RESEARCH ARTICLE

# What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli [version 1; peer review: awaiting peer review]

Tricia X. F. Seow[1,2], Tobias U. Hauser[1-4]

[1]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, England, WC1B 5EH, UK
[2]Wellcome Centre for Human Neuroimaging, University College London, London, England, WC1N 3AR, UK
[3]Department of Psychiatry and Psychotherapy, Eberhard Karls Universitat Tubingen, Tübingen, Baden-Württemberg, 72076, Germany
[4]German Centre for Mental Health (DZPG), Tübingen, Germany

**v1** **First published:** 19 Feb 2024, **9**:83
https://doi.org/10.12688/wellcomeopenres.20693.1
**Latest published:** 19 Feb 2024, **9**:83
https://doi.org/10.12688/wellcomeopenres.20693.1

## Abstract

### Background

Visual stimuli are integral to psychology and cognitive neuroscience research, with growing numbers of image repositories tagged with their affective information like valence and arousal. However, more specific affective domains such as anxiousness and harm have not been empirically examined and reported for visual stimuli, despite their relevance to task paradigms investigating common psychiatric disorders like anxiety and obsessive-compulsive disorder (OCD).

### Methods

In this study, we asked N = 80 participants to assess a set of 42 unique visual stimuli consisting of a variety of animals and tools on anxiety and harm scales. We then assessed the ratings' psychometric properties.

### Results

We found that animals were generally rated as more harm-perceiving and anxiety-inducing than tools, and were also higher in their inter-rater and test-retest reliabilities.

### Conclusions

**Open Peer Review**

**Approval Status** *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

With this, we provide a database of affective information for these stimuli, which allows for their use in affective task paradigms using psychometrically validated visual stimuli.

## Plain Language Summary

We often use images in experimental paradigms - as cues, signs of feedback, or even as stimuli to provoke emotions. For the latter, there has been efforts to collate reliable emotional ratings of unique images for future research use. However, the effective information of images are often examined in terms of valence and arousal, and not so much on more specific information like anxiousness or harm. Here, we investigated an array of images encompassing animals and tools, and asked participants to rate if the items were perceived as harmful or anxiety-inducing. We then assessed how reliable these ratings were. We found that animals were seen as more harmful and anxiety-inducing than tools, and their ratings were also more reliable. We provide a well annotated and validated database of images with anxiousness and harm information for future use.

## Keywords

affective ratings, visual stimuli, psychometric properties, reliability

**Corresponding author:** Tricia X. F. Seow (t.seow@ucl.ac.uk)

**Author roles: Seow TXF**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hauser TU**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

**Competing interests:** TUH consults for Limbic Ltd., and holds shares in the company, which is entirely unrelated to the current project. No other competing interests were disclosed.

**Grant information:** This work was supported by Wellcome (211155, https://doi.org/10.35802/211155). TXFS is a Sir Henry Wellcome Postdoctoral Fellow [224051, https://doi.org/10.35802/224051] of the Wellcome Trust, based at the Max Planck UCL Centre for Computational Psychiatry and Ageing Research. TUH is supported by a Sir Henry Dale Fellowship (211155; 224051) from Wellcome & Royal Society, a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, a 2018 NARSAD Young Investigator grant (27023) from the Brain & Behavior Research Foundation and a Philip Leverhulme Prize from the Leverhulme Trust (PLP-2021-040). The Max Planck UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from Wellcome [203147, https://doi.org/10.35802/203147]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Seow TXF and Hauser TU. **What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli [version 1; peer review: awaiting peer review]** Wellcome Open Research 2024, **9**:83 https://doi.org/10.12688/wellcomeopenres.20693.1

**First published:** 19 Feb 2024, **9**:83 https://doi.org/10.12688/wellcomeopenres.20693.1

## Introduction

Visual stimuli are a mainstay of cognitive neuroscience research, across many domains such as perception (Brosch *et al.,* 2010; Standing *et al.,* 1970), attention (Pessoa, 2005; Schupp *et al.,* 2007), memory (Kensinger *et al.,* 2007; Standing *et al.,* 1970) and decision making (Ahn & Picard, 2006; Koster *et al.,* 2016). They can hold affective content, being intrinsically rewarding or punishing (Blatter & Schultz, 2006; Bray & O'Doherty, 2007), and can evoke related autonomic (Klorman *et al.,* 1975) and cognitive responses (Begleiter *et al.,* 1967). Affective images are also often integral to experimental paradigms probing cognitive mechanisms underlying psychiatric disorders, especially those thought of as emotion dysregulation disorders like anxiety (Bradley *et al.,* 1999; Carretié, 2014), depression (Gotlib *et al.,* 2005; Zhang *et al.,* 2022) and obsessive-compulsive disorder (OCD) (Schienle *et al.,* 2005). With their widespread use, there has been a burgeoning effort to make stimuli repositories publicly available along with their rated affective levels of valence and arousal for free use (e.g., Geneva affective picture database (GAPED) (Dan-Glauser & Scherer, 2011), Open Affective Standardized Image Set (OASIS) (Kurdi *et al.,* 2017), EmoMadrid (Carretié *et al.*, 2019), etc.).

However, one limitation of these datasets is that they do not consider more specific affective domains, such as anxiety or harm, which are characteristic of common mental health disorders like generalised anxiety disorder and OCD. While a high level of anxiousness is a more obvious relevant symptom, harm avoidance is also common behaviour of anxiety (Markett *et al.,* 2016) and OCD (Hauser *et al.,* 2016), with hypothesized neural underpinnings. Especially for OCD, stimuli thought to be harmful have been shown to provoke different responses in patients (Da Victoria *et al.,* 2012; Tolin *et al.,* 2001). For instance, patients with OCD reported higher confidence in their memory for items they deemed as harmful as compared to those they thought of as safe (Tolin *et al.,* 2001). In the present study, we sought to characterise a range of images in terms of how harmful and anxiety-inducing the item was perceived. We chose two image categories to examine—animals, due to their familiarity with the general population, and tools, as household objects are commonly known to manifest as OCD-related harm triggers but unknown if they would be relevant in the general population.

A second aim of the study was to test the reliability of the reported anxiety and harm ratings when these visual stimuli are presented online. There is a growing awareness of the importance of psychometric properties of task designs like its reliability and validity for reproducible research (LeBel & Paunonen, 2011; Matheson, 2019). Stimuli utilised in these tasks should be no exception. Given the popularity of web-based testing, it is imperative that these stimuli are validated in online populations with reasonable test-retest reliability for reliable and effective use. Notably, online cohorts are typically more diverse than in-lab sample (Henrich *et al.,* 2010), and thus may provide greater ecological validity to these data.

Here, we used a web-based rating task where N = 80 participants rated 21 animal and 21 tool unique images over repeated presentations along anxiety and harm scales. The study had two main aims: to (i) gather a database of anxiety and harm ratings on these visual stimuli and to (ii) examine the psychometric properties (inter-rater similarity, internal consistency, intra-rater stimulus test-retest reliability, scale reliability test-retest and individual stimulus test-retest reliability) of these ratings. Participants also completed psychiatric questionnaires on trait anxiety, depression and obsessive-compulsive symptoms so that we could probe if symptom severities were linked to the level or reliability of reported ratings.

We found that the array of tested visual stimuli presented a range of anxiety and harm rating levels. On average, animal stimuli were perceived as more harmful and anxiety-provoking than tools. Animal stimuli also generally garnered more reliable ratings than tools in several measures. The rating levels and their reliability in both stimuli categories were not significantly associated to psychiatric severity scores, except for a trending effect of higher depressive scores and lower rating reliability for tool stimuli. Our findings support the use of these visual stimuli for online research, particularly in using animal stimuli for affective paradigms.

## Methods

### Participants

We recruited N = 100 participants via Prolific (https://www.prolific.co/). All participants were aged between 18–55 years old, residents of the United Kingdom, fluent in English and had a ≥ 90% approval rate on Prolific submissions. Informed consent for the study was provided online after reading the study information and consent pages. Participants were given 35 minutes for task completion and were reimbursed £5. The study procedures were approved by the UCL Research Ethics Committee (project ID number 15301\001, 29th November 2019), this ethics approval number covers a series of online studies that fall under the umbrella of this series (Dubois & Hauser, 2022; Loosen *et al.,* 2022; Seow & Hauser, 2022).

### Stimuli array

We selected an array of 42 images to contain a variation of small and large items from 2 categories: animals and tools (21 each category) (Figure 1A). The original images were taken from the publicly available image repository (https://konklab.fas.harvard.edu/ImageSets/AnimacySize.zip) (Konkle & Caramazza, 2013). For this study, we resized the images to 400 × 400 pixels, converted them to greyscale and matched them for luminesce with a custom MATLAB script.

### Procedure

After online consent, participants were directed to the task instruction pages. They were told that an array of images would be presented to them one by one. For each image, they were required to name the item in a text box and rate the extent to how the item was perceived on two rating scales: anxiety and harm (Figure 1B). We used 1 to 100 continuous scales. For anxiety, the

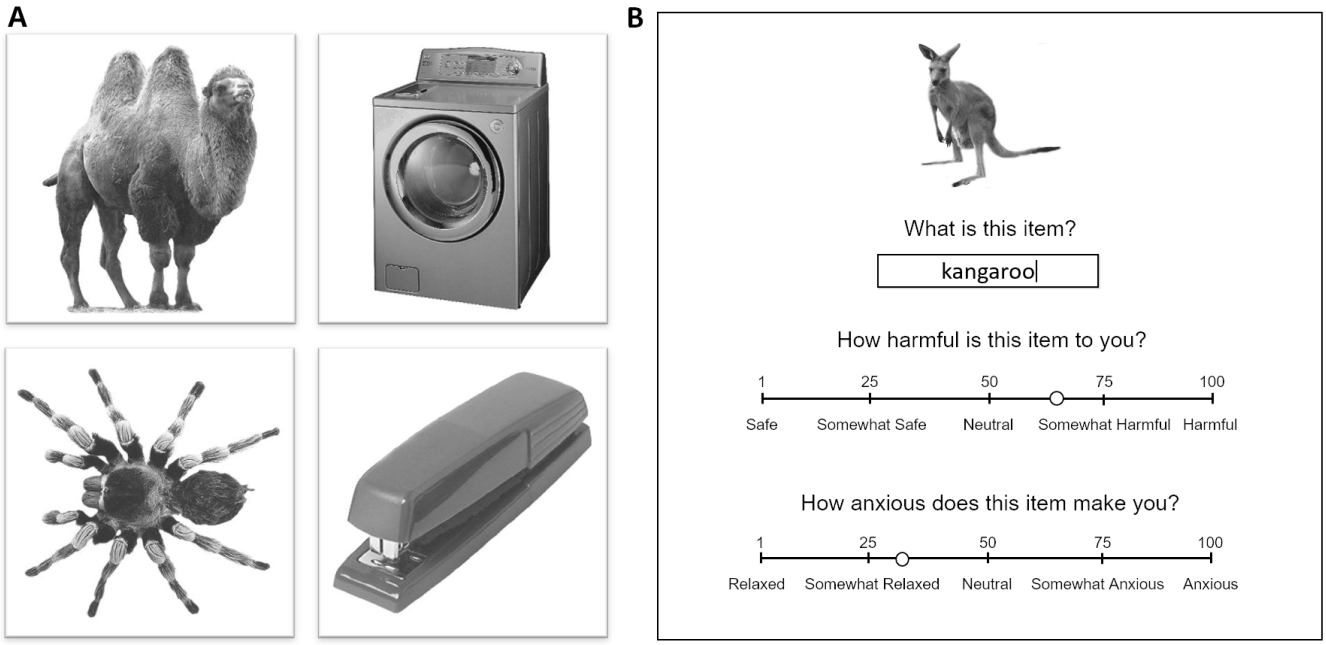


**Figure 1. Task stimuli and paradigm.** (**A**) Examples of animal and tool images were chosen as the stimuli, consisting of both small and large items in both categories. (**B**) On each trial, we asked participants to name the item of the image and rate how harmful it was to them or how anxious it made them on continuous scales from 1 to 100.

scale ranged from relaxed (1) to neutral (50) to anxious (100), while the harm scale ranged from safe (1) to neutral (50) to harmful (100). Thereafter, participants had to take a short quiz to ensure that they understood the task and how to use the scales correctly. They were only allowed to continue to the task if all questions were correct, else they were directed to the start of the instructions to try again (mean (M) = 1.48, standard deviation (SD) = 0.64). Each unique image was presented twice and in random order, resulting in n = 84 trials. The default value indicator for both scales was also randomized to begin between 40 and 60 for every presentation. Participants could not leave the text box blank and were required to adjust the slider on both scales before they could move on to name and rate the next image.

The experimental task was fully programmed with the JavaScript library React v.18.2.0 (https://react.dev/), developed in an app bootstrapped by Create React App (https://github.com/facebook/create-react-app), and hosted on Scalingo (https://scalingo.com/).

## Exclusion criteria

To ensure data quality, we predefined several exclusion criteria. Participants were excluded if they (i) failed the instruction quiz >5 times (N = 2), (ii) were unable to name the stimuli correctly, which was not due to variations of the item name or typo errors (N = 4), (iii) chose the same rating value for >50% of the trials (N = 1), (iv) had anxiety or harm ratings with r > 0.5 correlation with the initial default indicator value (N = 1), or (v) failed the attention check question (i.e., "Demonstrate your attention by selecting 'A lot'.") in the questionnaire section (N = 1). We also excluded participants with ratings of $\rho < 0.3$ intra-rater reliability (N = 5), $\rho$ threshold determined as 5 standard deviations away from the mean across participants' harm rating reliability of animal images (the most reliable category). Due to the remote nature of testing and data collection, participants with extraneous repeated stimuli presentation (N = 1) and incomplete dataset (N = 5) issues were excluded. We defined an initial N = 100 participants as the sample size according to prior stimuli affective rating studies also examining rating reliability (Seow & Hauser, 2022).

In total, 20 participants (20%) were excluded, leaving N = 80 participants for analysis. Of the remaining sample, participants were aged 20–55 years old (M = 37.25, SD = 8.66), with N = 37 (46.25%) identified as female.

## Questionnaires

Participants also provided basic demographic data (age and gender) and completed three self-report mental wellbeing questionnaires. We administered questionnaires assessing symptoms of depression using the Zung Self-rating Depression Scale (SDS) (Zung, 1965), trait anxiety using the trait portion of the State-Trait Anxiety Inventory (STAI-Y2) (Spielberger *et al.*, 1971) as well as obsessive-compulsive disorder (OCD) using the Obsessive–Compulsive Inventory–Revised (OCI-R) (Foa *et al.*, 2002). The presentation order of the questionnaires was randomised. The distribution of age and scores collected are depicted in Figure 2.

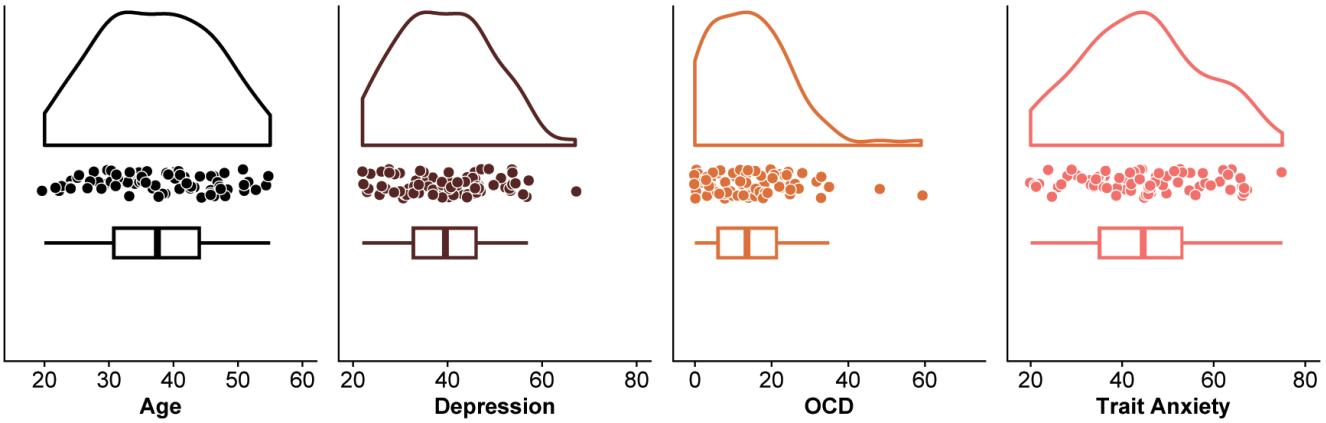


**Figure 2. Distributions of age and mental health symptom scores.** Circles represent individual ages or scores for each participant.

## Analyses

All analyses were conducted in R, version 3.6.0, via RStudio version 1.2.1335 (http://cran.us.r-project.org). We utilised the lmer() function (lmerTest package) to estimate mixed-effects models, lm() function (lme4 package) for general linear modelling, ICC() and alpha() functions (psych package) for the intraclass correlation (ICC) measure and the internal consistency measure Cronbach's alpha, and cor.test() function (stats package) for non-parametric Spearman's correlation tests to account for the non-normality of the data.

We detail the motivations for our specific analyses:

1. **Anxiety and harm ratings.** First, we expected that there may be some variation of reported anxiety and harm ratings of the stimuli, particularly between image categories. We tested if the degree of the ratings (RatingLevel) was influenced by stimuli category (Category: animal or tool) and rating scale type (ScaleType: anxiety or harm) using a mixed-effects model:

   RatingLevel ~ Category * ScaleType + Age + Gender + (1 + Category + ScaleType | Subject)

   The Category * ScaleType interaction term was not included as a random effect predictor as number of observations would be less than that of the random effects.

2. **Inter-rater similarity.** Next, we explored how similar the ratings reported by each unique stimulus were across participants. We used the two-way random-effects ICC measure to observe the agreement of ratings across participants for each stimuli presentation (i.e., images = 84) separately for anxiety and harm ratings.

3. **Internal consistency.** We also tested the internal consistency of the ratings with Cronbach's alpha, which reflected similarity of the ratings across all stimuli presentations (i.e., images = 84) separately for anxiety and harm scales.

4. **Intra-rater stimulus test-retest reliability.** To assess the consistency of ratings, we examined how reliable a participant rates each unique image across its repeated presentation. We correlated the individual ratings (anxiety or harm) of the first presentation of each stimuli with the rating of its identical, repeated presentation for each participant (i.e., images per time point = 42). We obtained a correlation measure for each rating scale type, for each participant. We then asked if this intra-rater item test-retest reliability measure (RatingReliability) was linked to stimuli Category or rating ScaleType in a mixed-effects model:

   RatingReliability ~ Category + ScaleType + Age + Gender + (1 + Category + ScaleType | Subject)

5. **Scale test-retest reliability.** We also investigated whether participants used the rating scales differently between the repeated image presentation batches. For each participant, we calculated the rating mean and standard deviation across all unique images (i.e., 42 images) by stimuli category and rating scale type, separately for the stimuli's first and second presentation. Then, we correlated these measures between the two presentations across participants for each stimuli category and rating scale type.

6. **Individual stimulus test-retest reliability.** Next, we asked if the images differed in test-retest reliability variance owing to their inherent characteristics (e.g., familiarity). For each image and each scale type, we correlated the ratings from its repeated presentations across all participants. We then probed if Category and ScaleType influenced the image reliability (StimuliReliability) using a mixed-effects model:

   StimuliReliability ~ Category + ScaleType + (1 | Stimuli)

7. **Mental health symptom associations with ratings and their reliability.** Lastly, we asked if interindividual levels of mental health symptoms was linked to the level

of anxiety or harm ratings, or to the individual's intra-rater item test-retest reliability measure. For each rating scale, we used a general linear model:

RatingLevel / RatingReliability ~ (Symptom + Age + Gender) * Category

## Results

### Anxiety and harm ratings of animal and tool stimuli

Across the participants, we generally observed that animal stimuli (anxiety: M = 47.26, SD = 14.26; harm: M = 47.37, SD = 11.75) were rated as more harmful and anxious than tools (anxiety: M = 19.91, SD = 14.14; harm: M = 16.08, SD = 10.04) (Figure 3) (Seow, 2023). We tested the statistical significance of these differences in a linear mixed model, and found a main effect of stimuli category, where ratings of tools were lower ($\beta$ = -31.30, SE = 1.68, $p < 0.001$) than animals. We also observed a rating scale and stimuli category interaction effect where tools had higher anxiety than harm ($\beta$ = 3.95, SE = 1.07, $p < 0.001$) ratings. We then ranked the stimuli according to anxiety and harm ratings in descending order by the averaged rating across repeated presentations of the same image (Figure 4). We find that the top (e.g., panther, bear, snake, tiger) and bottom (e.g., couch, wristwatch, sponge) rated stimuli were quite similar for both anxiety and harm scales.

### Inter-rater similarity

We then asked how similar the ratings given by our online sample were. We examined the degree of agreement amongst the ratings across the participants for all animal or tool stimuli with intraclass correlation measures (ICC). For all the images, we found relatively decent agreement between participants' ratings (ICC = 0.63, 95% Confidence Interval (CI) = [0.55 070]) with the harm ratings, but poor agreement between anxiety ratings (ICC = 0.45, 95% CI = [0.38 0.54]). The latter seemed to be driven by the disagreement of ratings of the tool category (anxiety: ICC = 0.19, 95% CI = [0.13 0.28]; harm: ICC = 0.31, 95% CI = [0.23 0.43]) more so than the animal category (anxiety: ICC = 0.39, 95% CI = [0.30 0.51]; harm: ICC = 0.64, 95% CI = [0.57 0.71]).

### Internal consistency

We also examined the internal consistency of the ratings for all stimuli for each scale type—whether each stimulus evoked a response that was similar compared to the rest of the other stimuli. The internal consistencies were found to be high for both rating scale types, and for both animal (anxiety: $\alpha$ = 0.95; harm: $\alpha$ = 0.94) and tools (anxiety: $\alpha$ = 0.98; harm: $\alpha$ = 0.96).

### Inter-rater stimulus test-retest reliability

Participants' consistency in their ratings were also tested in terms of test-retest reliability; their similarity over repeated trials. We examined this in our task design where each unique stimulus was presented twice. We calculated the Spearman's correlation between the two ratings of the same image for each participant and observed fairly good test-re-test reliability, where correlations across the ratings were positive for both rating scales in the animal (anxiety: M = 0.91, SD = 0.07; harm: M = 0.92, SD = 0.07) and tool (anxiety: M = 0.83, SD = 0.14; harm: M = 0.85, SD = 0.11) category. In a mixed-effects model, we quantified the statistical significance of the comparison between stimuli categories and rating scale types. We found a main effect of category, where ratings of tools were less reliable than animals ($\beta$ = -0.08, SE = 0.01, $p < 0.001$), and scale type, where harm ratings were more reliable than anxiety ones ($\beta$ = 0.02, SE = 0.01, p = 0.02) (Figure 5A). Overall, these findings suggest that animal stimuli provided more reliable ratings, especially in their evaluation of harm.

### Scale test-retest reliability

We also investigated how consistently participants utilized the rating scale over repeated stimuli presentations (Figure 5B) in terms of the average rating as well as the variance of the

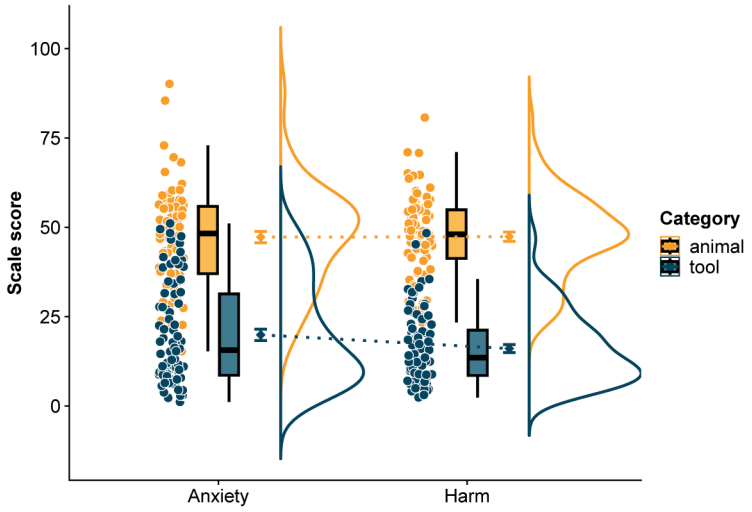


**Figure 3. Distributions of mean anxiety and harm ratings according to stimuli categories of animals and tools.** Circles represent the mean rating across all stimuli for its specific category for each participant. Animal stimuli were observed to have higher anxiety and harm ratings than tool stimuli.

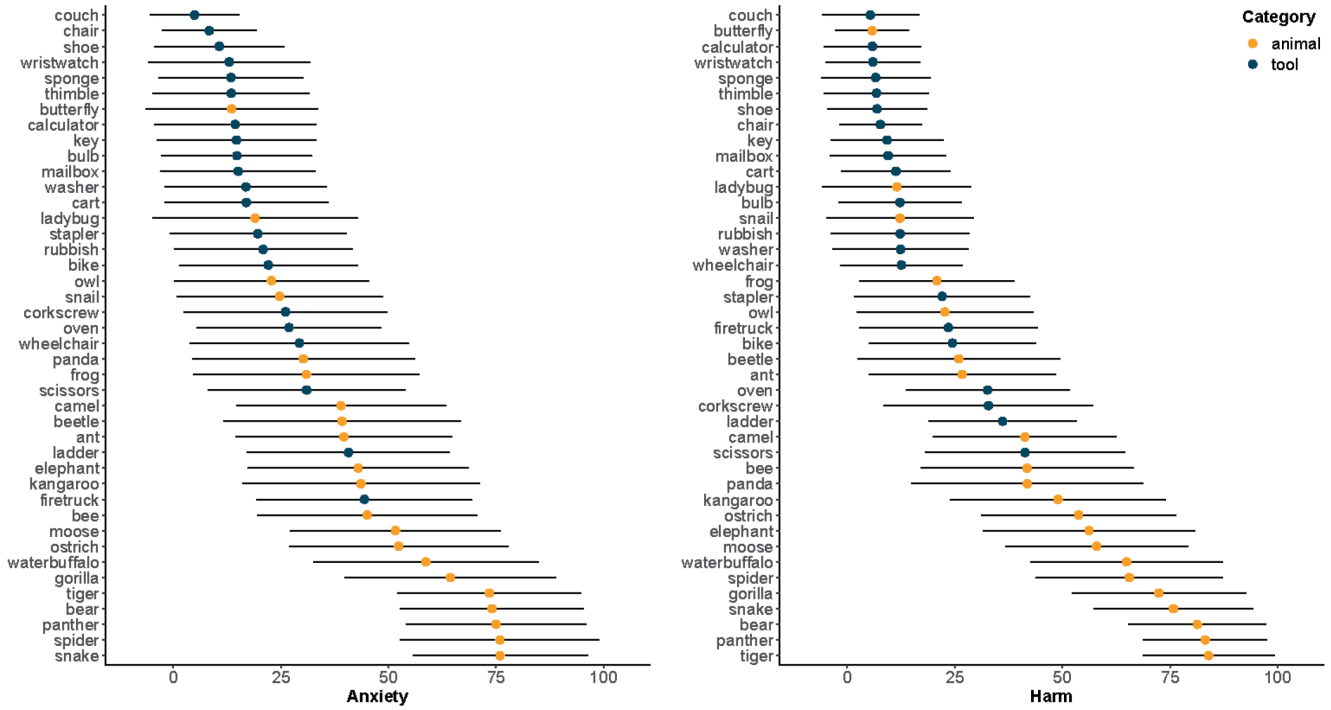


**Figure 4. Mean anxiety and harm ratings for the stimuli array across all participants.** We ranked the ratings in descending order by the averaged rating across its repeated presentations, with stimuli category identified by differing marker colour. Marker indicates rating mean and error bars indicate standard deviation for each image.

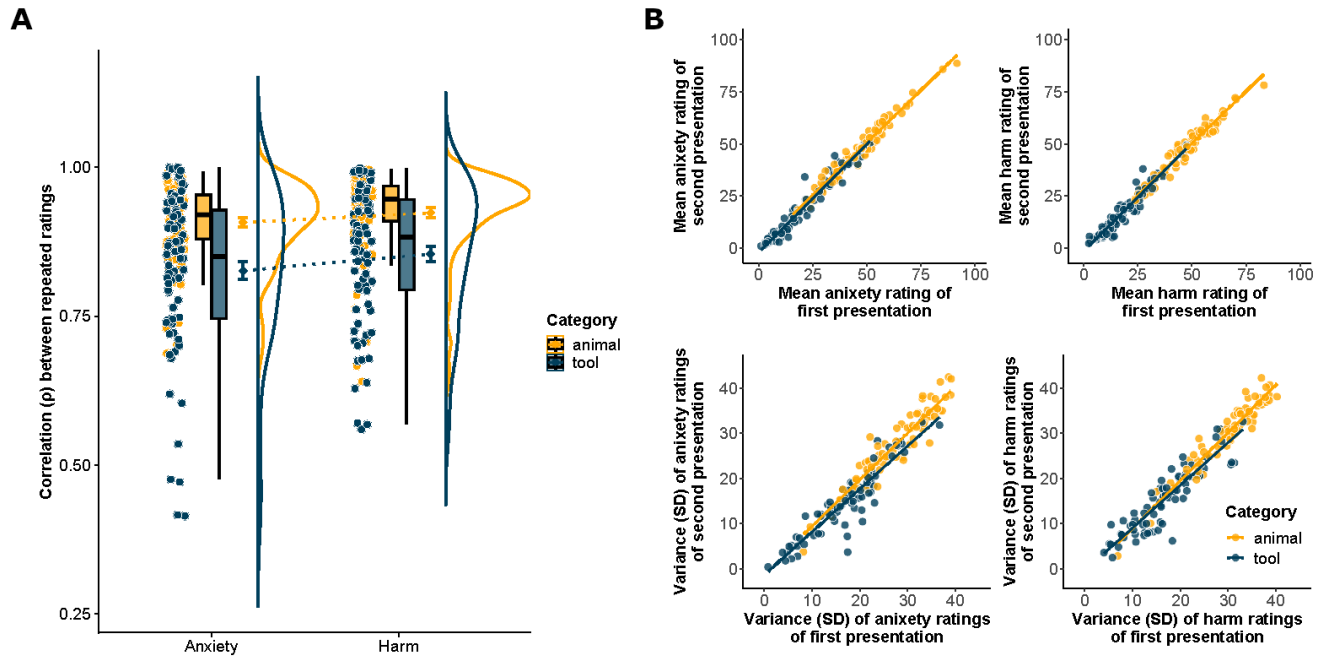


**Figure 5. Intra-rater and scale test-retest reliability.** (**A**) Distribution of Intra-rater stimulus test-retest reliability. Ratings for each stimulus were correlated with its repeated presentation per participant. Circles represent Spearman's correlation estimate across the repeated stimuli ratings, across all stimuli, for each participant. (**B**) Scale test-retest reliability by correlation of the means/standard deviation across all stimuli ratings per participant in the first (x-axis) versus second (y-axis) presentation. Circles represent the rating mean for each participant. Bold lines indicate the linear relationships.

ratings. We found that both rating scales were used quite consistently—correlations between the mean rating of all stimuli between the first and second image presentation across all participants were high for animals (anxiety: $\rho = 0.98$, $p < 0.001$; harm: $\rho = 0.96$, $p < 0.001$) and tools (anxiety: $\rho = 0.97$, $p < 0.001$; harm: $\rho = 0.95$, $p < 0.001$). The variances were also very similar across presentations, for both animals (anxiety: $\rho = 0.94$, $p < 0.001$; harm: $\rho = 0.94$, $p < 0.001$) and tools (anxiety: $\rho = 0.90$, $p < 0.001$; harm: $\rho = 0.91$, $p < 0.001$). Again, these results suggest a high reliability of the ratings reported over repeated presentations.

### Ranking of individual stimulus test-retest reliability

Across our array of 42 stimuli, we then wondered whether the individual images intrinsically differed in the consistency of their ratings across their repeated presentations. We investigated this through an analysis of the degree of correlation between the ratings from its first and second presentation, across all participants, for each stimulus. We ranked the test-retest reliability of the stimuli in descending order (Figure 6). With a linear mixed model, we found that tools garnered less reliable ratings than animals ($\beta = -0.05$, SE = 0.02, $p = 0.005$), as did the harm scale as compared to the anxiety ($\beta = -0.03$, SE = 0.01, $p < 0.005$).

### Mental health symptoms scores are not associated with affective ratings nor rating reliability

Finally, we tested if anxiety and harm ratings differed as a function of mental health symptoms. We found that none of the symptoms we tested (depression, trait anxiety, obsessive-compulsive symptoms; all total scores) showed significant links to rating scores (anxiety: $\beta > -0.01$, SE > 1.67, $p > 0.67$; harm: $\beta < 0.21$, SE > 1.37, $p > 0.39$), nor did we find a symptom interaction effect with stimuli category on rating score (anxiety: $\beta > -0.15$, SE > 1.93, $p > 0.18$; harm: $\beta > 1.34$, SE > 1.47, $p > 0.13$). We also examined if the reliability of the ratings was linked to mental health symptom severity (Figure 7). Similarly, we found that none of the symptoms significantly affected the reliability ratings (anxiety: $\beta > -0.11$, SE > 0.01, $p > 0.41$; harm: $\beta < -0.01$, SE > 0.01, $p > 0.23$), although depression had a trending effect for lower reliability ratings for the tool stimuli in both scales (anxiety: $\beta = -0.03$, SE = 0.02, $p = 0.09$; harm: $\beta = -0.03$, SE = 0.02, $p = 0.07$).

## Discussion

In this study, we explored how visual stimuli spanning various animals and tools were rated on levels of anxiety and harm. We found that animals on average were reported as more harm-perceiving and anxiety-provoking than tools. We also found that the ratings were generally reliable, especially in terms of test-retest reliabilities, with animal stimuli ratings being more reliable than tools. These data thus provide a validated database of visual stimuli with their rated anxiety and harm ratings for future use.

The first aim of our study was to observe the levels of anxiety and harm linked to the array of visual stimuli tested. Most stimuli databases have ratings on valence and arousal (Carretié *et al.*, 2019; Dan-Glauser & Scherer, 2011; Kurdi *et al.*, 2017), and sometimes dominance (Sutton *et al.*, 2019)—but not, to our knowledge, on anxiety or harm. We chose two item categories for investigation, animals and tools. Animals are well known to

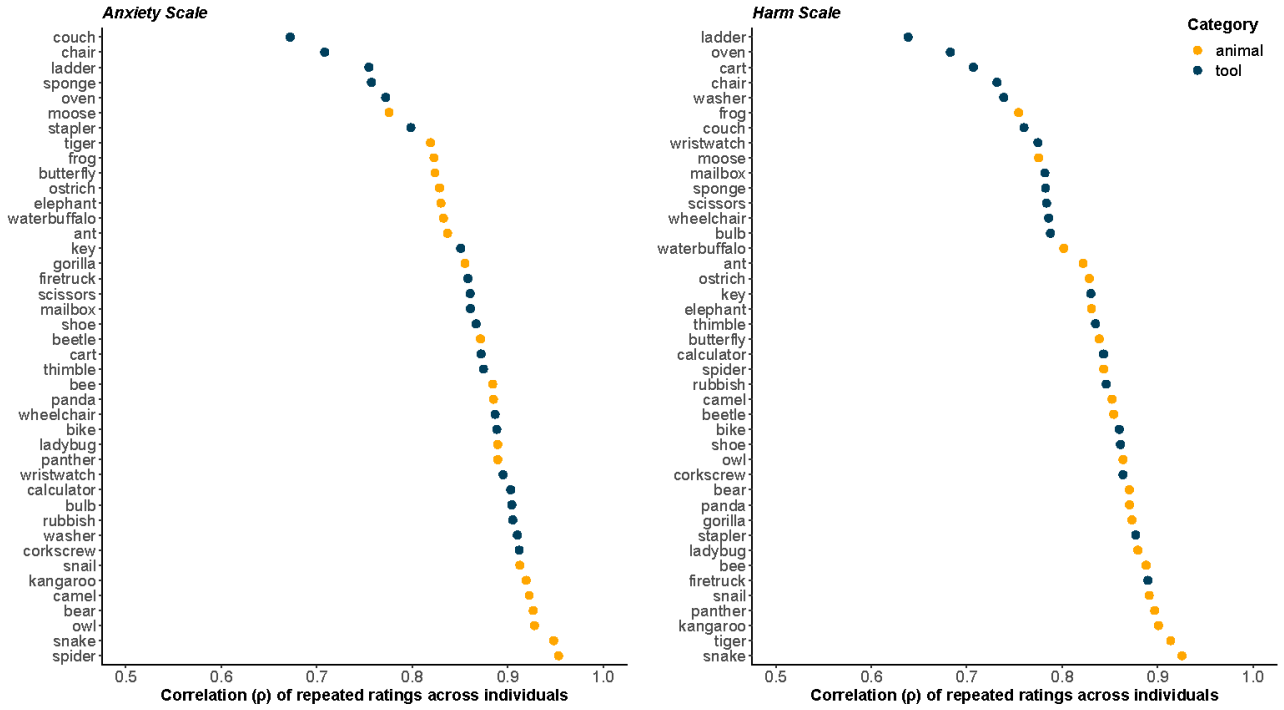


**Figure 6. Test-retest reliability of each stimuli, estimated via Spearman's correlation of anxiety or harm ratings across all participants.**

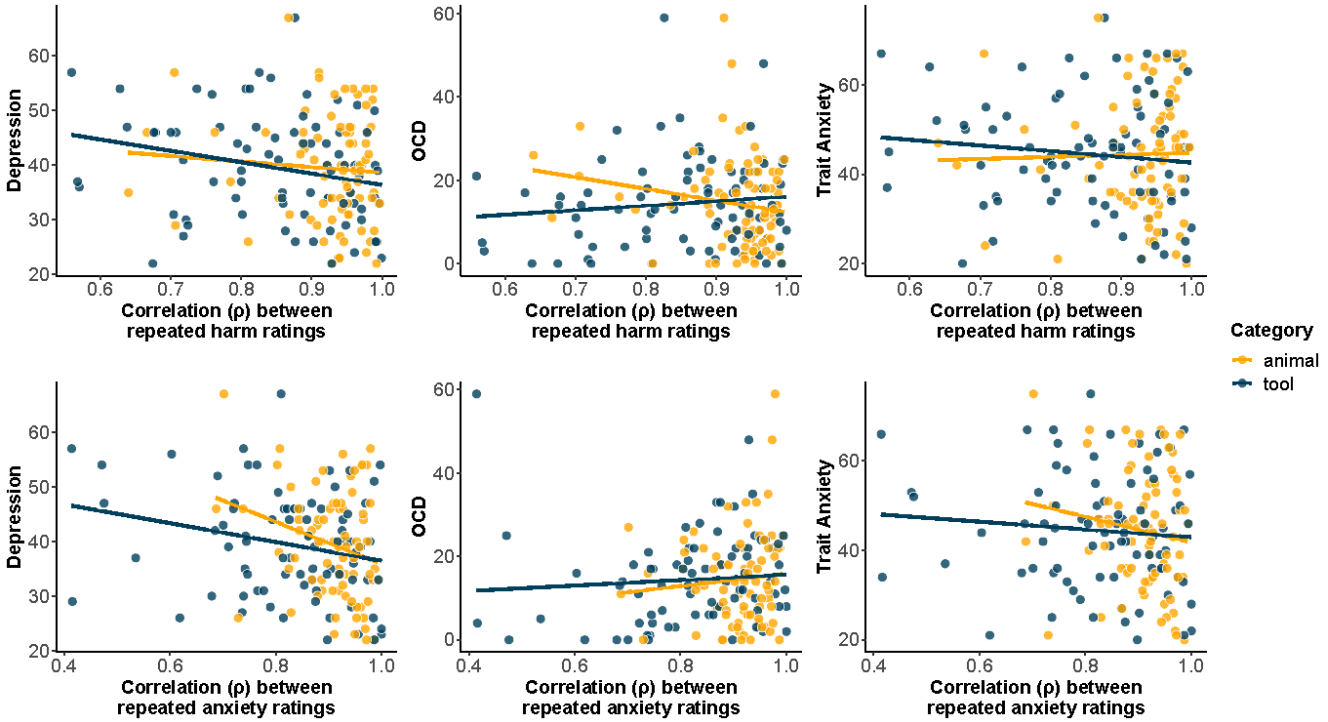


**Figure 7. Correlation of mental health symptoms and participant rating reliability.** In these figures, higher depressive scores are linked with lower harm (ρ = -0.29, p = 0.01) and anxiety (ρ = -0.23, p = 0.04) rating reliability for tools, however these associations do not pass significance when examined in a mixed-effects model controlled for age and gender. Circles represent the intra-rater item test-retest reliability measure for each participant. Bold lines indicate the linear relationship between symptom scores and reliability estimates.

provoke differences in emotion response (Possidónio *et al.,* 2019). While the emotional responses to tools are generally lower (Kurdi *et al.,* 2017), individuals with OCD are known to designate household objects as OCD-relevant (e.g. repeated checking of light switch) particularly in harm-related OCD (e.g., avoiding scissors or knives due to the fear of accidentally stabling someone). We specifically chose a variety of items in both categories, small to large, to cover a distribution of ratings. We found that animals were generally rated higher in both domains of anxiety and harm, with similar items as the top-rated stimuli, such as the panther, bear, snake or tiger, while the bottom rated items tended to be tools like the couch, wristwatch and sponge.

The second aim of our study was to test the reliability of these stimuli ratings. We examined a range of measures, finding good reliability for most of them, particularly for test re-test reliabilities. This supports the use of these visual stimuli in task paradigms, where participants are required give consistent responses over repeated presentations. Intra-rater similarity of the ratings, on the other hand, showed a difference in stimuli category. There was decent agreement of the ratings for the animal stimuli, but poor agreement for the tool stimuli. This suggests that animal stimuli elicited responses that were more similar between individuals, while the tool stimuli did not.

Finally, we characterised participants by several psychiatric symptom scores to examine if the ratings or their reliability would be linked to symptom severity. This is particularly important due to the use of these stimuli in psychiatric research, which is popularly conducted via online platforms (Chandler & Shapiro, 2016; Shapiro *et al.,* 2013). We found no consistent relationship of the ratings and their intra-rater reliabilities with symptom scores. This further supports the use of these stimuli in paradigms, where the ratings' mean level and reliability would not be confounded by mental wellbeing levels.

We also considered some limitations to this study. Firstly, we examined harm impressions from a general population sample. Diagnosed patients, particularly patients with OCD tend to have idiosyncratic OCD-relevant triggers, and 'harmful' objects may not be categorised in the same way in healthy individuals. Future studies can consider examining levels of evoked disgust, which is arguably a more universal experience and is also relevant to OCD (Bhikram *et al.,* 2017). Secondly, the test re-test reliability measure was probed by two repeated presentations relatively close in time. It remains to be seen if these ratings would be reliable between longer periods that would be important for longitudinal task designs. Lastly, a common critique of reported affective ratings is that participants may have provided ratings they recognised as

semantically characteristic of the stimuli (e.g. tigers are a dangerous animal) and not on their personal emotive evaluation (e.g. are tigers really harmful to me) (Coles *et al.,* 2023; Nichols & Maner, 2008). Nonetheless, the ratings reported are reliable and provide a starting point for affective information of anxiety and harm, which are rarely investigated.

Our current data provides a publicly available database for visual stimuli with anxiety and harm ratings with their reliabilities estimated in several measures. Particularly, we note that animal stimuli tended to have higher harm and anxiety ratings, as well as higher reliabilities of these ratings, in comparison to tool stimuli, suggesting that they may be useful for future task paradigm use.

## Data and software availability

Open Science Framework: What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli. https://doi.org/10.17605/OSF.IO/7MH63 (Seow, 2023).

All stimuli, data, and the analysis script to generate the figures in this manuscript are available in this project.



## Acknowledgements

We thank the participants for their contribution to this study.

## References

Ahn H, Picard RW: **Affective cognitive learning and decision making: The role of emotions**. 2006.
Reference Source

Begleiter H, Gross MM, Kissin B: **Evoked cortical responses to affective visual stimuli.** *Psychophysiology.* 1967; **3**(4): 336–344.
PubMed Abstract | Publisher Full Text

Bhikram T, Abi-Jaoude E, Sandor P: **OCD: obsessive-compulsive... disgust? The role of disgust in obsessive-compulsive disorder.** *J Psychiatry Neurosci.* 2017; **42**(5): 300–306.
PubMed Abstract | Publisher Full Text | Free Full Text

Blatter K, Schultz W: **Rewarding properties of visual stimuli.** *Exp Brain Res.* 2006; **168**(4): 541–546.
PubMed Abstract | Publisher Full Text

Bradley BP, Mogg K, White J, *et al.*: **Attentional bias for emotional faces in generalized anxiety disorder.** *Br J Clin Psychol.* 1999; **38**(3): 267–278.
PubMed Abstract | Publisher Full Text

Bray S, O'Doherty J: **Neural coding of reward-prediction error signals during classical conditioning with attractive faces.** *J Neurophysiol.* 2007; **97**(4): 3036–3045.
PubMed Abstract | Publisher Full Text

Brosch T, Pourtois G, Sander D: **The perception and categorisation of emotional stimuli: A review.** *Cogn Emot.* 2010; **24**(3): 377–400.
Publisher Full Text

Carretié L: **Exogenous (automatic) attention to emotional stimuli: a review.** *Cogn Affect Behav Neurosci.* 2014; **14**(4): 1228–1258.
PubMed Abstract | Publisher Full Text | Free Full Text

Carretié L, Tapia M, López-Martín S, *et al.*: **EmoMadrid: An emotional pictures database for affect research.** *Motiv Emot.* 2019; **43**: 929–939.
Publisher Full Text

Chandler J, Shapiro D: **Conducting clinical research using crowdsourced convenience samples.** *Annu Rev Clin Psychol.* 2016; **12**: 53–81.
PubMed Abstract | Publisher Full Text

Coles NA, Gaertner L, Frohlich B, *et al.*: **Fact or artifact? Demand characteristics and participants' beliefs can moderate, but do not fully account for, the effects of facial feedback on emotional experience.** *J Pers Soc Psychol.* 2023; **124**(2): 287–310.
PubMed Abstract | Publisher Full Text

Dan-Glauser ES, Scherer KR: **The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance.** *Behav Res Methods.* 2011; **43**(2): 468–477.
PubMed Abstract | Publisher Full Text

Da Victoria MS, Nascimento AL, Fontenelle LF: **Symptom-specific attentional bias to threatening stimuli in obsessive-compulsive disorder.** *Compr Psychiatry.* 2012; **53**(6): 783–788.
PubMed Abstract | Publisher Full Text

Dubois M, Hauser TU: **Value-free random exploration is linked to impulsivity.** *Nat Commun.* 2022; **13**(1): 4542.
PubMed Abstract | Publisher Full Text | Free Full Text

Foa EB, Huppert JD, Leiberg S, *et al.*: **The Obsessive-Compulsive Inventory: development and validation of a short version.** *Psychol Assess.* 2002; **14**(4): 485–96.
PubMed Abstract | Publisher Full Text

Gotlib IH, Sivers H, Gabrieli JD, *et al.*: **Subgenual anterior cingulate activation to valenced emotional stimuli in major depression.** *Neuroreport.* 2005; **16**(16): 1731–1734.
PubMed Abstract | Publisher Full Text

Hauser TU, Eldar E, Dolan RJ: **Neural mechanisms of harm-avoidance learning: a model for obsessive-compulsive disorder?** *JAMA Psychiatry.* 2016; **73**(11): 1196–1197.
PubMed Abstract | Publisher Full Text

Henrich J, Heine SJ, Norenzayan A: **The weirdest people in the world?** *Behav Brain Sci.* 2010; **33**(2-3): 61–83; discussion 83-135.
PubMed Abstract | Publisher Full Text

Kensinger EA, Garoff-Eaton RJ, Schacter DL: **Effects of emotion on memory specificity: Memory trade-offs elicited by negative visually arousing stimuli.** *J Mem Lang.* 2007; **56**(4): 575–591.
Publisher Full Text

Klorman R, Wiesknfeld AR, Austin ML: **Autonomic responses to affective visual stimuli.** *Psychophysiology.* 1975; **12**(5): 553–560.
PubMed Abstract | Publisher Full Text

Konkle T, Caramazza A: **Tripartite organization of the ventral stream by animacy and object size.** *J Neurosci.* 2013; **33**(25): 10235–10242.
PubMed Abstract | Publisher Full Text | Free Full Text

Koster R, Seow TXF, Dolan RJ, *et al.*: **Stimulus novelty energizes actions in the absence of explicit reward.** *PLoS One.* 2016; **11**(7): e0159120.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurdi B, Lozano S, Banaji MR: **Introducing the open affective standardized image set (OASIS).** *Behav Res Methods.* 2017; **49**(2): 457–470.
PubMed Abstract | Publisher Full Text

LeBel EP, Paunonen SV: **Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures.** *Pers Soc Psychol Bull.* 2011; **37**(4): 570–583.
PubMed Abstract | Publisher Full Text

Loosen AM, Seow T, Hauser TU: **Consistency within change: Evaluating the psychometric properties of a widely-used predictive-inference task.** *Psyarxiv.* 2022.
Publisher Full Text

Markett S, Montag C, Reuter M: **Anxiety and harm avoidance**. In: *Neuroimaging personality, social cognition, and character*. Elsevier, 2016; 91–112.
Publisher Full Text

Matheson GJ: **We need to talk about reliability: making better use of test-retest studies for study design and interpretation.** *PeerJ.* 2019; **7**: e6918.
PubMed Abstract | Publisher Full Text | Free Full Text

Nichols AL, Maner JK: **The good-subject effect: Investigating participant demand characteristics.** *J Gen Psychol.* 2008; **135**(2): 151–166.
PubMed Abstract | Publisher Full Text

Pessoa L: **To what extent are emotional visual stimuli processed without attention and awareness?** *Curr Opin Neurobiol.* 2005; **15**(2): 188–196.
PubMed Abstract | Publisher Full Text

Possidónio C, Graça J, Piazza J, *et al.*: **Animal images database: Validation of 120 images for human-animal studies.** *Animals (Basel).* 2019; **9**(8): 475.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Schienle A, Schäfer A, Stark R, *et al.*: **Neural responses of OCD patients towards disorder-relevant, generally disgust-inducing and fear-inducing pictures.** *Int J Psychophysiol.* 2005; **57**(1): 69–77.
**PubMed Abstract** | **Publisher Full Text**

Schupp HT, Stockburger J, Codispoti M, *et al.*: **Selective visual attention to emotion.** *J Neurosci.* 2007; **27**(5): 1082–1089.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Seow TXF: **What looks dangerous? Reliability of anxiety and harm ratings of animal and tool visual stimuli.** *OSF.* [Dataset]. 2023.
**http://www.doi.org/10.17605/OSF.IO/7MH63**

Seow TXF, Hauser TU: **Reliability of web-based affective auditory stimulus presentation.** *Behav Res Methods.* 2022; **54**(1): 378–392.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Shapiro DN, Chandler J, Mueller PA: **Using Mechanical Turk to study clinical populations.** *Clin Psychol Sci.* 2013; **1**(2): 213–220.
**Publisher Full Text**

Spielberger CD, Gonzalez-Reigosa F, Martinez-Urrutia A, *et al.*: **The state-trait anxiety inventory.** *Revista Interamericana de Psicologia/Interam J Psychol.* 1971; **5**(3 & 4).
**Reference Source**

Standing L, Conezio J, Haber RN: **Perception and memory for pictures: Single-trial learning of 2500 visual stimuli.** *Psychon Sci.* 1970; **19**(2): 73–74.
**Publisher Full Text**

Sutton TM, Herbert AM, Clark DQ: **Valence, arousal, and dominance ratings for facial stimuli.** *Q J Exp Psychol (Hove).* 2019; **72**(8): 2046–2055.
**PubMed Abstract** | **Publisher Full Text**

Tolin DF, Abramowitz JS, Brigidi BD, *et al.*: **Memory and memory confidence in obsessive-compulsive disorder.** *Behav Res Ther.* 2001; **39**(8): 913–927.
**PubMed Abstract** | **Publisher Full Text**

Zhang Z, Huang P, Li S, *et al.*: **Neural mechanisms underlying the processing of emotional stimuli in individuals with depression: An ALE meta-analysis study.** *Psychiatry Res.* 2022; **313**: 114598.
**PubMed Abstract** | **Publisher Full Text**

Zung WW: **A self-rating depression scale.** *Arch Gen Psychiatry.* 1965; **12**(1): 63–70.
**PubMed Abstract** | **Publisher Full Text**