Behavior Research Methods (2022) 54:378–392
https://doi.org/10.3758/s13428-021-01643-0

Reliability of web-based affective auditory stimulus presentation

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

Accepted: 2 June 2021 / Published online: 8 July 2021

Abstract

Web-based experimental testing has seen exponential growth in psychology and cognitive neuroscience. However, paradigms involving affective auditory stimuli have yet to adapt to the online approach due to concerns about the lack of experimental control and other technical challenges. In this study, we assessed whether sounds commonly used to evoke affective responses in-lab can be used online. Using recent developments to increase sound presentation quality, we selected 15 commonly used sound stimuli and assessed their impact on valence and arousal states in a web-based experiment. Our results reveal good inter-rater and test-retest reliabilities, with results comparable to in-lab studies. Additionally, we compared a variety of previously used unpleasant stimuli, allowing us to identify the most aversive among these sounds. Our findings demonstrate that affective sounds can be reliably delivered through web-based platforms, which help facilitate the development of new auditory paradigms for affective online experiments.

Keywords Affective auditory stimuli · Affective ratings · Web-based testing · Reliability

Introduction

Cognitive psychology and neuroscience researchers are increasingly turning to online worker platforms such as Prolific (https://prolific.ac/) and Amazon's Mechanical Turk (MTurk; https://www.mturk.com/) for recruiting participants to complete research studies (Stewart et al., 2017). Given the recent COVID-19 pandemic (Wang et al., 2020), online experimental studies may now even be necessary to circumvent restrictions in carrying out in-lab testing. Web-based testing is attractive not only for its convenience, but it also offers a wealth of advantages, ranging from its cheap and rapid data collection system allowing the collection of large sample sizes necessary for well-powered research (Gillan & Daw, 2016), to the availability of more diverse or underrepresented populations (Berinsky et al., 2012; Casey et al., 2018; Goodman et al., 2013; Levay et al., 2016; Majima et al., 2017; Shapiro et al., 2013). Moreover, there is strong evidence that web-collected data is qualitatively on par with data collected from traditional participant pools (Berinsky et al., 2012; Klein et al., 2014; Paolacci et al., 2010), with high internal reliability and test-retest reliability (Shapiro et al., 2013). Pioneering web-based studies across cognitive neuroscience, political science and mental health have indeed produced impactful and replicable findings (Rollwage et al., 2018; Rouault et al., 2018; Schulz et al., 2020; Seow & Gillan, 2020).

However, not all fields have enthusiastically adopted the online approach, owing to several criticisms of web-based testing, such as the lack of experimental control. In particular, experiments which utilize auditory stimuli must contend with limited control over audio volume adjustment as system sound settings are not accessible through the standard internet browser; a necessary security measure. Additionally, it is difficult to certify consistent quality of sound presentation across participants due to the variability in participants' audio delivery equipment and distractions in the listening environment. While the former issue is difficult to overcome, ensuring that participants wear headphones (including in-ear varieties, i.e., earphones) can help to increase sound quality and reduce interfering noises from the surrounding environment. Recently, headphone screening tests have been developed and validated online (Milne et al., 2020; Woods et al., 2017), allowing an improvement in web-based audio presentation quality.

Building on this important work, we were interested in testing the viability of sounds presented online to evoke affective responses. Affective stimuli like loud, unpleasant noises

✉ Tricia X. F. Seow
t.seow@ucl.ac.uk

[1] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square, London WC1B 5EH, UK

[2] Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London WC1N 3AR, UK

are commonly used in cognitive paradigms, for example to induce aversive states and reactions (Neumann & Waters, 2006; Oyarzún et al., 2012; Zald & Pardo, 2002). These paradigms are widely utilized, as behavior in these tasks are well-characterized with neural correlates (Büchel & Dolan, 2000; Zald & Pardo, 2002) and computational modeling (Malaka, 1999; Moutoussis et al., 2008; Tzovara et al., 2018), and are even central in investigating psychiatric disorders (Birbaumer et al., 2005; Duits et al., 2015; Hauser et al., 2016). Yet these tasks have not been adapted online (except one which designed a gamified avoidance paradigm using colliding asteroids as a threat to end the game; Wise & Dolan, 2020) as it is unknown whether affective sounds delivered through the web browser would be able to reliably and effectively evoke the expected emotional responses.

The second intention of the study was to examine which aversive sound stimulus would be most suitable for inducing negative affective responses. Known unpleasant noises vary greatly in semantic category, ranging from female screams (Lau et al., 2008; Morriss et al., 2015, 2016), metal screeches (Neumann & Waters, 2006; Zald & Pardo, 2002), and high-frequency tones (Mirz et al., 2000; Zald & Pardo, 2002) to a loud blast of white noise (Bacigalupo & Luck, 2018; Morris et al., 2001). Unpleasant noises chosen for use in previous research were often either selected by researchers *a priori*, picked from unpublished pilot studies, or chosen from sound databases with affective ratings collected from in-lab sessions that could afford a decent degree of experimental control (Bradley & Lang, 2007; Yang et al., 2018). As such, it is unknown how the valences of these previously utilized aversive stimuli compare against one another and whether biases in valence would be similar when presented online with less stringent experimental control.

Here, we capitalize on recent advances in auditory research and sound delivery measures using a headphone screening test and other technical adjustments to optimize online presentation of auditory stimuli. Our aims were twofold: to (i) investigate whether affective sounds presented through an online platform could garner reliable affective responses, and (ii) to identify an effective aversive stimulus by comparing the valence of previously utilized aversive noises. Notably, an advantage of collating ratings from an online audience facilitates the recruitment of a more diverse sample beyond that traditionally common to psychology studies (Henrich et al., 2010). This enables contact with participants from difficult-to-reach populations, allowing subjects that cannot attend in-lab experiments to participate, thus possibly giving these data greater ecological validity.

We conducted the study with a web-based task where $N = 84$ participants rated the emotional response that 15 different sounds evoked along valence and arousal dimensions over repeated presentations. Participants also completed psychiatric questionnaires on anxiety and obsessive-compulsive symptoms so that we could examine whether the severity of their psychiatric symptoms influenced the affective ratings reported or the reliability of the ratings. As these sound stimuli are commonly utilized in tasks examining these psychopathologies (Duits et al., 2015; Hauser et al., 2016), biased or poor reliability of ratings in these populations may then confound task behavioral readouts.

Overall, we found that the affective ratings had good reliability in several measures, were comparable to ratings from their original sound databases and were not associated with psychiatric severity scores. The most unpleasant sound in our array was a modified female scream (Bauer et al., 2020; Morriss et al., 2015, 2016, 2020). Our findings suggest that despite limitations of audio experimental control in web-based testing, affective responses can be reliably evoked in participants, validating the use of affective sounds for online research.

Materials and Methods

Participants Participants were recruited via Prolific (https://www.prolific.co/) ($N = 100$). They were aged between 18 and 40 years (mean [M] = 26.85, standard deviation [SD] = 5.96). We decided not to include older participants because declining sensitivity to high-frequency tones spreads to low-frequency tones from 40 years of age onwards (Moore et al., 2014). All individuals were residents of the UK and reported normal hearing. The latter meant that they had no past or current personal history of auditory/hearing difficulties including tinnitus, hearing sensitivity (e.g., hyperacusis), hearing loss, use of hearing aids, or current ear infections/inflammation. All participants provided informed consent online after reading the study information and consent pages. Participants were given 30 minutes for task completion and were reimbursed at £8.25/hour. All experimental protocols were conducted in accordance with guidelines approved by the UCL Research Ethics Committee (project ID number 15301\001).

Exclusion criteria Several predefined exclusion criteria were applied to ensure data quality. Participants were excluded (i) if their final chosen upper frequency threshold (see "Sound frequency calibration") was below 8000 Hz ($N = 11$), which indicated either faults in the auditory setup or a failure to understand instructions, or (ii) if they failed the attention check question (i.e., "Demonstrate your attention by selecting 'A lot'.") in the questionnaire section ($N = 2$). Participants with incomplete datasets due to remote data collection issues were also excluded ($N = 3$). In total, 16 participants (16%) were excluded, leaving $N = 84$ participants for analysis. Of the remaining sample, 43 (51.19%) identified as female, 40 (47.62%) as male and one (1.19%) as other gender. Note that

our sample size ($N = 84$) was more than three times that of prior rating studies (e.g., each sound was rated by $\geq N = 22$ in Yang et al., 2018).

Sound array A sound bite array of 15 sounds was selected to contain a range of (predominantly) unpleasant and pleasant noises. The majority of the unpleasant sounds were chosen from prior studies that had already identified and/or utilized them as an aversive stimulus, while the rest of the sounds were intended to balance out the sound assortment to avoid biased ratings. Across the array, we purposefully covered a variety of semantic categories, extending from signal noises to pure tones, to synthetic sounds and to naturalistic sounds.

The sound array was assembled from a variety of sources:

i) Four sounds from the open-source sound database Expanded Version of the International Affective Digitized Sounds (IADS-E) (Yang et al., 2018): female scream (ID: 0276), cicada (ID: 0335), sea wave (ID: 0921) and a piano melody (ID: 1360)
ii) Four sounds generated using the sound editor software Audacity (http://audacityteam.org/) version 2.4.2: pink noise, Brownian noise, 800 Hz sine tone and 5000 Hz sine tone, all with 1500 ms duration at 0.8 amplitude
iii) One sound from a collaborative database of Creative Commons Licensed sounds, Freesound (https://freesound.org): a dentist's drill (https://freesound.org/people/alexanderwendt/sounds/385680/)
iv) Three sounds from previous studies of aversive learning: white noise (Bacigalupo & Luck, 2018), a modified female scream (Bauer et al., 2020; Morriss et al., 2015, 2016, 2019, 2020) (which was altered from the original sound in the second version of the International Affective Digitized Sounds database (IADS-2) (Bradley & Lang, 2007), ID: 277), high-frequency tone and a metal screech (Zald & Pardo, 2002)
v) Two sounds of 2000 ms high-frequency tones generated in-browser from frequencies based on participants' input (see "Sound frequency calibration") with *react-tone* (https://www.npmjs.com/package/react-tone) version 1.1.1.

All sounds were cut to 2000 ms or 1500 ms in Audacity if the original file was longer. See Supplemental file 2 for a detailed list of all sounds and their specific sound length information. See Open Practices Statement for the availability of these stimuli.

Procedure The experimental task was programmed with ReactJS and bootstrapped with Create React App (https://create-react-app.dev/). The use of headphones (including earphones) was required in this study. After participants provided online consent, they were directed to adjust their volume settings, which was intended to help avoid presentation levels that would result in uncomfortably loud or inaudible sounds, before performing a headphone check test. Once through, participants were told to indicate their maximum audible frequency level (see "Sound frequency calibration") that would subsequently inform two sounds of the sound array (see "Sound array"). Participants then rated how each sound made them feel along valence and arousal dimensions, and completed three psychiatric questionnaires. Details of the procedures are further detailed below:

Volume calibration. Volume intensity depends on both the computer system and browser sound settings. As the former cannot be altered by the experimenter, we first instructed participants to adjust their computer sound setting to 30% of the system maximum. Thereafter, we relied on adjustments of the browser-based volume to manipulate loudness. To do this, participants were presented with a noise sample which could be adjusted in volume via browser sound settings from 0 to 100 along a logarithmic scale of lower bound 1 and higher bound 100 (as changes in perceived hearing are better described along a logarithmic scale). Participants were told to adjust the indicator of the scale (initial value = 80) until the sound bite volume was loud but comfortable, allowing as many repeated sound presentations as needed. The final set volume level was subsequently used for the rest of the task.
Headphone screening test. We required participants to use headphones, as this improves the control of sound delivery and attenuates environmental interference. Participants needed to pass a headphone check task (Woods et al., 2017) to ensure that they were wearing headphones. The task consisted of six questions where they had to discriminate the sound intensity of several tones (i.e., which was the quietest), where one was presented with a phase difference of 180° between stereo channels. The sound volumes are difficult to discriminate with loudspeakers (as phase-cancellation is imperfect), but easy with headphones. The sound stimuli of the headphone check task were presented in randomized order. Participants were required to answer at least 5/6 of the questions correctly to proceed to the next stage; otherwise they were directed back to the volume adjustment module to calibrate their sound settings before attempting the headphone test again. Most participants passed the headphone check on the first attempt ($N = 68$, 80.95%), while 15.48% of the participants took the test twice ($N = 13$), and the remainder ($N = 3$, 3.57%) took more tries, up to a maximum of five attempts. Participants were explicitly instructed not to adjust their sound settings prior to each of the subsequent audio sections.
Sound frequency calibration. This part of the experiment was intended to determine the individual's upper

frequency threshold, which was subsequently used to select high frequencies as unpleasant noise stimuli. Sensitivity to the high-frequency range tends to decline with age (Lee et al., 2012; Moore et al., 2014), and as such we sought to determine the maximum audible frequency for each individual which we could then derive subjectively high but audible frequency tones as part of the sound array. In this section, participants were shown a logarithmic (as frequency perception follows a logarithmic perception) scale that represented the frequency of a sine tone, with the indicator at an initial frequency of 8000 Hz. They were told to adjust the frequency of the tone three times (first scale: ranged from 5000 Hz to 20,000 Hz, second and third scales: 1000 Hz less than previously chosen frequency to 20,000 Hz; all logarithmic) to a pitch where they could just hear the tone. We did not find a significant influence of age on the final chosen frequency level in our relatively narrow participant age range (Supplementary Fig S1). Two frequencies were derived from this upper frequency threshold which were subsequently used as part of the sound bite array for the ratings task—one three-fourths (Frequency 1) and one half (Frequency 2) of the chosen frequency level set on a logarithmic scale of 10,000–20,000 Hz.

Auditory ratings. The participants rated how each of the sounds made them feel on two affective scales: valence and arousal. These dimensions are considered classic, primary dimensions of emotion that account for most of the variance in emotional judgments (Bradley & Lang, 1994) and were measured in prior rating studies (Bradley & Lang, 2007; Yang et al., 2018). We used 0 to 100 continuous scales. For valence, the scale ranged from very unpleasant (0) to neutral (50) to very pleasant (100), while the arousal scale ranged from very sleepy (0) to neutral (50) to very awake (100). Each of the 15 sound bites was presented four times (60 trials in total); twice at the same volume set at the beginning (100%) (see "Volume calibration"), and twice more at half the logarithmic scale (lower bound 1 and higher bound 100) of the original volume setting (50%). This was intended to enable test-retest reliability analyses comparing affective ratings across the repeated presentations of the specific sound stimuli. Sounds were presented in a randomized order.

The default value indicator for both affective scales was also randomized to begin between 35 and 65 for every presentation. Participants were required to adjust the slider on both scales before they could move on to rate the next sound. The participants had the option to play the sounds multiple times during rating, but sounds were generally played once (*M* per participant = 1.09, *SD* = 0.16). Participants could also indicate with a checkbox if they could not hear the sound; none of the sound presentations for any participant were signaled as such.

Questionnaires. Lastly, participants provided basic demographic data (age and gender) and completed three self-report psychiatric questionnaires. We administered questionnaires assessing symptoms of state (STAI-Y1) and trait anxiety (STAI-Y2) using both scales of the State-Trait Anxiety Inventory (Spielberger et al., 1983) as well as obsessive-compulsive disorder (OCD) symptoms using the Obsessive–Compulsive Inventory–Revised (OCI-R) (Foa et al., 2002). The presentation order of the questionnaires was randomized.

Analyses All analyses were conducted in R, version 3.6.0, via RStudio version 1.2.1335 (http://cran.us.r-project.org). The intraclass correlation (ICC) measure and the internal consistency measure, Cronbach's alpha, were estimated with the *ICC()* and *alpha()* functions, respectively, from the *psych* package. For correlation tests, we used nonparametric Spearman's correlation tests with no ties to account for the non-normality of the data, which were conducted with the *cor.test()* function from the *stats* package. All mixed-effects models were estimated with the *lmer()* function from the *lme4* package with the *lmerTest* package for statistical tests. Linear regression models were performed with the *lm()* function from the *stats* package. Paired *t* tests were calculated with the *t.test()* function from the *stats* package.

Reliability measures

1. *Inter-rater reliability.* First, we would expect that participants would rate each unique sound somewhat similarly, but with some variation reflecting interindividual differences. To examine the agreement of ratings between participants for each sound item presentation (i.e., sound items = 60), we estimated the intra-class correlation (ICC) of the ratings across participants with a two-way random-effects model, separately for both valence and arousal ratings.
2. *Intra-rater item test-retest reliability.* Next, we asked how reliable a participant rates each unique sound across its repeated presentation. We tested this by correlating the individual ratings (valence or arousal) of the first presentation of each sound with the rating of its identical, repeated presentation (same volume level) for each participant (i.e., sound items per time point = 30). From this, a correlation estimate was obtained for each specific scale type and volume level, for each participant. We then tested whether volume (*Volume*: low [50%] or high [100%]) or affective scale type (*ScaleType*: valence or arousal) were linked to this intra-rater item test-retest reliability correlation measure (*RatingReliability*). Both *Volume* and

ScaleType were taken as factors. We used a mixed-effects model in which *Volume, ScaleType* were fixed effect covariates, with them and the intercept taken as random effect predictors. The model was: *RatingReliability* ~ *Volume* + *ScaleType* + (1 + *Volume* + *ScaleType* | *Subject).*

3. *Sound test-retest reliability.* The sounds themselves may also differ in test-retest reliability variance owing to their inherent characteristics (e.g., emotional content). To probe whether a sound's features influenced its consistency of ratings between repeated presentations, we correlated the affective ratings from the two repeated presentations across all participants for each sound at its unique volume level. As such, correlation was examined for each sound, for each affective scale type and volume level. We additionally tested whether *Volume* and *ScaleType* influenced the sound reliability (*SoundReliability*) using a mixed-effects model where *Volume*, *ScaleType* were fixed effect covariates, with them and the intercept taken as random effect predictors. The model was: *SoundReliability* ~ *Volume* + *ScaleType* + (1 + *Volume* + *ScaleType* | *Sound*).
4. *Scale test-retest reliability.* We also asked whether there were differences in participants' ratings across the repeated sound presentations in terms of their overall use of the rating scale for each sound presentation batch. For this, we calculated the ratings' means and standard deviations by volume and affective scale type across all sounds for each participant, separately for the sounds' first and second presentation. We then examined the correlation between the means and standard deviations obtained across the two presentations across participants. Thus, a correlation measure for rating mean/standard deviation was obtained for each affective scale type and volume level.
5. *Internal consistency.* Internal consistency of the affective ratings was measured in prior studies (Bradley & Lang, 2007; Yang et al., 2018), which reflected similarity of the ratings across all sound items. Likewise, we measured the consistency of the ratings using Cronbach's alpha for all sound item presentations (i.e., sound items = 60) for both valence and arousal scales.
6. *Correlation with prior ratings.* To examine the robustness of our ratings which were garnered online versus in-lab rating studies, we compared the affective ratings of five sounds from the current online study (using the averaged rating across repeated presentation at 100% volume) with the ratings from their original lab-based study (Bradley & Lang, 2007; Yang et al., 2018). We used Pearson's correlation in this measure, as we were interested in the numerical rather than ordinal relationship between the ratings.

Ranking affective ratings We ranked the sounds in order of valence and arousal ratings. For these analyses, we organized them in descending order by the averaged rating across repeated presentations of the same sound at its specific volume. To test for significance in rating differences between sounds in order to identify the top- and bottom-ranked valence/arousal sounds, we conducted paired *t* tests between the two top/bottom-ranked sounds for each scale type. We also asked whether affective ratings were influenced by the volume of the presented sound by testing whether the averaged affective ratings across both presentations for a specific scale type (*ScaleRating*, for arousal or valence, which was the rating measure used in subsequent analyses) were associated with *Volume*, using a mixed-effects model for both affective scale types. The model was: *ScaleRating* ~ *Volume* + (1 + *Volume* | *Subject).* For further fine-grained examination of the impact of volume on each sound, we performed individual paired *t* tests between each volume pair for all sounds for each scale type. Note that *t* test analyses here were intended to serve as a broad illustration of rating differences between sounds/volume; the *t* test statistics were not corrected for multiple comparisons.

Unique high-frequency tones High-frequency tones are known to be unpleasant (Zald & Pardo, 2002). However, upper hearing frequency limits might differ for participants given the effect of age on hearing thresholds (Lee et al., 2012). Though we attempted to circumvent this issue with our age exclusion criteria (≤40 years), we also allowed participants to identify subjectively high frequencies for themselves (Frequency 1 and Frequency 2; see "Sound frequency calibration") in addition to having two separate sine wave stimuli with frequencies set at 800 Hz and 5000 Hz. Therefore, unique frequencies were played across the participants for Frequency 1 and Frequency 2 in the sound array. We thus tested whether the unique frequencies that participants heard (*Frequency*, z-scored) influenced their affective ratings (*ScaleRating*) for valence and arousal dimensions with the model: *ScaleRating* ~ *Frequency* + (1 + *Frequency* | *Subject*).

Questionnaires Finally, we asked whether mental health symptoms influenced the rating or the rating reliability of these stimuli, which could have ramifications for experiments with populations with high percentages of these psychopathologies. For this, we tested whether the affective ratings (*ScaleRating*) or intra-rater item test-retest rating reliability (*RatingReliability*) were associated with the psychiatric symptom severity, calculated as the total score for the individual questionnaire (*QuestionnaireScore*, z-scored), and whether that main effect was modulated by *Volume* for both valence and arousal dimensions. For this, we used a mixed-effects model: *ScaleRating* ~ *QuestionnaireScore* * *Volume* + (1 + *Volume* | *Subject*) and a linear model: *RatingReliability* ~ *QuestionnaireScore* * *Volume*. Also see Supplementary Fig.

S5 and Supplementary Fig. S6 for questionnaire score distributions and correlations, respectively.

Results

Volume and frequency adjustment

Volume adjustment First, to ensure that sounds presented in this study were appropriate (i.e., not dangerously loud), we let participants calibrate the browser sound volume to a level that was comfortable for them (see "Volume calibration"). We subsequently used the adjusted sound volume from the successful attempt of the headphone task (see "Headphone screening test") for the rest of the experiment. Participants generally chose a volume level on the scale that was close to the default of 80 (mean [*M*] = 76.49, standard deviation [*SD*] = 16.97), with a range of 27 to 100 (Fig. 1a).

Frequency adjustment Next, we attempted to identify frequency thresholds for each participant in order to select subjective high-frequency pitches for each participant as an unpleasant noise. Participants chose 16,141.15 Hz on average (*SD* = 2516.79 Hz, ranging from 8184 Hz to 22,000 Hz) (Fig. 1b) as the pitch where they could just hear the tone. These frequency levels are in line with hearing threshold studies where the 22–35 age group sample is thought to have approximately 16,000 Hz (at 63.03 dB, a relatively moderate to loud volume) as their upper frequency threshold (Lee et al., 2012). For the ratings task, we used sine tones of frequency levels which were slightly lower than their self-reported threshold for each participant (see "Sound frequency calibration"): Frequency 1, which was 70% (log-scaled) of the threshold (*M* = 8033.45 Hz, *SD* = 959.89 Hz, min = 4838.65 Hz, max = 10,158.21 Hz), and Frequency 2, which was 50% (log-scaled) of the threshold (*M* = 4004.45 Hz, *SD* = 326.87 Hz, min = 2860.77, max = 4690.42 Hz).

High reliability of affective ratings

Inter-rater reliability In this study, we sought to examine the reliability of the affective ratings by our online participant pool. First, we would expect that participants would give somewhat similar ratings for each unique sound, with some variation reflecting individual differences. We examined this degree of agreement amongst the ratings across participants for all sound items with an intraclass correlation measure (ICC) estimated for both valence and arousal ratings. We found that there was decent agreement between participants' ratings for all sounds in the valence domain (ICC = 0.73, 95% Confidence Interval (CI) = [0.64 0.77]), while arousal ratings were more diverse (ICC = 0.51, 95% CI = [0.44 0.59]). This suggested that participants indeed tended to rate the same sounds similarly, but there were still considerable individual differences, particularly in the arousal domain.

Intra-rater item test-retest reliability Next, we also aimed to measure how consistent participants were in their sound ratings over repeated trials. To do so, each of the sounds (at two volume levels, 50% and 100%) was played twice to examine how similarly the same sound was rated across two presentations, and we calculated the Spearman correlation (ρ) between the two ratings of the same sound played for each participant at its specified volume level. We found that the correlations across the ratings of these repeated sound presentations were positive and high for all participants in both valence (*M* =

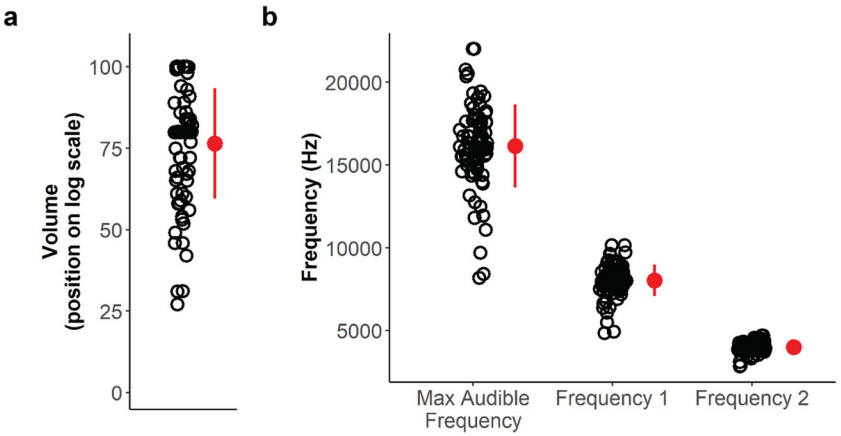

Fig. 1 Adjusted volume and frequency levels. **a** Distribution of adjusted volume of browser sound settings. Participants were instructed to keep their computer system sound settings at 30% of the maximum before calibrating the browser volume on a (log) scale to a level that was appropriate (not uncomfortably loud) for them. **b** Distribution of the maximum audible frequency selected, and the frequencies derived to use in the sound array for subsequent rating which were unique to each participant. Frequency 1 was derived from 75% (log-scaled), and Frequency 2 was 50% (log-scaled), of their threshold frequency. Circles in the graphs represent volume/frequency level per participant. Red marker indicates mean, and error bars indicate standard deviation

0.89, $SD = 0.08$) and arousal ($M = 0.79$, $SD = 0.14$) (Fig. 2a), showing good test-retest reliability. We generally observed that valence ratings had a stronger positive correlation between repeated presentations than arousal ratings ($\beta = 0.11$, $SE = 0.02$, $p < 0.001$) (Fig. 2a), indicating that participants gave more reliable valence than arousal ratings. There was no significant impact of volume on correlation strength between the repeated ratings ($\beta = 0.02$, $SE = 0.01$, $p = 0.10$). Overall, these findings suggest that participants rated sounds very similarly across their repeated presentation on both valence and arousal scales, confirming the ratings as a reliable measure of emotional response.

Sound test-retest reliability Across our array of 15 sounds, we also wondered whether the sounds individually intrinsically differed in the consistency of their ratings across their repeated presentations. Therefore, for each sound, we examined the degree of correlation between the ratings from its first and second presentation, across all participants. Overall, all sounds were able to reliably induce valence ratings at both 50% ($M = 0.68$, $SD = 0.06$) and 100% volume ($M = 0.70$, $SD = 0.07$) (Supplementary Fig. S2). At 100% volume, the sound which gave the most reliable ratings was the modified female scream of Morriss et al. ($\rho = 0.78$, $p < 0.001$), while the least reliable was the Pink Noise stimulus ($\rho = 0.56$, $p < 0.001$) (Supplementary Fig. S2). Similarly, arousal ratings were generally decently reliable across all sounds at both 50% ($M = 0.62$, $SD = 0.09$) and 100% volume ($M = 0.66$, $SD = 0.06$), with Frequency 1 ($\rho = 0.75$, $p < 0.001$) as the most reliable arousal rating, while the least reliable was Brownian Noise ($\rho = 0.54$, $p < 0.001$) at 100% volume (Supplementary Fig. S2). Similar to intra-rater item test retest-reliability, valence ratings had a stronger positive correlation between repeated presentations than arousal ratings ($\beta = 0.05$, $SE = 0.02$, $p = 0.01$), while the impact of volume on correlation strength between the repeated ratings did not reach significance ($\beta = 0.03$, $SE = 0.02$, $p = 0.06$).

Scale test-retest reliability We also investigated how consistently participants utilized the rating scale over repeated sound presentations. We first estimated the rating mean and standard deviation for all sounds for the first and second sound presentation separately for each participant, and then compared these over all participants. We found that participants gave comparable rating patterns over the repeated presentations in both valence and arousal scales. Correlations between the mean rating of all sounds between the first and second sound presentation across all participants were high (valence: $\rho = 0.87$, $p < 0.001$; arousal: $\rho = 0.87$, $p < 0.001$) and the variances were also very similar (valence: $\rho = 0.88$, $p < 0.001$; arousal: $\rho = 0.86$, $p < 0.001$) (Supplementary Fig. S3). Again, these results support the high reliability of participants providing affective ratings for the sounds presented online.

Internal consistency Prior studies have also reported good internal consistency of sound ratings in both valence and arousal affective dimensions (Yang et al., 2018). Similarly, we estimated Cronbach's alpha for all sounds per scale type. The top seven valence sounds (100% and 50% volume for Cicada, Sea Wave and Piano Melody, and 50% volume for Brownian Noise), which were the same as the bottom seven arousal sounds, were automatically identified as being negatively correlated with their respective scale types. This indicated that our sound array had a balanced range of sounds with

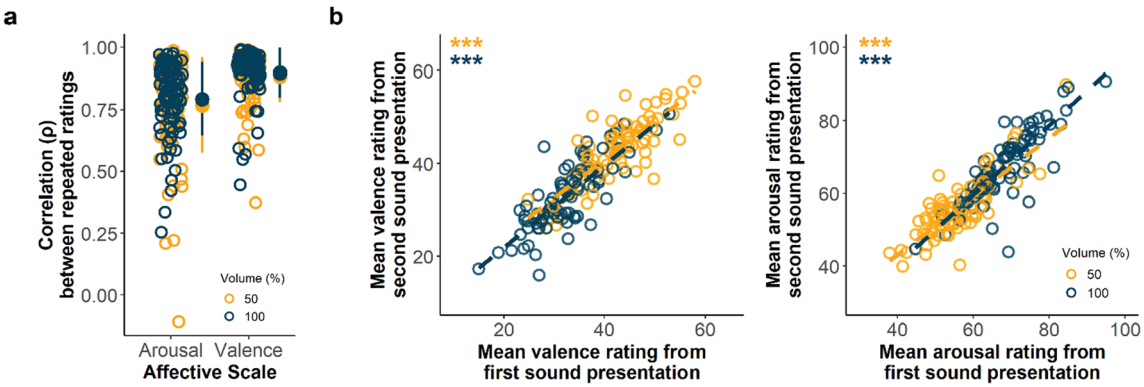

Fig. 2 Intra-rater and scale test-retest reliabilities of affective ratings. **a** Distribution of intra-rater item test-retest reliability. Affective ratings for each sound were correlated with those from its repeated presentation per participant as the intra-rater item reliability measure. These were found to be high for both arousal and valence in both volumes. Circles represent Spearman's rank correlation estimate ρ across repeated sound ratings for each participant across all sounds at a specified volume/scale. Marker indicates mean and error bars indicate standard deviation. **b** Scale test-retest reliability as measured by the comparison of the means across all sound ratings per participant in the first (x-axis) versus second (y-axis) stimulus presentation. Strong positive correlations between the measures for both arousal and valence indicate reliable rating patterns. Circles represent the rating mean for each participant at its specified volume, with the dashed line indicating the linear relationship between means. ***$p < 0.001$ (Spearman's correlation). See Supplementary Fig. S3 for the correlation between rating standard deviations in the first versus second sound presentation

differing valence and arousal. The internal consistencies were found to be high for both scale types after automatic sign reversion of those sounds in the estimation (valence: $\alpha = 0.93$, arousal: $\alpha = 0.94$).

Correlation with prior ratings Finally, we asked whether the ratings we collected online were comparable to those from in-lab studies to test whether the decreased degree of experimental control of web-based audio delivery impacted the affective ratings. We gathered the ratings of five sounds in our stimulus array (played at 100% volume) to compare them with those from their original studies, which were all conducted in traditional laboratory environments (Bradley & Lang, 2007; Yang et al., 2018). Pearson's correlations of the sounds across ours and the prior studies for both affective dimensions were high (arousal: $r = 0.98$, $p = 0.003$; valence $r = 0.96$, $p < 0.01$) (Fig. 3). These results suggest that the affective ratings we collected online were reliable and valid, and rated similarly as those measured in well-controlled in-lab environments.

Ranking of affective ratings

As part of the study, we aimed to identify the most unpleasant sound amongst a variety that have been utilized in prior research. Female screams were found to be the most unpleasant, with the modified female scream of Morriss et al. ($M = 8.48$, $SD = 12.09$) being rated as significantly more unpleasant than the scream from the IADS-E database (Yang et al., 2018) ($M = 11.24$, $SD = 12.68$) (at 100% volume for both sounds: $t = -1.99$, 95% CI = [−5.52, −0.008], $p < 0.05$) (Fig. 4). This was followed by high-frequency tones, metal sounds, complex noises, and finally natural pleasant noises. The most pleasant sound was the Piano Melody ($M = 88.57$, $SD = 10.88$), which was rated as significantly more pleasant than Cicada sounds ($M = 81.57$, $SD = 13.81$) (at 100% volume for both sounds: $t = 4.01$, 95% CI = [4.39, 13.02], $p < 0.001$) (Fig. 4).

We also observed that the ratings in valence and arousal dimensions for the current sound array were negatively correlated (Spearman: $\rho: = -0.83$, $p < 0.001$) (Supplementary Fig. S4). As such, the sounds followed almost the same (inverted) ranking pattern for arousal ratings. The Morriss et al. modified female scream was identified as the most arousing ($M = 90.35$, $SD = 11.51$), being more arousing than the IADS-E female scream ($M = 84.50$, $SD = 13.34$) (at 100% volume for both sounds: $t = 4.42$, 95% CI = [3.21, 8.48], $p < 0.001$) (Fig. 4). On the other hand, the 50% volume Piano Melody was rated as the least arousing sound ($M = 23.54$, $SD = 19.03$), being significantly lower than the Cicada sound ($M = 32.23$, $SD = 18.81$) (at 50% volume for both sounds: $t = 3.51$, 95% CI = [1.75, 6.33], $p < 0.001$).

Overall, the ratings suggest that the Morriss et al. modified female scream may be the most ideal aversive stimulus amongst those tested, as it was ranked as the most unpleasant and arousing sound.

Volume intensity increases arousal and unpleasantness

High sound intensity is often a factor in inducing aversive responses (Liberman et al., 2006; Neumann et al., 2008), and thus we sought to examine how volume influenced affective ratings. We found that both valence and arousal ratings were modulated by volume. Louder sounds were generally

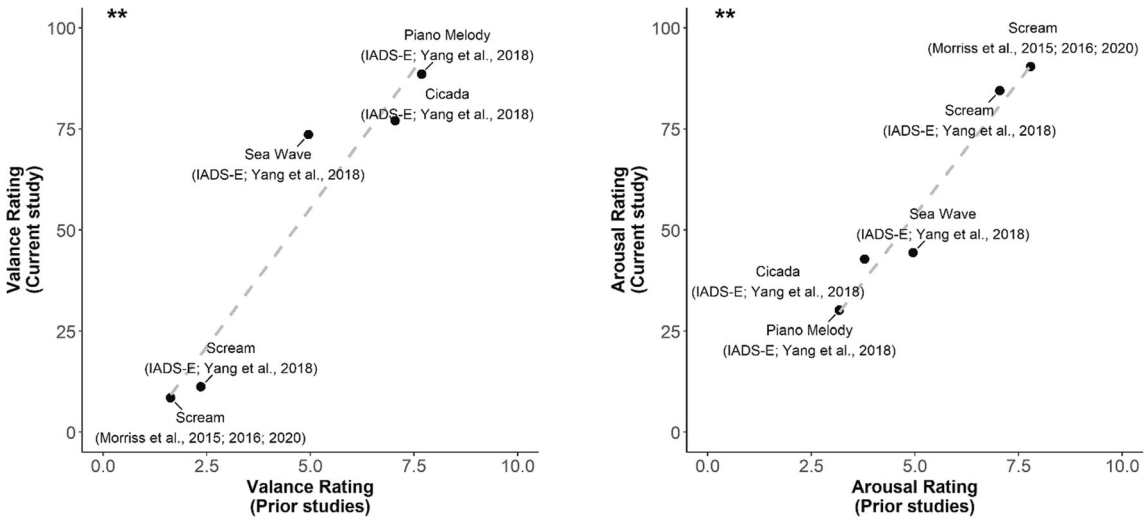

Fig. 3 Rating comparison with prior studies. Relationship between current affective ratings and those from prior studies (Bradley & Lang, 2007; Yang et al., 2018). Pearson's correlation across the five sounds in both rating dimensions was high (arousal: $r = 0.98$, $p = 0.003$; valence: $r = 0.96$, $p < 0.01$), indicating that our ratings collected online are very similar to those from traditional in-lab studies. (n.b. The Morriss et al. scream is a modified version of the original sound bite from the IADS-2 (Bradley & Lang, 2007), whose ratings were used for comparison here.) $^{**}p < 0.01$ (Pearson's correlation). Rating scale from prior studies ranged from 0 to 10 while that of the current study ranged from 0 to 100

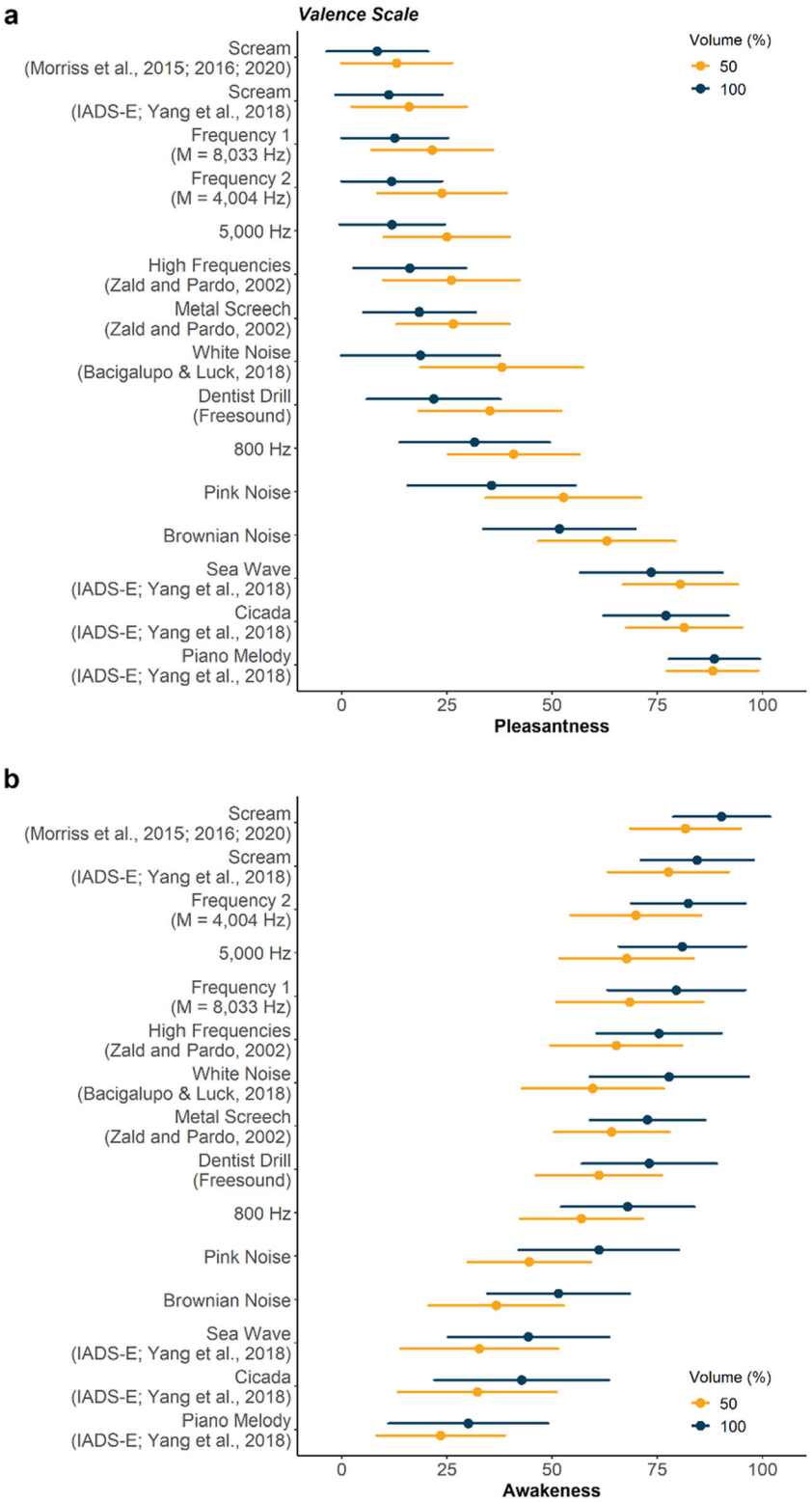

Fig. 4 Mean affective ratings for the sound array across all participants. We ranked the affective ratings from the **a** valence and **b** arousal scales to identify the most unpleasant sound in our array, split by volume. The Morriss et al. scream was found to be the most unpleasant and arousing sound. Frequency 1 and Frequency 2 were unique for every individual, based on their calibrated frequency threshold (see "Sound frequency calibration"), but the mean of the frequencies across all participants for those sounds are indicated in their labels. Marker indicates rating means and error bars indicate standard deviations for all sounds. Sounds are ranked by the averaged rating across volumes. See Supplementary file 2 for these data

rated as having lower valence (unpleasant) ($\beta = -11.98$, $SE = 2.33$, $p < 0.001$) and being more arousing (awakeness) ($\beta = 22.96$, $SE = 1.82$, $p < 0.001$). Separate paired t tests for ratings of every sound between the two volumes indicated that this was true for all stimuli (valence: all ts > 3.72, ps < 0.001, arousal: ts > 4.17, ps < 0.001), except Piano Melody in the valence dimension ($t = -0.49$, 95% CI = [−1.88, 1.14], $p = 0.62$) (Fig. 4). The sound that reported the largest volume-induced difference in both valence ($t = 11.73$, CI = [16.07, 22.63], $p < 0.001$) and arousal ($t = -10.11$, CI = [−21.69, −14.55], $p < 0.001$) ratings between the two volume levels was White Noise. These results suggest that affective ratings of sound stimuli are generally affected by volume intensity, but sounds may intrinsically differ in this influence.

Slight variation in high frequencies not associated with degree of unpleasantness

High-frequency tones are known to be unpleasant, and we observed that the 5000 Hz sine tone was rated significantly more unpleasant than the 800 Hz in both volumes (100%: $t = -9.20$, CI = [−22.83, −15.36], $p < 0.001$; 50%: $t = -9.27$, CI = [−19.29, −12.48], $p < 0.001$). As part of the sound array, participants were also presented with two high-frequency sine tones that were unique to their (self-reported) maximum audible frequency: Frequency 1, which was lower in pitch than their threshold by one-fourth, and Frequency 2, which was lower by half. Valence ($t = 0.53$, 95% CI = [−2.03, 3.50], $p = 0.60$) and arousal ($t = -1.96$, 95% CI = [−5.75, 0.04], $p = 0.05$) ratings did not differ significantly between these two sounds at 100% volume (also not at 50% volume: ts > −1.45, ps > 0.15) (Fig. 4). We considered that because participants heard objectively different frequencies in these sound bites, affective ratings for Frequency 1 and Frequency 2 might have been influenced by the absolute frequency that was chosen by each participant. However, we found no significant association between the ratings and the objective stimulus frequency (valence: $\beta = -0.44$, $SE = 0.70$, $p = 0.53$; arousal: $\beta = -1.07$, $SE = 0.70$, $p = 0.13$). Our findings suggest that while high frequencies are generally considered aversive, variation from about 3000 Hz to 10,000 Hz (the sample pool's minimum of Frequency 2 and maximum of Frequency 1, respectively) did not significantly impact valence ratings.

Psychiatric symptom scores are not associated with affective ratings or rating reliability

Lastly, we measured psychiatric symptom severity of state anxiety (STAI-Y1), trait anxiety (STAI-Y2) and obsessive-compulsive symptoms (OCI-R) to test whether psychiatric severity might affect the perception or rating reliability of these stimuli, which could in turn confound task behavior using these stimuli in samples with these psychopathologies.

For affective ratings, neither valence (STAI-Y1: $\beta = -0.96$, $SE = 1.89$, $p = 0.61$; STAI-Y2: $\beta = -0.21$, $SE = 0.03$, $p < 0.001$; OCI-R: $\beta = -0.96$, $SE = 1.89$, $p = 0.61$) nor arousal (STAI-Y1: $\beta = 2.49$, $SE = 1.53$, $p = 0.11$; STAI-Y2: $\beta = -0.96$, $SE = 1.89$, $p = 0.61$; OCI-R: $\beta = 2.49$, $SE = 1.53$, $p = 0.11$) ratings had a significant relationship with the questionnaire scores (Fig. 5a). There was also no interaction effect of psychiatric traits with volume in either dimension (arousal: STAI-Y1: $\beta = -2.03$, $SE = 1.83$, $p = 0.27$, STAI-Y2: $\beta = -2.03$, $SE = 1.83$, $p = 0.27$, OCI-R: $\beta = -2.03$, $SE = 1.83$, $p = 0.27$; valence: STAI-Y1: $\beta = 1.03$, $SE = 2.34$, $p = 0.66$, STAI-Y2: $\beta = 1.03$, $SE = 2.34$, $p = 0.66$, OCI-R: $\beta = 1.03$, $SE = 2.34$, $p = 0.66$) on the ratings (Fig. 5a).

For rating reliability, there was no significant relationship between the intra-rater item test-retest reliability measure and any questionnaire scores for arousal (STAI-Y1: $\beta = -0.006$, $SE = 0.02$, $p = 0.75$, STAI-Y2: $\beta = -0.004$, $SE = 0.02$, $p = 0.85$, OCI-R: $\beta = 0.006$, $SE = 0.02$, $p = 0.76$) or valence (STAI-Y1: $\beta = -0.008$, $SE = 0.01$, $p = 0.47$, STAI-Y2: $\beta = 0.006$, $SE = 0.01$, $p = 0.62$, OCI-R: $\beta = 0.006$, $SE = 0.01$, $p = 0.62$) ratings (Fig. 5b). Similarly, volume did not interact with any of these relationships (arousal: STAI-Y1: $\beta = -0.02$, $SE = 0.02$, $p = 0.33$, STAI-Y2: $\beta = -0.02$, $SE = 0.02$, $p = 0.20$, OCI-R: $\beta = -0.02$, $SE = 0.02$, $p = 0.15$; valence: STAI-Y1: $\beta = -0.01$, $SE = 0.03$, $p = 0.71$, STAI-Y2: $\beta = -0.01$, $SE = 0.03$, $p = 0.63$, OCI-R: $\beta = -0.03$, $SE = 0.03$, $p = 0.26$) (Fig. 5b). In sum, we did not observe an influence of psychiatric symptom severity on affective ratings or the reliability of those ratings.

Discussion

In this study, we examined the feasibility of inducing affective states in web-based online studies using sound stimuli. We find that with the right technical measures in place, we can reliably induce affective states using sound stimuli similar to in-lab studies, which demonstrates that affective audio stimuli can be used well in web-based tasks. Concretely, we found that the ratings were reliable, had good internal consistency and were comparable to those reported in prior studies which collected ratings in controlled in-lab environments. Moreover, we compared several unpleasant sounds consistently and found that amongst them, a modified female scream led to the most aversive state. Our study thus lays the groundwork to use affective sound stimuli for inducing negative affective states in online studies, a much needed means for cognitive online studies.

First, we took on the challenge of ensuring good auditory delivery on a web-based platform. In our procedure, participants were first screened to ensure they wore headphones (Woods et al., 2017) to reduce distractions in the listening environment. Thereafter, they were to set their computer system volume at a fixed level, to enable a more consistent sound

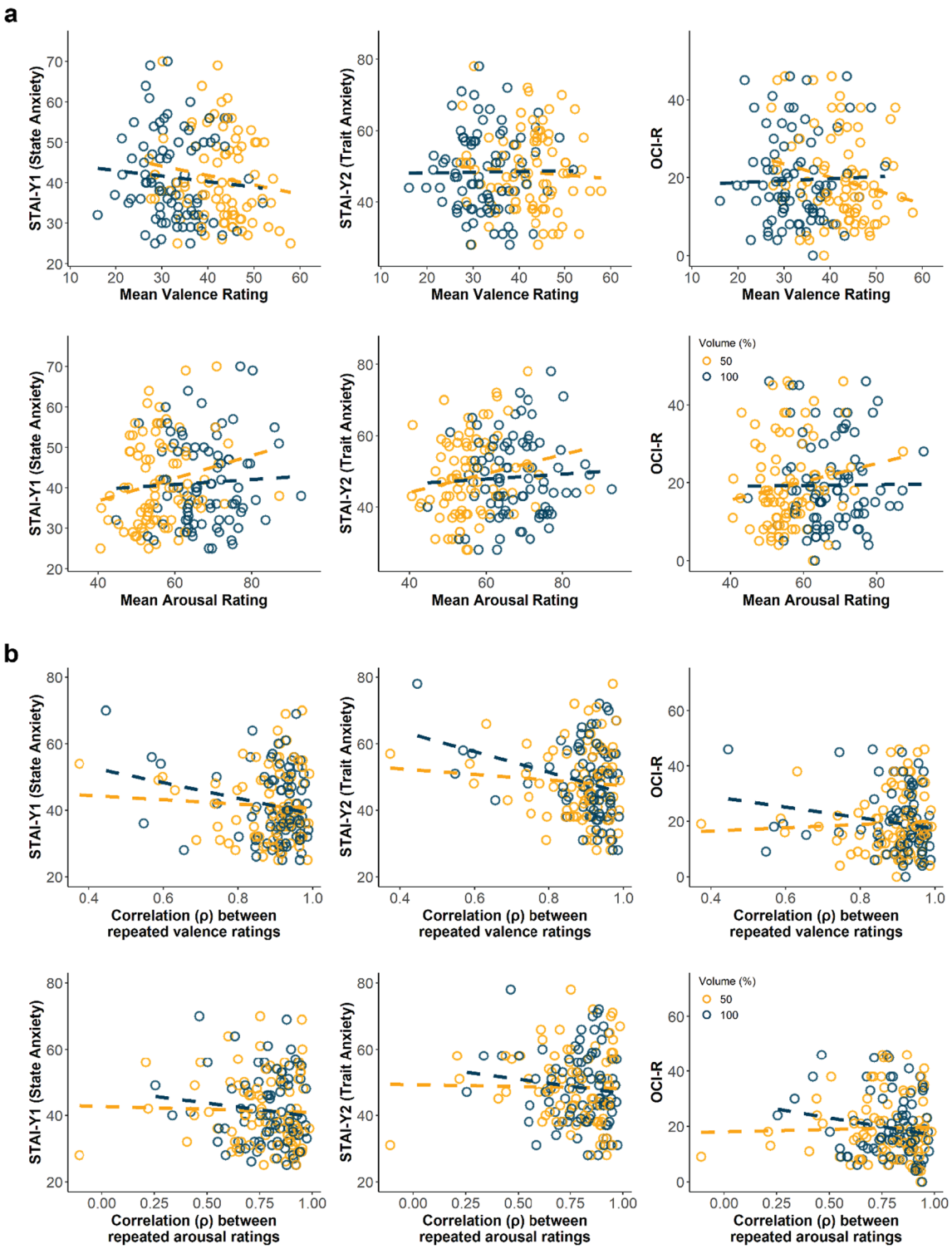

Fig. 5 Correlation of questionnaire scores with affective ratings or participant reliability. For illustration purposes, the scatter plots depict the relationship between **a** affective ratings or **b** ratings reliability with questionnaire total scores. None of the correlations were significant, indicating that neither affective rating nor its reliability was affected by psychiatric symptom severity. Circles represent either the mean affective rating for all sounds per participant or intra-rater item test-retest reliability measure, while the dashed lines represent the linear relationship between questionnaire score and mean rating/reliability estimate at its specified volume

intensity level across participants, before the final adjustment of the browser volume setting. While it is not possible to objectively track whether participants followed our instructions, due to browser-imposed security restrictions, the high pass rate (80.95%) of the headphone check task with one attempt signaled that participants were generally

compliant. Moreover, while there was wide variation in the final browser volume level chosen, the majority of the participants selected levels close to the default loud intensity. Lastly, none of the sounds was indicated as inaudible in the rating task, which might have occurred if volumes that were too low were chosen. Overall, we were confident that our procedure enabled a more consistent and reliable presentation of sound quality online.

We asked participants to rate how the sounds made them feel on two emotional scales, arousal and valence. Participants tended to rate each unique sound quite similarly in their valence ratings, showing decent agreement. Though their arousal ratings were more divergent, this is in line with other studies reporting that arousal ratings (of words) presented more variability than valence with other affective norms, or participants differing by age or language (Fairfield et al., 2017). To examine the test-retest reliability of sound ratings, we specifically designed the task such that participants gave repeated ratings of each sound. We found that participants reported very similar ratings across the two presentations of a particular sound, and they also gave comparable rating patterns in terms of their overall rating means and variances over the repeated presentations. Prior auditory affective rating studies tested in stricter, in-lab environments also examined reliability in terms of internal consistency—likewise, the internal consistency estimates we measured from our ratings were high and similar (e.g. $\alpha = 0.95$ for valence, $\alpha = 0.92$ for arousal in Yang et al., 2018, versus here: $\alpha = 0.93$ for valence, $\alpha = 0.94$ for arousal in our study). These results suggest that the affective ratings collected online were highly reliable. More fine-grained analyses show that the sounds intrinsically differed in their ratings reliability, but all of them (at 100% volume) evoked decent reliability of a correlation of $\rho > 0.60$ (Supplementary Fig. S2). More importantly, as several of our sounds were taken from an open-source sound database (Yang et al., 2018), we were able to compare our ratings that were presented online to ratings that were previously collected in a traditional in-lab setting that were afforded much stricter experimental control. Notably, our sample size was larger than the previous study ($N = 84$ participants rated each sound versus $N = 22$ in Yang et al.), affording greater statistical power, and we replicated the ratings of these sounds very closely. Overall, our results support the validity of a web-based presentation for sounds to evoke their expected affective responses.

A second aim of this study was to suss out the most unpleasant sound for use as an effective aversive stimulus. Unpleasant sounds come from a wide range of categories (Bradley & Lang, 2007; Yang et al., 2018), from naturalistic sounds such as female screams to more synthetic sounds like high-frequency tones or white noise. Varying sounds have been used in prior research as aversive stimuli to drive learning (Bacigalupo & Luck, 2018; Bauer et al., 2020; Morriss et al., 2015, 2016, 2019, 2020; Zald & Pardo, 2002), but it is unknown how they compare to each other given the different methods of selection. While some sounds have their affective ratings documented in sound databases, others are often chosen after unpublished pilot studies or selected *a priori*. Here, we specifically rated some of these unpleasant sounds together to compare their valence ratings. We found that white noise and metal screeches were mildly aversive, followed by high-frequency tones (> 4000 Hz), while more naturalistic sounds such as the female scream were rated as the most aversive. The Morris et al. scream was rated as the most unpleasant and arousing sound, indicating that it may be the most affective aversive stimulus amongst those that we tested.

In this same vein, we manipulated loudness intensity systemically (full and half volume), as the volume is often key in eliciting negative affect (Liberman et al., 2006; Neumann et al., 2008). One previous rating study reported no changes in affective ratings owing to sound intensity differences (Bradley & Lang, 2007), while another noted that volume negatively correlated with valence and positively correlated with arousal (Yang et al., 2018), though this was due to natural volumes in the environment rather than an intended task design. We replicated this same effect with our systematic volume manipulation here. Notably, the sounds intrinsically differed in the strength of influence of volume on their affective ratings—for instance, the ratings for White Noise varied the most depending on volume intensity amongst our array sounds. Nonetheless, volume did not seem to influence valence ratings more so than the type of sound itself (Fig. 3), suggesting that these sounds are valanced because of their inherent characteristics, and less because of the intensity at which they were perceived.

High frequencies are known to be unpleasant (Mirz et al., 2000; Zald & Pardo, 2002), but upper frequency thresholds are subjective to the individual and are heavily dependent on age (Fozard, 1990; Lee et al., 2012; Moore et al., 2014). We attempted to circumvent this issue by limiting the age in our participant pool to a maximum of 40 years. The final chosen upper frequency threshold levels of our sample did show variation, but this was not related to the participants' age (Supplementary Fig. S1). For rating, we presented high-frequency tones that were subjective to each participant by utilizing two frequencies that were slightly lower than their unique threshold frequency. In general, we found that the fixed high-frequency tone (5000 Hz) was more unpleasant than the low-frequency tone (800 Hz), but variation in the subjective high-frequency tones (from around 3000 Hz to 10,000 Hz) did not influence the degree of valence in ratings. Given the lack of relationship between high frequencies and unpleasantness, in addition to some participants having rather low upper frequency threshold levels (Supplementary Fig. S1), high frequencies may not be the most decisive aversive stimuli as compared to others like naturalistic screams, which were ranked as more unpleasant and are straightforward to use.

Lastly, we collected several psychiatric symptom questionnaires scores to ensure that affective ratings themselves were not linked to symptom severity, especially given that online populations are known to have increased mental illness severity (Chandler & Shapiro, 2016; Shapiro et al., 2013). We measured anxiety and obsessive-compulsive symptoms which are core components of two mental health disorders that often utilize aversive learning paradigms in research (Duits et al., 2015; Hauser et al., 2016). A prior in-lab rating study only collected state anxiety symptom data (Yang et al., 2018), which was akin to the severities of our current online sample. Overall, we found no significant relationship between the affective ratings and the psychiatric scores of the participant pool. Importantly, we also observed that the reliability of the ratings was not affected by the psychiatric symptoms, further supporting the use of these sounds in paradigms to study these illnesses. We acknowledge that because our results were drawn from a general population sample, their applicability to diagnosed patients may differ, but there is growing evidence that mechanisms captured with psychopathological variation in online general population findings are clinically relevant (Gillan et al., 2016, 2019).

There were some limitations to our study. Firstly, we examined the test-retest reliability of the sound ratings over two presentations that were relatively close in terms of time; additional trial presentations and/or a longer duration between data collection may result in greater variability between ratings. Secondly, though we attained a certain degree of experimental control with our procedure, the ability to monitor sound presentation quality delivered to participants is still limited. Commercially available headphones may vary in their frequency response, and we were not able to check whether participants had adjusted their system sound settings or had taken their headphones off after the screening test. Moreover, owing to security features of web browsers, it is impossible to track any computer system sound setting information—thus we could not record the objective perceived sound intensities experienced by our participants. Thirdly, the headphone screening test we utilized (Woods et al., 2017) relies on the ability of headphones to present separate signals to the two ears. As such, loudspeaker systems that broadcast a single channel/output of the combined audio could contribute to false positives. For future studies, recent developments of alternative headphone tests may help to resolve this issue (Milne et al., 2020). Lastly, though we asked participants to provide ratings based on the emotional response that they felt, they may have instead recognized and reported the emotional characteristic of the sound. Examining how physiological changes that are linked to emotional states (e.g., skin conductance) align with the affective ratings may bolster our interpretation. Despite the limitations, however, our findings show that affective responses are robust and can be evoked online to a level comparable to that collected in more controlled, in-lab environments.

Given that web-based crowdsourcing will become increasingly common for research, it is only reasonable to enable more paradigms to be translated to online platforms. While the degree of sound presentation control available may not be ideal, many web-induced limitations in other domains such as imprecision in measuring reaction times (Bridges et al., 2020; Plant, 2016) have been found not to compromise data substantially (Crump et al., 2013; Klein et al., 2014). Our findings present evidence that sounds presented through a web-based platform can evoke reliable affective ratings, which supports the translation of affective audio-based paradigms to be conducted online.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01643-0.

Acknowledgements The authors thank Sijia Zhao for the initial discussion of the study. We also thank David Zald, Felix Bacigalupo, Jayne Morris and Wanlu Yang for sharing sounds utilized in the audio array.

Funding This research was funded by a Wellcome Sir Henry Dale Fellowship (211155/Z/18/Z), a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, and Brain & Behaviour Research Foundation NARSAD Young Investigator grant (27023) to TUH. 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 the Wellcome Trust (203147/Z/16/Z).

Declarations

Disclosures The authors report no competing conflicts of interest.

References

Bacigalupo, F., & Luck, S. J. (2018). Event-related potential components as measures of aversive conditioning in humans. *Psychophysiology*, *55*(4), e13015.

Bauer, E. A., MacNamara, A., Sandre, A., Lonsdorf, T. B., Weinberg, A., Morriss, J., & Van Reekum, C. M. (2020). Intolerance of uncertainty

and threat generalization: A replication and extension. *Psychophysiology*, *57*(5), e13546.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., & Flor, H. (2005). Deficient fear conditioning in psychopathy: a functional magnetic resonance imaging study. *Archives of General Psychiatry*, *62*(7), 799–805.

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*(1), 49–59.

Bradley, M. M., & Lang, P. J. (2007). The International Affective Digitized Sounds (; IADS-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*.

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414.

Büchel, C., & Dolan, R. J. (2000). Classical fear conditioning in functional neuroimaging. *Current Opinion in Neurobiology*, *10*(2), 219–223.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk worker demographics. *Sage Open*, *7*(2), 1–15.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*, 53-81.

Crump, M. J. C., McDonnell, J. V, & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., Van Den Hout, M. A., & Baas, J. M. P. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety*, *32*(4), 239–253.

Fairfield, B., Ambrosini, E., Mammarella, N., & Montefinese, M. (2017). Affective norms for Italian words in older adults: age differences in ratings of valence, arousal and dominance. *PloS One*, *12*(1), e0169472.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496.

Fozard, J. L. (1990). Vision and hearing in aging. *Handbook of the Psychology of Aging*, *3*, 143–156.

Gillan, C. M., & Daw, N. D. (2016). Taking psychiatry research online. *Neuron*, *91*(1), 19–23.

Gillan, C. M., Kalanthroff, E., Evans, M., Weingarden, H. M., Jacoby, R. J., Gershkovich, M., Snorrason, I., Campeas, R., Cervoni, C., Crimarco, N. C., Sokol, Y., Garnaat, S. L., McLaughlin, N. C. R., Phelps, E. A., Pinto, A., Boisseau, C. L., Wilhelm, S., Daw, N. D., & Simpson, H. B. (2019). Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry*, *77*(1), 77–85.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *ELife*, *5*, e11305.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Hauser, T. U., Eldar, E., & Dolan, R. J. (2016). Neural mechanisms of harm-Avoidance learning a model for obsessive-compulsive disorder? *JAMA Psychiatry*, *73*(11), 1196-1197.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Klein, R. A., Ratliff, K., Vianello, M., Adams Jr, R. B., Bahník, S., & Bernstein, M. J. (2014). Investigating variation in replicability: a "many labs" replication project. *Social Psychology*, *45*(3), 143-152.

Lau, J. Y. F., Lissek, S., Nelson, E. E., Lee, Y., Roberson-Nay, R., Poeth, K., Jenness, J., Ernst, M., Grillon, C., & Pine, D. S. (2008). Fear conditioning in adolescents with anxiety disorders: results from a novel experimental paradigm. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(1), 94–102.

Lee, J., Dhar, S., Abel, R., Banakis, R., Grolley, E., Lee, J., Zecker, S., & Siegel, J. (2012). Behavioral hearing thresholds between 0.125 and 20 kHz using depth-compensated ear simulator calibration. *Ear and Hearing*, *33*(3), 315.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, *6*(1), 2158244016636433.

Liberman, L. C., Lipp, O. V, Spence, S. H., & March, S. (2006). Evidence for retarded extinction of aversive learning in anxious children. *Behaviour Research and Therapy*, *44*(10), 1491–1502.

Majima, Y., Nishiyama, K., Nishihara, A., & Hata, R. (2017). Conducting online behavioral research using crowdsourcing services in Japan. *Frontiers in Psychology*, *8*, 378.

Malaka, R. (1999). Models of classical conditioning. *Bulletin of Mathematical Biology*, *61*(1), 33–83.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Billig, A. J., & Chait, M. (2020). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*.

Mirz, F., Gjedde, A., Sdkilde-Jrgensen, H., & Pedersen, C. B. (2000). Functional brain imaging of tinnitus-like perception induced by aversive auditory stimuli. *Neuroreport*, *11*(3), 633–637.

Moore, D. R., Edmondson-Jones, M., Dawes, P., Fortnum, H., McCormack, A., Pierzycki, R. H., & Munro, K. J. (2014). Relation between speech-in-noise threshold, hearing loss and cognition from 40–69 years of age. *PloS One*, *9*(9), e107720.

Morris, J. S., Buchel, C., & Dolan, R. J. (2001). Parallel neural responses in amygdala subregions and sensory cortex during implicit fear conditioning. *Neuroimage*, *13*(6), 1044–1052.

Morriss, J., Bennett, K. P., & Larson, C. L. (2020). I told you it was safe: Associations between intolerance of uncertainty and different parameters of uncertainty during instructed threat of shock. *Journal of Behavior Therapy and Experimental Psychiatry*, *70*, 101620.

Morriss, J., Christakou, A., & Van Reekum, C. M. (2015). Intolerance of uncertainty predicts fear extinction in amygdala-ventromedial prefrontal cortical circuitry. *Biology of Mood & Anxiety Disorders*, *5*(1), 1–13.

Morriss, J., Christakou, A., & Van Reekum, C. M. (2016). Nothing is safe: Intolerance of uncertainty is associated with compromised fear extinction learning. *Biological Psychology*, *121*, 187–193.

Morriss, J., Saldarini, F., & Van Reekum, C. M. (2019). The role of threat level and intolerance of uncertainty in extinction. *International Journal of Psychophysiology*, *142*, 1–9.

Moutoussis, M., Bentall, R. P., Williams, J., & Dayan, P. (2008). A temporal difference account of avoidance learning. *Network: Computation in Neural Systems*, *19*(2), 137–160.

Neumann, D. L., & Waters, A. M. (2006). The use of an unpleasant sound as an unconditional stimulus in a human aversive Pavlovian conditioning procedure. *Biological Psychology*, *73*(2), 175–185.

Neumann, D. L., Waters, A. M., & Westbury, H. R. (2008). The use of an unpleasant sound as the unconditional stimulus in aversive Pavlovian conditioning experiments that involve children and adolescent participants. *Behavior Research Methods*, *40*(2), 622–625.

Oyarzún, J. P., Lopez-Barroso, D., Fuentemilla, L., Cucurell, D., Pedraza, C., Rodriguez-Fornells, A., & de Diego-Balaguer, R. (2012). Updating fearful memories with extinction training during reconsolidation: a human study using auditory aversive stimuli. *PloS One*, *7*(6), e38849.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*(1), 408–411.

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, *28*(24), 4014–4021.

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, *84*(6), 443–451.

Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, *117*(49), 31527–31534.

Seow, T. X. F., & Gillan, C. M. (2020). Transdiagnostic Phenotyping Reveals a Host of Metacognitive Deficits Implicated in Compulsivity. *Scientific Reports*, *10*(1), 2883.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science*, *1*(2), 213–220.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory*. Consulting Psychologists Press.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748.

Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human Pavlovian fear conditioning conforms to probabilistic learning. *PLoS Computational Biology*, *14*(8), e1006243.

Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, *395*(10223), 470–473.

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, *11*(1), 1–13.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Yang, W., Makita, K., Nakao, T., Kanayama, N., Machizawa, M. G., Sasaoka, T., Sugata, A., Kobayashi, R., Hiramoto, R., & Yamawaki, S. (2018). Affective auditory stimulus database: An expanded version of the International Affective Digitized Sounds (IADS-E). *Behavior Research Methods*, *50*(4), 1415–1429.

Zald, D. H., & Pardo, J. V. (2002). The neural correlates of aversive auditory stimulation. *Neuroimage*, *16*(3), 746–753.