***Table 1:*** Definitions of key terms (A) and data specifications applied across analyses (B).

**A)**

|  |  |
| --- | --- |
| **Term** | **Definition** |
| **Cross-sectional reliability** | In our study, cross-sectional reliability refers to the reliability of **conditioned responding within experimental phases** at both time points respectively. It provides information on the extent to which items – or in our case – trials measure the same construct (e.g., fear acquisition). We employed the split-half method by splitting odd and even trials and correlating them. |
| **Longitudinal reliability at the individual level** | Longitudinal reliability at the individual level indicates to which extent **responses within the same individuals are stable over time**. It takes into account the individual responses of participants, which are then related across time points. Longitudinal reliability at the individual level inherently includes the group level, as it is calculated for the sample as whole, but the individual responses are central to the calculation. |
| * ***Intraclass correlation coefficients (ICCs)*** | “ICC coefficients quantify the extent to which multiple measurements for each individual (within individuals) are statistically similar enough to discriminate between individuals” [cf. @aldridge2017]. Here, we calculated two types of ICCs, namely **absolute agreement** and **consistency**. To illustrate the difference between absolute agreement and consistency in a short example (@koo2016), consider an interrater reliability study with two raters: Consistency indicates the extent to which the score of one rater (y) is equal to the score of another rater (x) plus a systematic error (c) (i.e., y = x + c). In contrast, absolute agreement indicates to which degree y equals x. As “two raters” can be replaced by “two time points” and individual responses were taken into account here, we used ICCs to determine longitudinal reliability at the individual level. |
| * ***Within- and between subject similarity*** | Similarity analyses provide information on the extent to which trial-by-trial responses of one individual at one time point are comparable to responses of   * the same individual at a later time point (i.e., within-subject similarity) and * all other individuals at a later time point (i.e., between-subject similarity).   Comparisons of within- and between subject similarity were used here to determine longitudinal reliability at the individual level. |
| * ***Overlap at the individual level (applied for BOLD fMRI only)*** | Overlap at the individual level reflects the **degree of overlap of significant voxels** between both time points **for single subject level activation pattern**s. |
| **Longitudinal reliability at the group level** | Longitudinal reliability at the group level indicates to which degree **responses within the group as a whole are stable over time**. More precisely, longitudinal reliability at the group level relies on first averaging all individuals responses for each trial (for SCR) or voxel (for fMRI) yielding a group average for each trial/voxel. These are then related across time points, i.e. the calculation is carried out using the trial-wise (for SCR) or voxel-wise (for fMRI) group averages. |
| * ***Overlap at the group level (applied for BOLD fMRI only)*** | Overlap at the group level reflects the **degree of overlap of significant voxels** between both time points **for aggregated group level activations**. |

**B)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Measure** | **Cross-sectional reliability** | **Longitudinal reliability at the individual level** | | | **Longitudinal reliability**  **at the group level** | **Cross-phases predictability** |
|  |  |  | ***ICCs*** | ***Within- and between subject similarity*** | ***Overlap*** | ***Overlap (BOLD fMRI) or R squared (SCR)*** |  |
| **Included time points** | all | T0 and T1 separately | T0 and T1 | T0 and T1 | T0 and T1 | T0 and T1 | T0 |
| **Included stimuli** | SCR | CS+, CS-, CS discrimination, US | CS+, CS-, CS discrimination, US1 | CS+, CS-, CS discrimination, US | -- | CS+, CS-, CS discrimination, US | CS+, CS-, CS discrimination |
| Fear ratings | -- | CS+, CS-, CS discrimination, US1 | -- | -- | -- | CS+, CS-, CS discrimination |
| BOLD fMRI | -- | CS discrimination2 | CS discrimination2 | CS discrimination2 | CS discrimination2 | CS+, CS-, CS discrimination |
| **Phase operationalizations** | SCR | Entire phases (ACQ, EXT, RI-Test; except first trials of ACQ and EXT) | CS+, CS- and CS discrimination:  average ACQ  last 2 trials ACQ  1st trial EXT3  average EXT  last 2 trials EXT4  1st trial RI-Test3  US:  average RI | average ACQ5  average EXT | -- | average ACQ5  average EXT | average ACQ  last 2 trials ACQ  1st trial EXT3  average EXT  last 2 trials EXT4  1st trial RI-Test3 |
| Fear ratings | -- | CS+, CS- and CS discrimination:  post-pre ACQ  post ACQ  pre EXT  pre-post EXT  post EXT  1 trial RI-Test  US:  post RI-Test | -- | -- | -- | post-pre ACQ  post ACQ  pre EXT  pre-post EXT  post EXT  1st trial RI-Test |
| BOLD fMRI6 | -- | average ACQ  average EXT | average ACQ  average EXT | average ACQ  average EXT | average ACQ  average EXT | average ACQ  average EXT |
| **Transformations** | SCR | none  log-transformation7  log-transformation and range correction8 | none  log-transformation7  log-transformation and range correction8 | none9 | -- | none  log-transformation7  log-transformation and range correction8 | none  log-transformation7  log-transformation and range correction8 |
| Fear ratings | -- | none | -- | -- | -- | none |
| BOLD fMRI | -- | none | none | none | none | none |
| **Ordinal ranking** | SCR | no ranking | no ranking10 | no ranking | -- | no ranking | no ranking and ordinal ranking11 |
| Fear ratings | -- | no ranking10 | -- | -- | -- | no ranking and ordinal ranking |
| BOLD fMRI | -- | no ranking | no ranking | no ranking | no ranking | no ranking |

*Note.* The specifications we used here are exemplary and are not intended to cover all possible data specifications. Note that internal consistency, within- and between subject similarity and reliability at a group level could not be calculated for fear ratings due to the limited number of trials. ACQ = acquisition training, EXT = extinction training, RI = reinstatement, RI-Test = reinstatement-test.

1Non-pre-registered ICCs for SCRs to the USs and US aversiveness ratings were calculated as we considered these informative.

2For BOLD fMRI, ICCs were calculated only for CS discrimination and not for CS+ and CS- given the fact that the calculations are based on first level T contrast maps and contrasts against baseline are not optimal.

3Fear recall and reinstatement-test were operationalized as the first extinction training trial and the first reinstatement test-trial (since the reinstatement effect fades away relatively quickly [@haaker2014]) respectively.

4The operationalization of extinction training as the last two trials was not pre-registered and included for completeness. In cases where phase operationalizations included more than one SCR trial, trials were averaged.

5Note that reliability at a group level for SCRs during reinstatement-test was not calculated as correlations between two SCR data points are not meaningful.

6fMRI data for the reinstatement-test was not analyzed in the current study since data from a single trial do not provide sufficient power.

7Raw SCR amplitudes were log-transformed by taking the natural logarithm.

8Log-transformed SCR amplitudes were range corrected by dividing each individual SCR trial by the maximum SCR trial across all CS and US trials. Due to potentially different response ranges, the maximum SCR trial was determined separately for experimental days as recommended by @lonsdorf2017. Range correction was recommended to control for inter-individual variability [@lykken1971a; @lykken1972].

9We also carried out similarity analyses for log-transformed as well as for log-transformed and range corrected data. However, results were almost identical to the results from the raw data. For reasons of space, we therefore only report results for raw data.

10As opposed to what was pre-registered, we included in ICC analyses only non-ranked data as closer inspection of the conceptualization of ICCcon, we realized that it would be redundant to calculate both ICCabs and ICCcon with ranked and non-ranked data as ICCcon itself ranks the data.

11Ranks of SCRs were built upon raw, log-transformed as well as log-transformed and range corrected values.