Behav Res (2017) 49:184–197
DOI 10.3758/s13428-016-0704-6

BRIEF COMMUNICATION

Affective meanings of 1,469 Bengali concepts

Shibashis Mukherjee[1] · **David R. Heise**[1]

Published online: 8 March 2016

Abstract This article provides semantic differential ratings of 1,469 concepts in Bengali, a language spoken by about 250 million individuals in eastern India and Bangladesh. These data were collected from 20 male and 20 female Calcutta respondents who rated stimuli on three culturally universal affective dimensions: evaluation–potency–activity (EPA). This study employs pan-respondent component analyses as a means of examining the respondents' usage of the standard EPA scales. The pan-respondent component analyses indicate that some respondents used the rating scales in unexpected ways, recording their feelings about one component of concepts' EPA with ratings on a scale intended to measure a different dimension. When scores were based only on respondents who used the scales appropriately, several interesting patterns were found. For respondents of both genders, potency scores have a curvilinear relation with evaluation, such that very good and very bad concepts are mostly seen as very potent, whereas evaluatively neutral concepts are seen as somewhat impotent or just slightly potent. A moderate linear correlation exists between activity and evaluation, and a modest positive relation exists between potency and activity. Gender correlations are high on evaluation, .93, but much lower for potency scores, with a correlation of .55, and even lower for activity, .30. In this article we examine several explanations for why scales denoting potency and activity were reinterpreted as indicating goodness by certain respondents, and consider the matter of including data collected from respondents who used scales in this way.

Keywords Affective meaning · Scale usage · Bengali language

The selection of subjects for a study of cultural norms is determined by the cultural competence of the subjects and by the desired level of measurement accuracy rather than by statistical principles applying to the study of central tendencies in variable distributions. Thus, cultural norms can be assessed from relatively small samples of respondents, selected for their cultural competence rather than randomly, and individuals who give discrepant responses may be removed from analyses. These ideas about sampling in cultural studies were introduced by Romney, Weller, and Batchelder (1986) and have been elaborated mathematically and statistically in later publications (Batchelder & Romney, 1988; Heise, 2010; Romney, 1999; Romney, Batchelder, & Weller, 1987; Romney & Weller, 1984; Weller, 1987). This article applies the ideas in an area of growing interest—assessing the affectivity of words in natural languages. Affective features of words influence response latencies, pupil dilation, event-related potentials/ERPs, transcranial magnetic stimulation/TMS, and functional magnetic resonance imaging/fMRI (Schmidtke, Schröder, Jacobs, & Conrad, 2014).

The majority of studies assessing the affective associations of words measure three dimensions of affect established as universals in a cross-cultural study of more than 20 societies (Osgood, May, & Miron, 1975). That research program's

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0704-6) contains supplementary material, which is available to authorized users.

✉ Shibashis Mukherjee
smukherj@indiana.edu

[1] Department of Sociology, Indiana University, Ballantine Hall 744, 1020 E. Kirkwood Ave., Bloomington, IN 47405, USA

introductory study (Osgood, Suci, & Tannenbaum, 1957) named these dimensions *evaluation*, *potency*, and *activity*. Research emphasizing emotionality often uses other terms such as pleasure/valence, dominance/control, and arousal (Bradley & Lang, 1994).

Databases for words in English, German, Spanish, Portuguese, French, and Finnish have been compiled using the self-assessment manikin (SAM), introduced by Margaret M. Bradley and Peter J. Lang (1994). SAM represents the three affective dimensions with rating scales that are formed by series of hominoid figures. Valence is indicated by giving the figures facial expressions ranging from *happy* to *sad*. Dominance is symbolized by the size of figure, from tiny to oversize with crossed arms. Arousal is indicated by a belly design ranging from a small circle to a large asymmetric star. Respondents are instructed to rate how they themselves feel while reading each word, and the instructions expand the scale definitions for rating their self-feelings as follows (Bradley & Lang, 1999, pp. 2–3):

- **Valence** "At one extreme of this scale, you are happy, pleased, satisfied, contented, hopeful. . . . The other end of the scale is when you feel completely unhappy, annoyed, unsatisfied, melancholic, despaired, or bored."
- **Dominance** "At one end of the scale (point left) you have feelings characterized as completely controlled, influenced, cared-for, awed, submissive, or guided. . . . At the opposite end of this scale . . . you feel completely in control, influential, important, dominant, autonomous, or controlling."
- **Arousal** "At one extreme of this scale you are stimulated, excited, frenzied, jittery, wide-awake, or aroused . . . At the other end . . . you would feel completely relaxed, calm, sluggish, dull, sleepy, or unaroused."

Another approach to assessing the affective associations of words has been used to create databases for the USA, Canada, Northern Ireland, Germany, Japan, and China.[1] This approach measures the three affective dimensions with semantic differential scales anchored by one or more adjectives at either end. For example, Heise (2010, pp. 51–52) presented an Evaluation scale anchored with *bad–awful* at one end and *good–nice* at the other end; a Potency scale anchored with *powerless–little* versus *powerful–big*; and an Activity scale anchored with *slow–quiet–lifeless* versus *fast–noisy–lively*. The instructions tell respondents to rate their feelings about the concepts, and stimuli are presented in a manner that foregrounds the concept more than the respondent's personal emotions—for instance, "an athlete is." In this tradition, the task is framed as rating the connotations of stimuli rather than the emotional reactions of the rater to the stimuli.

Schmidtke et al. (2014) observed that the two approaches produce different data, especially with regard to ratings of personal dominance versus object potency. "The important difference between them resides in the perspective that the participant has to adopt toward the rated concept. In the case of dominance, the participant is asked to establish a relation toward the rated object and then to decide whether or not he or she can dominate the object. . . . In the case of potency, the concepts are rated independently of their relation to the participant, who has to evaluate what potency the object might have, as such (p. 1113)." Ratings of dominance correlate negatively with ratings of potency (–.35), and additionally dominance correlates positively with valence, whereas potency correlates positively with arousal. Schmidtke et al. concluded that "This pattern shows that both scales cannot be considered, and should not be used as if they were, interchangeable (p. 1116)."

This report provides a database based on semantic differential ratings of words in Bengali. The task is framed so as to elicit the connotations of stimuli, rather than the emotional responses of raters, so as to contribute to the long tradition of work on affect control theory (Heise, 2007). Bengali is an important contemporary language spoken by 83 million individuals in India (Jotwani, 2010) and more in Bangladesh. Mean ratings and standard deviations from this study can be downloaded as supplemental materials with this article. Beyond the database, this article focuses on examining respondent sampling issues. Each of 40 respondents, half female, rated 1,469 stimuli on all three affective dimensions. The unusually large amount of data obtained from each respondent facilitated multivariate analyses aimed at examining respondent quality. Findings were checked with American data, in which each respondent rated far fewer stimuli but there were many more respondents.

Data

Rating scales

The Evaluation scale was anchored with Bengali words for the adjectives *beautiful*, *lovely*, *kind*, and *superior* (সুন্দর, চমৎকার, দয়ালু, উত্তম) on one side, and *ugly*, *repulsive*, *cruel*, and *inferior* (কুৎসিত, বিশ্রী, নির্দয়, অধম) on the other side. The anchors for the Potency scale were the Bengali words for *huge*, *powerful*, *big*, and *strong* (বিশাল, প্রবল, বড়ো, জোর) versus *minute*, *powerless*,

[1] Mean ratings of 1,500 or more words by males and females in the six cultures can be retrieved from *Interact*, a social psychology simulation program, available at www.indiana.edu/~socpsy/ACT/interact.htm. Additionally ratings of 620 words obtained from high school males in the mid-Twentieth Century are available at www.indiana.edu/~socpsy/Atlas/ for Arabic (Beirut), Bengali (Calcutta, India), Dutch (Amsterdam and Haarlem), English (Illinois whites and blacks), Farsi (Teheran), German (Münster), Hebrew (Israel), Hindi (Delhi, India), Malay (Kelantan state), Portuguese (Portugal), Serbo-Croat (Belgrade), Spanish (Mexico City, Yucatan, Costa Rica), Thai (Bangkok), and Turkish (Istanbul). Heise (2010) describes these datasets in more detail.

little, and *weak* (ক্ষুদ্র, দুর্বল, ছোটো, কমজোর). The Activity scale was anchored by the Bengali words for *fast*, *industrious*, *alive*, and *thin* (দ্রুত, অনলস, জীবিত, পাতলা) versus *slow*, *lazy*, *dead*, and *thick* (ধীর, অলস, মৃত, পুরু).

The defining words for the three scales were derived in a cross-cultural study conducted during the 1960s and 1970s (Osgood et al., 1975). Heise (2007, pp. 10–11; 2010, pp. 29–33) summarized the research, noting that the cross-cultural study was designed explicitly to deal with translation issues and comparability of dimensions across cultures. In brief, the procedures were as follows. First, 100 universal concepts (like *mother*, *water*, and *moon*) were translated for each of the 21 language-culture groups involved in the study. All subsequent procedures of scale construction were performed entirely within each language–culture group, without further translations. The concepts were used to elicit 50 qualifiers (adjectives) and their opposites for use as anchors on seven-point bipolar rating scales, and those scales were used by indigenes to rate the 100 concepts. Ratings on the 50 scales were concatenated across the 21 cultures to obtain a pan-cultural correlation matrix, 1,050 × 1,050. The first three factors of these correlations were extracted, revealing the familiar evaluation–potency–activity dimensions, with scales from every culture contributing to each of the three factors. Finally, the scales best measuring each dimension in each culture were chosen using the factor analysis results.

The Bengali scale anchors used in this study were the ones defined in the pan-cultural study, and we believe they are the best available. At our study's beginning, we understood that some scale anchors undoubtedly relate to multiple dimensions, and thereby contribute to artifactual correlations between dimensions. However, this is the case with psychological measurements generally (e.g., measurements of intelligence, or of personality) and determining the extent to which correlations among dimensions are real versus artifactual is an ongoing problem in the discipline. Moreover our pan-respondent analyses below show that correlations between dimensions vary across respondents. That is, evaluation and activity are highly correlated for some respondents in our study, but not for others, and evaluation–potency correlations show a similar pattern.

We incorporated multiple scale anchors at the ends of each bipolar scale, following the practice of affect control theory researchers who collected data in the USA, Canada, Germany, Japan, and China (Heise, 2010). This practice has two benefits relative to the earlier practice of anchoring each side of a rating scale with a single word or phrase. First, it allows raters to induce the shared affective essence of multiple adjectives and ignore their varying denotations. Second, it reduces the number of ratings required to measure each dimension, enabling studies of many more stimuli than would be feasible if multiple scales were used to measure each dimension.

Concepts

Most stimuli were translations of 1,500 concepts rated by American respondents (Francis & Heise, 2006). However, the authors—one Bengali and one American—determined that 327 of the American concepts were not relevant to Bengali culture or lacked direct translation equivalents (such as concepts related to Christian religious practices and to unique aspects of American culture like Thanksgiving and Halloween). These concepts were dropped and replaced by concepts like একজন কমিউনিস্ট (a *Communist*) and একটা ক্রিকেট স্টেডিয়াম (a *cricket stadium*) that were relevant to Bengali/Indian culture. In all, the stimuli comprised 502 social identities, 480 interpersonal behaviors, 283 personal modifiers, 195 social settings, and nine other kinds of concepts. The English translations of the 1,469 stimuli were alphabetized within each group (identities, behaviors, modifiers, settings) and the resulting list was divided into sets of 98 (except for one set of 97) by starting at one of the first 15 identities and selecting every fifteenth stimulus thereafter. The sets of stimuli thereby defined were incorporated into 15 questionnaires, numbered one through 15. Respondents worked through the questionnaires in numerical order, about half beginning at number one and the rest at number eight in order to mitigate possible practice effects in the final dataset.

Data collection

Each electronic questionnaire in the online survey began with three demographic questions: sex, age, and geographic origin. Then, the first questionnaire presented to a respondent gave an interactive tutorial explaining how to use the Java applet to perform affective ratings of the stimuli. Thereafter, respondents rated the connotations of stimuli on three scales measuring evaluation, potency and activity. Within each questionnaire, the order of stimuli was randomized, the order of evaluation, potency, and activity scales was randomized, and the orientation of each scale (e.g., *Beautiful, Lovely, Kind, Superior* on the left vs. right) was randomized.

Because the Bengali language does not have word-for-word translations for key terms used in the applet (e.g., "skip" and "save as"), buttons were labeled in English, and the interactive tutorial was presented in English. (The Study Information Sheet that introduced respondents to the project also was in English.) Almost all college educated Bengalis can read, write, and speak English, which is one of the two official languages in West Bengal and is taught in government schools as a second language (Jotwani, 2010). All parts of the questionnaire except the tutorial were presented in Bengali. Specifically, the concepts the respondents had to rate were Bengali words written in Bengal fonts. The website http://rishida.net/tools/conversion/ was used to convert Bengal script into Unicode for the Java program.

At the completion of each survey, respondents provided their e-mail addresses and clicked a button to save their ratings over the Internet, on a server at Indiana University. The e-mail addresses were used to coordinate a respondent's answers on different forms into a single data set.

Respondents took from one to three weeks to complete all 15 forms. On completion of the entire survey each respondent was paid 1,000 rupees or approximately USD 20.

Respondents

Respondents were recruited with an ad posted on Facebook pages of student groups at the Jadavpur University in Calcutta. The initial respondents provided additional contacts, who also were recruited, in a snowball fashion. Interested parties contacted the Bengali-speaking researcher (who was physically in the USA) via Facebook or e-mail. The resulting sample is no random draw from the Calcutta population, but starting recruitment at a university does muster middle-class individuals who arguably are the best informants regarding a mainstream culture (Heise, 2010, pp. 2–3). In all, 20 males and 20 females, all native Bengali speakers, were obtained for the study. Heise (1966) examined whether such small nonrandom samples of respondents provide generalizable mean *EPA* scores by comparing scores derived from USA working class respondents with USA college student respondents. Some small differences between these groups were statistically significant, but no major variations existed between groups in the mean evaluation, activity, or potency ratings of the words considered.

Respondent variables consisted of sex, age, start date, total number of skipped concepts, and median number of minutes for questionnaire completion. Only one correlation among these variables was significant at the .05 level across the 40 respondents: number skipped compared with median completion time, $r = .61$. This positive association suggests that respondents who did not know the meanings of numerous concepts also were slow in relating the scale anchors to concepts that they did understand. However the association varies by sex, as discussed in the next section.

Missing data

A total of 77 individuals responded, but two factors reduced the number completing the project. First, data gathering began just when difficulties arose in running Java applets with Internet browsers because of security issues. Consequently, 26 individuals (34 %) were blocked from participation because they could not get their browsers to run the applet presenting the questionnaires. Second, 11 respondents (14 %) started but then quit after doing one or two questionnaires; we discarded the data from respondents who did not complete all 15 questionnaires, since they received no payment. Thus, the 40 respondents included in the study constitute 52 % of the total number who responded to our recruitment tactics.

Respondents were allowed to skip stimuli, though they were discouraged from doing so by a short delay before the next stimulus appeared. Table 1 shows the distribution of skipped stimuli over the 40 respondents.

Table 1 shows that male individuals tended to skip fewer stimuli than female individuals, and one female skipped about one in six stimuli (16 %). The correlations between skips and median time to complete a questionnaire—for males –.48, and for females .72—suggest that males who skipped sometimes may have passed over stimuli in a rush to complete the task, whereas females who skipped sometimes may have been challenged by the verbal stimuli to the point of viewing the stimuli as too difficult to rate. This would accord with the gender difference in Bengali culture, wherein males are more likely to be cognitive specialists and females more likely to be emotional specialists. This would accord with a gender difference in many societies, including Bengali, wherein instrumental and cognitive roles are mainly the province of males, whereas females are more likely to be assigned emotionally challenging care-taking roles, both in institutions (e.g., nurse, teacher, social worker) and at home (Alexander & Wood 2000).

With regard to the stimuli, 63 % of the stimuli were rated by all 20 male respondents, and 98 % of the stimuli were rated by 15 or more males. Three concepts were rated by just three males (*give too much indulgence to*, *obedient*, and *flea market*), and one concept (*host*) was rated by just nine males. For females, 73 % of the stimuli were rated by all 20 female respondents, and 99 % were rated by 15 or more females. The concepts rated by fewer than ten females were *obedient*, rated by two females; *flea market*, rated by three females; and *give too much indulgence to*, rated by six females.

Assuming that most respondents skipped a stimulus because they did not understand it, the data in this study are missing at random, following the terminology of Schafer and Graham (2002). The case here is similar to Schafer and Graham's example "How well do you get along with your siblings," which automatically generates missing data for those with no siblings. On the other hand, this study did not use ratings to assess individual respondents, but rather used ratings by individual respondents as items of information about the cultural meaning of a concept. In essence,

Table 1 Numbers of stimuli skipped of 1,469 presented, by sex

Number Skipped	Males	Females
0–10 (0 %–1 %)	5	3
11–30 (1 %–2 %)	7	6
31–70 (2 %–5 %)	4	5
71–150 (5 %–10 %)	4	5
231 (16 %)	0	1

respondents' ratings were treated as scale items for assessing cultural values. Schafer and Graham (2002, pp. 156–157) note that nonmissing data in a scale commonly are used to impute the overall scale value, and they present analyses indicating that this form of imputation is "reasonably well behaved." We took this approach to missing data in computing evaluation–potency–activity means for our database provided as supplemental materials with this article–that is, we replaced a missing datum with the relevant scale mean on the relevant concept from all available respondents.

Our analysis of pan-respondent correlations used a different approach to missing data, substituting a respondent's mean rating on a given scale across all rated concepts for that respondent's missing ratings on that scale. We decided that the exploratory component analysis we conducted would be less likely to be distorted by this procedure, which attenuates variances and covariances, than by imputing missing values from regression analyses, which exaggerates covariances (Schafer & Graham 2002).

Analyses

Heise (2014), analyzing vintage data from 17 cultures (Osgood et al., 1975), found that mean evaluations and Potencies of different concepts are correlated substantially across all cultures, and mean Activities of different concepts also are correlated cross-culturally but with clusters of cultures having higher than average activity correlations. Analyzing EPA ratings of pairs of societies from Europe, North America, and Asia, Heise (2001) found substantial cross-cultural correlations when examining social identities, and lesser correlations for social behaviors, in which Asian cultures separated from Western cultures, especially on the activity dimension. Using a different measurement technology, Schmidtke et al. (2014) found high correlations in EPA ratings in four European cultures.

With these precedents, we can expect to find substantial correlations between the Bengali mean ratings of concepts and the mean ratings obtained from USA respondents (Francis & Heise, 2006),[2] with the correlations probably being highest on the evaluation and potency dimensions, and less high for the activity dimension. Table 2 shows the results of the relevant analyses. Evaluations do correlate substantially across the two cultures (.78 males, .79 females). Potency ratings correlate positively, but at a much lower level (.48 males, .56 females). Only about 6 % of the variance in activity ratings is shared across cultures, corresponding to correlations of .24 for both males and females.

[2] The mean ratings of words by USA respondents were obtained from the social simulation program, *Interact* (Heise, 1997).

Table 2 Correlations between average ratings of 1,173 concepts rated in Bengali (Calcutta) and in the USA (Indiana)

	Bengali					
	Males			Females		
	E	P	A	E	P	A
USA Males E	.78	.52	.58	.77	.62	.64
USA Males P	.52	.48	.46	.52	.56	.48
USA Males A	.22	.21	.24	.24	.28	.27
USA Females E	.80	.53	.59	.79	.63	.65
USA Females P	.51	.48	.46	.50	.56	.46
USA Females A	.18	.18	.20	.19	.25	.24

The relatively low cross-cultural correlations for potency and activity are not a function of low reliabilities, because mean ratings based on even 20 respondents have high reliabilities (Heise, 2010), and this is demonstrated by the relatively high level of cross-gender correlations within the cultures: in Bengali, .93 on evaluation, .77 on potency, and .80 on activity; in the USA, .96 on evaluation, .90 on potency, and .90 on activity. The conclusion, then, is that the two cultures are substantially similar in evaluations (61 % or more shared variability); moderately similar in potency ratings (23 % or more shared variability); but barely similar in activity ratings (6 % shared variance). The next section provides a more detailed examination of the interrelations of the dimensions.

Pan-respondent component analyses

Osgood et al. (1975) introduced pan-cultural factor analyses in order to ascertain whether the three EPA dimensions were present cross-culturally and to determine which scales best measured the dimensions within each culture. The method involved concatenating concept measurements on all scales from all cultures in order to form an integrated correlation matrix showing how scale measurements clustered within cultures and also clustered across cultures. The method was adapted here in order to determine the extent to which EPA measurements were independent within respondents and correlated across respondents.

Data consisted of the ratings of 1,469 concepts on three scales by 20 male and 20 female respondents. Missing data were filled in by substituting the mean rating for available data in the corresponding column of the 1,469 × 120 matrix (i.e., the respondent's mean rating on evaluation, potency, or activity). Then pan-respondent correlations were computed by treating the matrix columns as 120 different variables.

Principal components of the correlation matrix were computed. Horn's parallel analysis indicated that nine components were significant. The first three components were general, with many evaluation ratings loading on the first dimension,

multiple potency ratings loading on the third dimension, and multiple activity ratings loading on the second dimension. Components beyond the first three mostly grouped different ratings of the same respondent or of two or three respondents. Only the first three components are considered here.

A varimax rotation was applied to the first three components, and the rotated component loadings of scale–respondents are given in Table 3. The table has separate sections for males and females (top and bottom), and also has separate sections (sets of columns) for the scales measuring evaluation, potency, and activity.

Most respondents' evaluation ratings have high loadings on Component 1, which accounts for 23 % of the total variance in ratings. The potency ratings of multiple respondents have relatively large loadings on Component 2, which accounts for 8 % of the total variance. The activity ratings of multiple respondents load on Component 3, which accounts for 5 % of the total variance. The rotated Components 1, 2, and 3 are labeled here as *evaluation*, *potency*, and *activity*, respectively.

Notwithstanding the patterns, examination of Table 3 reveals that in 18 out of 40 instances, the potency ratings load higher on the evaluation or activity component than on the potency component or else have loadings of .2 or less on potency. In 33 instances, respondents' activity ratings load higher on the evaluation or potency components than on the activity component or have loadings of .2 or less on activity. The loadings indicate that, for example, both the Potency and Activity scales were used essentially as Evaluation scales by Respondents 4, 6, 8, 9, 13, 14, 21, 24, 29, 31, 36, 38, and 40. Additionally, 13 respondents assessed activity with the Potency scale, and one respondent assessed potency with the Activity scale.

Respondents were graded on the appropriateness of their scale usages by assigning 0 if their ratings on a given scale loaded less than .2 on the corresponding component; 1 if the component loading was .2 or above but less than a loading on one of the other two components; and 2 if the loading was .2 or above and higher than the loadings on both of the noncorresponding components. The scores for the three scales were summed, giving a total score ranging from 0 to 6 that registered the appropriateness of a respondent's scale usages.

The actual range of the grades was 2 to 6, since all respondents used the Evaluation scale appropriately. Four respondents used only the Evaluation scale appropriately (a grade of 2). Eleven, with a grade of 3, used the Evaluation scale appropriately, and one of the other scales (usually Potency) semi-appropriately, in that the scale loaded on its corresponding component, while having a still higher loading on some other component. At the other end of the grading scale, five respondents had perfect grades of 6, having used all three scales appropriately. Six more got grades of 5 by using two scales appropriately and one semi-appropriately. The remaining 14 respondents, with a grade of 4, used either one scale completely inappropriately or two scales semi-appropriately. A total of 12 respondents used all three scales appropriately or semi-appropriately.

Males were slightly more likely to have higher appropriateness grades ($r = .28$, $p < .05$). The grades had no significant correlation with age, median time to complete a questionnaire, or number of stimuli skipped.

USA comparisons

This study employed pan-respondent component analyses as a means of examining respondents' usages of standard EPA scales, and for that reason it is of interest to know if the proportions of respondents with aberrant interpretations of scales is the same in places other than Calcutta. To address this matter, we examined an archive on the Web[3] that contains EPA ratings of 1,500 concepts by 1,028 Midwest USA respondents.

Each USA respondent rated just 100 concepts instead of all 1,500, so the USA material consists of 15 separate datasets rather than a single dataset, as in the Calcutta study. We conducted a pan-respondent principal component analysis within each of the 15 datasets, and the first three components in each of the 15 analyses were rotated by the varimax criterion. The first three rotated components in all 15 analyses were recognizable as evaluation, potency, and activity.

Loadings of each rating scale as used by each respondent were examined to determine whether that respondent had used the scale in a way that fit the dimension that the scale was supposed to measure. Ratings from a scale–respondent combination were deemed to be valid if the scale had a loading of at least .20 on the target dimension, and if the loading on the target dimension was higher than the loadings on either of the other two dimension. The results of these analyses are summarized in Table 4.

Across the 15 USA datasets, the fewest valid raters with respect to the evaluation dimension was 85 % in Study 7, and the most was 97 % in Study 11, with a median of 94 %. By comparison, 100 % of the Calcutta respondents gave valid evaluation ratings, when applying the same criteria as were used in constructing Table 4. Thus, the qualities of evaluation ratings were comparable among the Calcutta respondents and the USA respondents.

The percentage of USA respondents giving valid potency or activity ratings always was less than the percentages giving valid responses on evaluation. The fewest USA respondents giving valid potency ratings occurred in Study 10, in which only 44 % of respondents used the Potency scale in a valid way. The median percentage across all 15 datasets was a bit

[3] Francis and Heise (2006); information about the sample is provided at www.indiana.edu/~socpsy/ACT/PDF/ProjectNotes.pdf.

Table 3 Loadings on three varimax-rotated components of pan-respondent correlation matrix, with respondents separated by sex, and scales separated by presumed dimension measured

Scale–Respondent	Component			Scale–Respondent	Component			Scale–Respondent	Component		
	1	2	3		1	2	3		1	2	3
Males											
E40	.80	.22	.20	P39	–.13	.58	.16	A1	.19	.02	.43
E13	.75	.27	.16	P30	.26	.42	–.20	A37	.32	.10	.40
E29	.74	.21	.19	P13	.52	.38	–.03	A39	.42	.28	.35
E30	.73	.03	.04	P38	.37	.36	.12	A24	.43	.34	.29
E4	.73	.24	.29	P37	.02	.34	.27	A40	.55	–.11	.28
E6	.73	.26	.13	P31	.44	.33	–.12	A11	.16	.22	.22
E31	.72	.16	.00	P20	.21	.32	.01	A38	.41	.36	.19
E3	.68	.24	.09	P6	.53	.32	.14	A6	.55	.37	.14
E11	.66	.11	.18	P11	–.01	.30	.11	A29	.69	.23	.12
E23	.66	.11	.13	P4	.56	.28	.14	A18	.14	.41	.08
E9	.66	.29	.03	P9	.62	.28	–.01	A20	.17	.33	.06
E38	.65	.17	.14	P23	.08	.25	.05	A23	.08	.40	.06
E39	.65	.09	.23	P29	.69	.25	.16	A4	.46	.28	.06
E1	.64	.26	.29	P5	.03	.25	.04	A9	.55	.40	.05
E18	.64	.07	.25	P18	.22	.24	.15	A13	.54	.40	.03
E24	.63	.17	.30	P24	.46	.24	.32	A21	.08	.24	–.01
E20	.52	.21	.06	P40	.78	.16	.15	A3	–.22	.19	–.02
E5	.45	.24	.03	P1	–.19	.14	.05	A31	.35	.39	–.02
E37	.43	.08	.28	P21	.17	.14	.01	A5	.16	.19	–.06
E21	.27	.04	.13	P3	–.19	.09	–.02	A30	.39	.41	–.10
Females											
E8	.82	.24	.12	P27	.04	.47	.01	A26	.21	.18	.52
E7	.78	.20	.22	P25	–.02	.43	.24	A36	.48	.06	.45
E19	.77	.15	.05	P7	.23	.40	.06	A12	.09	.02	.43
E17	.73	.03	.18	P19	.34	.38	–.20	A25	.37	.21	.42
E28	.72	.08	.23	P32	.16	.38	.22	A34	.08	.11	.42
E25	.71	.12	.20	P22	.12	.37	.13	A10	.17	.34	.30
E27	.70	.08	.21	P12	–.23	.34	.12	A17	.11	.27	.30
E33	.70	.13	.24	P28	.29	.33	.26	A7	.53	.26	.29
E22	.69	.15	.11	P34	–.16	.32	.26	A32	.18	.42	.28
E10	.67	.22	.29	P10	.12	.31	.29	A28	.40	.35	.26
E34	.67	.02	.25	P15	.13	.31	.02	A14	.35	.27	.19
E15	.66	.10	.19	P14	.34	.29	.12	A22	.24	.40	.19
E36	.64	.16	.32	P16	.23	.27	.23	A15	.25	.37	.18
E14	.63	.14	.14	P17	–.02	.26	.04	A16	.14	.34	.18
E26	.63	.11	.29	P8	.70	.22	.00	A35	.21	.35	.14
E12	.61	.05	.27	P35	.14	.21	.15	A27	.04	.64	.05
E16	.61	.13	.25	P26	.20	.19	.55	A2	.13	.21	–.01
E2	.55	.22	.06	P36	.49	.19	.46	A33	–.02	.56	–.01
E32	.55	.14	.30	P2	.06	.14	.04	A19	.55	.32	–.14
E35	.52	.08	.12	P33	.55	.06	.04	A8	.51	.35	–.14

E, evaluation; P, potency; A, activity. For components, 1 = evaluation; 2 = potency; 3 = activity. The scale–respondent loadings are sorted high to low within sex and dimension

higher, at 57 %, and the maximum was 73 %, in Study 9. By comparison, 55 % of our Calcutta respondents used the Potency scale in a valid way, when applying the same criteria as in Table 4.

Table 4 Percentages of USA respondents who gave valid responses on each EPA scale when rating 100 stimuli (sexes combined)

	Number of Respondents	Percentage		
Subset		Evaluation	Potency	Activity
1	76	96	68	86
2	70	94	49	77
3	76	96	68	87
4	65	95	57	69
5	64	89	55	81
6	69	94	58	81
7	59	85	46	68
8	81	90	60	81
9	60	92	73	87
10	64	89	44	73
11	61	97	57	79
12	74	96	54	85
13	64	95	50	77
14	68	96	53	74
15	76	93	58	71
Minimum		85	44	68
Median		94	57	79

Proportionately more USA than Calcutta respondents gave valid responses with regard to activity. The USA minimum was 68 % in Study 7, the median percentage across all datasets was 79 %, and the maximum was 87 % in two studies, 3 and 9. Meanwhile, the percentage of Calcutta respondents with valid activity ratings was 18 %.

Overall, almost all Calcutta respondents used the Bengali Evaluation scale correctly, and almost all USA respondents similarly used the English Evaluation scale correctly. However, both Calcutta and USA respondents used the Potency and Activity scales in valid ways at much lower rates. Moreover, for the activity dimension, the incidence of appropriate use among the Calcutta respondents was lower than the incidence among USA respondents, and the cross-cultural difference was substantial.

Norms from selected respondents

These results raise a question about how best to estimate the EPA profile for a concept. Typically, the normative affective response to a stimulus is estimated by averaging all of the available respondent ratings on a given rating scale. However, the pan-respondent component analyses indicate that some respondents use the rating scales in unexpected ways, recording their feelings about concepts' evaluation, potency, or activity with ratings on a scale intended to measure a different dimension. Averaging these aberrant ratings along with others contaminates estimates of normative affect, and generates spurious empirical associations between affective dimensions. Instead, respondent ratings that do not contribute properly as measurements of a given dimension arguably should be culled, just as invalid items are culled from a psychological test before the test is used to assess individuals' traits (Clark & Watson, 1995; Nunnally, 1967).

A set of EPA measurements of concepts were derived based just on ratings by respondents who used a given scale to assess the affective dimension that the scale was intended to measure. A respondent's ratings on a scale were included in averages defining concepts' normative EPA values only if the scale was supposed to measure the given dimension, the respondent's use of the scale loaded at .32 or higher on the corresponding component, and the respondent's use of the scale loaded at .32 or lower on both of the other two components (the .32 breakpoint represents the point at which the component accounts for 10 % of the measurement item). Ratings met the criteria on Component 1 (evaluation) for 19 males—all except Respondent 21—and for all 20 females. Ratings met the criteria on Component 2 (potency) for four males—20, 30, 37, and 39—and for eight females—7, 12, 22, 25, 27, 28, 32, and 34. Ratings met the criteria on Component 3 (activity) for two males—1 and 37—and for three females—12, 26, and 34. The supplement to this article includes the male and female evaluation, potency, and activity statistics for the 1,469 concepts, based on ratings by these selected respondents.

Figure 1 shows a matrix of scatterplots in which the Calcutta evaluation, potency, and activity scores based on selected respondents are plotted against one another, within and across sexes. For both male and female respondents in Calcutta, potency scores have a curvilinear relation with evaluation, such that very good and very bad concepts mostly are seen as very potent, whereas evaluatively neutral concepts are seen as somewhat impotent or just slightly potent.[4] Regressing potency scores on evaluation scores and their squares captures the curvilinear shape of the relation and yields multiple correlation coefficients that are higher than the linear correlation coefficients, for both males and females, as is detailed in the note to Table 5. The two other Asian cultures for which parallel data are available—Japan (Smith, Matsuno, Ike, & Umino, 2006) and China (Smith & Cai, 2006)—have a similar evaluation–potency relation, but not as well-defined as the one obtained with Bengali sentiments, perhaps because the means in previous studies were based on all rather than selected respondents. The U-shaped pattern differs from the evaluation–potency relation in Germany (Schmidtke, Schröder, Jacobs, & Conrad, 2014) and the USA (Francis & Heise, 2006), where

[4] This relation is less pronounced when norms are computed over all respondents, as discussed in the next section. Interested readers can produce three-dimensional representations of all of the normative data by enhancing the spreadsheet in our appendix with an add-on available at www.doka.ch/Excel3Dscatterplot.htm.

plots of potency versus evaluation look similar to the plots of potency–activity in Fig. 1.

A moderate linear correlation exists between activity and evaluation—.42 males, .24 females.[5] A modest positive relation exists between potency and activity—.25 males, .26 females.[6]

Overall, the cloud of data points is shaped roughly like a cashew nut in the three-dimensional space, for both males and females. As we observed above and in the notes, the cloud's shape is shared by some other Asian cultures, but not by some Western cultures, suggesting that perhaps the Asian cultures may be less likely to denigrate weakness than the Western cultures.

The diagonal cells of the male versus female sectors in Fig. 1 indicate similarities of sentiments across genders. Male and female evaluations are quite similar in Calcutta, with a correlation of .93. Gender correspondence is much lower for Calcutta potency scores, with a correlation of .55. Gender correspondence is even lower for activity scores, with a correlation of .30.

One possible explanation for Calcutta respondents' modest gender correlations on potency and activity scores is methodological. These scores are based on small numbers of respondents—four males and eight females in the case of potency, and two males and three females in the case of activity. It is possible that these tiny samples do not represent the genders adequately, and the correlations might be higher with more respondents. Also, the reliabilities of the normative measures must be relatively low—about .70 for potency, according to Table 8.6 in Heise (2010), and around .40 for activity. Suppose that we use these figures as rough estimates of the reliabilities of the scales for both males and females and correct the gender correlations for the attenuation caused by unreliability. This yields gender correlations of .79 for potency and .75 for activity. These augmented values still are substantially lower than the gender correlation for evaluations, which lends support to a substantive explanation of the modest gender correlations on potency and activity among Calcutta respondents: Namely, there is actual gender variation among Calcutta males and females in potency and activity sentiments.[7]

Norms from all respondents

The database provided as supplemental materials with this article includes EPA measurements of concepts calculated with selected respondents, as detailed above, and also calculated in the traditional way of averaging ratings of each stimulus from all respondents. The two sets of EPA measurements correlate .99, .64, .36 for males, and .99, .71, .82 for females. One reason that male activity ratings correlate so much less than female activity ratings is that the males selected for the activity dimension were fewer in number and with lower component loadings than was the case for females.

Table 5 compares the two approaches with respect to interdimensional correlations and gender correlations. Correlations based on ratings from selected respondents appear in the lower triangle of Table 5's correlation matrix, and correlations based on ratings from all respondents appear in the upper triangle.

The three dimensions are substantially more correlated when based on all respondents than when based on selected respondents. Within the male norms, the potency–evaluation correlation is .78 as compared to .37, and within female norms, the respective correlations are .66 versus .37. (Since Fig. 1 shows that this relation is curvilinear, the note to Table 5 also reports correlation coefficients for a curvilinear model: .83 vs. .55 among males, and .77 vs. .66 among females.) Within male norms, the activity–evaluation correlation is .78 when norms are computed from all respondents, as compared to .42 when norms are computed from selected respondents; among females, the figures are .70 versus .24. In the case of the activity–potency correlation, the value among males is .80 when based on all respondents, and .25 when based on selected respondents; among females, the corresponding values are .77 and .26.

The correlations are higher when ratings from all respondents are used to define norms because many respondents assessed evaluation with the Potency and Activity scales. Consequently, incorporating their ratings into the norm

[5] The Calcutta pattern is similar to the one in the USA, where the evaluation–activity correlation is .38 for males and .27 for females. A linear relation also exists in Japan, but in the opposite direction: –.26 males, –.20 females. In China, the correlations are stronger than the Bengali relation: .66 for males and .71 for females. In Germany, evaluation and activity are uncorrelated (–.09 males, –.03 females).

[6] Similar but stronger relationships exist in China (.42 males, .48 females), Germany (.40 males, 054 females), and the USA (.60 males, .54 females). Potency and activity are unrelated in Japan (.08 males, .07 females).

[7] Concepts that are more potent for females (all with a gender difference greater than 3.3) include *killing someone*, *short-changing someone*, *offering someone a bribe*, *a pagan*, *being forgiving*, and *quarreling with someone*. Concepts that are more potent for males (all with a gender difference greater than 2.6) include a *White*, a *heroine*, *provoking someone*, *demeaning someone*, *an instructor*, and *being virtuous*. Concepts that are more active for females (all with a gender difference greater than 4.8) include a *fiancée*, *yelling at someone*, *a soul mate*, *a social scientist*, *grasping someone*, and a *finance minister*. Concepts that are more active for males (all with a gender difference greater than 5.2) include a *boy bachelor*, a *software engineer*, a *card game*, a *lady*, a *Hindu priest*, and a *playmate*.

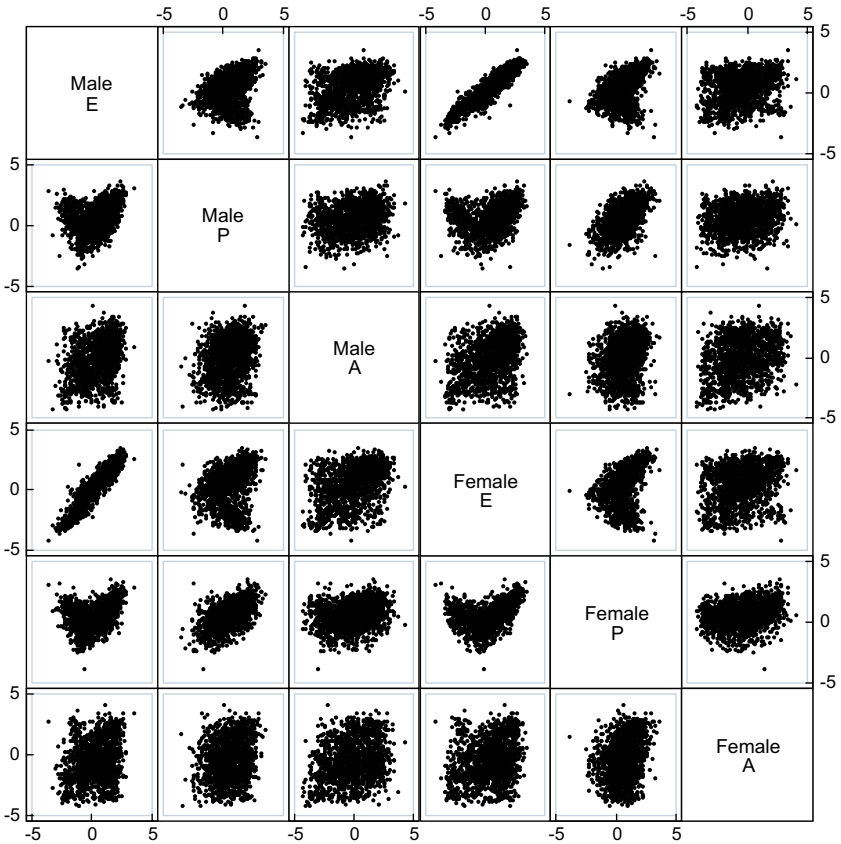

Fig. 1 Scatterplots showing the relations among evaluation, potency, and activity, over 1,469 concepts, computed from the ratings of selected Calcutta male and female respondents

estimates generated artifactual correlations of evaluation scores with the other two scores, and additionally created a spurious correlation between potency and activity. Also, a few respondents rated potency with the Activity scale or rated activity with the Potency scale, and including their ratings in the normative estimates also generated an artifactual correlation between potency and activity.

Much of the patterning of the potency–evaluation relation is lost when the ratings from all respondents are used to define the norms. In Fig. 1, the highest potency concepts are either very good or very bad, and there are numerous bad, potent concepts, such as *cheating on someone*, *killing someone*, *molesting someone*, *raping someone*, a *murderess*, and a *rapist*. A comparable graph based on norms from all respondents shows attenuation of the curvilinear relation between potency and evaluation relation with only a few bad-potent concepts remaining as outliers. This is because including potency ratings from all respondents contaminates the potency measurement with evaluation, and thereby bad-potent concepts move down on potency.

Gender correspondence on potency and activity seems greater when all respondents are used to define norms. The gender correlation for potency scores based on all respondents is .77, whereas the corresponding correlation for norms based on selected respondents is .55; the gender correlation for activity scores based on all respondents is .80, as opposed to the correlation of .30 when scores are based on selected respondents. Yet the higher levels of gender correspondence are misleading, since they are made up of the moderate levels of gender correspondence on potency and activity displayed in Fig. 1, combined with a high level of gender correspondence on evaluation, which gets imported to potency and activity by including respondents who assessed evaluation when making ratings with the Potency and Activity scales.

Table 5 Correlations among dimensional mean ratings of 1,469 concepts based on selected respondents (lower triangle) and all respondents (upper triangle), within and across sex

	Me	Mp	Ma	Fe	Fp	Fa
Male E (Me)	1.00	.78	.78	.93	.65	.70
Male P (Mp)	.37	1.00	.80	.71	.77	.72
Male A (Ma)	.42	.25	1.00	.72	.68	.80
Female E (Fe)	.93	.31	.39	1.00	.66	.70
Female P (Fp)	.39	.55	.23	.37	1.00	.77
Female A (Fa)	.23	.20	.30	.24	.26	1.00

With selected respondents, the multiple correlation of Mp with Me and Me^2 is .55; that of Fp with Fe and Fe^2 is .66. With all respondents, the multiple correlation of Mp with Me and Me^2 is .83; that of Fp with Fe and Fe^2 is .77.

Discussion

This article reports analyses of respondent quality in the measurement of affective norms, using procedures similar to item analyses in psychometrics.

Forty Calcutta respondents rated 1,469 stimuli in the Bengali language on an Evaluation scale, a Potency scale, and an Activity scale, with the Bengali adjective anchors for all three scales' endpoints having been defined in a comprehensive cross-cultural study in the mid-20th century. Each respondent's ratings on the three scales were correlated with other respondents' ratings, yielding a matrix of 120 pan-respondent correlations. The principal components of the pan-respondent matrix were computed, and the three largest components were identifiable as evaluation, potency, and activity.

The Evaluation scale was stable across respondents, with all respondents' ratings on this scale loading highest on the evaluation component. However, the Potency scale measured potency for only 55 % of respondents (applying the same criteria as were used in Table 4). That is, the ratings of only 22 respondents loaded appropriately on the potency component whereas most of the rest used the Potency scale as an Evaluation scale. In the case of the Activity scale, just seven respondents provided ratings that loaded appropriately on the activity component, 15 respondents provided ratings that related to evaluation instead of activity, and ratings of 15 respondents assessed potency rather than activity.

The data from an American study showed the same patterns, except that somewhat more respondents used Potency and Activity scales in intended ways. Typically about 94 % of American respondents used the Evaluation scale correctly, 57 % assessed potency with the Potency scale, and 79 % assessed activity with the Activity scale.

These results replicate those of a legacy study using a different methodology. Wiggins and Fishbein (1969) had 97 American college students rate the similarity of 15 semantic differential scales on a 7-point scale, and indicate which poles were related in the cases of nonzero ratings. For example, some individuals rated the *active–passive* scale as similar to the *good–bad* scale, and among these some thought that *active* related to *good* whereas others thought that *passive* related to *good*. The similarity ratings by all 97 raters were processed with a multidimensional scaling procedure that identified ten idealized individuals around a circumplex, along with the number of factors in their similarity ratings. One individual had no activity dimension, the Activity scales being absorbed into the evaluation dimension. Other individuals on the opposite side of the circumplex split the activity dimension or the evaluation dimension into two subdimensions. Some individuals along the other axis of the circumplex absorbed Potency scales into the evaluation or activity dimensions. Wiggins and Fishbein (p. 190) concluded that "scale indicants of *E*, *P*, and *A* are not substantively similar across individuals," even though a group-average analysis yielded the usual *EPA* structure. The Wiggins and Fishbein results, along with ours, indicate that diversity in the ways respondents use Potency and Activity scales is obscured by averaging ratings.

One reason why Potency and Activity scales functioned better in the American study might be that American respondents all were college students, who frequently experience testing situations. In contrast, the Calcutta respondents were selected from a general middle-class network, with just 32 % in the age range 18–20, whereas 68 % were 21 or older, and 20 % were above age 30. Thus most Calcutta respondents were long removed from structured testing situations that are common in academia. In this regard, reduced performance in the Calcutta study perhaps presages similar problems with respondents from other general populations. Results from several studies (Heise, 2010; Thomas & Heise, 1995; Wiggins & Fishbein, 1969) hint that the best informants are individuals who are socially integrated and academically successful.

Empirical studies typically show that the evaluation, potency, and activity dimensions are correlated as shown in Fig. 1. This study reveals that the correlations largely result from individual respondents who use a scale designed to measure one dimension as if it were a measure of another dimension. Then different dimensions correlate when mean ratings of stimuli include these respondents. In this study all three dimensions correlated highly when means were based on all respondents (Table 5). However, employing just respondents who used rating scales as intended substantially reduced the correlations between evaluation and activity and between potency and activity. At the same time, culling increased the relationship between evaluation and potency (Fig. 1), revealing a substantial non-linear correlation, of the kind that has received attention in previous studies of affective meanings (e.g., Schmidtke et al., 2014). Culling also reduced gender correlations on the potency and activity dimensions. Generally speaking, failure to cull ratings by respondents who use scales in unintended ways may cause EPA measurements to seem more interrelated than they really are, may obscure patterns of relationships among EPA measurements, and may lead to misconceptions about relations between EPA measurements and other variables like gender.

This article focused on respondents who used rating scales in a manner that measured the dimensions that scales were designed to measure, but the culled respondents also are of psychological interest. Below we consider three conjectures regarding these individuals.

Could it be that respondents who collapse potency or activity meanings to evaluation, or potency and activity meanings to each other, actually do not carry meanings of the collapsed dimensions at all? This seems doubtful because so much of personal and social life are dependent on all three dimensions. For example, emotions are distributed in a three-

dimensional space of evaluation–pleasantness, potency–control, and activation–arousal (Fontaine, Scherer, Roesch, & Ellsworth, 2007; MacKinnon & Keating, 1989; Morgan & Heise, 1988), and facial expressions of emotion express an individual's temporary state with respect to all three dimensions (Lively & Heise, 2014). The three dimensions also are integral to interpersonal communication and behavior, and to personality (Heise, 2007; Scholl, 2013). Thus individuals operating solely with evaluation probably would be too handicapped to manage proper interpersonal relations.

Do the culled respondents have idiosyncratic understandings of adjectives used to anchor rating scales, such that adjectives denoting potency and activity are reinterpreted as indicating goodness? Adjective anchors of rating scales never are pure measures of the dimension they are supposed to assess. For example, rating scales in English often employ a Potency scale with the adjective pair *strong–weak* at the scale endpoints. However, the EPA profile[8] of *strong* is 1.52, 1.65, 1.52; the profile for *weak* is –1.38, –2.52, –0.91; and the differences between the two are 2.90, 4.17, and 2.43. Thus, although these two adjectives contrast especially on the potency dimension, they also contrast on the other two dimensions, and therefore potency ratings with this scale inevitably are contaminated by evaluation and activity assessments. Wiggins and Fishbein (1969) provided evidence that individuals vary in the affective meanings attributed to the scale anchors, with some individuals seeing *strong–weak* as almost totally a potency contrast, whereas others see the two adjectives as mostly an evaluation contrast. In this conjecture individuals who use *strong-weak* mainly in an evaluative way do have potency and activity associations, but those associations have to be assessed on rating scales with other adjectives more tuned to the individuals' affective meaning systems. One might suppose that graphic rating scales have similar affective meanings across all respondents, allowing use of graphic scales to eliminate major differences among respondents and obtain uncontaminated assessments of affective meanings on each dimension. However, the case of Bradley and Lang's (1994) SAM suggests that this is not the case since pleasure and dominance correlate highly in their system (Bradley & Lang, 1994; Schmidtke et al., 2014). Another possibility might be to employ facial expressions of emotion as indicators of affective meanings since facial expressions have considerable universality (Ekman, 1971) and communicate states on the three affective dimensions (Lively & Heise, 2014). Respondents might draw cartoon expressions corresponding to what they feel, using software such as that presented by de Rooij, Broekens, and Lamers (2013) that simplifies the drawing process and gives the values on each affective dimension that underlie the facial expression.

[8] These EPA values are from the "North Carolina 1978" dictionary in the program Interact (Heise, 1997).

Still another interpretation of the culled respondents is it that they accurately judge potency and activity by their personal standards, but their affective associations on these dimensions largely are determined by their assessments of evaluation (or by the other dimension within the potency–activity pair). Thomas and Heise (1995) revealed that multiple affective norms exist for most concepts, with different groups of respondents providing different patterns of evaluation–potency–activity. For example, the stimulus *being mad feels* evokes powerful ratings among some respondents, but ratings of weakness among other respondents. Perhaps the first group corresponds to this study's selected respondents who assess potency apart from evaluation, whereas the second group corresponds to the culled respondents who infer impotency from negative evaluation. Each group presumably processes its assessments as authentic indicators of potency, for example with beliefs within the first group that anger empowers and within the second group that anger represents lack of control. As an interpersonal example, an individual encountering a negatively evaluated other might act acquiescent if the other were seen as powerful, but someone who equates badness with weakness might be confrontational with negatively evaluated others.

Though we culled them from measurement analyses, the respondents whose ratings on the Potency or Activity scale reflect one of the other dimensions cannot be disregarded because they represent substantial portions of respondent populations, and it would be strange to say we are measuring norms while dismissing these respondents entirely. Their variant interpretations of situations may lead to adjustments in behavior, much like introverts and extroverts bend norms of social interaction. (The variations probably do not correspond to subcultures or social structural variations like gender and race, which only influence affective meanings related to specific values or conditions within the divergent groups; Heise, 2007, 2010.) The Wiggins and Fishbein (1969) study indicates that such respondents might be identified fairly easily by asking them to rate the similarity of a few standard markers of the EPA dimensions. Such a measure would facilitate studies of their personality and social characteristics and allow their responses to be examined in experimental studies. Hypotheses for experimental research might be derived with affect control theory's simulation program (Heise, 1997)—for example, simulations could identify actions to be expected from an individual whose activity associations derived entirely from evaluations. Further research will be required to pursue these ideas and to examine the conjectures offered above for why individuals vary in their ratings of concepts.

Structure of the database

The database is an Excel file with the following fields for each of the 1,469 entries.

- ***English*** The English translation of the Bengali stimulus concept. Each English translation is prefixed by the grammatical form of the stimulus: *B_*, *I_*, *M_*, *S_*, *O_*, for *behavior*, *identity*, *modifier*, *setting*, and *other*, respectively.
- ***Bengali*** The Bengali word or phrase that was presented as a stimulus.

The following fields appear twice, once for the statistics based on all 20 males and all 20 females, and once for the statistics based only on selected respondents as defined above. Field names in the database are prefixed with @ when designating statistics based on all respondents, and prefixed with % when designating statistics based on selected respondents.

- ***mE, mP, mA*** Means on the Evaluation, Potency, and Activity scales by males.
- ***fE, fP, fA*** Means on the Evaluation, Potency, and Activity scales by females.
- ***mEN, mPN, mAN, fEN, fPN, fAN*** The number of respondents who did not skip the stimulus and whose ratings are the basis of means and standard deviations.
- ***mE_SD, mP_SD, mA_SD*** Male standard deviations on the Evaluation, Potency, and Activity scales.
- ***fE_SD, fP_SD, fA_SD*** Female standard deviations on the Evaluation, Potency, and Activity scales.

References

Alexander, M. G., & Wood, W. (2000). Women, men, and positive emotions: A social role interpretation. In A. H. Fischer (Ed.), *Gender and emotion: Social psychology perspectives*. New York, NY: Cambridge University Press.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika, 53,* 71–92.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings* (Technical Report No. C-1). Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

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

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

de Rooij, A., Broekens, J., & Lamers, M. (2013). Abstract expressions of affect. *International Journal of Synthetic Emotions, 4,* 1–31.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. In J. K. Cole (Ed.), *Nebraska Symposium on Motivation: 1971*. Lincoln, NE: University of Nebraska Press.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18,* 1050–1057.

Francis, C., & Heise, D. R. (2006) Mean affective ratings of 1,500 concepts by Indiana University undergraduates in 2002–2003 [Computer file]. Retrieved from the Affect Control Theory Website, Program Interact, www.indiana.edu/~socpsy/ACT/interact/JavaInteract.html

Heise, D. R. (1966). Social status, attitudes, and word connotations. *Sociological Inquiry, 36,* 227–239.

Heise, D. R. (1997). Interact On-Line (Java applet). Retrieved July 1, 2006, from www.indiana.edu/~socpsy/ACT/interact/JavaInteract.html

Heise, D. R. (2001). Project Magellan: Collecting cross-cultural affective meanings via the Internet. *Electronic Journal of Sociology, 5*(3).

Heise, D. R. (2007). *Expressive order: Confirming sentiments in social actions*. New York, NY: Springer.

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. Hoboken, NJ: Wiley Interscience.

Heise, D. R. (2014). Cultural variations in sentiments. *SpringerPlus, 3,* 170. doi:10.1186/2193-1801-3-170

Jotwani, N., (Ed.). (2010). *Forty-seventh report of the Commissioner for Linguistic Minorities (July 2008 to June 2010)*. Retrieved from nclm.nic.in/shared/linkimages/NCLM47thReport.pdf.

Lively, K. J., & Heise, D. R. (2014). Emotions in affect control theory. In J. E. Stets & J. H. Turner (Eds.), *Handbook of the sociology of emotions* (Vol. 2, pp. 51–75). New York, NY: Springer. doi:10.1007/978-94-017-9130-4_4

MacKinnon, N. J., & Keating, L. J. (1989). The structure of emotions: Canada–United States comparisons. *Social Psychology Quarterly, 52,* 70–83.

Morgan, R., & Heise, D. R. (1988). Structure of emotions. *Social Psychology Quarterly, 51,* 19–31.

Nunnally, J. C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana, IL: University of Illinois Press.

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

Romney, A. K. (1999). Culture consensus as a statistical model. *Current Anthropology*, *40*(Suppl.), S103–S115.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *American Behavioral Science, 31,* 163–177.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313–338.

Romney, A. K., & Weller, S. C. (1984). Predicting informant accuracy from patterns of recall among inidividuals. *Social Networks, 6,* 59–77.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

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,* 1108–1118. doi:10.3758/s13428-013-0426-y

Scholl, W. (2013). The socio-emotional basis of human interaction and communication: How we construct our social world. *Social Science Information, 52,* 3–33.

Smith, H. W., & Cai, Y. (2006). Mean affective ratings of 1,146 concepts by Shanghai undergraduates, 1999 [Computer file]. Retrieved from the Affect Control Theory Website, Program Interact, www.indiana.edu/~socpsy/ACT/interact/JavaInteract.html

Smith, H. W., Matsuno, T., Ike, S., & Umino, M. (2006). Mean affective ratings of 1,894 concepts by Japanese undergraduates, 1989–2002 [Computer file]. Retrieved from the Affect Control Theory Website, Program Interact, www.indiana.edu/~socpsy/ACT/interact/JavaInteract.html

Thomas, L., & Heise, D. R. (1995). Mining error variance and hitting pay-dirt: Discovering systematic variation in social sentiments. *The Sociological Quarterly, 36,* 425–439.

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist, 31,* 178–193.

Wiggins, N., & Fishbein, M. (1969). Dimensions of semantic space: A problem of individual difference. In J. G. Snider & C. E. Osgood (Eds.), *Semantic differential technique: A sourcebook* (pp. 183–193). Chicago, IL: Aldine.