HHS Public Access
Author manuscript
Psychiatry Res. Author manuscript; available in PMC 2019 May 10.

Published in final edited form as:
Psychiatry Res. 2018 March ; 261: 414–420. doi:10.1016/j.psychres.2017.12.068.

Development and validation of Image Stimuli for Emotion Elicitation (ISEE): A novel affective pictorial system with test-retest repeatability

Hanjoo Kim[a,*], **Xin Lu**[b,1], **Michael Costa**[c], **Baris Kandemir**[b], **Reginald B. Adams Jr.**[a], **Jia Li**[d], **James Z. Wang**[b], and **Michelle G. Newman**[a]

[a]Department of Psychology, The Pennsylvania State University, 378 Moore Building, University Park, PA 16802-3103, USA

[b]College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA

[c]Department of Computer Science, The Pennsylvania State University, University Park, PA, USA

[d]Department of Statistics, The Pennsylvania State University, University Park, PA, USA

Abstract

The aim of this study was to develop a novel set of pictorial stimuli for emotion elicitation. The Image Stimuli for Emotion Elicitation (ISEE), are the first set of stimuli for which there was an unbiased initial selection method and with images specifically selected for high retest correlation coefficients and high agreement across time. In order to protect against a researcher's subjective bias in screening initial pictures, we crawled 10,696 images from the biggest image hosting website (Flickr.Com) based on a computational selection method. In the initial screening study, participants rated stimuli twice for emotion elicitation across a 1-week interval and 1620 images were selected based on the number of ratings of participants and retest reliability of each picture. Using this set of stimuli, a second phase of the study was conducted, again having participants rate images twice with a 1-week interval, in which we found a total of 158 unique images that elicited various levels of emotionality with both good reliability and good agreement over time. The newly developed pictorial stimuli set is expected to facilitate cumulative science on human emotions.

Keywords

Emotion elicitation; Pictorial stimuli; Emotional dimensions; Test-retest analysis; Repeatability; Reliability; Agreement; Pearson correlation analysis; Bland-Altman analysis

*Corresponding author hzk136@psu.edu (H. Kim).
[1]Co-first author.

Appendix A. Supporting information
Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.psychres.2017.12.068.

1. Introduction

There has been a yearly increase in the number of studies using pictorial stimuli. For example, in 2016, there were 534 peer reviewed journal articles that involved picture stimuli, which was approximately two times more than the number of studies published in 2006 (PsychInfo, 2016). This increase may be due to the practicality of pictorial stimuli. Such stimuli can be useful in eliciting specific emotions and associated physiological responses (e.g., Lang, 1987; Lang et al., 2005; Lang et al., 1993). In addition, due to their ease of implementation, pictorial stimuli have been used in various types of experimental studies ranging from behavioral to neuroimaging. Picture stimuli are significantly easier to manipulate, rate, and validate compared to other methodologies such as video, audio, or narrative. Improvements in digital image processing allow for size, color, resolution, hue, saturation, contrast, brightness and sharpness of images to be more easily edited in an attempt to meet the goals of each experiment. Exposure time and intensity may also be easily controlled during an experiment. In addition, this increases the ease at which larger and varied sets can be created. Having a variety of pictorial stimuli sets increases the types of questions that can be answered. Also, use of pictorial stimuli is highly cost efficient since it does not require any sophisticated devices other than display materials. These practicalities are particularly important for experimental studies, which require the use of standardized methods.

Emotion may be one of the most malleable human characteristics changing its valence and intensity based on variation of the stimuli that evoke it. Thus, standardized methods are particularly important for emotion experiments as such methods allow for accurate comparison across different studies. Differences in study methodologies are also viewed as one of the biggest sources of error in meta-analytic comparisons (Flather et al., 1997). In this regard, the use of standardized stimuli can contribute to a cumulative science by making it possible to compare consistencies and inconsistencies across different results. In addition, application of standardized stimuli helps to reduce time and effort in recreating identical experimental materials, and procedures used in previous studies.

Among previously validated pictorial stimuli sets, the International Affective Picture System (IAPs) (Lang et al., 1999) is the most widely used in the psychology literature. Although there were other types of standardized sets developed prior to its development, such as Öhman's (1986) picture stimuli and Ekman and Friesen's (1977) emotional faces, these prior sets were used for more specific research purposes (e.g. elicitation of threat responses and facial expression recognition). The IAPS dataset consists of 1182 images, in which each image is associated with an empirically derived mean and variance of pleasure (i.e., renamed as 'valence' in later studies) and arousal. Images were evaluated using the Self-Assessment Manikin (Bradley and Lang, 1994), a culture and language-free instrument and initial development involved examining male and female emotional responses to color pictures with varying emotional content. IAPs development received a great deal of attention and welcome from many emotion researchers and it has encouraged scientific replication in emotion research.

Following IAPS development, researchers have attempted to improve upon methods used in its studies. The Geneva Affective Picture Database (GAPED) study, which was developed by a group of researchers in Switzerland (Dan-Glauser and Scherer, 2011), targeted the elicitation of negative emotions and selected images based on four negative categories including spiders, snakes, and scenes which depicted moral and legal norms. In this study, 754 images were rated according to valence, arousal, and the congruence of the represented scene with moral and legal norms and a final group of 730 pictures were selected to comprise the new database. In the Nencki Affective Picture System (NAPS), which was developed in Poland (Marchewka et al., 2014), researchers chose 1356 high-quality photographs based on five categories (people, faces, animals, objects and landscapes) and they were rated on valence, arousal and approach-avoidance dimensions. There is also another picture stimulus set that focuses on a military population. The Military Affective Picture System (MAPS) consists of 240 images which depict combat-relevant scenarios (Goodman et al., 2016). In the validation study, U.S. Army soldiers rated valence, arousal, and dominance of each of the MAPS images and ratings were compared to those of non-military participants.

Despite providing an important foundation for emotion studies, the IAPS and other pictorial systems have some potential shortcomings with respect to the methodology used to validate them. Most importantly, none of the previously developed picture systems have examined the retest reliability or agreement over time of their picture stimuli. Such examination of repeatability is particularly important for potential usage in experimental studies with a longitudinal design. For such studies, it is crucial to use stimuli which elicit the same degree of affect stably over time. Without good repeatability, it is not possible to trust that the data collected from an experimental manipulation is an accurate depiction of participants' responses rather than due to irrelevant artifacts during the experimental session.

Although the examination of retest reliability/agreement for one of the previously developed pictorial systems might be thought of as a potential remedy for this limitation, there are additional potential concerns with this. First, the sampling methods used for initial selection of prior pictorial systems had some potential biases. For example, IAPS stimuli were manually selected based on subjective judgment of emotional evocativeness by only a few researchers. Similarly, the selection and classification of initial images of GAPED also depended on researchers' judgement. In the NAPS study, authors selected photographs taken by coauthors of the study and from noncommercial stock photos of newspaper companies. The authors divided these photographs into five categories (people, faces, animals, objects, and landscapes) and this preliminary categorization was tested by three raters. Moreover, although it was stated that MAPS pictures were adapted from a set of images developed by the U.S. Army Aeromedical Research Laboratory, it is not clear how the authors chose the initial 240 images for the stimuli set. Nonetheless, manual selection and classification of pictures could potentially increase the possibility of contamination effects from subjective judgments of a very small sample of individuals. The use of subjective judgments can be minimized by employing non-human selection techniques and by separating photographers from researchers during the initial image selection. Furthermore, although presentation of the four latter systems used counterbalancing or pseudorandom order, the IAPS applied only three different randomization orders for presentation of stimuli. The limited number of randomization orders could risk the influence of order effects.

We aimed to improve upon previous selection and validation approaches, with the goal of systematic development of a new pictorial dataset called Image Stimuli for Emotion Elicitation (ISEE). In order to develop an even more valid and reliable set of stimuli, the current study applied new selection and validation methods. First, in order to remove possible subjective bias and to acquire more up-to-date images, we used a computational method to randomly sample images from 'Flickr.Com', one of the most well-known image hosting websites with a rich and diverse collection of photographs (Mislove et al., 2008). According to Flicker Statistics (Flickr, 2014), 586 million, 518 million, and 560 million images were uploaded in 2011, 2012, and 2013, respectively. As mentioned in a recent study (Kennedy et al., 2007), the population of Flickr users covers almost every corner of the world. Using images from this source ensured that our stimuli were consistent with visual experiences of the modern world. We automatically retrieved relevant images using 558 emotional words as queries, which enabled us to collect thousands of images as our starting point. Furthermore, to acquire categorical emotional labels of each image, the current initial dataset included more enriched information for the image affect analysis: valence and arousal, categorical emotions, and likability. In the first and second study, rating sessions were conducted with a 1-week interval and we examined each picture's within-participant agreement over time.

For the agreement test, we applied three different statistical approaches. Although correlational analyses have been commonly used for retest reliability studies, such analyses are mainly concerned with linearity between two measurements, as opposed to absolute agreement (Bland and Altman, 1986). Although high linearity is considered a necessary condition for high agreement, correlation coefficients may not be a sufficient standard for selecting pictures with equivalent levels of emotionality over time (Berchtold, 2016; Bland and Altman, 1986). Therefore, in order to test agreement across time, we used both correlational analyses and a method which accounted for equality between the two sets of measurement. In Study 1, we applied Pearson correlational analyses. Using the number of ratings and the correlation coefficient as our screening criteria, we selected the initial set of stimuli. In Study 2, we gathered more ratings from another larger group of participants. On top of our correlation coefficient analyses, we conducted the Bland-Altman Agreement Test to examine absolute agreement and variability within each picture (Bland and Altman, 1986). Furthermore, in order to ascertain stability between the two sets of measurements, we applied a *t*-test.

2. Study 1: Preliminary selection of pictorial stimuli

2.1. Methods

2.1.1. Materials

2.1.1.1. Image crawling methods.: In order to collect images that aroused emotions, we used the 558 emotional words summarized by Averill (1975) in an attempt to trigger the image search engine, Flickr.Com. To ensure a high correlation between images and the query, we included only images with high rankings among all returned results for each emotional word. The crawled images were generated by Web users and covered diverse visual scenarios. Next, we removed duplicate, illegal, and explicit images. For this study, a total of 10,696 images were selected. The monitors we used for the study were 19 in. standard Dell monitors whose resolution was set to 1280×1024. All pictures filled 90% of the monitor. The smallest image had a resolution of 500×217, which allowed for a reasonably clear image when enlarged on the monitor. Examples are shown in Fig. 1.

2.1.1.2. Study interface.: Inspired by the concept of semantic meaning, defined in Charles Osgood's Measurement of Meaning as "the relation of signs to their significance" (Osgood, 1957; p. 7), we asked participants to evaluate a series of images from three perspectives: (1) by making ratings along dimensional scales, (2) by selecting one or more categorical emotions if relevant, and (3) by selecting their level of like/dislike of every presented image (Lu, 2016).

In step 1 (presented as section I to participants), we adopted a dimensional approach as a means to understand the emotional characteristics that participants associated with the vast array of images. The dimensional approach was also used in the creation of the IAPS database (Lang et al., 2005). The strengths of this approach have been supported by recent studies (Bradley and Lang, 1994; Lang et al., 1990, 1998; Lindquist et al., 2012). In line with the IAPS study, we used the Self-Assessment Manikin (SAM) for ratings of valence and arousal on a 9-point Likert scale. Instead of the static SAM instrument used in the IAPS study, we implemented a dynamic version of SAM, which could be manipulated easily by sliding a solid bubble along a bar (Lu, 2016). This was motivated by previous IAPS researchers' indication that "SAM was presented to subjects as an oscilloscope display under the control of a joy-stick" on the arm of the chair and "An initial instruction program associated the visual anchors of each display with the polar adjectives used by Mehrabian to define the semantic differential scales" (Lang, 1980; p. 123). The bar and bubble used in the current study, allowed the user to change the facial expression of dynamic SAM in a fluid manner, contributing to a more natural rating experience for participants. As a result of making SAM dynamic, we were able to display only a single SAM figure for each dimension, minimizing the clutter that would otherwise exist with two rows of static SAM figures, varying slightly in expression. Other than providing participants with an intuitive rating system, our system was simplistic in appearance and operation (Lu, 2016).

In addition, we provided participants with the opportunity to apply categorical emotion labels. In step 2 (Section II) we targeted the general emotion(s) that individuals associated with any given image. We included the eight basic emotions discussed in Mikels et al. (2005), which are amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. The eight emotions were displayed with check-boxes next to each one, allowing participants to select as many discrete emotions as were applicable. More importantly, by including the extra step of having participants indicate if an emotion was felt, we were trying to minimize the problem of "leading the witness" as well as discourage emotion words from being selected solely because the emotion word was or was not presented. Participants could enter one or more emotions not included in the list provided by selecting the checkbox next to the word "other", whereby a blank text box would appear (Lu, 2016). However, although we gave them the opportunity to report on additional categorical emotions, these emotion selections were not included as part of the current data analyses. Therefore, the retest reliability/agreement of the categorical emotions were not tested in the current study.

Our goal in Section (III) was to measure the likability of an image. We included likability as an affective measure to determine the image aesthetics. The study of image aesthetics has long been an active research topic in visual arts (Carter, 1994; Miller, 2009; Niekamp, 1981). The notion was introduced to the computer and information sciences domain by Datta et al. (2008). It has since become an important area of focus in recent years (Geng et al., 2011; Marchesotti et al., 2011; Nishiyama et al., 2011; O'Donovan et al., 2011; Su et al., 2011; Wu et al., 2011, 2010). Image aesthetics is highly relevant to image affect and tends to be highly correlated with valence. To quantify likability, we used Peryam and Geridot's (1952) hedonic 7-point scale for participants to select: like extremely, like very much, like slightly, neither like nor dislike, dislike slightly dislike very much, and dislike extremely (Lu, 2016).

2.1.2. Participants—A total of 179 participants were recruited from the psychology subject pool at a public university located in the northeast region of the United States. Each study consisted of two one-hour sessions, which were held with a 1-week interval, and participants received two experimental credits for their participation. Of the 179 participants, 126 (70%) were women, 53 (30%) were men. Self-identified ethnic breakdown was 74 (41.34%) Caucasian, 25 (13.97%) Asian, 10 (6.59%) African American, 5 (2.79%) Latino, and 65 (36.31%) other. Participants' average age was 19, ranging from 18 to 30 and the standard deviation was 1.7.

2.1.3. Procedure—Human subject approval was attained for this study prior to its execution. All participants voluntarily signed up for two consecutive experimental slots on the subject pool website. Experimental sessions were held in three identical computer labs. After joining the study, participants were asked to access a website which was created specifically for this study. Once participants clicked the "agree" button on the consent form, instructions for participating were presented. After carefully reviewing instructions for 5–10 min, participants were asked again to consent to the study. In order to ensure that the accumulated number of ratings on pictures had some level of balance, we limited the maximum number of images shown to any one participant within each 50–55 min session to be 200 (minimum of 15 s to rate any one picture) for both study 1 and study 2. In addition, for study 1, based on the expected number of participants, the maximum number of participants who rated each picture was set up to be 11. In order to prevent confounding from possible order effects, we counterbalanced the order of each stimulus set presented to each participant. We also collected participants' demographics, including their age, gender, ethnic group, nationality, educational background, and income level. Each study session took about one hour. More detailed information about the procedure of the study is explained below (Lu, 2016).

Step 1. Once participants clicked the "Start" button, there was a 5 s delay before they were presented with the first image.

Step 2. Each image was displayed for 6 s

Step 3. A page with 3 sections was displayed. We allowed a minimum of 15 s for participants to complete all 3 sections. For Section (I), participants were asked to rate valence and arousal based on how they "actually felt" while they observed the image. They were asked to complete Section (II) only if they felt emotions by selecting one or more of the emotions they felt and/or by entering the emotion(s) they felt into "Other." For Section (III) they rated how much they liked or disliked the image. They clicked "Next" in the lower right-hand corner of the screen when they were finished with all 3 Sections.

Step 4. Steps 2 and 3 were repeated until a button with the word "Finish" was displayed.

Step 5. Participants clicked the "Finish" button.

In step 1, we intended to provide participants with a short period to make sure they were ready for the study. We set 6 s as the default value in step 2 for participants to view the image, because we aimed to collect their immediate affective responses without putting too much thought into it. This setting was in line with the original IAPS study (Lang et al., 1999). If participants needed to refer back to the image, they were allowed to click "reshow image" in the upper left part of the screen, and click "hide" to return to the three Sections.

After completion of the first session, all participants were asked to attend the second phase, which was held one week later at which time they rated the same images presented in the first session. In order to make sure each set of 200 pictures were replicated in the second session, we saved into the data server both participant ID numbers and the ID numbers of the pictures they rated. Images which were included in the initial set but not rated by the subject were not included in the second presentation. In addition, when presenting the images the second time, we changed the order of presentation of the images by randomizing them.

2.2. Results

In the first phase of the study, a total of 10,696 pictures were rated by 179 participants. Each image was rated along the three dimensions (valence, arousal and likability) by an average of 3.14 (ranging from 1 to 11) participants. Rating scores of each image varied from 1 to 9 for valence and arousal and 1–7 for likability where a higher score indicated a higher level of each emotional dimension. Using the raw rating scores at each session, retest reliability of images in both the first and second session were analyzed, based on their correlation-coefficients, which were tested using two-tailed tests.

Descriptive statistics showed that the mean of valence ratings was 5.16 (SD=1.23). The mean rating of arousal was 4.83 (SD=0.94) and that of likability was 4.21 (SD=0.90). On the other hand, results of correlation-coefficient analyses indicated that the mean absolute Pearson correlation coefficient for valence was 0.69 (SD=0.30) and that for arousal was 0.67 (SD=0.30) and for likability was 0.66 (SD=0.31) (see Table 1).

In order to select images that would be further examined in Study 2, and to ensure representation of images along the three dimensions of emotions (valence, arousal and likability), we broke down the mean rating scores into three categories (high, neutral and low) (see Table 2). The likability score was not considered in this combination due to its high correlation with valence ($r(10,696) = 0.82$, $p < 0.001$).

Given that the initial set had a limited number of ratings which could increase the variance in their retest reliability, we applied rigorous statistical criteria for initial screening. Based on guidelines suggested by Cicchetti (1994), correlation coefficients over 0.60 in each emotion domain were regarded to have “high” retest reliability ($p < 0.05$, two-tailed) in Study 1 and were used as screening criteria for selection of stimuli. In addition, each picture had to have at least 4 separate ratings in Study 1 to be included in Study 2. In order to ensure good representativeness of initially selected images, we sampled an equal number of pictures (180 images/combination) to create 9 combinations for a total final set of 1620 images.

3. Study 2: within participant repeatability test

3.1. Methods

3.1.1. Materials—In the current study, the initial set of 1620 images was re-examined with more participant ratings.

3.1.2. Participants—A total of 497 undergraduate participants who did not attend the initial rating study (Study 1) were recruited from the subject pool of the same university. As was the case in the first study, individuals participated in two one-hour sessions with a 1-week interval in between and were provided with two experimental credits as part of their course requirements. Gender distribution was 212 (42.66%) women and 285 (57.34%) men. Ethnic breakdown was 242 (48.69%) Caucasian, 74 (14.89%) Asian, 37 (7.44%) Latino, 19 (3.82%) African American, and 125 (25.15%) other. Average age was 19.28 ranging from 18 to 27 and the standard deviation was 1.29.

3.1.3. Procedure—The procedure of Study 2 was identical to that of Study 1. The same website and monitors with the same resolution and ratio were used in the same computer labs. Each of the first and second sessions took one hour with a 1-week interval. The 1620 pictures were rated along the 3 dimensions (valence, arousal and likability). The maximum number of pictures presented to each participant in each session was 200 as in study 1. However, in study 2, the maximum number of participants to rate any one image was set to be 120, considering the large sample of participants expected to be recruited. Each of the 497 participants rated a random subset of 200 images presented twice in counterbalanced order.

To test reliability, we calculated Pearson correlation-coefficients for valence, arousal, and likeability using each participant’s raw rating scores at each time point. A post hoc power analysis revealed that a correlation coefficient of 0.60 could be detected at $p < 0.05$, two-tailed at a power of greater than 0.90 and for this, a minimum of 25 ratings would be required (Faul et al., 2007). Based on these criteria, we screened in pictures with absolute correlation coefficients greater than or equal to 0.60 (i.e. positive correlation: greater than or equal to 0.60; negative correlation: less than or equal to −0.60) and number of ratings greater than or equal to 25 on each target emotion.

Second, for the test of agreement, we conducted Bland-Altman analyses. In this metric, the standard deviation (SD) of differences between two measurements is used to create upper and lower Limits of Agreement (LOA) (mean bias ± 1.96 SD) and agreement is tested by the magnitude of difference between the two LOAs and the data scatter within the two lines (Bland and Altman, 1986). Unlike a Pearson correlation coefficient for which strength can be interpreted by its reliability guideline, Bland-Altman agreement is interpreted based on visual judgement of how well two sets of measurements agree. In Bland-Altman analyses, researchers typically rely on a priori standards as their reference for acceptable limits, as previously defined by the literature of their interest. However, such a priori agreement standard does not exist in the literature of stimuli validation, and therefore, a visual test could not be performed in the current study. Instead, we used a method which allows a quantitative comparison. The magnitude between the upper and lower LOA can be quantified by the repeatability coefficient, which represents the value below which the absolute difference between two repeated ratings could lie with a 95% probability. The Repeatability coefficient was calculated as 2.77 times $(\mathrm{i.e.}, \sqrt{2} \times 1.96)$ the standard error of measurement (i.e., within-subject standard deviation). The standard error of measurement was calculated by estimating the mean of the individual variances and then taking the square root of this value $(\mathrm{S_w} = \sqrt{\frac{\sum \mathrm{s_i^2}}{n}}$, when there are only two measurements per subject, this also can be calculated as $\mathrm{S_w} = \sqrt{\frac{\sum \mathrm{d_i^2}}{2n}})$ (Bland and Altman, 1986; Vaz et al., 2013). The benefit of using this repeatability coefficient is that it allows a quantitative comparison across different pairs of measurements which consist of the same units (Vaz et al., 2013). A smaller repeatability coefficient indicates stronger agreement between two measurements. Next, based on the repeatability coefficient, we divided the pictures into two groups, one with high agreement and the other with low agreement. For this, we applied a 50th percentile cutoff, and screened in pictures if their repeatability coefficients were lower than the 50th percentile. Along with referencing correlation coefficients and repeatability coefficients as screening criteria, we also conducted a *t*-test as a follow-up analysis in order to ensure that the group means of the two time points for the final pictures did not yield any significant differences.

3.2. Results

In study 2, each picture was rated by an average of 58 participants. Based on the aforementioned screening criteria, a total of 193 pictures were selected as a valid set of pictures (valence: 75; arousal: 27; likability: 91). Results from the follow-up *t*-test indicated that none of the selected images had significant differences between the first and second ratings.

Next, in order to sort pictures, standard-score transformations were conducted for valence, arousal and likability rating scores. For the selection of high emotion pictures, images that fell in the 66.66th percentile (upper 33.33%) with respect to the degree of each target emotion rating were included (standard score greater than or equal to 0.4307). For low emotion pictures, images which fell into the 33rd percentile (lower 33.33%) were selected (standard score lower than or equal to −0.4307). For the screening of neutral emotion pictures, we included images that fell between the 33.33rd and 66.66th percentile (standard score between −0.4307 and 0.4307). Follow-up *t*-tests indicated that none of the pictures had significant differences in their ratings between time 1 and time 2, which confirmed the repeatability of the pictures.

Finally, a total of 193 images were determined to be highly repeatable stimuli which were suitable for elicitation of high, neutral, or low levels of each emotion domain (i.e., positive valence: 29 pictures, neutral valence: 25 pictures, negative valence: 21 pictures; high arousal: 10 pictures, neutral arousal: 7 pictures, low arousal: 10 pictures; high likability: 34 pictures, neutral likability: 27 pictures, low likability: 30 pictures). After removing duplicates, a total of 158 images were counted as unique images to comprise the ISEE dataset.

In the final ISEE stimuli set, mean rating of the 75 valence pictures was 5.51 (SD=1.13), the mean of correlation coefficient was 0.67 (SD=0.06), and the mean repeatability coefficient was 2.22 (SD=0.22). For the 27 arousal pictures, the mean rating was 5.22 (SD=0.64), the mean correlation coefficient was 0.65 (SD=0.04) and mean repeatability coefficient was 2.28 (SD=0.18). The mean of rating for the 91 likability pictures was 4.39 (SD=0.66), the mean correlation coefficient was 0.69 (SD=0.06) and the mean repeatability coefficient was 1.72 (SD=0.14) (see Table 3). The resolution of the final pictures ranged from 500×217 to 2048×1233. Sample images for each degree of emotion elicitation are provided in Fig. 2.

4. Discussion

The purpose of the present study was to develop a new set of reliable pictorial stimuli called “ISEE”, that used an unbiased selection method and would elicit target emotions with strong agreement over time. In the first study, 10,696 images were crawled from a non-commercial website based on their emotional labels. These images were rated twice by 179 participants, albeit with only a small number of participants rating each image. Based on initial retest reliability, number of ratings, and distribution across emotion combinations, a total of 1620 images were selected as the preliminary set of stimuli to be retested in Study 2. In Study 2, more participant ratings were conducted for each image. Based on retest reliability and agreement, and degree of emotion induction, a total of 158 images were chosen as a reliable set of pictorial stimuli.

Several limitations of the current study should be mentioned. First, participants were undergraduate students enrolled in the same university. Due to the high uniformity in their ethnicity, income, education level and cultural background, there might be restrictions in the generalizability of study findings. In order to ameliorate these limitations, more diverse samples need to be included in future studies. In addition, for better cross-cultural generalizability of the current study’s findings, more studies need to be conducted in different countries. Furthermore, through the screening process between the first and second phase of the studies, a large number of images were excluded due to the limited number of ratings in the first study. For Study 2, pictures less than or equal to three ratings were not included. In order to increase the number of images in this pictorial system, we intend to collect more ratings on those excluded images. Also, although we gave participants the option to reshow images, we failed to collect data on this and therefore, cannot speak to any stimuli that may have required longer average duration to process. Finally, even though it has been widely posited that emotion induction is accompanied by physiological responses, physiological assessments were not conducted in this study. Thus, it is important for future studies to measure physiological reactivity from these stimuli.

Regardless of these limitations, the ISEE stimuli set developed in the current study has advantages over previously developed stimuli sets. Although it contains fewer images than prior pictorial sets, the ISEE stimuli have stability over time. For studies that use longitudinal designs (e.g. multiple time series designs, equivalent time sample designs, counterbalanced designs, separate sample pretest-posttest designs, recurrent institutional cycle designs etc.), having stable emotion elicitation effects across different time points is particularly important. In order to accurately capture pure effects of target variables, contamination effects from time variance need to be controlled. Since the ISEE stimuli set consists of pictures with adequate retest reliability and agreement, it is expected that researchers will be able to more accurately assess emotional changes across time. High retest repeatability demonstrated in the ISEE images will reduce the chance of introducing error or confounding variables such as random response errors, pictorial ambiguities, transient error due to high temporal variances, etc. Another advantage of the pictorial system we developed, is that the images were selected based on computing methods. In this study, 558 emotion-relevant labels were applied to crawled images from the world's largest image hosting website, "Flickr.Com" and ratings of those images were acquired through computer-based studies. Through computer-based approaches, we tried to minimize human bias, which is likely to operate when initial images are selected by experimenters. This type of computer-based technology is expected to greatly enhance the objectivity of stimuli development studies and thus will be widely applied in the future. All images rated in this study will be shared using the study website (http://riemann.ist.psu.edu/emotion/kim2017.htm). In order to avoid any potential copyright infringement and ensure these images are used only for research purposes, we will obtain consent from researchers prior to sharing these images. In this website, rating scores, retest repeatability indices, and detailed instructions will be provided for better ease in using these stimuli.

References

Averill JR, 1975 A Semantic Atlas of Emotional Concepts, JSAS Catalogue of Selected Documents in Psychology American Psychological Association, pp. 330 (Ms. No. 421).

Berchtold A, 2016 Test–retest: agreement or reliability? Methodol. Innov 9 (2059799116672875).

Bland JM, Altman D, 1986 Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327 (8476), 307–310.

Bradley MM, Lang PJ, 1994 Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25 (1), 49–59. [PubMed: 7962581]

Carter D, 1994 Some Important Principles of Composition Georgetown University, Washington, DC.

Cicchetti DV, 1994 Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess 6 (4), 284–290.

Dan-Glauser E, Scherer K, 2011 The Geneva Affective Picture Database (GAPED): a new 730-picture database focusing on valence and normative significance. Behav. Res. Methods 43 (2), 468–477. [PubMed: 21431997]

Datta R, Joshi D, Li J, Wang JZ, 2008 Image retrieval: ideas, influences, and trends of the new age. ACM Comput. Surv 40 (2), 1–60.

Ekman P, Friesen WV, 1977 Facial Action Coding System: A Technique for the Measurement of Facial Movement Consulting Psychologists Press, Palo Alto, CA.

Faul F, Erdfelder E, Lang A-G, Buchner A, 2007 G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191. [PubMed: 17695343]

Flather MD, Farkouh ME, Pogue JM, Yusuf S, 1997 Strengths and limitations of meta-analysis: larger studies may be more reliable. Control Clin. Trials 18 (6), 568–579. [PubMed: 9408719]

Flickr, 2014 Flickr Statistics

Geng B, Yang L, Xu C, Hua X−S, Li S, 2011 The role of attractiveness in web image search, In: Proceedings of the 19th ACM International Conference on Multimedia ACM, Scottsdale, Arizona, pp. 63–72.

Goodman AM, Katz JS, Dretsch MN, 2016 Military Affective Picture System (MAPS): a new emotion-based stimuli set for assessing emotional processing in military populations. J. Behav. Ther. Exp. Psychiatry 50, 152–161. [PubMed: 26255051]

Kennedy L, Naaman M, Ahern S, Nair R, Rattenbury T, 2007 How flickr helps us make sense of the world: Context and content in community-contributed media collections, In: Proceedings of the 15th International Conference on Multimedia

ACM, Augsberg, Bavaria, Germany, pp. 631–640.

Lang PJ, 1980 Behavioral treatment and bio-behavioral assessment: computer applications. In: Sidowski JB, Johnson JH, Williams TA (Eds.), Technology in Mental Health Care Delivery Systems Ablex, Norwood, NJ, pp. 119–137.

Lang PJ, 1987 A reply to watts and blackstock. Cogn. Emot 1 (4), 407–426.

Lang PJ, Bradley MM, Cuthbert BN, 1990 Emotion, attention, and the startle reflex. Psychol. Rev 97 (3), 377–395. [PubMed: 2200076]

Lang PJ, Bradley MM, Cuthbert BN, 1998 Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. Biol. Psychiatry 44 (12), 1248–1263. [PubMed: 9861468]

Lang PJ, Bradley MM, Cuthbert BN, 1999 International Affective Picture System (IAPS): Instruction Manual and Affective Ratings. The Center for Research in Psychophysiology University of Florida, Gainsville.

Lang PJ, Bradley MM, Cuthbert BN, 2005 International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual University of Florida, Gainesville, FL Technical Report A-6.

Lang PJ, Greenwald MK, Bradley MM, Hamm AO, 1993 Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology 30 (3), 261–273. [PubMed: 8497555]

Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF, 2012 The brain basis of emotion: a meta-analytic review. Behav. Brain Sci 35 (03), 121–143. [PubMed: 22617651]

Lu X, 2016 Visual Characteristics for Computational Prediction of Aesthetics and Evoked Emotions The Pennsylvania State University.

Marchesotti L, Perronnin F, Larlus D, Csurka G, 2011 Assessing the aesthetic quality of photographs using generic image descriptors, In: IEEE International Conference on Computer Vision (ICCV) IEEE, Barcelona, Spain, pp. 1784–1791.

Marchewka A, Żurawski Ł, Jednoróg K, Grabowska A, 2014 The Nencki Affective Picture System (NAPS): introduction to a novel, standardized, wide-range, high-quality, realistic picture database. Behav. Res. Methods 46 (2), 596–610. [PubMed: 23996831]

Mikels JA, Fredrickson BL, Larkin GR, Lindberg CM, Maglio SJ, Reuter-Lorenz PA, 2005 Emotional category data on images from the International Affective Picture System. Behav. Res Methods 37 (4), 626–630. [PubMed: 16629294]

Miller J, 2009 Building blocks of visual design, Booklet #2: The Northern Virginia Alliance of Camera Clubs

Mislove A, Koppula HS, Gummadi KP, Druschel P, Bhattacharjee B, 2008 Growth of thE Flickr Social Network, First Workshop on Online Social Networks ACM, Seattle, WA, pp. 25–30.

Niekamp W, 1981 An exploratory investigation into factors affecting visual balance

Educ. Commun. Technol. J 29 (1), 37–48.

Nishiyama M, Okabe T, Sato I, Sato Y, 2011 Aesthetic quality classification of photographs based on color harmony, In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Providence, RI, pp. 33–40.

O'Donovan P, Agarwala A, Hertzmann A, 2011 Color compatibility from large datasets. ACM Trans. Graph 30 (4), 1–12.

Öhman A, 1986 Face the beast and fear the face: animal and social fears as prototypes for evolutionary analyses of emotion. Psychophysiology 23 (2), 123–145. [PubMed: 3704069]

Osgood CE, 1957 The Measurement of Meaning University of Illinois Press.

Peryam DR, Girardot NF, 1952 Advanced taste-test method. Food Eng 24 (7), 58–61. PsychInfo, 2016. PsychInfo Search Results.

Su H−H, Chen T−W, Kao C−C, Hsu WH, Chien S−Y, 2011 Scenic photo quality assessment with bag of aesthetics-preserving features, In: Proceedings of the 19th ACM International Conference on Multimedia ACM, Scottsdale, Arizona, pp. 1213–1216.

Vaz S, Falkmer T, Passmore AE, Parsons R, Andreou P, 2013 The case for using the repeatability coefficient when calculating test–retest reliability. PLoS One 8 (9), e73990. [PubMed: 24040139]

Wu O, Hu W, Gao J, 2011 Learning to predict the perceived visual quality of photos, IEEE International Conference on Computer Vision (ICCV) IEEE, Barcelona, Spain, pp. 225–232.

Wu Y, Bauckhage C, Thurau C, 2010 The good, the bad, and the ugly: Predicting aesthetic image labels, In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR) IEEE, Istanbul, pp. 1586–1589.

Fig. 1.

Image examples in the database. Images in the database were crawled from Flickr.Com, with 558 emotional words as crawling seeds.

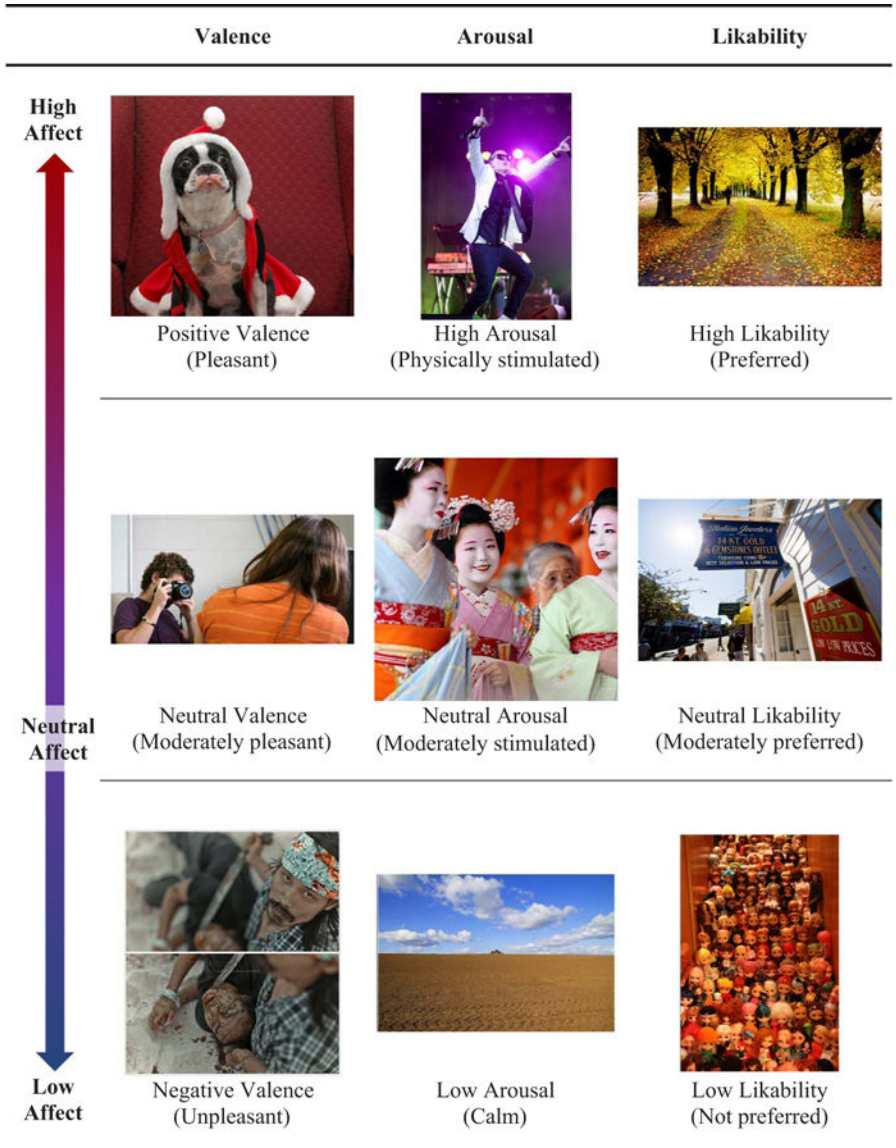

Fig. 2.
Example images which are suitable for the elicitation of each target emotion.

Table 1

Descriptive statistics of study 1 (n = 179).

Emotion Type	Mean (SD)	Pearson Correlation Coefficient
Valence	5.16 (1.23)	0.69 (0.30)
Arousal	4.83 (0.94)	0.66 (0.31)
Likability	4.21 (0.90)	0.67 (0.30)

Table 2

Number of pictures in the combination of valence and arousal rating scores (Study 1) (n = 10,696).

Case	Combination	Number of Pictures
1	High Valence, High Arousal	993
2	High Valence, Mid Arousal	960
3	High Valence, Low Arousal	1615
4	Mid Valence, High Arousal	1004
5	Mid Valence, Mid Arousal	1551
6	Mid Valence, Low Arousal	1015
7	Low Valence, High Arousal	1559
8	Low Valence, Mid Arousal	1077
9	Low Valence, Low Arousal	922

Table 3

Mean and repeatability statistics of final pictures (n = 497).

Picture Type (n)	Mean (SD)	Repeatability	
		Pearson Correlation Coefficient	Bland-Altman Repeatability Coefficient
Valence Pictures (75)	5.51 (1.13)	0.67 (0.06)	2.22 (0.22)
Arousal Pictures (27)	5.22 (0.64)	0.65 (0.04)	2.28 (0.18)
Likability Pictures (91)	4.39 (0.66)	0.69 (0.06)	1.72 (0.14)