Deleting “fear” from “fear extinction”: Estimating the individual extinction rate via non-aversive conditioning

Michelle Spix

Maastricht University, Maastricht

Miriam J. J. Lommen

University of Groningen, Groningen

Yannick Boddez*

Ghent University, Ghent

KU Leuven, Leuven

Michelle Spix, Department of Clinical Psychological Science, Maastricht University; Miriam J. J. Lommen, Department of Clinical Psychology and Experimental Psychopathology, University of Groningen; Yannick Boddez, Department of Experimental Clinical and Health Psychology, Ghent University and Center for the Learning of Psychology and Experimental Psychopathology, KU Leuven.

Yannick Boddez is supported by Ghent University grant BOF16/MET_V/002 awarded to Jan De Houwer.

*Correspondence concerning this article should be addressed to Yannick Boddez, Ghent University, Henri Dunantlaan 2 B-9000 Ghent Belgium. E-mail: yannick.boddez@ugent.be

Abstract

Individual differences in extinction learning have attracted ample attention of researchers and are under investigation as a marker for the onset of anxiety disorders and treatment response. Unfortunately, the common paradigm for obtaining the extinction rate, which entails aversive stimulus pairings, is subject to practical limitations. Therefore, the present study assessed whether the use of an aversive stimulus is actually needed to get a good estimate of the extinction rate. A total of 161 undergraduate students completed a conditioning task with both an aversive and a non-aversive stimulus. Using latent class growth analysis (LCGA), distinct trajectories, representing normal and stunted extinction learning, were identified for both these stimulus types. Participants' membership in these classes largely overlapped for aversive and non-aversive stimulus pairings and respective extinction indices were significantly correlated. Thereby, findings suggest that the use of a non-aversive stimulus could suffice for successfully capturing individual differences in extinction learning. However, future studies are needed to confirm that conditioning with a non-aversive stimulus may serve to predict clinically relevant outcomes.

Keywords: conditioning, extinction learning, non-aversive US, LCGA, inter-individual differences

Deleting "fear" from "fear extinction": Estimating the individual extinction rate via non-aversive conditioning

About 5 to 10% of the world population (Baxter, Scott, Vos & Whiteford, 2012) suffer from anxiety disorders, harming individuals and their loved ones, as well as the economy that they live in (Johnson & Casey, 2015). Cognitive-behavioral therapy is considered the gold standard treatment (Vervliet, Craske, & Hermans, 2013) and its efficacy has been supported by several meta-analyses (e.g., Olatunji, Cisler, & Deacon, 2010). However, only half of the treated patients show a clinically significant symptom reduction (Johnson & Casey, 2015**)** and changes are often not maintained in the long run (Craske & Mystkowski, 2006).

Research efforts are, therefore, invested in predicting the onset of anxiety disorders, leaving room for prevention, and in predicting treatment response (Johnson & Casey, 2015). Fear extinction has been claimed to be a promising tool in this respect. In a fear conditioning procedure, a neutral stimulus (conditional stimulus or CS) is paired with an aversive stimulus such as electric shock (unconditional stimulus or US). In the subsequent fear extinction phase, the CS is presented by itself, typically resulting in a decrease of the fear responses that were previously established by pairing the CS with the US. In line with this, deficits in extinction learning were found to predict post-traumatic stress disorder symptoms in a sample of fire fighters (Guthrie & Bryant, 2012) and Dutch Soldiers (Lommen, Engelhard, Sijbrandij, van den Hout & Hermans, 2013), as well as less favourable outcomes after exposure therapy in subclinical (Forcadell et al., 2017b) and clinical samples (Duits, 2016; Duits et al., under review[1]).

However, the use of aversive USs can bring about practical challenges. For example, being exposed to aversive stimuli might put an additional burden on vulnerable populations, such

[1] Please note that the study under review is also discussed in Duits (2016). This latter publication is accessible online (see reference list for a link).

as people suffering from anxiety. In addition, the presentation of aversive stimuli such as electrical shocks requires specific equipment and supervision by an experimenter, which renders the process of data gathering time and resource intensive. Surprisingly, the necessity of an *aversive* US for the manifestation of inter-individual differences in extinction performance, as well as a correct estimation of the extinction rate remains unclear (Vroling & de Jong, 2013). In other words, the possibility exists that a procedure with a non-aversive US (e.g. neutral image) might suffice to estimate an individual's extinction learning capacity, while easily circumnavigating the previously mentioned obstacles. Therefore, we examined whether extinction with a non-aversive US provides the same information as extinction with an aversive US.

The current focus on aversive conditioning could be justified if one assumes that a given individual will respond differently to aversive and to non-aversive stimulus pairings. In this respect, some authors have argued that different processes are recruited if aversive stimuli are used. For example, only learning about aversive events might be mediated by a reflexive and relatively uncontrolled process (LeDoux, 2014; LeDoux & Daw 2018). However, other theoretical perspectives are possible. For example, propositional learning theorists hypothesize that both learning about aversive and non-aversive events is mediated by inferential reasoning processes and the formation of propositional beliefs concerning how stimuli are related (De Houwer, 2020; Boddez, Moors, Mertens, & De Houwer, 2020; Vroling & de Jong, 2013). One possibility is that people show reduced extinction, because they expect to be tricked and infer that it is likely that the US will suddenly reappear in the course of the extinction phase (Boddez et al., 2020; Vervliet and Boddez, 2020). Such inference may affect learning about aversive and non-aversive events alike.

Given this theoretical debate (also see Mertens, Boddez, Sevenster, Engelhard & de Houwer, 2018), it seems worthwhile to study individual differences in extinction of a non-

aversive US. The existing research on conditioning with a non-aversive US, mainly from the field of human contingency learning (De Houwer & Beckers, 2002; Boddez, De Houwer & Beckers, 2017, Pineño & Miller, 2007) already showed that basic conditioning phenomena like acquisition, cue competition and extinction do occur when using non-aversive USs (e.g. Meulders, Boddez, Vansteenwegen & Baeyens, 2013; Boddez, Baeyens, Hermans & Beckers, 2011). However, the question whether extinction with an aversive US provides different information than extinction with a non-aversive US remains open. Given the clinical and scientific goal of predicting treatment response and the onset of anxiety disorders, it would be a valuable outcome if an individual's extinction rate could be successfully estimated only on the basis of their extinction performance after conditioning with a non-aversive US.

Present study

The current study therefore combines both aversive and non-aversive USs into a single differential conditioning task. Participants were presented with CS1 - aversive US (shock), CS2 - non-aversive US (neutral picture) and CS3 - no US pairings. We collected US-expectancy ratings, in line with previous studies in which fear extinction was found to predict posttraumatic stress (Lommen et al., 2013) and treatment success (Forcadell et al., 2017b; Duits, 2016; Duits et al., under review). These US-expectancy ratings were inspected with latent class growth analysis (LCGA), which allows the investigation of intra- as well as inter-individual differences by identifying heterogeneous trajectories in repeated measurement data (Galatzer-Levy, Bonanno, Bush & LeDoux, 2013). This method has previously proven successful in identifying trajectories in conditioning performance in various research designs, outcome measures and species (Duits, 2016; Duits et al., under review; Galatzer-Levy et al., 2013). We investigated whether it is possible to estimate participants' extinction trajectory membership for the aversive US from their extinction trajectory for the non-aversive US. In addition to the LCGA, we assessed participants' extinction performance by computing their extinction rate for the aversive and non-aversive US

and, then, tested whether the extinction rates for the two outcomes correlate. Thereby, this study could provide a first indication that using a non-aversive US suffices to obtain a useful estimate of an individual's extinction performance.

Method

Participants

A total of 161 first-year psychology students at the University of Groningen enrolled in the study in order to obtain course credits. Exclusion criteria were (1) a current psychiatric diagnosis (2) visual problems (if not corrected), (3) the use of drugs or medication interfering with memory or attention, (4) epilepsy, (5) a heart condition and (6) pregnancy. The presence of a current psychiatric diagnosis was assessed by directly asking participants 'Do you have a current psychiatric diagnosis?'. Answering 'Yes' led to exclusion from the study. Four participants were excluded from the analysis because of technical problems, prior knowledge about the conditioning paradigm, a current mental disorder and drop-out from the study because of anxiety regarding the shock administration. Additionally, data of one participant were left out of the analysis after the participant spontaneously told the experimenters that answers had not been given truthfully. The final sample consisted of 156 participants (95 females) with a mean age of 20.49 (*SD* = 1.89) years. Further information on the sample can be obtained from Table 1.

Apparatus and stimuli.

Unconditional stimuli. An electric shock (0 -5 mA), applied to the second and third digit of the left hand, served as the aversive US. We chose to use electrical stimulation because it has provided fairly consistent results across different conditioning studies (Lonsdorf et al., 2017). For the non-aversive US, a neutral picture of a basket was used. This decision was based on its rating in the IAPS system (Lang, Bradley & Cuthbert, 2008), with a mean valence rating of 4.94 (*SD* =

1.07) and a mean arousal rating of 1.76 (*SD* = 1.48) on a scale from 1 to 10. Both stimuli were administered for 500 ms.

Conditional stimuli. Three geometrical shapes (triangle, square and circle; 7 x 7 cm) were used as CSs. These were presented in a semi-random order, so that the same stimulus was not presented more than two times in a row. The assignment of geometrical shapes to CSs was counterbalanced so that every combination of geometrical shape - US pairings was used.

The task was programmed and run via the computer program E-Prime (version 2.0.10.356; Psychology Software Tools, Inc., Pittsburgh, PA).

Measures

US-expectancy ratings. In order to assess the course of acquisition and extinction learning, US-expectancy ratings were used. For an elaborate discussion regarding the validity of US-expectancy ratings see Boddez et al. (2013). In summary, studies comparing clinical and non-clinical populations provide support that these groups differ in their US-expectancy ratings (i.e., diagnostic validity). In addition, US-expectancy measures capture the expectancies that are crucial in learning theories of anxiety (i.e., construct validity). Finally, there is evidence that US-expectancies can predict symptom levels (i.e., predictive validity). For example, previous studies relating PTSD onset and treatment response to fear extinction learning relied on US-expectancies (e.g. Lommen et al., 2013; Forcadell et al., 2017b; Duits, 2016; Duits et al., under review).

Participants rated their expectancy of the aversive and non-aversive USs on a single visual analogue scale (“To what extent do you expect that either a shock or a picture will follow?”; 0 = "certainly not“; 100 = “certainly“). As a result of this formulation, they only indicated a low score, when they expected neither US and answered with a high score when expecting either of both.

Valence ratings. Participants were asked to indicate on a 10-point Likert scale (-5 = "very unpleasant"; 5 = "very pleasant") how unpleasant they found the conditional and unconditional stimuli. As evaluative conditioning seems to be relatively resistant to extinction learning (O'Malley & Waters, 2018; Baeyens, Diaz & Ruiz, 2005; Baeyens, Eelen, van den Bergh & Cromez, 1989), these ratings offer a possibility to check differential responding to the aversive CS+ and the other CSs after the extinction phase.

Contingency awareness. To assess whether participants learned the correct CS-US relations, after the conditioning task they indicated which figures were previously followed by the electrical stimulation, the neutral image or nothing (e.g., "Which figure(s) were followed by the electrical stimulation?"). Thus, participants had to answer three questions in total. Participants could provide more than one answer option (circle, square, triangle and none) per question. Individuals were regarded as contingency aware if they provided the correct answer for all three questions. This means that they needed to indicate that the aversive CS+ figure was followed by the electrical stimulation, the non-aversive CS+ figure by the neutral image and the CS- figure by nothing.

Procedure

The study was approved by the Ethical Committee of Psychology at the University of Groningen. After arrival in the laboratory, exclusion criteria were checked. Subsequently, the general set-up of the study was explained both orally and in writing. It was emphasized that participants could refrain at any time without any negative consequences. If they agreed to the terms of the study, participants provided informed consent and filled in the OQ45-symptomatic distress subscale (Lambert et al., 1996), the Attention Control Scale (Derryberry & Reed, 2002), the emotional attention control scale (Barry, Hermans, Lenaert, Debeer & Griffith, 2013) and the Spielberger Trait Anxiety Inventory (Spielberger & Gorsuch, 1983). These were included in the study as part

of other (master thesis) projects and will not be discussed in this manuscript. To determine the appropriate level of shock intensity for the conditioning task, a work-up procedure **(**Orr et al., 2000**)** was used until individuals described the shock as "highly annoying, but not painful". Subsequently, participants practiced giving judgements on a visual analogue scale by rating a picture of a banana on its pleasantness ("How pleasant is this picture?"; 0 = "very unpleasant"; 100 = "very pleasant"). Before the start of the conditioning task, they were instructed to answer quickly, as the rating scales and pictures would only be presented for a limited amount of time. Furthermore, participants were told to keep in mind that their ratings referred to both their expectancy of the shock (aversive US) and of the picture (non-aversive US).

The conditioning task consisted of an acquisition and an extinction phase. During the acquisition phase, the two CS+ were presented 8 times and were immediately followed by the aversive or the non-aversive US in 75% of the trials. While the CS- was also shown 8 times, it was never paired with an event. Thus, the acquisition phase consisted of 24 trials in total. During the extinction phase, each CSs was presented 12 times without being followed by any of the USs resulting in a total of 36 trials. Throughout the presentation of the CSs participants were asked to give ratings of their US-expectancy. CSs were displayed for 8 s on a computer screen (27 inches). During the following 5 s intertrial interval (ITI) a white blank screen was presented. A schematic overview of the task is given in Figure 1.

After the conditioning task, participants' valence ratings of the CSs and USs as well as their contingency awareness were assessed. Lastly, participants were debriefed and received their course credit.

Extinction indices

Researchers have used different indices to quantify extinction learning (e.g., Lommen et al., 2013; Forcadell, Torrents-Rodas, Treen, Fullana & Tortella-Feliu, 2017a; ; Lenaert et al., 2014;

Pineles et al., 2016) and extinction retention (Lonsdorf, Merz & Fullana, 2019). We decided before data inspection to rely on three different extinction indices in this study. By analyzing different extinction indices, we tried to prevent that conclusions become biased due to the choice of a particular extinction index. The three indices were calculated separately for the aversive and non-aversive CS+.

First, extinction learning was defined as the overall level of US-expectancy across the extinction phase. For this, we calculated the area under the curve (AuC) with the trapezoid rule. This means that an individual's CS+ expectancy ratings across extinction were connected with an imaginary line to form a learning curve. The space under this learning curve was then divided into trapeziums, which were eventually added together ((expectancy at trial 1 + expectancy at trial 2)/2 * (time point 2 – timepoint 1) … + (expectancy at trial 11 + expectancy at trial 12)/2 * (time point 12 – timepoint 11); for related approaches see Forcadell, Torrents-Rodas, Treen, Fullana & Tortella-Feliu, 2017a; Lenaert et al., 2014). Lower values are assumed to represent greater extinction of the conditional response.

In addition, we operationalized extinction learning as the difference in US-expectancy between the first and the fourth trial of the extinction phase (100 minus (CSs+ trial 1 minus CSs+ trial 4); Lommen et al., 2013), as well as the difference score between extinction trial one and eight (100 minus (CSs+ trial 1 minus CSs+ trial 8); Lommen et al., 2013). We subtracted the difference scores from 100 in order to ensure that lower scores on all three indices represent a greater reduction in US-expectancy and therefore greater extinction learning.

Data reduction and analysis

LCGA (for a detailed description see Jung & Wickrama, 2008) was used to investigate inter-individual differences and intra-individual changes in extinction learning across trials. Missing data was imputed with the full-information maximum likelihood. This analysis was conducted in

Mplus (Version 8), data exploration was done in IBM SPSS Statistics (Version 23) and additional tests were carried out using RStudio (Version 1.1.463).

The LCGA was conducted separately for conditioning with the aversive and non-aversive US by using the US-expectancies of the respective trials in the acquisition and extinction phase. We decided to jointly consider acquisition and extinction, because learning patterns during extinction are difficult to interpret without taking learning performance during acquisition into account. Moreover, we decided to not include the CS- in our analysis as CS- responding might not constitute a neutral baseline, but possibly represents additional processes such as generalization and safety learning (Haddad, Pritchett, Lissek & Lau, 2012). Thus, including the CS- in the analyses could have rendered the interpretation of the results less definite. The current approach has been adopted by previous studies that used LCGA to investigate heterogeneity in extinction learning (Duits et al, under review; Duits, 2016; Galatzer-Levy et al., 2013). As heterogeneity between and within growth profiles was expected across the conditioning procedure, the intercept and slope were not fixed to a certain parameter. Instead, the algorithm was allowed to estimate them freely in order to achieve the best fit for the data. Common issues with LCGA are the identification of local solutions and non-convergence during the log-likelihood estimation (Jung & Wickrama, 2008). To reduce these risks and to improve the reliability of the log-likelihood estimation, the number of random sets on starting values was determined to be 800 and the number of final optimizations was put to 200.

In order to identify the best representation of the data, models with one to six trajectories were compared based on four criteria. Reductions in the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), entropy scores, as well as theoretical considerations were taken into account (similar to Duits, 2016; Duits et al., under review). For the latter, we compared the models to trajectories identified in the existing literature (Duits, 2016; Duits et al., under review, Galatzer-Levy et al., 2013) and checked whether the trajectories for the k class

models were distinct enough to represent clinically and theoretically relevant differences. Entropy scores nearing 1 are considered as a sign of satisfactory delineation between classes (Celeux & Soromenho, 1996). Additionally, we tested whether the step from a model with k classes to a model with k - 1 classes lead to a significant reduction on the Lo, Mendell and Rubin likelihood ratio test (LMR-LRT) statistic. After deciding on a model, individuals were assigned to one of the identified classes based on their respective probability scores.

In addition, correlations between the extinction indices after conditioning with the aversive and non-aversive US were calculated. Here the statistical significance level was set at $\alpha = .05$.

Results

Trajectories of US-expectancy for the aversive US

To answer the first research question whether similar patterns of individual differences in extinction learning can be found for conditioning with an aversive and non-aversive US, LCGAs for both USs were run.

Two distinct trajectories of aversive US expectancy ratings were identified (see Figure 2). The decision for a two-class model was based on statistical as well as theoretical arguments. First, the two-class model presented with a satisfactory entropy score of .944 . Second, the highest reduction in BIC and AIC scores could be observed, when moving from a one-class model to a two-class model (see Table 2 and Figure 3). Third, the two-class solution overlapped with previous findings (e.g. Duits, 2016; Duits et al., under review).

The two classes were marked by distinct courses of learning. The larger trajectory contained 61.53% of all participants ($n = 96$) and depicted the typical course of conditioning and extinction. Participants started with US-expectancy ratings about halfway up the scale, which increased during acquisition and reduced after the onset of the extinction phase. We labelled this pattern ‘ normal extinction’. Figure 2 seems to suggest that members of the second

trajectory (38.46% of all participants; $n = 60$) showed a comparable increase of US-expectancy during acquisition, while their rate and magnitude of extinction learning appeared slower and generally smaller. Therefore, this trajectory was labelled 'poor extinction'.

Trajectories of US-expectancy for the non-aversive US

For the US-expectancy ratings of the non-aversive US, three distinct trajectories were identified using LCGA (see Figure 4). The decision for the three-class model was based on different considerations. The model showed a high reduction in BIC and AIC, as well as the highest entropy score from the six computed models (see Table 2 and Figure 5). When additionally checking the LMR-LRT statistic comparing the three-class model with a two-class solution, a significant reduction in log-likelihood from the former to the latter was found, $\chi^2\,(21) = 491.07$, $p = .004$. Lastly, when inspecting the trajectories of the three-class model; they seem to represent distinguishable courses of learning that are relevant from a theoretical and clinical perspective.[2] The largest class entailed 57.06% ($n = 89$) of the participants and was labelled 'normal' as US-expectancy ratings increased during acquisition and decreased again at the beginning of the extinction phase. The trajectory of the second largest class (24.35%; $n = 38$) showed a stunted increase of US-expectancies during acquisition followed by a steady reduction of expectancy ratings across the extinction phase and was, therefore, labelled 'poor acquisition'. Twenty-nine individuals (18.58%) were assigned to the third class labelled 'poor extinction', which was marked by a steady increase of expectancy ratings during acquisition and a delayed onset of US-expectancy reduction during extinction. Furthermore, Figure 4 indicates that this decrease occurred slower and was smaller than for the 'normal' class.

[2] When visually inspecting the n-class models, the n > 3 class models resulted in highly overlapping trajectories, which did not capture inter-individual differences in learning. The trajectories of the 3-class model, on the other hand, showed little overlap and represented distinguishable courses of learning. Therefore, they might be the most informative for clinicians and researchers.

The overlap between trajectories

After identifying the classes for aversive and non-aversive conditioning, we examined conditional probabilities indicating the chances of aversive US class membership given an individual's non-aversive US trajectory, and vice versa (see Table 3). Additionally, we computed correlations between extinction indices for aversive and non-aversive US conditioning. The results pointed towards a high overlap in extinction performance. First, most participants in the 'normal' and 'poor extinction' class for non-aversive US conditioning belonged to a similar class for the aversive US (i.e. 73.03% and 79.31%; see Figure 5). Second, the aversive and non-aversive extinction indices were significantly correlated, AuC $r(145) = .48$, $p < 0.001$, Extinction 1 – 4 $r(151) = .35$, $p < 0.001$, Extinction 1 – 8 $r(151) = .51$, $p < 0.001$. However, the overlap between categories can obviously not be perfect due to the additional ' poor acquisition' trajectory for non-aversive US conditioning.

In order to assure that the results were not caused by a failure to learn the correct CS – US relations (i.e., participants simply mixing up the aversively and non-aversively conditioned CSs), we repeated the analysis with contingency-aware individuals only ($n = 77$). The findings resembled the complete-sample findings concerning class membership (see Table 4) and correlations between extinction indices with AuC $r(70) = .67$, $p < 0.001$, Extinction 1 – 4 $r(73) = .36$, $p = 0.001$, Extinction 1 – 8 $r(73) = .46$, $p < 0.001$. Additionally, we compared the valence ratings of the three CSs and the two USs. The aversive US was rated as significantly less pleasant than the non-aversive US, $t(155) = 24.28$, $p < .001$. Similarly, the aversive CS+ was rated as significantly more unpleasant than the non-aversive CS+, $t(155) = 13.15$, $p < .001$, and the CS-, $t(155) = 13.19$, $p < .001$. The non-aversive CS+ and the CS- showed no significant difference in valence ratings, $t(155) = -0.99$, $p = .319$. Thus, results were in line with the learning pattern that we expected.

Discussion

In the present study, we investigated whether similar patterns of extinction learning appear for aversive and non-aversive US conditioning and to what extent individuals' extinction learning performance overlapped for these types of conditioning. With this, we intended to gather information on whether the use of neutral outcomes in conditioning procedures might suffice to obtain a good estimate of the extinction rate.

In line with previous research findings (e.g., Duits, 2016; Duits et al., under review), we identified two trajectories, marked by normal and stunted extinction learning performance for the aversive US conditioning. The findings underline the stability of these two classes. LCGA analysis for the non-aversive US conditioning revealed an additional third group that showed a limited learning of the relationship between the CS and the non-aversive US. Reduced acquisition for a non-aversive US can be explained by associative learning models. These models suggest that learning is a function of US salience, so that stronger conditioning effects are expected for more intense USs compared to more neutral ones (De Houwer & Hughes, 2020; Rescorla & Wagner, 1972). At the same time, these models leave unexplained why the reduced acquisition is found in some individuals, but not others.

Importantly, most participants in the 'normal' and 'poor extinction' class for non-aversive US conditioning belonged to a corresponding class for the aversive US. In addition, the aversive and non-aversive extinction indices were significantly correlated, with effect sizes ranging from medium to large. This suggests that researchers and clinicians who aim to predict anxiety disorder onset and treatment response by the use of fear extinction rates (Lommen et al., 2013; Forcadell et al., 2017b) might suffice with a measurement procedure that solely relies on non-aversive USs.

Nevertheless, we want to highlight that this study only constitutes a first step in understanding the value of non-aversive US extinction learning and that our findings need to be

interpreted cautiously. The overlap between the aversive and non-aversive US trajectories was not perfect and the correlation between extinction rates ranged from medium to large (r = .35 to r = .51). Therefore, it remains possible that what is unique to the extinction of an aversive US plays an important role in the prediction of anxiety disorder onset (Lommen et al., 2013; Guthrie & Bryant, 2012) or treatment response (Duits, 2016; Duits et. al., under review; Forcadell et. al., 2017b). Future research, therefore, needs to investigate whether extinction for non-aversive USs can predict relevant outcomes to a similar extent as extinction after aversive US conditioning. So, we hope that our data may serve to invite researchers who aim to predict clinical outcomes to include our non-aversive conditioning task in their studies and put its value to the test.

Some theoretical and methodological aspects of the study deserve further attention. First, some theorists might not agree to use the term "conditioning" for a procedure with a neutral US, perhaps because they want to reserve that term for learning about biologically significant stimuli (Öhman & Mineka, 2001). Other theorists, however, define conditioning as a change in behavior (including a change in US-expectancies) due to stimulus pairings (De Houwer and Hughes, 2020), irrespective of the used stimuli. In line with this latter view, we have used the term conditioning for learning about both USs in our design.

Second, it is important to note that US-expectancies served as our only outcome measure and that no brain data or physiological data were collected. We decided to focus on US-expectancies, because these measures have been previously employed in studies that aimed to predict PTSD onset (Lommen et al., 2013) and treatment response (Forcadell et al., 2017b). Moreover, when aspiring to eventually have a procedure that is broadly and easily applicable for clinical purposes, US-expectancy ratings might be the measure of choice compared with physiological assessments that require specialized equipment. Nonetheless, future research comparing conditioning phenomena for differently vale nced USs might include several outcomes measures (e.g., skin conductance responding) in order to allow more fine-grained

statements on possible similarities and differences.

Third, the within-subjects design, combining aversive and non-aversive US conditioning into one task, allowed us to investigate the overlap in individual's extinction trajectories. However, one could argue that the high overlap between aversive and non-aversive US trajectories represents a failure in discrimination learning. For example, participants could have mistakenly believed that the aversive US followed both CSs (CS – shock and CS – neutral image) resulting in corresponding learning trajectories and extinction indices. Two arguments go against this interpretation though. First, the evaluative conditioning data indicate that participants acquired the correct CS – US relationships as the CS paired with the aversive US was rated as significantly less pleasant compared to the other CSs. Second, our findings regarding the overlap between learning trajectories, as well as the correlation between extinction rates, were similar when only including participants with complete contingency awareness in the analysis. Nonetheless, we cannot exclude the possibility that the found overlap has something to do with the inclusion of shock in the experiment at large (Robinson, Letkiewicz, Overstreet, Ernst & Grillon, 2011) or that learning about the aversive and non-aversive US influenced each other. Future studies could present the aversive and non-aversive conditioning trials in separate sessions. When accounting for possible confounders (Lonsdorf et al., 2017), rates of fear learning proved relatively stable over time (Zeidan et al., 2012). Alternatively, existing datasets of non-aversive US conditioning could be reanalyzed using LCGA in order to see whether trajectories replicate when dropping the aversive US from the task.

In summary, our findings showed considerable (although imperfect) overlap in extinction performance for aversive and non-aversive US conditioning. We hope that these findings may encourage researchers to start to evaluate the predictive validity of non-aversive US conditioning with respect to the onset of anxiety disorders and treatment response.

Declaration of competing interest

The authors declare no conflicts of interest.

References

Baeyens, F., Díaz, E., & Ruiz, G. (2005). Resistance to extinction of human evaluative conditioning using a between-subjects design. *Cognition & Emotion*, *19(2)*, 245–268. http://doi.org/10.1080/02699930441000300

Baeyens, F., Eelen, P., van den Bergh, O., & Crombez, G. (1989). Acquired affective-evaluative value: Conservative but not unchangeable. *Behaviour Research and Therapy*, *27(3)*, 279–287. http://doi.org/10.1016/0005-7967(89)90047-8

Barry, T. J., Hermans, D., Lenaert, B., Debeer, E., & Griffith, J. W. (2013). The eACS: Attentional control in the presence of emotion. *Personality and Individual Differences*, *55(7)*, 777–782. https://doi.org/10.1016/j.paid.2013.06.014

Baxter, A. J., Scott, K. M., Vos, T., & Whiteford, H. A. (2013). Global prevalence of Anxiety Disorder: A Systematic Review and Meta-Regression. *Psychological Medicine*, *43(5)*, 897-910.

Boddez, Y., Baeyens, F., Hermans, D., & Beckers, T. (2011). The hide-and-seek of retrospective revaluation: Recovery from blocking is context dependent in human causal learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 230-240. https://psycnet.apa.org/doi/10.1037/a0021460

Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., & Beckers, T. (2013). Rating data are underrated: Validity of us expectancy in human fear conditioning. *Journal of Behavior Therapy and Experimental Psychiatry*, *44(2)*, 201-206. https://doi.org/10.1016/j.jbtep.2012.08.003

Boddez, Y., De Houwer, J., & Beckers T. (2017). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford UK: Oxford University Press

Boddez, Y., Moors, A., Mertens, G., & De Houwer, J. (in press). Tackling fear: Beyond associative memory activation as the only determinant of fear responding. *Neuroscience & Biobehavioral Reviews*.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, 13, 195-212.

Craske, M. G., & Mystkowski, J. L. (2006). Exposure Therapy and Extinction: Clinical Studies. In M. G. Craske, D. Hermans, & D. Vansteenwegen (Eds.), *Fear and learning: From basic processes to clinical implications* (p. 217–233). American Psychological Association. https://doi.org/10.1037/11474-011

De Houwer, J. (2020). Conditioning is More Than Association Formation: On the Different Ways in Which Conditioning Research is Valuable for Clinical Psychology. *Collabra: Psychology*, *6(1)*. http://doi.org/10.1525/collabra.239

De Houwer, J. & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology: Section B*, *55(4)*, 289-310. https://doi.org/10.1080/02724990244000034

De Houwer, J., & Hughes, S. (2020). The psychology of learning: An introduction from a functional-cognitive perspective. Boston, MA: The MIT Press.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, *111(2)*, 225–236. https://doi.org/10.1037/0021-843X.111.2.225

Duits, P. (2016). Fear less. Individual differences in fear conditioning and their relation to treatment outcome in anxiety disorders (Doctoral dissertation, Utrecht University). https://dspace.library.uu.nl/bitstream/handle/1874/339826/Duits.pdf?sequence=1&isAllowed=y

Duits, P., Baas, J. M., Engelhard, I. M., Richter, J., Huisman-van Dijk, H. M., Limberg-Thiesen, A., ... Cath, D. C. (2018). Latent class growth analyses reveal overrepresentation of dysfunctional fear conditioning trajectories in patients with anxiety disorders compared to controls. Manuscript submitted for publication.

Forcadell, E., Torrents-Rodas, D., Treen, D., Fullana, M. A., & Tortella-Feliu, M. (2017a). Attentional control and fear extinction in subclinical fear: An exploratory study. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01654

Forcadell, E., Torrents-Rodas, D., Vervliet, B., Leiva, D., Tortella-Feliu, M., & Fullana, M. (2017b). Does fear extinction in the laboratory predict outcomes of exposure therapy? A treatment analog study. *International Journal of Psychophysiology*, *121*, 63-71. https://doi.org/10.1016/j.ijpsycho.2017.09.001

Galatzer-Levy, I. R., Bonanno, G. A., Bush, D. E. A., & LeDoux, J. E. (2013). Heterogeneity in threat extinction learning: Substantive and methodological considerations for identifying individual differences in response to stress. *Frontiers in Behavioral Neuroscience*, *7*.

Guthrie, R. M., & Bryant, R. A. (2012). Extinction learning before trauma and subsequent posttraumatic stress. *Psychosomatic Medicine*, *68*, 307–311. https://doi.org/10.1097/01.psy.0000208629.67653.cc

Haddad, A. D., Pritchett, D., Lissek, S., & Lau, J. Y. (2012). Trait anxiety and fear responses to safety cues: Stimulus generalization or sensitization?. Journal of Psychopathology and Behavioral Assessment, 34(3), 323-331

Johnson, D. C., & Casey, B. J. (2015). Easy to remember, difficult to forget: The Development of Fear Regulation. *Developmental Cognitive Neuroscience*, *11*, 42-55.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and

growth mixture modeling. *Social and Personality Psychology Compass*, *2(1),* 302–317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., Yanchar, S. C. (1996). The reliability and validity of the outcome questionnaire. *Clinical Psychology & Psychotherapy*, *3(4),* 249-258. 10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

LeDoux, J. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences of the United States of America*, *111(8),* 2871-2878. https://doi.org/10.1073/pnas.1400335111

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience*, *19(5),* 269–282. https://doi.org/10.1038/nrn.2018.22

Lenaert, B., Boddez, Y., Griffith, J. W., Vervliet, B., Schruers, K., & Hermans, D. (2014). Aversive learning and generalization predict subclinical levels of anxiety: A six-month longitudinal study. *Journal of Anxiety Disorders*, *28(8),* 747-753. https://doi.org/10.1016/j.janxdis.2014.09.006

Lommen, M., Engelhard, I., Sijbrandij, M., Hout, M., Hermans, D. (2013). Pre-trauma individual differences in extinction learning predict posttraumatic stress. *Behaviour Research and Therapy*, *51*, 63-67. 10.1016/j.brat.2012.11.004.

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J. … Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition,

extinction, and return of fear. *Neuroscience and Biobehavioral Reviews*, *77*, 247– 285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., Merz, C. J., Fullana, M. A., (2019) Fear Extinction Retention: Is It What We Think It Is?, *Biological Psychiatry*, *85(12),* 1074 - 1082, https://doi.org/10.1016/j.biopsych.2019.02.011.

Mertens, G., Boddez, Y., Sevenster, D., Engelhard, I., & De Houwer, J. (2018). A review on the effects of verbal instructions in human fear conditioning: Empirical findings, theoretical considerations, and future directions. *Biological Psychology*, *137*, 49-64. https://doi.org/10.1016/j.biopsycho.2018.07.002

Meulders, A., Boddez, Y., Vansteenwegen, D., & Baeyens, F. (2013). Unpredictability and context conditioning: Does the nature of the US matter? *The Spanish Journal of Psychology*, *16*.

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, *108*(3), 483–522. https://doi.org/10.1037/0033-295x.108.3.483

Olatunji, B. O., Cisler, J. M., & Deacon, B. J. (2010). Efficacy of cognitive behavioral therapy for anxiety disorders: A review of meta-analytic findings. *Psychiatric Clinics of North America*, *33(3)*, 557-577.

O'Malley, K. R., & Waters, A. M. (2018). Attention avoidance of the threat conditioned stimulus during extinction increases physiological arousal generalization and retention. *Behaviour Research and Therapy*, *104*, 51–61. https://doi.org/10.1016/j.brat.2018.03.001

Orr, S. P., Metzger, L. J., Lasko, N. B., Macklin, M. L., Peri, T., & Pitman, R. K. (2000). De novo conditioning in trauma-exposed individuals with and without posttraumatic stress

disorder. *Journal of Abnormal Psychology*, *109*, 290-298. https://doi.org/10.1037/0021-843X.109.2.290

Pineles, S. L., Nillni, Y. I., King, M. W., Patton, S. C., Bauer, M. R., Mostoufi, S. M., Gerber, M. R., Hauger, R., Resick, P. A., Rasmusson, A. M., & Orr, S. P. (2016). Extinction retention and the menstrual cycle: Different associations for women with posttraumatic stress disorder. *Journal of abnormal psychology*, *125*(3), 349–355. https://doi.org/10.1037/abn0000138

Pineño, O., & Miller, R. R. (2007). Comparing associative, statistical, and inferential reasoning accounts of human contingency learning. *The Quarterly Journal of Experimental Psychology*, *60*(3), 310-329.

Rescorla R. A., & Wagner A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robinson, O. J., Letkiewicz, A. M., Overstreet, C., Ernst, M., & Grillon, C. (2011). The effect of induced anxiety on cognition: Threat of shock enhances aversive processing in healthy individuals. *Cognitive, Affective & Behavioral Neuroscience*, *11(2),* 217–227. https://doi.org/10.3758/s13415-011-0030-5

Spielberger, C. D., & Gorsuch, R. L. (1983). Manual for the State-Trait Anxiety Inventory (Form Y) "Self- Evaluation Questionnaire". Palo Alto, CA: Consulting Psychologists Press.

Vervliet, B., & Boddez, Y. (in press). Aversive stimulus pairings are an unnecessary and insufficient cause of pathological anxiety. *Biological Psychiatry*.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: State of the art. *Annual Review of Clinical Psychology*, *9*, 215–248. https://doi.org/10.1146/annurev-clinpsy-050212-185542

Vroling, M., & de Jong, P. (2013). Belief bias and the extinction of induced fear. *Cognition & Emotion*, *27(8),* 1405-20. https://doi.org/10.1080/02699931.2013.792245

Zeidan, M. A., Lebron, M. K., Thompson, H. J., Im, J. J. Y., Dougherty, D. D., Holt, D. J., Orr, S. P., & Milad, M. R. (2012). Test–retest reliability during fear acquisition and fear extinction in humans. *CNS Neuroscience & Therapeutics*, *18(4),* 313–317. https://doi.org/10.1111/j.1755-5949.2011.00238.x

Tables

Table 1

Descriptives for the complete sample and the learning trajectories for the aversive and non-aversive US conditioning.

		Aversive US		Non-aversive US		
	Complete sample (*N* =156)	Normal (*n* = 96)	Poor extinction (*n* = 60)	Poor acquisition (*n* = 38)	Normal (*n* = 89)	Poor extinction (*n* = 29)
Demographics						
Age	20.49 (1.89)	20.42 (1.67)	20.60 (2.22)	20.24 (1.65)	20.54 (1.66)	20.66 (2.74)
Gender (female, *n*, %)	95 (60.89 %)	56 (58.33 %)	39 (65.00 %)	20 (52.63 %)	55 (61.79 %)	20 (68.96 %)
Conditioning-related						
Valence ratings						
Aversive US	-2.92 (2.04)	-2.94 (1.86)	-2.88 (2.33)	-2.82 (2.08)	-2.92 (2.02)	-3.03 (2.15)
Non-aversive US	1.06 (2.19)	1.23 (2.24)	0.80 (2.10)	1.29 (2.38)	1.09 (2.23)	0.69 (1.79)
CS-	1.24 (2.23)	1.06 (2.28)	1.53 (2.14)	1.24 (2.30)	1.31 (2.64)	1.03 (2.13)
Aversive CS+	-1.02 (2.14)	-0.63 (2.13)	-1.65 (2.02)	-1.26 (2.19)	-0.80 (2.19)	-1.38 (1.92)

CONDITIONING WITH A NON-AVERSIVE US						
Non-aversive CS+	1.08 (1.99)	1.20 (1.96)	0.90 (2.06)	1.47 (2.08)	1.15 (2.01)	0.38 (1.72)
Contingency-awareness						
Aversive CS+ (*n,* %)	110 (70.51 %)	63 (65.62 %)	47 (78.33 %)	28 (73.68 %)	60 (67.41 %)	22 (75.86 %)
Non-aversive CS+ (*n*, %)	102 (65.38 %)	68 (70.83 %)	34 (56.66 %)	27 (71.05 %)	53 (59.55 %)	22 (75.86 %)
Both (*n*, %)	77 (49.35 %)	48 (50,00 %)	29 (48.33 %)	21 (55.26 %)	38 (42.69 %)	18 (62.06 %)

Table 2

Overview of the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and entropy scores for the six models estimated with LCGA for aversive and non-aversive US conditioning.

		1 class	2 classes	3 classes	4 classes	5 classes	6 classes
BIC							
US expectancy	Aversive US	29173	28533	28328	28250	28200	28174
	Non-aversive US	29433	28756	28371	28208	28152	28102
AIC							
US expectancy	Aversive US	29051	28347	28078	27936	27822	27732
	Non-aversive US	29321	28563	28112	27894	27774	27659
Entropy							
US expectancy	Aversive US	NA.	.944	.974	.953	.966	.962
	Non-aversive US	NA.	.991	.974	.966	.962	.966

Table 3

Overlap in membership for aversive and non-aversive US conditioning trajectories in percentages and absolute numbers (in brackets) for the complete sample (N = 156)

Type of US		Aversive US		Non-aversive US		
		'Normal'	'Poor extinction'	'Poor acquisition'	'Normal'	'Poor extinction'
Aversive US	'Normal'			26.04 (25)	67.70 (65)	6.25 (6)
	'Poor extinction'			21.66 (13)	40.00 (24)	38.33 (23)
Non-aversive US	'Poor acquisition'	65.78 (25)	34.21 (13)			
	'Normal'	73.03 (65)	26.96 (24)			
	'Poor extinction'	20.68 (6)	79.31 (23)			

Table 4

Overlap in membership for aversive and non-aversive US conditioning trajectories in percentages and absolute numbers (in brackets) for participants with complete contingency awareness (n =77).

Type of US		Aversive US		Non-aversive US		
		'Normal'	'Poor extinction'	'Poor acquisition'	'Normal'	'Poor extinction'
Aversive US	'Normal'			31.25 (15)	64.58 (31)	4.16 (2)
	'Poor extinction'			20.68 (6)	24.13 (7)	55.17 (16)
Non-aversive US	'Poor acquisition'	71.42 (15)	28.57 (6)			
	'Normal'	81.57 (31)	18.42 (7)			
	'Poor extinction'	11.11 (2)	88.88 (16)			

Figures [all figures should be printed in black and white in a print version of the manuscript]

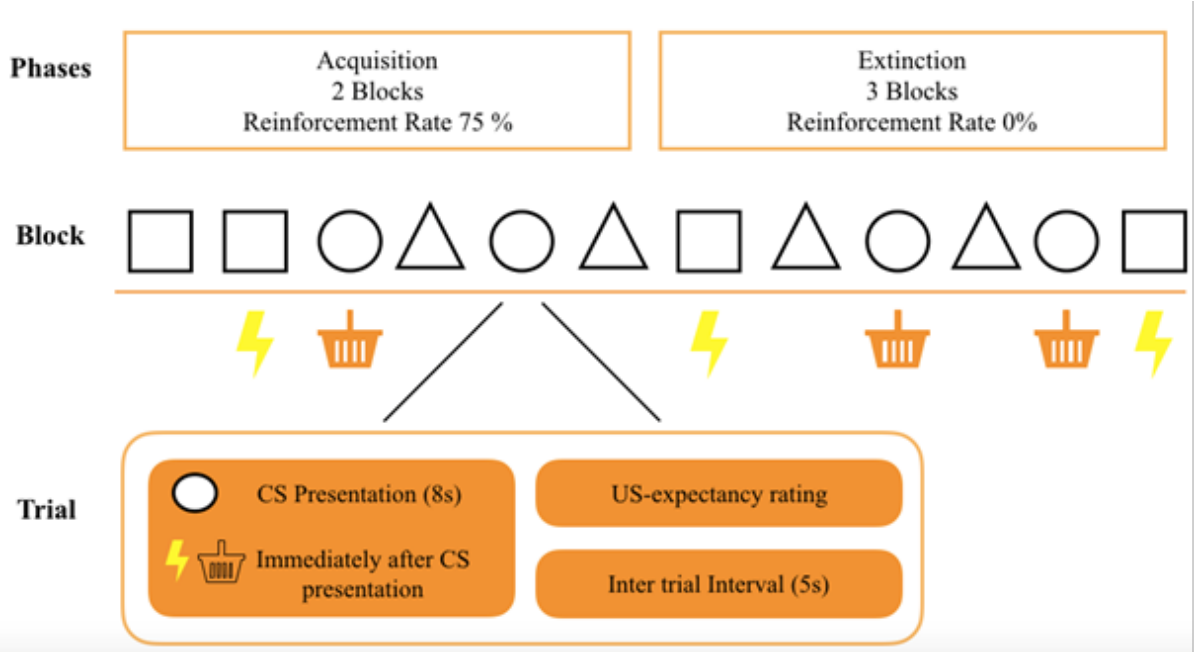

Figure 1. Schematic overview of the conditioning procedure. Each CS was presented four times during each block, resulting in a total of 12 trials per block. Trials proceeded similarly in the acquisition and extinction phase, with the only difference that during extinction no USs were presented. The lightning bolt indicates the aversive (shock) and the basket the non-aversive (picture) US.

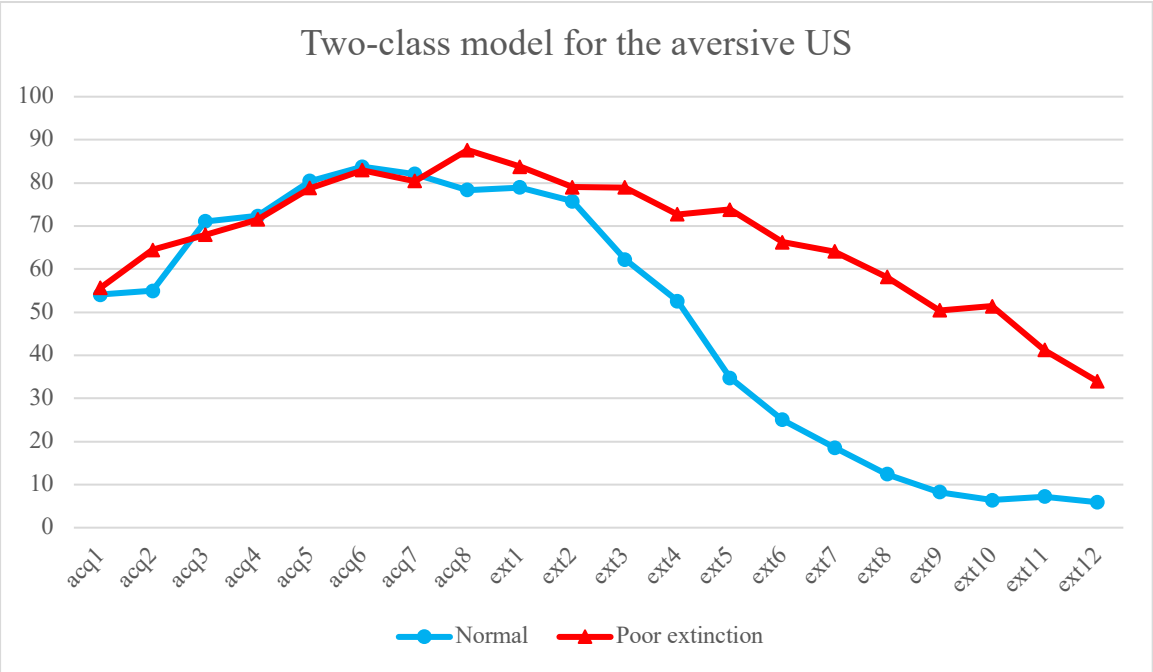

Figure 2. Two-class model of US-Expectancy ratings for conditioning with the aversive US.

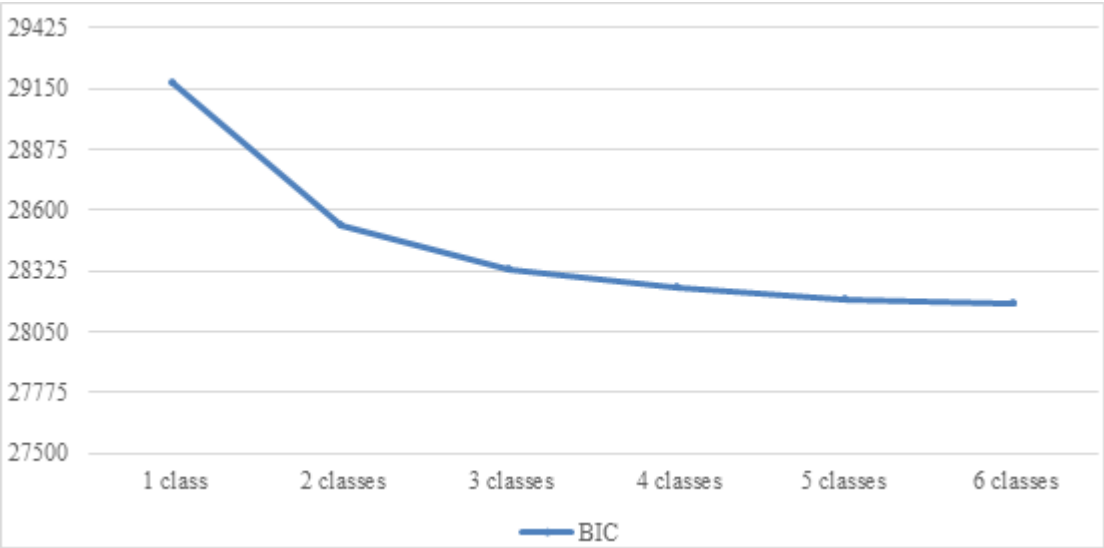

Figure 3. Bayesian Information criterion for the six different models estimated with LCGA for the aversive US-expectancy ratings.

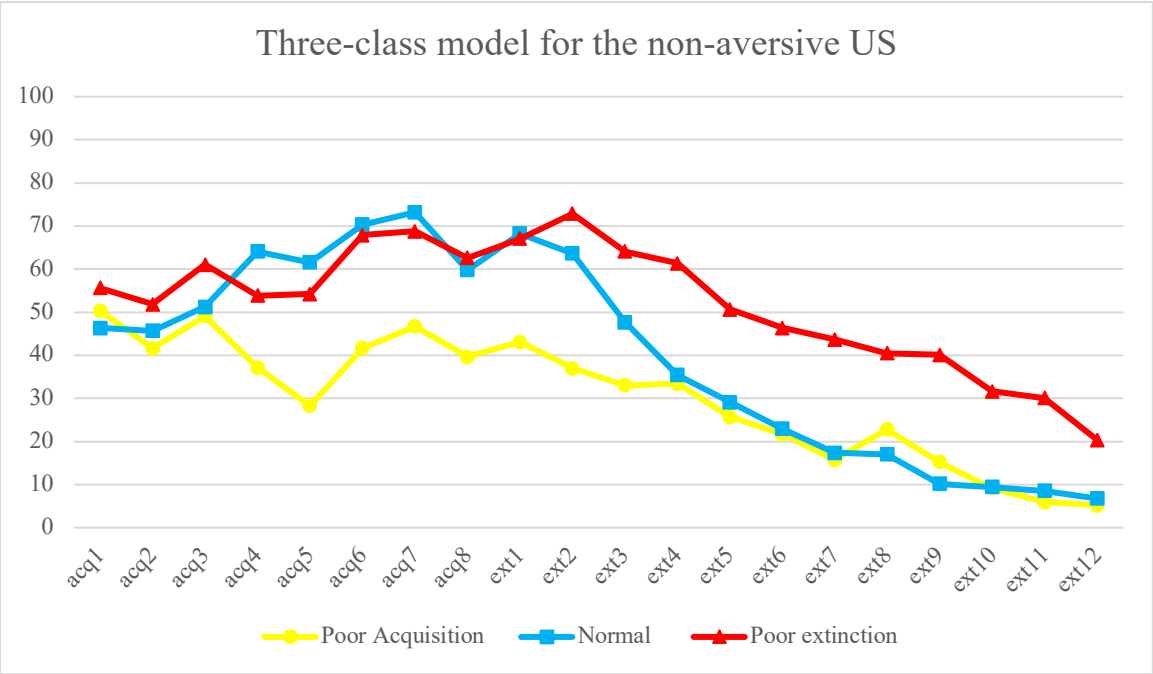

Figure 4. Three-class model of US-Expectancy ratings for conditioning with the non-aversive US.

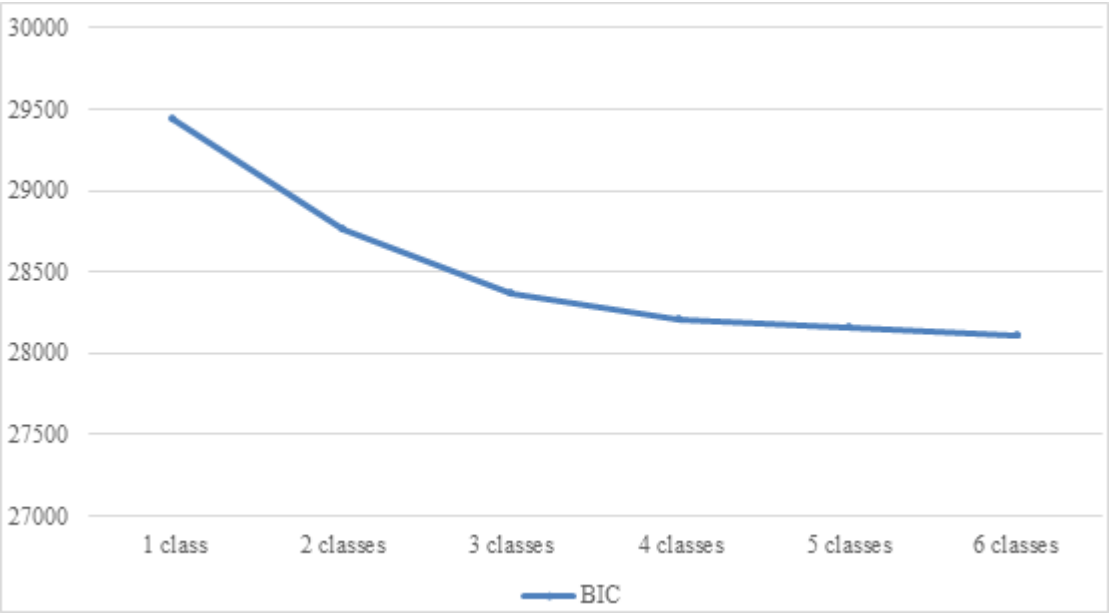

Figure 5. Bayesian Information criterion for the six different models estimated with LCGA for the non-aversive US-expectancy ratings

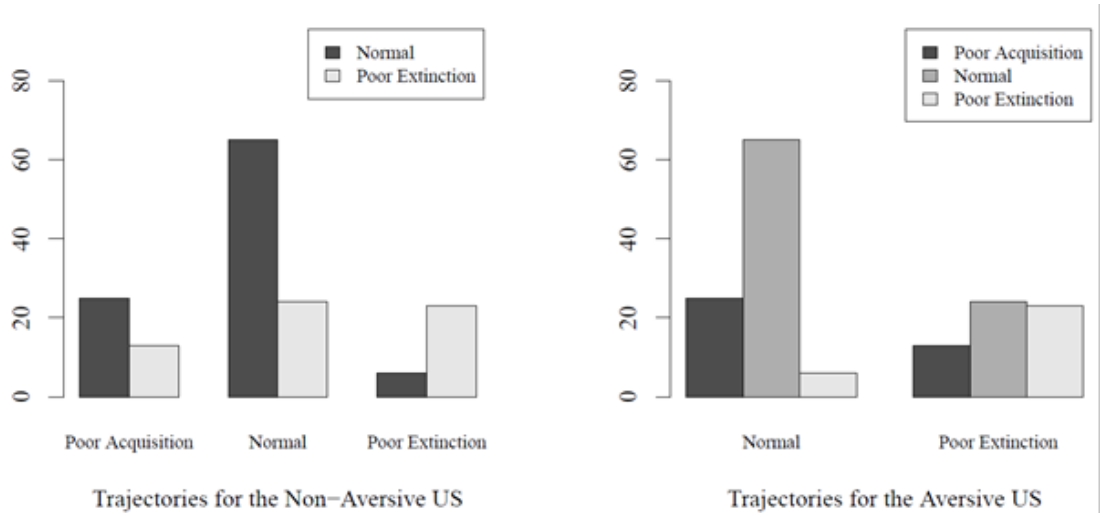

Figure 6. Bar plot representing the distribution of individuals in the aversive US classes depending on their non-aversive US trajectory and vice versa.