval for the difference between the observed and replicated chi-square values was [−19.013, 39.189], with the Posterior Predictive P-value of .250, indicating a good fit. The coefficients for the effects of PE (Wave 2) and intention (Wave 1) on behavior (Wave 3) were constrained to be zero. We made this specification to create a parsimonious model which is consistent with theoretical and empirical literature on the integrative model of behavioral prediction (12) and PE (4). Testing the mediation model without these two constraints yielded similar results. The results reported here also remained similar when running longer chains with a fixed number of Bayes iterations (e.g., 10,000 & 20,000).

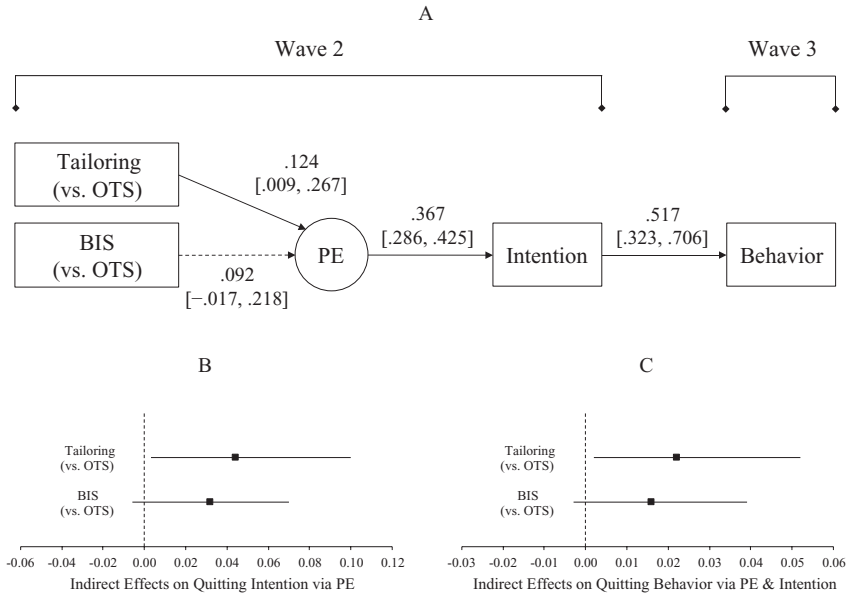

Figure 6. Indirect effects of recommendation methods on smoking cessation intention and behavior. (A) shows the mediation model operating at the between-participant level. (B) and (C) present the indirect effects of recommendation methods on quitting intention (through PE) and behavior (through PE and intention), respectively. Values represent Bayesian posterior median estimates of unstandardized coefficients and 95% credibility intervals.

perceived persuasiveness of the PSAs, and further exerts a small but significant effect on behavior – in this case quitting attempts by smokers. Recommendation algorithms whose function is to tailor message selection holds promise for effective health communication. The rank ordering of effects from low to high was messages selected OTS, selected as BIS, and tailored by algorithm.

One might argue that BIS is the preferred option because it is cheaper to conduct than tailoring and is almost as effective. We offer some counters to this position below. It should be immediately noted that the testing context involved a very large set of messages with the BIS set topping out a group of some 200 PSAs, a sample most campaign projects would never have. So our BIS is more like the "very best in show" offering a very tough comparison against the tailoring option. In more realistic circumstances with 20 to 50 messages available for testing, best in show might not quite be very best in show.

The test conducted was not just a test that an algorithm could be created to predict message selection but a test of the (self-reported) behavioral consequences of messages so selected. If algorithmic recommendations are to be useful in human behavioral studies, evaluation of those algorithms cannot simply be that selection is successfully predicted but must include the possibility that the selection has behavioral consequences of concern to campaign planners. Hence, our test was a behavioral one ultimately and, while only small effects result, the behavior affected–quit attempts–is one of consequence and one that is difficult to move.

One objection that could be raised here is that the effects are too small to be important. This is not an unreasonable argument but in the domain of tobacco consumption where rates in adults have been plummeting, the remaining set of smokers is going to offer difficult targets for change. Any increments to change in the remaining cadre of smokers is a welcome occurrence even if the increments are small ones.

Algorithms are computational tools and seemingly divorced from the human social and psychological processes that would ordinarily be on the surface of behavior change work. So what could the algorithmic procedures represent socially and psychologically? Collaborative filtering is clearly based on a kind of social influence in that those messages that others similar to the target person find persuasive, the target person will also find persuasive. But, in the algorithm case, the social influence is influence without social interaction. Recommendations made by the algorithm are not recommendations made by people through the normal mechanisms of social interaction. Instead, the recommendations are implicit ones made by influencers unknown to the person being influenced. De facto the implicit recommendations made without social interaction appear to have an effect not unlike recommendations made with social interaction. So, recommendation algorithms at least in part mimic social influence processes.

Unsurprisingly the OTS approach does least well in predicting intentions and outcomes. At one point, we labeled this approach the "random selection" approach but that characterization is unfair and not descriptive. Messages are not truly randomly selected because the initial set of messages in our test are ones that have been relatively carefully developed for use in tobacco cessation campaigns with adults. They are relatively effective messages in general no matter which are selected. They are just not as effective as ones chosen specifically for an individual.

The purpose of this research has *not* been to find the best algorithm for anti-tobacco messages tailored to adult smokers. Tailoring algorithms are *not* compared to one another in our study. Instead, approaches to message

selection as a part of campaign planning and design are compared to see which approach has the best predictive value in message evaluation and ultimately which is most predictive of behavioral change. Our results are certainly suggestive that algorithmic tailoring is as good or better than two other standard approaches to message selection. Whether a more predictively powerful algorithm can be developed in comparison to ours awaits subsequent inquiry.

Researchers might view the methodology we followed as very data heavy and therefore an unrealistic methodology in practical applications. We offer three responses to this concern. First, other tailoring work in message selection is also data intensive and requires a large amount of psychographic information from sources in order to be implemented and achieve its benefits. So tailoring is in general data intensive whether algorithmic or psychographic. Second, the work we report is based on individualized message choices and not on a criterion such as micro-targeting very small groups of individuals. Micro-targeting very small, homogeneous groups of individuals would reduce data burdens by treating the micro groups as if they were individuals selecting the same message for everyone in the micro cluster but reducing data gathering burdens substantially. Whether micro-targeting is as effective as tailoring is an empirical question. Third, in this paper we have not explored alternatives to collaborative filtering such as using demographic and psychographic procedures to find clones to recommend effective messages. This alternative to the sparse data problem might be effective enough for future applications of recommendations systems. Future research will tell the value of these other approaches.

Our purpose in this manuscript was not to test various computational engines against one another but rather to derive one that would predict – in a true sense of prediction over time and not simply variance explained – behaviorally reported changes in a consequential behavior from messages previously unseen but selected on computational criteria that are analogous to social interaction and to content similarity. We carried out this test not just with the algorithm on its own as has been done in other work but by contrasting it to two other standard and well-established – but simple -- prediction procedures.

The algorithmic approach to message recommendation is worth serious, long-term investment by the research community. First, the availability of large numbers of eligible messages through archiving and data sharing and through social media sources allows, and even invites, extensive message testing across targeted populations. For example, websites for the “Wounded Warrior Project” (www.woundedwarriorproject.org) and

"Doctors without Borders" (www.doctorswithoutborders.org) each have hundreds of videos and many thousands of visitors. Navigating sites like these could be made more efficient with the help of an effective algorithm after even just a few video views per visitor. These large archives make it impossible for everyone to see or hear every message being considered. So, message evaluation needs to be shared across the population. The success of tailoring approaches in various applications suggests that tailoring works. But tailoring suffers from the need for extensive private information from individuals and from a lack of consensus about what kinds of data from each person would be useful in message selection. The algorithmic approach focusing as it does on behavioral tailoring and mechanization of message selection makes possible an optimal balance of working with large groups of people while selecting messages which appeal to specific persons. The promise and potential of mass audiences with tailored appeals is great enough to continue to explore, model and compute message recommendation systems.

Acknowledgements

The authors acknowledge the funding support of the National Cancer Institute at the National Institutes of Health (R01CA160226). The content is solely the responsibility of the authors and does not necessarily represent the views of the funding agency.

Notes

1 The tailoring algorithm-based prediction model's log-likelihood was –7685.01 (*df* = 69; AIC = 15508.02; BIC = 15993.89). The intercept-only model's log-likelihood was –8798.74 (*df* = 3; AIC = 17603.49; BIC = 17624.61). The tailoring model was developed using a combination of (a) a priori specification and (b) a build-up strategy. First, we obtained the set of individual-level and message-level predictors from the analysis of the 200-PSA data from which the current 80 PSAs were sampled. Second, we estimated a multilevel model with random intercepts and slopes (with unstructured error covariance structure), while testing one slope at a time. Third, we removed non-significant interaction terms (i.e., non-significant predictors of the slopes; one at a time). Finally, we removed random-effects terms for zero-residual-variance slopes (one at a time). Additional information about the tailoring model is available upon request from the authors.

2 We did not include message-level random intercepts or slopes because doing so caused the model to fail to converge. The non-convergence problem occurred probably because the model already included the UBCF-generated similarity-weighted average PE ratings of PSA *j* provided by other participants at Wave 1, which possibly explained all of the

variance at the message (given the relatively small number of messages [n = 80], compared with the number of participants [n = 1,057]).

3 To our knowledge, the Bayes estimator is the best-established estimation method for the analysis of a cross-classified multilevel mediation model. A desirable property of the Bayes estimator is that it does not need to assume symmetry in the distributions of indirect effects. At the time of this writing, we are not aware of any other better estimators (e.g., maximum likelihood) that can be used to examine indirect effects in cross-classified multilevel data.

References

Adomavicius, G. & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 734-749. doi:10.1109/TKDE.2005.99

Aggarwal, C. C. (2016). *Recommender systems: The textbook*. New York, NY: Springer.

Amatriain, X. & Basilico, J. (2015). Recommender systems in industry: A Netflix case study. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (2nd ed., pp. 385-419). New York, NY: Springer.

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education and Behavior*, *31*, 143-164. doi:10.1177/1090198104263660

Biener, L. & Abrams, D. B. (1991). The Contemplation Ladder: Validation of a measure of readiness to consider smoking cessation. *Health Psychology*, *10*, 360-365. doi:10.1037/0278-6133.10.5.360

Bigsby, E., Cappella, J. N., & Seitz, H. H. (2013). Efficiently and effectively evaluating public service announcements: Additional evidence for the utility of perceived effectiveness. *Communication Monographs*, *80*, 1-23. doi:10.1080/03637751.2012.739706

Buhmann, A., Likely, F., & Geddes, D. (2018). Communication evaluation and measurement: Connecting research to practice. *Journal of Communication Management*, *22*, 1–7. doi:10.1108/JCOM-12-2017-0141

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306-307. doi:10.1207/s15327752jpa4803_13

Cappella, J. N. (2018). Perceived message effectiveness meets the requirements of a reliable, valid, and efficient measure of persuasiveness. *Journal of Communication*, *68*, 994-997. doi:10.1093/joc/jqy044

Cappella, J. N., Yang, S., & Lee, S. (2015). Constructing recommendation systems for effective health messages using content, collaborative, and hybrid algorithms. *Annals of the American Academy of Political and Social Science*, *659*, 290-306. doi:10.1177/0002716215570573

Carey, M. A. (1994). The group effect in focus groups: Planning, implementing, and interpreting focus group research. In Morse, J.M. (Ed.), *Critical issues in qualitative research methods* (pp. 225-241). Thousand Oaks, CA: Sage.

Dillard, J. P., Shen, L., & Vail, R. G. (2007). Does perceived message effectiveness cause persuasion or vice versa? 17 consistent answers. *Human Communication Research*, *33*, 467-488. doi:10.1111/j.1468-2958.2007.00308.x

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach* New York, NY: Psychology Press.

Hastie, T., Mazumder, R, Lee, J. D., & Zadeh, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, *16*, 3367-3402.

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerström, K.-O. (1991). The Fagerström Test for Nicotine Dependence: A revision of the Fagerström Tolerance Questionnaire. *British Journal of Addiction*, *86*, 1119-1127. doi:10.1111/j.1360-0443.1991.tb01879.x

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2011). *Recommender systems: An introduction*. New York, NY: Cambridge University Press.

Kang, Y., Cappella, J.N., Strasser, A.A., & Lerman, C. (2009). The effect of smoking cues in antismoking advertisements on smoking urge and psychophysiological reactions. *Nicotine & Tobacco Research*, *11*, 254-261. doi:10.1093/ntr/ntn033

Kelder, S. H., Pechmann, C., Slater, M. D., Worden, J. K., & Levitt, A. (2002). The national youth anti-drug media campaign. *American Journal of Public Health*, *92*, 1211-1212. doi:10.2105/AJPH.92.8.1211

Kim, M. & Cappella, J. N. (in press). Reliable, valid and efficient evaluation of media messages. *Journal of Communication Management*.

Koren, Y. & Bell, R. (2015). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (2nd ed., pp. 77-118). New York, NY: Springer.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*, 30-37. doi:10.1109/MC.2009.263

Krebs, P., Prochaska, J. O., & Rossi, J. S. (2010). A meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine*, *51*, 214-221. doi:10.1016/j.ypmed.2010.06.004

Kreuter, M. W., Farrell, D., Olevitch, L., & Brennan, L. (2000). *Tailoring health messages: Customizing communication with computer technology*. Mahwah, NJ: Earlbaum.

Lang, A. (2006). Using the limited capacity model of motivated mediated message processing to design effective cancer communication messages. *Journal of Communication*, *56*, S57-S80. doi:10.1111/j.1460-2466.2006.00283.x

Luo, W. (2017). Testing mediation effects in cross-classified multilevel data. *Behavior Research Methods*, *49*, 674-684. doi:10.3758/s13428-016-0723-3

Lustria, M. L. A., Noar, S. M., Cortese, J., Van Stee, S. K., Glueckauf, R. L., & Lee, J. (2013). A meta-analysis of web-delivered tailored health behavior change interventions. *Journal of health communication*, *18*, 1039-1069. doi:10.1080/10810730.2013.768727

Macnamara, J. (2018). A review of new evaluation models for strategic communication: Progress and gaps. *International Journal of Strategic Communication*, *12*, 180-195. doi:10.1080/1553118X.2018.1428978

Morgan, S. E., Palmgreen, P., Stephenson, M. T., Hoyle, R. H., & Lorch, E. P. (2003). Associations between message features and subjective evaluations of the sensation value of antidrug public service announcements. *Journal of Communication*, *53*, 512-526. doi:10.1111/j.1460-2466.2003.tb02605.x

Ning, X., Desrosiers, C., & Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook*, In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (2nd ed., pp. 37-76). New York, NY: Springer.

Noar, S. M., & Harrington, N. G. (2016). Tailored communications for health-related decision-making and behavior change. In M. A. Diefenbach, S. Miller-Halegoua, & D. J. Bowen (Eds.), *Handbook of health decision science* (pp. 251-263). New York, NY: Springer.

Nonnemaker, J., Farrelly, M. C., Kamyab, K., Busey, A., & Mann, N. (2010). *Experimental study of graphic cigarette warning labels*. Research Triangle Park, NC: RTI International.

Parvanta, S., Gibson, L., Forquer, H., Shapiro-Luft, D., Dean, L., Freres, … & Cappella, J. N. (2013). Applying quantitative approaches to the formative evaluation of antismoking campaign messages. *Social Marketing Quarterly*, *19*, 242-264. doi:10.1177/1524500413506004

Petty, R. E., Barden, J., & Wheeler, S. C. (2009). The elaboration likelihood model of persuasion: Developing health promotions for sustained behavioral change. In R. J. DiClemente, R. A.

Crosby, & M. C. Kegler (Eds.), *Emerging theories in health promotion practice and research* (2nd ed., pp. 185-214). San Francisco, CA: Jossey-Bass.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Resnick, P. & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*, 56-58. doi:10.1145/245108.245121

Rimer, B. K., & Kreuter, M. W. (2006). Advancing tailored health communication: A persuasion and message effects perspective. *Journal of Communication*, *56*, S184-S201. doi:10.1111/j.1460-2466.2006.00289.x

Sadasivam, R. S., Cutrona, S. L., Kinney, R. L., Marlin, B. M., Mazor, K. M., Lemon, S. C., & Houston, T. K. (2016). Collective-intelligence recommender systems: Advancing computer tailoring for health behavior change into the 21st century. *Journal of medical Internet research*, 18, e42. doi:10.2196/jmir.4448

Sarwar, B. Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, 285-295.

Slater, M. D. (2006). Specification and misspecification of theoretical foundations and logic models for health communication campaigns. *Health Communication*, *20*, 149-157. doi:10.1207/s15327027hc2002_6

Zhao, X., Alexander, T. N., Hoffman, L., Jones, C., Delahanty, J., Walker, M., Berger, A. T., & Talbert, E. (2016). Youth receptivity to FDA's the real cost tobacco prevention campaign: Evidence from message pretesting. *Journal of Health Communication, 21*, 1153-1160. doi:10.1080/10810730.2016.1233307

Zhao X., Strasser, A., Cappella, J. N., Lerman, C., & Fishbein, M., (2011). A measure of perceived argument strength: reliability and validity. *Communication Methods and Measures*, *5*, 48-75. doi:10.1080/19312458.2010.547822

About the authors

Hyun Suk Kim: Department of Communication, Seoul National University, Seoul, 08826, Republic of Korea.

Sijia Yang: School of Journalism and Mass Communication, University of Wisconsin-Madison, Madison, WI 53706, USA.

Minji Kim: Center for Tobacco Control and Research, University of California, San Francisco, San Francisco, CA 94143, USA.

Brett Hemenway: Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

Lyle Ungar: Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

Joseph N. Cappella: Annenberg School for Communication, University of Pennsylvania, Philadelphia
Correspondance address: Joseph N. Cappella, Annenberg School for Communication, University of Pennsylvania, 3620 Walnut St., Philadelphia, PA 19104. Phone: 215-898-7041. Email: joseph.cappella@asc.upenn.edu