**Spring break or heart break?**

Nicholas Harp, Catherine Brown, Nathan Petro, Maital Neta

# introduction

## Decision-making under uncertainty is ubiquituous in daily life (e.g., economics, politics, emotions), and a growing body of work has linked appraisals of emotional ambiguity in some social cues (e.g., surprised expressions predict both positive and negative outcomes) to mental health (i.e., depression, anxiety) and well-being (i.e., stress) outcomes.

**Emotinal ambiguity is frequently studied using facial expressions (surprise) and scenes, but reliance on these ambiguous cues overlooks the inherent ambiguity in language.**

**Ambiguity is a core feature of language, but previous work has failed to capitalize on the dual-valence ambiguity that characterizes some words in the English language (e.g., “break”).**

**Here we tested if appraisals of emotional ambiguity (valence bias) in faces and scenes extends to words by first identifying words with emotional ambiguity and then testing the relationship of appraisals among the different stimuli categories in a new sample.**

# Methods

## Participants

Workers on Amazon’s Mechanical Turk (MTurk) were invited to participate in an eligibility screener with the option to earn a bonus if they met the requirements and completed the entire study. The Workers clicked a hyperlink that directed them to the study. The screener task included demographic questions and one block of word ratings that included 5 instances of the word “negative” and 5 instances of the word “positive” (see Procedure below for full details). Workers were invited to complete the entire study if they indicated that they were over 18 years old, had English as their native language, had no history of psychological or neurological disorder, and correctly rated the words “positive” and “negative” as positive or negative with at least 80% accuracy. Of the 145 Workers who completed the screener, 119 met the eligibility requirements, and 103 (54.37% female, 45.63% male) chose to complete the entire study. The final sample was 3.88% Asian, 5.83% Black, 2.91% Hispanic or Latino, 85.44% White, and 0% Other, with a mean(sd) age of 37.16(10.60).

## Stimuli

We compiled an initial set of 59 words that we believed had two distinct definitions, one clearly positive definition and one clearly negative definition. To create lists of clearly positive and clearly negative words, we first created a master list of words that were included in both the study by Warriner, Kuperman, and Brysbaert (2013), for valence and arousal ratings, and the English Lexicon Project online word query (Balota et al., 2007), for lexical characterisic measurements. We then elimiated any words with a mean arousal rating that was greater than 1 standard deviation away from the mean arousal of the list of 59 ambiguous words. We classified “positive” words as those with a mean valence > 7 on the 1-9 scale used by Warriner et al. (2013); “negative” words had mean valence < 3. To ensure that all words shared similar lexical characteristics, we eliminated any words from the master list whose lexical characteristics did not fall within the minimum and maximum values of the 59 ambiguous words’ lexical characteristics. The following were used for the cutoffs: length, the frequency of a word as reported by the Hyperspace Analogue to Language (HAL) study (Lund & Burgess, 1996), the log of HAL frequency, number of phonemes, number of syllables, number of morphemes, lexical decision reaction time and accuracy, and naming reaction time and accuracy. The final list of pilot words included 59 ambiguous, 267 positive, and 304 negative words.

## Procedure

After giving informed consent, participants first answered demographic questions about their gender, age, race, native language, and whether they had ever been diagnosed with a psychological or neurological disorder. They were then shown a brief self-guided instructional walkthrough of the task before completing the screener.

Using a random seed, we selected 20 positive and 20 negative words from the final pilot list for use in the screener task. These 40 words, along with 5 instances of the word “positive” and 5 instances of the word “negative” were presented randomly, one at a time, each following a 250 ms fixation cross. Each word remained on screen until the participant indicated that they thought it was positive or negative by pressing A or L on their keyboard (key pairing randomized across participants). If no response was made after 2000ms, a reminder appeared on screen, “Please respond as quickly as you can! A = POSITIVE. L = NEGATIVE.” Participants who rated the words “positive” and “negative” with less than 80% accuracy were compensated for their time but were not invited to complete the rest of the study. Participants were also excluded at this point if they indicated that they were younger than 18, that English was not their native language, or that they had been diagnosed with a psychological or neurological disorder.

The remaining 590 words from the final pilot list were randomly presented across 10 blocks of 59 words using the same button-press procedure as the screener block.

All tasks were created and presented using Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2019). The study was only accessible to participants using a computer (not a phone or tablet) within the United States.

## Data Analysis

All of the calculations described in this section were scripted using R version and are available in the **Supplementary Information**.

# Results

Trials with a response time faster than 250ms were removed from the data prior to analysis, as well as trials with a reaction time greater than 3 SDs above the mean reaction time averaged across all trials.

We assessed average reaction time to identify the ambiguous words within the range of 35%-65% average negative rating, suggesting low response consensus.

Previous work has shown that ambiguous faces and images are associated with longer reaction times in a forced-choice valence classification task (Neta & Whalen, 2013). **Figure 1a** shows that 29 amibugous, 5 negative, and 6 positive words surpassed a reaction time threshold of 875ms (Why did we use 875? Just visual inspection?).

![A screenshot of a cell phone

Description automatically generated]()![A picture containing bird, flock, group, large

Description automatically generated]()

These 40 words were considered for inclusion in a final list of ambiguous words. We removed 7 words that did not have both a clearly positive and clearly negative definition (“recession”, “faceless”, “headstone”, “inherit”, “abundant”, “cosmic”, “receive”), as well as 1 word that was redundant to another ambiguous word that we included (“courtroom”), resulting in a final list of 32 ambiguous words.

As shown in **Figure 1b**, visual inspection of the average valence ratings revealed two distinct groups of words with high response consensus: one with a clearly negative meaning (n = 18, mean valence rating > 75% negative) and one with a clearly positive meaning (n = 20, mean valence rating < 10% negative). We removed the words “positive” and “negative” from each list (explain). Because the valence bias task requires an equal number of ambiguous (50%) and clearly-valenced (25% positive, 25% negative) stimuli, we included the 16 words with the fastest reaction time for the positive and negative word lists, respectively.

## STUDY 1 DISCUSSION Many of the words showed variability in the percent negative categorizations across people, but analysis of additional features of ambiguity (slower reaction times) revealed a subset of words (clearly positive, clearly negative, and ambiguous) to test in the follow up study.

**METHODS**

**Participants**

Amazon MTurk Workers were again invited to participate in the study through a hyperlink. After completing the same eligibility screener used in Study 1, eligible participants proceeded to complete a valence bias task, described below. XX of the XX participants were eligible to participate, and XX chose to complete the study (XX male). The final sample was 6.09% Asian, 9.14% Black, 5.58% Hispanic or Latino, 76.65% White, and 2.54% Other, and included a wide range of ages (18-76).

**Stimuli**

Three task blocks (faces, scenes, and words) were used to assess valence bias. As in previous work (Neta, Norris, & Whalen, 2009), the face and scene task blocks included 24 ambiguous images, 12 positive images, and 12 negative images. The facial expressions were selected from the NimStim (Tottenham et al., 2011) and Karolinska Directed Emotional Faces (Lundqvist, Flykt, & Öhman, 1998) sets, and the scenes were selected from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2008). For the words block, the 32 ambiguous, 16 positive, and 16 negative words identified in Study 1 were used. All words were presented in all capital letters in plain black font on a white background.

**Procedure**

Participants were randomly assigned to pseudorandom presentation orders of the faces, scenes, and words blocks. Within each block, all stimuli were preceded by a 2000 ms fixation cross and then presented for XX ms. If participants did not make a response within 2000 ms, no response was recorded and the task advanced to the next trial. Participants responded by pressing either the “A” or “L” key on their keyboard, with the positive and negative response keys counterbalanced across participants. Valence bias for each stimulus category was calculated as the percent of negative responses for the ambiguous stimuli. As in Study 1, the task was administered using Gorilla Experiment Builder (Anwyl-Irvine et al., 2019), and was only accessible to participants in the United States through a computer.

**Data analysis**

Preregistration is available at the Open Science Framework website (osf.io/LINK). All data cleaning, analyses, and visualizations were completed using R (Version 3.6.0; R Core Team, 2017). Packages needed to reproduce analyses include (a bunch… list here). Prior to calculating our measure of valence bias (i.e., percent negativity for ambiguous stimuli), trials with reaction times less than 250 ms were removed (*note that this was set to 200 ms in the data cleaning script. I’ve since updated to 250, but our previous results were likely at the 200 ms cutoff*), as in Study 1. Additionally, only participants’ first response during each stimulus presentation was retained for analysis, and participants that failed to respond to 75% or more of the trials or did not correctly rate the clearly valenced stimuli greater than 60% of the time (n = 6) were removed prior to the statistical analyses. After, we calculated the proportion of trials in which each stimulus category was categorized as negative to measure valence bias. Linear mixed effects models were used to test for differences in percent negativity and reaction time across the stimulus and valence categories. Partial correlations were used to assess whether valence bias towards the ambiguous words was related to that of the faces and scenes, while controlling for gender and age. Where applicable, non-parametric tests (Spearman’s correlations) were used for data failing to meet normality assumptions.

**RESULTS**

**Manipulation check**

Confirming our prediction that response consensus for the ambiguous words would be lower than that of the clearly valenced (positive and negative) words, a paired sample t-test showed that standard deviations of ambiguous words (M = .45 SD = .13) were higher than those of the clearly valenced words (M = .11 SD = .20; *t*(31) = 8.90, *p* < .001).

**Stimulus Effects on Percent Negativity and Reaction Time**

In a repeated measures ANOVA, there was a significant main effect of Valence on percent negativity for positive (*M* = 5.38%), negative (*M* = 89.63%), and ambiguous (*M* = 48.33%) images (*F*(2, 1,576.00) = 6,113.69, *p* = 0), such that negative images were rated more negatively than both positive (*t*(1,584.04) = 110.29, *p* = 0) and ambiguous images (*t*(1,584.04) = -54.07, *p* = 0), and ambiguous images were more negative than positive images (*t*(1,584.04) = 56.22, *p* = 0). Additionally, there was a significant main effect of Stimulus on percent negativity for the faces (*M* = 49.74%), scenes (*M* = 46.14%), and words (*M* = 47.45%; (*F*(2, 1,576.00) = 11.41, *p* = 0.00), such that faces were rated more negatively than scenes *t*(1,584.04) = 4.71, *p* = 0.00) and words (*M* = 49.74%) and words (*t*(1,584.04) = 3.00, *p* = 0.00), but words were not significantly different from scenes (*t*(1,584.04) = -1.71, *p* = 0.09). These main effects were qualified by a significant interaction of Valence x Stimulus (*F*(4, 1,576.00) = 6,113.69, *p* = 0.00), such that negative images were rated as more negative than both positive and ambiguous images in all three stimulus categories (all *p’s* < .001), but there were also differences across stimulus categories within each valence condition. Specifically, negative words were rated more negatively than both faces (*t*(1,584.04) = -3.57, *p* = 0.00; Bonferroni corrected significance for these analyses p < 0.00) and scenes (*t*(1,584.04) = -5.90, *p* = 0.00), but faces and scenes did not differ after correcting for multiple comparisons (*t*(1,584.04) = 2.33, *p* = 0.02). Further, ambiguous faces were rated more negatively than both scenes (*t*(1,584.04) = 5.35, *p* = 0.00) and words (*t*(1,584.04) = 7.40, *p* = 0.00), but scenes and words did not differ after correcting for multiple comparisons (*t*(1,584.04) = 2.05, *p* = 0.04). There were no significant differences in negativity across stimulus categories for positively valenced stimuli (all *p’s* > .172).

In the same model (Stimulus x Valence) with reaction time (RT) as the dependent variable, there were significant effects of Stimulus (F(2, 1576) = 296.13, p < .001), Valence (F(2, 1576) = 218.21, p < .001), and their interaction (F(4, 1576) = 7.01, p < .001). Comparing the effects of Valence revealed that RTs were slower for ambiguous stimuli (M = 781.11, S.E. = 9.37) than positive (M = 682.64, S.E. = 9.37; t(1584) = 23.65, p < .001) and negative stimuli (M = 712.23, S.E. = 9.37; t(1584) = 16.55, p < .001). Additionally, RTs for negative stimuli were slower than those for positive stimuli (t(1584) = 7.11, p < .001). Comparing the effects of Stimulus revealed that RTs for faces (M = 678.52, S.E. = 9.37) were faster than both scenes (M – 733.32, S.E. = 9.37; t(1584) = -13.16, p < .001) and words (M = 764, S. E. = 9.37; t(1584) = -20.57, p < .001), and that RTs for scenes were faster than for words (t(1584) = -7.41, p < .001). Finally, a Stimulus x Valence interaction revealed that RTs for ambiguous faces (M = 725.27, S.E. = 10.26) were faster than RTs for ambiguous scenes (M = 780.82, S.E. = 10.26; t(1584) = -7.70, p < .001) and ambiguous words (M = 837.26, S.E. = 10.26; t(1584) = -15.53, p < .001), and RTs for ambiguous scenes were faster than ambiguous words (t(1584) = -7.83, p < .001). Among the negatively valenced stimuli, faces (M = 666.79, S.E. = 10.26) were faster than scenes (M = 728.13, S.E. = 10.26; t(1584) = -8.51, p < .001) and words (M = 741.77, S.E. = 10.26; t(1584) = -10.40, p < .001), but scenes and words did not differ (t(1584) = -1.89, p = .059). Among the positively valenced stimuli, RTs for faces (M = 643.50, S.E. = 10.26) were faster than both scenes (M = 691.02, S.E. = 10.26; t(1584) = -6.59, p < .001) and words (M = 713.42, S.E. = 10.26; t(1584) = -9.70, p < .001), and RTs for scenes were faster than words (t(1584) = -3.11, p = .002; *maybe not after correction*). Within each stimulus category, ambiguous stimuli were rated slower than both positive and negative stimuli (all p’s < .001) and negative stimuli were rated slower than positive stimuli (all p’s < .001).

**Relationships among the stimulus categories**

Next, we tested for relationships among valence bias for faces, scenes, and words within participants while controlling for age and gender. Replicating previous work (Neta, Kelley, & Whalen, 2013), there was a positive relationship between categorizations of ambiguous stimuli for the faces and scenes (*rs* = .34, *p* < .001), faces and words (*rs* = .26, *p* < .001), and scenes and words (*r*(195) = .44, *p* < .001) while controlling for both age and gender (Figure 2a-c). **![A close up of a map

Description automatically generated]()**

**Exploratory analysis of word characteristics**

In exploratory analyses, we assessed differences between the ambiguous and clearly valenced words on linguistic characteristics which might contribute to emotional ambiguity. For instance, comparing frequency (HAL; Lund & Burgess, 1996) between the clear and ambiguous words revealed that the ambiguous words are used more frequently (*t*(58) = 2.08, p = .042).

**GENERAL DISCUSSION**

**Here we identified ambiguous and clearly valenced words, showed that valence bias measured with face and scene stimuli extends to words, and explored characteristics of ambiguous words.**

**The development and validation of this new stimulus set provides a novel method for measuring valence bias, and shows that valence bias generalizes to words with valid positive and negative interpretations.**

**This task offers leverage beyond that of many other “ambiguity” resolution tasks in the language literature, which often rely on resolution of neutral-positive or neutral-negative rather than dual valence (positive-negative) ambiguity.**

**One limitation of this work is that the words are presented in isolation (to prevent contextual clues from a sentence disambiguating the words), and future work should aim to expand upon these findings by assessing appraisals of sentences with dual valence emotional ambiguity.**

**Altogether, this work builds on a growing body of work aiming to understand individual differences in appraisals of emotional ambiguity, and opens the door for future research to capitalize on this new class of emotionally ambiguous stimuli.**

# Key….Leave in place

# Heading 1

## Heading 2

### Heading 3

#### Heading 4

##### Heading 5

###### Heading 6

Body text