Behavior Research Methods (2022) 54:26–41
https://doi.org/10.3758/s13428-021-01607-4

Valence and arousal ratings for 11,310 simplified Chinese words

Xu Xu[1] · Jiayin Li[1] · Huilin Chen[1]

Accepted: 29 April 2021 / Published online: 7 June 2021

Abstract
This study reports valence and arousal ratings for 11,310 simplified Chinese words, including 9774 two-character words, 949 three-character words, and 587 four-character words. These affective ratings are validated through comparisons with prior ratings of smaller word samples. All but four words included in this study are from the MEgastudy of Lexical Decision in Simplified CHinese (MELD-SCH) database. As age-of-acquisition ratings and concreteness ratings have recently become available for large portions of words in the MELD-SCH, the affective ratings not only further enrich the database as a valuable research tool, but also allow us to gain insight into a range of psycholinguistic constructs based on normative ratings of a large set of Chinese words. Cross-language comparisons of the valence ratings between Chinese words and English words appear to indicate cultural and sociopolitical influences reflected in affect representations.

Keywords Valence · Arousal · Concreteness · Age-of-acquisition · Chinese words · Cross-language comparison

Affective experiences constitute a significant part of our lives, and the ability to understand and express these experiences plays a critical role in our physical and psychological well-being. Clinical research has shown that impairment in affect communication is linked to a range of health issues, including but not limited to cardiac disease, chronic pain, depression, diabetes, eating disorders, morbid obesity, and substance dependence (e.g., Bird & Cook, 2013; Lumley et al., 2007; Ricciardi et al., 2015; Taylor et al., 1997). The relationship of language and emotion therefore has drawn great attention from researchers across disciplines. On the one hand, researchers attempt to, through the lens of language, peek into the inner workings of affect processing in both healthy and clinical populations. On the other hand, investigative efforts are directed to assessing both the affective qualities of natural language data and the impacts of affect on language comprehension. Recently, Hinojosa et al. (2020) have proposed the theoretical framework of affective neurolinguistics based on the interplay of language and emotion to bridge research between the two fields.

In these long lines of research, affective ratings of words have played an important role. Taking English as an example, the Affective Norms for English Words (ANEW; Bradley & Lang, 1999) and a later substantial expansion of the database (Warriner et al., 2013) are widely utilized for the purposes of experimental stimuli construction (e.g., Duyser et al., 2020; Louwerse & Qu, 2017; Lund et al., 2019; Madan et al., 2012; Mordecai et al., 2017), sentiment analysis and opinion mining (e.g., Crossley et al., 2017; Islam & Zibran, 2018; Reagan et al., 2017; Wrobel, 2020), and automated affective lexicon expansion (e.g., Ahmed et al., 2020; Gatti et al., 2016; Palogiannidi et al., 2015; Wu & Tsai, 2014), as well as serving as templates to develop databases of affective norms in other languages, for example, Spanish (Redondo et al., 2007; Stadthagen-Gonzalez et al., 2017), Italian (Montefinese et al., 2014), Polish (Imbir, 2015), French (Monnier & Syssau, 2017), German (Schmidtke et al., 2014), Turkish (Torkamani-Azar et al., 2019), Croatian (Ćoso et al., 2019), and European Portuguese (Soares et al., 2012). As an illustration, Fig. 1 presents the number of journal articles since the early 2000s that have cited ANEW or Warriner et al. as a research tool in two different fields: clinical psychology and natural language processing.

As indicated above, affective norms for words are now available for many languages. Among them, large-scale databases can be found for English and Spanish. Warriner et al. (2013) published affective ratings for 13,915 English words, mostly nouns, adjectives, and verbs. Stadthagen-Gonzalez et al. (2017) published affective ratings for 14,031 Spanish words, covering almost all word categories. In addition, EmoFinder, a web-based search engine, provides affective

✉ Xu Xu
xu2xu3@gmail.com

[1] School of Foreign Languages, Shanghai Jiao Tong University, 800 Dongchuan Rd, Shanghai 200240, China

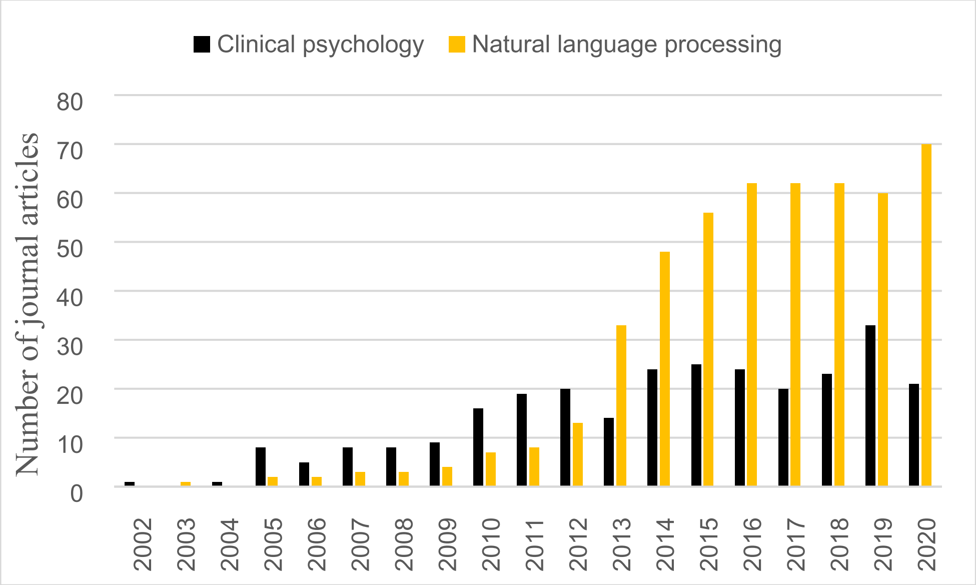

Fig. 1 Number of journal articles in clinical psychology and natural language processing citing ANEW (Bradley & Lang, 1999) or its expansion (Warriner et al., 2013)

ratings for 16,375 Spanish words (Fraga et al., 2018). Besides these large scale databases, affective norms for over 2000 words can also be found for Dutch (4300 words; Moors et al., 2013), German (2902 words by Võ et al., 2009; 2654 nouns by Lahl et al., 2009; and 2592 nouns by Grandy et al., 2020), Polish (2902 words; Riegel et al., 2015), Croatian (3022 words; Ćoso et al., 2019), and Turkish (2031 words; Kapucu et al., 2021).

There are a few databases of affective norms for Chinese words, both in simplified Chinese (Xu et al., 2008 for 334 words; Yao et al., 2017 for 1100 words; and Wang et al., 2008 for 1500 words) and in traditional Chinese (Ho et al., 2015 for 160 words; Yee, 2017 for 292 words). However, a review of the literature shows that there seem to be a greater number of experimental studies (e.g., Chen et al., 2015; Ding et al., 2016; Luo et al., 2020; Wang et al., 2019; Wang & Fu, 2011; Xu et al., 2017) that collected their own affective ratings for word stimuli than those (e.g., Wei et al., 2016; Yao et al., 2019; Zhang et al., 2017) that retrieved affective ratings from existing databases, an indicator that a larger database of affective norms for Chinese words is in demand.

The present study reports affective ratings for 11,310 simplified Chinese words. Similar to the ANEW (Bradley & Lang, 1999) and its expansion by Warriner et al. (2013), we collected both valence (i.e., positive versus negative affect conveyed by a word) and arousal (i.e., degree of arousal evoked by a word) ratings. Different from the two studies, we omitted dominance ratings (i.e., extent of feeling in control versus feeling being controlled as suggested by a word) as, in general, valence and arousal are considered the two primary dimensions of affect representation, whereas dominance has a much less salient dimension (e.g., Bradley & Lang, 1999; Osgood et al., 1957; Russell, 1980). In fact, rating analyses of both studies show a strong linear correlation between valence and dominance ($r = .84$ in Bradley & Lang, 1999; $r = .72$ in Warriner et al., 2013). Warriner et al. have further pointed out that they exhibit a similar relation with arousal, suggesting a considerable overlap between valence and dominance. As there has been ample theoretical discussion about these constructs in many studies cited above, we will defer further discussion until after the analysis of the affective ratings collected for simplified Chinese words in the present study.

Method

Participants

Participants took part in this study anonymously over the Internet. They were randomly assigned to complete either a valence rating task or an arousal rating task. A total of 2949 participants completed the study. Among these, 1444 provided valence ratings, while 1505 provided arousal ratings. They received monetary compensation at the end for participation.

In line with previous research (e.g., Warriner et al., 2013), we took into account two factors to ensure that all raters were native speakers of Mandarin Chinese. Specifically, in addition to self-identification, the participants also indicated whether they had spent most of the first seven years of their lives in mainland China. All but two participants, who were excluded from data analysis, were native speakers of Mandarin Chinese. Based on the data screening criteria (see Results section), we further excluded 526 participants from data analysis. In the end, there were 2421 participants included to calculate mean valence ratings (N = 1232, 55.1% women) and mean arousal ratings (N = 1189, 54.2% women). Their age ranged from 18 to 62. Education level ranged from middle school to graduate school, with 96.9% college-level or above college-level education. Figure 2 illustrates the geographic distribution of the raters for the two ratings tasks.

Word sample

The word sample was retrieved from the MEgastudy of Lexical Decision in Simplified CHinese (MELD-SCH; Tsang et al., 2018). Tsang et al. (2018) sampled 20,000 one- to four-character items from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). After screening proper nouns, their sample consisted of 12,578 items, including 1020 characters: 10,022 two-character words, 949 three-character words, and 587 four-character words. In two previous studies, we reported age-of-acquisition ratings (Xu et al., 2020) and concreteness ratings (Xu & Li, 2020) for most words contained in this database. In this study, we chose the MELD-SCH again for affective ratings to further enrich the database and to make it possible to investigate the properties and relations of these psycholinguistic constructs based on ratings of a large set of Chinese words.

In the previous study (Xu et al., 2020), we found that many of the characters in the MELD-SCH were either extremely uncommon or nonwords, including characters that are rarely seen or used (e.g., 蜱, 胍, and 潞), characters that seem to only appear in Classical Chinese (e.g., 炷, 孛, and 乜) or in translations of foreign names (e.g., 圭, 堺, and 樋), and characters that are typically used in combination with other characters to form a word (e.g., 蒟, 珐, and 噌). In addition, as is well known, lexical ambiguity is pervasive among Chinese characters (e.g., Liu et al., 2007), which would negatively impact rating reliability. We therefore excluded characters from our sample.

Next, we further removed the following: (1) homographs that are obviously ambiguous in pronunciation and in meaning, e.g., 穿着 "outfit" as a noun or "wearing" as a verb, (2) nonwords, e.g., 那是 "that is," and (3) words unknown to five or more participants in the other two rating studies conducted in our lab (Xu et al., 2020; Xu & Li, 2020). In the end, there were 11,306 words that remained.

Among the remaining words, we identified 21 two-character words that denote emotions, e.g., 快乐 "happy" and 害怕 "afraid," corresponding to some of the English emotion words extensively researched in the literature of affect representation (e.g., Russell, 1980; Russell et al., 1989; Tsai et al., 2006). Based on the lists of emotion words from these prior studies, we added four more words (气愤 "angry," 痛苦 "distressed," 厌烦 "bored," and 抑郁 "depressed"). This list of 25 emotion words served to enhance rating reliability in our study (see Procedure section). As a result, the final word sample consisted of 11,310 words, including 9774 two-character words, 949 three-character words, and 587 four-character words.

Procedure

We consulted previous rating studies conducted both in Mandarin Chinese (e.g., Wang et al., 2008; Yao et al., 2017; Yee, 2017) and in English (e.g., Bradley & Lang, 1999; Warriner et al., 2013), and adopted the essence of the instructions from these studies. Specifically, the instructions for the valence rating task were first illustrated with examples of positively valenced words (e.g., 金牌 "gold medal") versus negatively valenced words (e.g., 勾当 "criminal dealing"). Similarly, the instructions for the arousal rating task described high-arousal words (e.g., 台风 "typhoon") versus low-arousal words (e.g., 文书 "paperwork"). The instructions further explained to the participants that word valence or word arousal

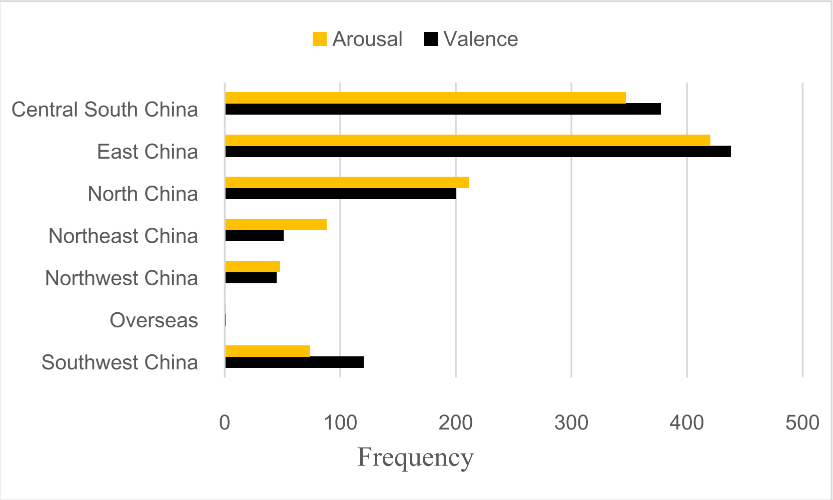

Fig. 2 Geographic distribution of the raters for valence and arousal

can vary on a continuum, with some words falling in between the two extremes. Then, the numerical rating scales were introduced, on which participants indicated their assessment of valence or arousal. For valence ratings, participants were asked to rate on a seven-point scale provided below each word, where "−3" was labeled as "extremely negative," "0" as "neutral," and "+3" as "extremely positive." For the arousal rating task, participants were asked to rate on a five-point scale, where "0" was labeled as "very low arousal" and "4" as "very high arousal." To enhance rating validity for both tasks, an additional option "N" was also provided for each word in case that any participants felt that they did not know the meaning of the word (e.g., Brysbaert et al., 2014). Finally, the instructions stated that there would be no right or wrong answers, and that they should make a quick assessment based on their first reaction upon seeing the word.

For the following reasons, the valence and arousal ratings scales in the present study were different from the 9-point scale, with numerical markers of 1 through 9, utilized by most previous affective norming studies. First, citing Laming (2004), Brysbaert et al. (2014) argued that it could be difficult for participants to consistently differentiate more than five categories. We therefore reduced the number of scale markers from nine to seven and five for valence and arousal, respectively. Second, as shown by previous research, the construct of valence appeared to be bipolar in nature, more so than arousal (more on this point in the Discussion section). When rating valence, participants might go through a two-step process: judging polarity, and then assessing degree of positivity or negativity. Therefore, a seven-point valence scale with polarity clearly marked should be able to provide sufficient resolving power without overtaxing participants' working memory. The third and additional reason to mark the valence scale with both positive and negative numbers was to invoke participants' numerical knowledge and ensure correct application of the scale. Warriner et al. (2013) reported that, with the 9-point scale marked with all positive numbers ranging from 1 to 9, some participants' ratings had to be reversed, as they correlated negatively with the mean ratings, an indication that they had misapplied the scale.

The 25 emotion words in our word sample were presented to every participant. The remaining 11,285 words were divided into 48 lists of 235–236 words, roughly matching on word frequency retrieved from the SUBTLEX-CH corpus (Cai & Brysbaert, 2010). Each participant was randomly assigned to complete a word list in addition to the 25 emotion words. The purpose of repeating the emotion words across word lists was twofold. First, as there were a large number of words on each list, participants' rating criteria might shift along the way. The valence and arousal values of these emotion words had been researched and established for different languages including English, Estonian, Greek, Polish, and Mandarin Chinese (e.g., Russell et al., 1989; Yik, 2009; Yik & Russell, 2003). As a result, these randomly embedded emotion words would help participants recalibrate their criteria during the process, and thus increase rating reliability. Second, and related, calculating inter-list correlations based on ratings of these words would help us assess and determine rating reliability and validity.

The total of 260–261 words on each word list were presented in random order to each participant. Participants rated either valence or arousal. We asked them to provide a response to each word on the list. However, they were free to withdraw from the study at any time that they chose to do so. After the word list, demographic questions, presented in a fixed order, asked participants to report gender, age, education level, native language, and the place where they spent most of the first seven years of their lives (Warriner et al., 2013). It took the participants on average 26 minutes to complete the valence rating task and 29 minutes to complete the arousal rating task.

Results

Data screening

We tabulated the frequencies of all possible responses, including −3 to +3 for valence rating and 0–4 for arousal rating, as well as the response "N" (i.e., "I don't know the word"). First, we removed 203 (14.1% of 1444) participants in the valence rating task and 100 (6.6% of 1505) participants in the arousal rating task who had 15% or more "N" responses, which demonstrated either inattentiveness to the task or limited vocabulary knowledge unfit to provide credible assessment (Yee, 2017). Next, we removed five (0.3% of 1444) participants of the valence rating task and four (0.3% of 1505) participants of the arousal rating task who had 85% or more same responses across the entire word list, as such a low variation in their responses suggested noncompliance with the instructions (Yao et al., 2017). Finally, we removed four (0.3% of 1444) participants of the valence rating task and 210 (14.0% of 1505) participants of the arousal rating task, whose ratings correlated poorly (<. 10) with the rest of the participants assigned to the same word list (Warriner et al., 2013). As indicated earlier, the criteria resulted in the removal of a total of 526 participants, or 17.8% of the original 2949 participants.

Before we counted the number of valid ratings and calculated the mean valence and arousal rating for each word, we removed all "N" responses ($n = 16{,}526$), or 2.6% of the total ratings from the 2421 participants. As a result, the total number of valid ratings was 613,180 (311,252 valence ratings and 301,928 arousal ratings), and the number of valid ratings for each word ranged from 13 to 34 (15–34 for valence and 13–31 for arousal). Excluding the 25 emotion words repeated across lists, words with 18 or more valid ratings accounted for more than 99.9% and 97.9% of the remaining 11,285 words for the valence rating task and the arousal rating task, respectively. Only 11 words had 17 or less valid valence ratings, whereas 234 words had 17 or less valid arousal ratings. We provide the number of valid ratings for each word in the database. Mean valence and arousal ratings were then computed for each word.

Table 1 summarizes central tendency and variability of mean valence and arousal ratings of the 11,310 words. In addition to ratings based on all participants, we also calculated ratings of men and women, respectively (Table 1). In the literature, there has been some evidence, albeit inconsistent, for gender differences in valence and arousal ratings. The inconsistency might be attributable to the fact that, after breaking down by gender, the ratings became less stable due to much smaller numbers of valid ratings for some words. We therefore ran paired-sample t-tests on words that received 10 or more valid ratings from each gender group (N = 7929 for analysis of valence ratings; N = 6698 for analysis of arousal ratings). Results showed that women (M = .11, SD = 1.03) rated the words more positively than did men (M = .10, SD = 1.01), t (7928) = 4.30, $p < .001$, whereas men (M = 2.14, SD = .59) rated words higher in arousal than did women (M = 2.09, S = .58), t (6697) = 9.40, $p < .001$. However, the magnitudes of gender differences in both valence and arousal ratings were quite small.

Furthermore, a finding relatively consistent across languages is that, compared to men, women tended to rate positive words more positively, and negative words more negatively (Monnier & Syssau, 2014 in French; Riegel et al., 2015 in Polish; Soares et al., 2012 in European Portuguese). We found the same results. For positive words (i.e., overall mean valence ratings greater than 0; N = 5124), women assessed significantly higher ratings (M = .71, SD = .59) than did men (M = .67, SD = .57), t (5123) = 7.58, $p < .001$. For negative words (N = 2660), women assess significantly lower ratings (M = −1.03, SD = .72) than did men (M = −1.01, SD = .72), t (2659) = 2.21, $p = .03$. Again, the magnitudes of these differences were small. For the remaining 145 neutral words with an overall mean valence rating of 0, there was no gender difference, $|t| < 0$. We provide the mean valence and arousal ratings of men and women in our database. Numbers of valid ratings are included for future researchers to use these ratings at their own discretion.

Data analysis

Reliability and validity To assess reliability, we first calculated inter-rater reliabilities of valence and arousal ratings for each of the 48 word lists. For valence ratings, Cronbach's alphas were highly desirable, ranging from .97 to .99, with a mean of .98 (SD = .01). For arousal ratings, Cronbach's alphas ranged from .79 to .92, with a mean of .86 (SD = .03). Next, we computed split-half reliabilities. Correlations between ratings of odd- and even-numbered participants were .95 and .74, which yielded split-half reliabilities of .97 and .85 for valence and arousal ratings, respectively. Finally, we also assessed inter-list consistency by evaluating ratings of the 25 emotion words that were repeated across lists. Correlation coefficients of valence ratings for the 25 emotion words across the 48 word lists ranged from .98 to 1.00 (mean = .99, SD = .003). For arousal ratings, they ranged from .74 to .98 (mean = .91, SD = .04). In sum, both valence ratings and arousal ratings demonstrated good reliabilities. Consistent with past reports from research on various languages (e.g., Spanish by Guasch et al., 2016; French by Monnier & Syssau, 2014; English by Warriner et al., 2013; Chinese by Yee, 2017), reliabilities of valence ratings were more desirable than arousal ratings. We will return to this point when discussing characteristics of valence versus arousal ratings.

To evaluate validity, we compared the present valence and arousal ratings with ratings collected in two past studies on Chinese words (Wang et al., 2008; Yee, 2017). There were 996 and 184 words on our word lists in common with Wang et al. and Yee, respectively. Table 2 presents correlation

Table 1 Summary of valence and arousal ratings of 11,310 words

		Overall		Men		Women	
		Valence	Arousal	Valence	Arousal	Valence	Arousal
Mean		0.11	2.08	0.10	2.11	0.11	2.05
Median		0.22	2.05	0.20	2.10	0.21	2.00
Mode		0.00	2.00	0.00	2.00	0.00	2.00
Standard deviation		0.97	0.55	0.98	0.60	1.00	0.59
Min		−2.88	0.61	−3.00	0.29	−3.00	0.29
Max		2.68	3.81	2.90	3.88	2.75	3.91
Percentile	25th	−0.32	1.67	−0.36	1.67	−0.33	1.62
	75th	0.65	2.46	0.67	2.55	0.69	2.46

Note. Valence was rated on a 7-point scale (−3 to +3), and arousal on a 5-point scale (0–4).

Table 2 Correlations of valence and arousal ratings between the present study and two prior studies

	Valence	Arousal	*N*
Valence (Wang et al., 2008)	.933**	−.056	996
Arousal (Wang et al., 2008)	−.254**	.751**	996
Valence (Yee, 2017)	.868**	.104	184
Arousal (Yee, 2017)	−.192**	.608**	184

** $p < .01$ (two-tailed)

coefficients of ratings between studies, which support the validity of the present ratings. Based on Steiger's (1980) method of testing differences between correlation coefficients, the present study showed a significantly greater alignment with past studies in terms of valence ratings than arousal ratings, $z = 15.6$ for comparison with Wang et al. (2008) and $z = 6.04$ for comparison with Yee (2017). Figures 3 and 4 illustrate these correlations.

Relations of valence, arousal, and other lexical and semantic variables We examined the relation of valence and arousal ratings, and their relations with SUBTLEX-CH word frequency (Cai & Brysbaert, 2010), zRT (i.e., standardized reaction time) and error rate retrieved from the MELD-SCH (Tsang et al., 2018). In addition, we analyzed the relations of the valence and arousal ratings with age-of-acquisition ratings (Xu et al., 2020) and concreteness ratings (Xu & Li, 2020) that recently became available for large portions of words in the MELD-SCH database. Table 3 presents the results of bivariate correlation analysis on these variables.

Figure 5 shows a curvilinear relationship between valence and arousal ratings, which is consistent with reports from research on other languages, including English (Bradley & Lang, 1999; Warriner et al., 2013), French (Monnier & Syssau, 2014), Polish (Riegel et al., 2015), Spanish (Stadthagen-Gonzalez et al., 2017), Turkish (Kapucu et al., 2021), etc. That is, the strength of a word's valence, be it negative or positive, was associated with its level of arousal. The quadratic term of valence was significantly correlated with arousal, $r = .52$, $p < .0001$. Correlations of valence and arousal with other variables were modest. As an illustration, Fig. 6 shows the strongest among them, i.e., the correlation between arousal and concreteness, $r = .20$, $p < .001$.

We also assessed the degree to which valence and arousal could contribute to the efficiency of lexical processing as measured by zRT and error rate. Regression analyses indicated that, similar to another semantic variable, i.e., concreteness, neither valence nor arousal made a substantial contribution to predicting zRT or error rate. Valence (or its quadratic term) accounted for less than 1% and arousal approximately 1% of additional variance above and beyond word frequency and age-of-acquisition, two variables reliably predictive of performance on lexical decision tasks (Kuperman et al., 2012; Xu et al., 2020). These findings were consistent with similar analysis on a sample of 12,658 English words where, after controlling for other factors, valence and arousal together accounted for approximately 2% of the variance in lexical decision latency (Kuperman et al., 2014). We had reported the details of the regression models on zRT and error rate in two recent studies (Xu et al., 2020; Xu & Li, 2020). For the sake of succinctness, they were omitted from the current report.

Rating variabilities of valence, arousal, concreteness, and age-of-acquisition scales As indicated above, concreteness, age-of-acquisition, valence, and arousal ratings are now available for a large portion of words in the MELD-SCH database. All were collected with numeric rating scales constructed based on commonly accepted conceptualizations of these variables. Pollock (2018) recently conducted an in-depth investigation into the concreteness numeric rating scale, and highlighted the importance to evaluate rating variability among individual raters. Specifically, he showed that words from the middle

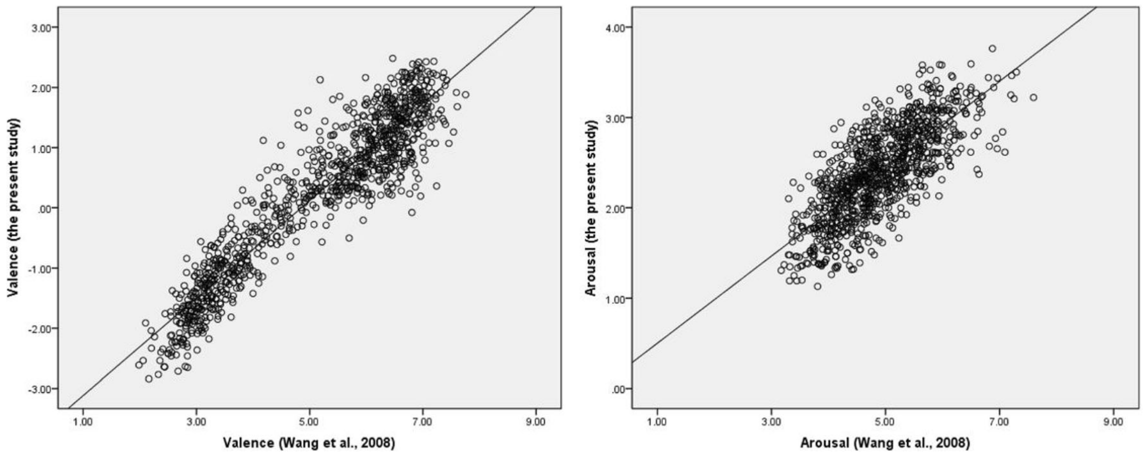

Fig. 3 Cross-study consistency of valence ratings versus arousal ratings between the present study and Wang et al. (2008)

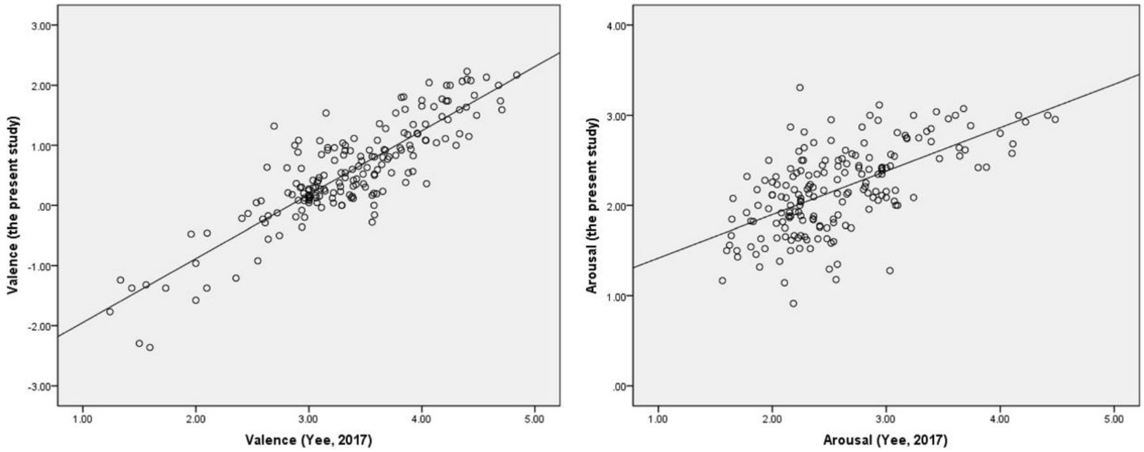

Fig. 4 Cross-study consistency of valence ratings versus arousal ratings between the present study and Yee (2017)

portion of the concreteness scale had higher rating variabilities than words from the two ends (very concrete and very abstract), and cautioned against utilizing words from the middle portion of the scale while investigating representational and processing differences between concrete and abstract words. We corroborated Pollock's (2018) observation about variability of concreteness ratings with Chinese words (Xu & Li, 2020). Below, we took the same approach to examine and compare rating variabilities for valence, arousal, concreteness, and age-of-acquisition in order to gain more insight into these psycholinguistic constructs.

To place the variables on an even ground for comparison, the following plots and analysis were generated based on 9770 words in the MELD-SCH, for which all four types of ratings are available. Figure 7 shows that, much like the previous report with a slightly larger sample of words (N = 9877; Xu & Li, 2020), ratings of words in the middle portion of the concreteness/abstractness scale had higher standard deviations (SDs) than those from either end of the scale. That is, there were words more consistently perceived to be either concrete or abstract by the raters, whereas there were other words that did not seem to evoke the same level of consensus.

Figure 8 displays the change of rating SDs along the valence scale. Similar to the concreteness scale, many words from the middle portion of the scale had high rating variabilities. However, unlike the concreteness scale, a cluster of words near the center of the scale showed high levels of rating consistency. In fact, of the 11,310 words in our database, the 29 words with a rating SD of zero were all located at the midpoint (0) of the valence scale.

Figure 9 exhibits rating SDs along the arousal scale. Again, words in the middle portion of the scale seemed to show high

Table 3 Bivariate correlations between valence, arousal, and other variables

		1	2	3	4	5	6
1. Valence							
2. Arousal	Pearson's *r*	−.100**					
	N	11,310					
3. Frequency (log)	Pearson's *r*	.030**	.142**				
	N	11,310	11,310				
4. zRT	Pearson's *r*	−.103**	−.154**	−.613**			
	N	11,305	11,305	11,305			
5. Error	Pearson's *r*	−.031**	−.154**	−.400**	.682**		
	N	11,305	11,305	11,305	11,305		
6. AoA	Pearson's *r*	−.083**	.090**	−.381**	.408**	.284**	
	N	11,309	11,309	11,309	11,305	11,305	
7. Concreteness	Pearson's *r*	.008	.202**	−.009	.085**	.082**	.356**
	N	9770	9770	9770	9769	9769	9770

** $p < .01$ (two-tailed). Frequency(log): logarithmic-transformed word frequency (Cai & Brysbaert, 2010); zRT: standardized reaction time (Tsang et al., 2018). Error: error rate (Tsang et al., 2018); AoA: age-of-acquisition ratings (Xu et al., 2020); Concreteness: concreteness ratings (Xu & Li, 2020).

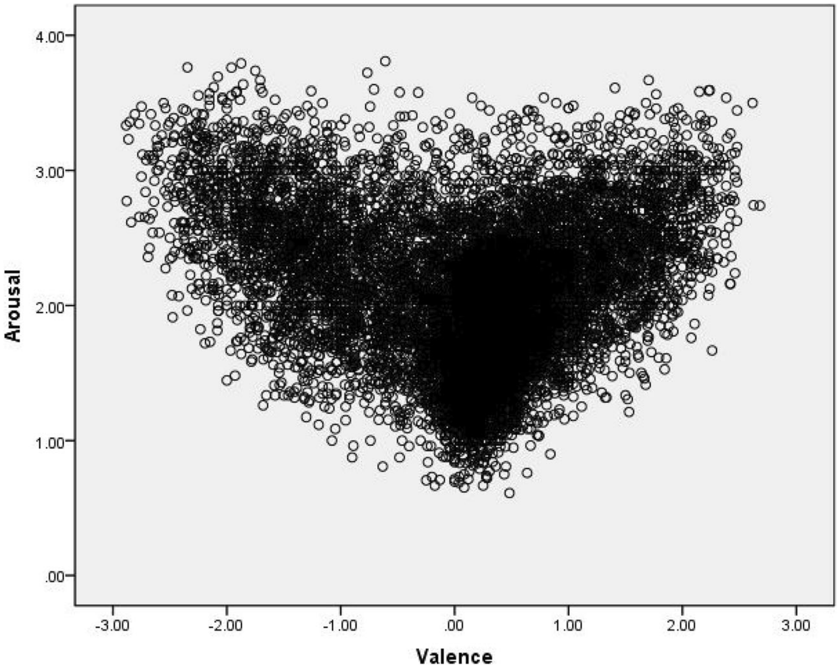

Fig. 5 Relationship of valence and arousal ratings

levels of rating variabilities. However, the distribution appeared conspicuously different from that of the concreteness/abstractness ratings. Whereas rating variabilities of concreteness/abstractness demonstrated a gradual change from one extreme to the other, rating variabilities of arousal seemed generally large for most words, and only a small portion of words close to the top of the scale were relatively consistently rated to be high in arousal.

Finally, Fig. 10 plots rating SDs of age-of-acquisition. Different from concreteness, valence, and arousal, there appeared to be a linear pattern, $r = .27$, $p < .0001$. Words acquired at younger ages showed the lowest rating variabilities, which increased for words acquired at older ages.

Comparing affect representations in Chinese versus English

We examined the valence and arousal ratings of Chinese words versus English words in order to explore potential cultural similarities and differences in mental representations of affect. First, we plotted the 25 emotion words on our list in a two-dimensional plane (valence and arousal) to compare with Russell's circumplex model of affect (1980). Russell (1980) conducted a series of analyses on 28 English emotion words, and all solutions yielded from these analyses indicated that, among native speakers of English, affective experiences appeared to be systematically organized in a two-dimensional space. Figure 11 shows Russell's model generated based on a multidimensional scaling approach and a unidimensional

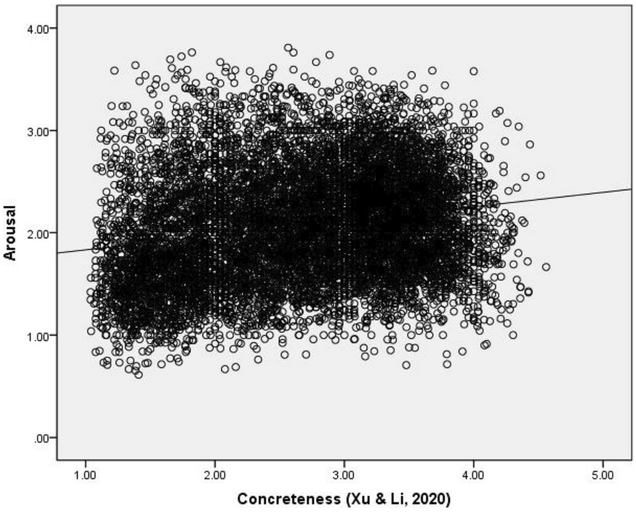

Fig. 6 Relationship of concreteness (1 = "very concrete"; 5 = "very abstract") and arousal ratings

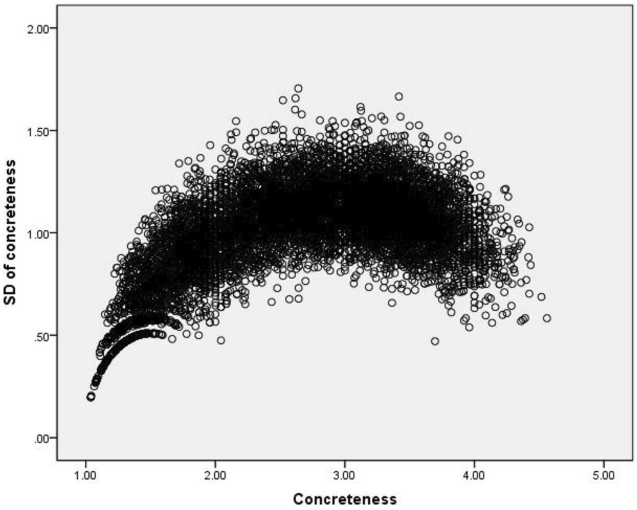

Fig. 7 Variability of concreteness ratings on a 5-point scale

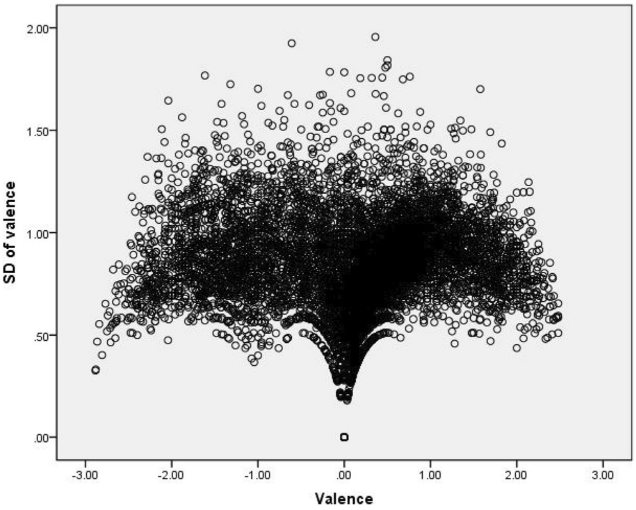

Fig. 8 Variability of valence ratings on a 7-point scale

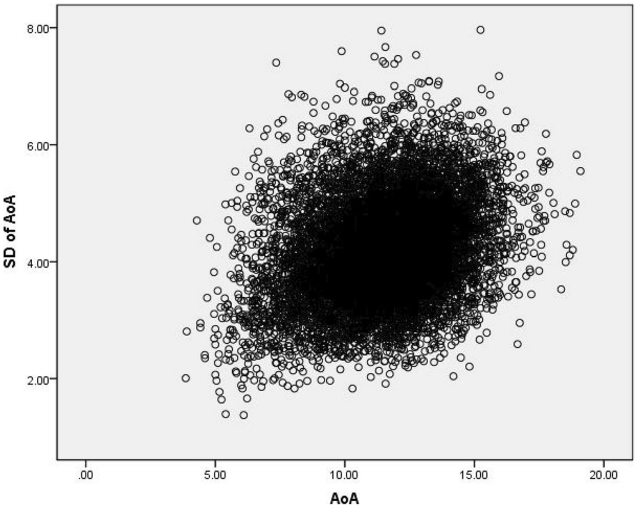

Fig. 10 Variability of age-of-acquisition ratings

scaling approach. Figure 12 shows, in the present study, the two-dimensional distribution of 25 English equivalents of Chinese emotion words based on their valence and arousal ratings. The layouts of the 24 common words between the two studies roughly matched, indicating similar representations of emotion concepts between speakers of the two languages. To quantitatively verify this similarity, we also ran bivariate correlations on these 25 emotion words with valence and arousal ratings retrieved from the present study and Warriner et al. (2013). The results corroborated what was demonstrated by Figs. 11 and 12, r (25) = .90, $p < .001$ for valence and r (25) = .70, $p < .001$ for arousal. (Note that Warriner et al. does not contain ratings for *alarmed* or *at ease*, which we replaced with ratings of their close equivalents, *frightened* and *comfortable*, respectively, for the correlation analyses.)

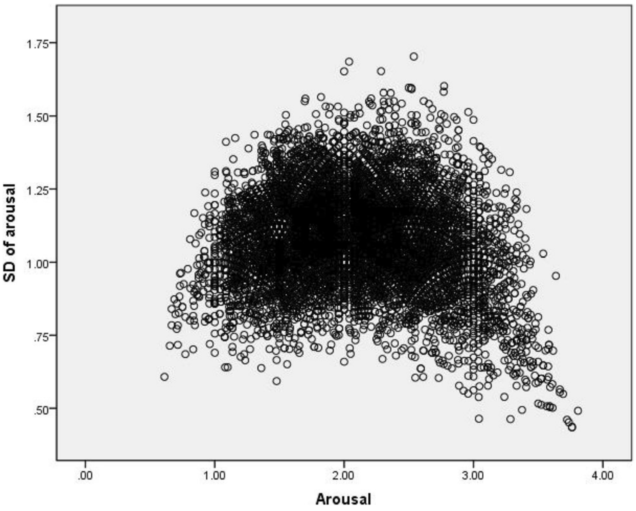

Fig. 9 Variability of arousal ratings on a 5-point scale

Next, we examined the Chinese words and the English words that fell on the two extremes of the valence rating scales. We retrieved Chinese words with the ten highest valence ratings and the ten lowest valence ratings on the seven-point scale utilized in the present study. They are listed in Table 4 along with English translations. We initially also retrieved English words with the ten highest and the ten lowest ratings on the 9.0 valence scale utilized by Warriner et al. (2013). However, we found that, whereas all Chinese words listed in Table 4 had rating SDs less than .70, four of the 20 English words (i.e., *happy*, *Christmas*, *free*, and *leukemia*) had rather high rating variabilities, with SDs ranging from 1.28 to 1.61, and the remaining 16 words had more desirable rating variabilities, with SDs less than 1.00. To ensure that the words we chose to compare were reflective of a reasonable level of raters' consensus on affect evaluation, we replaced these four English words with words that had the next highest ratings (for *happy*, *Christmas*, and *free*) or the next lowest rating (for *leukemia*) with SDs less than 1.00 (Table 4).

At the positive end of the scale, the top-rating English words were mainly positive emotion words (e.g., *happiness*, *enjoyment*, and *delight*) and an event or action associated with the emotions (e.g., *vacation* and *hug*). In contrast, the top-rating Chinese words appeared more diverse, including virtues (e.g., *act heroically for what is right*, *filial piety*, and *reverence*), achievements (e.g., *success*, *bright future*, and *triumphant return*), national ideals and aspirations (e.g., *the country prospers; the people are at peace* and *(government officials) love the people like children*), and medical professional (i.e., *angel in white*). At the negative end of the scale, the English words could be placed into two categories: hideous crimes, particularly sex crimes, as well as the person who commits the crime (e.g., *pedophile*, *rapist*, and *murder*), and potentially deadly diseases, symptoms, and treatment (e.g., *AIDS*, *chemo*, and *asphyxiation*). The Chinese words also

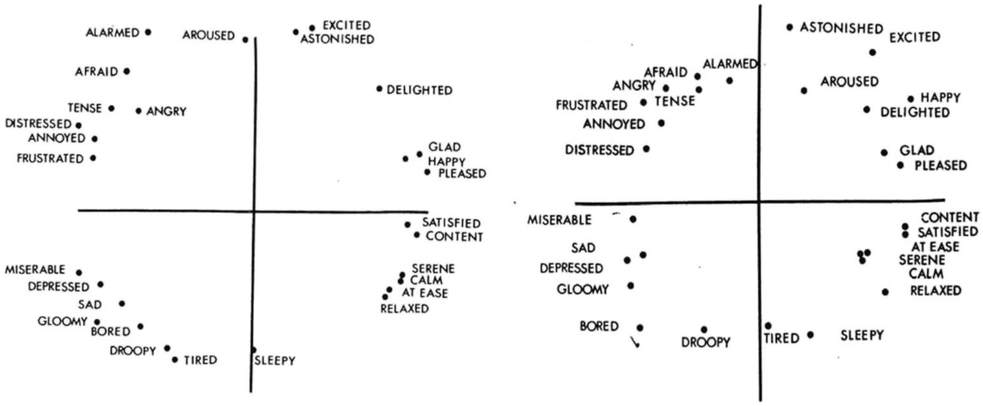

Fig. 11 Two-dimensional distributions of emotion words presented in Figure 3 (left; outcome of a multidimensional scaling approach) and in Figure 4 (right; outcome of a unidimensional scaling approach) by Russell (1980)

included hideous crimes and the person who commits the crime. However, instead of disease-related words, there were words referring to shameful action or crime against the nation and the person who commits the crime (e.g., *forfeit sovereignty and humiliate the country* and *treasonous traitor*).

Finally, we had tried to compare the English and the Chinese words that fell on the two extremes of the arousal rating scales, but found that it seemed less meaningful as mean arousal ratings were generally high in variabilities and thus less representative of affect evaluation at a collective level. This was the case for both English words and Chinese words. The SDs of the English words with the ten highest and the ten lowest arousal ratings ranged from .88 to 2.47 on a 9.0 scale, while the SDs of the Chinese words with the ten highest and the ten lowest arousal ratings ranged from .41 to .95 on a five-point scale. However, both mean ratings of arousal and SDs are provided for Chinese words in our database and for English words in the database by Warriner et al. (2013) for interested users to conduct similar cross-language comparisons.

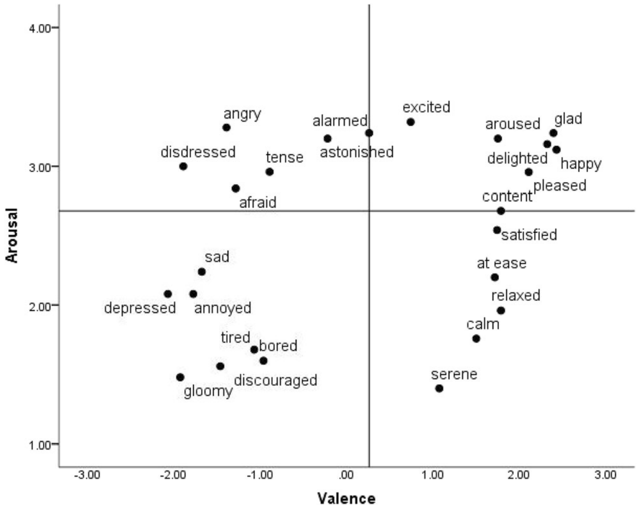

Fig. 12 Two-dimensional distribution of the 25 emotion words in the present study. (The plane is divided into four regions by the medians of valence and arousal ratings.)

Discussion

This study collected valence and arousal ratings for 11,310 simplified Chinese words. With these ratings, we examined representational properties of valence and arousal in comparisons to other semantic variables. In addition, we evaluated the words that denote emotions and the words rated to be most positively/negatively valenced by Mandarin Chinese speakers versus English speakers to explore cross-language differences in affect representation.

Representational properties of valence and arousal

Relative to the arousal ratings, the valence ratings in the present study demonstrated greater inter-rater reliability, greater split-half reliability, and greater alignment with the valence ratings collected by previous studies on Chinese words. That is, valence of Chinese words has consistently shown, relative to arousal, a greater level of rater consensus, i.e., a lower level of rating variability, which is in line with reports on other languages (e.g., Guasch, Ferré, & Fraga, 2016; Monnier & Syssau, 2014; Warriner et al., 2013). These findings suggest that, between the two primary theoretical dimensions of affect representation, valence seems to be a better defined and more salient feature of our affective experiences than arousal, which should be attributable to the significance of valence evaluation to our physical and psychological well-being. That is, we constantly evaluate and judge, either consciously or subconsciously, objects and events surrounding us in terms of their values with respect to our interests in order to guide our thoughts and actions, particularly to avoid harm and danger and to protect safety and health. It is this bipolar nature of the valence scale that renders valence ratings across studies a greater level of consistency. Specifically, people generally

Table 4 Highly valenced words in Chinese versus in English

Highly valenced words in Chinese				Highly valenced words in English		
Word	English translation	Valence rating	SD	Word	Valence rating	SD
Positive						
国泰民安	The country prospers; the people are at peace	2.68	.69	Vacation	8.53	.77
前程似锦	Bright future	2.63	.49	Happiness	8.48	.81
见义勇为	Act heroically for what is right	2.62	.57	Enjoyment	8.37	.96
鼓励	Encourage	2.48	.51	Fun	8.37	.96
成功	Success	2.48	.59	Fantastic	8.36	.79
孝敬	Filial piety	2.48	.65	Lovable	8.26	.99
敬重	Reverence	2.48	.59	Hug	8.23	.87
爱民如子[a]	(Government officials) love the people like children	2.46	.51	Magical	8.23	.97
白衣天使[a]	Angel in white (meaning “medical professional”)	2.46	.58	Delight	8.21	.92
凯旋[a]	Triumphant return	2.46	.58	Joyful	8.21	.98
长命百岁[a]	Longevity	2.46	.71			
Negative						
轮奸	Gang rape	−2.88	.33	Pedophile	1.26	.65
弑父	Father killing	−2.88	.33	Rapist	1.30	.73
拐卖	Human trafficking	−2.86	.47	AIDS	1.33	.80
杀人犯	Murderer	−2.85	.60	Torture	1.40	.82
贩毒	Drug trafficking	−2.84	.55	Racism	1.48	.68
强奸	Rape	−2.81	.40	Murder	1.48	.81
丧权辱国	Forfeit sovereignty and humiliate the country	−2.77	.51	Molester	1.48	.98
虐待	Abuse	−2.77	.65	Homicide	1.50	.92
卖国贼	Treasonous traitor	−2.75	.51	Chemo	1.50	.95
荡妇	Slut	−2.75	.52	Asphyxiation[b]	1.53	.84
				HIV [b]	1.53	.90

[a] The 8th to the 11th Chinese words at the high end of the valence scale had the same mean ratings. [b] The 10th and the 11th English words at the low end of the valence scale had the same mean ratings. SD: standard deviation of valence ratings. Highly valenced English words along with their valence ratings and standard deviations were retrieved from Warriner et al. (2013).

agree on what should be judged to be positive versus negative, for example, what can cause loss versus what may bring profit, even though they may not agree on the degree of positivity or negativity entailed by an object or event. In contrast, the arousal scale essentially prompts raters to indicate the degree to which they perceive an object or event to be arousing. Individuals naturally differ in their perceptions of arousal due to differences in their interests, temperament, their current state of mind, experiences, and many other aspects of their lives.

We do not believe that this difference between valence and arousal ratings in the present study should be attributed to the different methods of scale construction for the two variables. Specifically, in the present study, while the valence rating scale varied from −3 to +3, the arousal rating scale ranged from 0 to 4, which might have indicated or even exaggerated to the raters the bipolar nature of the valence scale compared to the arousal scale. However, in many past studies cited above that revealed similar contrasts between valence and arousal, researchers utilized the same numeric rating scale for both variables. For example, Yee (2017) utilized a five-point scale (1–5), and Warriner et al. (2013) a nine-point scale (1–9). In addition, both studies adopted a slightly modified version of the rating instructions by Bradley and Lang (1999), which exemplified and contrasted the two extremes of the valence and arousal scales with many pairs of antonyms, e.g., *happy* and *unhappy*, *hopeful* and *despaired*, *excited* and *bored*, *aroused* and *unaroused*, etc. Therefore, the greater rating variabilities of arousal relative to valence revealed in the present and past studies seem to reflect inherent representational differences between the two variables.

The differences between the two variables are further revealed by comparing variability distribution along their respective rating scales. As presented earlier in the Results section, two recent studies have shown that rating variabilities along the concreteness/abstractness rating scale exhibit an arch-shaped distribution (Pollock, 2018; Xu & Li, 2020). From the concrete extreme of the scale (e.g., *rabbit*, *egg*), rating variabilities gradually increase with the decrease of concreteness, reach a plateau in the middle of the scale (e.g., *Yama*, *zombie*), and then gradually decrease with the increase of abstractness (e.g., *ideal*, *destiny*). Plotting rating variabilities along the valence scale, however, reveals a bouquet-shaped distribution. A cluster of words at the center (0) of the scale boast the highest level of rater consensus, even higher than words close to the extremes of the scale, which are more or less consistently considered to be positive or negative. That is, there are words that people clearly perceive to be neutral, for example, *diameter*, *location*, and *appearance*, which is consistent with what was revealed by the valence ratings of English words (Warriner et al., 2013). In contrast, distribution of rating variabilities along the arousal scale does not reveal such rater consensus with regard to words in the middle of the scale, and people seem to agree, to an extent, that a small cluster of words close to the high end of the scale to invoke high arousal, for example, *infectious disease*, *murder*, and *anger*. Further, no words received a mean arousal rating of 0, representing little or no arousal, and the lowest mean arousal rating was above .60. In contrast, as indicated earlier, some words were reliably rated to be neutral (mean valence ratings of 0) by 100% raters, which attests that valence and arousal should be considered two separate constructs despite the correlation between strength of valence and level of arousal. Pollock (2018) pointed out the importance of taking into account rating variabilities in word sampling and statistical analysis. More importantly, we think that analysis of rating variabilities helps to enhance our understanding about the psychological reality of these theoretical constructs. Further supporting this point, unlike all three variables above, i.e., concreteness/abstractness, valence, and arousal, the rating variabilities of AoA display a linear pattern. They increase with the increase of AoA, reflecting the presumptive outcome of cognitive constraints at earlier stages and diversified trajectories at later stages of vocabulary development among individuals.

Cross-language comparison of affect representations

Analyses of valence and arousal ratings in the present study and past studies have revealed some common properties of affect representations between Mandarin Chinese speakers and English speakers. First, valence seems a more salient and clearly defined feature of affective experiences, evidenced by its greater rating validity and reliability than arousal (Warriner et al., 2013; Yee, 2017). Second, there is a curvilinear relationship between valence and arousal, with greater strength of valence being associated with higher arousal (Warriner et al., 2013). Finally, speakers of both languages seem to similarly organize common emotion words in a semantic space where valence and arousal emerge as two primary dimensions (Russell, 1980). These commonalities not only lend support to expanding theories on affect representations across languages, but also lay the necessary foundation for clinical practice, with proper validations, to transfer effective intervention programs for affect disorders, e.g., alexithymia, from one language to another.

Despite similar conceptualizations of common emotion words between speakers of the two languages, an analysis contrasting valence assessment of specific Chinese versus English words has found some cross-language differences. Specifically, the top ten positive English words mainly contain emotion words (e.g., *happiness*), whereas the negative English words contain several illness-related words, often life-threatening illness (e.g., *AIDS*). In contrast, both the highly positive and the highly negative Chinese words contain words concerning national interests and dignity (e.g., *the country prospers*; *the people at peace*, and *forfeit sovereignty and humiliate the country*), which are missing from the list of highly valenced English words. On the one hand, this contrast can be considered evidence for the widely accepted notion of individualism versus collectivism between the cultures (Triandis, 1988). On the other hand, these top-rating Chinese words expressing strong patriotic sentiments may also be a reflection of the lasting impacts left by the nation's modern history on its people's collective conscience. Similarly, the word *racism* is one of the top-rating negative English words, reflecting a heightened level of awareness of this social justice issue among English speakers.

Further, some of the highly valenced Chinese and English words seem to represent events experienced or challenges faced by the participants when the studies were conducted. For example, *angel in white*, nickname for "medical professional," appears among the top-rating positive Chinese words, likely a product of the fact that the present study was conducted in the era of a pandemic. Likewise, the word *torture* appears among the top-rating negative English words reported by Warriner et al. (2013). At the time of the study, the word had been in the discourse of social, political, and legal debates for a few years following the exposure, investigation, and policy change in the U.S. about "enhanced interrogation techniques" (Apuzzo et al., 2014; Mayer, 2009).

To make sure that the aforementioned differences were not simply due to difference in word sampling between studies, we examined the word lists of the present study and Warriner et al. (2013). Most top-rating positive and negative words in one study could find their equivalents or close equivalents in the other study. For example, there were plenty of positive emotion words (e.g., 快乐 *joyful*) and illness-related words (e.g., 癌症 *cancer*) on the Chinese word list. The word 种族歧视 (*racial discrimination*) can also be found in the present study. Likewise, there were many words on the English word list representing virtues (e.g., *reverence*) and achievements (e.g., *triumph*), as well as nouns referring to medical professionals in Warriner et al. One possible exception was that many of the four-character Chinese words,

Table 5 Valence rankings of highly valenced English words versus valence rankings of their Chinese translations

Highly valenced English word	Ranking in English	Chinese translation	Valence rating in Chinese	Ranking in Chinese
Positive				
Vacation	1	度假	1.5769	669
Happiness	2	幸福	2.2917	45
Enjoyment	3	享受	1.0000	1684
Fun	3	乐趣	1.9583	235
Fantastic	5	奇妙	1.2692	1155
Lovable	6	可爱	1.9615	223
Hug	7	拥抱	1.2308	1217
Magical	7	奇幻[a]	0.8966	1995
Delight	9	开心	2.3214	36
Joyful	9	快乐	2.4286	14
Negative				
Pedophile	1	恋童癖[b]	-	-
Rapist	2	强奸[a]	−2.8077	6
AIDS	3	艾滋病[b]	-	-
Torture	4	酷刑	−2.1538	231
Racism	5	种族歧视[a]	−2.5714	36
Murder	5	谋杀	−2.5385	39
Molester	5	性骚扰[a]	−2.6000	32
Homicide	8	凶杀案[a]	−2.6818	16
Chemo	8	癌症[a]	−2.2000	194
Asphyxiation	10	窒息	−2.0385	314
HIV	10	艾滋病毒[b]	-	-

[a] Translation close equivalent; [b] Chinese valence rating is not available for translation equivalent or close equivalent.

e.g., 国泰民安 (*the country prospers; the people are at peace*) and 丧权辱国 (*forfeit sovereignty and humiliate the country*), which might be more aptly categorized as phrases or even sentences, could be difficult to find their word equivalents in English. However, Warriner et al. did contain English words related or partially related to the meanings of these four-character Chinese words. For example, *prosperity*, *treason*, and *sovereignty* were all included.

As an illustration, Table 5 contrasts valence rankings of the top-rating English words and their Chinese translation equivalents or close equivalents. As can be seen, the same concepts were ranked differently by speakers of the two languages. As the raters of the present study were native speakers of Mandarin Chinese, with at least seven years (mostly likely more based on geographic distribution of the raters) of imprinting within the same cultural environment, these cross-language differences revealed by valence rankings should to some extent reflect cultural and societal influences on affect representations. The findings thus demonstrate the potential for affect rating tasks with word stimuli to be utilized to investigate the influences of cultural variations and societal changes on individual mentality, to identify commonalities and differences, and to enhance mutual understanding and collaboration. Similarly, we think that these simple tasks may also be utilized in developmental and clinical research in order to detect psychological changes over one's life span or during the course of a treatment program. Further research is certainly needed to explore the potentials of these tasks to address both theoretical and practical questions.

Conclusion

This study provides affective norms for 11,310 simplified Chinese words. A large-scale database of affective norms like this can serve as a useful tool for researchers of different fields such as clinical psychology, natural language processing, and affective neuroscience. These affective ratings also further enrich the MELD-SCH (Tsang et al., 2018), a psycholinguistic database for simplified Chinese words. This comprehensive database allows researchers to gain more insight into the semantic representations of Chinese words, and to conduct more systematic cross-language investigations between the Chinese language and other languages.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01607-4.

References

Ahmed, M., Chen, Q., & Li, Z. (2020). Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Computing and Applications*, *32*(18), 14719–14732. https://doi.org/10.1007/s00521-020-04824-8

Apuzzo, M., Park, H., & Buchannon, L. (2014). Does torture work? The CIA's claims and what the committee found.

Bird, G., & Cook, R. (2013). Mixed emotions: The contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry*, *3*(May), 1–8. https://doi.org/10.1038/tp.2013.61

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings(Technical Report No. C-1).* Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46, 904-911.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6). https://doi.org/10.1371/journal.pone.0010729

Chen, P., Lin, J., Chen, B., Lu, C., & Guo, T. (2015). Processing emotional words in two languages with one brain: ERP and fMRI evidence from Chinese–English bilinguals. *Cortex*, *71*, 34-48. https://doi.org/10.1016/j.cortex.2015.06.002

Ćoso, B., Guasch, M., Ferré, P., & Hinojosa, J. A. (2019). Affective and concreteness norms for 3,022 Croatian words. *Quarterly Journal of Experimental Psychology (2006)*, *72*(9), 2302–2312. https://doi.org/10.1177/1747021819834226

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Ding, J., Wang, L., & Yang, Y. (2016). The dynamic influence of emotional words on sentence comprehension: An ERP study. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(3), 433-446. https://doi.org/10.3758/s13415-016-0403-x

Duyser, F. A., van Eijndhoven, P. F. P., Bergman, M. A., Collard, R. M., Schene, A. H., Tendolkar, I., & Vrijsen, J. N. (2020). Negative memory bias as a transdiagnostic cognitive marker for depression symptom severity. *Journal of Affective Disorders*, *274*(March), 1165–1172. https://doi.org/10.1016/j.jad.2020.05.156

Fraga, I., Guasch, M., Haro, J., Padrón, I., & Ferré, P. (2018). EmoFinder: The meeting point for Spanish emotional words. *Behavior Research Methods*, *50*(1), 84–93. https://doi.org/10.3758/s13428-017-1006-3

Gatti, L., Guerini, M., & Turchi, M. (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, *7*(4), 409–421. https://doi.org/10.1109/TAFFC.2015.2476456

Grandy, T. H., Lindenberger, U., & Schmiedek, F. (2020). Vampires and nurses are rated differently by younger and older adults—Age-comparative norms of imageability and emotionality for about 2500 German nouns. *Behavior Research Methods*, *52*(3), 980-989. https://doi.org/10.3758/s13428-019-01294-2

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. Behavior Research Methods, 48, 1358-1369. https://doi.org/10.3758/s13428-015-0684-y

Hinojosa, J. A., Moreno, E. M., & Ferré, P. (2020). Affective neurolinguistics: towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, 35, 813-839. https://doi.org/10.1080/23273798.2019.1620957

Ho, S. M. Y., Mak, C. W. Y., Yeung, D., Duan, W., Tang, S., Yeung, J. C., & Ching, R. (2015). Emotional valence, arousal, and threat ratings of 160 Chinese words among adolescents. *PLoS ONE*, *10*(7), 1–13. https://doi.org/10.1371/journal.pone.0132294

Imbir, K. K. (2015). Affective norms for 1,586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, *47*(3), 860–870. https://doi.org/10.3758/s13428-014-0509-4

Islam, M. R., & Zibran, M. F. (2018). SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, *145*(August), 125–146. https://doi.org/10.1016/j.jss.2018.08.030

Kapucu, A., Kılıç, A., Özkılıç, Y., & Sarıbaz, B. (2021). Turkish Emotional Word Norms for Arousal, Valence, and Discrete Emotion Categories. *Psychological Reports*, *124*(1), 188-209. https://doi.org/10.1177/0033294118814722

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/a0035669

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavioral Research Methods*, 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods*, *41*(1), 13–19. https://doi.org/10.3758/BRM.41.1.13

Laming, D. (2004). *Human judgement: The eye of the beholder*. Thompson Learning.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese. *Behavior Research Methods*, *39*(2), 192–198. https://doi.org/10.3758/BF03193147

Louwerse, M., & Qu, Z. (2017). Estimating valence from the sound of a word: Computational, experimental, and cross-linguistic evidence. *Psychonomic Bulletin and Review*, *24*(3), 849–855. https://doi.org/10.3758/s13423-016-1142-2

Lumley, M. A., Neely, L. C., & Burger, A. J. (2007). The assessment of alexithymia in medical settings: Implications for understanding and treating health problems. *Journal of Personality Assessment*, *89*(3), 230–246. https://doi.org/10.1080/00223890701629698

Lund, T. C., Sidhu, D. M., & Pexman, P. M. (2019). Sensitivity to emotion information in children's lexical processing. *Cognition*, *190*(January), 61–71. https://doi.org/10.1016/j.cognition.2019.04.017

Luo, Y., Liu, C., Zheng, L., & Chen, X. (2020). Attachment and autobiographical memory retrieval: Event-related potential evidence from strategic information processing. *Consciousness and Cognition*, *83*. https://doi.org/10.1016/j.concog.2020.102980.

Madan, C. R., Caplan, J. B., Lau, C. S. M., & Fujiwara, E. (2012). Emotional arousal does not enhance association-memory. *Journal of Memory and Language*, *66*(4), 695–716. https://doi.org/10.1016/j.jml.2012.04.001

Mayer, J. (2009). Behind the executive orders. *The New Yorker.*

Monnier, C., & Syssau, A. (2014). Affective norms for french words (FAN). *Behavior Research Methods 46*(4), 1128–1137

Monnier, C., & Syssau, A. (2017). Affective norms for 720 French words rated by children and adolescents (FANchild). *Behavior Research Methods*, *49*, 1882–1893. https://doi.org/10.3758/s13428-016-0831-0

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, *46*(3), 887–903. https://doi.org/10.3758/s13428-013-0405-3

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A. L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169–177. https://doi.org/10.3758/s13428-012-0243-8

Mordecai, K. L., Rubin, L. H., Eatough, E., Sundermann, E., Drogos, L., Savarese, A., & Maki, P. M. (2017). Cortisol reactivity and emotional memory after psychosocial stress in oral contraceptive users. *Journal of Neuroscience Research*, *95*(1–2), 126–135. https://doi.org/10.1002/jnr.23904

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Palogiannidi, E., Iosif, E., Koutsakis, P., & Potamianos, A. (2015). *Valence, arousal and dominance estimation for English, German, Greek, Portuguese and Spanish lexica using semantic models* [Conference presentation]. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. Behavior Research Methods, 50, 1198–1216. https://doi.org/10.3758/s13428-017-0938-y

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, *6*(1). https://doi.org/10.1140/epjds/s13688-017-0121-9

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of anew (Affective Norms for English Words). *Behavior Research Methods*, *39*(3), 600–605. https://doi.org/10.3758/BF03193031

Ricciardi, L., Demartini, B., Fotopoulou, A., & Edwards, M. J. (2015). Alexithymia in Neurological Disease: A Review. *The Journal of Neuropsychiatry and Clinical Neuroscience*, *27*(3), 179–187.

Riegel, M., Wierzba, M., Wypych, M., Żurawski, Ł., Jednoróg, K., Grabowska, A., & Marchewka, A. (2015). Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, *47*(4), 1222–1236. https://doi.org/10.3758/s13428-014-0552-1

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, *57*(5), 848–856.

Schmidtke, D. S., Schröder, T., Jacobs, A. M., & Conrad, M. (2014). ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, *46*(4), 1108-1118. https://doi.org/10.3758/s13428-013-0426-y

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods*, *44*(1), 256–269. https://doi.org/10.3758/s13428-011-0131-7

Stadthagen-Gonzalez, H., Imbault, C., Sánchez, M. A. P., & Brysbaert, M. (2017). Norms of Valence and Arousal for 14,031 Spanish Words. *Behavior Research Methods*, *49*(1), 111–123.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Taylor, G. J., Bagby, M. R., & Parker, J. D. A. (1997). *Disorders of affect regulation: Alexithymia in medical and psychiatric illness*. Cambridge University Press.

Torkamani-Azar, M., Kanik, S. D., Vardan, A. T., Aydin, C., & Cetin, M. (2019). Emotionality of Turkish language and primary adaptation of affective English norms for Turkish. *Current Psychology*, 273–294. https://doi.org/10.1007/s12144-018-0119-x

Triandis H. (1988). Collectivism v. individualism: A reconceptualisation of a basic concept in cross-cultural social psychology. In: Verma G.K., Bagley C. (eds.) *Cross-Cultural Studies of Personality, Attitudes and Cognition*. Palgrave Macmillan, .

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology*, *90*(2), 288–307.

Tsang, Y. K., Huang, J., Lui, M., Xue, M., Chan, Y. W. F., Wang, S., & Chen, H. C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, *50*(5), 1763–1777. https://doi.org/10.3758/s13428-017-0944-0

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Wang, B., & Fu, X. (2011). Time course of effects of emotion on item memory and source memory for Chinese words. *Neurobiology of Learning and Memory*, *95*(4), 415-424. https://doi.org/10.1016/j.nlm.2011.02.001

Wang, X., Wang, B., & Bi, Y. (2019). Close yet independent: Dissociation of social from valence and abstract semantic dimensions in the left anterior temporal lobe. *Human Brain Mapping*, *40*(16), 4759-4776. https://doi.org/10.1002/hbm.24735

Wang, Y., Zhou, L., & Luo, Y. (2008). The Pilot Establishment and Evaluation of Chinese Affective Words System. *Chinese Mental Health Journal*, *22*(8), 608–612.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wei, P., Wang, D., & Ji, L. (2016). Reward expectation regulates brain responses to task-relevant and task-irrelevant emotional words: ERP evidence. *Social Cognitive and Affective Neuroscience*, *11*(2), 191-203. https://doi.org/10.1093/scan/nsv097

Wrobel, M. R. (2020). The Impact of Lexicon Adaptation on the Emotion Mining from Software Engineering Artifacts. *IEEE Access*, *8*, 48742–48751. https://doi.org/10.1109/ACCESS.2020.2979148

Wu, C. E., & Tsai, R. T. H. (2014). Using relation selection to improve value propagation in a ConceptNet-based sentiment dictionary. *Knowledge-Based Systems*, *69*(1), 100–107. https://doi.org/10.1016/j.knosys.2014.04.043

Xu, S., Yin, H., & Wu, D. (2008). Initial Establishment of the Chinese Affective Words Categorize System uesd in Research of Emotional Disorder. *Chinese Mental Health Journal*, *22*(10), 770–774.

Xu, X., Kang, C., Sword, K., & Guo, T. (2017). Are emotions abstract or concrete? An ERP study on affect representations. *Experimental Psychology*, *64*(5), 315–324. https://doi.org/10.1027/1618-3169/a000374

Xu, X., & Li, J. (2020). Concreteness / abstractness ratings for two-character Chinese words in MELD-SCH. *PLOS ONE*, *15*(6). https://doi.org/10.1371/journal.pone.0232133

Xu, X., Li, J., & Guo, S. (2020). Age of acquisition ratings for 19 , 716 simplified Chinese words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01455-8

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, *49*(4), 1374–1385. https://doi.org/10.3758/s13428-016-0793-2

Yao, Z., Xuan, Y., & Zhu, X. (2019). Effect of experience information on emotional word processing in alexithymia. *Journal of Affective Disorders*, *259*, 251-258. https://doi.org/10.1016/j.jad.2019.08.068

Yee, L. T. S. (2017). Valence, arousal, familiarity, concreteness, and imageability ratings for 292 two-character Chinese nouns in

Cantonese speakers in Hong Kong. *PLoS ONE*, *12*(3), 1–16. https://doi.org/10.1371/journal.pone.0174569

Yik, M. (2009). Studying Affect Among the Chinese: The Circular Way. *Journal of Personality Assessment*, *91*(5), 416–428.

Yik, M. S. M., & Russell, J. A. (2003). Chinese affect circumplex: Structure of recalled momentary affect. *Asian Journal of Social Psychology*, (6), 185–200.

Zhang, H., Fu, Y., Zhang, X., & Shi, J. (2017). The effect of item similarity and response competition manipulations on collaborative inhibition in group recall. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-12177-x

Open Practices Statement The database created in this study can be found in the electronic supplemental file attached to this article. The study was not preregistered.

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