



ARTICLE



<https://doi.org/10.1057/s41599-022-01126-3>


OPEN

Gender differences in emotional connotative meaning of words measured by Osgood's semantic differential techniques in young adults

Robert M. Chapman¹, Margaret N. Gardner¹ & Megan Lyons¹

Semantic differential techniques are a useful, well-validated tool to assess affective processing of stimuli and determine how that processing is impacted by various demographic factors, such as gender. In this paper, we explore differences in connotative word processing between men and women as measured by Osgood's semantic differential and what those differences imply about affective processing in the two genders. We recruited 94 young participants (47 men, 47 women, ages 18–39) using an online survey and collected their affective ratings of 120 words on three rating tasks: Evaluation (E), Potency (P), and Activity (A). With these data, we explored the theoretical and mathematical overlap between Osgood's affective meaning factor structure and other models of emotional processing commonly used in gender analyses. We then used Osgood's three-dimensional structure to assess gender-related differences in three affective classes of words (words with connotation that is Positive, Neutral, or Negative for each task) and found that there was no significant difference between the genders when rating Positive words and Neutral words on each of the three rating tasks. However, young women consistently rated Negative words more negatively than young men did on all three of the independent dimensions. This confirms the importance of taking gender effects into account when measuring emotional processing. Our results further indicate there may be differences between Osgood's structure and other models of affective processing that should be further explored.

¹Department of Brain and Cognitive Sciences and Center for Visual Science at the University of Rochester, Rochester, NY, USA.

email: rmc@cvs.rochester.edu

Introduction

Word meaning is often studied as denotative meaning, which is a word's specific, direct meaning as it functions in a language to produce expository information. Less analyzed is the connotative or affective (emotional) meaning of words. One of the useful aspects of examining the affective meaning of words is that most individuals already possess common understanding of the emotional aspects of words because both connotative and denotative comprehension is acquired through normal language learning. More importantly, there is a well-formed theory of affective word meaning with a strong quantitative basis that is measured by connotative differential ratings of words on semantic scales defined by adjective-pairs of antonyms, such as "good–bad", "strong–weak", and "fast–slow". This theory is based on the behavioral ratings and multivariate (principal components analysis) measurement methods of Osgood (Chapman et al., 1980; Osgood, 1969a; Osgood et al., 1957). Differential affective scores lend empirically validated methods of quantifying something that is normally subjective and difficult to discern: the processing of emotions and emotional stimuli. This in turn has widespread research utility, from better modeling of emotions to increasing understanding of how demographic differences interact with affective processing to detecting changes in emotions due to neuropsychiatric and neurodegenerative diseases.

The core of Osgood's semantic differential technique lies in its definition of the affective meaning of words along three underlying dimensions, which are named Evaluation, Potency, and Activity (E, P, A). Osgood developed these dimensions by having subjects rate words using a wide variety of adjective pairs and then performing factor analysis to derive this dataset's underlying structure. The Evaluation factor is comprised of adjective pairs that describe the word's value, such as "good–bad" and "happy–unhappy". The Potency factor had loadings for adjective pairs such as "strong–weak" and "dominant–submissive". Finally, the Activity factor was defined by adjective pairs such as "fast–slow" and "excited–calm" and can be thought of as how active or intense the word is. With this model, each word then can be placed within this three-dimensional space via its mathematical relationship with each of the three dimensions (its factor loadings). For example, the word "coward" might have rating scores of -0.5 on Evaluation, -0.7 on Potency, and $+0.2$ on Activity, which quantitatively captures the affective meanings: "Quite Bad", "Very Weak", and "Slightly Active" (Osgood and McGuigan, 1973). Such a measurement becomes a quantitative index of the emotional aspects of this word that allows comparisons among the words and among the people rating them in an experimental task.

Since Osgood's seminal publication on factor analysis and the measurement of meaning, a great deal of work has been done concerning the emotional processing of various stimuli. Mehrabian and Russell (1974) developed a similar model of semantic differentials using the emotional context of pictures (rather than words). Like Osgood, they asked participants to rate the emotional aspects of pictures using adjective pairs. Also, like Osgood, they derived three factors, which they termed Pleasure (sometimes also termed Valence), Arousal, and Dominance. These factors by interpretation are not dissimilar from those defined by Osgood (Bakker et al., 2014; Russell, 1980; Sereno et al., 2015). The Evaluation/Pleasure dimensions is seen as describing the general positivity or negativity of the word's emotional characteristics. The Potency/Dominance factor has been described as the strength or power of the word's emotional meaning. There is significant dissonance in the interpretation of this factor (Bakker et al., 2014). Finally, again with some disagreement, Activity/Arousal often refers to the activity (physical, emotional, or

mental) caused by or implied by the word's meaning. This is not surprising; in both Osgood's and Mehrabian and Russell's work with factor analysis, the Evaluation/Valence factor has tended to represent the most variance in the affective rating dataset, with the other two factors accounting for less and sometimes switching their factor order in the solution. Still, these three measures of emotion in language are relatively constant and universal (Jackson et al., 2019; Majid, 2019). The affective three-dimensional space has been replicated across many languages and cultures and across different stimulus modalities and demographic groups (Mukherjee and Heise, 2017; Osgood, 1980; Osgood et al., 1975; Skrandies, 2014; Skrandies and Chiu, 2003), leading to the conclusion that "without a single exception, E, P, and A have appeared as dominant factors" in young adults (Osgood and McGuigan, 1973).

Therefore, despite some disagreements concerning the interpretation of these factors, they have been extensively employed in psychological and linguistics research, and there exist multiple different lexicons presenting affective ratings for words. One of the commonly employed is the ANEW lexicon (Bradley and Lang, 1999), which presents word ratings along the Valence, Dominance, and Arousal dimensions. These ratings have provided a basis for multiple investigations into demographic differences in affective processing. If the dimensional structure is constant across demographic factors, such as gender, how might the words' locations within that space be influenced by those factors? This is a useful question, as the word-space itself becomes a common metric by which to measure changes in affective processing due to demographic variability or other conditions, such as mood or language disorders. Some work suggests Osgood's semantic differential word-space is sensitive to gender differences (Hall and Matsumoto, 2004; King, 2001; MacKinnon and Keating, 1989; Skrandies, 2014). Additionally, there is a body of evidence that similar findings concerning gender disparities have been discovered when using the ANEW and other Valence/Arousal/Dominance based lexicons (Soares et al., 2012; Söderholm et al., 2013; Warriner et al., 2013). Specifically, women may tend to produce more extreme ratings of words than men. This can have implications for the study of affective processing in general, as well as the treatment of multiple neuropsychiatric and even neurodegenerative disorders.

However, in order for findings such as these to be useful, they must be generalizable and validated. As some have moved away from Osgood's original dimensional structure, some have postulated that only two of the three dimensions (Valence and Arousal) are truly meaningful in discussing demographic differences (Bradley and Lang, 1999; Söderholm et al., 2013), ignoring the Potency/Dominance dimension. Also, Mehrabian and Russell's factor analysis used an oblique rotation, rather than the orthogonal rotation Osgood originally employed. While this may be closer to the reality of a complex situation, it can lead to complications in interpreting that dimensional space, as no two dimensions are independent. If the dimensions are correlated, the stimulus ratings along those dimensions may also be, which would require a more sophisticated, multivariate approach to tease out group differences.

In this article, we examine what gender differences, if any, exist in the semantic word ratings and what those gender differences may imply about affective processing in young men and women. While there is a large body of research on this topic using Mehrabian and Russell's semantic differential factors, Osgood's semantic differential factors have been less explored. Naturally, this question leads into how well Osgood's original semantic structure truly relates to the semantic differential techniques proposed by Mehrabian and Russell, which we will also investigate. Finally, with

careful selections of stimulus words to reduce correlations among the stimuli and multivariate methodologies, we aim to determine if the Potency/Dominance dimension can add useful information into the discussion of gender differences between young men and young women.

Methods

Materials

Stimuli. A list of the 120 words used in this study, including Osgood's and Heise's original factor loadings as well as our own experimental ratings, appears in the Supplementary Materials. The word frequency, or the commonality of the usage of the word in everyday language, was measured for each of the 120 words using the Corpus of Contemporary American English (Davies, 2009, 2010, 2021).

Survey. Our web-based survey consisted of three sections presented in this order: (1) questions concerning the participant's demographical information, (2) the Connotative Meaning paradigm described below, and (3) short surveys to assess cognition and mood. These surveys were derived from the PROMIS bank of surveys to assess anger, depression, anxiety, general life satisfaction, and positive affect (PROMIS Bank v1.1). The PROMIS Bank v2.0 cognitive function survey was used to assess general cognition (Northwestern University, 2021).

The survey was operated with a Qualtrics (Qualtrics, 2020) backend and Amazon's Mechanical Turk (MTurk) (Amazon Mechanical Turk) service as a recruitment method. MTurk provides an internet-based, crowd-sourcing platform conducive to psychological research; Ratcliff and Henderson (Ratcliff and Hendrickson, 2021) found data obtained through MTurk matched carefully controlled data obtained from similar in-person methods. In this present study, control questions (consisting of a clear direction for the participant to press a particular answer) were periodically and randomly included to ensure participants were paying attention to the survey questions rather than "clicking through".

Participants. To complete the survey, participants needed to have MTurk accounts, speak English as a primary language, and be 18 years or older. Participants also had to have access to an internet connection and a computing device (PC, tablet, or smart phone). We gathered survey results from a total of 300 human adults after immediately discarding those participants who (1) failed to answer the control questions, (2) completed the survey much more quickly than their peers (average completion time was 38.7 min), or (3) gave the same response on every word. To create a baseline of gender effects in a wide variety of young adults, we selected individuals from the ages of 18–39. This range was chosen because it encompasses early adulthood before many of the biological and maturational shifts of middle life have occurred without too narrowly confining the age-range to limited socio-economic groups, such as only college students. Selecting this age group resulted in 193 participants, which we then divided into gender groups through their self-reported gender. This resulted in 74 young women and 119 young men. We also discarded individuals who self-reported they have been diagnosed with a language or learning disability as it is not clear how a language disorder (such as autism spectrum disorder) may impact the connotative processing of words. Removing participants with these conditions reduced the likelihood that these covariates could influence the results, given there is generally a higher prevalence of language and learning disabilities in boys (Zablotsky et al., 2019). 21 young men and 12 young women were discarded in this step. In addition, because mood disorders may

Table 1 Demographical information of the young participants used in the connotative meaning analyses.

	Male (n = 47)	Female (n = 47)
Mean age in years (SD)	26.51 (3.68)	31.42 (5.09)
Mean years of education (SD)	15.70 (2.03)	15.30 (1.79)
Number of subjects with exposure to other languages	11 (23%)	8 (17%)

Participants were anonymously recruited using Amazon's Mechanical Turk crowdsourcing service. Gender group assignments were based on participant self-report.

also influence affective processing of stimuli, we discarded individuals whose scores on the PROMIS Anxiety or Depression surveys indicated moderate to severe anxiety or depression. If scores on these surveys were not available, we discarded individuals based on a self-report of a mood disorder diagnosis. An additional 18 young men and 15 young women were discarded due to the occurrence of anxiety and/or depression.

Finally, to keep the gender groups balanced, we reduced the size of the larger male group by randomly discarding male participants until our groups of male and female respondents totaled 47 individuals each, resulting in a total sample size of 94. The demographical information for these groups appears in Table 1. Despite efforts to age-match, there was a significant age difference between the groups ($t(1, 91) = 5.36, p < 0.0001$) such that the young men were roughly 5 years younger than the young women on average. There was no significant difference in years of education ($t(1, 91) = -1.02, p = 0.15$). Some individuals in both groups self-reported proficiency in a second and sometimes a third language (8 in the female group and 11 in the male group). This is not statistically different ($\chi^2 = 0.59, p = 0.44$).

Procedures

Connotative meaning paradigm. Using Osgood's and Heise's work (Chapman, 1978a; Heise, 1971; Osgood et al., 1975) on semantic analysis and dimensionality as a basis, we designed a paradigm where participants evaluate words based on three semantic scales: Evaluation (E), Potency (P), and Activity (A) (Chapman et al., 1980). Each trial contained a single word presented on a computer screen and a 7-point rating scale corresponding to one of the three rating Tasks (E, P, or A). The participant was asked to rate the meaning of this word within the supplied adjective pair and do so by clicking or tapping an answer on the screen. For the E Task, for example, the adjective pair was Good–Bad, so the participant was presented with the scale:

- Extremely GOOD
- Quite GOOD
- Slightly GOOD
- Neither GOOD nor BAD
- Slightly BAD
- Quite BAD
- Extremely BAD

For the P Task, the adjective pair was Strong–Weak, and for the A Task, the adjective pair was Fast–Slow. These adjective pairs were selected based on Osgood's initial work with factor analyses which empirically linked each of these adjective pairs to the underlying dimension (Osgood et al., 1957). The scale is represented numerically from -3 to $+3$ with 0 equating to neutral.

When using these measures to assess differences in rating patterns between groups, it can be helpful to identify words that are more purely representative of one of the three dimensions. Doing so can ease interpretations later, as well as reduce

Words in the Three-Dimensional Evaluation (E), Potency (P), and Activity (A) Word Space

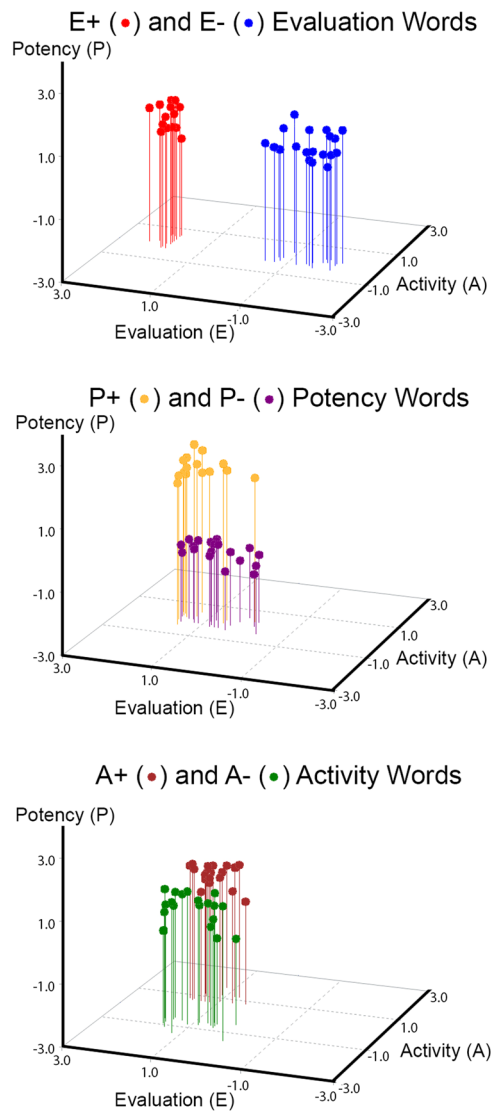


Fig. 1 Osgood's Evaluation, Potency, and Activity (E, P, A) word-space, based on a list of 1551 words compiled by Heise (1971) and Snider and Osgood (1969). Each word exists in this space based on its relationship to these three (E, P, and A) orthogonal dimensions. We selected 120 words that had a rating score strongly either positively or negatively on one dimension while remaining relatively neutral on the other two. This led to the six Word-Classes (20 words each) plotted here.

covariance effects among the dimensions. This can be tricky, as this is a three-dimensional word space, cubic and words are unlikely to be located precisely along one of the dimensions. To accomplish this, we used factor loadings derived from young adults that were supplied by Osgood (Osgood et al., 1975) and compiled by Heise (1965, 1971). Of the 1551 words, we selected 120 words that, on average, would produce an “extreme” (positive or negative) rating on only one of the three dimensions (Chapman et al., 1977) (Supplemental Table 1). This leads to an assignment of an affective Word-Class for each word as E+, E-, P+, P-, A+, or A-, depending on the directionality of its extreme rating. Averaging by Word-Class can therefore yield “extreme” measures of that class where each class has only a directionally large rating on one dimension while its ratings on

the other two dimensions are nearer zero. This also produces average “neutral” words; A+ words should, on average, be rated roughly neutral on the E dimension, for example. We term this property of the Word-Class its Word Relevance (Positive, Neutral, Negative).

We presented 120 trials of single words, 20 in each of the six Word-Classes, for each rating Task. All 120 words were the same for each rating Task but presented in randomized order. Given the randomization per Task, the large number of words in each Task, and the long time between presentations of the same word (~10 min per Task), word repetition effects should be minimal. Each participant completed the E Task first, the P Task second, and the A Task third. This Task order was fixed because the E and P scales account for the most variance in Osgood's factor structure (Jackson et al., 2019; Kissler et al., 2006; Osgood et al., 1957), increasing the chance of the more important data being collected before participant fatigue might occur. Using the same 120 words in different random orders for each task allowed each Word-Class to be rated along its dimension in addition to the two other dimensions, which then permitted us to study separately the effects of the stimulus Word-Class and the rating Task on the behavioral responses.

Statistical analysis. Both univariate and multivariate methods were used in this study of Osgood's semantic differential structure and gender differences. All data were analyzed in SAS 9.4 (SAS Institute Inc., 2017) with the MEANS and SCATTER. For comparisons to other lexicons of affective meaning word ratings, we used the CORR and FACTOR procedures. We performed mixed linear effects analyses with the MIXED procedure and post-hoc mean comparisons with the PLM procedure.

See the Supplementary Materials for more information on the statistical procedures used in this paper.

Results

Verification of Osgood's dimensions. To assess how well the Osgood structure appeared in our data, we first compared word ratings from our participants to the ratings listed for each word as derived and compiled by Osgood (Snider and Osgood, 1969) and Heise (1971), where gender was not separated. Each of the 120 words is plotted in Fig. 1 by its Word-Class. The three-dimensional affective-meaning word-space along the Evaluation (E), Potency (P), and Activity (A) dimensions can be visualized in the figure, and each word's ratings along those three dimensions is symbolized by its position within this cubic space. For clarity's sake, each Word-Class pair (e.g., E+ and E- words, which are words with extreme high or low ratings on the Evaluation scale) is separately plotted. Figure 1 clearly depicts the positive/negative polarity of these words along one dimension; their noisy distribution along the other dimensions in this figure is reduced later through averaging in our analyses, leaving a cleaner distinction between words that are strongly positive or strongly negative on one dimension and limiting the influence of the other two dimensions (Supplemental Fig. 1).

Additionally, we found strong correlations between our own data and the Word-Class constructs within Osgood's connotative dimensionality. We averaged the ratings for the 20 words in each of six Word-Classes and three rating Tasks, resulting in 18 mean measures per participant. We then further averaged the data across the genders. For E+ and E- words rated on our E Task (Good vs. Bad) on the Evaluation dimension, our young participants produced ratings for each of the 40 words that significantly correlated with Osgood's original ratings for these words ($n = 40$, $r = 0.9995$, $p < 0.0001$). For the P+ and P- words rated by our P Task (Strong vs. Weak) on the

Table 2 Comparisons of word relevance (Positive Words, Neutral Words, Negative Words) within each rating Task (Evaluation, Potency, Activity).

Word Relevance	Mean	SD	Significance
<i>E Rating Task (Good vs. Bad): F(2, 561) = 472.79, p < 0.0001</i>			
Positive Words: E+	1.70	0.62	* Different from Neutral, Negative
Neutral Words: P+, P-, A+, A-	0.97	0.67	* Different from Positive, Negative
Negative Words: E-	-1.27	0.95	* Different from Positive, Neutral
<i>P Rating Task (Strong vs. Weak): F(2, 561) = 66.03, p < 0.0001</i>			
Positive Words: P+	1.29	0.66	* Different from Neutral, Negative
Neutral Words: E+, E-, A+, A-	0.79	0.83	* Different from Positive, Negative
Negative Words: P-	-0.06	0.94	* Different from Positive, Neutral
<i>A Rating Task (Fast vs. Slow): F(2, 561) = 23.63, p < 0.0001</i>			
Positive Words: A+	0.83	0.61	* Different from Neutral, Negative
Neutral Words: E+, E-, P+, P-	0.43	0.79	* Different from Positive, Negative
Negative Words: A-	0.03	0.93	* Different from Positive, Neutral

The Positive and Negative word groups contained 20 words each. The Neutral word group contained 80 words. ANOVAs were performed by each rating Task. All 94 participants were used in this analysis with gender not separated. Each analysis was corrected for multiple comparisons using the Bonferroni adjustment. Mean estimates refer to predicted population margins (estimates of the marginal means over a balanced population with the covariance structure derived in the mixed linear model). For all analyses, the standard error was 0.09.

Potency dimension, there was also a significant correlation between our young participants' word ratings and Osgood's ratings ($n = 40, r = 0.9765, p < 0.0001$). Finally, we found a third significant correlation for the A+ and A- words rated on the A Task (Fast vs. Slow) along the Activity dimension ($n = 40, r = 0.9257, p < 0.01$).

Given the significant correlations, we combined experimental conditions to create measures of Word Relevance. Word Relevance refers to whether the words, on average by Word-Class, were relevant to individual semantic rating tasks: Evaluation, Potency, and Activity. By design, the ratings for each of the Word-Classes should be a function of the Task; therefore, reorganizing the data this way reduces structural complexity of the dataset while maintaining the useful variation related to the word rating paradigm. Positive words ($n = 20$) were those Word-Classes with the most extreme positive rating on that rating Task (for example, the E+ words on the E rating Task). Negative words ($n = 20$) were the Word-Classes with the most extreme negative rating on that rating Task (E- words on the E Task, for example). The Neutral words ($n = 80$) were the average of the remaining Word-Classes that have, on average, ratings near zero for that Task (the P+, P-, A+, and A- words on the E Task). This reorganization was applied to each rating Task. Note, again, that the same words change Word Relevance depending on the Task; this served to reduce the possibility that word repetition effects, as well as covariates based upon the words themselves, would influence the Gender results.

Before continuing, we confirmed that these averages were correct and applicable to the data. We used ANOVA to examine differences among Positive, Neutral, and Negative words within each Task across both genders (Table 2). First, each of the Tasks has a significant Word Relevance effect. Second, for all three tasks, the Neutral words were, on average, rated between the Positive and Negative words. This was the anticipated result, and so we continued with our Word Relevance groupings.

Modeling Osgood's dimensions and gender. To examine Gender differences with our Word Relevance measures, we began with a linear mixed model procedure using repeated measures. This approach permits the possibility of random, demographic and subject-related effects that are not of particular interest to this study, such as age, to influence the analysis. The model consisted of word rating as a function of the fixed effects of Gender and Word Relevance (Positive, Neutral, Negative) and their interaction. We also introduced Age and Education into the model as

random effects, as well as random effects of Word Concreteness, Word Frequency, and Word Age of Acquisition (Supplemental Table 1). While these do not vary with subject effects, they do vary with Word-Class and hence with Word Relevance. There is also some evidence these effects, while not a focus of this study, interact with affective processing (Warriner et al., 2013). The random effect of Education was not found to significantly contribute to or interact with the model. In addition, Concreteness, Frequency, and Age of Acquisition did not provide meaningful covariance in this model (Table 3). Age, however, had a significant Wald's Z statistic ($Z = 4.69, p < 0.0001$). The residual was also significant, suggesting the model can be improved.

Given all random effects, aside from Age, did not improve the model, we opted to remove them and use the simpler model with fixed effects of Gender and Word Relevance and one random effect of Age. With this, we found a significant main Gender effect ($F(1, 769) = 5.38, p < 0.05$). We also found a significant Word Relevance effect ($F(2, 769) = 385.03, p < 0.0001$), which again serves to confirm Osgood's dimensionality within the dataset. Finally, there was a significant two-way interaction of Gender \times Word Relevance ($F(2, 769) = 9.55, p < 0.0001$).

Examining gender differences. The significant main Gender effect and Gender \times Word Relevance interaction suggest meaningful gender differences within our word rating data. For an initial view of gender differences in the three-dimensional word-space, we plotted each of the 120 words by their Word-Class for our two groups of 47 men and 47 women. This was done for each of the rating Tasks (E, P, and A) (Fig. 2). Women's word ratings appear in red, and men's word ratings are blue. In addition, the positive Word-Classes (E+, P+, A+) are filled-in circles, while the negative Word-Classes (E-, P-, A-) are empty circles. Each of the 20 words within a Word-Class are shown. These scatter plots are a bit noisy, but they do depict some gender differences. Particularly for the negative Word-Classes, women seem to rate the words more negatively than men do. This is especially obvious for the E Task.

To plot this effect more fully, we derived mean Word-Class scores for each gender separately by averaging across the 20 words within each Word-Class. Since Word Relevance combined both Word-Class and rating Task (E, P, and A), we graphed these means for each Task separately (Fig. 3). Again, data is shown in red for females and blue for males. Again, we can see Osgood's dimensionality in these plots, confirming once more the experimental design combining both the Word-Class and the Task.

Table 3 Results of repeated measures mixed linear model analysis of word rating data ($n = 94$ participants).

Covariance parameter estimates - random effects				
Parameter	Estimate	Standard error	Wald's Z	p
Intercept (Age)	0.19	0.04	4.69	<0.0001
Intercept (Education)	0.05	0.04	1.17	0.12
Word frequency	0.01	0.01	0.68	0.25
Word concreteness	3.25	4.89	0.67	0.25
Word age of acquisition	0.07	0.10	0.69	0.25
Residual (participants)	0.38	0.02	19.38	<0.0001
Fixed effects				
Effect	DF	F	P	
Gender (Male, Female)	1, 769	5.38	0.02	
Word Relevance (Positive, Neutral, Negative)	2, 769	385.03	<0.0001	
Gender \times Word Relevance	2, 769	9.55	<0.0001	

Gender (Male, Female) and Word Relevance (Positive, Neutral, Negative stimulus words) were entered into the analysis as fixed effects. Participant variability was represented in the residual. Two random effects related to participants (Age, Education) and three random effects (Concreteness, Frequency, Age of Acquisition) related to the word stimuli were represented as covariance parameters. Note that the Fixed Effects were calculated after removing the random effects that did not improve the model. DF = degrees of freedom (numerator, denominator).

The difference between a pair of positive and negative Word-Classes is largest on those Word-Classes' matching their rating dimension (which can be noticed looking at the diagonal on Fig. 3 from the E Task on the top left to the A Task on the bottom right). For example, the difference between E+ words (which were quite positive) and E- words (which were quite negative) was far larger on the E Task than on either of the other two Tasks (P, A). This difference for E+ and E- words was also the largest difference on the E rating Task (observed by comparing the three Word-Class differences horizontally within the E rating Task). The large difference between a matching pair of Word-Classes on the corresponding rating Task shows the semantic differential, which is the result we would expect from applying Osgood's semantic dimensionality to these word rating data.

Figure 3 also shows more clearly the gender difference seen in Fig. 2, specifically that young women may rate words more negatively than young men do. The magnitude of the difference varies, but it appears for nearly every Word-Class and rating Task. The effect appears to be mainly true of the negative Word-Class on its corresponding task. For example, a large disparity can be observed between young women (more negative) on P- words on the P task than young men (more positive). Compared to the P+ words, this P- gender difference appears quite sizeable. This is also true for A+ and A- words on the A Task (though there is also a noticeable difference between the P+ and P- words as well).

These effects are quantified in the postfitting statistical analyses performed using the mixed linear model. The main Gender effect is quite obvious in Fig. 3; across nearly all the Word-Classes and Tasks, women seem to rate words more negatively. However, this difference is only significant for Negative words. For each Task show in Fig. 3, the Negative words are rated more negatively by women. This corresponds to a significant Gender difference for Negative words ($F(1, 769) = 19.25, p < 0.0001$) (Table 4).

Mapping evaluation, potency, and activity onto valence, dominance, and arousal. We wanted to assess how much overlap there may be between Osgood's connotative dimensions (1957) and the dimensions developed by Mehrabian and Russell (1974) since a great deal of the recent research has focused on this model rather than Osgood's original work. There appear to be some interpretative similarities in the literature (Bakker et al., 2014; Bradley and Lang, 1999), but we were unable to find a direct mathematical comparison. Therefore, we decided to examine how

well the three dimensions correlate across varying datasets to get a clearer picture.

Though many studies of affective word meaning make use of the Mehrabian Valence/Arousal/Dominance model and the ANEW lexicon (Bradley and Lang, 1999), we selected the list of affective word ratings provided by Warriner et al. (2013) as an example set of data. These data were collected in a manner similar to our own, where words were presented to individuals through an online survey during which the participant rated the word using an adjective pair. As a simple test, we correlated our raw word rating data with their raw rating data for young adults along each dimension for the 110 words that overlapped between our two stimuli sets. We did so by interpretation of the dimensions. For Warriner's Valence and our Evaluation dimensions, the correlation was strong ($n = 110, r = 0.89, p < 0.0001$), accounting for 79% of the variance between the datasets. However, for the other two pair-wise comparisons, the strength of the correlation dropped (Warriner's Arousal with our Activity: $r = 0.37, p < 0.0001$; Warriner's Dominance with our Potency: $r = 0.09, p = 0.36$), with large drops in accounted variance.

This suggested there may be some fundamental differences between word ratings in our study and theirs. Warriner (Warriner et al., 2013) suggests that the Dominance dimension, as defined in the ANEW lexicon (Bradley and Lang, 1999), is correlated with the Valence dimension. We found evidence of this as well. We also submitted Warriner's ratings as well as our own rating data for each of the same 110 words to a principal components analysis, hypothesizing that, if our studies are truly measuring the same independent dimensions, the procedure should group them together by correlation (Valence with Evaluation, etc.). Instead, we found after Varimax rotation that Factor 1 was comprised of our Evaluation, Warriner's Valence, and Warriner's Dominance. Warriner's Arousal and Activity were strongly loaded on Factor 2. Factor 3, however, contained only our Potency ratings (Supplemental Table 2). Again, this implied that Dominance and Valence (as measured by Warriner) are correlated. It also suggests Dominance and Potency, despite a common conceptual interpretation, are not strongly mathematically related. The variance explained by each factor after rotation was 2.59, 1.53, and 0.77, respectively.

Discussion

Osgood's semantic differential techniques provide a well-validated method of examining the connotative meaning of words.

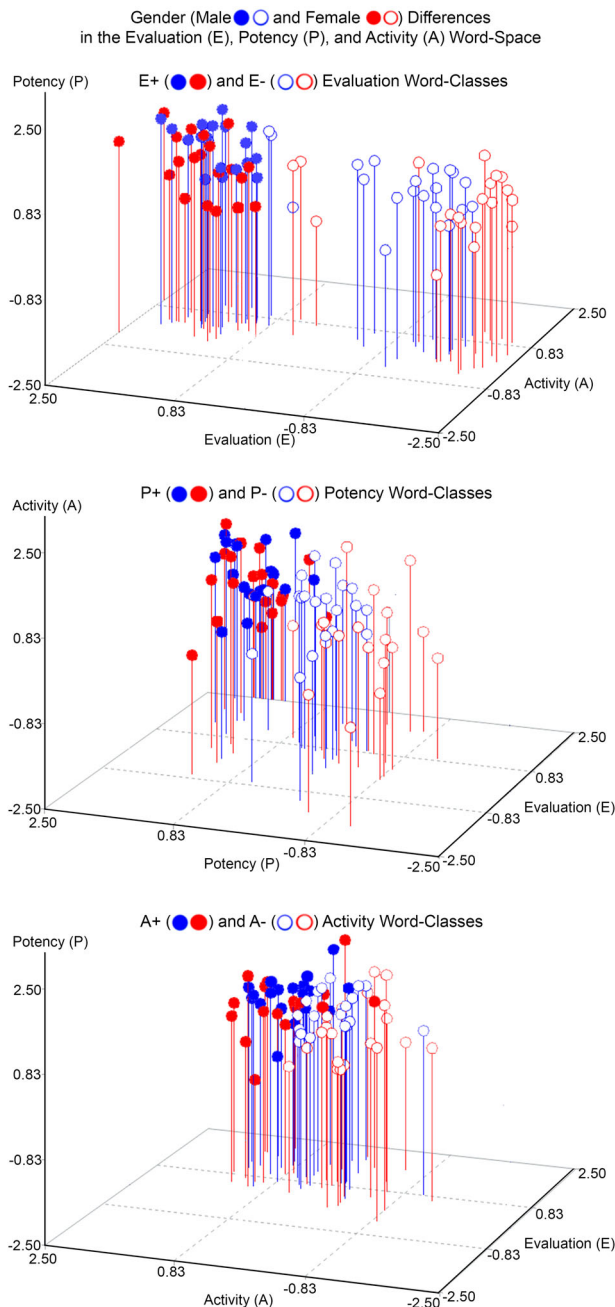


Fig. 2 Scatter plots of gender differences (averaged for $n = 47$ participants in each gender) in the three-dimensional EPA word-space. Note that the three axes are rotated to place the axis corresponding to the rating Task along the X-axis. This was done to better visualize gender differences in the three-dimensional space. Particularly for negative WordClasses (E-, P-, and A-), women tended to rate words more negatively.

This approach allows the quantification of emotions and affective processing, which otherwise could be highly subjective and difficult to measure. While semantic and other affective differentials have been studied in many different research contexts, defining fundamental gender differences in a young population concerning affective processing can lay important groundwork across multiple disciplines of human cognition and behavior.

The connotative meaning of words. Clearly the semantic differentials Osgood originally developed (Osgood et al., 1957

Gender Differences (Male ● and Female ●) in Connotative Word Meaning

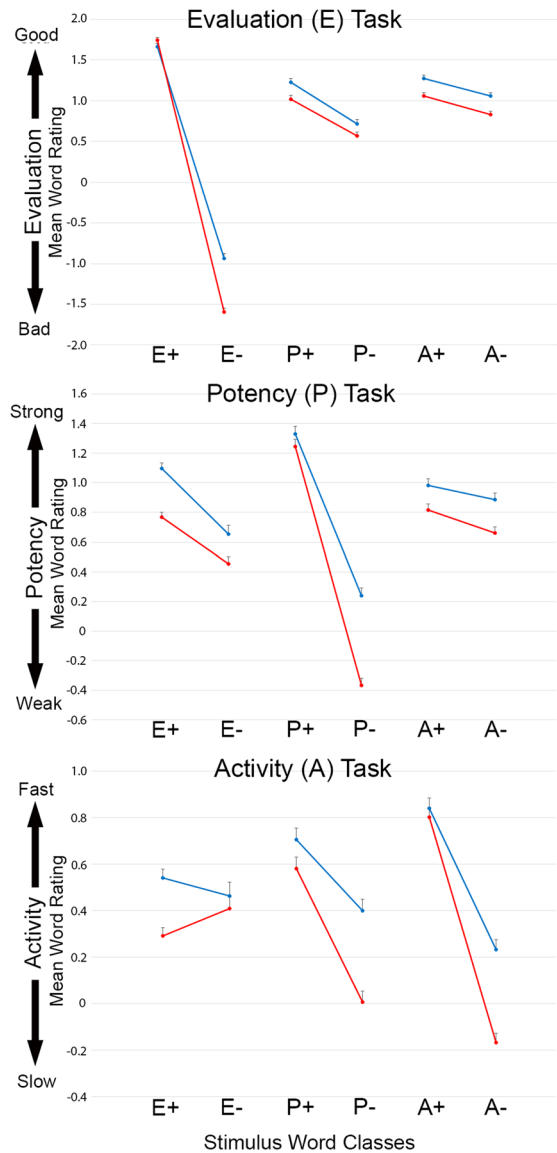


Fig. 3 Word rating means on each rating Task by gender in our group of 94 young adults. The words are grouped into Word-Classes by their largest positive or negative connotative word rating on each of the three word rating Tasks (E, P, and A) as compiled by Heise and Snider and Osgood (Chapman, 1978a). Error bars represent standard error of the mean (SEM).

(Fig. 1) and others have studied (Chapman, 1978b, 1979; Chapman et al. 1977, 1978, 1980; Heise, 1971; Skrandies, 2014) appear in our word rating dataset. Our manipulation of Osgood's and Heise's words into our Word-Classes and then into different types of Word Relevance (Positive, Neutral, Negative) depending on the rating Task is empirically supported. When we combined our word ratings into our Word-Classes, our participants' responses correlated highly with Osgood's three semantic differential dimensions. Additionally, our mean word rating results by Word-Class show that both genders exhibited the expected semantic differential. Between the positive and negative Word-Classes on the corresponding differential Task, there are marked differences in ratings such that the positive Word-Class (E+ words on the E task, for example) had more positive ratings and

Table 4 Comparisons of gender within each word relevance (Positive Words, Neutral Words, Negative Words) with the rating tasks (E, P, and A) combined.

Gender	Mean estimate	Significance
<i>Positive Words: $F(1, 769) = 0.15, p = 0.70$</i>		
Male	1.22	Not significant
Female	1.26	
<i>Neutral Words: $F(1, 769) = 1.84, p = 0.18$</i>		
Male	0.77	Not significant
Female	0.62	
<i>Negative Words: $F(1, 769) = 19.25, p < 0.0001$</i>		
Male	-0.38	Significant gender difference
Female	-0.71	

This table represents comparisons within the Gender \times Word Relevance interaction (see Table 3 for the complete mixed linear model). Each analysis was corrected for multiple comparisons using the Bonferroni adjustment. Mean estimates refer to predicted population margins (estimates of the marginal means over a balanced population with the covariance structure derived in the mixed linear model). For all analyses, the standard error was 0.12. NS = Not significant.

the negative Word-Class (E- words on the E task) was rated more negatively. Thus, connotative aspect of these words holds true: the word “baby”, for example, elicited a generally “good” emotional response, while the word “war” produced a generally “bad” one. Our word rating data (Fig. 3) also showed that the other Word-Classes that should be more neutral on average on the Task in question (e.g., A- words on the E Task) showed a more neutral semantic differential than the Word-Class pair corresponding to that Task. This effect is supported by our main effect of Word Relevance ($p < 0.001$).

Our six average Word-Classes (Supplemental Fig. 1) organize the data in a meaningful way that constructs a “purer” positive or negative rating on each of the three orthogonal, semantic dimensions (E, P, A) than single word ratings. This approach also provides “control words” that should not produce a strong semantic differential on that rating Task (Fig. 3). Finally, because the same stimulus words are organized as either relevant (Positive or Negative) or irrelevant (Neutral) to each Task, we can reduce the influence of the words themselves affecting connotative ratings. We believe that focusing on words that are more meaningful to each Task and comparing those ratings to the same words that are Neutral to the other Tasks is an important part of our approach. This method reduces the noise inherent in examining a great deal of ratings of “random” words offered in a lexicon and focuses on comparing words that should elicit a strong directional effect on one task dimension and words that should not produce much of an average response to the other task dimensions. This is not an approach we have seen much in the research on affective meaning.

In addition, given that much of the current research on affective meaning and gender differences focuses on Valence, Arousal, and Dominance (and often times, omitting Dominance entirely), the question of how well that model overlaps with Osgood’s is an important one. We posit that there appears to be a number of differences in the experimental design and interpretations between Mehrabian and Russell (1974) and Osgood et al. (1957). First, Osgood structured his experiments such that subjects were asked to make semantic judgments on the meaning of words (to rate the words based upon the properties of the word’s affective meaning given strict adjective pairs). Conversely, Mehrabian and Russell (1974) and research that stems from their work typically involves asking the participant to make a judgment based on how that word (or picture stimulus) makes the participant feel (Bradley and Lang, 2002; Warriner et al., 2013). These may be two different constructs (one of language

and the other of self-assessment), which could occasionally be correlated but not always depending on the words selected and the subjects involved. For example, given the word “baby” and a scale of “happy vs. unhappy”, a majority of people may rate this word as “happy” when asked to judge the meaning of the word. However, when asked to rate how the word makes them feel, some may rate this word less strongly “happy”, depending on their own perceptions, experiences, and emotional states. It is possible that conflating these constructs (measuring a semantic property of a word’s meaning versus what emotional response a word generates) has led to some confusing results when examining word ratings and how different demographic groups process language.

Second, the factor structure underlying the Valence, Arousal, and Dominance model is not likely to be orthogonal. Mehrabian and Russell (1974) and Bradley and Lang in their repetition of this work (1994) conducted factor analyses that did not use an orthogonal rotation. This then produces factors that are not mathematically independent from each other. Warriner et al. (2013) found in examining their words ratings between the Valence and Dominance dimensions that these values were significantly correlated. We also determined that the correlation between their rating results for Valence and ours for Evaluation is significant and accounts for most of the variance in the dataset, but correlations between our Potency and their Dominance ratings and between our Activity and their Arousal ratings are not nearly so large or meaningful. Finally, our factor analyses of word ratings using words that occur in both datasets produced factors where Warriner’s Dominance dimension appeared on the same factor as our Evaluation dimension and Warriner’s Valence dimension. All of this evidence lends credence to the possibility that, despite common theoretical underpinnings, these two constructs of affective meaning are perhaps not mathematically the same. This could in turn affect interpretations of results using these differing measurement systems.

Women and affective processing. Our work with these affective rating tasks suggests that women tend to rate the emotional aspects of stimuli more negatively than men. Our mixed linear models analysis produced a main Gender effect ($p < 0.05$) and a significant interaction between Gender and Word Relevance ($p < 0.0001$) (Table 3). These effects can be visualized in the three-dimensional word-space in Fig. 2 and in the mean word ratings in Fig. 3. Women generally tended to rate words more negatively than men do, leading to the main Gender effect and women’s word ratings surrounding men’s word ratings in Fig. 2. Figure 3 shows this pattern of gender differences for nearly all Word-Classes and regardless of the connotative word rating Task (Evaluation, Potency, or Activity). Importantly, the difference between women and men is only significant on the Negative words (E- words on the E Task, P- words on the P Task, A- words on the A task) ($F(1, 764) = 17.33, p < 0.0001$). This creates a larger, steeper semantic differential for women than men on all three rating Tasks.

Literature review of brain imaging and other studies has indicated there are neurological underpinnings for different emotion regulatory strategies in men and women that may lead to behavioral differences in how women make judgments concerning emotional stimuli (Bradley and Lang, 2002; Whittle et al., 2011). Additionally, it appears that females show greater neural activation to negative emotional stimuli (particularly involving the amygdala) than men. The fact that women react more negatively to negative stimuli has been shown in quite a few studies (Belleza et al., 1986; Grunwald et al., 2010; Söderholm et al., 2013; Warriner et al., 2013). Marogna et al. (2016) noted

that with university students women's semantic judgments tended to fall at either end of each bipolar continuum, and their negative perceptions tended to be extremely negative. This inclination to rate items more on the extreme of the differential scale has been mirrored in other studies exploring facial expressions and internet culture (Hall and Matsumoto, 2004; King, 2001; MacKinnon and Keating, 1989; Skrandies, 2014). Vasa (2006) suggested that girls rated positive and threat words more extremely than boys did, but there was no gender difference on neutral words. There is specific evidence that girls and women may react more strongly to negatively valenced stimuli, particularly in cases of high stimulus arousal (Gabert-Quillen et al., 2015; Markovits et al., 2018; Söderholm et al., 2013; Vasa et al., 2006). It would be of interest to determine if gender-based analyses on words that have both high ratings on the Evaluation and Arousal dimensions would produce a similar result.

It is also interesting that the Potency dimension, which has become less featured in studies of affective processing, shows the same gender effect as the more commonly explored Evaluation and Activity dimensions. Often there has been focus on a bi-dimensional model of affective processing, tying Valence/Evaluation in a curvilinear fashion to Activity/Arousal (Soares et al., 2012; Söderholm et al., 2013; Teismann et al., 2020). We believe our results do show that the Potency dimension is not negligible. The same pattern of gender differences we found on the Evaluation and Activity Tasks also appeared on the Potency Task. Since, in Osgood's structure, these are independent factors, we believe this adds meaningful context to how the two genders process emotional stimuli differently. In other studies of affective word processing, if Dominance and Valence are correlated, it may be difficult to tease out differences related to underlying dimensions, whereas Osgood's approach, focusing on the connotative property of the word itself rather than the emotion it personally elicits, begins with uncorrelated dimensions and words organized by empirical factor loadings.

Regardless of these differences, there is clearly a gender difference in the semantic processing of word connotative meaning. MacKinnon (MacKinnon and Keating, 1989) postulated that women are generally more closely in touch with their feelings and more affectively expressive. This, in turn, leads to greater discrimination and variation in the cognitive labels used to represent connotative meaning. Therefore, women may show a greater range in affective decision-making and rating. This is strongly evident in our measure of a steeper semantic differential in women than in men (Fig. 3). Also, work in facial emotion recognition suggests that men perhaps have less sensitivity to the emotional aspects of stimuli than women do (Montagne et al., 2005). All of these important results in emotional processing lay important groundwork for assessing gender equalities and differences in psychology and cognition. Research suggests aging and neurological disorders alike may occur differently in men and women, and understanding these mechanisms can help build more targeted therapeutics tailored to the different genders (Baizabal-Carvallo and Jankovic, 2020; Chapman et al., 2011; Georgiev et al., 2017; Matin et al., 2017; Young and Pfaff, 2014).

Limitations and additional considerations. Our results are somewhat constrained by the anonymous nature of our data collection. This study was conducted via the internet, which expands its generalizability greatly since the sample is less homogeneous. For a baseline gender analysis of a well-validated and easily employed measurement tool, this is an important advantage. However, we are relying on the subjects' self-report for their classifying information, such as their gender and diagnosis

with language, cognitive, and other disorders that may impact their processing of connotative meaning. While we administered anxiety and depression questionnaires to screen individuals suffering from mood disorders that may have influence on affective processing which may themselves have gender differences, this analysis would benefit from replication with clinical evaluation. Also, we can only make interpretations regarding gender differences, rather than biologically based sex differences, since this study used gender self-report. Correlating these behavioral results with biological differences is a useful step.

Also, this study is not controlling for cultural differences. All participants were located within the United States, but the diverse cultural regions of the country may introduce effects that could impact our findings. This, as well as exposure to other languages, should be studied, although Osgood's semantic differential has been shown to be constant across many different languages and cultures and related gender differences may also be consistent (Ellis et al., 1994; Gibson, 1995; Moore et al., 1999). In addition, there is a small but significant age difference between our young men and young women groups such that women were on average roughly 5 years older than men. While research has indicated that age-related changes in affective processing are important (Skrandies, 2014; Teismann et al., 2020; Warriner et al., 2013), we accounted for any covariance due to age in our mixed linear model. The significant random effect (Table 3) furthers our belief that age requires research using Osgood's approach.

Finally, it is possible that some of these effects could be explained by a gender bias in the original dataset of word ratings built by Osgood (Snider and Osgood, 1969) and Heise (1971), particularly the nearly significant effect that women rated words more negatively almost "across the board". Heise made a careful effort through multivariate methods, including PCA and regression, to remove the effects of age, gender, and many other covariates from his list of 1551 words which served as the source of the 120 words used in this study. Still, if the original dataset was comprised of mostly young men, it could subtly shift the neutral to be more positive, which would then make it appear that women are reacting more negatively. This is slightly apparent in Supplemental Fig. 1, where the "neutral" aspects of some of the average Heise ratings for each Word-Class appear more Evaluation positive. One of the datasets used in Heise's original work was gathered from naval recruits and its participants were likely male. Osgood has also indicated some of his early work was conducted with young boys (Osgood, 1969b). Again, this slight bias perhaps introduced by the fact these foundational studies used primarily boys and men in their samples may influence our results today and otherwise dampen or skew our measurement of gender differences within the three-dimensional affective word-space. However, the affective meaning of words along the three, orthogonal dimensions is still strongly supported by our work utilizing an equal number of men and women, and the gender difference of the placement of words within that space is large and significant.

Conclusion

This study confirms that there are gender differences in the affective processing of word stimuli in emotionally healthy young adults. Rating words based on their connotative meanings are sensitive to gender effects such that young women showed a significantly larger semantic differential (the difference between positive and negative words on a particular rating scale) than young men did. It remains to be seen if this difference remains constant with aging and in the presence of mood disorders. Our efforts here also suggest that Osgood's semantic structure and

approach may be different enough from the Valence/Arousal/Dominance model that comparisons between them should be made tentatively. More work is really required to determine in a modern population how comparable these dimensional constructs truly are.

Data availability

Data has been made publicly available free of charge on the University of Rochester's UR Research repository (<https://urresearch.rochester.edu/>).

Received: 8 September 2021; Accepted: 10 March 2022;

Published online: 06 April 2022

References

- Amazon Mechanical Turk Inc. (2020) Amazon mechanical turk <https://www.mturk.com/>. Accessed 30 Apr 2020
- Baizabal-Carvallo JF, Jankovic J (2020) Gender differences in functional movement disorders. *Mov Disord Clin Pract* 7(2):182–187
- Bakker I, van der Voordt T, Vink P et al. (2014) Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Curr Psychol* 33:405–421
- Belleza F, Greenwald AG, Banaji MR (1986) Words high and low in pleasantness as rated by male and female college students. *Behav Res Methods Instrum Comput* 18(3):299–303
- Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. *J Behav Ther Exp Psychiatry* 25:49–59
- Bradley MM, Lang PJ (1999) Affective norms for English words (ANEW): instruction manual and affective ratings. University of Florida, Gainesville, FL
- Bradley MM, Lang PJ (2002) Measuring emotion: behavior, feeling, and physiology. In: Lane RD, Nadel L (Eds.) *Cognitive neuroscience of emotion*. Oxford University Press, pp. 242–276
- Chapman RM (1978a) Language and evoked potentials. In: Otto DA (Ed.) *Multidisciplinary perspectives in event-related brain potential research*. US Government Printing Office, Washington, DC, pp. 245–249
- Chapman RM (1978b) Method of EP analysis in linguistic research. In: Otto DA (Ed.) *Multidisciplinary perspectives in event-related brain potential research*. US Government Printing Office, Washington, DC, pp. 265–266
- Chapman RM (1979). Connotative meaning and averaged evoked potentials. In: Begleiter H (Ed.) *Evoked brain potentials and behavior*. Plenum Publishing Corporation, pp. 171–196
- Chapman RM, Bragdon HR, Chapman JA et al. (1977) Semantic meaning of words and average evoked potentials. In: Desmedt JE (Ed.) *Progress in clinical neurophysiology, language and hemispheric specialization in man: cerebral event-related potentials*. Karger, Basel, pp. 36–47
- Chapman RM, Mapstone M, Gardner MN et al. (2011) Women have farther to fall: gender differences between normal elderly and Alzheimer's disease in verbal memory engender better detection of AD in women. *J Int Neuropsychol Soc* 17:654–662. <https://doi.org/10.1017/S1355617711000452>
- Chapman RM, McCrary JW, Chapman JA et al. (1978) Brain responses related to semantic meaning. *Brain Lang* 5:195–205
- Chapman RM, McCrary JW, Chapman JA et al. (1980) Behavioral and neural analyses of connotative meaning: word classes and rating scales. *Brain Lang* 11:319–339
- Davies M (2009) The 385+ million word corpus of contemporary American English (1990–present). *Int J Corpus Linguist* 14(2):159–190. 10.1075%2Fijcl.14.2.02dav
- Davies M (2010) The corpus of contemporary American English as the first reliable monitor corpus of English. *Digit Sch Humanit* 25(4):447–465. 10.1093%2Fllc%2Ffqq018
- Davies M (2021) Corpus of contemporary American English. <https://www.english-corpora.org/coca/>. Accessed 1 Apr 2020
- Ellis BB, Kimmel HD, Diaz-Guerrero R et al. (1994) Love and power in Mexico, Spain, and the United States. *J Cross Cult Psychol* 25(4):525–540
- Gabert-Quillen CA, Bartolini EE, Abravanel BT et al. (2015) Ratings for emotion film clips. *Behav Res Methods* 47(3):14
- Georgiev D, Hamberg K, Hariz M et al. (2017) Gender differences in Parkinson's disease: a clinical perspective. *Acta Neurol Scand* 136(6):570–584
- Gibson CB (1995) An investigation of gender differences in leadership across four countries. *J Int Bus Stud* 26(2):255–279
- Grunwald IS, Borod JC, Obler LK et al. (2010) The effects of age and gender on the perception of lexical emotion. *Appl Neuropsychol* 6(4):226–238
- Hall JA, Matsumoto D (2004) Gender differences in judgments of multiple emotions from facial expressions. *Emotion* 4(2):201
- Heise DR (1965) Semantic differential profiles for 1,000 most frequent English words. *Psychol Monogr* 79(8):1–31
- Heise DR (1971) Evaluation, potency, and activity scores for 1,551 words: a merging of three published lists. University of North Carolina, Chapel Hill, NC
- Jackson JC, Watts J, Henry TR et al. (2019) Emotion semantics show both cultural variation and universal structure. *Science* 366(6472):1517–1522
- King AB (2001) Affective dimensions of Internet culture. *Soc Sci Comput Rev* 19(4):414–430
- Kissler J, Assadollahi R, Herbert C (2006) Emotional and semantic networks in visual word processing: insights from ERP studies. In: *Progress in Brain Research*. Elsevier, Amsterdam, pp. 147–183
- MacKinnon NJ, Keating LJ (1989) The structure of emotions: Canada-United States comparisons. *Soc Psychol Q* 52(1):70–83
- Majid A (2019) Mapping words reveals emotional diversity. *Science* 366(6472):1444–1445
- Markovits H, Trémolière B, Blanchette I (2018) Reasoning strategies modulate gender differences in emotion processing. *Cognition* 170:76–82
- Marogna C, Caccamo F, Salcuni S et al. (2016) University students and semantic differential: a pilot study comparing subjects who sought psychological help with subjects who did not. *Test Psychom Methods Appl Psychol* 23(3):319–333
- Martin N, Young SS, Williams B et al. (2017) Neuropsychiatric associations with gender, illness duration, work disability, and motor subtype in a US functional neurological disorders clinic population. *J Neuropsychiatry Clin Neurosci* 29(4):375–382
- Mehrabian A, Russell JA (1974) *An approach to environmental psychology*. MIT, Cambridge, MA
- Montagne B, Kessels RPC, Frigerio E et al. (2005) Sex differences in the perception of affective facial expressions: do men really lack emotional sensitivity? *Cogn Process* 6:136–141
- Moore CC, Romney AK, Hsia TL et al. (1999) The universality of the semantic structure of emotion terms: methods for the study of inter- and intra-cultural variability. *Am Anthropol* 101(3):529–546
- Mukherjee S, Heise DR (2017) Affective meanings of 1,469 Bengali concepts. *Behav Res Methods* 49:184–197
- Northwestern University (2021) Health Measures: intro to PROMIS. <https://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis>. Accessed 30 Nov 2019
- Osgood CE (1969a) On the whys and wherefores of E, P, and A. *J Pers Soc Psychol* 12(3):194–199
- Osgood CE (1969b) Semantic differential technique in the comparative study of cultures. In: Snider JG, Osgood CE (Eds.) *Semantic differential technique, a sourcebook*. Aldine, Chicago, pp. 303–332
- Osgood CE (1980) *Lectures on language performance*. Springer-Verlag New York, Inc., New York, NY
- Osgood CE, May WH, Miron MS (1975) *Cross-cultural universals of affective meaning*. University of Illinois Press, Urbana, IL
- Osgood CE, McGuigan FJ (1973) Psychophysiological correlates of meaning: essences or tracers? In: McGuigan FJ, Schoonover R (Eds.) *Psychophysiology of thinking*. Academic Press, New York, pp. 449–492
- Osgood CE, Suci GJ, Tannenbaum PH (1957) *The measurement of meaning*. University of Illinois Press, Urbana, Chicago, and London
- Qualtrics (2020) qualtricsXM <https://www.qualtrics.com/>. Accessed 5 May 2020
- Ratcliff R, Hendrickson AT (2021) Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behav Res Methods* 53:2302–2325
- Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39:1161–1178
- SAS Institute Inc. (2017) SAS 9.4 product documentation. <http://support.sas.com/documentation/94/> Accessed 15 June 2021
- Sereno SC, Scott GG, Yao B et al. (2015) Emotion word processing: does mood make a difference? *Front Psychol* 6:1–13
- Skrandies W (2014) Electrophysiological correlates of connotative meaning in healthy children. *Brain Topogr* 27:271–278
- Skrandies W, Chiu MJ (2003) Dimensions of affective semantic meaning—behavioral and evoked potential correlates in Chinese subjects. *Neurosci Lett* 341:45–48
- Snider JG, Osgood CE (1969) Semantic atlas for 550 concepts. In: Snider JG, Osgood CE (Eds.) *Semantic differential technique: a sourcebook*. Aldine, Chicago, pp. 625–636
- Soares AP, Comesaña M, Pinheiro AP et al. (2012) The adaptation of the affective norms for English words (ANEW) for European Portuguese. *Behav Res* 44:256–269
- Söderholm C, Häyry E, Laine M et al. (2013) Valence and arousal ratings for 420 Finnish nouns by age and gender. *PLoS ONE* 8(8):1–10
- Teismann H, Kissler J, Berger K (2020) Investigating the roles of age, sex, depression, and anxiety for valence and arousal ratings of words: a population-based study. *BMC Psychol* 8:1–14. <https://doi.org/10.1186/s40359-020-00485-3>
- Vasa RA, Carlino AR, London K et al. (2006) Valence ratings of emotional and non-emotional words in children. *Pers Individ Dif* 41(6):1169–1180

- Warriner MB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res* 45:1191–1207
- Whittle S, Yücel M, Yap MBH et al. (2011) Sex differences in the neural correlates of emotion: evidence from neuroimaging. *Biol Psychol* 87:319–333
- Young LJ, Pfaff DW (2014) Sex differences in neurological and psychiatric disorders. *Front Neuroendocrinol* 35(3):253–254
- Zablotsky B, Black LI, Maenner MJ et al. (2019) Prevalence and trends of developmental disabilities among children in the United States: 2009–2017. *Pediatrics* 144(4) <https://doi.org/10.1542/peds.2019-0811>

Acknowledgements

We thank R. Emerson, S. Theis, D. Theis, and W. Vaughn for their technical contributions; SRS Rao Poduri for statistical advice; S. Chapman for help in writing; our undergraduate assistants and research assistants; and the many voluntary participants in this research. Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Numbers P30-AG08665, R01-AG041313, and R01-AG065314. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

The authors declare no competing interests.

Ethical approval

This study was approved by the University of Rochester Research Subjects Review Board (IRB) as exempt due to the anonymous nature of the data collection (4/27/2020). All ethical and IRB guidelines were followed.

Informed consent

Study participants were shown an IRB-approved informed consent document to which they agreed before beginning the study.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-022-01126-3>.

Correspondence and requests for materials should be addressed to Robert M. Chapman.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022, corrected publication 2022