Published in final edited form as:
Cognit Ther Res. 2018 December ; 42(6): 747–757.

Personalizing Affective Stimuli Using a Recommender Algorithm: An Example with Threatening Words for Trauma Exposed Populations

Andrea N. Niles[1,2], **Aoife O'Donovan**[1,2]

Andrea N. Niles: andrea.niles@ucsf.edu

[1]Department of Psychiatry and Weill Institute for Neurosciences, University of California, San Francisco, 4150 Clement Street, San Francisco, CA, USA

[2]San Francisco Veterans Affairs Medical Center, 4150 Clement St. (116C1), San Francisco, CA, 94121, USA

Abstract

Experimental paradigms used in affective and clinical science often use stimuli such as images, scenarios, videos, or words to elicit emotional responses in study participants. Choosing appropriate stimuli that are highly evocative is essential to the study of emotional processes in both healthy and clinical populations. Selecting one set of stimuli that will be relevant for all subjects can be challenging because not every person responds the same way to a given stimulus. Machine learning can facilitate the personalization of such stimuli. The current study applied a novel statistical approach called a recommender algorithm to the selection of highly threatening words for a trauma-exposed population (N = 837). Participants rated 513 threatening words, and we trained a user–user collaborative filtering recommender algorithm. The algorithm uses similarities between individuals to predict ratings for unrated words. We compared threat ratings for algorithm-based word selection to a random word set, a word set previously used in research, and trauma-specific word sets. Algorithm-selected personalized words were more threatening compared to non-personalized words with large effects (*d*s = 2.10–2.92). Recommender algorithms can automate the personalization of stimuli from a large pool of possible stimuli to maximize emotional reactivity in research paradigms. These methods also hold potential for the personalization of behavioral treatments administered remotely where a provider is not available to tailor an intervention to the individual. The word personalization algorithm is available for use online (https://threat-word-predictor.herokuapp.com/)

Correspondence to: Andrea N. Niles, andrea.niles@ucsf.edu.

Compliance with Ethical Standards

Conflict of Interest Drs. Niles and O'Donovan declare that they have no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Animal Rights Statements No animal studies were carried out by the authors for this article.

Keywords

Anxiety; Posttraumatic stress disorder; Emotion; Affect; Recommender algorithm; Trauma; Personalization; Personalized medicine

Introduction

Laboratory and computer-based paradigms used in affective and clinical science frequently employ evocative images, scenarios, videos, and words to elicit an emotional response in study participants. For example, the International Affective Picture System (IAPS; Lang et al. 1999), a picture set that includes positive, neutral, and negative images, is widely used in clinical and affective science and has been cited over 3130 times as of March 7, 2018. Researchers typically select a set of stimuli that they believe will be relevant for the phenomenon they are studying (e.g. fear, disgust, excitement, sadness) and hope that the stimuli do in fact reliability elicit the intended emotion in most or all study participants. However, although a stimulus may strongly elicit a certain emotion *on average*, the experience of that stimulus will differ depending on the individual. For example, ratings of the IAPS images have a mean standard deviation of 1.6 for valence and 2.2 for arousal on a 1–9 scale, meaning that it is quite common for participants to rate the stimuli 1.5–2 points above or below the mean rating. Thus, methods to select personalized stimuli for study participants can facilitate greater precision when activating emotional responses in the lab and reduce noise resulting from individual differences in responses to stimuli. Further, such methods could be used to personalize the content of behavioral treatments administered via Internet and mobile applications. In the current study, we developed and tested an algorithm to select emotionally evocative stimuli that are tailored to an individual participant. In the current demonstration, we select personalized threatening words for use with trauma-exposed populations.

Recommender systems are machine-learning algorithms that predict how a person will rate an item or product (e.g. movies, music, news, books, research articles). They can be used to match a person with potential jobs, collaborators, restaurants, and even romantic partners. Recommender algorithms underlie the personalization of social media content such as one’s Facebook, Twitter, or Instagram feed and guide the selection of personalized advertisements that appear in one’s web browser. These systems are ubiquitous in technology because they work extremely well to personalize what a person interacts with online. Personalization improves the experience of the user by displaying only content that he or she will be most likely to enjoy or find useful. Prior researchers have yet to apply recommender systems to select personalized stimuli for use in psychological research.

One type of emotionally evocative stimuli that has been used to assess and modify affective responses, cognitive biases, and attention allocation, is words. Researchers have employed the use of positive, negative, and neutral words to better understand affective responses in both healthy and clinical populations. The affective norms for English words (ANEW; Bradley and Lang 1999), which provides valence and arousal ratings for thousands of English words, has been used in a wide variety of studies and has been cited approximately

2400 times as of March 7, 2018. Evocative words have been used, for example, to assess neural responses to emotional stimuli in depression (Siegle et al. 2002), memory for emotional stimuli (Kensinger and Corkin 2003), performance for sustained attention and attention switching (Chambers et al. 2008), and assessment and modification of attentional bias for threat words in anxiety disorders (Amir et al. 2009).

Although the majority of studies apply the same set of word stimuli for all study participants, a few investigators have attempted to select personally relevant words for use in computerized attention training paradigms for anxiety and alcohol use disorders. Amir and colleagues (2009, 2016) used what they call an "idiographic" approach to stimuli selection. Participants are asked to rate word stimuli on a scale, and the researchers select a subset of those words rated most highly by the participant for inclusion in training programs. A similar method involves showing participants a set of words and asking them to select the most relevant subset (McGeary et al. 2014). Another approach involves participants spontaneously producing personally relevant words (Fridrici et al. 2013). Finally, for a study using personalized words for individuals with PTSD, clinical interviews conducted with the participant were transcribed, then the interviewer selected words from the transcription based on emotional value, distinctiveness, and number of repetitions in the interview (Schoorl et al. 2014). For research in clinical populations, the approach used most frequently to match stimuli to the individuals in the study, is to recruit a specific patient population and select stimuli relevant for that population (e.g. angry faces in social phobia, contamination-related stimuli in obsessive compulsive disorder, sad faces in depression, combat images for veterans with combat-trauma).

These personalization methods have limitations. For the idiographic approach, there is a limit to how many stimuli a participant can rate or consider, limiting the total pool of stimuli that can be used. The spontaneous production method requires the user to understand what type of word or stimulus is needed and puts a significant burden and responsibility on the user to produce an appropriate response. Further, the task of spontaneously producing words can be limited by cognitive impairment or attention deficit hyperactivity disorder (Shao et al. 2014) and the words produced using this approach may not actually represent the most personally relevant words. For clinical populations, selecting words derived from a clinical interview requires significant time and personnel investment including availability of highly trained interviewers, transcription of the interviews, and a protocol for word selection from the interviews. This method also relies on a patient's ability to produce a wide variety of the type of stimuli that are needed for a given study during the interview. Methods that allow selection from a large set of words and that require minimal effort by both researchers and users could significantly facilitate our ability to select personalized word stimuli within research paradigms and increase the relevance and emotional salience of word stimuli for every study participant. Finally, although focusing on a specific patient population may be ideal depending on the goals of the researcher, the selection of stimuli for research on transdiagnostic disorder mechanisms or treatments becomes more challenging.

The current study aimed to develop and test the accuracy of a recommender algorithm for identifying personalized threatening words. Because a number of prior studies have used words to assess various cognitive and affective processes in trauma-exposed individuals, a large number of word stimuli relevant for such populations (in addition to the words from the ANEW dataset) were available (Beck et al. 2001; Blix and Brennen 2011; Bunnell 2007; Ehring and Ehlers 2011; Roberts et al. 2010; Swick 2009). Therefore, we chose to develop the recommender algorithm in a trauma-exposed population and used words selected for previous studies to create our pool of word stimuli. To develop the algorithm, we tested different values of model parameters to optimize performance. Then, to compare the threat level of the words selected using the algorithm approach to that of non-personalized words, we tested whether the algorithm outperformed: (a) random word selection; (b) a standard word set that has been used in prior studies (MacLeod et al. 2002); and (c) word selection based on trauma type.

Methods

Participants

The study protocol was approved by the UCSF Institutional Review Board. Participants were recruited via Amazon’s Mechanical Turk (MTurk) platform (Buhrmester et al. 2011). Participants completed a screening questionnaire to assess trauma exposure (N = 1605), and only those who had been exposed to trauma that “really bothered [them] emotionally” were invited to complete the remainder of the study (N = 1223). Eligible participants had a mean age of 34 years old (SD = 10.7). In terms of gender, 55.8% identified as female, and 0.4% as non-binary or third gender. Participants received 10 cents for the screener and $4.50 for completing the word ratings. Of all eligible participants, 881 (72%) participants rated all the words as requested, 231 (19%) participants rated some, but not all words, and 111 (9%) participants did not rate any words. Participants who rated some of the words were given partial payment.

To identify people who may have been responding randomly or whose data would not inform the algorithm, we used a number of tactics. We removed participants who rated positive or neutral words as highly threatening (N = 34), those who had very low variability (SD < 1) in their responses across the words (N = 10), participants who rated fewer than 50 words (N = 155), and participants who met more than one of these criteria (N = 76). Participants included in the final dataset (N = 837) were 34.8 years old (SD = 11.0, Range = 19–71) and 60.1% female. Participants reported trauma exposure as shown in Table 1. The average number of different trauma types reported was 4.2 (SD = 2.5) with a range from 1 to 11.

Sample Size Determination

Data were collected in five stages to allow us to check that participants were providing sufficient variability of responses and to check the performance of the algorithm. Because we had no way to estimate the total sample size needed ahead of data collection, we planned to track the algorithm performance with each stage of data collection and to end data collection when inclusion of additional participants no longer improved the performance of the algorithm. Algorithm performance reached an asymptote (and even slightly decreased) at approximately 700 participants (as shown in Fig. 1) providing evidence that our sample size of 837 was sufficient to achieve optimal model performance.

Materials

Trauma History Screen

The Trauma History Screen (Carlson et al. 2011) is a very brief self report measure used to assess exposure to events associated with significant and persisting posttraumatic stress. The measure assesses for the occurrence of 14 traumatic or stressful events such as "a really bad car, boat, train, or airplane accident," and "attack with a gun, knife, or weapon." The measure then asks, "did any of these things really bother you emotionally?" In the current study, to increase the likelihood of identifying trauma involving threat to life or physical integrity, we eliminated the "other," "sudden move or loss of home and possessions," and "suddenly abandoned by spouse, partner, parent or family" items. Further, to reduce participant burden we only asked participants to indicate whether or not they experienced each event, but did not ask them to report the number of times they experienced events or to elaborate on events experienced. Temporal stability of the measure is good to excellent, it shows convergent validity with longer measures and it is well understood by people with low reading levels (Carlson et al. 2011).

Word Selection

Threatening words were selected based on previous studies that have used words to measure or modify cognitive processes in PTSD and trauma-exposed individuals (Beck et al. 2001; Blix and Brennen 2011; Bunnell 2007; Ehring and Ehlers 2011; Roberts et al. 2010; Swick 2009), as well as from the affective norms for English words (ANEW) standard word rating set (Bradley and Lang 1999), and a study examining cognitive processing of emotional words (Kousta et al. 2009). ANEW words had previously been rated on dimensions of valence and arousal using the self-assessment mannequin (Bradley and Lang 1999). Words from ANEW were chosen if they had a valence rating less than or equal to 5 on a 1–9 scale and an arousal rating greater than or equal to 5 on a 1–9 scale (262 words). Duplicate words were removed. Further, when pairs or groups of words had the same stem but a different suffix, we selected one of the words using the following criteria. We retained (a) the more negative/arousing word if both had been rated in ANEW, (b) the word that had been rated in ANEW if only one was rated, and (c) the shorter word if neither was rated. Also, words or acronyms that were very specific to the military (e.g. IED, Falluja, APC, Kirkuk) were removed. This resulted in a total of 513 words. To help us identify random responders, we also included four highly positive words chosen from ANEW (champion, delight, joyful, lucky) and four fruit words (apple, orange, pear, tangerine).

Word Rating Scales

Participants were asked to rate how threatening each word was on a scale from 1 to 9 where 1 was "not threatening" and 9 was "very threatening." In the instructions, threat was defined as how much each word represents "the potential for harm, loss or damage for you". A subset of participants also rated the words on valence and arousal using the SAM rating scales (Bradley and Lang 1994), which are non-verbal, pictorial scales that have previously been used to rate affective stimuli (N = 143).

Although prior research has used the SAM rating scales to assess valence and arousal based on theories that emotion falls along these two dimensions (Carroll et al. 1999), we wanted to identify words that participants would identify as "threatening." For consistency with prior approaches to rating stimuli, for the first 143 participants, each word was rated for all three dimensions: valence, arousal, and threat. We then examined the relationship between our threat measure and measures of valence, arousal, and a composite of the two. Threat was strongly correlated with arousal ($r = .85$), with valence ($r = -.81$), and very strongly with the composite ($r = .96$). Thus, to reduce participant burden and because threat had greater face validity for our construct of interest, we proceeded with only the measure of threat for additional data collection.

Procedure

Data were collected using the Qualtrics Research Platform, and participants were provided a link to the Qualtrics survey through the MTurk system. Participants first reviewed a consent form and then completed the Trauma History Screen to determine eligibility. Eligible participants were then sent to the word rating portion of the study. They were first provided with instructions for how to rate the words. Then, words were presented one at a time, and participants provided ratings for each word. Participants who completed the survey received a code for entry into the MTurk interface, which would allow the researchers to verify completion of the survey and to provide payment.

Algorithm Development

The recommender system was developed using R (R Core Team 2013). The code and data are available at https://github.com/andreaniles/threat_word_recommender.

User–User Collaborative Filtering System Development

For a target user for whom recommendations are needed, the user–user collaborative filtering algorithm identifies similar users in a dataset and makes recommendations based on the item preferences of those similar users. Aggarwal (2016) describes the algorithm as follows (for the current recommender algorithm we developed, "items" are words). For an $m \times n$ rating matrix R = [r_{uj}] with m users and n items, let I_u represent the set of items rated by user u. If u has rated the 1st, 4th and 5th item, $I_\mathrm{u} = \{1,4,5\}$. The set of items rated by another user v is I_v, and the overlapping items rated by users u and v is indicated by $I_u \cap I_v$. Ratings by each user on this set of items are used to compute the similarity between users u and v, Sim(u,v). Pearson's correlation coefficient r is frequently the measure of similarity between users, and we have used r for the present recommender algorithm. Pearson's correlation is computed between each user in the dataset and the target user (the person for whom recommendations are needed) on the set of common items rated. The correlation coefficients, representing the similarity of each user in the dataset to the target user, are sorted from largest to smallest. A subsample of k similar users or neighbors with the highest r values are then selected, and their ratings are used to predict the ratings of items not rated by the target user. The word ratings of the k similar users are weighted by r so that the ratings of more similar users are weighted more heavily in the prediction of unrated items for the target user. Predicted ratings for the target user on each item are comprised of a weighted average of the ratings of the k similar users.

Optimizing the Model

To optimize the performance of the algorithm, various parameters can be modified or 'tuned' to achieve optimal model performance. Potential tuning parameters for recommender algorithms include the number of neighbors (k), what information to obtain from new users (termed the cold-start problem), and precisely how to calculate the similarity metric between users. The current study assessed multiple values for each tuning parameter and, for our final model, we selected values that resulted in the greatest model accuracy.

We used two methods to measure model accuracy. (1) We calculated the percentage of the mean of the actual top rated words that was achieved by the mean of the recommended top rated words for each participant. This is termed % max throughout the paper and is also referred to as "precision" in computer science. (2) We calculated the root mean squared error (RMSE) of the predicted word ratings for each person in the sample n. Calculation of these accuracy metrics was possible because all participants rated all words, and actual word ratings could be compared to the predicted word ratings based on the algorithm.

$$\%\mathrm{max} = \frac{\sum A}{\sum B} \times 100 \quad RMSE = \frac{\sum_{i=1}^{n}\sqrt{\left(\hat{y}_i - y_i\right)^2}}{n}$$

For RMSE, $\hat{y}_i$ represents the predicted rating for word i, y_i is the actual rating for word i, and n is the total number of words. For % max, A is the set of word ratings for words recommended to the user and B is the set of word ratings for the actual top rated words. For example, if we wanted to recommend 4 highly rated words to the user, and the recommender algorithm selected words with ratings 3, 4, 5 and 5 (out of 5), A would be {3, 4, 5, 5}. If the user had actually given four words ratings of 5 out of 5, B would be {5, 5, 5, 5}, and % max would be (3 + 4 + 5 + 5)/(5 + 5 + 5 + 5)*100 or 85%. For the current demonstration, we recommended 60 words.

Cold-Start Problem—The cold-start problem concerns the absence of data for a new user entering the recommender system. If the goal is to make recommendations to a new user, as was our goal for the current project, information must be obtained from that new user in order to identify similar users in the training dataset. In the "ask to rate" technique, new users are asked to rate items until sufficient information has been obtained to make a recommendation. Nadimi-Shahraki and Bahadorpour (2014) reviewed and compared different methods to select items for a new user to rate. To select the words rated by a new user from the pool of 513 words, we calculated a score that weighted the entropy of words (i.e. variability of ratings across the threat scale) and the average threat level of the words (i.e. popularity). This method is one of the most effective methods for making accurate recommendations (Nadimi-Shahraki and Bahadorpour 2014; Rashid et al. 2002), and has been termed a "balanced" method because it uses information about both the entropy and popularity of words. Entropy and popularity were calculated for each word as follows:

$$\text{Entropy} = \sum_{i=1}^{9} p_i \log\left(p_i\right) \quad \text{Popularity} = \sqrt{\sum_{i=1}^{n} y_i^2}$$

For entropy, p_i refers to the probability of rating i on the 1–9 rating scale. For popularity, y_i is the rating for user i, and n is the total sample of users. Although one approach (Nadimi-Shahraki and Bahadorpour 2014) is to determine a balance score by combining entropy and popularity using the equation log(popularity) × entropy, we experimented with different weighting methods to identify the optimal weighting of entropy and popularity for our dataset. We chose to use different amplification parameters α within the equation popularity$^{\alpha}$ * entropy in order to select the set of words a new user should rate to produce the most accurate word recommendation. We examined different values of α ranging from 0 to 3 in increments of .5.

We also needed to identify the number of words a new user would rate before we could make a recommendation. The more words a new user rates, the more accurate the recommendation will be. However, we needed to balance the new user burden with the accuracy achieved for different numbers of initial words rated. We determined that we did not want users to rate more than 70 words, and preferably fewer than 70. We then examined the accuracy of the recommender system based on the number of words provided ranging from 20 to 70 by increments of 5.

Neighbors—Neighbors are the number of similar users (k) whose ratings are averaged to predict ratings for a target user. We examined the accuracy of the system with 10 to 80 neighbors increasing by increments of 10.

Similarity Cutoff—When selecting the neighbors used to make a recommendation for a new user, the ratings of users whose ratings correlate most strongly with the new user are averaged. In addition to selecting the number of neighbors to be used in calculating the average rating, a minimum similarity value can be identified such that users with similarity values less than the cut off are not used for the rating. This cut off removes users who have very low or negative similarity values. We examined the accuracy of the model for different similarity cutoffs from 0 to .4 in increments of .05.

Optimization Order

Although ideally we would examine every combination of each of the tuning parameters described above to select the optimal model, this would result in a 7 by 11 by 8 by 9 dimensional array (a total of 5544 possible cells). Because each cell required approximately 1 h of computing time, the total computation time for this analysis would be 231 days. Instead, we examined the different values for each tuning parameter fixing all other tuning parameters at one value. We tuned the parameters in the order in which each entered into the recommender algorithm as follows: (1) amplification parameter, (2) number of words given, (3) number of neighbors, (4) similarity cut off. Once the optimal value was determined for each parameter, subsequent parameters were tuned at the optimal value of the previous parameter. Initial values for the number of words given, number of neighbors, and similarity cut off were 50, 30, and 0 respectively.

Statistical Analysis

Because accuracy values were nested within individuals (i.e. accuracy was calculated for multiple methods of word selection for each person in the dataset), we used a multi-level model (MLM) in Stata 14 (StataCorp 2015) to compare the accuracy of words selected using different methods. Our outcome was % max and the predictor was word selection method. We included random intercepts in the model. We could not compare methods on RMSE because the calculation of RMSE requires a “predicted” rating, which was not available for word sets other than those from the recommender algorithm. We compared the algorithm method to three word selection methods: (1) set of 60 words (randomly selected from the 94 words) from the MacLeod (2002) word set (coded as 0), (2) random selection of 60 words out of 513 words (coded as 1), and (3) trauma-group specific word sets (coded as 2). Significant omnibus tests were followed up with pairwise tests of simple effects using the “margins” command in Stata. For the trauma-specific word selection, we identified five studies that selected words for specific trauma groups including sexual abuse (Blix and Brennen 2011), physical abuse (Bunnell 2007), accident (Beck et al. 2001; Ehring and Ehlers 2011), and combat trauma (Swick 2009). For combat trauma, some words from the prior study were not included in the larger word set because they were very specific (as described in the “Word Selection” section above). Thus, for combat trauma, 26 out of the 40 total words from Swick (2009) were used for the comparison. For this analysis, we selected only participants reporting each type of trauma and calculated the mean rating given to the trauma-specific words. We then compared the accuracy of the trauma-specific words to the accuracy of the algorithm-selected words for people in each trauma group using paired samples t-tests.

Results

Selecting Model Parameters

Graphs showing algorithm performance based on % max and RMSE across different values of the tuning parameters are shown in Figs. 2 and 3 respectively. For some of the parameters, the optimal algorithm performance differed depending on whether we optimized based on % max or RMSE. However, because our goal was to select a set of the most threatening words rather than obtain the most accurate prediction for all words, we prioritized optimizing % max over RMSE. Final parameters selected are shown in Table 2. For the selected parameter values, % max was 91.4% and RMSE was 1.33.

Comparing Accuracy of Different Types of Word Selection

Algorithm Versus Random and Standard Words—Results are shown in Fig. 4. Using MLM, we examined the effect of word selection method on model accuracy (% max). The overall effect of word selection method was significant χ^2 (2, n = 837) = 9517.58. Tests of simple effects revealed that the algorithm-selected words (*M* = 91%, SD = 9) were significantly more accurate compared to the standard word set (*M* = 61%; SD = 14; *b* = .31, 95% CI .30, .32, $p < .001$, $d = 2.42$) and compared to a random word set ($M = 61\%$, SD = 12; $b = .31$, CI .30, .31, $p < .001$, $d = 2.92$). In contrast, the standard word set was not significantly more accurate compared to the random word set ($p = .797$).

Algorithm Versus Trauma–Specific Words

Results are shown in Fig. 4. For sexual trauma (n = 359), the accuracy for words selected by the algorithm ($M = 91\%$, $SD = 9$*)* was significantly higher than the accuracy for trauma-specific words ($M = 63\%$, $SD = 13$), t(358) = 45.5, $p < .001$, $d = 2.50$. For physical abuse (n = 446), the accuracy for words selected by the algorithm ($M = 91\%$, $SD = 9$) was significantly higher than the accuracy for trauma-specific words ($M = 62\%$, $SD = 14$), $t(445) = 47.7$, $p < .001$, $d = 2.39$. For accident-related trauma (n = 466), the accuracy for words selected by the algorithm ($M = 91\%$, $SD = 9$) was significantly higher than the accuracy for trauma-specific words ($M = 68\%$, $SD = .15$), $t(465) = 40.6$, $p < .001$, $d = 2.10$. Finally, for combat trauma (n = 69), the accuracy for words selected by the recommender system ($M = 92\%$, $SD = 7$) was significantly higher than the accuracy for trauma-specific words ($M = 70\%$, $SD = 12$), $t(68) = 20.4$, $p < .001$, $d = 2.92$.

Discussion

The goal of the current study was to develop and test an algorithm-based approach to personalizing negative affective word stimuli. Specifically, we used a collaborative-filtering recommender algorithm to personalize threatening words for trauma-exposed populations. Algorithm-based word selection identified words that were 50% more threatening compared to: (a) a random threatening word set, (b) a standard word set, and (c) words selected by researchers based on trauma type. Cohen's *d* effect sizes for these comparisons were large, ranging from 2.10 to 2.92, supporting the superiority of this system over non-personalized word selection. Further, the data collection necessary to train the recommender system was fast (approximately 200 participants in 24 h) and simple on MTurk, indicating that this approach could be applied rapidly and widely to train recommender systems for many types of affective stimuli.

To our knowledge, this study is the first to apply an algorithm-based machine learning approach to the selection of personalized stimuli, and it overcomes significant limitations of prior stimuli personalization efforts. Previous personalization approaches have required participants to rate all available stimuli, asked participants to spontaneously produce stimuli, used transcription of clinical interviews and subsequent stimuli selection by experts, or required limitation of research samples to very specific groups to allow tailoring of stimuli. The recommender algorithm we developed requires the user to rate only 55 words, and can then select a personalized word set of 60 words (although any number can be requested) from a pool of 513 words. This approach greatly increases the number of stimuli available without requiring significant effort and time on the part of the participant to rate all possible stimuli. Further, algorithm-based word selection does not require trained personnel to help the user select personalized words, thus allowing for an entirely automated process. Finally, the recommender system can be used to personalize words for individuals with any type of trauma exposure, eliminating the necessity to focus on a particular trauma group in order to select relevant stimuli.

Not only does the algorithm-based personalization approach resolve limitations of previous methods, but it also showed remarkable accuracy at identifying a highly personalized set of words. Algorithm-based word selection produced words that were much more threatening on average compared to a randomly selected word set and greatly outperformed a common set of words from MacLeod et al. (2002) that has been widely used (e.g. Amir et al. 2009; Amir and Taylor 2012; Colin et al. 2007; Hazen et al. 2009). Most remarkably, the set of words selected by the algorithm was far more threatening on average compared to words that were selected by researchers for specific trauma groups. This suggests that there is a large amount of variability in the types of stimuli that people find threatening, even among a relatively homogenous sample of people who have experienced the same type of trauma. Thus, when the same set of stimuli are used for every participant, even within a homogeneous population, the stimuli may not be personally relevant for a large portion of people. Thus, our ability to evoke emotional responses may be greatly improved by selecting personalized stimuli for each study participant, thereby reducing variability in participants' responses and increasing our ability to detect psychological and biological correlates of affective responding. The heterogeneity of trauma exposure in our sample is a strength of the current study because it demonstrates that the recommender algorithm approach can be used to make recommendations for individuals with a wide array of traumatic experiences. Thus, using a recommender algorithm to select affective stimuli can facilitate the study of individuals with PTSD resulting from many different types of trauma.

The current study has some limitations. The algorithm we developed is based on the ratings of people with trauma exposure. Thus, the present algorithm may not be directly applicable for selection of words for other populations such as those with anxiety disorders, obsessive–compulsive disorders, or depression. However, additional algorithms relevant to other patient populations and for other types of stimuli, such as images, can be easily developed. Further, a challenge of this type of data collection is random responding by participants, although research suggests that random responding is no more of an issue on MTurk than with convenience samples of college students (Fleischer et al. 2015). We used a number of methods to identify people who may have been responding randomly, but it is possible that some random responders were included in our training dataset. Future studies could use other indicators of random responding, such as reaction time. Another potential limitation is that participants may have provided inaccurate responses during the screening questionnaire to be eligible for the study. We had no way to detect such malingering. However, prior research indicates that an estimated 90% of people have experienced a traumatic event (Kilpatrick et al. 2013), which is higher than the percentage of screened participants who were eligible for the present study (76%). Thus, if some participants did malinger in an effort to be included in the study, it was likely a small number of subjects. Inclusion of these individuals in the training data could, however, reduce the accuracy of the recommender algorithm. Finally, the current collaborative filtering approach does not incorporate changes in ratings over time. Methods such as matrix factorization can model temporal dynamics more effectively than collaborative filtering (Koren et al. 2009). Thus, future research could incorporate longitudinal measurement of ratings and utilize matrix factorization in an effort to more accurately model temporal changes and make predictions of future ratings.

In conclusion, the current study proposes a novel algorithm-based approach to the personalization of evocative stimuli that can maximize emotional reactivity for every participant in a given study. We have demonstrated that algorithm-based personalization results in a set of words that is far more threatening to each individual trauma-exposed person compared to non-personalized word selection approaches. Further, this method resolves significant limitations to previous methods used to select personalized words. The recommender algorithm approach may be used to select personalized and highly evocative stimuli for lab-based or neuroimaging paradigms in affective and clinical science. The selection of personalized stimuli can maximize affective responding in study participants, potentially making psychological and biological sequelae of emotion easier to detect. Our study demonstrates just one example of the successful application of algorithm-based personalization for threat word selection. Artificial intelligence may hold tremendous potential for personalization of evocative stimuli in psychological science.

Acknowledgements

The current project was funded by the UCSF Digital Mental Health Resource Allocation Program. Dr. Niles is funded by the San Francisco VA Advanced Women's Health Fellowship and Dr. O'Donovan is funded by a K01 Award from the National Institute of Mental Health (K01MH109871) and a University of California Hellman Award. We also acknowledge Vivien Li for her significant contributions during the data collection phase of this project.

Funding Funding was provided by University of California, San Francisco (Grant No. 7710 133105 7500634 45), the San Francisco VA Advanced Women's Health Fellowship, a K01 Career Development Award (K01MH109871) and a University of California Hellman Award.

References

Aggarwal C (2016). Recommender systems: The textbook. Yorktown Heights, NY: Springer.

Amir N, Beard C, Burns M, & Bomyea J (2009). Attention modification program in individuals with generalized anxiety disorder. Journal of Abnormal Psychology, 118(1), 28–33. 10.1037/a0012589. [PubMed: 19222311]

Amir N, Kuckertz JM, & Strege MV (2016). A pilot study of an adaptive, idiographic, and multi-component attention bias modification program for social anxiety disorder. Cognitive Therapy and Research, 40(5), 661–671. 10.1007/s10608-016-9781-1. [PubMed: 27795598]

Amir N, & Taylor CT (2012). Combining computerized home-based treatments for generalized anxiety disorder: An attention modification program and cognitive behavioral therapy. Behavior Therapy, 43(3), 546–559. 10.1016/j.beth.2010.12.008. [PubMed: 22697443]

Beck JG, Freeman JB, Shipherd JC, Hamblen JL, & Lack-ner JM (2001). Specificity of Stroop interference in patients with pain and PTSD. Journal of Abnormal Psychology, 110(4), 536–543. [PubMed: 11727943]

Blix I, & Brennen T (2011). Intentional forgetting of emotional words after trauma: A study with victims of sexual assault. Frontiers in Psychology. 10.3389/fpsyg.2011.00235.

Bradley M, & Lang P (1994). Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25(1), 49–59. [PubMed: 7962581]

Bradley M, & Lang P (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Gainesville: University of Florida

Buhrmester M, Kwang T, & Gosling SD (2011). Amazon’s Mechanical Turk. Perspectives on Psychological Science, 6(1), 3–5. 10.1177/1745691610393980. [PubMed: 26162106]

Bunnell S (2007). The impact of abuse exposure on memory processes and attentional biases in a college-aged sample. Ann Arbor Dissertations & Theses.

Carlson EB, Smith SR, Palmieri PA, Dalenberg C, Ruzek JI, Kimerling R, … Spain DA (2011). Development and validation of a brief self-report measure of trauma exposure: The trauma history screen. Psychological Assessment, 23(2), 463–477. 10.1037/a0022294. [PubMed: 21517189]

Carroll JM, Yik MSM, Russell JA, & Barrett LF (1999). On the psychometric principles of affect. Review of General Psychology, 3(1), 14–22.

Chambers R, Lo BCY, & Allen NB (2008). The impact of intensive mindfulness training on attentional control, cognitive style, and affect. Cognitive Therapy and Research, 32(3), 303–322. 10.1007/s10608-007-9119-0.

Colin M, Lih YS, Elizabeth MR, & Lynlee WC (2007). Internet-delivered assessment and manipulation of anxiety-linked attentional bias: Validation of a free-access attentional probe software package. Behavior Research Methods, 39(3), 533–538. [PubMed: 17958165]

Ehring T, & Ehlers A (2011). Enhanced priming for trauma-related words predicts posttraumatic stress disorder. Journal of Abnormal Psychology, 120(1), 234–239. 10.1037/a0021080. [PubMed: 21058753]

Fleischer A, Mead AD, & Huang J (2015). Inattentive responding in MTurk and other online samples. Industrial and Organizational Psychology, 8(2), 196–202.10.1017/iop.2015.25.

Fridrici C, Leichsenring-Driessen C, Driessen M, Wingenfeld K, Kremer G, & Beblo T (2013). The individualized alcohol Stroop task: No attentional bias toward personalized stimuli in alcohol-dependents. Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors, 27(1), 62–70. 10.1037/a0029139. [PubMed: 22747499]

Hazen RA, Vasey MW, & Schmidt NB (2009). Attentional retraining: A randomized clinical trial for pathological worry. Journal of Psychiatric Research, 43(6), 627–633. 10.1016/j.jpsychires.2008.07.004. [PubMed: 18722627]

Kensinger EA, & Corkin S (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? Memory & Cognition, 31(8), 1169–1180. 10.3758/BF03195800. [PubMed: 15058678]

Kilpatrick DG, Resnick HS, Milanak ME, Miller MW, Keyes KM, & Friedman MJ (2013). National estimates of exposure to traumatic events and PTSD prevalence using DSM-IV and DSM-5 criteria. Journal of Traumatic Stress, 26(5), 537–547. 10.1002/jts.21848. [PubMed: 24151000]

Koren Y, Bell R, & Volinsky C (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37. 10.1109/MC.2009.263.

Kousta S-T, Vinson DP, & Vigliocco G (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. Cognition, 112(3), 473–481. 10.1016/j.cognition.2009.06.007. [PubMed: 19591976]

Lang PJ, Bradley MM, & Cuthbert BN (1999). International affective picture system (IAPS): Technical manual and affective ratings. Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

MacLeod C, Rutherford E, Campbell L, Ebsworthy G, & Holker L (2002). Selective attention and emotional vulnerability: Assessing the causal basis of their association through the experimental manipulation of attentional bias. The Journal of Abnormal Psychology, 111(1), 107–123. 10.1037/0021-843X.111.1.107. [PubMed: 11866165]

McGeary JE, Meadows SP, Amir N, & Gibb BE (2014). Computer-delivered, home-based, attentional retraining reduces drinking behavior in heavy drinkers. Psychology of Addictive Behaviors, 28(2), 559–562. 10.1037/a0036086. [PubMed: 24955674]

Nadimi-Shahraki M-H, & Bahadorpour M (2014). Cold-start problem in collaborative recommender systems: Efficient methods based on ask-to-rate technique. Journal of Computing and Information Technology, 22(2), 105–113. 10.2498/cit.1002223.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna: R Development Core Team

Rashid AM, Albert I, Cosley D, Lam SK, McNee SM, Kon-stan JA, & Riedl J (2002). Getting to know you: Learning new user preferences in recommender systems In Proceedings of the 7th international conference on intelligent user interfaces (pp. 127–134). New York, NY, USA: ACM 10.1145/502716.502737.

Roberts KE, Hart TA, & Eastwood JD (2010). Attentional biases to social and health threat words in individuals with and without high social anxiety or depression. Cognitive Therapy and Research, 34(4), 388–399. 10.1007/s10608-009-9245-y.

Schoorl M, Putman P, Mooren TM, Van Der Werff S, & Van Der Does W (2014). Attentional bias modification in Dutch veterans with posttraumatic stress disorder: A case series with a personalized treatment version. Journal of Traumatic Stress, 27(2), 240–243. 10.1002/jts.21896. [PubMed: 24700603]

Shao Z, Janse E, Visser K, & Meyer AS (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. Frontiers in Psychology. 10.3389/fpsyg.2014.00772.

Siegle GJ, Steinhauer SR, Thase ME, Stenger VA, & Carter CS (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biological Psychiatry, 51(9), 693–707. 10.1016/S0006-3223(02)01314-8. [PubMed: 11983183]

StataCorp. (2015). Stata statistical software: Release 14. College Station, TX: StataCorp LP.

Swick D (2009). The separate and cumulative effects of TBI and PTSD on cognitive function and emotional control (Annual Report No. W81XWH-08-2-0086). Martinez, CA: Veterans Health Administration.

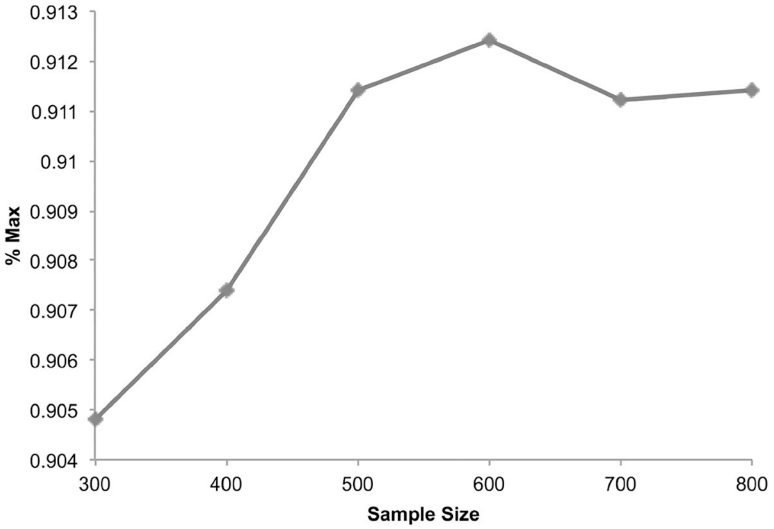

Fig. 1.

Model performance by sample size

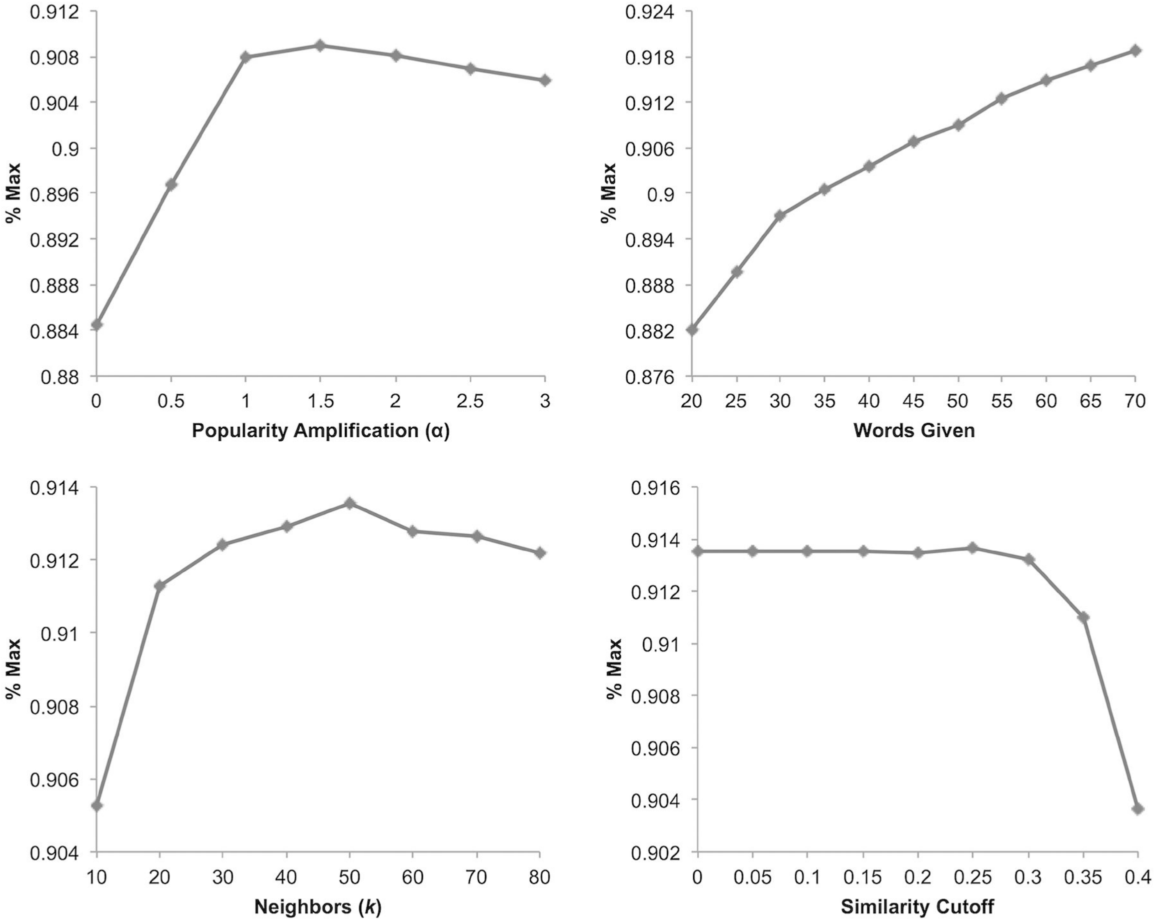

Fig. 2.

Model performance (% maximum rating) for different values of tuning parameters used to optimize the recommender algorithm. Higher values indicate better model performance. Amplification of popularity versus entropy (top left), words given to the recommender algorithm (top right), number of neighbors used to predict ratings (bottom left), and cut off for low similarity values (bottom right)

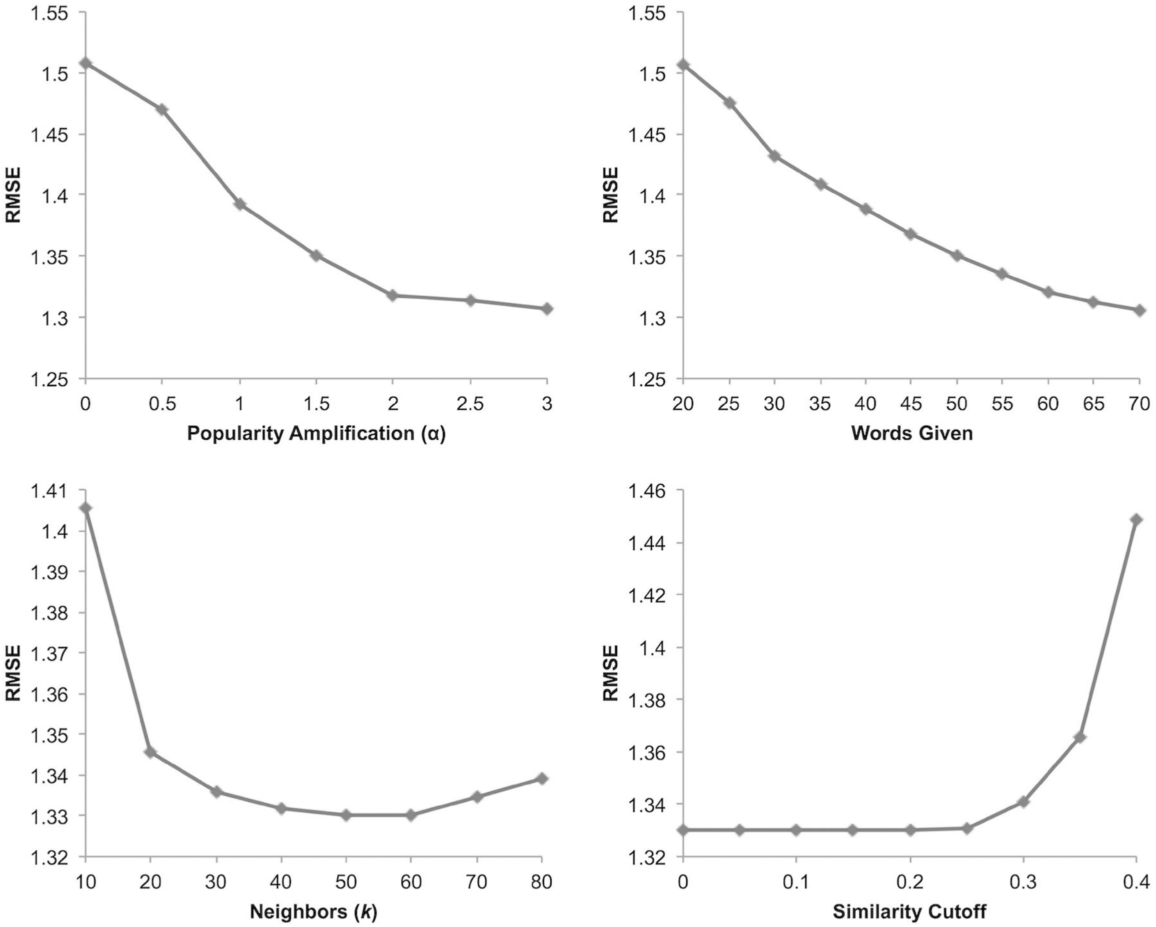

Fig. 3.

Model performance (root mean squared error) for different values of tuning parameters used to optimize the recommender algorithm. Lower values indicate better model performance. Amplification of popularity versus entropy (top left), words given to the recommender algorithm (top right), number of neighbors used to predict ratings (bottom left), and cut off for low similarity values (bottom right)

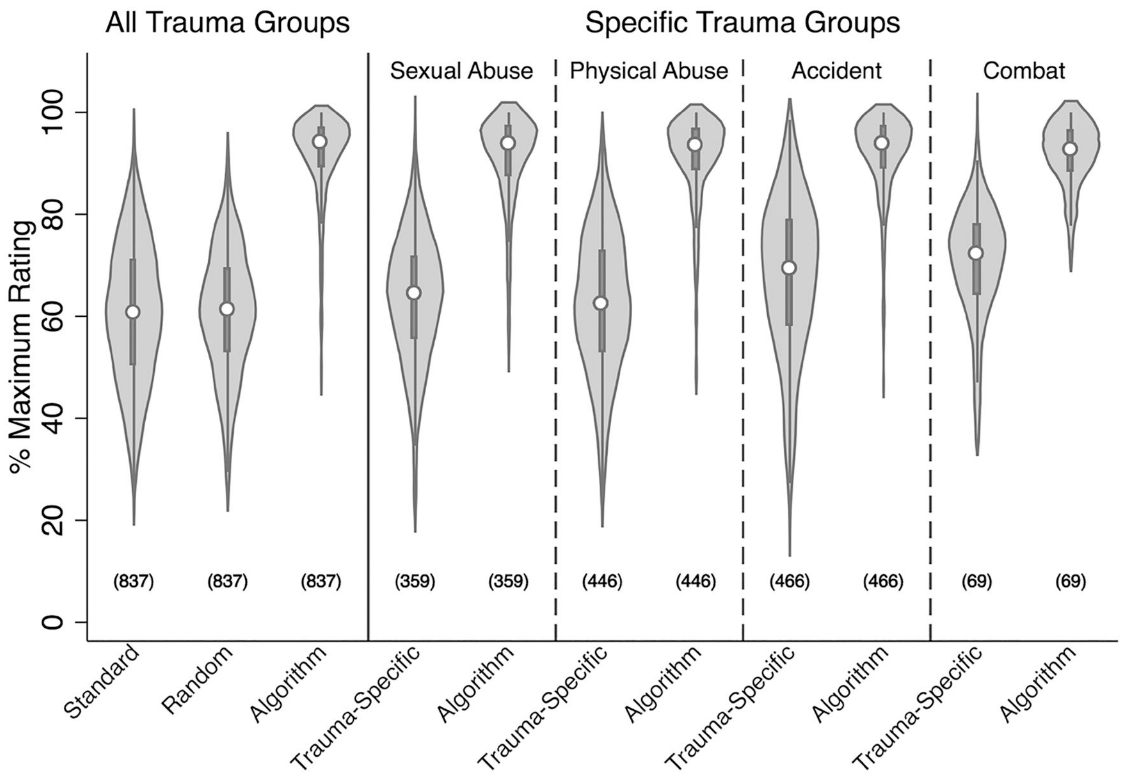

Fig. 4.

Violin plot of percent of maximum rating densities (i.e. accuracy) by method of word selection. Higher % maximum rating values indicate better accuracy. *Note* numbers shown in parentheses are the sample sizes included in the density plot

Table 1

Trauma type frequency by gender

Trauma type	Frequency (%)	
	Male	Female
Car, boat, train or airplane accident	155 (46.4)	203 (40.5)
Accident at work or home	125 (37.4)	151 (30.0)
Hurricane, flood, earthquake, tornado, or fire	178 (53.3)	256 (50.9)
Hit or kicked hard enough to injure (child)	164 (49.1)	182 (36.2)
Hit or kicked hard enough to injure (adult)	121 (36.2)	183 (36.4)
Forced or made to have sexual contact (child)	66 (19.8)	156 (31.0)
Forced or made to have sexual contact (adult)	60 (19.9)	185 (39.2)
Attacked with gun, knife or weapon	89 (26.7)	96 (19.1)
Seeing something horrible or being badly scared in military	33 (9.9)	36 (7.2)
Sudden death of close family or friend	274 (82.0)	401 (79.7)
Seeing someone die suddenly or get badly hurt or killed	160 (47.9)	222 (44.1)

Table 2

Final values selected for each tuning parameter in the recommender system

Parameter	Final value
Popularity amplification (α)	1.5
Words given	55
Neighbors (*k*)	50
Similarity cutoff	.25