Habitual avoidance in trait anxiety and anxiety disorders

Habituelles Vermeidungsverhalten bei Ängstlichkeit und Angststörungen

Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences,

Julius-Maximilians-Universität Würzburg,

Section Neuroscience

submitted by

Valentina Glück

from Berlin

Würzburg, 2023

Members of the Thesis Committee

Chairperson:	Prof. Dr. Matthias Gamer
Primary Supervisor:	Prof. Dr. Andre Pittig
Second Supervisor:	Prof. Dr. Paul Pauli
Third Supervisor:	Prof. Dr. Grit Hein
Fourth Supervisor:	Prof. Dr. Roland Deutsch

Submitted on: December 19th 2023

Acknowledgements

This work would not have been possible without continuous support and collaboration. First, I would like to thank my supervisor, Prof. Dr. Andre Pittig, for continuously standing behind my work. Thank you for granting me time and space to do research on this topic, for being always open for discussions, and for your reliability and patience. I am grateful to have been part of your research group during the last few years. I would also like to thank my supervisors, Prof. Dr. Paul Pauli, Prof. Dr. Grit Hein, and Prof. Dr. Roland Deutsch, for providing valuable feedback and help during our meetings.

A huge thank you to Juliane Boschet-Lange and Paula Engelke. You have been the most reliable, kindest colleagues I could have wished for. I am grateful for having met you in the research group and to have shared an office with you for many years. Thank you for your ideas and your interest in my thoughts. I am also grateful to Alex Wong and Menghuan Chen for sharing your perspectives on my work and motivating me to move forward. Thank you so much, Tomko Settgast, for enlightening discussions. Thanks to my friends outside of research for your support. Also, thank you to my colleagues at the Department of Psychology I for continuous feedback, help, and a vital sense of community.

I also want to thank the Human Dynamics Center and the Gender Equality Programme at the Faculty for Humanities at the University of Würzburg for funding the experiments.

Lastly, I would like to thank my family. Thank you for always wanting to learn with me.

Table of Contents

Summary

Maladaptive avoidance behaviors can contribute to the maintenance of fear, anxiety, and anxiety disorders. It has been proposed that, throughout anxiety disorder progression, extensively repeated avoidance may become a habit (i.e., habitual avoidance) instead of being controlled by internal threat-related goals (i.e., goal-directed avoidance). However, the process of the acquisition of habitual avoidance in anxiety disorders is not yet well understood. Accordingly, the current thesis aimed to investigate experimentally whether trait anxiety and anxiety disorders are associated with an increased shift from goal-directed to habitual avoidance.

The aim of *Study 1* was to develop an experimental operationalization of maladaptive habitual avoidance. To this end, we adapted a commonly used action control task, the *outcome devaluation paradigm*. In this task, habitual avoidance was operationalized as persistent responses after extensive training to avoid an unpleasant stimulus when the aversive outcome was devalued, i.e., when individuals knew the aversive outcome could not occur anymore. We included indicators for costly and low-cost habitual avoidance, whereby habitual avoidance was associated with a monetary cost, while low-cost habitual avoidance was not associated with monetary costs. In Experiment 1 of Study 1, a pronounced costly and non-costly outcome devaluation effect was observed. However, this result may have partly resulted from trial-and-error learning or a better-safe-than-sorry strategy since not instructions about the stimulus-response-outcome contingencies after the outcome devaluation procedure had been provided to the participants. In Experiment 2 of Study 1, instructions on these stimulus-response-outcome contingencies were included to prevent the potential confounders. As a result, we observed no indicators for costly habitual avoidance, but evidence for low-cost habitual avoidance, potentially because competing goal-directed responses could easily be implemented and inhibited costly habitual avoidance tendencies.

In *Study 2*, the strength of habitual avoidance acquisition was compared between participants with and without anxiety disorders, using the experimental task of Experiment 1 in Study 1. The results indicated that costly and low-cost habitual avoidance was not more pronounced in participants with anxiety disorders than in the healthy control group. However, in an exploratory subgroup comparison, panic disorder predicted more substantial habitual avoidance acquisition than social anxiety disorder.

In *Study 3*, we investigated whether trait anxiety as a risk factor for anxiety disorders is associated with a specific increased shift from goal-directed to habitual avoidance and approach. The task from the Experiment 1 of Study 1 was adapted to include parallel versions for operationalizing habitual avoidance and habitual approach responses. Using a within-subjects design, the individuals – pre-screened for high and low trait anxiety – took part in the approach and the avoidance outcome devaluation task version. The results suggested stronger non-costly habitual responses in more highly trait-anxious individuals independent of the task version, and suggested a tendency towards an impact of trait anxiety on costly habitual approach rather than on costly habitual avoidance.

In summary, individuals with high trait anxiety or anxiety disorders did not develop habitual avoidance more readily than individuals with low trait anxiety or without anxiety disorders. Therefore, this thesis does not support the assumption that an increased tendency to acquire habitual avoidance contributes to persistent maladaptive avoidance in anxiety disorders. The thesis also contributes to the discourse on the validity of outcome devaluation studies in general by highlighting the impact of task features, such as the instructions after the outcome devaluation procedure or the task difficulty in the test phase, on the experimental results. Such validity issues may partly explain the heterogeneity of findings in research with the outcome devaluation paradigm. We suggest ways towards more valid operationalizations of habitual avoidance in future studies.

Zusammenfassung

Vermeidungsverhalten ist an der Aufrechterhaltung von Furcht, Angst und Angststörungen beteiligt. Maladaptives Vermeidungsverhalten ist in Bezug auf die objektiv vorliegende Bedeutung einer Bedrohung unverhältnismäßig und kann selbst in Abwesenheit von Furcht oder Angst anhalten. In ätiologischen Modellen zu maladaptiver Vermeidung wurde zur Erklärung solcher anhaltenden Vermeidung vorgeschlagen, dass sich Vermeidung über die Zeit von einer geplanten, zielgerichteten zu einer gewohnheitsmäßigen, habituellen Vermeidung entwickeln könnte. Die Rolle habituellen Vermeidungsverhaltens in der Entstehung und Aufrechterhaltung von Angststörungen ist bisher nicht ausreichend verstanden und untersucht worden. Die vorliegenden Studien prüfen, ob Trait-Ängstlichkeit als Risikofaktor für Angststörungen sowie bereits vorliegende Angststörungen mit einer verstärkten Tendenz zur Ausbildung habitueller Vermeidung einhergehen.

In der *ersten Studie* wurde eine häufig verwendete experimentelle Aufgabe zur Untersuchung von zielgerichteter und habitueller Handlungssteuerung, das *Ergebnis-Devaluations-Paradigma*, weiterentwickelt. Gewohnheitsmäßige Vermeidung wurde hierbei als jene Tendenz operationalisiert, ein Vermeidungsverhalten, nachdem es ausführlich trainiert worden war, auch dann noch auszuführen, wenn die entsprechende Bedrohung nicht mehr bedeutsam, d.h. devaluiert war. Die fortgeführte, habituelle Vermeidung konnte – bei sogenannter kostspieliger habitueller Vermeidung – mit finanziellen Kosten verbunden sein oder – bei der sogenannten kostenarmen habituellen Vermeidung – keine finanziellen Kosten verursachen. Im ersten Experiment der ersten Studie wurden nach dem ausführliche Vermeidungstraining sowohl kostspielige als auch kostenarme fortgeführte Vermeidung beobachtet. Diese Effekte konnten jedoch möglicherweise auf Versuch-und-Irrtum-Lernen erklärt werden, auf das die Versuchsteilnehmer_innen möglicherweise zurückgriffen, da sie nach dem Vermeidungstraining keine ausreichenden Informationen darüber erhalten hatten, welche Reaktionen zu Kosten oder keinen Kosten führten (Stimulus-Reaktions-Ergebnis-Zusammenhänge). Im zweiten Experiment der ersten Studie wurden die Stimulus-Reaktions-Ergebnis-Zusammenhänge explizit erklärt, um Versuch-und-Irrtum-Lernen zu verhindern. Kostenreiche habituelle Vermeidung wurde nun nicht mehr beobachtet, dafür jedoch ein Hinweis auf kostenarme Vermeidung. Dies könnte mit der niedrigeren Aufgabenschwierigkeit und der damit verbundenen Erleichterung von zielgerichtetem Handeln erklärt werden, wodurch möglicherweise kostenreiche habituelle Tendenzen abgeschwächt werden konnte. In der *zweiten Studie* wurde die experimentelle

Aufgabe aus dem ersten Experiment der ersten Studie erneut verwendet, um die Stärke habitueller Vermeidung zwischen Personen mit und ohne Angststörungen zu vergleichen. Im Ergebnis zeigte sich keine verstärkte kostenreiche oder kostenarme habituelle Vermeidung bei Personen mit Angststörungen im Vergleich mit der gesunden Kontrollgruppe. In einer explorativen Analyse zeigte sich zwar stärkere habituelle Vermeidung bei Personen mit Panikstörung als bei Personen mit sozialer Angststörung, ein mögliches Versuch-und-Irrtum-Lernen erschwerte jedoch wieder die Interpretation dieser Ergebnisse. Die experimentelle Aufgabe wurde in der *dritten Studie* deshalb weiter angepasst und um eine parallele Version zur Untersuchung habitueller Annäherung erweitert. Personen mit hoher und niedriger Trait-Ängstlichkeit nahmen bearbeiteten sowohl die Annäherungs- als auch der Vermeidungsversion. Die Trait-Ängstlichkeit der Teilnehmer_innen sagte eine stärkere Tendenz zu kostenarmen habituellen Reaktionen vorher. Zudem fanden wir einen Hinweis auf stärkere kostenreiche habituelle Annäherung, nicht jedoch habituelle Vermeidung bei Personen mit höherer Trait-Ängstlichkeit. Auch unabhängig von Trait-Ängstlichkeit wurde eine stärkere habituelle Annäherung als eine habituelle Vermeidung beobachtet. Möglicherweise hingen diese Unterschiede erneut mit unerwarteten Aufgabeneffekten zusammen.

Zusammenfassend zeigten Personen mit hoher Trait-Ängstlichkeit oder Angststörungen in den vorliegenden Studien keine stärkere habituelle Vermeidung als Personen mit niedriger Trait-Ängstlichkeit oder ohne Angststörungen. Die verstärkte Entstehung habitueller Vermeidung erschien daher nicht als relevanter Faktor in der Aufrechterhaltung maladaptiven Vermeidungsverhaltens bei Personen mit Angststörungen. Die Arbeit zeigt zudem deutlich den Einfluss von Aufgabendetails auf die experimentellen Ergebnisse in Ergebnis-Devaluations-Paradigmen auf. Die Diskussion dieser Ergebnisse könnte zur Erhöhung der Validität in zukünftigen Ergebnis-Devaluations-Studien beitragen.

1. General Introduction

The dictum “fear is a bad advisor” conveys a seemingly straightforward message: Fear can influence decision-making in unfavorable ways. However, when contemplating the sentence, it becomes apparent that it does not address *how* fear is assumed to impact decisions. At least two general routes seem plausible. First, fear or anxiety may lead to suboptimal decisions via an inflexible preference for safe options, which can be accompanied by a reduced exploration of different available opportunities and actions (Paulus & Yu, 2012). Consistent avoidance of slight risks, for example, can reduce the exploration of options that are associated with potential rewards (e.g.., Pittig & Scherbaum, 2020). When fear influences behavior strongly, decisions can, thus, become unbalanced. Therefore, fear can influence *what* an individual chooses to do. Second, however, the dictum may also mean that fear and anxiety change *how* we make decisions. Perceived threats often imply urgency to take action to control that threat, potentially leading to hasty decisions. Besides potential freezing and an inability to act in highly fearful states (e.g., Roelofs, 2017), the fear-related pressure to resolve a threatening situation may not allow for careful consideration of the potential consequences of all available behavioral options. Fear may thus lead to a preference for fast solutions that reduce fearful states, falling short of integrating less salient, potentially rewarding outcomes (e.g., LeDoux & Daw, 2018). Relatedly, fear may influence how we make decisions via a preference for repeating well-known, extensively rehearsed actions (Hartley & Phelps, 2012). However, being able to form flexible decisions that do not permanently restrict behavioral opportunities is central to navigating complex, ambiguous environments. This thesis aims to examine the influence of anxiety on the flexibility of the regulation of avoidance. Such research may contribute to understanding how avoidance becomes inflexible and, sometimes, restrictive.

One interrogative theme in action control research centers around the assumption of different action control processes, their potential definitions, and interactions. These assumptions and debates will be briefly outlined to provide some background for the theoretical and methodological approach of the thesis. First, the associative dual-process model of action control and its delineation of habitual and goal-directed action control components will be elucidated. Second, experimental operationalizations of habitual and goal-directed responses, the current debates about their validity, and the appropriate interpretations of experimental results obtained from outcome devaluation studies will be discussed. Third, it will be outlined why action control studies are relevant for explaining persistent avoidance behaviors in anxiety disorders. The current evidence on associations between

anxiety and shifts from goal-directed to habitual decision-making components will be briefly summarized, highlighting a potentially intensified shift toward habitual avoidance. As will be explained, the current evidence, however, is insufficient to evaluate these claims. The introduction closes with an outline of the research questions addressed in this thesis.

1.1 Action control models

The question of how humans control their actions has a long history in philosophical and psychological theories that have often featured a balance between two processes. For example, a theory involving an interaction between intuitive and reasoning-based processes in action control was already apparent in ancient Greek philosophy (see Keren & Schul, 2009). Early experimental psychology also produced various action control theories delineating different forms of response control. For example, William James described the ideomotor theory in 1890, which was influenced by 19th-century theories from philosophy and physiology. The ideomotor theory states that after a phase of learning, the cognitive associations between actions and their consequences would be so strong that the mere idea or mental representation of a consequence would directly, reflex-like, elicit an associated motor response (see Shin et al., 2010). Therefore, the representation of an action outcome was very closely and causally related to the elicitation of the action. Of note, James (1890/2021) presumed this direct association between outcomes and actions to be a ubiquitous mechanism that guides actions, while some of the earlier theories on which James built the ideomotor theory had assumed such reflex-like ideomotor actions to be apparent only in disordered states with compromised willful action control (Stock & Stock, 2004). The ideomotor theory was criticized by Edward Thorndike (1913), who, in the wake of behaviorist theory, rejected the assumption that representations or "ideas" were necessary or even possible elicitors of behavior (Stock & Stock, 2004). Thorndike highlighted the view that responses were elicited by the perception of environmental stimuli if an external reinforcer had followed the response regularly and had thereby strengthened the stimulus-response association ("Law of Effect"). Another early empirical theory of action control by Narziß Ach (1910) assumed an interaction between stimulus-dependent, learned habits and willful actions (Hommel, 2019). Ach derived experimental results from controlled designs to provide a quantitative measure of habitual actions resembling experimental designs still in use today. First, stimulus-response associations were extensively trained; then, the tendency to repeat these associations was tested (Ach, 1910). Further examples of early accounts of action control stem from Kurt Lewin (1922a, 1922b), who proposed that habitual actions were embedded within goal-directed

planning, and Edward Tolman (1948), who centered the role of represented action consequences in action control (see Hommel, 2019). To summarize, whether actions are performed to reach a specified goal or external stimuli can directly elicit behavior has been central in many action control theories and remains debated in action control research today. Central points in the current discourse on action control in the experimental literature are going to be presented in the following.

1.1.1 Action control in the associative dual-process model

Associative learning models, in general, assume that the performance of actions is deeply intertwined with associative learning processes and that identifying patterns in the environment is a necessary cognitive function for being able to generate flexible, adaptive responses (Rescorla & Wagner, 1972; for reviews, see Colwill et al., 2022; Dymond, 2019). The beginning of empirical work in the context of the current associative dual-process model of action control dates back to a seminal study by Adams and Dickinson (1981), which introduced the outcome devaluation paradigm becoming formative for experimental action control research. In the experiment by Adams and Dickinson (1981), hungry rats first learned that they could obtain two different food outcomes when pressing one of two levers. Thus, they acquired the associations between the lever pressing and the two foods, forming response-outcome associations (R-O associations) in an extensive training phase over several days that comprised at least 600 hundred responses per lever. After this instrumental training phase, one of the two food outcomes was selectively devalued. The outcome devaluation procedure consisted of the injection of a nausea-inducing substance in a context where the rat could access only one of the two food outcomes. This temporal pairing between the aversive stimulus and the food outcome was assumed to reduce the reinforcing strength of the food outcome (i.e., devalued outcome). In contrast, the reinforcing strength of the other food outcome, which had not been paired with nausea, was assumed to be intact (i.e., still-valued outcome). Lastly, the rats were placed in a context where they could again press both levers freely. If associations between the responses and the outcomes (i.e., R-O associations) were responsible for the responses, it was predicted that the outcome devaluation procedure should reduce the frequency of the response associated with the devalued outcome. In other words, the rats would be expected to be less inclined to produce the food that had been associated with nausea. In contrast, if direct associations between the visual perception of the lever and the response guided the response (S-R associations), the outcome devaluation should have not affected the responses. The perception of the lever would then directly elicit the

extensively trained response that it had associated with the outcome regardless of whether the outcome had been paired with nausea. The study revealed that the rats pressed the lever which was associated with the still-valued outcome more frequently than the lever which was associated with the devalued outcome, and, thus, adjusted their responses to the pairing with the nauseating substance. Adams & Dickinson (1981) inferred that the instrumental behavior was sensitive to the value of the reinforcers instead of being merely elicited by the environmental stimulus. In subsequent rodent studies, however, they observed an increase in outcome insensitivity as a function of training duration. Thus, after more extensive training, the sensitivity of the behavior to changes in outcome values declined (Adams, 1982; Dickinson, 1985; Dickinson et al., 1995). In summary, the associative dual-process framework states that instrumental actions can be controlled habitually and goal-directedly, whereby goal-directed action control is based on R-O contingencies and is sensitive to outcome values, and habitual action control is based on direct S-R associations that develop via extensive repetition (Balleine & Dickinson, 1998; Wood & Rünger, 2016). The dual-process associative model of action control and, correspondingly, the outcome devaluation paradigm have guided action control research for several decades and are still influential today (Watson & Wit, 2018).

Within the framework of the dual-process associative model of action control, different conceptualizations of the interplay or arbitration between the two processes have been brought up. For instance, some models assume that habitual processes are controlling behavior by default unless goal-directed processes intervene and, thus, downregulate the default habitual responses (e.g., Evans & Stanovich, 2013). Hierarchical models assume more complex collaborations between habitual and goal-directed processes, such as the chunking of actions into sequences (i.e., habits) that are controlled goal-directedly (Balleine & Dezfouli, 2019). The assumption that some extensively rehearsed actions are not selected based on complex goal-directed decision-making seems plausible given the high number of actions continuously performed by humans and animals in naturalistic environments. If each action initiation required constant updates of stimulus-response-outcome (S-R-O) contingency knowledge and comparisons between different outcome values as an integration of contextual cognitive, affective, social, and psychophysiological information, an overload for the information processing system could be the result (Ernst & Paulus, 2005; Wood & Rünger, 2016).

A central assumption in the associative dual-process model of action control concerns the role of repetition in the development of habitual responses (Balleine & Dickinson, 1998).

Functionally, the acquisition of automatic control in stable environments can be considered beneficial because exploiting regularities in the environment saves cognitive resources (Evans & Stanovich, 2013; Moors & de Houwer, 2006). Several animal studies supported the assumption that habitual control develops as a function of repetition of the respective behavior in stable contexts (see Yin & Knowlton, 2006). However, some animal studies reported no evidence for an association between the degree of habitual responses and the amount of training (Colwill & Rescorla, 1985) or did not find evidence for habitual responses even after extensive training durations (Colwill & Rescorla, 1985; Garr et al., 2021). The evidence on the assumed direct association between the amount of instrumental training and the strength of outcome insensitivity in humans is even more inconclusive. Only one study with humans demonstrated a direct association between training duration and habitual responses (e.g., Tricomi et al., 2009), a finding that was not replicated in two direct replication attempts (Gera et al., 2023; Pool et al., 2022). Two additional comprehensive studies in humans did not report more pronounced habits after more pronounced training (de Houwer et al., 2018; de Wit et al., 2018). Another study found an association between training duration and habit strength, but only in a subgroup of stressed individuals, which may suggest that the association between training duration and outcome insensitivity underlies boundary conditions (Pool et al., 2022). The exact conditions for goal-directed and habitual control are therefore still a matter of debate.

Empirical evidence supporting the dual-process model assumptions comes from neuroimaging studies in humans demonstrating that a shift from goal-directed to habitual control was accompanied by a shift from activation in the ventromedial prefrontal cortex and the ventral striatum to the dorsal striatum (e.g., Tricomi et al., 2009; Valentin et al., 2007; Zwosta et al., 2018; for a review see Peak et al., 2019). Since ventral striatal areas, which include the nucleus accumbens, are involved in learning the values of stimuli, this shift was proposed to suggest a reduced involvement of goal values in habitual responses (Peak et al., 2019). However, two direct replication attempts of one of the milestone studies on the neural correlates of habit acquisition in humans (Tricomi et al., 2009) did not replicate that a shift towards the ventral striatum was involved in habit acquisition (Gera et al., 2023; Pool et al., 2022). Of note, there is also substantial evidence for discernable neural patterns associated with habitual and goal-directed responding in rodents (for reviews, see Cain, 2019; Yin & Knowlton, 2006).

1.1.2 The outcome devaluation paradigm in humans: a matter of debate

The outcome devaluation paradigm has been translated to human research rather recently (Valentin et al., 2007). The underlying rationales in animal and human outcome devaluation studies are similar in many ways. They rest on the assumptions that firstly, goal-directed, but not habitual responses, are sensitive to changing R-O contingencies or outcome values and, secondly, habitual responses are acquired as a function of repetition (Vandaele & Janak, 2018). Concerning the experimental design (i.e., the sequence of extensive training, outcome devaluation procedure, and test phase), however, several adjustments in experiments on human action control are noticeable. First, in both animal and human research, the response options are simple and predefined by the experimental design. Second, in animal studies, the outcomes are often strongly rewarding (e.g., food for hungry rats in Balleine & Dickinson, 2005). In contrast, instead of direct primary reinforcers (e.g., immediate food availability), studies on humans frequently use secondary reinforcers such as pictures symbolizing food outcomes (e.g., Gera et al., 2023). Third, instead of induced sickness, outcome devaluation procedures in humans consist of specific satiation procedures (i.e., the participants consume one appetitive outcome until the desire for it recedes, e.g., Schwabe & Wolf, 2009; Tricomi et al., 2009; Valentin et al., 2007), instructions (i.e., instructed unavailability of rewards, e.g., Gillan et al., 2015), taste aversion (i.e., pairing one food outcome with a bad taste, e.g., Buabang, Boddez, et al., 2023), or the removal of device the which had delivered the outcome during the training (i.e., removing one of two electrodes for electrotactile stimulations, e.g., Gillan et al., 2014). Besides direct outcome value manipulations, response-outcome contingency degradation procedures involving changes of the R-O relationship can also be used to test for outcome sensitivity in humans (e.g., Vaghi et al., 2019). Concerning the operationalization of habitual and goal-directed responses in outcome devaluation tasks, the frequencies of the two responses to the still valued and the devalued outcomes are compared both in animal and human research. Goal-directed control is inferred when R-O contingency or outcome value changes are followed by adjusted response frequencies. In contrast, habitual responding is inferred when the response frequency is not affected by outcome contingency or outcome value changes since such non-adjusted responses are interpreted to result from outcome insensitivity.

The assumption of two processes controlling actions has been criticized for various reasons, and several action control models have been brought up which assume only one process (e.g., Kruglanski et al., 2006; Moors et al., 2017) or more than two processes (e.g., Melnikoff & Bargh,

2018). One recurrent critique regarding the associative dual-process model is the inconsistent use of definitions in the literature. The definitions have been criticized for being too vague, with, for example, one process broadly being described as "reflective" and the other as "reflexive" without precise descriptions of these terms (see de Houwer et al., 2022; Keren & Schul, 2009). Evans (2008), for example, listed 14 different definitions of the two processes in the associative dual-process literature. Speculatively, these imprecise definitions may be partly caused by the long history of dual-process accounts that heuristically distinguish two processes involved in decision-making, memory, or attention. In this regard, many dichotomous models in psychological research share a "common philosophical core" (Collins & Cockburn, 2020, p. 579) that consists of the assumption that two processes work together to control the phenomenon of interest, of which one is explicit, slow, and computationally heavy, while the other is implicit, fast, and computationally undemanding (e.g., Kahneman & Tversky, 1979). The specific definitions and operationalizations of these processes, however, vary widely between and within research areas. For example, in memory research, spatial navigation as driven by dorsal striatum-dependent S-R navigation versus hippocampus-dependent navigation via cognitive maps have been contrasted (Schwabe & Wolf, 2013). Additionally, declarative and procedural memory or goal-directed and habitual memory have been conceptualized as complementary processes (Schwabe & Wolf, 2013). In theories on attention, automatic processes, which develop through extensive, consistent training and pose low demands on working memory have been contrasted with cognitively controlled processes, which load heavily on working memory capacity (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). These two attentional processes were later reconceptualized as a bottom-up system that facilitates the detection of salient stimuli without feedback from higher-order processes and a top-down system that enables goal-directed focused attention (Corbetta & Shulman, 2002). These similar, yet not identical, terminologies and concepts in different research areas may have added to the lack of concise definitions of goal-directed and habitual action control.

Critically, different experimental paradigms have been employed to operationalize the habitual versus goal-directed action control dichotomy, such as the slips-of-action task (e.g., de Wit et al., 2018; Wit et al., 2007) and the already outlined outcome devaluation paradigm (Valentin et al., 2007). Importantly, within the research that used outcome devaluation tasks, various implementations of training phases, outcome devaluation procedures, and test phases exist. The impacts of these variations on the test phase responses have, with the exception of research on different training durations (de Houwer et al., 2018; de Wit et al., 2018; Gera et al., 2023; Pool et al., 2022; Tricomi et al., 2009), not yet been systematically

compared. An additional inconsistency in the area of action control research is that the term *habit* can denote relatively complex actions such as eating habits or exercising habits (e.g., Evans & Stanovich, 2013) but can also describe also more straightforward behavioral routines and heavily trained motor sequences that do not require explicit cognitive control or controlled attention, such as driving a car or playing a musical instrument (Du et al., 2022). This heterogeneity of the to-be-explained actions that are subsumed under the same terms arguably further complicates the comparability of results from different studies action control research.

As the last critical point of research on habits in the associative dual-process literature, the term habit is used interchangeably to describe observable behaviors (i.e., habitual behavior) or the assumed underlying mechanisms of action control (i.e., habitual control driven by S-R associations). Such identification of the process to be explained with the explanatory process may create a circular structure that makes the disconfirmation of central dual-process model assumptions impossible (de Houwer, 2019). It was argued that, since behavior may virtually in every case result from goals that may be unconscious and therefore not easily observed, the absence of goals could never be confirmed (de Houwer, 2019). Thus, observed seemingly goal-independent behavior does not always necessarily reflect goal-independent underlying processes (de Houwer, 2019). Therefore, observed behavior should not be understood as if directly reflecting the presumed underlying goal-directed and habitual processes (de Houwer, 2019). Distinguishing between, for example, the term *habit* to describe the assumed underlying control process and the term *habitual behavior* to describe the observed behavior, if used consistently, may help to eliminate the ambiguity between terms for the explanandum and the explanans in habit research (de Houwer, 2019; Rebar et al., 2018). As potential solutions to these problems, it was suggested that studies should present explicit definitions of habitual and goal-directed processes and avoid equating behavioral observations (i.e., habit indicators derived from experimental tasks) and the processes assumed to underlie them (see de Houwer et al., 2022).

1.1.3 Debates on the sensitive detection of goals

Critical accounts of the validity of the operationalizations of habitual and goal-directed processes in outcome devaluation paradigms have gathered strength recently. Internal validity of an experimental task denotes that the data derived from the experiment allow for valid conclusions about the phenomenon under investigation (e.g., Lundh, 2019). Thus, if an outcome devaluation task is not valid, the derived results would not be informative about the phenomena under question, namely habitual and goal-directed responses (e.g., Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023; de Houwer et al., 2022; Moors et al., 2017). The validity of an of

outcome devaluation task does, then, depend especially on its sensitivity to detect the potential goal-directed processes influencing responses in the test phase. Relatedly, outcome devaluation tasks have been criticized for not directly measuring habitual control. Since habitual control in outcome devaluation paradigms is indirectly inferred from reduced or absent goal-directed control, processes that impair goal-directed control can lead to erroneous classifications of such responses as habitual (de Houwer et al., 2018). Therefore, it can be considered to be even more important to account for all potential goals in the interpretation of non-adjusted responses as indicating habitual responses (de Houwer, 2019; de Houwer et al., 2018; Moors et al., 2017; Vandaele & Janak, 2018). In other words, non-adjusted responses after an outcome devaluation may not necessarily indicate habitual responses since they can theoretically also result from goals that are not sensitively detected by the experimental structure (de Houwer, 2019; de Houwer et al., 2018; Moors et al., 2017).

Numerous goals and strategies that may not be anticipated by a researcher may lead to non-adjusted and, thus, seemingly outcome-insensitive responses in outcome devaluation paradigms. For example, an outcome devaluation manipulation may be ambiguous when only one of two electrodes that had been used to deliver aversive outcomes during the training phase is removed (e.g., Gillan et al., 2015; Gillan et al., 2014). In this case, some participants might expect to receive aversive stimulations after the outcome devaluation from the still attached electrode and may, therefore, continue to avoid it. Importantly, non-adjusted responses in reaction to the outcome devaluation would then be goal-directed. Of note, such an expectation of ongoing aversive stimulations may especially arise in participants with pronounced goals of seeking safe states, potentially leading them to avoid persistently in this situation. Such a difference in expectations may bias group comparisons, for example, in studies with patients with obsessive-compulsive disorder or anxiety disorders who may prefer to avoid unnecessarily rather than risk encounters with aversive outcomes. Similarly, if one study group has stronger expectations about the potential ineffectiveness of the aversive outcome devaluation, the devaluation procedure would not effectively devalue the outcome in this group. The non-adjusted responses in this group may then be erroneously interpreted as habitual responses while they would, in fact, reflect stronger goal-directed avoidance. Systematic differences concerning goals or strategies in the test phase resulting from the participants' expectations about the potential ineffective devaluation of the aversive outcome devaluation procedure may, thus, bias clinical studies. Further examples of goal-directed strategies that may lead to unadjusted responses are strategies to reduce cognitive effort or to regulate stress (Buabang, Boddez, et al., 2023). For example, individuals who are

unmotivated to comply with the experimental task logic and primarily aim to receive money or credit points for their participation may continue to perform the overtrained responses after the outcome devaluation to reduce cognitive effort. As an example of a stress regulation strategy, participants may continue to consume a devalued food because eating makes them feel comfortable (see Buabang, Boddez, et al., 2023). It has thus been argued that persistent, non-adjusted responses after outcome devaluation can only validly be interpreted as habitual responses if all potential goals that may drive these persistent responses are being ruled out (Moors et al., 2017). To eliminate alternative goals and goal-directed strategies in outcome devaluation paradigms, complete and robust outcome devaluation procedures have been called for (Buabang, Köster, et al., 2023).

The problem of undetected goals that are potentially influencing the participants' behavior may be amplified when outcome evaluation paradigms do not incorporate disadvantages or costs for a non-adjustment of responses after the outcome devaluation procedure. In this case, there is no disadvantage to the persistent performance of the previously adaptive responses. The continuation of the devalued response due to an explicit strategy may then be beneficial because it reduces potential risks while creating no costs. Without costs for non-adjustment, non-adjustment can even be considered a rational and objectively advantageous choice. However, to the best of our knowledge, no outcome devaluation study has, so far, included costs for the continuation of previously trained responses in the test phase. The incorporation of costs has been discussed to potentially enhance the external validity of experimental avoidance research (Krypotos et al., 2018; Pittig et al., 2020).

1.1.4 Another framework: reinforcement learning models

Action control research in humans today does not only include experimental paradigms in the realm of the associative dual-process framework but has relatively recently been enriched by another experimental approach to action control, i.e., reinforcement learning models (e.g., Daw et al., 2011; Dayan, 2009). These models discern two classes of computational strategies underlying learning and, relatedly, decision-making, namely model-free reinforcement learning and model-based reinforcement learning (e.g., Daw et al., 2011; Dayan, 2009). The concept of model-free learning describes that contingencies between responses and outcomes are learned based on state-dependent assignments of values to response options. The response value is then used for action selection without any need for inferences about potential outcomes (Cushman & Morris, 2015; Drummond & Niv, 2020). In that regard, model-free reinforcement learning

resembles habitual control (Dayan, 2016). Model-based reinforcement learning, in contrast, is conceptualized to enable flexible planning based on complex representations of internal and external states, the transition probabilities between the states based on different actions, and the outcomes in each state (i.e., state values). In line with assumed similarities between habitual control and model-free reinforcement learning as well as between goal-directed control and model-based reinforcement learning (e.g., Collins & Cockburn, 2020; Daw et al., 2011; Dayan, 2009; Friedel et al., 2014; Perez & Dickinson, 2023; Vandaele & Janak, 2018; Wood & Rünger, 2016), it was observed that both goal-directed control and model-based planning involve sensitivity to the contingencies between stimuli, actions, and outcomes, as well as to changes of outcome values (Drummond & Niv, 2020). Model-free planning and habitual control, on the other side, share the disconnectedness of response values from the potential volatility of future state values (Drummond & Niv, 2020). In addition to these conceptual similarities, there is some evidence that model-free control is positively correlated with the strength of habitual responses (Friedel et al., 2014). Thus, although model-free and model-based reinforcement learning processes are not congruent to habitual and goal-directed action control processes, the concepts resemble each other in several regards. Therefore, results from both frameworks are taken into consideration in this thesis.

Reinforcement learning models have been described as successors of the associative dual-process models, with enhanced transparency of the interpretation of the experimental results and more precise hypotheses about the interplay between the two processes (Dolan & Dayan, 2013), while its parameters can well be analyzed in conjunction with psychophysiological data (Akam et al., 2015). Model-free and model-based reinforcement learning are often experimentally operationalized with two-step sequential decision tasks, which usually feature two stages in which participants can select one of two visual stimuli. Depending on their choice in the first stage, the participants are presented with different second stages that again feature two visual stimuli and the task is once more to choose one of them (e.g., Daw et al., 2011). The transitions between the two stages and the stimuli's values, which are usually points signaling monetary rewards earned in the task, vary over time (Cushman & Morris, 2015). The strength to which the participants use information on the transitions and state values in their choices is then modeled. However, similar to outcome devaluation tasks, the validity of two-step sequential tasks has been appraised critically: if participants apply a model-based strategy that does not align with the expected model-based strategy, they may be misclassified as following a model-free or hybrid strategy (Da Feher Silva & Hare, 2018; Feher da Silva & Hare, 2020). In addition, minor

modifications to the task structure, such as the number of trials (Akam et al., 2015), the specific instructions (Feher da Silva & Hare, 2020), and data modeling choices (Toyama et al., 2019) can significantly change the obtained parameters. Thus, the disadvantages of the outcome devaluation tasks are, in several ways, mirrored in sequential decision tasks.

1.1.5 Action control models in clinical research

The associative dual-process model has informed clinical research aiming to explain why dysfunctional behaviors persist despite creating considerable distress and long-term harm (Huys et al., 2015; Voon et al., 2015; Voon et al., 2017). Habitual, outcome-insensitive control is hypothesized to contribute to the persistence of these behaviors by reducing the ability of the individual to inhibit these behaviors (e.g., Huys et al., 2015). The definition of habitual responses as behavior that is decoupled from current goals provides a rationale with high face validity for assuming a potential role of habits in the etiology of psychological disorders characterized by maladaptive, repetitive responses (LeDoux & Daw, 2018; LeDoux et al., 2017). Of note, persistent maladaptive behaviors can also be explained with goal-directed processes, for example, by a competition between a highly valued, but maladaptive short-term goal with a weaker valued, long-term adaptive goal (Voon et al., 2017). For instance, an individual may value the short-term fear reduction resulting from avoiding a feared stimulus more highly than the potential rewards that may result from approaching the stimulus. However, habitual control may be one additional mechanism to explain behaviors unaligned with explicit goals.

An overreliance on habits at the expense of goal-directed control has been demonstrated in various psychological and neurological disorders, such as Tourette syndrome (Delorme et al., 2016; Scholl et al., 2022), Autism spectrum disorder (Alvares et al., 2014; Alvares et al., 2016), Bulimia nervosa (Berner et al., 2023), and Anorexia nervosa (Uniacke et al., 2018). Two research areas with a pronounced emphasis on habits are research on dependency disorders and obsessive-compulsive disorder. Habitual drug use is a component in current models of drug dependency (e.g., Everitt & Robbins, 2016; Wise & Koob, 2014, but see the critical perspectives by Buabang, Köster, et al., 2023; Field & Kersbergen, 2020). An increased shift towards habitual control has also been reported in individuals with obsessive-compulsive disorder when compared to healthy controls and was interpreted to underlie compulsive behaviors (Gillan et al., 2015; Gillan et al., 2016; Gillan et al., 2014; Gillan et al., 2011; Gillan & Robbins, 2014; Verhoeven & Wit, 2018). Transdiagnostically, the Research Domain Criteria, a taxonomy for psychiatric research that focuses on processes instead of symptoms, includes the habit construct twice, first, under the

domain of sensorimotor systems, and, second, as a subconstruct of reward learning under the domain of positive valence systems (National Institute of Mental Health, 2023).

The observable association between stress and maladaptive behavioral symptoms may be mediated by a shift from goal-directed towards habitual responses (e.g., Schwabe & Wirz, 2013; Huys et al., 2014). From a diathesis-stress perspective, stress-related behavioral adjustments are central in buffering or exacerbating the negative impact of stress on mental health (Folkman, 2013). Several studies have demonstrated elevated habitual and model-free control following acute stress before outcome devaluation tasks or sequential tasks (Quaedflieg et al., 2019; Schwabe & Wolf, 2010). Elevated habitual responses have also been demonstrated when stressful events happened during the task (i.e., after the outcome devaluation procedure; Schwabe & Wolf, 2010). Chronic stress (Pool et al., 2022; Radenbach et al., 2015) and early-life stress (Patterson et al., 2019) have also been associated with more pronounced habitual control.

One explanation for the association between stress and habitual control is a shift from a hippocampus-dependent to a dorsal striatum-dependent memory system following a stress-related release of glucocorticoids (e.g., Schwabe & Wolf, 2013; Wirz et al., 2018). In individuals with low working memory capacity, acute stress predicted a shift towards habitual control more strongly than in individuals with high working memory capacity (Otto, Gershman, et al., 2013; Quaedflieg et al., 2019). These results suggest that working memory capacity moderates the association between stress and habitual control and implicate that executive functioning is involved in the observed stress-related shift towards habitual control. However, the general effect of acute stress on habitual responses was not replicated in recent studies, leading to doubts on the internal validity of the measurement (Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023). Potentially, participants used goal-directed strategies to regulate stress in the original studies, which produced a non-adjustment of responses in the outcome devaluation task and may have led to false positive classifications of such goal-directed responses as habitual responses (Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023). Future studies on the effects of stress on habitual control may consider such stress-induced goal-directed processes. Nevertheless, some evidence indicates a shift towards habitual control as a function of stress, suggesting that shifts from habitual to goal-directed responses may be clinically relevant.

1.2 Anxiety disorders, avoidance, and fear (reduction)

Fear and anxiety have been defined and measured in different ways (e.g., Endler & Kocovski, 2001; McNaughton, 2018). They can, for example, be separated depending on their temporal characteristics. Proximate, short-term reactions to acute threats are commonly described as *fear* (e.g., Mobbs et al., 2009). The term *anxiety* instead refers to prolonged reactions to more distal, anticipated threat encounters, i.e., states in which an individual experiences a form of uncertainty where potential threats are apparent, but their probability and exact dangerousness are unknown (i.e., Mobbs et al., 2009; Remmers & Zander, 2018). In contrast, the construct *trait anxiety* refers to an enduring tendency to experience anxious episodes frequently (Elwood et al., 2012). Trait anxiety, as a discernable pattern of cognitive, behavioral, physiological, and neural responses to threat, has been discussed as a risk factor and precursor of anxiety disorders (Knowles & Olatunji, 2020).

1.2.1 Definitions of anxiety disorders and maladaptive avoidance

Anxiety disorders are a group of disorders characterized by frequent, high levels of fear, anxiety, or avoidance that are disproportionate to the objective level of threat (American Psychiatric Association, 2013; Craske et al., 2017). Anxiety disorders are common mental disorders, with one-year prevalence estimates between 10% and 20% and lifetime prevalence estimates between 15% and 34% (Bandelow & Michaelis, 2015; Kasper, 2006; Somers et al., 2006). An estimated 25 million individuals in the European Union were affected in 2018 (OECD, 2018). Anxiety disorders can be associated with substantial reductions in quality of life (Olatunji et al., 2007) and are frequently comorbid with a range of other mental disorders, e.g., depressive disorders (Brown et al., 2001; Kaufman & Charney, 2000) and substance use disorders (Marmorstein, 2012). Additionally, anxiety disorders frequently follow a chronic course. For example, one study reported that nearly one in four individuals with a remitted anxiety disorder at baseline measurement fulfilled the criteria for at least one anxiety disorder two years later (Scholten et al., 2013). Other studies reported even higher relapse rates. For example, in a prospective study by Bruce and colleagues (2005), 39% of individuals who had recovered from social anxiety disorder relapsed within 12 years. For individuals with remitted generalized anxiety disorder, the recurrence rate was 45%, and 58% for individuals with panic disorder with agoraphobia (Bruce et al., 2005). On an individual functional level, anxiety disorders were associated with lower income and lower educational and professional success (e.g., Kasper, 2006; McCurdy et al., 2022; Tolman et al., 2009). The economic impact of anxiety disorders in

the European Union was estimated to be approximately 40 billion € in 2004, which was comparable to the economic impact of dementia (Andlin-Sobocki et al., 2005). Although psychotherapy for anxiety disorders is effective in reducing symptoms one year after treatment (Keefe et al., 2014; Olatunji et al., 2010; van Dis et al., 2020), the relapse rates remain relatively high, with reported relapse rates of up to 42% within two years after treatment (Lorimer et al., 2021).

In addition to the symptoms of intense and frequent episodes of fear and anxiety, avoidance is a central symptom of anxiety disorders, specifically of specific phobias, agoraphobia, and social anxiety disorder (American Psychiatric Association, 2013). Fear and anxiety-related avoidance can range from avoiding specific objects or activities to avoiding entire classes of activities or objects, and from relatively simple responses to complex patterns of behaviors (Craske et al., 2017). Avoiding threatening stimuli is not maladaptive per se: identifying and avoiding threat-predicting stimuli is an important psychological function that reduces the frequency of dangerous encounters and, thus, supports the survival of the organism (e.g., Öhman & Mineka, 2001). Therefore, avoidance can be considered adaptive when the perception of the threat is realistic and the effort or costs of avoidance are justified by the threat. Avoidance can, however, become maladaptive when it is excessive and disproportionate to the actual threat posed by the avoided stimulus (Dymond, 2019) or when it is performed regardless of disproportionately high associated costs (i.e., Arnaudova et al., 2017; Aupperle & Paulus, 2010). Thus, avoidance can become maladaptive when it persistently restricts activities and functions or is performed in the face of objectively harmless stimuli that do not pose an objective threat to the individual's well-being (Arnaudova et al., 2017). Such maladaptive features can explain why avoidance is associated with functional impairments and reduced quality of life (Hendriks et al., 2016; Mendlowicz & Stein, 2000; Wilmer et al., 2021). For example, avoiding burglary by locking one's apartment door will not be associated with considerable costs and can be considered an adaptive and beneficial behavior. However, avoiding unpleasant social encounters by averting social gatherings, for example, will likely be associated with significant costs since it may result in reduced positive social interactions or conflicts with friends or family members. Because excessive avoidance can come with a reduced number of encounters with pleasant and rewarding outcomes, avoidance has also been discussed to promote depressive symptoms via a reduction of rewards (Trew, 2011). In line with this, avoidance has been demonstrated to mediate the relationship between anxiety and depression symptoms, suggesting that avoidance is transdiagnostically relevant (Carvalho & Hopko, 2011; Jacobson & Newman, 2014).

Approach-avoidance theories have formulated the behavioral and neural processes involved when animals or humans make decisions in situations where appetitive and aversive outcomes compete (for reviews, see Corr, 2013; Kirlic et al., 2017; Loijen et al., 2020; Talmi & Pine, 2012). Avoidance is frequently associated with sacrificing positive outcomes in naturalistic environments, and a flexible trade-off between risks and rewards can be considered adaptive (Kirlic et al., 2017). In experiments on the regulation of such costly avoidance, participants can typically choose to avoid a threatening outcome; however, the avoidance is associated with a monetary loss (e.g., Bublatzky et al., 2017; Pittig, Brand, et al., 2014; Pittig, Schulz, et al., 2014. Generally, healthy individuals seemed to show a reduction of avoidance behaviors when avoidance was directly paired with competing monetary rewards, pointing towards a flexible downregulation of avoidance in healthy individuals (Pittig, 2019; Pittig, Boschet, et al., 2021; Pittig & Dehler, 2019; Pittig, Hengen, et al., 2018; Pittig & Scherbaum, 2020). However, in anxiety disorders, the experimental studies suggest a deficit in avoidance inhibition in the face of competing rewarding outcomes. Individuals with anxiety disorders adjusted avoidance less to approach competing rewards and more frequently avoided when no aversive outcomes were present anymore as compared to individuals without anxiety disorders (Pittig, Boschet, et al., 2021). Similarly, individuals with high trait anxiety showed a stronger tendency to avoid a threatening outcome in the face of high competing rewards compared to less anxious individuals (Pittig & Scherbaum, 2020). In the same study, however, low-cost avoidance (i.e., avoidance with low competing rewards) did not differ between low and highly anxious individuals (Pittig & Scherbaum, 2020). Therefore, anxious psychopathology may not per se be associated with excessive avoidance, but more specifically with dysregulated avoidance that is not adjusted flexibly in the face of competing rewards.

Due to the manifold interactions between fear, anxiety, and avoidance, the reduction of maladaptive avoidance is one of the central targets in psychotherapy for anxiety disorders (i.e., Arnaudova et al., 2017; Craske et al., 2014); a target that is not easily met, since maladaptive avoidance can be temporally stable and persistent even in the absence of perceived threat, as has already been explained. Amongst other effects of avoidance, pronounced, inflexible avoidance behavior can impede the progress of therapies for anxiety disorders, too, especially when interventions include exposure-based elements that rely on the ability of the individual to approach feared stimuli (Mesri et al., 2017; Porter & Chambless, 2015; Telch et al., 1995). Thus, even though the processes that maintain dysregulated costly avoidance behavior are not

completely understood, they seem to be relevant for the development of novel psychotherapy approaches targeted at reducing avoidance to treat anxiety disorders.

1.2.2 The two-factor model of avoidance

The mutual influences between fear, anxiety, and avoidance are commonly emphasized in the literature (Pittig et al., 2020). An influential model that has guided avoidance research proposed two distinguishable learning processes in the acquisition of avoidance, assuming an intricate connection between the learning of fear and avoidance (Mowrer, 1951; Mowrer & Lamoreaux, 1946; Rescorla & Solomon, 1967; for a review, see Krypotos et al., 2015). The model states that avoidance is acquired in two learning stages. First, a previously neutral stimulus becomes associated with an aversive unconditioned stimulus (US), such as a loud noise or a painful stimulation. The neutral stimulus thus becomes a conditioned stimulus (CS+) that elicits fear as a conditioned response. Second, upon this Pavlovian fear acquisition, instrumental avoidance can develop when allowed to prevent encountering the CS+, stopping the CS+ presentation and thereby preventing the US occurrence, or using the CS+ as a warning signal to flee, consequently avoiding encountering the aversive US. Of note, avoidance behaviors in the strictly defined sense describe the avoidance of the CS+ and are contrasted with safety behaviors performed to avoid the US without CS+ avoidance and escape behaviors performed to terminate the CS+ presentation (see Krypotos et al., 2015). A naturalistic example of avoidance in social anxiety disorder may be an individual who avoids meeting other people, an example of a safety behavior may be carrying anxiolytic medication, and an example of escape behavior may be leaving social interactions early (see Pittig et al., 2020). The instrumental process in the second stage of the two-factor model can be understood as negative reinforcement learning (i.e., the avoidance is reinforced by the omission of the aversive US; see Hofmann & Hay, 2018). Thus, the fear reduction associated with the avoidance response was assumed to increase the likelihood of future performances of the according behavior. This instrumental training phase conceptualization is in line with drive reduction theories (e.g., Hull, 1943; see Krypotos et al., 2015) since goal-directed actions are assumed to result from the need to regulate internal homeostasis. In the case of fear, this need is to reduce fear (see Krypotos et al., 2015). However, more recent models understand instrumental avoidance as being positively reinforced by further processes such as the relief following the avoidance (e.g., Perez & Dickinson, 2023; Pittig et al., 2020).

One advantage of the two-factor theory was that the assumed processes could be operationalized directly in controlled experimental settings. For example, in one of the early experimental designs, the so-called *shuttle box paradigm*, dogs were presented with a neutral stimulus (e.g., a light), which preceded an aversive US (i.e., an electric shock), to learn that the neutral stimulus predicted the aversive US (e.g., Solomon et al., 1953). Thereby, the light became a CS+, which signaled a threat. After this classical fear conditioning procedure, the animals learned to prevent the occurrence of the aversive US by shuttling to a specific cage area during the CS+ presentation. This design provided the foundation for experimental research on the learning mechanism involved in avoidance in animals, and it was also used translationally in research on avoidance learning in humans (Krypotos et al., 2015).

1.2.3 Shortcomings of the two-factor theory of avoidance

Despite the advantage of the two-factor theory in terms of the feasible experimental operationalization and its considerable explanatory value, it became apparent that avoidance behavior demanded a more complex conceptualization. One critical finding that the model could not explain was that fear reduction did not lead to reduced avoidance under extinction schedules where no aversive USs were presented anymore even if the avoidance response was not performed (i.e., under an extinction schedule; see Krypotos et al., 2015). Such a decoupling of threat and avoidance was observed in animals (e.g., Seligman & Campbell, 1965; Solomon et al., 1953) and humans (e.g., Levis & Boyd, 1979; Malloy & Levis, 1988). Under extinction, the two-factor model would predict that conditioned fear of the CS+ recedes after repeated pairings of the avoidance response with the CS+. Due to the avoidance, the CS+ is not followed by the aversive US anymore and would lose its threat-predictive value. Consequently, the CS+ termination resulting from the avoidance response should produce less relief, and fear and the frequency of avoidance of the CS+ should decrease (Dinsmoor, 1954; Krypotos et al., 2015; Maia, 2010). However, accumulated evidence indicates that fear reduction and US omission do not always lead to a reduction of avoidance, which has been identified early as a significant inadequacy of the two-factor theory (e.g., Pittig & Scherbaum, 2020; Rachman, 1976). The two-factor theory could not explain why the empirical findings did not align with this prediction, suggesting that mechanisms other than the reinforcement by fear reduction were involved in the regulation of avoidance.

In addition to the problem of avoidance continuation under extinction schedules, recent human studies demonstrated a continuation of avoidance even after successful fear extinction,

which is also not in line with the presumed causal role of fear for avoidance in the two-factor theory (van Uijen et al., 2018; Vervliet & Indekeu, 2015; Xia et al., 2017). Theoretically, such extinction-resistant avoidance may result from incomplete fear extinction procedures since avoidance can hamper opportunities to learn that the CS+ is not followed by the US anymore (e.g., Krypotos et al., 2018). However, extinction-resistant avoidance was also reported after extinction with response prevention, when participants were forced to experience the CS+ without it being followed by the US (e.g., Vervliet & Indekeu, 2015). Thus, the non-occurrence of the US in association with the CS+ was actively experienced, and complete fear extinction can be assumed. When the participants were, however, allowed to perform the avoidance response again, avoidance re-emerged (Vervliet & Indekeu, 2015). One way to explain such observed extinction-resistant avoidance in experimental settings is by understanding the re-availability of avoidance as a context change (Engelhard et al., 2015; Vervliet & Indekeu, 2015). In this regard, new context elements after fear extinction – e.g., the re-availability of the avoidance behavior in the paradigm – can elicit an increase in fear and threat-related expectations (Lonsdorf et al., 2017). Seemingly extinction-resistant avoidance may, thus, be motivated by a return of threat expectancy elicited by the re-availability of the avoidance response in the perceived new context (Engelhard et al., 2015; Vervliet & Indekeu, 2015). Thus, despite a need to investigate these findings further, overall, the empirical evidence supports a role of fear reduction in acquiring avoidance and less so in maintaining avoidance, suggesting that conditioned fear of the CS+ seems to extinguish more readily than the respective acquired instrumental avoidance responses (see Arnaudova et al., 2017; Hofmann & Hay, 2018; Krypotos et al., 2015; LeDoux et al., 2017).

Given the inability to explain the maintenance of persistent avoidance with the two-factor theory, several mechanisms were suggested in addition to the initial fear reduction proposed by the theory (e.g., Pittig et al., 2020). Cognitive models have highlighted the role of explicit threat expectations that rely on inferences about the likely outcomes of avoidance and non-avoidance (e.g., Lovibond et al., 2000; Seligman & Johnston, 1973), whereby the contingencies between actions and outcomes are assumed to be explicitly represented (Lovibond et al., 2008; Lovibond et al., 2009). For example, avoidance can preserve explicit threat expectations by reducing opportunities to experience that a feared CS+ is no longer associated with the aversive US under an extinction schedule (Lovibond et al., 2000; Lovibond et al., 2008; Lovibond et al., 2009). Although threat expectancy drops off short-term as an immediate result of the avoidance, the threat expectancy is not updated to account for the omission of the aversive US because no new association between the CS+ and safety can be formed (i.e., inhibitory learning; see Craske et al.,

2014) due to the avoidance. Avoidance, therefore, prevents fear extinction learning. Consequently, the preserved threat expectation and the corresponding fear can, again, motivate avoidance behavior when confronted with the CS+. The avoidance then, again, prevents fear extinction learning, creating a vicious cycle (see Seligman & Johnston, 1973). Additionally, avoidance in the presence of a CS+ (i.e., safety behavior) can block extinction learning since the association between CS+ and safety (i.e., no US occurrence) is then learned only conditionally of the avoidance response (Krypotos et al., 2015). As a result, performing the avoidance response in the face of the CS+ decreases the expectancy of a US occurrence when being confronted with the CS+, and the avoidance response, instead of the CS+, becomes a safety signal (Treanor & Barry, 2017). In contrast, an adaptive process would be to update US expectancies when a CS+ does no longer predict a threat (i.e., fear extinction; see Craske et al., 2014), and, thus, to adapt threat and safety expectations flexibly in changing environments.

Besides cognitive interventions to reduce explicit threat expectancies (e.g., Rief et al., 2022), the prevention of avoidance has also been discussed as one way to support updating threat expectancies (e.g., Baum, 1970). However, as already mentioned, when individuals are allowed to avoid again after response prevention, avoidance behavior can re-emerge even if the fear of the CS+ has been extinguished during the previous forced CS+ confrontation (Vervliet & Indekeu, 2015). Also, avoidance prevention is not always a practical option outside of controlled experimental settings, reducing its therapeutic potential. Experimental interventions to reduce avoidance that aim to modulate memory processes via psychopharmacological agents are still under investigation (Treanor & Barry, 2017).

Avoidance can, thus, impair the flexible updating of threat expectancies and, relatedly, the regulation of approach and avoidance that is necessary to adapt in naturalistic environments with complex and involve various option possibilities and ambiguous stimuli that can predict rewards and threats simultaneously (Aupperle et al., 2023). It has been suggested that avoidance studies may benefit from incorporating approach-avoidance decisions, since experimental designs that restrict behavioral choices to avoidance while precluding any competing approach may lack external validity for researching maladaptive avoidance (see Krypotos et al., 2018). Costly avoidance (i.e., avoiding while sacrificing competing rewards) may thus be a more externally valid operationalization of maladaptive avoidance that is disproportionate to the avoided threat than low-cost avoidance (i.e., avoiding without a loss of rewards; Krypotos et al., 2018). Thus, including competitions between aversive and rewarding outcomes and between avoidance and

approach responses in avoidance studies has been discussed to enable a more complex and, therefore, more naturalistic perspective on the processes that maintain maladaptive avoidance (Krypotos et al., 2018; Pittig et al., 2020). The benefit of including costs for avoidance in experimental studies is corroborated by the already mentioned evidence on a stronger association of anxious psychopathology with costly than low costly or non-costly avoidance (Pittig, Boschet, et al., 2021; Pittig & Scherbaum, 2020).

1.3 An action control perspective on avoidance in anxiety disorders

The avoidance models described in the previous paragraphs explicitly or implicitly assumed that goals underlie avoidance behavior. For instance, the two-factor theory of avoidance assumes that fear reduction drives avoidance. Although the reinforcement of avoidance by fear reduction does not necessarily require explicit representations of goals, outcomes, or action-outcome contingencies, the assumed reinforcement by fear reduction implicates a close relation between avoidance responses and their outcomes. The concept of avoidance acquisition in operant conditioning models does not necessarily involve the assumption of a deliberate planning process. However, reinforcement is assumed to depend on response outcomes, implying sensitivity of the reinforced behaviors to changes in outcome values or response-outcome contingencies (Bouton, 2018). Cognitive theories of avoidance, in contrast, assume that avoidance is guided by goals that are connected to threat expectations (e.g., Lovibond et al., 2000; Lovibond et al. 2009). Thus, despite representing different conceptualizations, instrumental learning models and cognitive models of avoidance share the assumption that avoidance depends on goals. Maladaptive avoidance is, thus, conceptualized as the result of biased goal-directed processes (Vandaele & Janak, 2018). However, as has already been discussed, habitual processes are being increasingly discussed as potential contributors to maladaptive responses in general (e.g., Huys et al., 2015; Schwabe & Wolf, 2009, 2013), and evidence on increased tendencies to acquire habitual responses has been demonstrated in a range of disorders such as obsessive-compulsive disorder (e.g., Gillan et al., 2015; Gillan et al., 2016; Verhoeven & Wit, 2018), and substance use disorders (e.g., Everitt & Robbins, 2016; Wise & Koob, 2014).

Several theoretical accounts on avoidance have proposed habitual avoidance as one explanation of the persistence of maladaptive avoidance that has been so difficult to explain in the two-factor model and cognitive models (Arnaudova et al., 2017; Cain, 2019; Hofmann & Hay, 2018; Ilango et al., 2014; LeDoux & Daw, 2018; LeDoux et al., 2017; Pittig et al., 2015; Pittig, Treanor, et al., 2018; Pittig et al., 2020). Some of these proposals have explicitly suggested

a three-factor avoidance learning model by adding a habitual stage to the first two learning processes postulated in the two-factor theory (i.e., fear acquisition and instrumental avoidance acquisition). The proposals were often based on the evidence of distinguishable neural circuits associated with goal-directed and habitual control in rodents (Cain, 2019; Ilango et al., 2014; LeDoux & Daw, 2018; LeDoux et al., 2017). Most prominently, it was suggested that, over time, goal-directed avoidance in anxiety disorders would be replaced by habitual avoidance as a function of avoidance repetition. The two first stages of the two-factor model would then be followed by a third stage in which goal-directed avoidance would be transformed into habitual avoidance that is performed without threat expectations or S-R-O contingency learning. In this regard, habitual avoidance was presented as a solution to the theoretical problems with the two-factor theory, especially concerning the explanation of persistent avoidance under extinction schedules (Hofmann & Hay, 2018; LeDoux & Daw, 2018; LeDoux et al., 2017). The transition from goal-directed, fear-related avoidance to habitual, outcome-insensitive avoidance may result simply due to frequent repetitions of avoidance responses in anxiety disorders (Ilango et al., 2014; LeDoux & Daw, 2018; LeDoux et al., 2017). There is currently, to my best knowledge, no evidence of a longitudinal increase of habitual avoidance during the development and maintenance of anxiety disorders. However, a rapid development of habitual avoidance (i.e., an increased shift from goal-directed to habitual avoidance) may also be a specific risk factor in the etiology of anxiety disorders. Individuals with a more pronounced tendency toward developing habitual avoidance may be at higher risk for avoiding inflexibly. Due to the negative consequences of avoidance on fear preservation and general functioning, such a shift towards a more pronounced acquisition of habitual avoidance may facilitate the development of an anxiety disorder. Individual characteristics that are associated with a stronger shift towards habitual avoidance may then put individuals at risk for developing anxiety disorders via a strengthened formation of inflexible avoidance (Arnaudova et al., 2017; Pittig et al., 2020).

Several studies have investigated the impact of trait anxiety and clinical anxiety on action control, but they did not produce a conclusive pattern of results. In two studies by Alvarez et al. (2014, 2016), trait anxiety was associated with stronger habitual approach. However, in three experiments by Gillan et al. (2021), state anxiety was not associated with model-free reinforcement learning. There is, currently, no study on the distinct effect of trait anxiety on the acquisition of habitual avoidance. Studies with clinically anxious individuals are also rare and inconclusive: elevated habitual approach was found in individuals with social anxiety disorder compared to healthy control participants, and within the clinical group, stronger habitual control

was associated with less symptom improvement in psychotherapy (Alvares et al., 2014). However, no elevated habitual avoidance was found in generalized anxiety disorder (Roberts et al., 2022). Thus, the evidence on the effects of trait anxiety or anxiety disorders on the acquisition of habitual avoidance is currently unclear.

Habitual responses may be favored in individuals with high trait anxiety or anxiety disorders due to anxiety-related effects on attention and perception. A robust literature indicates that anxious states are associated with faster detection of and orientation toward threat-related stimuli and less effective disengagement of attention away from threat-related stimuli (for reviews, see Bar-Haim et al., 2007; Sussman et al., 2016). Such threat-related alterations of attentional processes in anxious states are expected to influence more complex, cognitive forms of information processing, such as the formation of expectations related to safety and threat (see Mogg & Bradley, 1998). Besides the associations of anxiety with attentional biases, the control of attentional processes has also been proposed to be affected by anxiety and, specifically, trait anxiety. Attentional control refers to several attention-related functions related to processing efficiency, such as flexible shifts of attention, the suppression of task-irrelevant information, and updates of the information in working memory (e.g., Berggren & Derakshan, 2013). One hypothesis is that trait anxiety impairs the efficiency of these attentional control processes. In other words, trait anxiety may interfere with the flexible adjustment of attention, which is necessary to represent changing environments (e.g., Berggren & Derakshan, 2013). Anxiety, in this concept, fosters a broader, scanning-like distribution of attention with a higher sensitivity to threat-related stimuli in the environment and a lower impact on goal-directed processes (Berggren & Derakshan, 2013; Eysenck & Calvo, 1992; Eysenck & Derakshan, 2011). Potentially, a shift towards inflexible, more stimulus-driven attentional control resulting from inefficient regulation of cognitive control may also contribute to inflexible, stimulus-driven habitual responses. Put differently, since goal-directed control depends on flexible, precise representations of the environment to predict outcomes of the available behavioral options, goal-directed processes may also be diminished when cognitive control efficiency is reduced. However, no empirical evidence on this potential mechanism underlying effects of trait anxiety on habitual control is available so far, and an increased acquisition of habitual avoidance is currently an interesting but completely speculative potential mediator between trait anxiety and anxiety disorders.

When the flexible tracking of S-R-O contingencies is impeded because of high task difficulty or high working memory load, habitual responses may become more pronounced. Task difficulty and working memory capacity have been demonstrated to moderate the relationship between trait anxiety and habitual control (Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013). This working memory-dependent effect of trait anxiety may be explained by the capability of highly trait-anxious individuals to compensate for attentional control deficits by investing more cognitive effort under low cognitive load (Berggren & Derakshan, 2013; Eysenck & Derakshan, 2011). The effectiveness of such efficiency deficit compensation may be limited under high cognitive load, producing performance deficits in highly trait-anxious individuals under high-load conditions only (see Berggren & Derakshan, 2013; Eysenck & Derakshan, 2011). Interestingly, the role of memory processes in moderating the association between trait anxiety and habitual responses aligns with the assumption that habitual responses are functionally explained by limited cognitive resources (i.e., Moors & de Houwer, 2006). If trait anxiety induces a load on cognitive resources (Berggren & Derakshan, 2013; Eysenck & Derakshan, 2011), high trait anxiety may potentially also favor a shift from cognitively complex goal-directed control to cognitively more simple habitual control. Thus, many questions on the effects and boundary conditions of trait anxiety on habitual control have not been solved yet. However, there is a theoretical basis for hypothesizing that trait anxiety may be associated with habitual action control. Further elucidating the effects of trait anxiety and clinical anxiety on habitual control processes may allow us to better understand aberrations of avoidance regulation in individuals at risk for anxiety disorders.

1.4 Objectives

Maladaptive avoidance is a central symptom in anxiety disorders that contributes to the maintenance of fear and anxiety, and can hamper psychotherapy progress and create significant functional impairment. The emotional and cognitive mechanisms involved in the persistence of goal-directed maladaptive avoidance are relatively well understood, but the role of a shift from goal-directed to habitual avoidance remains debated. Theoretical accounts suggesting habitual avoidance in trait anxiety and anxiety disorders have yet to be underpinned by robust empirical evidence. The current dissertation aims to address this research gap by experimentally examining the impact of trait anxiety and anxiety disorders on the acquisition of habitual avoidance.

Study 1 aimed to develop a variant of a commonly used experimental paradigm in action control research, the outcome devaluation paradigm. The most important variation was the

inclusion of costs for habitual responses to address three main critique points that may threaten the internal validity of the outcome devaluation paradigm: First, the task variation should allow inferring habitual responses without relying on null effects. Specifically, by including costs for habitual responses, different experimental conditions could be created in the task that allowed to evade such null difference testing. Second, the inclusion of costs should rule out the performance of habitual responses as an advantageous option for participants, and therefore, reduce the risk of seemingly habitual responses that are the result of goal-directed strategies, for example, to reduce cognitive effort. Third, costs for habitual avoidance were included to support the external validity of the paradigm for researching maladaptive persistent avoidance. Therefore, the paradigm established in Study 1 should support a more internally and externally valid operationalization of maladaptive habitual avoidance that could be used in the subsequent studies to investigate the effect of trait anxiety and anxiety disorders on the acquisition of habitual avoidance.

In Study 2, we intended to examine the impact of anxiety disorders on the acquisition of habitual avoidance. Maladaptive avoidance is a central symptom of most anxiety disorders. Theoretical models have recently proposed that a more pronounced acquisition of habitual avoidance in individuals with anxiety disorders may explain the frequent persistence and inflexibility of maladaptive avoidance in these disorders. Additionally, many studies indicate that a range of disorders featuring maladaptive inflexible behaviors as symptoms (e.g., addiction disorders) are associated with elevated habitual response acquisition. However, empirical evidence on the acquisition of habitual avoidance in individuals with anxiety disorders is scarce. Therefore, we aimed to investigate whether the acquisition of habitual avoidance was more pronounced in individuals with anxiety disorders than in healthy, age- and gender-matched, control participants.

Study 3 aimed to compare the effects of trait anxiety – a known risk factor for anxiety disorders – on the acquisition of approach and avoidance habits. The available evidence for a specific tendency to shift from goal-directed to habitual avoidance in individuals with higher trait anxiety is ambiguous. Higher trait anxiety may put individuals at risk for specifically developing habitual avoidance. However, trait anxiety may also predict the acquisition of approach and avoidance habits unspecifically, or predict neither approach nor avoidance habit acquisition. Differentiating the impact of trait anxiety on habitual approach and avoidance acquisition may help to develop more precise models on action control aberrations in highly trait-anxious

individuals in the future. Additionally, if trait anxiety predicted the acquisition of habitual avoidance specifically, this would suggest a role of habitual avoidance acquisition in the etiology of anxiety disorders. Such evidence may, thus, build a ground for deriving more specific hypotheses on the role of habitual avoidance in the etiology of anxiety disorders. To gain first evidence on these potential associations, we compared the influence of trait anxiety on the acquisition of habitual approach and habitual avoidance in a within-subjects design using two parallel versions of the outcome devaluation task from Study 1. We included heart rate variability as a psychophysiological indicator of trait anxiety.

2. Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm

Valentina M. Glück[1], Katharina Zwosta[2], Uta Wolfensteller[2], Hannes Ruge[2] & Andre Pittig[1,3*]

[1]Department of Psychology (Biological Psychology, Clinical Psychology, and Psychotherapy), University of Würzburg, Germany

[2]Department of Psychology, Technical University Dresden, Germany

[3]Translational Psychotherapy, Department of Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

*Corresponding author: Andre Pittig, Georg-August-University of Goettingen, Translational Psychotherapy, Kurze-Geismar-Str. 1, 37073 Goettingen, Germany. Email: andre.pittig@uni-goettingen.de

Published as:

Glück, V. M., Zwosta, K., Wolfensteller, U., Ruge, H., & Pittig, A. (2021). Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm. *Behaviour Research and Therapy*, *146*, 103964. https://doi.org/10.1016/j.brat.2021.103964

Abstract

Avoidance habits potentially contribute to maintaining maladaptive, costly avoidance behaviors that persist in the absence of threat. However, experimental evidence about costly habitual avoidance is scarce. In two experiments, we tested whether extensively trained avoidance impairs the subsequent goal-directed approach of rewards. Healthy participants were extensively trained to avoid an aversive outcome by performing simple responses to distinct full-screen color stimuli. After the subsequent devaluation of the aversive outcome, participants received monetary rewards for correct responses to neutral object pictures, which were presented on top of the same full-screen colors. These approach responses were either compatible or incompatible with habitual avoidance responses. Notably, the full-screen colors were not relevant to inform approach responses. In Experiment 1, participants were not instructed about post-devaluation stimulus-response-reward contingencies. Accuracy was lower in habit-incompatible than in habit-compatible trials, indicating costly avoidance, whereas reaction times did not differ. In Experiment 2, contingencies were explicitly instructed. Accuracy differences disappeared, but reaction times were slower in habit-incompatible than in habit-compatible trials, indicating low-cost habitual avoidance tendencies. These findings suggest a small but consistent impact of habitual avoidance tendencies on subsequent goal-directed approach. Costly habitual responding could, however, be inhibited when competing goal-directed approach was easily realizable.

Key words: Costly avoidance, Habit, Goal-directed behavior, Devaluation paradigm

2.1 Introduction

Learning to avoid threatening stimuli and situations is crucial for organisms' survival and well-being as it prevents harm and danger. Goal-directed avoidance enables flexible behavioral adaptations to ever-changing environments. However, persistent, inflexible, and intense avoidance in the absence of threat constitutes a core symptom of anxiety disorders (American Psychiatric Association, 2013) and results in functional impairments and quality of life reductions (e.g., Hendriks et al., 2016; Pittig, Brand, et al., 2014). Moreover, inflexible avoidance can impede learning that a formerly threat-predicting stimulus or situation is now safe and may be approached without harm (Lovibond et al., 2009; Pittig, 2019). Inflexible avoidance therefore plays an important role in maintaining anxiety disorders (e.g., Krypotos et al., 2018; Pittig, Treanor, et al., 2018; Pittig et al., 2020). As avoidance can persist in the absence of threat and even in the absence of fear (i.e., after successful extinction of fear; Vervliet & Indekeu, 2015; Xia et al., 2017), it has been suggested that avoidance is not exclusively maintained and reinforced by fear reduction (Krypotos et al., 2015; Pittig et al., 2020). One potential further explanation for the maintenance of avoidance is that avoidance behavior may acquire habitual features over the course of individual learning histories (Arnaudova et al., 2017; LeDoux & Daw, 2018; LeDoux et al., 2017; Pittig et al., 2020).

A core characteristic of habitual behavior is its insensitivity to changes in response-outcome contingencies. Specifically, habitual behaviors are assumed to be insensitive to degraded response-outcome contingencies and reversed outcome values (Wood & Rünger, 2016). The insensitivity to outcome changes develops gradually through the extensive repetition of reinforced behavior (e.g., Adams & Dickinson, 1981, but see de Wit et al., 2018). In early learning stages, outcomes strongly modulate instrumental responses towards stimuli (S-R-O behavior). Through extensive repetition, outcomes guide action control less and less, while stimuli preceding the behavior gain more importance (S-R behavior, see Tricomi et al., 2009, for a review, see Balleine & O'Doherty, 2010). Habitual responses are adaptive in stable environments because they reduce the amount of cognitive control needed to maximize rewards and to minimize harm (e.g., concerning the encoding and retrieval of outcome values, see Dolan & Dayan, 2013). However, habitual behavior may impede the effective adaption of behavior when outcome values or response-outcome contingencies change. A potential over-reliance on habitual relative to goal-directed action control may thus help explain the persistence of maladaptive behaviors such as persistent maladaptive avoidance (e.g., Wood & Rünger, 2016).

The mechanisms involved in the reduction of maladaptive habits are important topics in habit research (Luque & Molinero, 2020).

Experimental evidence about habitual action control originates mainly from contingency degradation paradigms and outcome devaluation paradigms (Balleine & O'Doherty, 2010; Voon et al., 2017). In comparison with the wealth of paradigms available for the study of habitual approach, only a few paradigms examined habitual avoidance (see de Wit et al., 2018; Flores et al., 2018; Gillan et al., 2015; Gillan et al., 2014; Zwosta et al., 2018). Typical outcome devaluation paradigms measuring habitual avoidance implement Pavlovian fear conditioning, followed by instrumental avoidance acquisition. During Pavlovian fear acquisition, two formerly neutral conditioned stimuli (CSs+) are repeatedly paired with aversive unconditioned stimuli (USs; e.g., electrical stimulation or loud noise). Afterward, two instrumental avoidance responses preventing the aversive USs are extensively trained. Next, one of the USs is devalued by either minimizing its intensity (Flores et al., 2018) or eliminating the possibility of US deliverance (e.g., by removing the electrode delivering the stimulation; see de Wit et al., 2018; Gillan et al., 2014). In the final devaluation test, which is usually carried out in extinction, both extensively trained behavioral options are still available. Importantly, performing the instrumental response to the stimulus which had before predicted the now devalued aversive outcome is now unnecessary. The strength of habitual responding is then analyzed by comparing response rates to the CS that predicts the now devalued outcome with response rates to the CS that still predicts the valuable outcome. Habitual responding is assumed when the rate of responding to the former CS is *not lower* than to the latter CS. Habits are thus inferred from null effects (i.e., no difference between devalued and devalued stimuli, see de Houwer et al., 2018). Of note, devaluation studies often did not find completely outcome-insensitive responding. Stronger habitual responding has been inferred in a group of participants even when a devaluation effect was apparent (i.e., less responding to devalued than to valued stimuli), but was less pronounced than in another group (Gillan et al., 2014, 2015). Alternatively, individual response rate differences between devalued and valued stimuli have been associated with other psychological measures (Flores et al., 2018). These studies, thus, usually did not apply the criterion of complete outcome insensitivity and instead used the relative difference between responding to valued and devalued stimuli as a continuous measure of habit strength.

While devaluation paradigms have been widely adopted, their measurement specificity has been questioned (de Houwer et al., 2018; de Wit et al., 2018). First, the use of null effects to infer habitual behavior may lead to erroneous classifications of behaviors as habitual when, in

fact, the behaviors are driven by goals that are not captured by the experimental design (de Houwer et al., 2018), or by an aberrant goal-directed process (Balleine & Dezfouli, 2019). Second, in habitual avoidance paradigms, avoidance in the devaluation test is commonly operationalized as low-cost avoidance for which the sole cost is the motor action needed to press a button (see de Wit et al., 2018; Flores et al., 2018; Gillan et al., 2015; Gillan et al., 2014; Zwosta et al., 2018). While selective non-responding to devalued CSs may enable participants to obtain some goals (e.g., refraining from redundant responses), it also involves cognitive costs (Pezzulo et al., 2013; Shenhav et al., 2017). Participants might explicitly decide against investing the cognitive effort needed to adjust responding to changed outcome values, hence displaying a behavioral pattern resembling habitual avoidance. Additionally, selective non-responding introduces some risks in paradigms with real aversive outcomes. For example, participants may conceive the devaluation procedure as potentially ineffective or expect that aversive outcomes might still be administered by the experimenter (e.g., when participants are merely instructed about outcome changes). Such perceived potential risks may be especially problematic for the investigation of habitual avoidance in risk-aversive, anxious individuals who may decide to follow a "better safe than sorry" strategy (i.e., sticking to actions that have been safe in the past rather than exploring potentially more rewarding behaviors, see Schulz et al., 2016). Although resulting from an explicit, goal-directed cognitive strategy, such responding would produce behavioral patterns resembling habitual avoidance (i.e., continuing avoidance after devaluation). In conclusion, when devaluation paradigms involve low-cost avoidance behaviors only, various possible causes may explain rigid post-devaluation responding, habitual behavior being only one of them.

When investigating maladaptive habitual avoidance, critique concerning the validity of low-cost avoidance for researching maladaptive avoidance processes should also be considered (Krypotos et al., 2018; Pittig et al., 2020). One recent study demonstrated that trait anxiety is related to costly avoidance but not to low-cost avoidance (Pittig & Scherbaum, 2020), supporting the notion that specifically costly avoidance, and not avoidance per se, is associated with anxious psychopathology. Accordingly, costly avoidance paradigms have been suggested to increase the external validity of experimental avoidance research approaches (Krypotos et al., 2018). Habitual avoidance can be expected to persist in the absence of threat and even in the presence of concurring rewards due to its property of outcome insensitivity. However, manifest costs of habitual avoidance have not been implemented in the experimental literature so far. Based on the study of Zwosta et al. (2018), which tested the strength of habitual behavior

in direct competition with an assigned goal-directed behavior, we aimed to investigate whether avoidance habits impact competing approach responses, i.e., whether habitual avoidance to an aversive US persists despite competing incentives to not avoid.

To this end, we developed a modified habit-goal competition paradigm in which habitual avoidance is acquired and subsequently competes with the goal-directed approach of rewards, based on the study of Zwosta et al. (2018). This procedure was previously effective in detecting habitual tendencies after extensive training of approach or avoidance. Initially, avoidance of an aversive electrical US was extensively trained for habit acquisition. Next, the aversive US was devalued by removing the electrode delivering the US. In a subsequent habit-goal competition phase, participants were instructed to approach rewards, thereby implementing meaningful goals in the test phase (see Zwosta et al., 2018). Taken together, we aimed to examine a) habitual avoidance without relying on null effects and b) the degree of low-cost and costly habitual avoidance in the presence of rewards. Here, low-cost habitual avoidance was defined as habitual responding that did not result in monetary costs, indicated by either slower reaction times in habit-compatible than in habit-incompatible responses to approach rewards, or by a preference of habitual responding when all stimulus-related responses predicted the same outcome. Costly habitual avoidance was defined as habitual responses resulting in monetary costs, indicated by impaired accuracy in goal-directed, habit-incompatible approach of rewards. We hypothesized that, as a result of extensive avoidance training, habitual avoidance behavior would impair the competing goal-directed approach of rewards in both a costly and a low-cost way.

2.2 Experiment 1

Methods

Participants

The experiment was conducted in accordance with the latest update of the Declaration of Helsinki (WHO, 2001) and was approved by the local ethics committee. Participants were recruited via an online recruitment platform run by the Department of Psychology at the University of Würzburg. Exclusion criteria were age under 18 or over 55 years, any current self-reported psychological and/or psychiatric disorder (including substance abuse), cardiovascular or respiratory diseases, medical advice to avoid stressful situations, current psychopharmacological medication, central nervous system medication, and pregnancy.

Volunteers participated in return for course credit or 9 € compensation per hour and a monetary bonus dependent on their performance during the task. We estimated a sample size of $N = 55$ for Experiment 1 based on a previous study (Zwosta et al., 2018). Sixty-six participants took part in the experiment. Eleven participants had to be excluded from all analyses: Five due to technical failures, five due to deviations from the standard operation procedure, and one due to insufficient response accuracy during the first phase of the experiment (46 %), which was considered to preclude habit acquisition. The final sample consisted of 55 participants: 39 females (70.9 %), seven left-handed (12.7 %), and mean age 24.9 years ($SD = 6.9$, range: 18 – 51 years).

Procedure and materials

All experimental sessions were conducted with one single participant at a time. After having provided written informed consent and sociodemographic data, participants completed a standardized calibration procedure to adjust the individual intensity of the aversive unconditioned stimulus (US). Participants were instructed to choose a US level that was "unpleasant, but not painful", corresponding to a rating of "4" on a rating scale from "0" (*no sensation*) to "5" (*painful sensation*). The US was an electro-tactile stimulation consisting of 125 consecutive stimulations with a duration of 3 ms each and a temporal distance of 2 ms between consecutive stimulations (i.e., total duration 625 ms). The US was delivered using a bar electrode (diameter 8 mm, spacing 30 mm) that was attached to the participant's non-dominant forearm. USs were generated by a Digitimer DS7R stimulator (Digitimer Ltd). The mean US intensity was 0.7 mA ($SD = 0.4$). After the US calibration procedure, the habit-goal competition task was completed. Participants were seated in front of an HDMI monitor (resolution: 1080x1920 px, diameter: 24 inches) on which all instructions and the experimental paradigm were presented. Centrally in front of the monitor, a customary computer keyboard with two marked buttons (left and right Windows button) was positioned for participants to navigate through instructions as well as to respond during the experiment. The experiment was programmed and delivered, and data were recorded, with Presentation 18.1 (Neurobehavioral Systems, Berkeley, USA).

Habit-goal competition task

The paradigm consisted of two phases: 1) Extensive avoidance training, 2) competition phase (see Figure 1). Between extensive training and competition, the experiment was paused for an outcome devaluation procedure. Exploratorily, we added two further phases (i.e.,

revaluation and reinstatement phase) after the competition phase. These exploratory phases were not included in the main analysis of habitual responding. Further methodological information and data analyses concerning these exploratory experimental phases can be found in the Supplemental Material.

Extensive avoidance training. The goal of this phase was to extensively train avoidance responses to the aversive US. Avoidance responses consisted of button presses (left or right) in response to one of two different full-screen color stimuli (orange and blue). Both full-screen color stimuli were presented 100 times each. The trial sequence was pseudo-randomized so that each full-screen color stimulus was presented four times in eight consecutive trials. Participants were instructed that the aversive US would occur after each presentation of a full-screen color stimulus but that they could avoid the US by pressing one of the two designated keyboard buttons as fast as possible following the onset of the full-screen color stimulus. Participants were also instructed that they needed to find out by trial-and-error learning which of the two response buttons worked to prevent the US for each of the two color stimuli. Each full-screen color stimulus was presented for a maximum duration of 1000 ms or until a response was performed. Afterward, an outcome was presented depending on the response. Correct responses within 1000 ms were followed by no US, and US omission was highlighted with the presentation of an image of a grey, crossed-out lightning on a white background (1000 ms). Incorrect responses or misses (i.e., no response within 1000 ms) were followed by the US in combination with the presentation of a picture of yellow lightning on a white background (1000 ms). A black fixation cross on a white background was presented for 2000 ms between trials (inter trial interval, ITI). The associations between correct response buttons (left and right) and full-screen color stimuli (orange and blue) did not change within an individual participant and were counterbalanced across participants.

Outcome devaluation. Following the extensive avoidance training, the US outcome was devalued by removing the US electrode from the participant's arm, rendering subsequent US administration impossible. The electrodes were then put away from the participant's table. This removal of the electrode by the examiner was clearly visible for the participants. To ensure that all participants were aware of the devaluation of the aversive US, the removal was emphasized both verbally by the examiner and by on-screen instructions to assure that no participant missed the devaluation procedure (i.e., "I will now remove the electrode").

Habit-goal competition phase. The main question of this phase was whether extensively trained avoidance behavior influenced subsequent goal-directed responses even in the absence of threat (i.e., after devaluation). At the beginning of this phase, participants were instructed that they needed to press the same two buttons as in the previous phase and that they could now gain rewards that would be converted into real money and paid at the end of the experiment. Participants were paid 1 cent per correct approach response but were unaware of this ratio during the experiment. Participants were instructed to continue responding as fast as possible. Participants were not instructed about the new stimulus-response-outcome contingencies or about the different trial type conditions.

The trial sequence was similar to the trial sequence in extensive avoidance training. Each trial consisted of an ITI (black mid-screen fixation cross on a white background, 2000 ms), followed by the presentation of a compound stimulus consisting of a full-screen color with the addition of one of nine neutral object pictures in the middle of the screen (see Figure 1). Compound stimuli were presented for a maximum duration of 1000 ms or until a response was made and were followed by a response-dependent outcome. Following correct responses, a reward outcome (picture of realistically colored 50 cent coin, duration 1000 ms) was presented. Following incorrect responses and misses, a no-reward outcome (picture of a grey, crossed-out 50 cent coin, duration 1000 ms) was presented. The neutral object pictures were simple, black, symmetric depictions (anchor, ball, car, cow, house, lungs, scissors, snowflake, and tree; all widths: 2.1-2.4 cm, all heights: 2.0-2.5 cm), presented vertically and horizontally centered on top of one of three full-screen colors (orange, blue, and green as novel color). The compound stimuli had two important properties: First, two of the full-screen colors of the compound stimuli were identical to the colors in the preceding extensive avoidance training phase (i.e., orange and blue). Second, the full-screen colors were irrelevant for correct responding, i.e., whether the left or the right button was the correct response was determined by the respective object pictures. Thus, the full-screen colors that had been used in extensive avoidance training may elicit habitual responding, either facilitating (i.e., in habit-compatible trials) or impairing (i.e., in habit-incompatible trials) goal-directed responses to the object pictures. Additionally, a third full-screen color (i.e., green) was used, which had not previously been associated with an avoidance response. Responses to this novel full-screen color were not influenced by prior training (i.e., neutral control trials).

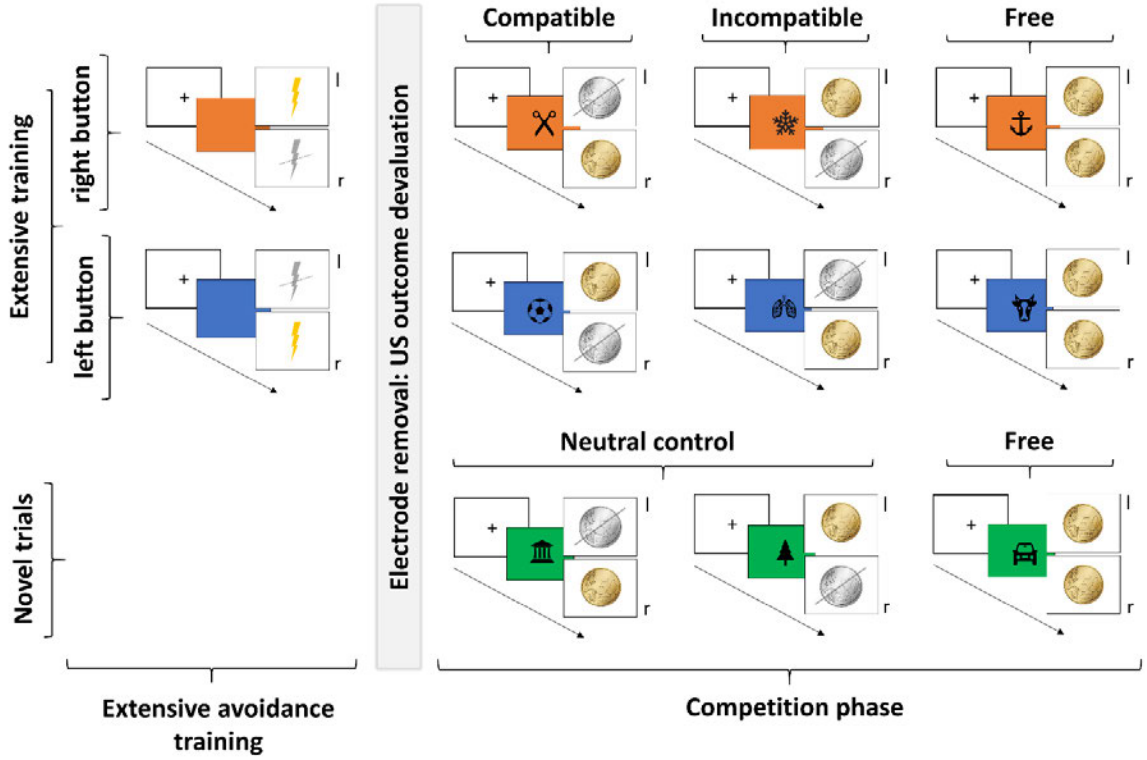
Each full-screen color was paired with three different objects, resulting in nine different compound stimuli in four conditions: 1) habit-compatible trials, 2) habit-incompatible trials, 3)

neutral control trials, and 4) free choice trials. For the compound stimuli with colors that were used during extensive avoidance training (i.e., orange and blue), the correct response for *habit-compatible trials* was the same as the previously extensively trained avoidance response. One distinct object stimulus was presented in compound with each previously used full-screen color stimulus to create two habit-compatible compound stimuli per participant. The correct response for *habit-incompatible trials* was not the same as the previously extensively trained avoidance response, i.e., the other button was correct. Again, one distinct object stimulus was presented in compound with each previously used full-screen color stimulus to create two habit-incompatible compound stimuli per participant. In *neutral control trials*, a novel full-screen color (green) was used. Here, stimulus-response-outcome associations needed to be learned in the absence of competing extensively trained responses. The neutral control condition was added to exploratorily create a reference category for individual learning without prior extensive training. For one neutral control compound stimulus, pressing the left button led to the reward, and for the other neutral control compound stimulus, pressing the right button led to the reward. Neutral control trials were included in the experiment to explore whether extensive training impaired or facilitated subsequent responding as compared with responding without previous training and were used for explorative analyses only. In *free choice trials*, pressing any of the two buttons within the time window of 1000 ms led to the reward outcome. One compound stimulus was implemented for each of the three full-screen background colors. Summarized, the nine different compound stimuli included: two habit-compatible (one for each previously used full-screen color), two habit-incompatible (one for each previously used full-screen color), two neutral control (left and right as correct response for the novel full-screen color), and three free choice stimuli (one for each full-screen color). The trial sequence was pseudo-randomized so that each compound stimulus was presented twice in 18 consecutive trials. The competition phase consisted of 270 trials in total.

Costly habitual avoidance was operationalized as accuracy difference between compatible and incompatible trials (i.e., accuracy compatibility effect), which indicates a failure to inhibit habitual avoidance even when habitual responding produces costs in incompatible trials. Low-cost avoidance habits (i.e., habitual responding without monetary costs) were operationalized as a) reaction time differences between compatible and incompatible trials (i.e., reaction time compatibility effect) and b) extensively trained responses in free trials with the colors that had been presented in the extensive training phase (i.e., orange and blue).

Figure 1

Experimental phases and stimuli



Note. The mapping of colors to correct responses in the extensive training phase and the matching of object symbol stimuli and full-screen colors in the competition phase were counterbalanced across participants. l = left-side button, r = right-side button.

After the experiment, participants rated the perceived unpleasantness of the US (M = 62.56, SD = 21.25), their motivation to avoid the US (M = 83.65, SD = 17.67) and their motivation to approach the rewards (M = 84.84, SD = 18.78) (all Visual Analogue Scales from 0 to 100), and were debriefed.

Statistical analyses

Accuracy data were recorded trial-wise as either correct response, incorrect response, or miss. The mean percentage of correct responses was then calculated block-wise in each condition. Reaction times were recorded on every trial as the interval between stimulus onset and button press. Only reaction time data from correct responses were analyzed. Reaction times lower than 100 ms were excluded from the data analysis. Average reaction times were calculated block-wise for each condition.

Main analyses. In the overtraining phase, accuracy and reaction time data were analyzed in five blocks with 40 consecutive trials each, using repeated measures ANOVAs with the factor Block (block 1, block 2, block 3, block 4, block 5), and subsequent *t* tests for dependent samples.

Data in the competition phase were analyzed in three blocks with 90 consecutive trials each. We hypothesized a priori that 1) accuracy in habit-compatible trials would be higher than accuracy in habit-incompatible trials, 2) reaction times in habit-compatible trials would be lower than in habit-incompatible trials, and 3) habit-compatible responses in free trials would be more frequent than chance. The hypotheses were generated and the sample size was based on an earlier devaluation study (Zwosta et al., 2018), and were tested one-sided with an α-level of .05. We tested the first two hypotheses including potential changes of the effects over time with repeated measures ANOVAs with factor Trial Type (compatible and incompatible) and Factor Block (block 1, block 2, block 3). The third hypothesis was tested with a Wilcoxon test and change of the strength of the effect over time was tested with a repeated measures ANOVA with factor Block (block 1, block 2, block 3). Data were checked for the assumption of normality with Shapiro-Wilk tests. When non-normality was assumed, nonparametric tests were conducted. We report matched rank biserial correlations r_{bs} as effect sizes for the Wilcoxon test, η^2 for ANOVAs, and Cohen's *d* for *t* tests. Data were tested for sphericity with Mauchly's tests. When sphericity could not be assumed, we report Greenhouse-Geisser corrected degrees of freedom. When multiple comparisons were performed, we report Bonferroni-Holm corrected *p* values. Data were aggregated with IBM SPSS Statistics 24 and analyzed with JASP 0.13.1.0.

Exploratory analyses. We added three exploratory analyses for effects for which we could not formulate a priori assumptions. First, as participants' preferred handedness may impact responding in free trials (see Zwosta et al., 2018), we compared the proportion of preferred hand usage in habitual vs. non-habitual responses in free trials with a Wilcoxon test. Second, to investigate whether extensive avoidance facilitated or impaired subsequent goal-directed responding (i.e., relative to goal-directed responding without extensive training), we compared accuracy and reaction times between learning control trials and compatible and incompatible trials, using ANOVAs and *t* tests. A facilitation effect through extensive training would be indicated by higher accuracy or faster response times in habit-compatible than in neutral control trials, while an impairment effect would be indicated by lower accuracy or higher response times in habit-incompatible as compared with neutral control trials. We calculated *t* tests to compare overall accuracy and response times in neutral control trials with habit-compatible

trials and habit-incompatible trials. Third, we exploratorily examined associations between low-cost and costly habitual avoidance. We therefore calculated Pearson's correlations between the accuracy difference scores (i.e., accuracy difference between habit-compatible and habit-incompatible trials, in %; indicating costly habitual avoidance) and the two indicators of low-cost habitual avoidance (i.e., a) reaction time difference scores as the difference in reaction time between habit-compatible and habit-incompatible trials, in ms) and b) the proportion of habit-compatible responses in free trials, in %).

Results

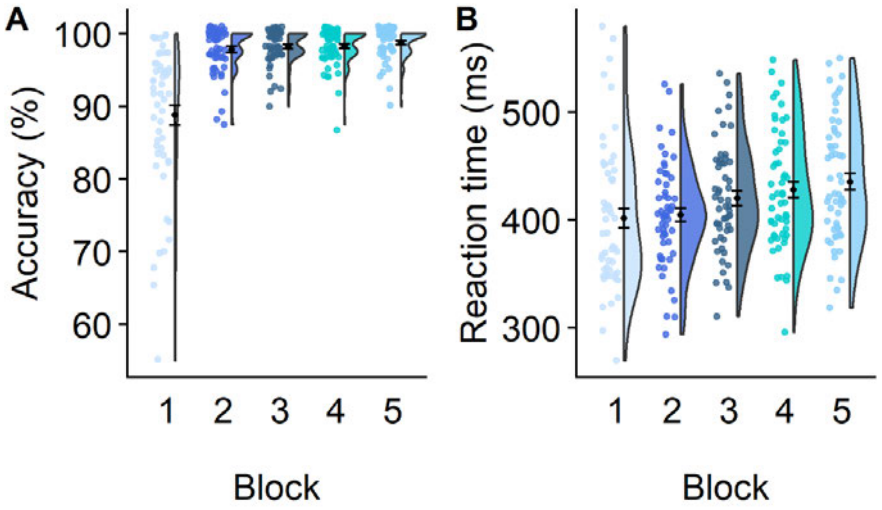
Extensive avoidance training

Accuracy. The mean accuracy in the extensive avoidance training phase was high, M = 96.34% (SD = 3.01), with individual accuracies ranging from 88% to 100%. The average accuracy rate differed significantly between the five blocks, $F(1.359, 73.402) = 47.895$, $p < .001$, $\eta^2 = .470$. Pairwise comparisons revealed a lower average accuracy rate in the first block (M = 88.77%, SD = 10.15) than in all other blocks, $ps_{holm} < .001$, $ds \geq 0.957$, but no differences between the second (M = 97.82%, SD = 3.05), third (M = 98.14%, SD = 2.32), fourth (M = 98.27%, SD = 2.40) and fifth block (M = 98.68%, SD = 2.14), all $ps_{holm} \geq .272$, $ds \leq 0.276$. In sum, accuracy increased early during avoidance overtraining and then remained at a consistently high level (see Figure 2A).

Reaction times. The average overall reaction time during extensive avoidance training was very fast, M = 418.01 ms (SD = 49.15), with a range from 297.35 ms to 516.69 ms. The average reaction times differed significantly between blocks, $F(2.176, 117.512) = 12.754$, $p < .001$, $\eta^2 = 0.191$. Pairwise comparisons between blocks showed that the average reaction times in the first (M = 401.67 ms, SD = 67.66) and second block (M = 404.76 ms, SD = 47.99) were lower than in the third (M = 420.11 ms, SD = 50.09), fourth (M = 427.81 ms, SD = 55.30) and fifth block (M = 435.59 ms, SD = 57.30), $ps_{holm} \leq .042$, $ds \geq 0.358$. Also, the average reaction time was faster in the third than in the fifth block, $t = 3.299$, $p_{holm} = .010$, $d = 0.445$. Reaction times did not differ between the first and second, third and fourth, and fourth and fifth block, $ps_{holm} \geq .131$, $ds \leq 0.358$. These findings indicate that reaction times slowed down over the course of the extensive avoidance training (see Figure 2B), presumably because participants noticed that slightly slower responses were still effective to avoid the US.

Figure 2

Mean accuracy rates (A) and mean reaction times (B) during extensive avoidance training



Note. Black dots depict average values. Error bars depict standard errors. Points in color depict individual data points.

Habit-goal competition

Accuracy. The ANOVA with the factors Block (block 1, block 2, and block 3) and Trial Type (habit-compatible and habit-incompatible) yielded a significant interaction between Trial Type and Block, F (1.479, 91.529) = 8.094, p = .002, η^2 = .016 (see Figure 3A). Post-hoc tests revealed a larger compatibility effect in the first block than in the second (t(54) = 3.029, p_{holm} = .009, d = 0.408) and third block (t(54) = 3.085, p_{holm} = .009, d = 0.416), but no difference between the second and the third block, $t(54)$ = 3.296, p_{holm} = .647, d = 0.062. Thus, costly avoidance as indexed by the accuracy compatibility effect decreased after the first block and remained at a constant level afterward. Follow-up one sample t tests revealed higher average accuracy rates in habit-compatible trials than in habit-incompatible trials within each block, all ps$_{holm} \leq .022$, all ds $\geq$ 0.400 (one-sided testing), indicating that costly habitual responding prevailed until the end of the competition phase.

In free trials with color stimuli which had been presented in the extensive avoidance training (i.e., orange and blue), the previously reinforced response was performed in 55.70% (SD = 29.44), which missed to differ significantly from chance, W = 803.500, p = .055, r_{bs} = .123 (one-sided testing). The proportion of low-cost habitual responding in free trials did not significantly differ between blocks, F(1.510, 81.559) = 0.369, p = .633, η^2 = .007. Therefore, a

low-cost avoidance habit as operationalized by responding in free trials cannot be confirmed, although there was a trend towards significance.

Habitual responses in free trials were carried out predominantly with individuals' preferred hands. Of all habitual responses in free trials, M = 61.12% were carried out with the preferred hand (SD = 32.03), and M = 38.88% were carried out with the non-preferred hand (SD = 32.03). This difference was significant, W = 402.500, p = .024, r_{bs} = .369. The percentage of preferred handedness responses did not differ between free trials in which preferred handedness converged with extensively trained response choices (e.g., for right-handed participants in trials where the right-side button had been extensively trained, M = 61.82%, SD = 38.51) and free trials in which preferred handedness diverged from extensively trained response choices (e.g., for right-handed participants in trials where the left-side button had been extensively trained, M = 49.27%, SD = 42.19), W = 819.500, p = .080, r_{bs} = .285. In neutral free choice trials (i.e., green color), participants tended to press the button corresponding to their preferred hand more often than chance (M = 57.79%, SD = 40.36), W = 1004.00, p = .050, r_{bs} = .457.

Accuracy in habit-compatible trials was higher than in neutral control trials (M = 80.70%, SD = 13.27), $t(54)$ = 3.724, p_{holm} = .002, d = 0.502. Accuracy did not differ between habit-incompatible trials and neutral control trials, W = 417.500, p_{holm} = .081, r_{bs} = .290. These results may indicate that, relative to novel learning, habit-compatible responding was facilitated through extensive training, while habit-incompatible responses were not impaired.

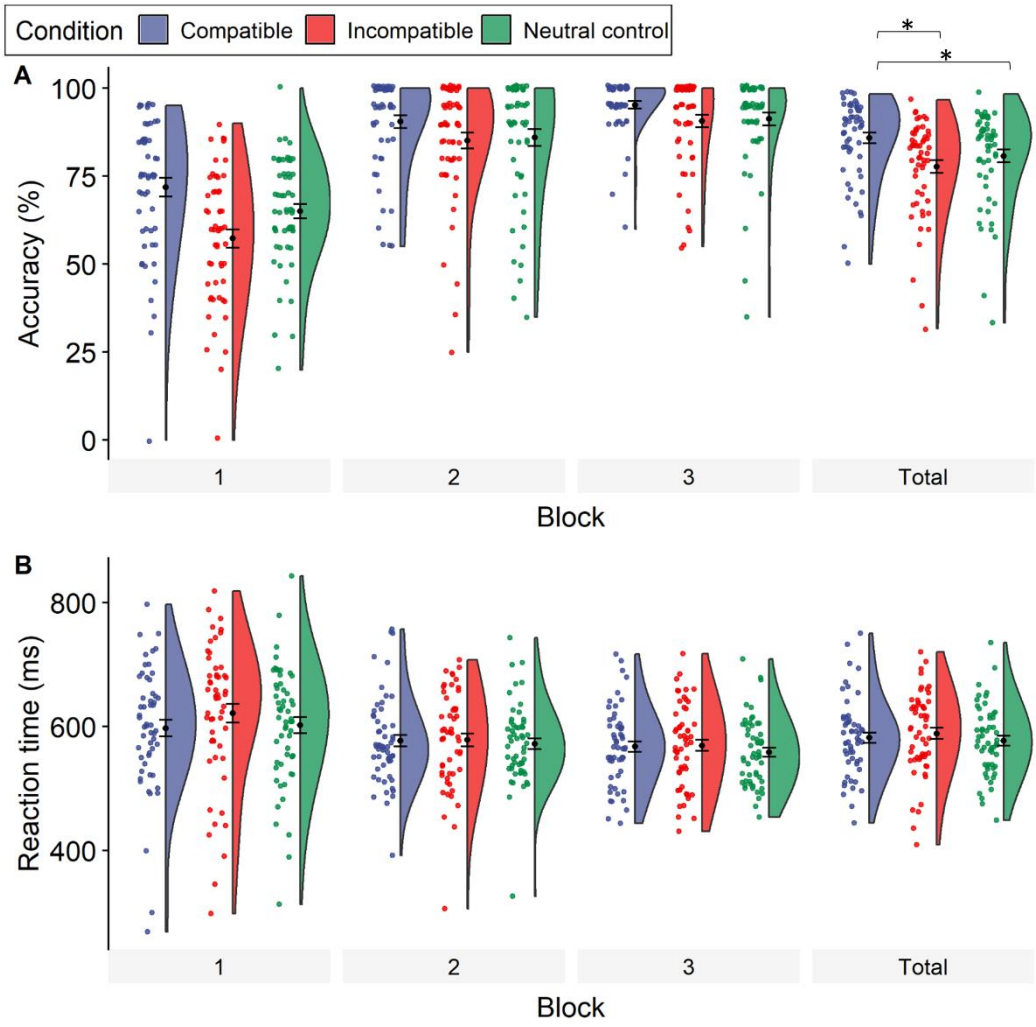
Reaction times. Reaction time did not differ between habit-compatible (M = 581.93 ms, SD = 61.95) and habit-incompatible trials (M = 588.98 ms, SD = 66.54), as indicated by a non-significant effect of Trial Type (see Figure 3B), $F(1, 52)$ = 3.746, p = .058, η^2 = .010, in an ANOVA with factors Block (block 1, block 2, block 3) and Trial Type (compatible and incompatible). There was no significant interaction between Block and Trial Type, $F(1.762, 71.172)$ = 2.202, p = .123, η^2 = .006, indicating that the strength of the reaction time compatibility effect did not vary over time (see Figure 3B). These findings indicate no low-cost avoidance habit as operationalized by reaction time differences.

Reaction time in neutral control trials (M = 577.19 ms, SD = 55.98) did not differ from habit-compatible trials, $t(54)$ = 0.895, p_{holm} = .375, d = 0.121, or habit-incompatible trials, $t(54)$ = 2.277, p_{holm} = .054, d = 0.344.

Associations between costly and low-cost avoidance measures were exploratorily tested with Pearson's correlations. Costly habitual avoidance correlated positively with both indicators of low-cost habitual avoidance, i.e., with the proportion of habitual responding in free trials, $r(55) = .628$, $p_{holm} = .002$, and with the reaction time difference between incompatible and compatible trials, $r(55) = .633$, $p_{holm} = .002$, suggesting that low-cost and costly habitual avoidance measures are associated.

Figure 3

Data distribution and means for accuracy (A) and reaction times (B) during the competition phase



Note. Black dots depict average values. Error bars depict standard errors. Points in color depict individual data points. *$p < .05$, **$p < .001$.

Discussion of experiment 1

We used a modified outcome devaluation paradigm to capture costly and low-cost avoidance habits. To this end, behavior was tested in the presence of positive rewards that could be obtained by habitual (habit-compatible trials) or non-habitual responses (habit-incompatible trials). Goal-directed approach responses were more often correct in habit-compatible trials than in habit-incompatible trials in all three competition phase blocks. Our results, therefore, indicate that extensively trained instrumental avoidance altered subsequent goal-directed approach and created monetary costs. The extensive training effect was strongest in the first block after devaluation, which may hint at stronger avoidance habits directly following extensive avoidance training, but may also be a by-product of the reliance on exploitation during initial stages of trial and error learning, which was strongest at the beginning of the competition phase and declined after S-R-O contingencies were encountered and learned. As participants needed to explore new S-R-O contingencies in the first trials after the devaluation through trial and error learning, the measurement of habitual behavior may be confounded with the exploitation of previously learned associations between stimuli and avoidance responses. Such exploitation can be expected to be most pronounced at the beginning of the competition phase, and should decline after the new S-R-O contingencies are encountered and learned. In other words, higher accuracy in habit-compatible relative to habit-incompatible trials at the beginning of the competition phase may reflect the use of the explicit strategy to exploit the previously beneficial (i.e., preventing the aversive US) habit-compatible responses. Such exploitation of the pre-devaluation contingencies may be an advantageous strategy, as participants had no available information guiding their response choices in the first post-devaluation trials. Importantly, however, the accuracy compatibility effect was observed throughout the entire competition phase, including its last block. The high overall response accuracy in the third block of the competition phase (i.e., 91% in incompatible and neutral control trials vs. 95% in compatible trials) indicates that S-R-O contingencies were successfully acquired, while, simultaneously, an influence of the extensive avoidance training persisted. Arguably, therefore, the accuracy differences in the competition phase did not result solely from trial-and-error learning, but reflect habitual responding. The habit effect in the third block after devaluation may therefore indicate that avoidance habits prevailed. However, only the exclusion of strategic exploration at the beginning of the competition phase would justify to confidently infer an impact of habitual avoidance. To this end, we conducted a second

experiment in which contingencies in the habit-goal competition phase were explicitly instructed.

2.3 Experiment 2

To disentangle potential habitual responding from strategy-driven responding, such as the deliberate exploitation of previously advantageous contingencies, we used a similar design as in Experiment 1. However, we added post-devaluation instructions about stimulus-response-outcome contingencies in the subsequent habit-goal competition phase.

Methods

Participants

Eighty participants took part in the experiment in exchange for 9€ and additional monetary rewards obtained during the experiment. The recruitment process and the exclusion criteria were the same as in Experiment 1. Participants who had taken part in Experiment 1 were not eligible for participation in Experiment 2 and vice versa. The estimation of the expected effect size was based on the accuracy compatibility effect size of d = 0.502 in Experiment 1 and on the assumption that the compatibility effect in Experiment 2 may be smaller due to the lower task difficulty. We, therefore, estimated the sample size based on an expected effect size of d = 0.3 for the main compatibility effect (i.e., mean accuracy difference between compatible and non-compatible trials), one-sided hypotheses testing and a statistical power of .80, resulting in an estimated sample size of N = 71. From the 80 participants tested in the laboratory, seven participants were excluded from the statistical analysis: one due to technical failure, five due to deviations from the standard operation procedure, and one due to self-reported frequent substance use. The final sample consisted of 73 participants: 55 female (75.3%), 66 right-handed (90.4%), mean age of 24.7 years (SD = 6.1, range: 18-53 years).

Procedure, materials, task, and data analysis

The procedure, design, and materials were the same as in Experiment 1. The paradigm consisted of the same phases as in Experiment 1, except that we explicitly instructed stimulus-response-outcome contingencies after outcome devaluation to preclude exploratory trial-and-error learning during the competition phase. Participants were instructed on-screen that they would see symbolic depictions of objects which belonged into three categories. For each category, one correct button press response was assigned. During the presentation of objects in the first category, right-side button presses produced the reward. During the presentation of

objects in the second category, left-side button presses produced the reward. During the presentation of objects in the third category, pressing any of the two buttons produced the reward; participants were asked to spontaneously decide which button they pressed in these trials. The assignment between categories and buttons (left/right/both) was counterbalanced across individuals. All object pictures were presented in the instruction on-screen to eliminate ambiguity about their assignment to the categories. To be able to classify the pictures into meaningful categories, we used different object stimuli than in Experiment 1. Object stimuli were nine symbolic, symmetrical, black depictions of neutral objects (category "furniture": bed, chair, cupboard; category "transport": airplane, car, train; category "tools": driller, pliers, wrench; all heights: 0.8 cm – 2.5 cm, all widths: 1.5 cm – 2.3 cm). After the experiment, we assessed the perceived general unpleasantness of the US (M = 69.86, SD = 13.82), the motivation to avoid the US (M = 86.44, SD = 12.57), and the motivation to approach the reward (M = 79.89, SD = 20.78) as in Experiment 1. Average US intensity was 0.78 mA (SD = 0.57). Data were aggregated and analyzed in the same manner as in Experiment 1.

Results

Extensive avoidance training

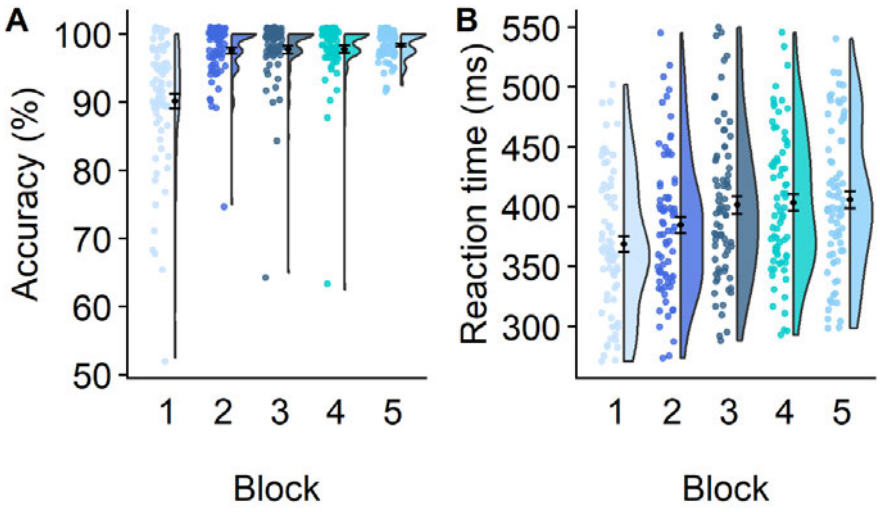
Accuracy. As expected, average accuracy during extensive avoidance training was high, M = 96.33% (SD = 3.87, range 73% - 100%). Average accuracy rates differed significantly between the five blocks, $F(1.770, 127.416) = 45.51$, $p < .001$, $\eta^2 = .387$. Pairwise comparisons revealed that accuracy in the first block (M = 90.16%, SD = 9.23) was lower than in all other blocks, $ps_{holm} < .001$, $ds \geq 1.198$. However, accuracy did not differ between the second (M = 97.57%, SD = 3.89), third (M = 97.71%, SD = 4.82), fourth (M = 97.77%, SD = 4.85) and fifth block (M = 98.36%, SD = 2.09), all $ps_{holm} \geq .999$, all $ds \leq 0.127$. These findings indicate that accuracy rates increased early during extensive avoidance training and then remained at a consistently high level (see Figure 4A).

Reaction time. The mean overall reaction time in extensive avoidance training was 392.73 ms (SD = 52.99, range 286.71 ms - 518.31 ms). The average reaction time differed between blocks, $F(2.415, 173.914) = 19.63$, $p < .001$, $\eta^2 = .214$. Pairwise comparisons revealed that the average reaction time in the first block (M = 368.58 ms, SD = 57.19) was shorter than in all other blocks, $ps_{holm} \leq .006$, $ds \geq 0.376$. Additionally, the average reaction time in the second block (M = 384.77 ms, SD = 58.91) was shorter than in the third, fourth, and fifth block, all $ps_{holm} \leq .006$, all $ds \geq 0.382$. Reaction times did not differ between the third (M = 401.24 ms,

SD = 62.75), fourth (*M* = 403.48 ms, *SD* = 58.48) and fifth (*M* = 405.55 ms, *SD* = 60.48) block, $p_{s_{holm}} \geq .999$, $ds \leq 0.382$. These differences indicate faster responses in the first 80 trials than in the last 120 trials. Similarly to Experiment 1, responses decelerated during the course of the avoidance overlearning phase (see figure 4B), potentially as participants noticed that slightly slower responses were still effective to avoid the US.

Figure 4

Data distribution and average accuracy rates (A) and reaction times (B) during extensive avoidance training



Note: Black dots depict average values. Error bars depict standard errors. Points in color depict individual data points. Error bars depict standard errors.

Habit-goal competition

Accuracy. The ANOVA with factor Block (block 1, block 2, and block 3) and factor Trial Type (habit-compatible and habit-incompatible) yielded no significant main effect of Trial Type, $F(1, 72) = 0.192$, $p = .663$, $\eta^2 < .001$, and no interaction between Block and Trial Type, $F(1.844, 132.770) = 0.157$, $p = .855$, $\eta^2 < .001$. Thus, accuracy in habit-compatible trials (*M* = 96.83%, *SD* = 3.48) did not differ from habit-incompatible trials (*M* = 96.60%, *SD* = 3.46), indicating no costly avoidance habit (see Figure 5A). Given the high accuracy rates, contingency instructions seemed to produce a ceiling effect.

In free trials with full-screen color stimuli which had been presented in extensive avoidance training (i.e., orange and blue), habitual responses were significantly more frequent than chance (51.69%, *SD* = 8.68), *W* = 884.000, *p* = .019, r_{bs} = .345 (one-sided testing). The proportion of

this low-cost habitual responding did not change over the course of the competition phase, $F(2,144) = 0.660$, $p = .518$, $\eta^2 = .009$. These findings indicate stable low-cost habitual avoidance in free trials.

Accuracy in neutral control trials (M = 95.89%, SD = 5.28) did not differ from habit-compatible trials, $W = 963.500$, $p_{holm} = .352$, $d = 0.207$, or from habit-incompatible trials, $W = 1084.000$, $p_{holm} = .212$, $d = 0.185$.

Looking closer at the habitual responses in free trials, we found that habitual responses were carried out mostly with individuals' preferred hands. Of all habitual responses in free trials, M = 71.60% (SD = 32.73) were performed with the preferred hand, while M = 28.40% (SD = 32.73) were performed with the non-preferred hand. This difference was significant, $W = 2174.500$, $p < .001$, $r_{bs} = .610$. The proportion of responding with the preferred hand did not differ between free trials in which preferred handedness and extensively trained responding converged (M = 72.24%, SD = 32.09) and trials where preferred handedness and extensively trained responding diverged (M = 68.86%, SD = 33.74), $W = 861.100$, $p = .063$, $r_{bs} = .299$. In neutral free choice trials (i.e., with green background color), participants used their preferred hand in 71.14% (SD = 33.68), which was above chance level, $W = 2073.00$, $p_{holm} = .003$, $r_{bs} = .767$.

Reaction times. Responding in habit-compatible trials (M = 513.99 ms, SD = 47.63) was significantly faster than responding in habit-incompatible trials (M = 519.85 ms, SD = 49.32), as indicated by a significant effect of Trial Type, $F(1,72) = 4.473$, $p = .038$, $\eta^2 = .012$ (see Figure 5B). There was no significant interaction between Block and Trial Type, $F(2, 144) = 1.244$, $p = .291$, $\eta^2 = .003$, indicating that the strength of the reaction time compatibility effect did not vary over time.

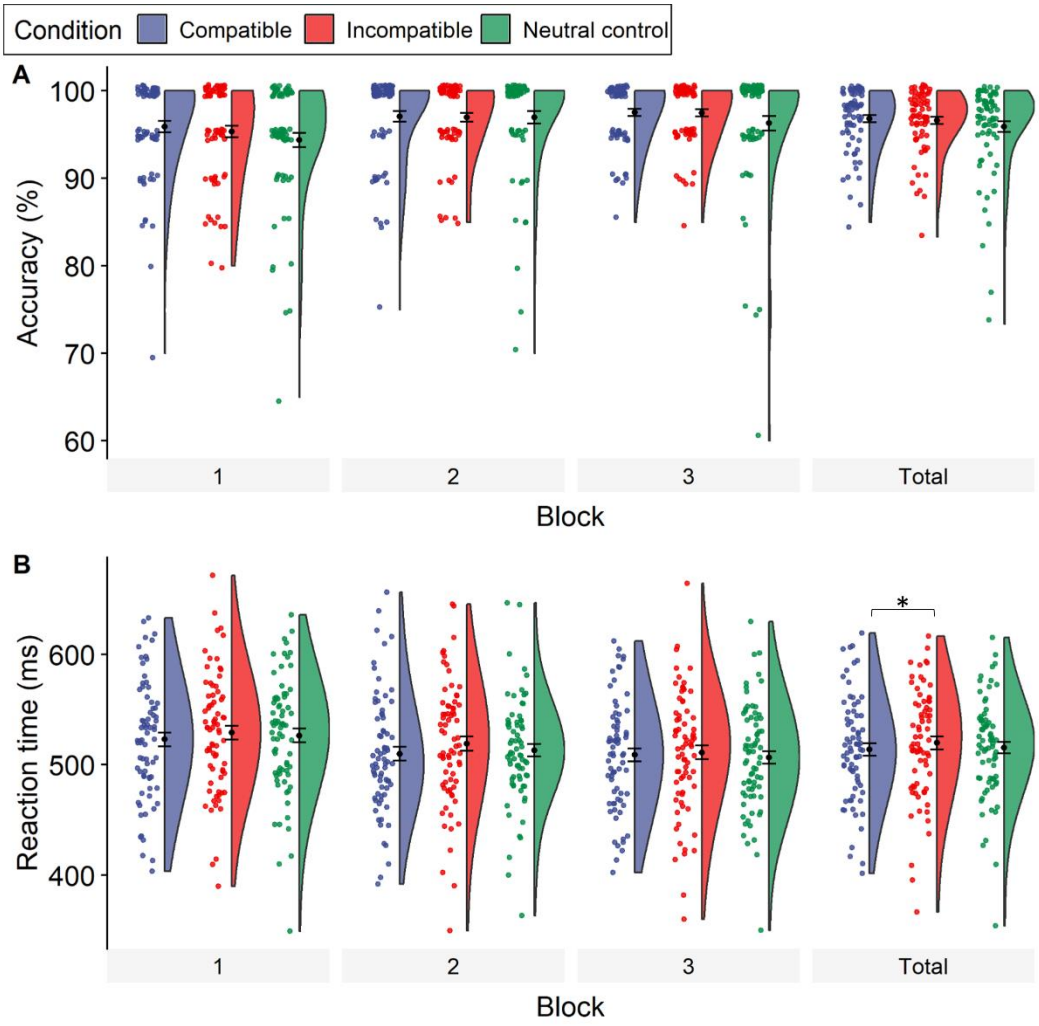
Response time did not differ between neutral control trials (M = 515.41 ms, SD = 44.93) and habit-compatible trials, $t(72) = 0.605$, $p_{holm} = .547$, $d = 0.071$, or between neutral control trials and habit-incompatible trials, $W = 1593.000$, $p_{holm} = .366$, $r_{bs} = .180$.

Costly habitual avoidance (i.e., accuracy differences between habit-compatible and habit-incompatible trials) correlated positively with low-cost habitual avoidance (i.e., reaction time differences between habit-incompatible and habit-compatible trials), $r(73) = .424$, $p_{holm} = .002$, indicating that low-cost and costly avoidance habits were associated. The low-cost habit effect

in free trials did not correlate with the reaction time compatibility effect, $r(73) = .205$, $p_{holm} = .081$.

Figure 5

Data distribution and means for accuracy (A) and reaction times (B) during the competition phase



Note: Black dots depict average values. Error bars depict standard errors. Points in color depict individual data points. *$p < .05$.

Discussion of experiment 2

Explicit instructions about stimulus-response-outcome contingencies after extensive avoidance training and outcome devaluation produced compatibility effects that differed from those in Experiment 1. In contrast to Experiment 1, we did not observe any impact of the extensive training on the accuracy rates in habit-compatible versus habit-incompatible trials.

However, habit-incompatible responses were modestly slower than habit-compatible responses, indicating reaction time costs due to extensive avoidance training. Interestingly, an additional low-cost avoidance habit effect in free choice trials emerged. These findings again suggest an impact of avoidance habits on the subsequent goal-directed approach of rewards.

2.4 General discussion

Habitual responding is assumed to play a pivotal role in maintaining maladaptive avoidance behavior, but the evidence for this assumption is scarce. Using a modified devaluation paradigm, the present study tested the strength of costly and low-cost avoidance habits following extensive training of avoidance responses. Costly habits were defined when extensively rained avoidance influenced goal-directed approach at the cost of monetary reward (i.e., response accuracy in habit-compatible vs. habit-incompatible trials). Low-cost habits were defined as extensively trained avoidance influencing goal-directed approach without inflicting costs (i.e., reaction time compatibility effect and habitual response preference in free trials). The main results indicate that extensively trained avoidance affected subsequent goal-directed approach differentially in the two experiments. Evidence for costly, but not low-cost avoidance was found in Experiment 1, whereas evidence for low-cost habitual avoidance tendencies but not costly habitual avoidance was found in Experiment 2. As both experiments differed only in the instructions about and complexity of the stimulus-response-reward contingencies during the competition phase, the differential effects cannot be explained by differences in avoidance training. The strength of inflexible avoidance was thus modulated by the complexity and ambiguity of the competing goal-directed approach behavior.

Our findings demonstrate that extensive avoidance training resulted in low-cost habitual avoidance under specific circumstances. First, low-cost habitual tendencies as indicated by a reaction time compatibility effect emerged in Experiment 2, but not in Experiment 1. This reaction time delay in habit-incompatible trials is in line with data from earlier devaluation studies (Luque et al., 2019; Zwosta et al., 2018), corroborating the notion that reaction time compatibility effects can be used as an indicator for competing habitual and goal-directed action control when habitual control is not strong enough to generate accuracy habit effects (see Luque et al. 2019). However, it should be noted that the reaction time compatibility effect in Experiment 2 was relatively small. As a separate indicator for low-cost avoidance habits, we used the percentage of habit-compatible responses, which was significantly higher than chance in free trials in Experiment 2. This result is in line with previous devaluation studies on low-

cost avoidance habits (e.g., Flores et al., 2018; Gillan et al., 2014; Gillan et al., 2015; Zwosta et al., 2018, but see de Wit et al., 2018). Low-cost habitual avoidance can be assumed to adaptively guide actions when stimulus-response-outcome contingencies are stable, i.e., when extensively trained stimulus-response associations still lead to a favorable outcome and do not create costs. In such stable environments, the exploitation of previously trained responses (i.e., habitual responding) reduces the cognitive costs of action control and therefore may be beneficial (Kane & Engle, 2003).

Importantly, evidence for costly avoidance habits was found under specific circumstances. In Experiment 1, extensive training resulted in costly avoidance (i.e., lower accuracy in habit-incompatible than habit-compatible trials). The accuracy compatibility effect in the last block of Experiment 1 suggests that this costly avoidance was not the sole product of trial-and-error learning and goal-directed exploitation of previously beneficial stimulus-response-outcome contingencies, but had habitual characteristics. In Experiment 2, however, no costly habitual avoidance was found (i.e., accuracy rates did not differ). Thus, although our study suggests that extensively trained avoidance responses may impair subsequent goal-directed approach, this effect seems to depend on the characteristics of the competing goal-directed behavior.

Indeed, goal-directed responding's complexity differed between the two experiments, and these differences may help explain the diverging effects. In Experiment 1, participants had to explore and learn stimulus-response-outcome contingencies for nine stimuli through trial and error. In contrast, in Experiment 2, the contingencies between only three object categories and the corresponding correct responses needed to be learned and were explicitly instructed. This resulted in more frequent and faster correct responses during the competition phase in Experiment 2 as compared with Experiment 1 (e.g., accuracies in habit-incompatible trials > 95% vs. < 85%, average reaction time in habit-incompatible trials 518 ms vs. 589 ms). These differences in complexity of the stimulus-response-outcome associations guiding the goal-directed behavior may have affected the expression of habitual avoidance.

Goal-directed control more effectively inhibits habitual behavior when this inhibition requires low effort (see Otto, Gershman, et al., 2013). When stimulus-response-outcome contingencies are complex rather than simple, goal-directed responding can be expected to require more cognitive resources for monitoring behaviorally relevant stimuli and their associations with outcomes and advantageous responses. Goal-directed control may, then, less effectively inhibit competing, disadvantageous, habitual responses elicited by environmental

stimuli. Our results show that costly avoidance (i.e., less accuracy in compatible versus incompatible trials) was apparent in the difficult task version (i.e., in Experiment 1), but not in the easy task version (i.e., in Experiment 2). Potentially, costly avoidance may be more pronounced when goal-directed control requires more cognitive control, such as under more complex environmental contingencies, or when subjective uncertainty about stimulus-response-outcome contingencies is high, for example, early in trial-and-error learning.

The findings indicate a mixture of costly and low-cost habitual responding after extensive training. Although any costly effect of extensive training on subsequent goal-directed responding was eliminated in Experiment 2, low-cost habitual tendencies were found. This indicates that habitual responding was acquired to some extent but could be inhibited when interfering with goal-directed control (see also Hardwick et al., 2019; Luque et al., 2019). Of note, costly and non-costly avoidance measures positively correlated in both experiments, potentially suggesting a general individual propensity to respond habitually after extensive training. In summary, a costly impact of habitual avoidance on subsequent approach seems more likely when stimulus-response-outcome contingencies guiding goal-directed behavior are complex or ambiguous. In consequence, costly habitual avoidance may be reduced by competing rewards when the goal-directed approach responses to obtain these rewards are simple and clearly instructed.

Different mechanisms may cause habitual responding in this devaluation paradigm. Theoretically, extensive training can facilitate habit-compatible responding, impede habit-incompatible responding, or facilitate habit-compatible responding and impede habit-incompatible responding simultaneously. We observed higher accuracy in habit-compatible compared with neutral control trials in Experiment 1, potentially suggesting a facilitating effect of extensive training, which is not in line with previous findings (Zwosta et al., 2018). No further differences in comparison with neutral control trials were found, rendering it difficult to draw conclusions about potential mechanisms underlying reaction time and accuracy compatibility effects. Further studies may shed light on these mechanisms, for example, by trying to boost differences between conditions by adding time pressure (see Luque et al., 2019) or by using more complex instructed S-R-O contingencies.

Low-cost avoidance in free trials was predominantly carried out with participants' preferred hands. When the preferred handedness effect superimposed on habitual tendencies acquired through extensive training, the combination of these two habitual tendencies may thus have led

to elevated habitual responding. Responding with the preferred hand itself may include a habitual component that interacts with other habitual response tendencies. For example, habits are amplified when carried out with individuals' dominant hand, whereas behaviors performed with the non-dominant hand tend to be less influenced by habitual tendencies (Neal et al., 2011). A dominance of responding with the dominant hand in free trials has accordingly been found in one earlier study (Zwosta et al., 2018). Future devaluation studies may include more systematic examinations of the impact of preferred handedness on habitual responding.

Devaluation paradigms are commonly implemented standard procedures to evaluate habit strength. Devaluation procedure implementations vary widely in the literature and include the satiation with favorable outcomes (e.g., Schwabe & Wolf, 2010), the mere instruction of stimulus-response-outcome contingency changes (e.g., Luque et al., 2017; Luque et al., 2019; Zwosta et al., 2018), and the removal of electrodes used to deliver electrical stimulations (e.g., Gillan et al., 2015; Gillan et al., 2014). The removal of the electrodes for electrical stimulations has been shown to reliably decrease participants' avoidance motivation (see Gillan et al., 2015). In our experiments, we did not measure the self-reported effect of the removal of the electrodes on participants' subsequent avoidance motivation. As the electrodes were completely removed, and the removal was emphasized prior to the competition phase verbally and on-screen, however, any remaining expectation of shock during the competition phase is very unlikely. Future studies may, however, control for the change of threat expectancy or avoidance motivation after the devaluation.

The reduction of costly habitual avoidance behavior through a clearly instructed approach of rewards introduces some potential clinical implications. Previous research demonstrated that goal-directed avoidance can be reduced by reinforcing avoidance-incompatible approach behavior (Pittig, 2019). Our findings suggest that competing rewards can similarly reduce costly habitual avoidance. As habit-incompatible rewards reduced costly avoidance only when the contingencies for obtaining these rewards were clear (i.e., in Experiment 2), potentially, simple and transparent approach goals and reward contingencies facilitate the reduction of avoidance habits. For example, patients and therapists may identify such simple responses for obtaining motivating rewards, which are incompatible with avoidance habits, in the context of exposure-based therapies. In general, early interventions to support goal-directed approach may be especially effective since goal-directed action control can be expected to be stronger in early learning stages. Such early interventions may be feasible for some patients seeking treatment in the first year after disorder onset (Christiana et al., 2000; Kessler et al., 1998). However,

studies with clinically relevant samples are needed to investigate the role of costly habitual avoidance in anxiety disorders to provide evidence-based recommendations. The robustness of the effectiveness of rewards in reducing costly habitual avoidance in different tasks and environments needs to be confirmed in further studies.

In conclusion, our experimental findings indicate that in partly habitual avoidance, some degree of goal-directedness remained and could overrule habitual tendencies to obtain concurrent rewards. Thus, the potential for adaptive behavior change remained in partly habitual avoidance. Implementing rewards for habit-incompatible responses may therefore be applicable to counteract costly habitual, maladaptive avoidance.

3. Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm

Valentina M. Glück[1], Juliane M. Boschet-Lange[1], Roxana Pittig[1], Andre Pittig[2]

[1]Department of Psychology (Biological Psychology, Clinical Psychology, and Psychotherapy), University of Wuerzburg, Germany

[2]Translational Psychotherapy, Institute of Psychology, University of Goettingen, Göttingen, Germany

*Corresponding author: Andre Pittig, Georg-August-University of Goettingen, Translational Psychotherapy, Kurze-Geismar-Str. 1, 37073 Goettingen, Germany. Email: andre.pittig@uni-goettingen.de

Published as:

Glück, V. M., Boschet-Lange, J. M., Pittig, R., & Pittig, A. (2023). Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm. *Behaviour Research and Therapy*, *170*, 104417. https://doi.org/10.1016/j.brat.2023.104417

Abstract

Background: A habitual avoidance component may enforce the persistence of maladaptive avoidance behavior in anxiety disorders. Whether habitual avoidance is acquired more strongly in anxiety disorders is unclear. *Methods:* Individuals with current social anxiety disorder, panic disorder and/or agoraphobia (n = 62) and healthy individuals (n = 62) completed a devaluation paradigm with extensive avoidance training, followed by the devaluation of the aversive outcome. In the subsequent test phase, habitual response tendencies were inferred from compatibility effects. Neutral control trials were added to assess general approach learning in the absence of previous extensive avoidance training. *Results:* The compatibility effects indicating habitual control did not differ between patients with anxiety disorders and healthy controls. Patients showed lower overall approach accuracy, but this effect was unrelated to the compatibility effects. *Conclusions:* In this study, anxiety disorders were characterized by reduced approach but not stronger habitual avoidance. These results do not indicate a direct association between anxiety disorders and the acquisition of pervasive habitual avoidance in this devaluation paradigm.

Key words: Anxiety disorders, avoidance, approach, habits

3.1 Introduction

Avoiding feared stimuli is a key symptom of anxiety disorders (e.g., American Psychiatric Association, 2013; Heeren & McNally, 2018). Avoidance behaviors contribute to the chronification of fear and anxiety by limiting opportunities for fear extinction (e.g., Lovibond et al., 2009), and negatively impact general psychosocial functioning (Moitra et al., 2008; Wittchen et al., 2000). The relationship between avoidance and fear can take the form of a vicious circle: fear motivates avoidance, while avoidance causes and maintains fear via several mechanisms (e.g., Craske et al., 2017; Krypotos et al., 2015; Lovibond et al., 2009; Pittig et al., 2020; Struijs et al., 2018). Reducing avoidance behaviors is an important psychotherapeutic target, which is often not easily met because avoidance can strongly persist. This persistence has been demonstrated experimentally by studies showing that avoidance behavior can prevail even after successful fear extinction (Pittig & Wong, 2022; van Uijen et al., 2018; Vervliet & Indekeu, 2015), and this effect is especially pronounced in anxious individuals (Wake et al., 2021). These findings point towards anxiety-related biases in decision-making and action regulation, both regarding the effects of anxiety on *what* is chosen (e.g., approach or avoidance) and regarding *how* anxiety affects response choices (e.g., biases in habitual and goal-directed action control).

A shift from goal-directed to habitual avoidance has been proposed as a potential mechanism contributing to the persistence of avoidance (Arnaudova et al., 2017; Hofmann & Hay, 2018; LeDoux & Daw, 2018; LeDoux et al., 2017; Pittig et al., 2020). Habitual behavior, as conceptualized in the associative dual-process framework (e.g., Adams & Dickinson, 1981; Dickinson, 1985; Watson et al., 2022), reflects a direct association between the perception of an external or internal stimulus and a motor response (i.e., S-R association) that emerges through frequent repetitions of responses in environments with stable contingencies. Habitual control of behavior is contrasted with goal-directed control, which enables organisms to pursue internally generated goals (e.g., Daw, 2015). This distinction is related to, although not congruent with, action control frameworks such as the reinforcement learning model, in which model-free and model-based planning are distinguished (e.g., Drummond & Niv, 2020; Vandaele & Janak, 2018). Model-free planning relies directly on recent learning experiences with a stimulus, thereby conceptually resembling habitual control. In contrast, model-based planning relies on representations of the environment which enables flexibility in volatile environments, thereby resembling goal-directed control (see Lloyd & Dayan, 2016). Findings from the reinforcement learning model can thus inform research in the associative dual-process framework. Although the processes involved in allocating habitual and goal-directed control

are still being debated (e.g., Balleine & Dezfouli, 2019; Wood et al., 2022), and despite powerful critique of the methods and the interpretations of experimental results (e.g., Moors et al., 2017) the framework has been frequently applied in clinical research. For example, amplified habitual response tendencies have been demonstrated in obsessive-compulsive disorder (e.g., Gillan et al., 2014; Gillan et al., 2011) and substance use disorder (e.g., Everitt & Robbins, 2016), among other disorders. However, the evidence about amplified habitual response tendencies in anxiety disorders is scarce and inconclusive.

Biases in habitual control in high trait anxiety and anxiety disorders have been examined in several studies. These studies used outcome devaluation tasks derived from the associative dual-process framework or reinforcement learning tasks derived from the reinforcement learning model. Trait anxiety as a potential risk factor for acquiring habitual responses was not associated with more pronounced habitual responding: Trait anxiety was not associated with habitual responding (Gillan et al., 2014; Patterson et al., 2019) or was predictive of habitual responding only when not controlling for intolerance of uncertainty (Flores et al., 2018). Trait anxiety was also not associated with more pronounced model-free planning in a reinforcement learning task (Gillan et al., 2016). Furthermore, three studies examining the association between anxiety and stronger habitual vs. goal-directed responding resulted in ”no evidence that anxiety impairs goal-directed control in human subjects” (Gillan et al., 2021, p. 1467). Thus, these studies do not support the view that trait anxiety is a risk factor directly contributing to the acquisition of habitual responding. The few available studies on habitual control in anxiety disorders are inconsistent: Alvares et al. (2014) reported stronger habitual responses in participants with social anxiety disorder than in healthy controls in an outcome devaluation task and replicated this finding in an independent sample (Alvares et al., 2016). In another study, generalized anxiety disorder diagnosis did not predict stronger habitual responses in a devaluation task (Roberts et al., 2022). The current evidence on the relationship between sub-clinical and clinical anxiety and habitual avoidance is, thus, inconclusive.

A well-established experimental procedure to differentiate habitual and goal-directed action control is the outcome devaluation task, which originated from rodent research (e.g., Adams & Dickinson, 1981) and was only recently translated into research in humans (Schwabe & Wolf, 2009). A typical devaluation paradigm consists of three parts. First, in a training phase, participants are trained to perform two instrumental responses that lead to a rewarding outcome (i.e., to test approach habits) or to the cancellation of an aversive outcome (i.e., to test avoidance habits) to two external stimuli (such as two geometrical shapes). One of the previously presented outcome values is then devalued. The different devaluation procedures used in the

literature with human participants vary substantially and can include the satiation with a reward (e.g., offering unlimited chocolate consumption to devalue the outcome of chocolate milk in Schwabe & Wolf, 2010), the direct manipulation of the outcome value (i.e., pairing one of two liquids with an aversive taste in Buabang, Boddez, et al., 2023), the mere instruction of a reduced value of the outcome (e.g., informing the participants that no shocks would be delivered anymore in Gillan et al., 2015), or the removal of the outcome (e.g., taking off one electrode which had delivered the outcome aversive electrotactile stimulation in Gillan et al., 2014). These procedures differ in whether the outcome value (i.e., desirability or undesirability of the outcome) or the outcome contingency (i.e., likelihood of the occurrence of the outcome) is changed. The term outcome devaluation has first been introduced for direct outcome value reductions (i.e., Valentin et al., 2007), but outcome contingency reductions are usually also described as devaluation procedures, although they resemble instructed extinction procedures (see Luck & Lipp, 2016) when the outcome can technically still occur (e.g., no removal of the electrode used for deliverance of an aversive outcome). The various outcome devaluation practices share the purpose of testing whether responding in the subsequent test phase is sensitive or insensitive to the change of outcome valence or outcome contingency. The subsequent test phase usually features the same stimuli and response choices as the training phase but is usually carried out in extinction (i.e., no outcomes are presented anymore). Typically, the participants are not instructed about this removal of all outcomes but only about the devaluation of one of the stimuli. Finally, the difference between the response rates to the still valued and the now devalued stimulus is used to indicate the impact of the outcome devaluation on action control. Larger differences are assumed to indicate a stronger impact of outcomes on behavior and more pronounced goal-directedness. Reversely, more equal response rates to the valued and the devalued stimulus are assumed to indicate stronger habitual responding, i.e., less impact of outcomes on behavior.

Despite their wide use, devaluation paradigms have been criticized regarding several methodological limitations. Firstly, they are often not well suited to detect goal-directed behavior when participants pursue goals that differ from those implied by the experimental design logic, leading to false-positive classifications of such responses as habitual (Buabang, Boddez, et al., 2023; de Houwer et al., 2022; de Houwer et al., 2018; Moors et al., 2017). For example, in designs where habitual responding in the test phase is not linked to a disadvantage for the participant and only one of two aversive outcomes is devalued, participants may use a "better safe than sorry strategy" for their decisions in the test phase because they doubt the effectiveness of the devaluation of the aversive stimulation and thereby still avoid

goal-directedly. As another example, the goal to save cognitive resources may lead to task disengagement and the strategic non-adjustment of responses especially when non-adjustment is not associated with costs. Avoidance without costs can be seen as an adaptive response that is also frequent in healthy individuals, even without previous extensive avoidance training (e.g., Pittig, 2019; Pittig, Boschet, et al., 2021; Pittig & Wong, 2021). Persistent responses to devalued stimuli in the test phase may therefore be strategy-driven and goal-directed but may frequently erroneously be interpreted as habitual in traditional devaluation paradigms when the potential use of a better-safe-than-sorry strategy or other strategies is not considered.

One possibility to circumvent this ambiguous interpretation of persistent responses to devalued stimuli is to include costs for habitual avoidance in the test phase. To this end, we recently introduced a variation of the traditional devaluation paradigm (Glück et al., 2021). In this design, habitual responses in incompatible trials (i.e., test phase trials that are incompatible with the extensively trained response) lead to monetary losses to motivate participants to adjust their responses goal-directedly and, thereby, to create a valid competition between habitual and goal-directed response tendencies. The competing rewards are thought to minimize the risk of pursuing arbitrary goals that may produce seemingly habitual responses. For example, the goal to save cognitive resources may lead to task disengagement and the strategic non-adjustment of responses. Additionally, in this design, the devaluation of the aversive outcome is unambiguous (i.e., both electrodes for the electrical stimulations are removed) to reduce uncertainty about the effectiveness of the devaluation and thus to make the application of a goal-directed “better safe than sorry” strategy unlikely.

Using this adapted outcome devaluation paradigm, we examined differences in the acquisition of habitual avoidance between individuals with anxiety disorders and healthy control participants to gain more evidence on habitual avoidance aberrances in anxiety disorders. Specifically, we hypothesized that individuals with anxiety disorders were more prone to consistently repeat extensively trained avoidance responses after the devaluation of the aversive outcome than individuals without anxiety disorders.

3.2 Material and methods

Participants

71 patients with a primary diagnosis of panic disorder with or without agoraphobia or social anxiety disorder and 71 matched participants without current psychological diagnoses (i.e., healthy controls) took part in the experiment. The controls and the patients were matched on

age (± 3 years) and gender. Data from one patient and five controls had to be excluded prior to data analysis due to technical problems during data collection. Additionally, data from three controls were excluded because of low accuracy in the training phase (i.e., < 70%), which was considered too low for reliable habit acquisition. As the respective matched participants were also excluded from data analysis, the final matched sample consisted of n = 62 participants in each group (see Table 1). As no prior research was available to estimate an expected effect size, the sample size was estimated loosely based on an earlier study (Glück et al., 2021).

The main inclusion criterion for the anxiety group was a primary ICD-10 diagnosis of agoraphobia (N = 10; 16.1%), panic disorder (N = 4, 6.45%), panic disorder with agoraphobia (N = 23; 27.1%), and/or social anxiety disorder (N = 25; 40.3%). Thirteen patients (21.0%) fulfilled the criteria for two inclusion diagnoses. The primary diagnoses were confirmed prior to study participation by licensed psychotherapists with the Mini-DIPS, a structured clinical interview for the assessment of ICD-10 and DSM-5 diagnostic criteria (Margraf & Cwik, 2017). Participants with comorbid disorders were included if the primary diagnosis was the anxiety disorder diagnosis. Twenty-two patients (35.5%) were diagnosed with one comorbid disorder, which was not a second anxiety diagnosis, and six patients (9.7%) were diagnosed with two comorbid disorders, which were not anxiety diagnoses. Current psychopharmacological medication was not an exclusion criterion if the medication had been used for at least four months with a stable dosage for at least four weeks prior to the study participation. In the anxiety group, 22 participants (35.5%) used at least one psychopharmacological medication. A control analysis with only unmedicated patients and their respective matched control participants yielded no change in the pattern of results (see Supplement B). The patients completed the experimental task as a part of their participation in a psychotherapy trial, which included laboratory assessments before the treatment. They were recruited via the outpatient psychotherapy clinic at the University of Würzburg and advertisements in local doctor's offices. The control participants were recruited from the general local population via advertisements in social media and an online recruitment platform of the University of Würzburg. Psychology students were generally allowed to take part as matched participants in the control group, but only during their first two study semesters. Exclusion criteria in both groups included acute suicidality, current substance use disorder, a lifetime diagnosis of bipolar, psychotic, or borderline personality disorder, current pregnancy, serious physiological conditions, age under 18 years or over 65 years, as well as insufficient proficiency in German to understand the task instructions. In the control group, additional exclusion criteria included a self-reported lifetime

diagnosis of psychiatric or neurological disorders and current psychopharmacological medication.

All participants provided written informed consent prior to their study participation. The ethics committee at the University of Würzburg approved all procedures (GZEK2018-20) and the experiment was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013). Each participant took part in two laboratory sessions one week apart for approximately 3 h each. The experimental task described here was scheduled on the second day. All participants received a fixed reimbursement of 10 € per hour and were informed that they could gain up to 5 €, depending on their performance in the experimental tasks.

Procedure

All experimental procedures were conducted with one participant at a time and were identical in both groups. At the beginning of the session, the participants were fitted with electrodes to record skin conductance responses and an electrocardiogram; these data were recorded for purposes other than for the current study. Next, the participants were asked to fill in several psychometric questionnaires. They were placed seated in front of a desktop computer screen on which the experimental task stimuli and instructions were presented. The participants' computer keyboard for the behavioral responses was positioned centrally in front of the 24'' screen on which the task was presented. Before the experimental task started, the intensity of the aversive electrotactile stimulus was calibrated with a standardized procedure. The stimulation was presented in ascending intensity until it was rated as "unpleasant, but not painful", corresponding to a rating of "4" on a scale from "0" (no sensation) to "5" (painful sensation); this intensity was used in the experimental tasks. Each aversive stimulation consisted of 125 separate, consecutive electrotactile stimulations with a duration of 2 ms each and a temporal distance of 3 ms between them (total stimulus duration: 625 ms). The electrotactile stimulations were delivered with a bar electrode (diameter 8 mm, spacing 30 mm) attached to the participant's non-dominant forearm and were generated with a Digitimer DS7R stimulator (Digitimer Ltd). On the day of the experiment, the participants completed two experimental tasks involving electrotactile stimulation immediately before the experimental task described here.

Devaluation paradigm

The experimental task was identical to an earlier study (i.e., Experiment 1 in Glück et al., 2021) with the only exception that the test phase was shortened from 270 trials to 180 trials

(i.e., ten instead of 15 blocks with 18 trials each were presented) to reduce the task duration. The experimental task consisted of three parts: *extensive avoidance training*, *outcome devaluation*, and *test phase*. Two exploratory phases were presented after the test phase (i.e., a *revaluation* and a *reinstatement phase*). These exploratory phases cannot be interpreted in the devaluation paradigm framework, and no a priori hypotheses were formulated. Further information about the exploratory phases, including the data analysis, can be found in Supplement D. The total task duration was approximately 25 min. The experiment was programmed and presented, and the data were recorded with Presentation 18.1 (Neurobehavioral Systems, Berkeley, USA).

Extensive avoidance training. Two simple avoidance responses were extensively trained in this phase. The avoidance responses were button presses (i.e., left or right button) to prevent the occurrence of the electrotactile stimulation in response to one of two different full-screen background colors (i.e., orange and blue, see Figure 1). Before this phase, the participants received written instructions that the aversive electrotactile stimulation would occur after each presentation of the background color unless they pressed the correct button during the background color presentation. The participants were also instructed to find out by trial-and-error which of the two response buttons was effective in canceling the aversive electrotactile stimulus for each background color and to respond as fast as possible.

Each trial started with the presentation of the background color and ended with the button press or if no button was pressed, after 1000 ms. Correct and timely (i.e., reaction time < 1000 ms) responses prevented an aversive stimulation, as highlighted by the visual presentation of a crossed-out, grey lightning bolt centered on a white background (1000 ms). Incorrect or missing responses (i.e., no response within 1000 ms) were followed by the electrotactile stimulation and the visual presentation of a yellow lightning centered on a white background on the screen (1000 ms). The associations between outcome (stimulation or no stimulation), response button (left and right), and background color (orange and blue) did not change within an individual participant but were counterbalanced across participants. The trials were separated by a black fixation cross on a white background which was presented for 2000 ms (i.e., inter-trial-interval). The training phase consisted of 200 trials in a pseudo-randomized order within 25 blocks with eight trials each (i.e., four trials with blue and four trials with orange background were presented in a randomized order in each block, and the blocks were presented seamlessly).

Outcome devaluation. The aversive outcome was devalued by removing the stimulation electrode from the participant's arm. To ensure that all participants were aware of this

procedure, the removal was emphasized verbally by the experimenter and by on-screen instructions. We, thereby, did not directly manipulate the outcome contingency (i.e., the deliverance of the aversive outcome was rendered impossible by the devaluation procedure).

Test phase. Before the test phase, the participants were informed that the two response buttons they had used in the training phase would still be available in the upcoming phase but that the electrotactile stimulations could not be delivered anymore. They were informed that they could instead receive monetary rewards that would be converted into real money and paid at the end of their study participation. They were also instructed to respond as fast as possible and to collect as many monetary rewards as possible. The instructions did not explain the stimulus-response-outcome contingencies or inform about the different conditions (i.e., compatible, incompatible, neutral, and free trials). Each trial consisted of an inter-trial-interval (black mid-screen fixation cross on a white background, 2000 ms), a compound object-color stimulus, and a response-dependent outcome stimulus.

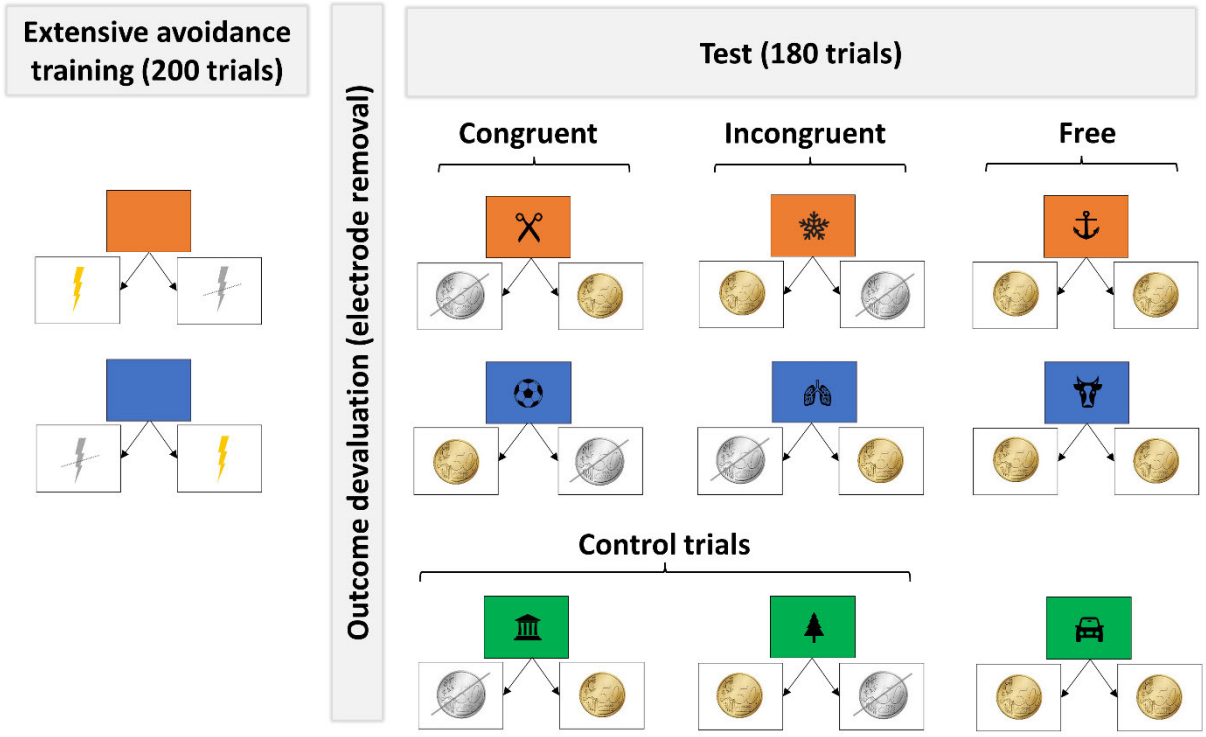
Nine object-color stimuli were used in the test phase, which each consisted of one of nine neutral object pictures displayed horizontally and vertically centered on top of one of three background colors, two of which had already been presented in the training phase (i.e., orange and blue). The third background color (i.e., green) had not been presented before and thus had not been associated with any avoidance response (see Figure 1). The object pictures were simple, black, symmetric depictions of objects (anchor, ball, car, cow, house, lungs, scissors, snowflake, and tree; all widths: 2.1-2.4 cm, all heights: 2.0-2.5 cm). Each of the three background colors was paired with three object pictures. The combinations between colors and pictures were counterbalanced between participants to prevent idiosyncratic stimulus effects. Importantly, the background colors did not predict the outcomes during the test phase. Instead, the outcomes were associated with the object pictures. The object-color stimuli were presented in each trial until a button was pressed, or, if no button was pressed, for a maximum duration of 1000 ms. Following correct responses, a realistic picture of a 50 Euro cent coin on a white background was presented for 1000 ms (i.e., reward outcome). Following incorrect responses and misses (i.e., no response within 1000 ms), a grey, crossed out 50 Euro cent coin was depicted for 1000 ms (i.e., no reward outcome). After the experiment, the participants were paid 1 Euro cent for each correct approach response but were unaware of this ratio during the experiment.

The nine object-color stimuli were grouped into four conditions depending on the compatibility of the rewarded response with the previously extensively trained responses (see

Figure 1): 1) compatible trials (two object-color stimuli; orange and blue background), 2) incompatible trials (two object-color stimuli; orange and blue background), 3) neutral control trials (two object-color stimuli; green background), and 4) free choice trials (three object-color stimuli; orange, blue, and green background). In *compatible* trials, the reward was presented when the participant pressed the button that had prevented the aversive stimulation for the background color during avoidance training (i.e., responses to gain rewards were compatible with responses to avoid the aversive stimulation during avoidance training). In *incompatible* trials, the reward was presented when participants pressed the button which had *not* prevented the aversive stimulation for this background color during the avoidance training. In *free choice* trials, the reward was presented following any timely response (i.e., pressing any of the two buttons). The rate of compatible responses in free choice trials with background colors which had already been presented during the training phase (i.e., orange and blue) was assumed to indicate the non-costly effect of the training on responding (i.e., habitual responses were not linked to costs). Free choice trials with the green background color were included to balance out the conditions and background colors and were not analyzed. Lastly, in *neutral control* trials, which featured the new full-screen color (i.e., green), the reward was presented when participants pressed the correct button in response to the presented object picture. As the new background color had not previously been associated with an avoidance response, the responses in neutral control trials were not influenced by the prior extensive training. The neutral control trials were included to measure the general acquisition of the stimulus-response-outcome associations in the test phase. The trials were presented in a pseudo-randomized order in ten blocks with 18 trials each (i.e., four compatible, four incompatible, four neutral, and six free trials were presented in randomized order within each block, and the blocks were presented seamlessly).

Figure 1

Schematic depiction of the experimental phases



Note: The arrows depict left and right button presses. The object color stimuli and the aversive stimulation and reward symbols are depicted larger than in the experiment to enhance readability.

Behavioral measures

The dependent variables, 1) reaction time (i.e., the time between the color-object stimulus presentation onset and button press, in milliseconds) and 2) accuracy (i.e., correct response, incorrect response, or missing response), were recorded trial-wise. Reaction time was analyzed as a continuous variable. Accuracy was analyzed as a binary variable (i.e., correct yes/no, whereby "no" includes incorrect responses and response misses). Only reaction times in trials with correct responses were analyzed.

The training phase analysis comprised two dependent variables: response accuracy and reaction time. The test phase responses were analyzed using four dependent variables: First, we compared the accuracy in compatible vs. incompatible trials (i.e., accuracy compatibility effect). A larger accuracy compatibility effect (i.e., higher accuracy in compatible than incompatible trials) was conceptualized to indicate a stronger reliance on the extensively trained responses and, therefore, stronger habitual control. As the accuracy compatibility effect was associated with a monetary loss (i.e., due to lower accuracy in incompatible trials), we

conceptualized it as a costly effect of the extensive training. Second, the reaction times were compared between incompatible and compatible trials (i.e., reaction time compatibility effect). A larger reaction time compatibility effect (i.e., slower responding in incompatible trials) was conceptualized to indicate a stronger tendency to respond with the extensively trained responses. The reaction time compatibility effect was not associated with a direct reduction of money since the amount of the monetary reward was not directly tied to the response speed, unless the response time exceeded 1000 ms (i.e., then, the no-reward outcome was presented). We therefore conceptualized it as a low-cost effect of the extensive training. Third, free trial responses were analyzed regarding their compatibility with the extensively trained responses (i.e., compatible or incompatible with the extensively trained response). The compatibility effect in free trials (i.e., the percentage of compatible responses) can also be described as a low-cost effect since incompatible responses in free trials were not associated with monetary costs. Fourth, responses in neutral control trials were analyzed to assess potential group differences in the approach performance without any compatibility effects.

Questionnaires

For all questionnaires, individual sum scores were obtained. We substituted single missing items with the individual's average of the available questionnaire items if no more than 25% of the questionnaire's items were missing. If more than 25% of items were missing for a participant in a questionnaire, the individual item sum score was treated as missing. Cronbach's α was calculated where appropriate (i.e., for questionnaires with more than one item) using the arsenal R package (v.3.6.3; Heinzen et al., 2021). The anxiety symptom questionnaires were answered at the beginning of the first laboratory session (i.e., one week before the devaluation task).

Patient-Reported Outcome Measurement Information System – Anxiety – Short Form (PROMIS-A-SF). The PROMIS-AS-F v1.0 Anxiety 8a scale assesses the frequency of common anxiety symptoms during the preceding seven days (HealthMeasures, 2019). On eight items, participants indicate how often they experienced the respective anxiety symptom on a scale from 1 ("never") to 5 ("always"). The sum of all items indicates the current frequency of anxiety symptoms. The reliability of the scale has been reported to be excellent ($\alpha = .93$) (Pilkonis et al., 2011). The internal consistency in our sample was excellent ($\alpha = .95$). We used the German translation (Wahl et al., 2011).

DSM 5 Cross-Cutting Dimensional Scale for Anxiety Disorders (DSM–Cross D). The DSM-Cross D measures the severity of general anxiety-related feelings, thoughts, and

behaviors in the last month (LeBeau et al., 2012). On twelve items, participants indicate how often they encountered the respective symptoms during the last four weeks. Good test-retest reliability (ICC = .85) has been reported (Knappe et al., 2014), and the internal consistency in our sample was excellent (α = .94). We used the German translation (Knappe et al., 2014).

Depression, anxiety, and stress scales (DASS-21). The DASS measures the severity of depression, anxiety, and stress during the previous seven days (Lovibond & Lovibond, 1995). On 21 items, participants indicate how often they experienced the respective symptom on a scale from 0 ("not at all") to 4 ("very strongly or most of the time"). The sum scores for the three subscales, depression, anxiety, and stress, indicate the severity of depression and anxiety symptoms and stress load. Good internal consistency has been reported for the anxiety subscale (α = .78 - .82), the depression subscale (α = .91), and the stress subscale (α = .81 - .89; Nilges & Essau, 2015). The internal consistency in our sample was α = .89 for the anxiety subscale, α = .87 for the depression subscale, and α = .86 for the stress subscale. We used the German translation (Nilges & Essau, 2015).

Karolinska Sleepiness Scale (KSS). The KSS is a one-item questionnaire assessing subjective sleepiness (Akerstedt & Gillberg, 1990), which was assessed before and after the experiment. Participants are asked to indicate how sleepy they feel at the current moment, using a ten-point scale from 1 ("extremely alert") to 10 ("very sleepy, cannot stay awake"). The KSS has demonstrated sensitivity to objective individual state variations such as sleep deprivation or night work and correlates with EEG sleepiness indicators (Akerstedt et al., 2014). We used the German version by Popp et al. (2011).

Arousal rating. Subjective ratings of arousal were obtained using self-assessment manikins (Bradley & Lang, 1994) before and after the experiment. The participants are asked to rate their current state on a 9-point scale with five illustrative symbolic figures as anchors, illustrating arousal states from very low (i.e., a score of 1) to very high arousal (i.e., a score of 9).

Motivation ratings and subjective aversiveness ratings. Self-reported motivation to avoid the aversive stimulus (i.e., avoidance motivation) and to approach the rewards (i.e., approach motivation) as well as to respond fast were assessed immediately after the experiment using unvalidated items, each on a visual analogue scale ranging from 0 ("Not motivated at all") to 100 ("Extremely motivated"). The subjective electrotactile stimulus' aversiveness was rated after the experiment on a visual analog scale ranging from 0 ("not unpleasant at all") to 100 ("extremely unpleasant"). The items had been constructed for this study.

Affective responses to monetary gains. We assessed the general emotional responsiveness towards monetary rewards to preclude a bias in the group comparison. On a ten-point Likert scale, the participants indicated how happy, pleasant, and aroused they imagined feeling in the hypothetical scenario of winning two specific amounts of money (1 € and 10 €). Each item was answered on a scale from 1 ("very unhappy/unpleasant", "not at all excited") to 10 ("very happy/pleasant", "very excited"). The six item's scores were averaged to obtain the final individual score. The items had been constructed for this study. The internal consistency across all six items was $\alpha = .89$.

Table 1: *Sample characteristics*

	Patients with anxiety disorder (N = 62)[6]	Healthy controls (N = 62)	Test statistic	p[3]	ES[4]
Age	26.71 (8.60)	26.56 (8.36)	1934[1]	>.999	<0.01[4]
Gender	-	-	0[2]	>.999	0.02[5]
(% female gender)	41 (66%)	41 (66%)	-	-	-
(% male gender)	21 (34%)	21 (34%)	-	-	-
Subjective aversiveness (post) (VAS, 0-100)	52.50 (23.43)	51.23 (21.17)	1998[1]	>.999	0.04[4]
Anxiety symptoms (DASS, 0-21)	9.23 (4.75)	1.89 (2.32)	3578[1]	**<.001**	0.75[4]
Anxiety symptoms (PROMIS, 0-72)	18.15 (6.39)	4.98 (3.87)	3600[1]	**<.001**	0.78[4]
Anxiety symptoms (DSM Cross-D, 0-48)	19.23 (9.52)	4.63 (4.30)	3611[1]	**<.001**	0.76[4]
Depression symptoms (DASS, 0 - 21)	5.72 (4.60)	2.68 (2.50)	2740[1]	**<.001**	0.37[4]
Arousal (pre) (SAM, 0 - 9)	4.55 (1.58)	3.63 (1.58)	2523[1]	**.020**	0.28[4]
Arousal (post) (SAM, 0 - 9)	4.10 (1.72)	3.60 (1.79)	2252[1]	.672	0.15[4]
Sleepiness (pre) (KSS, 1 - 10)	5.42 (1.71)	5.10 (1.86)	2086[1]	>.999	0.08[4]
Sleepiness (post) (KSS, 1 - 10)	5.50 (1.77)	4.90 (2.06)	2264[1]	.672	0.16[4]
General reward sensitivity (pre) (VAS, 1 - 10)	6.12 (1.95)	6.66 (1.64)	1598[1]	.836	0.13[4]
Avoidance motivation (post) (VAS, 0 - 100)	65.65 (30.72)	67.94 (27.30)	1928[1]	>.999	<0.01[4]
Approach motivation (post) (VAS, 0 - 100)	60.44 (31.98)	84.24 (18.05)	1030[1]	**<.001**	0.41[4]

Notes. [1]Wilcoxon's *W*. [2]Pearson's Chi-squared test. [3]Bonferroni-Holm corrected. [4]Wilcoxon effect size for independent samples. [5]Phi effect size. [6]N = 61 for the general reward sensitivity score and for the PROMIS sum score.

Hypotheses

We tested the main hypotheses that in the entire test phase, 1) the patient group would display a larger accuracy compatibility effect than the control group (i.e., stronger costly habitual avoidance), 2) the patient group would display a larger reaction time compatibility effect than the control group (i.e., stronger low-cost habitual avoidance), and 3) the patient group would display a higher proportion of training-compatible responses in free trials than the control group (i.e., stronger low-cost habitual avoidance). Although we did not formulate a priori hypotheses about the responses in the training phase, we also analyzed potential group differences in accuracy and response times during the training phase to rule out systematic training differences.

Exploratory analyses and replication analysis

In addition to the group comparison hypotheses, we first compared the subgroup of unmedicated patients and their matched controls to rule out systematic effects of current psychopharmacological treatment (see Supplement B). Secondly, we explored the hypotheses with a dimensional anxiety symptom strength measure instead of group as a predictor variable (see Supplement C). Thirdly, we exploratorily compared the responses of patients with a primary social anxiety disorder diagnosis with those with a primary panic disorder diagnosis (each $n = 21$; see Supplement E).

We also explored whether the compatibility effects were intercorrelated and whether accuracy in neutral control trials was systematically associated with the compatibility effects. We calculated the correlations between the accuracy in neutral trials, the accuracy compatibility effect, the reaction time compatibility effect, and the compatibility effect in free trials as Spearman correlations with Bonferroni-Holm corrections.

We additionally aimed to replicate the results from an earlier study with a very similar experimental task (see Experiment 1 in Glück et al., 2021). Details and results of this analysis can be found in Supplement F.

Data analysis

Data exclusion criteria. Trials with reaction times below 100 ms were excluded from data analysis (i.e., responses in these trials were treated as missing data) as they were not assumed to reflect voluntary movements; this applied to 38 trials or 0.05% of all trials.

Statistical modeling. To account for the longitudinal clustering of the responses within individual participants, the data were analyzed with linear mixed effects models (LMMs) and generalized linear mixed effects models (GLMMs). Mixed models can be described as a hierarchical system of regression equations with randomly varying coefficients (i.e., random effects) and randomly varying regression slopes (i.e., random slopes; Hox et al., 2018). As the data were clustered within blocks and individual participants, we added a random intercept for participant and a random linear and quadratic slope for block in each model, which was considered the maximal random effects structure (Barr et al., 2013). When convergence problems were encountered (i.e., in several models with the continuous anxiety predictor), the random slopes for block were removed (see Brauer & Curtin, 2018; Meteyard & Davies, 2020), resulting in a simple random effects structure with a random intercept for each participant. When this simple random effects structure was applied, this is stated in the respective description of the analysis in the supplementary information. Reaction times (i.e., an approximately normally distributed continuous outcome) were modeled with LMMs. Accuracy data (i.e., a binary outcome) were modeled with GLMMs.

Data processing and analyses were conducted using R statistical software (v.4.2.2; R Core Team, 2022). All statistical models were estimated using the restricted maximum likelihood approach (REML) with the optimizer bobyqa (Bound Optimization by Quadratic Approximation) and a maximum number of $2e^5$ iterations using the lme4 R package (v.1.1-28; Bates et al., 2015). The significance of single regression coefficients was tested using Type III sums of squares ANOVAs (i.e., Wald test) using the car R package (v.3.1-2, Fox & Weisberg, 2019). Collinearity and the normality of the random effects were assessed in each model with the performance R package (v.0.10.4; Lüdecke et al., 2021). The normality of the distribution of the residuals was assessed for each LMM by visually inspecting the QQ plot and for each GLMM by visually inspecting the binned residual plot (see Gelman & Hill, 2006) using the performance R package (v.0.10.4; Lüdecke et al., 2021) and the arm R package (v.1.13-1; Gelman et al., 2022). In the models for the responses in the training phase, we included the fixed effects *block* (i.e., 1 to 25) and *group* (i.e., patients vs. controls). In the models for the responses in the test phase, we included the fixed effects *condition* (i.e., compatible vs. incompatible), *block* (i.e., 1 to 10), *group* (patients vs. controls), and the respective two- and three-way interactions. In all models, we modeled the differences across blocks with a linear and a quadratic trend and the differences between the conditions and groups with sum contrasts as recommended for the direct comparison of two conditions or groups (see Schad et al., 2020). Post-hoc comparisons of the estimated marginal means and the linear and quadratic trends were

performed with the emmeans R package (v.1.8.8; Lenth et al., 2023). The significance of the contributions of the predictors is reported in detail in Supplement A. To improve the interpretability of null effects, we added a post-hoc Bayesian analysis (see Krypotos et al., 2017). Using the brms R package (v.2.20.1; Bürkner, 2017), we estimated models for each outcome with the same fixed and random effects as in the LMMs and GLMMs. In each model, we used an uninformative prior for all parameters, with a Cauchy distribution, mean of 0, and standard deviation of 1. The models were estimated with the default of 2000 iterations per chain and four chains. After model estimation, Bayes' factors were estimated for the main effects and interaction effects of *group*, *condition*, and *block* (linear and quadratic trend). We report Bayes' Factors comparing the probability of the H0 (i.e., the effect is equal to zero) to the H1 (i.e., the effect is unequal to zero). Bayes' Factors can take values between approximately 0 and infinitely positive. A Bayes factor (BF_{01}) of 1 indicates that the data are equally likely under the H0 and the H1, which is very low evidence for either of the hypotheses. A BF_{01} of 5, for example, indicates that the data are five times more likely under the H0 than under the H1. A BF_{01} of 0.2 (i.e., 1/5) indicates that the data are five times more likely under the H1 than under the H0. In addition to the Bayes' factors, we report parameter estimates with 95% credibility intervals (see Bürkner, 2017; Krypotos et al., 2017). The data and the analysis code are available at https://osf.io/nr28s/.

3.3 Results

Training phase

Accuracy

In the GLMM with the predictors *block* and *group* and their interaction, the *block* x *group* interaction was a significant predictor, $\chi^2(2) = 7.59$, $p = .023$, with a stronger linear trend in the anxiety group than the control group (difference estimate = 0.05, $SE = 0.02$, $p = .020$, $BF_{01} = 0.97$, estimate = -2.22, 95% CI [-21.70; 3.87]), while the quadratic trend did not differ between the groups (difference estimate = -0.002, $SE = 0.001$, $p = .138$, $BF_{01} = 1.15$, estimate = 0.74, 95% CI [-3.84; 11.25]). A visual inspection (see Figure 2) indicated that the accuracy differed most strongly between the groups in the first block of training ($M_{\text{Anxiety group}} = 62.5\%$, $SD_{\text{Anxiety group}} = 37.1$, $M_{\text{Control group}} = 75.0\%$, $SD_{\text{Control group}} = 18.5$). Post-hoc comparisons of the estimated marginal means indicated that the overall accuracy was not significantly lower in the anxiety group than in the control group ($OR = 0.92$ [0.73, 1.17], $BF_{01} = 1.70$, estimate = 0.31, 95% CI

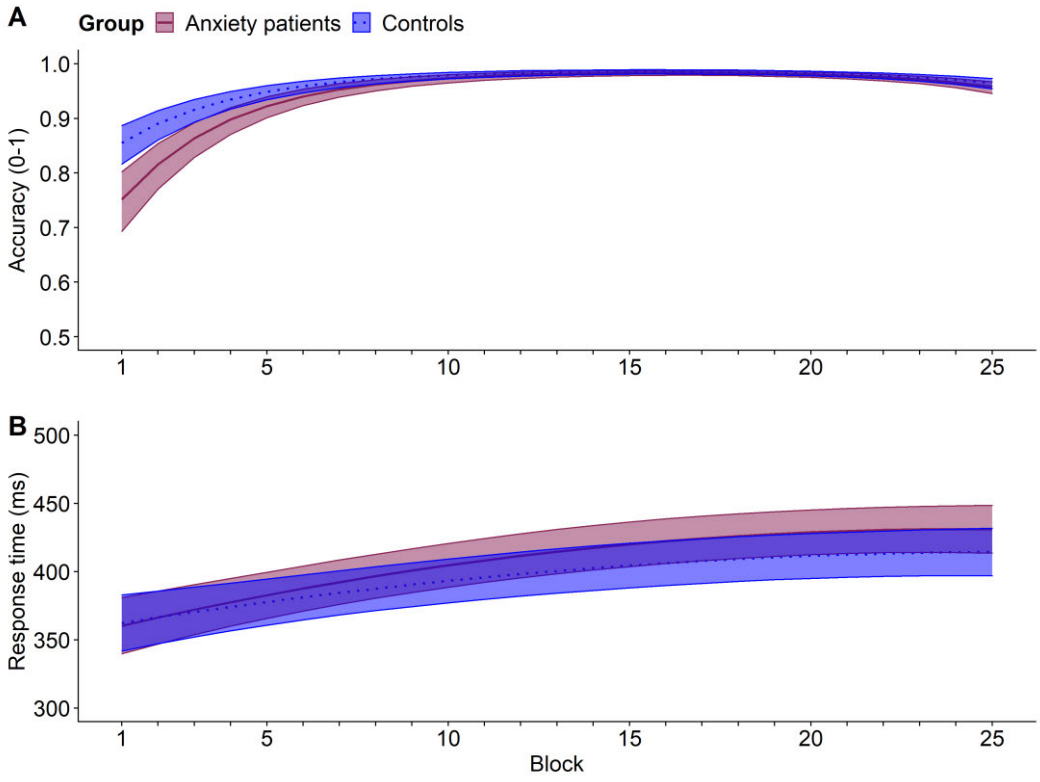
[-0.03; 0.64]). The descriptive overall accuracy was *M* = 92.9% (*SD* = 6.24) in the anxiety group and *M* = 94.6% (*SD* = 5.70) in the control group.

Reaction time

The LMM with the predictors *block*, *group,* and their interaction resulted in a significant prediction of *block*, $\chi^2(2) = 99.23$, $p < .001$, with a significant linear trend (estimate = 2902, 95% CI [2330; 3474], $BF_{01} < 0.01$), and quadratic trend (estimate = -787, 95% CI [-1187; -386], $BF_{01} = 0.23$), indicating that the initial deceleration of the responses decreased throughout the test phase (see Figure 2). The interaction of *group* and *block* did not significantly predict reaction times in the LMM, $\chi^2(2) = 2.31$, $p = .315$. However, the Bayesian analysis indicated moderate evidence for a group effect which increased over time (*group* x *block (linear):* $BF_{01} = 0.34$ (estimate = 0.58, 95% CI [-9.13; 15.89]), *group* x *block (quadratic)*: $BF_{01} = 0.36$, estimate = 0.12, 95% CI [-9.69; 10.97]). *Group* as a single factor did not significantly predict reaction times, χ^2 (1) = 1.07, $p = .300$, $BF_{01} = 0.35$, estimate = -0.86, 95% CI [-10.83; 4.59].

Figure 2

Estimated marginal means in the models to predict accuracy (A) and response time (B) during the training phase



Note. Significance bands display 95% confidence intervals.

Test phase

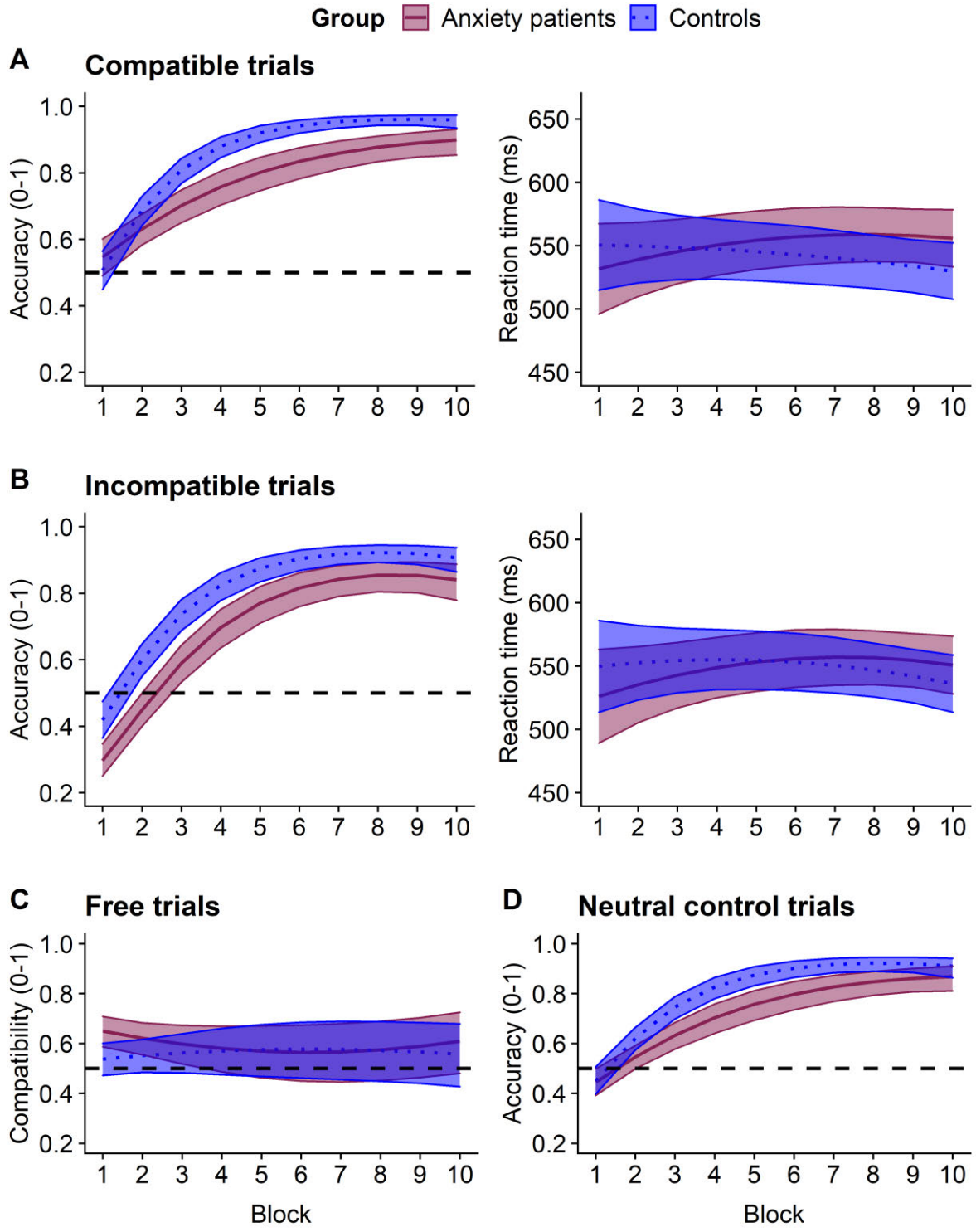
Accuracy

In the GLMM with *condition* (compatible vs. incompatible), *group* (anxiety group vs. control group), *block* (1 - 10), and their interactions as fixed factors, the three-way interaction between *group, block,* and *condition* was a significant predictor, indicating that the accuracy compatibility effects in the two groups varied differently throughout the task, $\chi^2(2) = 17.32$, $p < .001$, BF_{01} (linear effect of block) = 0.42, estimate = -2.11, 95% CI [-17.59; 2.49], BF_{01} (quadratic effect of block) = 0.41, estimate = 0.99, 95% CI [-5.55; 16.88]. Following this significant three-way interaction, we computed separate models with the fixed effects *group, block,* and their interaction for each of the two conditions. The interaction between *group* and *block* significantly predicted accuracy in compatible trials, $\chi^2(2) = 13.37$, $p < .001$, but not incompatible trials, $\chi^2(2) = 0.58$, p .749, indicating that the trajectory of accuracy differed between the groups within the compatible but not within the incompatible condition (see Figure 3). Post-hoc comparisons of the estimated marginal means indicated lower accuracy in the anxiety group both in compatible trials (*OR* = 0.69, 95% CI [0.52; 0.90]) and incompatible trials (*OR* = 0.70, 95% CI [0.54; 0.92]). As an additional follow-up analysis on the significant three-way interaction between *group, block,* and *condition,* we computed one separate model with the predictors *block, condition,* and their interaction for each of the two groups. The interaction between *condition* and *block* was a significant predictor within the anxiety group, $\chi^2(2) = 19.50$, $p < .001$, but not within the control group, $\chi^2(2) = 3.99$, $p = .136$. To pinpoint the blocks where the compatibility effects significantly differed between the groups, we performed post-hoc Bonferroni-Holm corrected Wilcoxon tests with the average accuracy difference between compatible and incompatible trials in each *block* as the dependent variable and *group* as the independent variable. The only group difference approaching significance was found in the second block, where the accuracy compatibility effect tended to be larger in the anxiety group ($Median_{Anxiety\ group}$ = 25.0%, $Median_{Control\ group}$ = 0%, $W = 2466.5$, $p = .055$). In all other blocks, the compatibility effect did not significantly differ between the groups, all $W < 2178$, all $p > .999$. The Bayesian analysis indicated robust evidence for the null effect of the *group* x *condition* interaction ($BF_{01} = 5.12$, estimate = -0.04, 95% CI [-0.24; 0.17]), and for the effects of *group* ($BF_{01} = 0.12$, estimate = 0.43, 95% CI [0.11; 0.82]) and *condition* ($BF_{01} < 0.01$, estimate = -0.45, 95% CI [-0.59; -0.32]). These results indicate a generally higher accuracy in the control group both in compatible and incompatible trials and a tendency towards a slightly

more substantial accuracy compatibility effect in the anxiety group at the beginning of the phase, but no overall stronger compatibility effect in the anxiety group.

Figure 3

Estimated means for the LMMs and GLMMs to predict responses in compatible trials (A), incompatible trials (B), free trials (C) and neutral control trials (D)



Note. Significance bands display 95% confidence intervals. Horizontal dashed lines indicate 50% accuracy.

Reaction times

In the LMM for reaction times in the test phase with *group*, *condition*, *block*, and all their interactions as predictors, none of the predictors was significant, all *p*s ≥ .068 (see Figure 3). Thus, our hypothesis that participants in the anxiety group would display a larger reaction time compatibility effect in the entire test phase was not supported. The Bayesian analysis indicated inconclusive evidence for all predictors (i.e., three-way interaction between *group*, *condition*, and *block* (linear effect): BF_{01} = 0.97, estimate = 0.35, 95% CI [-13.45; 13.40]; three-way interaction between *group*, *condition*, and *block* (quadratic effect): BF_{01} = 1.15, estimate = -0.45, 95% CI [-23.60; 12.09]; interaction between *group* and *condition*: BF_{01} = 0.77, estimate = 2.24, 95% CI [-1.56; 10.31]; interaction between *group* and *condition*: BF_{01} = 1.05, estimate = -0.81, 95% CI [-11.67; 4.89]; effect of *group:* BF_{01} = 1.30, estimate = 0.47, 95% CI [-2.12; 3.95]. Therefore, the GLMM and the Bayesian analysis produced comparable results, but the Bayesian analysis indicated that the evidence for the null effects was inconclusive.

Free trials

In the GLMM to model the compatibility of responses in free trials with *block*, *group*, and their *interaction* as predictors (see Figure 3), none of the predictors was significant, all *p*s ≥ .166. The Bayesian analysis indicated inconclusive results (i.e., BF_{01} = 0.54, estimate = 0.26, 95% CI [-3.83; 5.96] for the interaction between *group* and *block* (linear effect); BF_{01} = 0.44, estimate = -1.25, 95% CI [-7.92; 2.20] for the interaction between *group* and *block* (quadratic effect); BF_{01} = 1.01, estimate = -0.30, 95% CI [-0.68; 0.11] for the effect of *group*. These results do not indicate that the compatibility effect in the anxiety group was larger than in the control group. However, the Bayesian analysis indicated inconclusive evidence for the null effects.

Neutral control trials

General learning differences between the groups were tested with a GLMM with accuracy in the neutral control trials as criterion and *group*, *block*, and their *interaction* as predictors (see Figure 3). The interaction between *group* and *block* was a significant predictor, $\chi^2(2) = 7.90$, $p = .019$, with a stronger quadratic trend in the control group than in the anxiety group (difference estimate = 0.30, *SE* = 0.01, $p = .009$, BF_{01} = 0.09), but without a group difference in the linear trend (difference estimate = -0.046, *SE* = 0.037, $p = .217$, BF_{01} = 0.48), indicating that accuracy increased more sharply in the control group than in the anxiety group and then reached a plateau (see Figure 3). Post-hoc comparisons of the estimated marginal means indicated that the general accuracy was lower in the anxiety group than in the control group (*OR* = 0.76, 95% CI [0.60,

0.92], BF_{01} = 0.27). Therefore, the GLMM and the Bayesian analyses indicated robust evidence for a more pronounced quadratic trend in the control group (see Figure 3) and moderate evidence for a group difference.

Correlations between compatibility effects

The accuracy in neutral control trials did not correlate with the accuracy compatibility effect, $r(122) = -0.05$, $p > .999$, with the reaction time compatibility effect, $r(122) = .08$, $p > .999$, or with the compatibility effect in free trials, $r(122) = .04$, $p > .999$. The compatibility effects intercorrelated highly: the accuracy compatibility effect correlated with the reaction time compatibility effect, $r(122) = .57$, $p < .001$, and with the overtraining effect in free trials, $r(122) = .64$, $p < .001$. The reaction time compatibility effect and the overtraining effect in free trials were also correlated, $r(122) = .44$, $p < .001$. These results indicate that, while the compatibility effects overlapped, the performance in neutral trials was not associated with the individual tendency towards the repetition of the previously favorable responses.

Comparison between social anxiety disorder and panic disorder

The exploratory subgroup comparison revealed a larger accuracy compatibility effect as indicated by a significant interaction between *condition* and *subgroup*, $\chi^2(1) = 4.97$, $p = .026$ (see Supplementary Table E.4), BF_{01} = 0.54, estimate = -0.34, 95% CI [-0.64; -0.03], and a larger compatibility effect in free trials as indicated by a significant effect of *subgroup*, $\chi^2(1) = 34.30$, $p < .001$ (see Supplementary Table E.6), $BF_{01} < 0.01$, estimate = 0.58, 95% CI [0.39; 0.77], in the panic disorder subgroup than the social anxiety disorder subgroup (both $n = 21$). The panic disorder subgroup also showed generally slower responses both in the training and the test phase as indicated by significant effects of *subgroup* in the LMMs, $BF_{01} < 0.01$ (estimate = 28.35, 95% CI [23.28; 33.43]) and BF_{01} = 0.49 (estimate = 4.74, 95% CI [-1.60; 19.90]), respectively (see Supplementary Tables E.3 and E.5 for the LMM results). Neither the general accuracy in compatible and incompatible trials (BF_{01} = 0.94, estimate = 0.24, 95% CI [0.02; 0.46]) nor the accuracy in neutral trials (BF_{01} = 7.96, estimate = 0.05, 95% CI [-0.16; 0.26]) or during the training (BF_{01} = 2.35, estimate = 0.12, 95% CI [-0.06; 0.29]) differed between the panic disorder subgroup and the social anxiety disorder subgroup (see Supplementary Tables E.4 and E.7 for the GLMM results). Also, the two subgroups did not differ in any psychological symptom measures or their motivation to avoid and approach. However, the average age in the panic disorder subgroup (M = 31.33 years, SD = 11.72) was significantly higher than in the

social anxiety subgroup (M = 23.38 years, SD = 5.32), W = 88, p = .013 (see Supplementary Table E.1).

3.4 Discussion

In this outcome devaluation study, extensive avoidance training impacted the approach of small monetary gains after the devaluation of the aversive outcome. Specifically, correct approach responses after the devaluation were more frequent when compatible with the extensively trained avoidance responses than when incompatible with it. This compatibility effect persisted during the entire test phase. We observed sustained carry-over effects of the extensive avoidance training on the post-devaluation approach responses and substantial correlations between the different habit indicators, thereby replicating critical results from a previous study (Glück et al., 2021), underlining the effectiveness of the extensive training to induce a stable tendency to respond with the previously favorable responses.

The general effect of the extensive avoidance training was not constantly stronger in patients with social anxiety disorder, agoraphobia, or panic disorder than in healthy participants. Although the effect of the extensive avoidance training on the accuracy compatibility effect tended to be slightly more pronounced in the patient group briefly after the devaluation (i.e., significant three-way interaction between group, block, and condition), this was only transient and not supported by other indices for habitual responding. Neither the frequency of compatible responses in free trials nor the reaction time compatibility effect differed between the groups. The patient group showed consistently reduced accuracy in compatible, incompatible, and neutral learning trials, suggesting a generally lower level of approach. However, the level of general approach was unrelated to the compatibility effects, suggesting that a lower level of approach was not related to stronger habitual control. In support of the robustness of the analyses, we found the same effects in the subsample of unmedicated participants and a very similar picture in an analysis using anxiety as a dimensional variable.

Briefly after the outcome devaluation, we observed a descriptive but non-significant tendency towards a stronger accuracy compatibility effect in the anxiety group. However, this transient tendency may reflect both a stronger use of habitual action control and a stronger strategical reliance on the previously favorable responses in the anxiety group. This ambiguity (i.e., in interpreting the compatibility effects shortly after the devaluation) results from the lack of instructions about the correct response for each object-color stimulus after the devaluation procedure. Therefore, the participants needed to learn the correct responses by trial and error.

In this situation of high ambiguity and uncertainty, participants in the anxiety group may have been more prone to strategically rely on the previously successful response (e.g., Smith et al., 2016; Wong & Lovibond, 2021). The descriptive accuracy compatibility effect group difference at the beginning of the test phase, therefore, cannot be interpreted as a clear sign of more habitual avoidance in the anxiety group. Additionally, to conclude that anxiety disorders were associated with more strongly acquired habitual avoidance, more pronounced group differences or effects of the continuous anxiety score on the compatibility effects would be expected throughout the entire test phase. In later trials of the test phase, when the stimulus-response-outcome contingencies are acquired, any compatibility effects should not result from an explicit strategy to rely on the learned associations between colors and responses in the training phase due to trial-and-error learning but instead can be expected to reflect habitual tendencies to press the extensively trained button when encountering the respective color. As we did not observe any compatibility effects at the end of the test phase, we conclude that the most likely interpretation of the results is that they provide no evidence for a more substantial development of habitual avoidance in anxiety disorders in this outcome devaluation task.

The lower accuracy in neutral trials and the general lower accuracy in compatible and incompatible trials in patients with anxiety disorders may reflect a general impairment of learning the stimulus-response-outcome contingencies, potentially due to effects of anxiety on attention or memory processes (e.g., Eysenck & Calvo, 1992; Eysenck et al., 2007). However, the general accuracy differences between the groups in the test phase may also result from an overall reduced pursuit of rewards in individuals with anxiety disorders (e.g., Pittig, Boschet, et al., 2021), potentially due to motivational factors. In support, the task-specific self-rated approach motivation was lower in the anxiety group than in the control group, which may have negatively affected the overall approach performance in the anxiety group in the test phase. The task-specific approach motivation was, however, assessed after the task and may therefore be confounded, for example, by the perception of the test phase as demanding or unpleasant. Of note, the general reward sensitivity, which was measured before the task, did not differ between the groups, indicating that the reduced approach in the anxiety group is unlikely caused by differences in general reward sensitivity. Notably, the general accuracy in the test phase did not correlate with the compatibility effects. Thus, the general approach performance in neutral trials was unrelated to the specific compatibility effects resulting from the extensive training. Therefore, it is unlikely that the general learning ability differences systematically biased the compatibility effects. Notably, the anxiety group also showed a lower accuracy in the beginning

of the training phase, which indicates a slightly lower acquisition of the stimulus-response-outcome associations during the avoidance training. This group difference, however, disappeared in later stages of the test phase, and therefore unlikely systematically impacted the test phase responses in the end of training. Additionally, the overall accuracy during the training phase was high in both groups, indicating a stable acquisition of the avoidance responses in the entire sample.

The null differences in habitual avoidance between the groups in this study and the inconclusive evidence landscape do not indicate a clear, direct association between anxiety and habitual avoidance. Anxiety may, however, impact habitual avoidance in a more complex form. Several potentially moderating or mediating factors that may impact the potential relationship between anxiety and the acquisition of habitual avoidance can be concluded from the existing literature. For example, trait anxiety has been demonstrated to impair general working memory task performance (Moran, 2016; Ward et al., 2020), potentially reducing the working memory capacity available for goal-directed control processes. Working memory capacity may thus moderate or mediate the relation between anxiety and habitual control. Taking a similar perspective, Berggren and Derakshan (2013) proposed that anxiety may be associated with a less efficient inhibition of automatic cognitive processes. They proposed that, in easy tasks, higher anxious individuals may compensate for this inefficiency by more strongly engaging in the task leading to uncompromised performance effectiveness. However, in cognitively demanding tasks, highly anxious individuals may be unable to compensate, and therefore show reduced performance (Berggren & Derakshan, 2013). This assumption is backed up by findings indicating that working memory moderates the detrimental effect of acute stress on model-based action control (Otto, Raio, et al., 2013) and goal-directed control (Quaedflieg et al., 2019). In these studies, acute stress in individuals with low working memory capacity was especially detrimental to goal-directed control (but see Patterson et al., 2019, where distraction during the task did not moderate the association between early life stress and habitual avoidance). The task difficulty in the test phase of the current study was relatively low, as it involved only two response options and nine stimuli-response-outcome associations with stable contingencies. Therefore, working memory load during the test phase may have been low, potentially enabling the participants in the anxiety group to compensate for deficits in inhibiting the extensively trained responses. Systematic variations of task difficulty or a baseline working memory assessment may help to assess working memory dependent effects of anxiety on action control.

The accumulating null findings regarding elevated habitual responding in patients with elevated trait anxiety (Gillan et al., 2021; Patterson et al., 2019) or an anxiety disorder (Roberts et al., 2022) may encourage adaptations of the existing theoretical approaches concerning the role of habitual avoidance in anxiety disorders (e.g., LeDoux & Daw, 2018; Pittig et al., 2020). So far, the evidence does not imply a stronger tendency to shift from goal-directed towards habitual avoidance as an essential factor in developing or maintaining maladaptive avoidance in anxiety disorders. However, the theoretical propositions did not explicitly state a faster acquisition of avoidance habits in individuals with anxiety disorders. Indeed, habitual avoidance in individuals with anxiety disorders may also develop from a long individual history of goal-directed avoidance. From this point of view, habitual avoidance in anxiety disorders would not develop due to an anxiety-specific tendency to develop habitual avoidance but instead result from the elevated frequency of goal-directed avoidance in these disorders (e.g., Pittig, Boschet, et al., 2021). Furthermore, maladaptive avoidance behaviors do not always need to result from habitual control, as maladaptive responses can also result from goal-directed processes. Even objectively maladaptive avoidance behaviors may, for example, result from a goal-directed use of a short-term strategy to regulate distressing emotions (see Buabang, Köster, et al., 2023). Relatedly, the maladaptive avoidance of a stimulus may be the time-dependent or context-dependent residual of avoidance which was adaptive in another situation or time (see Moors et al., 2017). Another alternative hypothesis may state that habitual avoidance is associated differentially with different types of avoidance. Specifically, habitual control may be involved in maintaining excessive active avoidance but not excessive passive avoidance. Roberts et al. (2022) have already discussed this distinction for generalized anxiety disorder. Passive avoidance may, for example, prevail in anxiety disorders with comorbid depressive symptoms, which are characterized by passive rather than active avoidance (e.g., Ferster, 1973; Spielberg et al., 2011). So far, the evidence of an aberrant acquisition of habitual responding is consistent primarily for active maladaptive behaviors such as found in obsessive-compulsive disorder (Gillan et al., 2016; Voon et al., 2015) or substance use disorders (Everitt & Robbins, 2016). Systematic research on habitual control in anxiety patients with active and passive avoidance profiles may enable a more specific understanding of the role of habitual control in avoidance behavior.

In this study, the participants with a diagnosed panic disorder on average showed more pronounced compatibility effects than the participants with a social anxiety disorder diagnosis. This finding should be interpreted cautiously due to the exploratory nature of the analysis and

the small subsample sizes. Additionally, the reaction time differences between the subgroups complicate the straightforward interpretation of the accuracy differences between the groups. During training and in compatible and incompatible trials, the participants in the panic disorder subgroup on average responded more slowly than the participants in the social anxiety disorder subgroup. These reaction time differences may have affected the accuracy compatibility effects (see Proctor et al., 2011). Despite these limitations, which call for a cautious interpretation of the subgroup differences, the compatibility effect differences between the subgroups may be of some relevance for future research since, to our knowledge, this is the first investigation of habitual control in panic disorder. One potential explanation for the more pronounced compatibility effects in panic disorder in our study may lie in the use of electrotactile stimulations as the aversive outcome, which may constitute a more relevant aversive outcome for participants with panic disorder who frequently display a high sensitivity for changes in bodily processes than for participants with social anxiety disorder (e.g., Rudaz et al., 2010). In contrast, participants with social anxiety disorder may react more strongly to aversive outcomes with social content (e.g., Ishikawa et al., 2021). Interestingly, the perceived aversiveness of the electrotactile stimulation did not differ between the two subgroups. However, we did not assess general feelings of stress or negative affect induced by the task, which may also have produced compatibility differences (see Buabang, Boddez, et al., 2023). Larger studies with homogenous samples and disorder-specific aversive outcomes may help to identify potential mechanisms underlying disorder-specific responding in devaluation tasks.

Several limitations characterize the current study. First, as already discussed, the tendency toward larger compatibility effects in the anxiety patient group briefly after the outcome devaluation may indicate both a stronger habitual response and stronger goal-directed use of the extensively trained stimuli-response-outcome associations. Second, we could only exploratorily differentiate between different anxiety disorders due to the relatively low sample size within each diagnostic group. Third, we did not analyze data about the participants' ethnic identification, their cultural background, or their socioeconomic status, which may potentially limit the generalizability of the findings. Fourth, reaction times and accuracy were analyzed separately. Future devaluation studies may implement data analytic strategies that allow modelling reaction times and accuracy jointly to account for potential interactions and to differentiate underlying processes contributing to the adjustment of non-adjustment of responses after outcome devaluation procedures (e.g., drift-diffusion models; for a review see Ratcliff and McKoon, 2008). Fifth, the external validity of experimental paradigms, such as the

outcome devaluation paradigm used in this study, with its predefined response choices and unambiguous outcome contingency schedules, has been questioned both in avoidance research (e.g., Krypotos et al., 2018) and in action control research (Pezzulo & Cisek, 2016). Aiming at a richer understanding of the associations between environmental cues, behavioral responses and outcomes, and anxiety, future studies may, for example, include ecological momentary assessments, virtual reality, or qualitative methods. Sixth, although the Bayesian analyses indicated relatively robust evidence for the absence of a group difference in the accuracy compatibility effect, the evidence for the null effects concerning group differences in the reaction time compatibility effect and the compatibility effect in free trials was inconclusive. Potential reasons for the inconclusive results in these two habit indicators may be a low reliability of these compatibility effects (see Enkavi & Poldrack, 2021) or an insufficiently large sample size. The low conclusiveness may, thus, reflect issues on the reliability of outcome devaluation paradigm measures (see Buabang, Köster, et al., 2023). Potentially, the field may benefit from systematic investigations of the reliability of the indicators used in outcome devaluation studies to enhance the planning of adequate sample sizes.

In summary, this study did not confirm the hypothesis that anxiety disorders are characterized by a stronger tendency to repeat specific, simple avoidance responses habitually. Although the clinical sample in general showed lower levels of approach, various control conditions provided insights that this difference was unrelated to habitual control. Therefore, the study adds a null finding to the literature on a the potential influence of anxiety on the formation of avoidance habits, calling a simple relation between anxiety and the habitual control of avoidance into question. The stronger compatibility effects in patients with panic disorder compared to patients with social anxiety disorder may inform future studies on disorder-specific mechanisms in avoidance control. Also, future studies on the relationship between anxiety and avoidance control may take moderating or mediating variables into account or utilize more naturalistic research designs.

4. The impact of trait anxiety on approach and avoidance habits in an outcome devaluation paradigm

Valentina M. Glück[1] & Andre Pittig[2]

[1]Department of Psychology (Biological Psychology, Clinical Psychology, and Psychotherapy), University of Wuerzburg, Germany

[2]Translational Psychotherapy, Institute of Psychology, University of Goettingen, Göttingen, Germany

Manuscript in preparation

Abstract

Background: Persistent avoidance as a central symptom in anxiety disorders is involved in the maintenance of fear and can decrease individual functioning. Trait anxiety may play a role in a faster acquisition of habitual, inflexible avoidance responses. However, empirical evidence on this mechanism is scarce, and it is unclear whether trait anxiety specifically influences avoidance or inflexible avoidance and approach responses in general. *Methods:* Ninety-five healthy individuals pre-selected for high and low trait anxiety participated in two outcome devaluation tasks with extensive training of approach (i.e., approach task) and avoidance (i.e., avoidance task) in a two-day design. Heart rate variability during a resting period was recorded before each task and included as a exploratory predictor of habitual responding. *Results*: Higher trait anxiety task-independently predicted stronger habitual responding in one of three indicators (i.e., the reaction time compatibility effect), which was task-independent. Trait anxiety also tended to predict stronger approach than avoidance habits on another indicator (i.e., the accuracy compatibility effect) but this effect only approached significance. Independently of trait anxiety, habitual responses were generally apparent in the test phase of the approach task but not the avoidance task. Self-reported retrospective stress levels were higher after the approach task's test phase than for the avoidance task's test phase. Heart rate variability did not predict habit strength. *Discussion:* Trait anxiety was not associated with a specific increase of avoidance habits, but predicted a small and unspecific increase of low-cost habitual responses. Additionally, trait anxiety tended to predict stronger costly approach habits as indicated by the accuracy compatibility indicator. Overall, habitual control was less pronounced in the avoidance task phase than in the approach task. The stronger habit effect in the approach task may result from higher perceived stress levels in this task's test phase. The results underline that analyzing reaction times in action control experiments on trait anxiety may be beneficial, and that outcome devaluation paradigms are sensitive to slight variations of the experimental design.

Key words: Trait anxiety, avoidance, approach, habits

4.1 Introduction

Flexibly approaching rewards and avoiding threats is essential in volatile environments that humans live in. The underlying action control processes can involve internal goals, i.e., can be based on response-outcome associations, or can be independent of internal goals, as is assumed by the associative dual-process model of action control (Adams & Dickinson, 1981). In this model, habitual responses are performed without the involvement of internal goals and are instead the result of underlying direct, specific associations between stimuli and responses. Since habitual action control is based on direct stimulus-response associations, it is assumed to put a low load on cognitive resources (e.g., working memory or attentional control) and to produce adaptive responses in environments where the associations between stimuli, responses, and outcomes are stable (Wood & Rünger, 2016). However, behavior guided by direct stimulus-response associations is assumed to produce inflexible responses in more volatile environments. This inflexibility has been frequently discussed as potentially contributing to maladaptive approach (e.g., Belin et al., 2013; Everitt & Robbins, 2016) and avoidance (e.g., Arnaudova et al., 2017; Ball & Gunaydin, 2022; LeDoux et al., 2017; Pittig et al., 2020). Evidence on individual characteristics which may contribute to the acquisition of habitually controlled behaviors may help to develop more precise clinical models involving action control processes (Verhoeven & Wit, 2018).

Persistent, inflexible avoidance influences the development and maintenance of anxiety disorders via several pathways (Krypotos et al., 2015; Pittig et al., 2020). Avoidance is not always controlled by explicit fear evaluations (van Uijen et al., 2018; Vervliet & Indekeu, 2015), and the flexibility of avoidance behaviors is reduced in anxiety disorders (Pittig, Boschet, et al., 2021). More pronounced goal-directed avoidance of threats despite loss of rewards (i.e., costs) and more persistent goal-directed avoidance when the aversive outcome was not presented anymore characterized individuals with anxiety disorders in one study (Pittig, Glück, et al., 2021). Avoidance can be performed irrespective of anxiety levels (Pittig, Glück, et al., 2021), indicating that the inflexible regulation of avoidance is not always driven by fear. One potential mechanism to explain this inflexibility may lie in a transition from goal-directed to habitual avoidance in anxiety disorders (Arnaudova et al., 2017; LeDoux et al., 2017; Pittig et al., 2020).

If a stronger transition to habitual avoidance is a risk factor for anxiety disorders, elevated transition may already be evident in individuals at risk who have not yet developed a disorder.

We were therefore interested in examining whether a well-established risk factor for anxiety disorders, trait anxiety, elevates the acquisition of habitual control in general, or is specifically associated with elevated acquisition of habitual avoidance. High trait anxiety is defined as an individual's propensity to frequently experience episodes of acute anxiety and is a vulnerability factor in the etiology of anxiety disorders (Mineka & Oehlberg, 2008).

Habitual approach and avoidance acquisition in highly trait-anxious individuals has been examined in several experimental studies involving various operationalizations of habitual avoidance and trait anxiety. Trait anxiety, as measured with the State-Trait Anxiety Inventory-Trait (STAI-T, Spielberger et al., 1983) was not associated with a stronger model-free approach in a sequential reinforcement learning task (Gillan et al., 2016). No effects of experimentally induced acute anxiety or general anxiety as indicated by self-reported recent panic attacks on habitual approach were reported in another study (Gillan et al., 2021). Studies investigating habitual avoidance in trait-anxious individuals as measured with the STAI-T reported null results (Gillan et al., 2014; Patterson et al., 2019) or only found an effect of trait anxiety when not controlling for intolerance of uncertainty (Flores et al., 2018). Thus, the evidence of an impact of trait anxiety on action control is mixed for habitual avoidance, while the null results for habitual approach are relatively consistent. In addition, two recent studies reported no elevated habitual avoidance in generalized anxiety disorder (Roberts et al., 2022) or social anxiety disorder, panic disorder, or agoraphobia (Glück et al., 2023). So far, no study has systematically assessed potential effects of trait anxiety on approach and avoidance habit acquisition using parallel tasks. Additionally, no experimental study has compared the acquisition of approach and avoidance habits.

As one physiological correlate of trait anxiety, lower heart rate variability has been reported in higher trait anxiety (Riganello et al., 2012) and in individuals with anxiety disorders (Alvares et al., 2013; Chalmers et al., 2014; Pittig et al., 2013; Riganello et al., 2012). Heart rate variability (i.e., variations of time intervals between adjacent heartbeats) is regulated by the autonomic nervous system (Kim et al., 2018), which is central to adaptations to environmental demands (Friedman, 2007). Therefore, a high heart rate variability is considered an adaptive individual characteristic (Chalmers et al., 2014; Friedman, 2007). To date, no studies on a potential association between heart rate variability and habitual control are available.

The balance between habitual and goal-directed control can be examined with variations of the outcome-devaluation paradigm (e.g., Adams & Dickinson, 1981; Valentin et al., 2007). Devaluation paradigms typically consist of three phases: First, two simple instrumental

behaviors that produce defined outcomes in response to two stimuli are extensively trained to establish direct stimulus-response associations. Second, one of the outcomes is devalued by manipulating the outcome value or the response-outcome contingency (Wood & Rünger, 2016). Third, the stimuli from the first phase are presented again, and the participants are free to respond or not to respond to them. Suppose the response rate to the still devalued outcome is higher than to the devalued outcome. In that case, action control is assumed to be sensitive to the environmental contingencies (i.e., for devaluations that manipulate the contingency) or the outcome value (i.e., for devaluations that manipulate the outcome value). In that case, goal-directed action control is inferred. In contrast, if the response rate is *not* adjusted (i.e., same observed number of responses to valued and devalued stimuli), habitual, outcome-insensitive control is inferred (at least to some extent) (Adams & Dickinson, 1981). The validity of this operationalization has, however, been questioned (e.g., de Houwer et al., 2022; Moors et al., 2017; Watson & Wit, 2018): non-adjusted responses after the devaluation, which are assumed to indicate habitual control, do not necessarily result from outcome insensitivity, since they can be guided by goals that are not adequately addressed by the devaluation procedure. For example, non-adjusted responses can arise when participants follow a strategy to save cognitive resources in the test phase when non-adjustment does not create any disadvantages (e.g., de Houwer et al., 2018). When outcome devaluation paradigms do not include any costs for habitual responses, adjusting responses is objectively unnecessary, and alternative goals leading to non-adjustment are especially likely.

To address the potential threats to the internal validity of the task, we used an adapted outcome devaluation paradigm in which the test phase is not carried out in extinction (i.e., without any consequences of actions) but includes preferable outcomes for non-habitual and non-preferable outcomes for habitual responses. Thus, habitual responses in this paradigm are associated with costs. Therefore, the participants can be expected to follow the goal to obtain favorable outcomes in the test phase, which, in incompatible trials, conflicts with potentially acquired habitual response tendencies. The various goals that may induce non-adjustment of responses can, thus, be expected to be narrowed. An additional benefit of outcomes in the test phase is that they allow to infer habitual control from observed behavioral compatibility effects, and not indirectly from a lack of evidence for goal-directed control (i.e., from a null difference between valued and devalued response frequency). Thus, the adjusted paradigm does not rely on null hypothesis testing, which may benefit to the validity of habitual control measures (e.g., de Houwer et al., 2018; Watson & Wit, 2018).

Concerning the potential impact of trait anxiety on action control, three potential associations seem plausible. Firstly, higher trait anxiety may be associated with a more substantial acquisition of habitual approach and avoidance. Such generally more pronounced habit acquisition may, for example, result from unspecific anxiety-related reductions in cognitive flexibility (e.g., Wang et al., 2019) or attentional control inefficiency (Berggren & Derakshan, 2013; Eysenck & Calvo, 1992). An association between trait anxiety and habitual control (i.e., for both approach and avoidance) would support a general role of habitual control in the etiology of anxiety disorders. Secondly, higher trait anxiety may be explicitly associated with a stronger habitual avoidance acquisition. Such a specific pattern may, for example, result from trait anxiety-related biases in the attentional processing of threat-related stimuli (e.g., Cisler & Koster, 2010) and an enhancement of threat-related biases in working memory processing in higher trait-anxious individuals (e.g., Berggren & Eimer, 2021). A specific association between trait anxiety and habitual avoidance would suggest that the association between trait anxiety and anxiety disorders may, potentially, partly be mediated by an elevation of habitual avoidance. Thirdly, trait anxiety may not be directly linked to either habitual avoidance or approach acquisition.

In this study, we aimed to separate the effects of trait anxiety on appetitive and aversive habit acquisition. Specifically, we examined whether trait anxiety enhances the acquisition of habitual avoidance more strongly than the acquisition of habitual approach. In a within-subjects design, participants who had been pre-screened for high interindividual variance in trait anxiety took part in two outcome devaluation experiments, which are commonly used to infer the strength of habitual control. Due to accumulating critique of the standard outcome devaluation paradigm, we used an adapted version that does not rely on null hypothesis testing and incorporates costs for habit-compatible responses in the test phase. We investigated three indicators for habitual responses in the test phase of the devaluation paradigm: 1) the difference in accurate responses in habit-compatible as compared to habit-incompatible trials, 2) the reaction time difference in habit-compatible and habit-incompatible trials, and 3) the proportion of habit-compatible responses in free trials where habit-compatible and habit-incompatible responses both lead to the same outcome. We hypothesized that, independent of trait anxiety, we would observe habitual control as operationalized with these three indicators in both devaluation tasks as a result of the extensive approach and avoidance training. Moreover, we hypothesized that trait anxiety would predict a stronger acquisition of habitual avoidance than

habitual approach. To our knowledge, this is the first study directly comparing habitual approach and avoidance responses in a within-subjects design.

4.2 Methods

Subjects

An a priori sample size estimation had yielded that $N = 98$ participants would be needed to detect a medium accuracy compatibility effect difference ($\eta^2 = 0.25$) between high and low anxious participants in a within-subjects ANOVA with a statistical power of $1- \beta = .80$ and $\alpha = .05$. 104 participants took part in the study. Four participants were excluded because they dropped out after the first experimental session. One participant was excluded because of a training accuracy below 70 % on one of the days. This cut-off was consistently used in earlier studies with similar experimental designs to ensure an adequate intensity of the extensive training (Glück et al., 2021). Two additional participants were excluded because they responded exceptionally inaccurately to at least one animal category in the test phase of the experiment (i.e., ≥ 3 interquartile ranges below the average sample accuracy for one animal category during the test phase), as this was assumed to indicate a general impairment of recognizing the animal category independent of the experimental conditions. The final sample consisted of 95 participants (72% female, age: *M* = 25.54 years, *SD* = 6.43 years, range = 18 to 48 years).

The participants were recruited via advertisements in social media and via an online participant recruitment system at the University Würzburg. We included participants aged between 18 and 55 years. Exclusion criteria were self-reported psychological, psychiatric, neurological, or cardiovascular disorders, self-reported hearing impairment, or current psychopharmacy use or pregnancy. Psychology students were allowed to take part only during their first two study semesters. The participants received 18 € compensation or two hours of course credit for psychology students. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013) and the ethics committee of the University of Würzburg approved of the procedure (GZEK2018-20).

The participants were pre-screened before taking part in the experiment using the N1 trait anxiety subscale of the NEO-Personality Inventory – Revised (NEO-PI-R, Costa & McCrae, 1992) in German (Ostendorf & Angleitner, 2004, *M* = 16.3, *SD* = 6.8, range = 4 to 28). On eight items, the participants are asked to indicate how accurately each item describes their personality on a Likert scale from 0 (”strongly disagree”) to 4 (“strongly agree”). We invited participants with sum scores ≤ 14 or ≥ 20. These thresholds correspond to percentiles of 27.5 (i.e., low

anxiety) and 72.6 (i.e., high anxiety) in norm data from Germany (Ostendorf & Angleitner, 2004). In total, *N* = 1249 participants completed the online questionnaire. *N* = 834 participants fulfilled the screening criteria concerning NEO-PI-R N1 sum scores and age, of whom *N* = 361 provided contact information and were invited to participate in the study. This recruitment strategy aimed at ensuring a broad range of trait anxiety in the study sample.

Procedure

Each participant individually took part in two identically structured experimental sessions on two consecutive days at the same time of day with a duration of one hour each. The participants were seated at a desk with a 24‘‘ sized computer screen and a customary computer keyboard during the experiment. At the beginning of each experimental session, the electrocardiogram electrodes were attached to the participant’s chest and the participants answered psychometric questionnaires (see section 2.4). Afterwards, the resting state electrocardiogram was recorded (see section 2.5). The intensity of the electrotactile stimulation, which was used as an aversive outcome for the avoidance training, was then adjusted with a standardized procedure designed to reach an individual intensity level of “unpleasant, but not painful” intensity, corresponding to a rating of “4” on a scale from 0 (“I do not feel the electrotactile stimulation at all”) to 5 (“painful stimulation”). Each aversive stimulation consisted of 125 separate, consecutive electrotactile stimulations with a duration of 2 ms each and a temporal distance of 3 ms between them (total stimulus duration: 625 ms). The electrotactile stimulations were delivered with a bar electrode (diameter 8 mm, spacing 30 mm) attached to the participant’s non-dominant forearm. The electrotactile stimulations were generated with a Digitimer DS7R stimulator (Digitimer Ltd, n.d.). Afterward, the respective experimental task (i.e., approach training version or avoidance training version) was completed, which took about 35 minutes. To support a standardized task completion, the participants were instructed that all information needed to complete the task would be presented on the computer screen, and that they should refrain from talking to the experimenter if possible. The experiment was controlled and data were recorded with Presentation 21.1 (Neurobehavioral Systems, Berkeley, USA). After task completion, the participants again answered several psychometric questionnaires, and after the second day, they were debriefed.

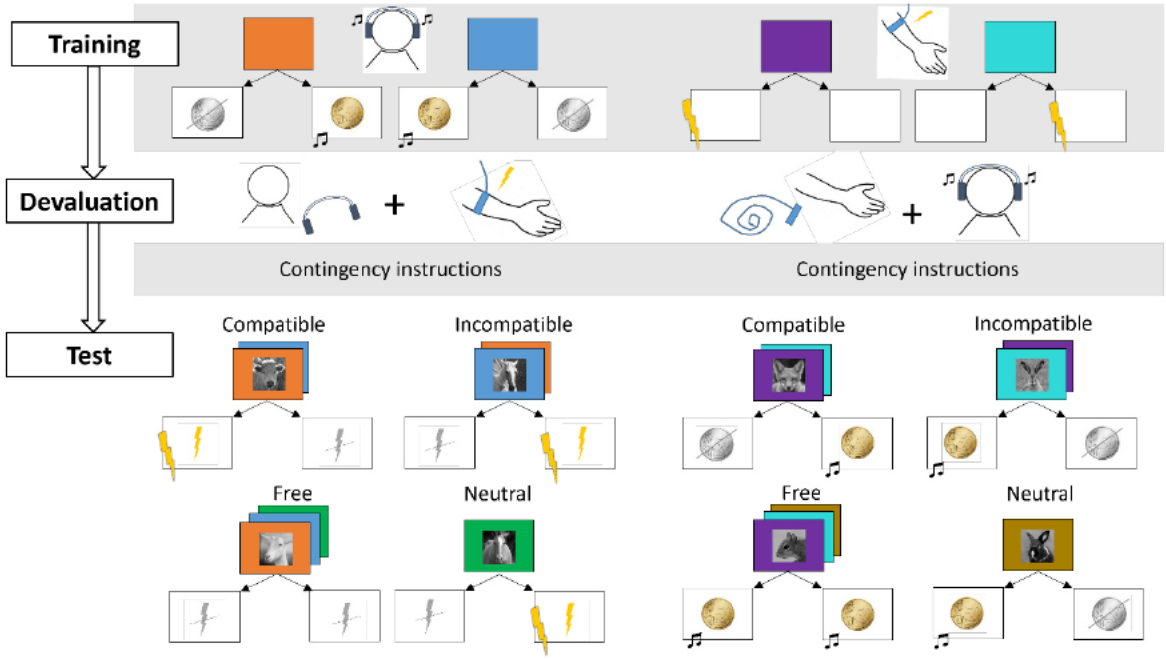
Experimental paradigm

Each participant completed two versions of a devaluation paradigm, one version per day. One of these versions was designed to test approach habits, the other to test avoidance habits.

Each task was presented on a separate day to avert spill-over effects. The order of versions was counterbalanced between the participants. Both tasks consisted of three phases, a training phase, a devaluation phase, and a test phase (see Figure 1), resembling the task used in an earlier study (i.e., Experiment 2 in Glück et al., 2021).

Figure 1

Structure of the approach training task (A) and the avoidance training task (B)



Note. The task order was counterbalanced and the pairings between the tasks, background colors, and picture categories were randomized between the participants.

Training phase

This phase consisted of an extensive training to avoid the electrotactile stimulation in the avoidance training task or to approach the monetary reward in the approach training task. The responses were made by pressing one of two keyboard buttons in response to two different full-screen colors per task (i.e., orange and blue, or purple and turquoise). The full-screen colors (i.e., orange/blue and purple/turquoise) were counterbalanced between the task versions. Each training phase consisted of 200 trials that were pseudo-randomly presented in 25 blocks with eight trials each (i.e., four trials per color in each block). Each trial consisted of the presentation of the respective full-screen color and the subsequent presentation of the response-dependent outcome. The trials were separated by a white fixation cross presented in the middle of the black screen (i.e., inter-trial-interval). The duration of the fixation cross in each trial was random but always between 1000 ms and 3000 ms to prevent the acquisition of a steady rhythm of

responding that may reduce variability in reaction times. Reaction times (i.e., time between color onset and button press, in ms) and accuracy (i.e., correct response or no correct response) were recorded trial-wise.

Approach training task. Before the approach training, headphones were placed on the participant's ears, which were used to signal the monetary rewards. Participants were informed that a soft, non-startling sound may occur. The bar electrode delivering the aversive outcome were not attached to the participant's arm. The participants were informed that they would see two different colors on the screen and that in response to each color, one of the two designated keyboard buttons would produce the reward (i.e., the correct button), while the other button would lead to no reward (i.e., incorrect button), and they would need to find by trial and error which of the two keyboard buttons produced the monetary reward outcome for each color. They were informed that they should aim to gain as many rewards as possible and to respond as fast as possible. The monetary rewards were converted into actual money and paid at the end of the study. After each correct response, a picture of a 50 Euro Cent coin (2.1 x 2.1 cm) was shown on the screen for 1000 ms while a soft sound of dropping coins (30 dB, duration: 450 ms) was presented via the headphones. After each incorrect response or after delayed responses (i.e., ≥ 1000 ms), a picture of a grey, crossed out 50 Euro Cent coin (2.1 x 2.1 cm) was shown on the screen for 1000 ms without any sound. Headphones and sounds were used to keep the outcome similar to the avoidance training and to emphasize later devaluation (i.e., removing headphones).

Avoidance training task. Before the avoidance training phase, the bar electrode was attached to the participant's non-dominant forearm, and the headphones were placed out of sight. The participants were informed that they would see two different screen colors and that, for each color, pressing one of the two buttons would effectively prevent the electrotactile stimulation (i.e., correct response), while pressing the other button would not omit the stimulation for this color (i.e., incorrect response). The participants were informed that they would need to learn the associations between the colors and the respective correct avoidance response by trial and error, and that they should aim to receive as few stimulations as possible while also responding as fast as possible. After each correct and timely response (i.e., ≤ 1000 ms), a picture of a grey, crossed-out lightning bolt (2.4 cm x 2.4 cm) was presented on the screen for 1000 ms without presentation of the electrotactile stimulation. After each incorrect or delayed response (i.e., ≥ 1000 ms), a picture of a yellow lightning bolt (2.4 cm x 2.4 cm, 1000 ms) was shown while the aversive electrotactile stimulation was presented.

Outcome devaluation. After completion of the training phase, the experimenter took off the headphones from the participant's ears in the *approach training task*. The stimulation electrode were then attached to the participant's nondominant forearm. The participants were then informed on-screen that no monetary rewards would be available in the next part of the experiment and, instead, electrotactile stimulations could be delivered. In the *avoidance training task*, the experimenter removed the stimulation electrode from the participant's nondominant forearm and placed the headphones on the participant's head. The participants were then informed on-screen that no aversive stimulations could be delivered anymore and that they could instead receive monetary rewards which would be converted into real money and paid out at the end of the experiment. The experimenter additionally commented that they would take off the electrode and attach the headphones or vice versa to ensure that the participants were aware of the outcome devaluation.

Test phase

Immediately after the devaluation, the participants were instructed on-screen that they would see pictures of three types of animals (i.e., cow, goat, and horse, or rabbit, squirrel, and fox) on the screen in the next part of the experiment, and that they should aim to respond to the pictures as accurately and fast as possible using the same two buttons which they had already used during the subsequent training phase. In the *avoidance training task*, the participants were instructed that pressing the correct button would produce the reward, while pressing the incorrect button would produce no reward. In the *approach training task*, the participants were instructed that pressing the correct button would omit the electrotactile stimulation, while pressing the incorrect button would produce electrotactile stimulation. The instructions included specific information about the association between the correct button for each stimulus type. One of the two response buttons would be correct for one animal type (e.g., "If you see a cow, pressing the left button is correct"), while the other button would be correct for another animal type (e.g., "If you see a goat, pressing the right button is correct"). For the third animal type, both buttons would be correct (e.g., "If you see a horse, then pressing any of the two buttons is correct"). These clear instructions were provided to prevent alternative goals in the paradigm (e.g., participants´ exploring whether the same response would be correct).

The animal pictures were presented centered on top of the two full-screen colors that had already been shown in the training phase (i.e., orange and blue or purple and turquoise) and one additional color (i.e., green or brown). Each visual stimulus consisted of one animal picture and one background color (i.e., picture-color stimuli). To prevent the development of new stimulus-

response associations in the test phase (i.e., between repeatedly presented pictures and correct responses), a new animal picture was presented in every trial (i.e., 60 pictures per animal category). The animal pictures were selected from the picture databases *MemCat* (Goetschalckx & Wagemans, 2019) and *CalTech 256* (Griffin, Holub, & Perona, 2022), rendered black-and-white and cut so that the head area of the animal was clearly recognizable, and were presented with a size of 2.1 x 2.1 cm. The gaze direction of the animals was balanced within each animal category (i.e., equal number of pictures in which the animals gazed left and right). In a pilot study, we compared the difficulty of classifying the pictures to the animal categories and selected pictures that were roughly comparable in difficulty (see Supplement E). Both sets of animals (i.e., set 1: fox, squirrel, and rabbit; set 2: horse, cow, and goat) were presented to all participants. We aimed to balance the two sets of animal types between the task versions so that each set was presented approximately equally often in each task version. The associations between the picture-color stimuli and the correct responses were stable within each participant, but the combinations between the background colors, pictures, and assigned correct buttons were randomized and roughly counterbalanced between the participants. The test phase consisted of 180 trials which were pseudo-randomized within ten blocks. Each block contained four compatible, four incompatible, four neutral, and six free trials in randomized order.

The *compound stimuli* (animal and background color) were grouped into four experimental trial conditions: First, in *compatible* trials, the reward was presented (i.e., avoidance training task version), or the aversive stimulus was omitted (i.e., approach training task version) when the participant pressed the button which had previously produced the desired outcome for the respective background color (i.e., prevented the aversive stimulation during avoidance training or produced the reward in approach training). Second, in *incompatible* trials, the reward was presented when participants pressed the button which had previously produced unwanted outcome for this background color in the training phase (i.e., aversive stimulation or missing reward). Third, in *neutral* trials a new background color was used and the reward was presented or the aversive stimulus was omitted when participants responded correctly (i.e., as instructed), respectively. The neutral control trials were included to assess the acquisition of instructed stimulus-response-outcome contingencies without previous training with the background colors. Fourth, in *free choice* trials, the reward was presented or the aversive stimulus was omitted following any timely response (i.e., pressing any of the two buttons within 1000 ms after stimulus onset).

Behavioral measures

The dependent variables, 1) reaction time (i.e., the time between the color-object stimulus presentation onset and button press, in milliseconds) and 2) accuracy (i.e., correct response, incorrect response, or missing response), were recorded trial-wise in the training phase and test phase. Only reaction times in trials with correct responses were analyzed.

Three operationalization of habitual responses were used: 1) *Accuracy compatibility effect* as the accuracy difference between compatible and incompatible trials. Higher accuracy in compatible than incompatible trials (i.e., a larger accuracy compatibility effect) was conceptualized to indicate a stronger reliance on the extensively trained responses and, therefore, stronger habitual control. Because the accuracy compatibility effect was associated with a monetary loss or more frequent aversive outcomes (i.e., lower accuracy in incompatible trials), we conceptualized it as a costly effect of the extensive training. 2) *Reaction time compatibility effect*, as the reaction time difference between incompatible and compatible trials. A larger reaction time compatibility effect (i.e., slower responding in incompatible trials) was conceptualized to indicate stronger habitual responding. Since the reaction time compatibility effect was not associated with a direct reduction of money or more frequent aversive outcome (i.e., outcomes were not directly tied to response speed unless the response time exceeded 1000 ms), we conceptualized it as a low-cost effect of the extensive training. 3) *Compatibility effect in free trials*. Free trial responses were analyzed regarding their compatibility with the extensively trained responses (i.e., compatible or incompatible with the extensively trained response). The compatibility effect in free trials can also be described as a low-cost effect since incompatible responses in free trials were not associated with monetary costs or more frequent aversive outcomes. Additionally, accuracy in neutral control trials was analyzed to assess the approach and avoidance performance without any compatibility effects.

Questionnaires

All questionnaire sum scores were computed after substituting missing item's scores with the average of the participant's remaining respective questionnaire's items' scores if ≤ 25% of the items were missing. If ≥ 25% of items were missing, the respective individual sum score was treated as missing score. The correlations between the trait anxiety related questionnaires and histograms of the sum scores are reported in Supplement A.

Trait anxiety. As the primary trait anxiety indicator, we used the 7-item subscore of the State-Trait Anxiety Inventory-Trait (STAI-T, Spielberger et al., 1983) as reported by Bados et

al. (2010). The participants answered the complete German version of the STAI-T at the beginning of the first session (Laux et al., 1981), indicating on 20 items how often they generally experienced each described emotion on a scale from 1 ("almost never") to 4 ("almost always"). Acceptable test-retest reliability has been reported for the STAI-T (r = .77 - .90; Laux et al., 1981). As the STAI-T sum score was repeatedly described as a general measure of negative affectivity (i.e., Bieling et al., 1998; Knowles & Olatunji, 2020), we used the STAI-T anxiety subscore as a more valid operationalization of trait anxiety (Bados et al., 2010). This trait anxiety score (sample average M = 14.29, SD = 4.85, range = 7 to 25) correlated substantially with all other anxiety questionnaires (all $r \geq .52$; see Supplementary Table A.1). The internal consistency was $\alpha = 0.93$ for the STAI-T and $\alpha = 0.89$ for the anxiety subscale.

Trait anxiety related questionnaires. At the start of the first experimental session, the participants rated the *Anxiety Sensitivity Inventory* (ASI-3, Taylor et al., 2007) in German (Kemper et al., 2011, M = 21.2, SD = 12.5, range = 0 to 55). The ASI-3 measures anxiety sensitivity, which describes the individual propensity to fear anxiety symptoms (e.g., Reiss, 1991; Taylor, 1995). A proportion of 58.1% of the participants in our sample showed a sum score ≥ 17, which has been reported as a threshold for moderate anxiety sensitivity (Allan et al., 2014). The internal consistency of the anxiety subscale was $\alpha = 0.88$. Additionally, at the start of the first experimental session, the participants rated the *Depression, Anxiety, and Stress Scales* (DASS-21, Lovibond & Lovibond, 1995; German translation by Nilges & Essau, 2015, M = 4.25, SD = 5.95, range = 0 to 30), which measures the frequency of depression, anxiety, and stress symptoms during the last even days with 21 items on a scale from 0 ("not at all") to 4 ("very strongly or most of the time"). We multiplied the sum score by two to achieve equivalence with a 42-item version of the questionnaire (Lovibond & Lovibond, 1996). A DASS anxiety sum score of 10 has been reported to differentiate participants with anxiety disorder diagnoses (Tran et al., 2013), while another study reported a sum score of 14 has been reported to identify individuals with moderate anxiety symptoms (Lovibond & Lovibond, 1996). A proportion of 12.6% in our sample exhibited a DASS anxiety scum score of at least 10, and a proportion of 9.8% of participants scored at least with a sum of 14, indicating a relatively small proportion of participants with high levels of anxiety symptoms. The internal consistency for the anxiety subscale was $\alpha = 0.79$, for the stress subscale $\alpha = 0.82$, and for the depression subscale $\alpha = .87$.

State questionnaires. State anxiety levels were assessed before and after each task with the *State-Trait Anxiety Inventory* (STAI-S, Spielberger et al., 1983) in German (Laux et al., 1981).

The participants are asked to indicate their current experiential strength of 20 state anxiety symptoms on a scale from 1 (“not at all”) to 4 (“greatly”). The internal consistency of the four STAI-S measurements in our sample varied between $\alpha = 0.87$ and $\alpha = 0.90$. The pre- and post-task state anxiety levels did not significantly differ between the two tasks (see Table 1)**.** Furthermore, the participants rated their current sleepiness level before and after each task with the Karolinska Sleepiness Scale (KSS, Akerstedt & Gillberg, 1990; German version by Popp et al., 2011) on a scale from 1 (“extremely alert”) to 10 (“very sleepy, cannot stay awake”). The pre- vs. post-task sleepiness changes did not differ between the two tasks (see Table 1).

Motivation and stimulus evaluation ratings. After the task was completed, the participants rated several task-specific items which had been constructed for this study without previous validation. They rated their retrospective motivation to avoid the aversive stimulus (i.e., avoidance motivation), to approach the reward (i.e., approach motivation), and to respond as fast as possible on visual analog scales from 0 (“Not motivated at all”) to 100 (“Extremely motivated”). Additionally, a retrospective, subjective rating of the level of stress elicited by the devaluation procedure was assessed after each task. For this, the participants were asked to recall which level of stress they had experienced during the devaluation procedure and to indicate it on a visual analog scale from 0 (“Not stressed at all”) to 100 (“Extremely stressed”). The participants also indicated the retrospective subjective aversiveness of the electrotactile stimulation on a visual analog scale ranging from 0 (“not unpleasant at all”) to 100 (“extremely unpleasant”) and the perception of the auditory stimulus on a visual analog scale ranging from 0 (“extremely unpleasant”) to 100 (“extremely pleasant”). On average, the electrotactile stimulation was rated as relatively unpleasant ($M = 67.41$, $SD = 21.29$), and the auditory stimulus as rather pleasant ($M = 73.62$, $SD = 21.38$). Avoidance motivation was rated higher than approach motivation after the approach training task, $W = 1758.5$, $p < .001$, and after the avoidance training task, $W = 1692.0$, $p = .001$, but the approach and avoidance motivations did not differ between the tasks (see Table 1). The only difference between tasks was a higher self-reported stress level as induced by the devaluation procedure in the approach training task (i.e., removing headphones and attaching electrodes) than in the avoidance training task (i.e., removing electrodes and attaching headphones; see Table 1). The STAI trait anxiety score correlated with self-reported stress as induced by the outcome devaluation procedure in the approach training task, $r = .37$, $p < .001$, but not in the avoidance training task, $r = .05$, $p > .999$. The STAI trait anxiety score did not correlate with approach or avoidance motivation or with the perception of the outcomes in any of the tasks (all $r \leq .18$, all $p \geq .607$).

Table 1

Self-reported subjective states, motivational tendencies, and subjective outcome stimulus perception in both tasks

	Approach training task		Avoidance training task				
	M	*SD*	*M*	*SD*	Wilcoxon's *W*	p^1	Wilcoxon effect size *r*
Approach motivation (post, 0-100)	79.82	19.94	78.39	21.37	4636.5	> .999	.09
Avoidance motivation (post, 0-100)	88.34	16.67	86.22	19.97	4758.0	> .999	.13
Motivation to respond fast (post, 0-100)	79.01	24.56	76.55	24.28	4881.5	> .999	.07
Electrotactile stimulus aversiveness (post, 0-100)	65.95	22.15	68.57	20.31	4185.5	> .999	.15
Auditory stimulus perception (post, 0-100)	74.46	19.49	72.56	23.23	4462.5	> .999	.09
Stress evoked by devaluation procedure (post, 0-100)	47.53	32.21	14.85	19.47	6964.5	> .001	.67
State anxiety (pre, 20-80)	35.31	8.00	34.52	7.37	4759.0	>.999	.03
State anxiety (post, 20-80)	37.45	9.91	34.45	8.48	5284.5	.264	.30
Sleepiness (pre, 1-10)	3.59	1.53	3.52	1.32	4449.5	> .999	.12
Sleepiness (post, 1-10)	3.48	1.67	3.98	1.88	3786.0	.533	.47

Note. $N = 95$. [1] Bonferroni-Holm corrected.

Heart rate variability measurement and preprocessing

An electrocardiogram was recorded at the beginning of each experimental session during a resting period for a duration of 4:30 minutes during which the participants were sitting upright and watched a soundless videoclip with a non-arising content (i.e, a clip taken from a documentary about the cleaning of an airplane; Graf, 2019). The same videoclip was presented on both sessions. The participants were instructed to seat comfortably and to avoid bodily movements during the resting period. We used a three-electrode system with disposable Ag/AgCl electrodes (i.e., below the right collarbone, on the lower right ribcage, and on the left collarbone). The electrocardiogram was recorded with *BrainVision Recorder* (Version 1.2, Brain Products GmbH, 2018b) with a sampling rate of 1000 Hz.

The electrocardiogram data were preprocessed with *BrainVision Analyzer* (Version 2.1 Brain Products GmbH, 2018a). The data were filtered with a low-cutoff filter of 3.2 Hz, a high-cutoff filter of 40.0 Hz, and a notch filter of 50.0 Hz. The first 30 seconds of each measurement were discarded to remove movement artifacts, resulting in 240 s per measurement. Subsequently, the recordings were visually inspected and artifacts were removed. The R peaks were then detected automatically using the *BrainVision Analyzer* package *EKG markers* (Version 1.11). Incorrectly detected R peaks were manually corrected before the R-R intervals were computed in *BrainVision Analyzer* (Version 2.1, Brain Products GmbH, 2018). We then calculated the standard deviation of the inter-beat-intervals of normal sinus beats (SDNN), a commonly used indicator for short measurement durations. A higher SDNN indicates higher heart rate variability, while a lower SDNN indicates lower heart rate variability (Shaffer & Ginsberg, 2017). We obtained the average individual heart rate and SDNN as recorded in each laboratory session with the software *Kubios HRV Standard* (Version 3.5.0, Tarvainen et al., 2014). Individual SDNN values which deviated from the sample average with three or more standard deviations wee dubstituted with the grand average score; this applied to two measurement points. Seven individual SDNN and heart rate scores (i.e., 3.7% of all measurements) were missing due to technical problems during the data collection and were substituted with the grand average score.

The heart rate and SDNN measurements as obtained on the first and second session correlated positively (heart rate: $r(94) = .74$, $p < .001$; SDNN: $r(94) = .77$, $p < .001$), indicating a relatively high temporal stability of both measures. The individual average heart rate and SDNN correlated negatively within each measurement (first session: $r(94) = -.46$, $p < .001$, second session: $r(94) = -.47$, $p < .001$), as was expected due to the inherent dependency of the

SDNN computation on heart rate (see Sacha, 2014; Shaffer & Ginsberg, 2017). The individual STAI-T anxiety sum score did not correlate with heart rate (first session: $r(94) = -.03$, $p > .999$, second session: $r(94) = .08$, $p > .999$), or SDNN (first session: $r(94) = -.08$, $p > .999$, second session: $r(94) = -.10$, $p > .999$).

Hypotheses

To test whether higher trait anxiety would predict a stronger acquisition of avoidance habits than approach habits, we hypothesized that the strength of a) the accuracy compatibility effect, b) the reaction time compatibility effect, and c) the compatibility effect in free trials would be predicted more pronouncedly by the individual trait anxiety score in the avoidance training task than in the approach training task. We exploratorily tested whether lower heart rate variability (i.e., a lower individual SDNN) would predict the strength of the three indicators of habitual avoidance or approach.

Data analysis

Data exclusion criteria. Trials with reaction times lower than 100 ms were excluded from data analysis (i.e., responses in these trials were treated as missing data) as they were not assumed to reflect voluntary movements; this applied to 11 trials (0.01% of all trials). .

Statistical modeling. The data were analysed with mixed effects models to account for the clustering of the response data within individual participants. Mixed models can be described as a hierarchical system of regression equations with randomly varying coefficients (i.e., random effects; see Hox et al., 2018). All mixed effects models included random intercepts for each participant, for each animal category, and for each randomization version. Reaction times (i.e., a continuous outcome) were modelled with linear mixed models (LMMs) and accuracy data (i.e., a binary outcome) were modelled with generalized linear mixed models (GLMMs). All LMMs and GLMMs were estimated using the restricted maximum likelihood approach (REML) with the Nelder-Mead optimizer with a maximum of 2^{10} iterations. The significance of the single regression coefficients in all LMMs and GLMMs were tested using Type III sums of squares ANOVAs with the *R* package *cars* (Fox & Weisberg, 2019). A potential collinearity of the predictors within each statistical model was checked with the *R* package *performance* (Lüdecke et al., 2021). The fit of each model was visually assessed by inspecting the normality of the residuals in the QQ plot for the LMMs, or the residuals' distributions in the binned residual plots for the GLMMs (see Gelman & Hill, 2006) with the *R* package *performance*

(Lüdecke et al., 2021). Individual *SDNN*, *heart rate*, and *STAI-T anxiety* scores were included as *z*-standardized predictors.

The training phase responses were analyzed to check for effects during the training which may have potentially biased the subsequent test phase responses. Two models were estimated (i.e., one GLMM for accuracy and one LMM for reaction times), which included the fixed factors *task version*, *trait anxiety score*, *block*, their two-way and three-way interactions, and *session number* (i.e., first and second session to account for order effects). We exploratorily compared the predictive value of the STAI-T anxiety score with the predictive value of the DASS anxiety subscale score and the ASI-3 score by modeling training accuracy separately with each of these predictors (see Supplement B). As a result, the three scales predicted the training accuracy in a similar way. As the rational of the STAI anxiety score maps more closely than the other two scales to the construct of interest, trait anxiety, we kept the STAI-T anxiety scale as predictor in the LMMs and GLMMs. To explore the impact of heart rate variability on training performance, we added *SDNN* (i.e., z-standardized average as measured during the resting period on each day), the *SDNN x task version* interaction, and heart rate (i.e., to control for a potential confounding effect on the effect of SDNN) to both models (see Supplement C).

Three models were calculated to analyze the impact of trait anxiety on the compatibility effects: First, the effect of *trait anxiety* on the reaction time compatibility effect was analyzed with a LMM with reaction time as the outcome. Second, and third, the effect of *trait anxiety* on the strength of the accuracy compatibility effect and on the compatibility effect in free trials were analyzed with one GLMM each. The LMM for reaction time and the GLMM for accuracy in compatible and incompatible trials included the fixed effects *condition* (i.e., compatible and incompatible condition), *task* (i.e., approach or avoidance training task version), *trait anxiety score* (z-standardized STAI-T anxiety factor sum score), *block (i.e., 1 to 10)*, and the two-way, three-way, and four-way interactions between these four factors. The GLMM for the effect of anxiety on the compatibility effect in free trials included these same predictors and interactions, but the predictor *condition* was omitted. The effects of *condition*, *task*, and *session number* were modelled as sum contrasts, while *trait anxiety score*, *block*, *SDNN*, and *heart rate* were modelled as linear predictors. To explore the impact of heart rate variability on the test phase responses, we added *SDNN* (i.e., z-standardized average as measured during the resting period on each day), the *SDNN x condition* interaction, the *SDNN* x *task* interaction, the *SDNN* x *condition* x *task* interaction, and *heart rate* (i.e., to control for a potential confounding effect on the effect of SDNN) (see Supplement C). Post-hoc pairwise comparisons were performed with

the R package *emmeans* (Lenth et al., 2023). All GLMMs and LMMs are reported in Supplement B. To improve the interpretability of null effects and to assess the level of confidence in the results, we additionally computed Bayes' Factors with models for each outcome with the same fixed and random effects as in the LMMs and GLMMs using the brms R package with the default of 2000 iterations per chain and four chains (v.2.20.1; Bürkner, 2017). For all parameters, we used an uninformative prior and a Cauchy distribution with a mean of 0 and a standard deviation of 1. Bayes' factors were estimated by comparing the probability of the H1 (i.e., effect is equal to zero) to the H0 (i.e., effect is unequal to zero). Alongside the Bayes' factors, we report parameter estimates with 95% credibility intervals (see Bürkner, 2017; Krypotos et al., 2017). The data and the analysis code are available at https://osf.io/7gr9b/.

4.3 Results

Training phase

Accuracy

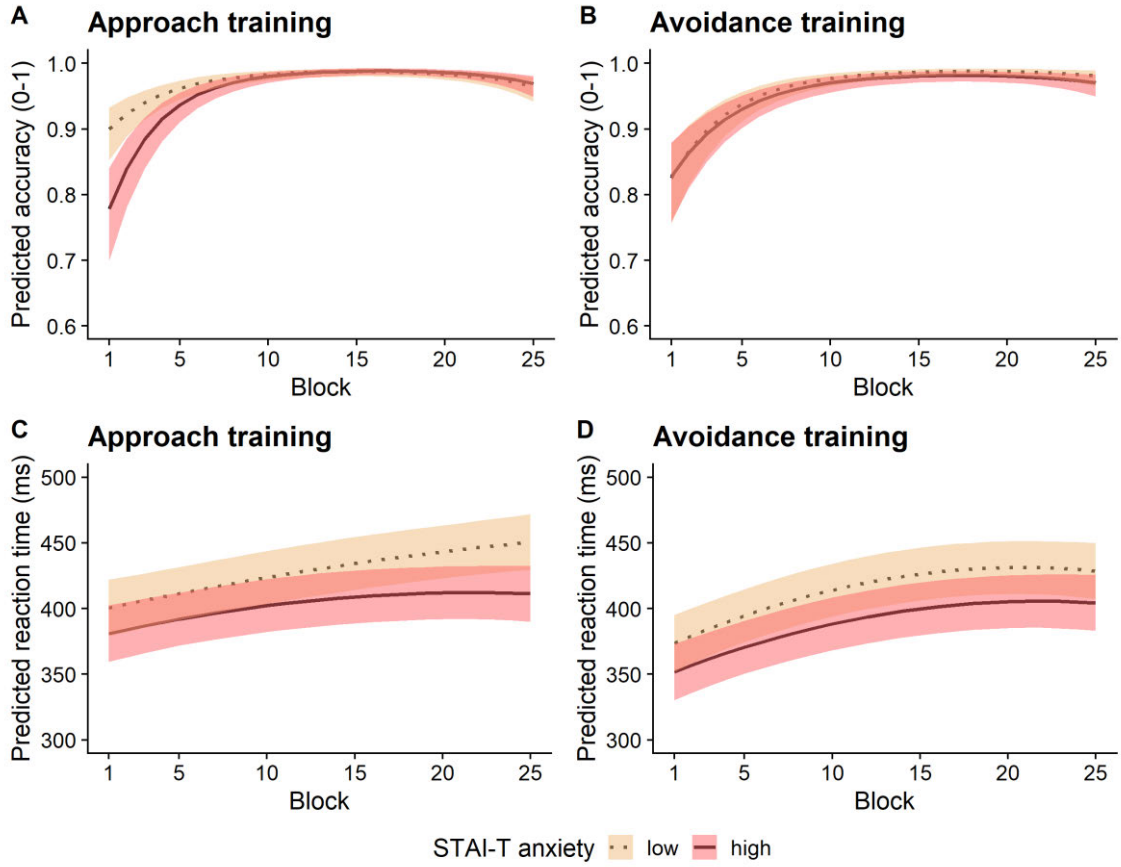
The GLMM to predict accuracy in the training phase yielded a significant interaction between *trait anxiety*, *task version*, and *block*, $X^2(1) = 11.05$, $p = .004$, BF_{01} (linear trend for block) = 0.29, estimate = 6.23, 95% CI [-1.70; 23.15]), BF_{01} (quadratic trend) = 0.74, estimate = -1.71, 95% CI [-13.50; 2.94], indicating lower accuracy in the beginning of the approach task's training phase and a stronger subsequent increase in higher anxious individuals (see Figure 2). Additionally, *day* was significant predictor, indicating that the accuracy was slightly lower on the first day (estimated hit probability = 97.66 %, 95 % CI [97.03; 98.16]), than on the second day (estimated hit probability = 98.92 %, 95 % CI [98.61; 99.16]), $BF_{01} < 0.01$, estimate = 0.78, 95% CI [0.67; 0.89]. *Task version* predicted accuracy, $X^2(1) = 7.45$, $p = .006$, $BF_{01} = 0.19$ (estimate = 0.19, 95% CI [0.07; 0.31]), with a slightly higher estimated accuracy in the approach training (estimated marginal mean = 98.64 %, 95 % CI [98.24; 98.95]) than in the avoidance training (estimated marginal mean = 98.13%, 95 % CI [97.60; 98.24]). The exploratory GLMM indicated that higher *SDNN* predicted lower accuracy, $X^2(1) = 8.49$, $p = .004$, $\beta = -0.143$, SE = 0.067; this effect did not interact with *task version*, $X^2(1) = 2.11$, $p = .146$.

Reaction times

The LMM to predict reaction times in the training phase yielded a significant interaction between *task version* and *block*, $X^2(2) = 31.57$, $p < .001$, BF_{01} (linear trend of *block*) = 0.53, estimate = 0.34, 95% CI [-7.48; 12.23], BF_{01} (quadratic trend of *block*) = 0.53, estimate = 0.34, 95% CI [-7.48; 12.23], indicating a slightly stronger linear trend and slightly stronger negative quadratic trend in the avoidance task (estimated linear slope = 2.18, [95 % CI: 1.99; 2.37], estimated quadratic slope = -0.14, 95 % CI [-0.16; -0.11]) than in the approach task (estimated linear slope = 1.65, 95% CI [1.47; 1.84], estimated quadratic slope = -0.05, 95 % CI [-0.08; -0.02]; see Figure 2). *Trait anxiety* did not predict reaction times as a single predictor, $X^2(1) = 2.16$, $p = .142$, $BF_{01} = 0.49$, estimate = -1.14, 95% CI [-7.40; 2.35]) or in interaction with *task version*, $X^2(1) = 0.02$, $p = .895$, with *block*, $X^2(2) = 2.48$, $p = .289$, or with both, $X^2(2) = 2.48$, $p = .289$. The estimated average reaction time in both task versions was fast (estimated average reaction time for approach training = 410 ms, 95 % CI [399; 420]; estimated average reaction time for avoidance training = 419 ms, 95 % CI [409; 429] with a difference between the task versions, $X^2(2) = 189.24$, $p < .001$, $BF_{01} < 0.01$, estimate = 13.28, 95% CI [11.30; 15.25]. The exploratory LMM indicated that higher *SDNN* significantly predicted faster reaction times, $X^2(1) = 9.06$, $p = .004$, $\beta = -4.86$, $SE = 1.64$; this effect did not interact with *task version*, $X^2(1) = 1.82$, $p = .177$. Higher *heart rate* predicted faster reaction times, $X^2(1) = 21.43$, $p < .001$, $\beta = -6.87$, $SE = 1.50$.

Figure 2

Estimated marginal means in the models to predict accuracy and response time during both training phases



Note. Significance bands display 95 % confidence intervals. Low STAI-T anxiety indicates two standard deviations below the sample average, high STAI-T anxiety indicates two standard deviations above the sample average.

Test phase

Accuracy compatibility effect

The interaction between *condition* and *task version* significantly predicted accuracy, $X^2(1) = 7.53$, $p = .006$, $BF_{01} = 0.69$, estimate = -0.26, 95% CI [-.51; -0.03]. Pairwise comparisons of the estimated marginal means indicated an accuracy compatibility effect in the approach training task version, $OR = 1.35$, $SE = 0.13$, $p = .001$, but not in the avoidance training task version, $OR = 0.95$, $SE = 0.09$, $p = .530$. Additionally, the *condition* x *task version* x *trait anxiety* interaction approached significance, $X^2(1) = 3.75$, $p = .053$, $BF_{01} = 1.67$, estimate = -0.11, 95% CI [-0.43; 0.03]. Post-hoc comparisons indicated that *trait anxiety* tended to predict a stronger accuracy compatibility effect in the approach task (estimate: 0.292, SE = 0.141, 95% CI [-0.568;

-0.015]), but not in the avoidance task (estimate: -0.149, SE = 0.140, 95% CI [-0.125; 0.424]) (see Figure 3). The four-way interaction between *condition, task version, trait anxiety, and block* was a nonsignificant predictor, $X^2(1) = 1.17$, $p = .557$, BF_{01} (linear effect of *block*) = 0.73, estimate = -0.11, 95% CI [-5.18; 4.65], BF_{01} (quadratic effect of *block*) = 0.67, estimate = -0.01, 95% CI [-5.93; 5.68]. Descriptively, the impact of *trait anxiety* seemed to be stronger in the end of the approach task's test phase than in the beginning (e.g., estimate in first block: 0.083, estimate in last block: 0.274; see Figure 3). *SDNN* and *heart rate* did not predict accuracy or the accuracy compatibility effect, all *p*s ≥ .115.

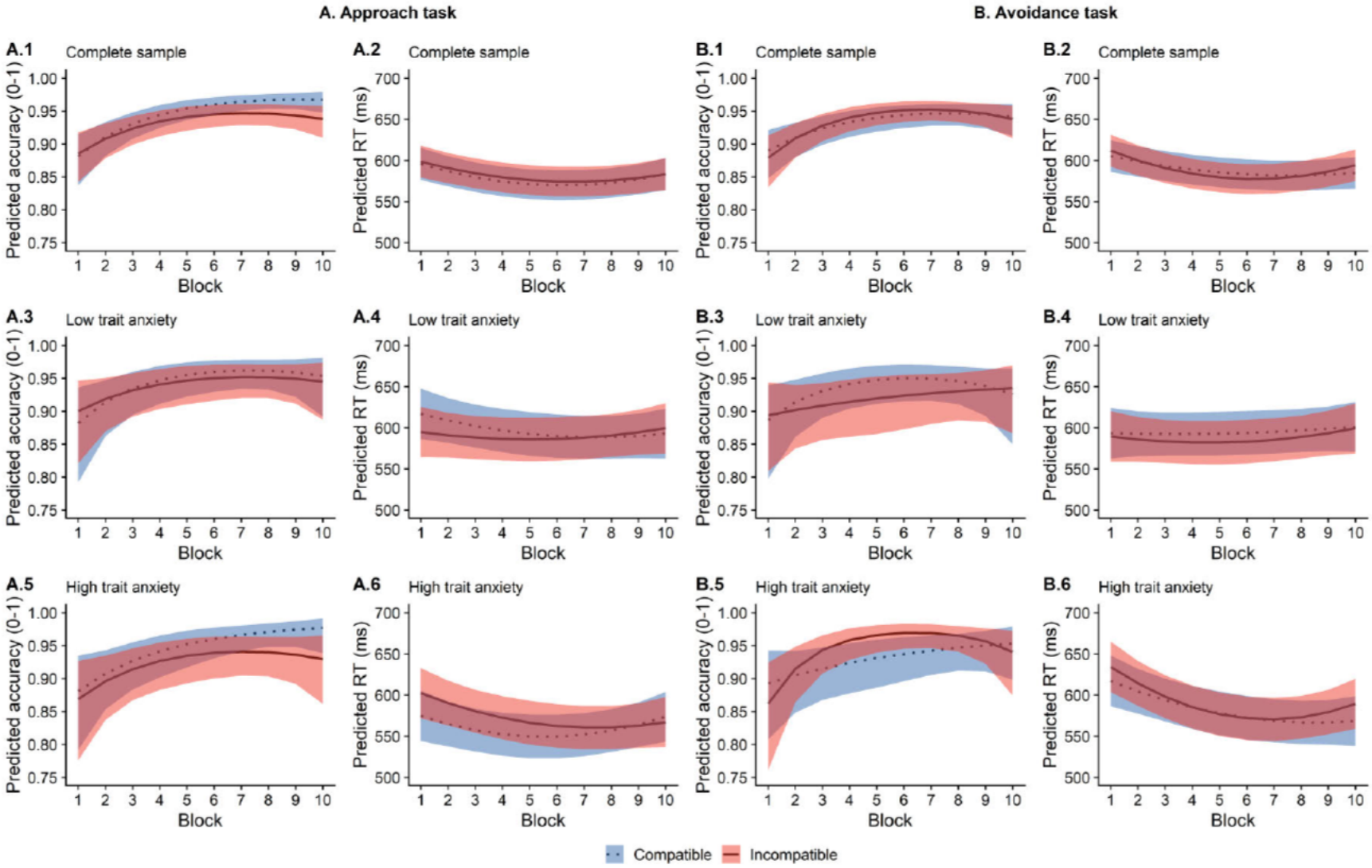
Reaction time compatibility effect

The interaction between *trait anxiety* and *condition* significantly predicted reaction times in compatible and incompatible trials, $X^2(1) = 6.64$, $p = .010$, $BF_{01} = 0.51$, estimate = 2.30, 95% CI [-0.51; 6.35]. Post-hoc comparisons indicated that higher *trait anxiety* predicted a generally stronger reaction time compatibility effect (i.e., estimated reaction time compatibility effect for high trait anxiety (i.e., 2 *SD* above sample average) = 7.01 ms, *SE* = 5.74, 95% CI [-3.09; 19.38]); estimated reaction time compatibility effect for low trait anxiety (i.e., 2 *SD* below sample average) = -8.14 ms, *SE* = 5.73, 95% CI [-18.25; 4.23]). Additionally, the interaction between *trait anxiety* and *task version* predicted reaction times in compatible and incompatible trials, $X^2(1) = 10.19$, $p = .001$, $BF_{01} = 0.13$, estimate = -4.21, 95% CI [-8.47; -0.24], indicating that higher *trait anxiety* predicted lower reaction times in the approach training task (β = -8.27, *SE* = 4.55, 95% CI [-17.20; 0.65]) but not in the avoidance training task (β = -3.62, *SE* = 4.56, 95% CI [-12.60; 5.30]). A significant interaction between *trait anxiety* and *block*, $X^2(2) = 11.14$, $p = .004$, BF_{01} (linear effect of *block*) = 1.04, estimate = -2.44, 95% CI [-31.79; 7.55], BF_{01} (quadratic effect of *block*) = 1.26, estimate = 1.08, 95% CI [-7.10; 19.61] indicated that higher *trait anxiety* predicted faster responses more strongly in later blocks (e.g., β = -6.12, *SE* = 4.67, 95% CI [-15.26; 3.03] in the last block) than in earlier blocks (e.g., β = 1.88, *SE* = 4.70, 95% CI [-7.32; 11.09] in the first block). The three-way interaction between *trait anxiety*, *condition*, and *task version* was an insignificant predictor, $X^2(1) = 0.10$, $p = .754$, indicating no specific effect of *trait anxiety* on the reaction time compatibility effect in any of the two tasks, albeit with a low clarity, $BF_{01} = 1.29$, estimate = 0.37, 95% CI [-2.93; 4.36]. The interaction between *condition* and *task version* was also insignificant, indicating no difference in habit strength between approach and avoidance, although the evidence was again rather inconclusive, $X^2(1) = 1.04$, $p = .310$, $BF_{01} = 1.21$, estimate = 0.72, 95% CI [-1.99; 4. 96]. In the exploratory model, higher *SDNN* significantly predicted higher reaction times, $X^2(1) = 5.21$, $p = .022$, β = 6.39, *SE*

= 2.80, but *SDNN* did not interact with *condition*, $X^2(1) = 1.75$, $p = .187$, or *task version*, $X^2(1) = 0.13$, $p = .716$.

Figure 3

Estimated marginal means in the models to predict accuracy and response times in compatible and incompatible trials



Note. Significance bands display 95 % confidence intervals.

Accuracy and reaction times in free trials

None of the predictors in the GLMM significantly predicted the compatibility effect in free trials, all $p \geq .206$. There was no general accuracy compatibility effect in the avoidance task (estimated marginal mean = 49.5%, *SE* = 0.81, 95% CI [47.7; 51.3]), or in the approach task (estimated marginal mean = 50.3%, *SE* = 0.81, 95 % CI [48.4; 52.1]; see Figure 4). However, the Bayesian analyses indicated moderate evidence for an interaction between *task version*, *trait anxiety* and *block*, BF_{01} (linear trend of *block*) = 0.31, estimate = 1.22, 95% CI [-1.75; 6.74], BF_{01} (quadratic trend of *block*) = 0.33, estimate = -0.86, 95% CI [-5.47; 1.92]. Additionally, the Bayesian analysis indicated strong evidence for a null effect of *trait anxiety* as a single predictor, BF_{01} = 8.89, estimate = 0.03, 95% CI [-0.04; 0.09], or in interaction with *task version*, BF_{01} = 8.19, estimate = -0.02, 95% CI [-0.11; 0.07]. To analyze whether some participants invariantly pressed the same button in the free trials, we obtained the consistency of same-button choices per participant for each task (i.e., 97.5% of same-button choices, equivalent to at least 39 same-button responses of 40 responses in free trials within one task version). In 31.05 % of tasks, the participants chose the same button with such consistency, indicating the potential use of a strategy for the responses in free trials. When the GLMM was exploratorily estimated after the exclusion of these participants, a tendency toward a *trait anxiety* x *block* interaction effect emerged, $X^2(2) = 5.69$, $p = .058$. Post-hoc comparisons suggested that *trait anxiety* tended to predict a more negative quadratic trend for the proportion of habit-compatible responses over blocks (β = -0.005, *SE* = 0.005, 95% CI [-0.016, 0.006], BF_{01} = 1.49), but did not predict a linear trend over blocks (β = - 0.005, *SE* = 0.014, 95% CI [0.034, 0.023], BF_{01} = 0.65). *SDNN* and *heart rate* did not predict the accuracy or the accuracy compatibility effect, all *p*s ≥ .207.

Reaction times in free trials were predicted by the *task* x *trait anxiety* interaction, $X^2(2) = 4.57$, $p = .033$, BF_{01} = 0.30, estimate = -2.13, 95% CI [-7.82; 0.89], indicating that *trait anxiety* predicted faster responses in the approach task (i.e., β = -7.92, *SE* = 5.68, 95 % CI [-19.10; 3.22]), but not in the avoidance task (i.e., β = 0.47, *SE* = 5.69, 95 % CI [-10.7; 11.62]; see Figure 4). Reaction times were also predicted by *session number*, $X^2(2) = 4.92$, $p = .027$, BF_{01} = 0.31, estimate = -1.96, 95% CI [-7.76; 1.06], with slightly slower responses on the first session (estimated marginal mean = 591 ms, SE = 7.98, 95 % CI [575; 606]) than on the second session (estimated marginal mean = 585 ms, SE = 7.98, 95 % CI [569; 600]). We exploratorily modeled reaction times in free trials with an additional predictor indicating whether the response was habit-compatible or habit-incompatible (see Supplementary Table B.9). The results indicated that reaction times did not differ between habit-compatible and habit-incompatible responses, $X^2(1) = 0.98$, $p = .322$, BF_{01} = 1.27, estimate = -0.41, 95% CI [-3.65; 2..05], and were not

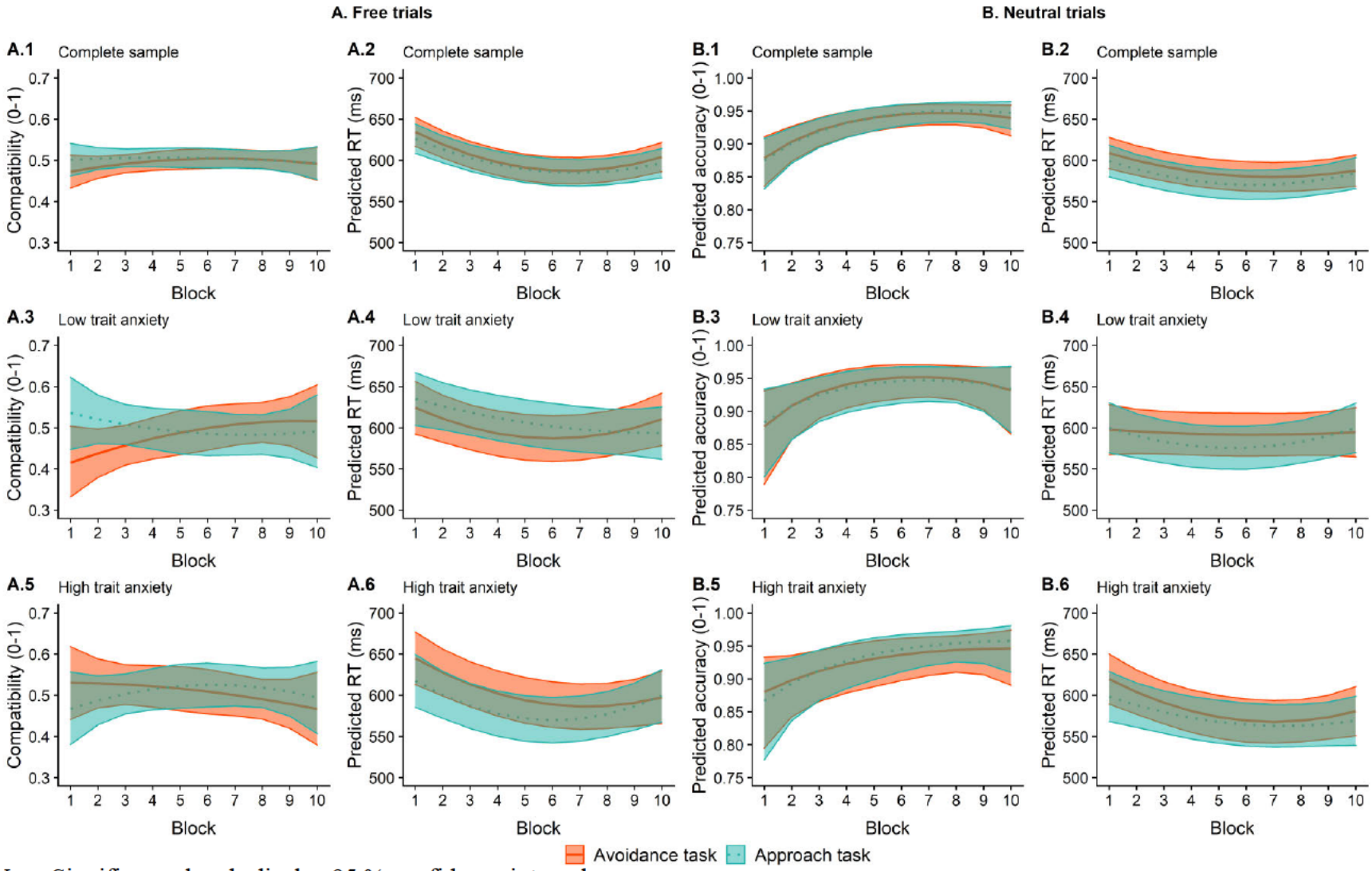
predicted by *trait anxiety*, $X^2(1) = 0.20$, $p = .658$, $BF_{01} = 1.11$, estimate = -0.05, 95% CI [-4.51; 3.97]. In the exploratory model with SDNN and heart rate, reaction times were predicted by the *task* x *SDNN* interaction, $X^2(1) = 7.08$, $p = .008$, $BF_{01} = 0.05$, estimate = -9.43, 95% CI [-16.96; -0.56], indicating that higher *SDNN* predicted faster responses in the approach training task ($\beta = -17.55$, *SE* = 4.11, 95 % CI [-25.60; -9.50]), but not in the avoidance training task ($\beta = -7.59$, *SE* = 4.66, 95 % CI [-16.70; 1.56]). Furthermore, higher *heart rate* predicted generally slower responses in free trials, $X^2(1) = 5.77$, $p = .016$, $\beta = 5.77$, *SE* = 0.02, $BF_{01} = 0.46$, estimate = -2.95, 95% CI [-10.25; 1.02].

Accuracy and reaction times in neutral trials

Block significantly predicted accuracy, $X^2(2) = 65.71$, $p < .001$, indicating that the accuracy in neutral trials increased linearly ($\beta = 0.095$, *SE* = 0.015, $BF_{01} < 0.01$, estimate = 22.69, 95% CI [13.70; 30.81]) and quadratically ($\beta = -0.021$, *SE* = 0.006; $BF_{01} = 0.03$, estimate = -8.96, 95% CI [-18.55; 0.06], see Figure 4). None of the other predictors significantly predicted accuracy, all $p \geq .599$. Reaction times in neutral trials were predicted by *task version*, indicating faster response times in the avoidance task (estimated marginal mean: 588 ms, S*E* = 8.86) than in the approach task (estimated marginal mean: 579 ms, *SE* = 8.87), $X^2(1) = 9.35$, $p = .002$, $BF_{01} = 0.11$, estimate = -5.07, 95% CI [-10.71; -0.02]). In the exploratory model, higher *SDNN* additionally predicted higher reaction times, $X^2(1) = 8.97$, $p = .003$, $\beta = 10.91$, *SE* = 3.68, $BF_{01} = 0.09$, estimate = 5.34, 95% CI [-0.59; 13.98].

Figure 4

Estimated marginal means in the models to predict accuracy and response times in free and neutral trials



Note. Significance bands display 95 % confidence intervals.

Correlations of the compatibility effects. The compatibility effects versions did not significantly correlate between the two tasks (all $r \leq .05$, all $p \geq .999$, see Supplementary Table A.2), implying that either the effects of the extensive training on the subsequent responses were not temporally stable within individuals, or that different effects were measured in the two tasks. The mean accuracy in neutral control trials, descriptively, correlated positively between the two tasks, $r(95) = .23$, $p = .104$. The compatibility effects within the tasks were not significantly correlated, all $r \leq .20$, all $p \geq .306$, indicating that they did not measure strongly overlapping underlying processes (see Supplementary A).

4.4 Discussion

In this direct experimental comparison between approach and avoidance habit acquisition, participants with high interindividual variance in trait anxiety completed two outcome devaluation tasks involving extensive approach and avoidance training, respectively. In each task version, habitual responses were operationalized with three habit indicators (i.e., reaction time compatibility effect, accuracy compatibility effect, and compatibility effect in free trials). Trait anxiety, as indicated by STAI-T anxiety subscale scores (see Bados et al., 2010; Spielberger et al., 1983), did not predict a stronger acquisition of habitual approach or avoidance. On a trend level, however, higher trait anxiety predicted stronger habitual approach but not stronger habitual avoidance (i.e., regarding the accuracy compatibility effect). Additionally, trait anxiety predicted more pronounced low-cost habit tendencies in both tasks (i.e., a more pronounced reaction time compatibility effect). Independently from trait anxiety, costly habitual approach was generally more pronounced than costly habitual avoidance (i.e., regarding the accuracy compatibility effect). The low-cost habit effects did not differ between the tasks. In general, retrospective self-reports indicated that the test phase in the approach task was perceived as more stressful than the test phase in the avoidance task. Additionally, stress in the approach task was perceived as higher in higher trait-anxious participants. This unexpected task difference is a limitation of the study since it implies that the two task versions were not effectively parallel operationalizations of habitual avoidance and approach.

The increased task-independent reaction time compatibility effect in higher trait-anxious participants indicates a generally stronger impact of the previous extensive training in highly trait-anxious individuals. The current study was, to the best of our knowledge, the first study on the association between trait anxiety and habitual control that analyzed reaction times. Although reaction time compatibility effects may not be readily interpreted as definite

indicators of habitual control on their own, available evidence indicates a sensitivity of reaction times to response conflicts in outcome devaluation tasks. Luque et al. (2019) demonstrated that reaction time compatibility effects increased as a function of training and were more pronounced when the conflict between two response choices was more prominent due to higher competing rewards. Reaction time compatibility effects allowed detecting competing goal-directed and habitual response tendencies in the absence of effects on overt response choices (Hardwick et al., 2019; Luque et al., 2019). The inclusion of reaction times can also increase the reliability of model-free and model-based parameter estimation in two-step sequential tasks (Shahar et al., 2019). Earlier studies reporting null associations between trait anxiety and habitual responses had analyzed accuracy effects, but not reaction time effects (i.e., Flores et al., 2018; Gillan et al., 2014; Gillan et al., 2021). However, analyzing accuracy in earlier studies may have prevented the detection of subtle effects of trait anxiety that may not have been transferred into accuracy compatibility effects. This result thus indicates a small and unspecific effect of trait anxiety on habitual response tendencies.

The current results indicate that the non-costly influence of trait anxiety on reaction times compatibility effects may apply to extensively trained approach and avoidance responses. Therefore, the study does not confirm null effects of trait anxiety on habitual approach (Gillan et al., 2016; Gillan et al., 2021) and habitual avoidance (Gillan & Robbins, 2014; Patterson et al., 2019). However, in our study, the indicator associated with trait anxiety was a low-cost habit indicator, and the effect was only apparent when analyzing reaction times. As already mentioned, potentially, the analysis of reaction times in the current studies enabled the detection of smaller effects than in analyses confined to accuracy in earlier studies (see Hardwick et al., 2019; Luque et al., 2019). Of note, trait anxiety did not influence accuracy or reaction times in neutral trials. Thus, trait anxiety selectively impacted responses in trials where extensively trained responses and goal-directed responses conflicted.

Of note, we observed stress differences between the tasks without an explicit stress induction procedure. Potentially, the elevated stress in the approach training task's test phase may have generally increased habitual responses, as was indicated by the more pronounced accuracy compatibility effect in the approach task than in the avoidance task. The retrospective self-reports indicated that the outcome devaluation procedure in the approach training task (i.e., the application of the electrodes and removal of the headphones) was generally perceived as more stressful than the devaluation procedure in the avoidance training task (i.e., the application of the headphones and removal of the electrodes). Combined with the limited time to respond

and the high number of different picture stimuli, the approach task's test phase may have been perceived as a threatening context with a high likelihood of aversive stimulations. In contrast, the outcome devaluation procedure in the avoidance training task signaled subsequent rewards and safety from the aversive stimulations., and may have been perceived as a safe context (e.g., Sjouwerman et al., 2015). Potentially, the elevated stress in the approach training task's test phase may have generally increased habitual responses, as was indicated by the more pronounced accuracy compatibility effect in the approach task than in the avoidance task.

In action control studies, stress before an instrumental training phase (Schwabe & Wolf, 2009), after the devaluation of an appetitive outcome (Schwabe & Wolf, 2010), or before a sequential decision task (Quaedflieg et al., 2019) has been associated with more pronounced habitual approach (but see the failed replication attempts by Buabang, Boddez, et al., 2023; Smeets et al., 2023). These stress effects have been attributed to alterations in memory processes due to a release of glucocorticoids (e.g., Schwabe & Wolf, 2013; Wirz et al., 2018). In contrast, a goal-directed account of stress-induced alterations in outcome devaluation tasks proposed that seemingly habitual responses may result from stress regulation strategies (Buabang, Boddez, et al., 2023). Given our data, we cannot corroborate any of these two potential explanations for the stress effects in our study. Future studies incorporating more nuanced self-reports of the motivational processes during the test phase may further elucidate how task characteristics influence outcome devaluation effects via alterations in task-induced stress.

Of note, in the current study, higher trait anxiety was associated with stronger self-reported stress after the outcome devaluation procedure in the approach task but not the avoidance task, suggesting that highly trait-anxious individuals experienced more stress when the electrodes were attached to their arm than low trait-anxious participants. This amplification of stress in high trait-anxious individuals may speculatively have resulted from stronger sensitivity to threatful contexts (e.g., Aylward et al., 2019; Robinson et al., 2013) or preferential processing of threatening stimuli (e.g., Aupperle et al., 2023; Aupperle & Paulus, 2010; Cisler & Koster, 2010; Corr, 2013). Higher stress perceived by high trait-anxious participants due to the approach task devaluation may also have potentially have increased the perceived time pressure in the approach task's test phase, which has been suggested to amplify habitual responses (Raio et al., 2020). In support, higher trait anxiety was associated with faster responses in free, compatible, and incompatible trials in the approach task but not in the avoidance task. To facilitate the interpretation of such potential speed-accuracy trade-offs, future studies may

analyze reaction time data in conjunction with accuracy data using drift-diffusion models (e.g., Ratcliff & McKoon, 2008).

More highly trait-anxious individuals showed a descriptive tendency towards more pronounced accuracy compatibility effects in the approach task compared to the avoidance task. Speculatively, this difference may result from the already mentioned self-reported elevated stress perception in the test phase of the approach task in more highly trait-anxious individuals. This result suggests that the association between trait anxiety and overt habitual responses may underly boundary conditions, such as elevated stress. Relatedly, it has been reported that the impact of stress on habitual control depended on working memory capacity. Thus, stress accelerated the transition from goal-directed to habitual control more strongly in individuals with reduced working memory capacity (see Otto, Raio, et al., 2013; Quaedflieg et al., 2019). Trait anxiety has also been associated with impaired attentional control deficits only under high cognitive load since compensation for efficiency deficits in more highly trait-anxious individuals may then come to a limit (Berggren & Derakshan, 2013). Speculatively, the more stressful approach task phase burdened working memory processes in individuals with higher trait anxiety more strongly, and, thus, caused stronger habitual tendencies. Future studies may investigate this hypothesis, bridging attentional control theory (Berggren & Derakshan, 2013) and action control research.

The results of the avoidance training task in this study can directly be compared to a previous outcome devaluation study on habitual avoidance (Experiment 2 in Glück et al., 2021). Both studies were nearly identical except that, in the current study, a considerably higher number of picture stimuli was shown in the test phase (i.e., 180 instead of 9 different pictures within the three categories), which, arguably, increased the task difficulty in the current study. In support, the average accuracy in compatible, incompatible, and free trials in the current study was reduced by approximately 4% compared to the earlier study. Additionally, the reaction time compatibility effect and the compatibility effect in free trials that were apparent in the previous study were not replicated in the current study's avoidance training task. Potentially, in the previous study (Glück et al., 2021), the low task difficulty in the test phase allowed the participants to adjust their overt responses to the post-devaluation contingencies to not lose monetary rewards. Such adjustment may have eliminated the accuracy compatibility effect but preserved low-cost compatibility effects in reaction times and free trials. A speculative explanation may be that in the current avoidance training task, the higher task difficulty may have increased the attentional focus on the pictures in the center of the screen, potentially

reducing the impact of the background colors. A second difference in the results is that the correlations between the compatibility effects in the current study were absent. In contrast, in the previous experiment (Glück et al., 2021), the accuracy compatibility effect and the reaction time compatibility effect correlated positively. Potentially, the presentation of sixty different pictures per animal category caused secondary variance that may have reduced the reliability of each of the habit indicators (see Nebe et al., 2023). A third difference is that the accuracy compatibility effect tended to be more pronounced in later than earlier blocks of the test phase in the current study. However, in the previous study, it was most pronounced immediately after the outcome devaluation. This may indicate that in the current study, overt habitual responses may have been inhibited in favor of goal-directed responses at the beginning of the test phase. Perceived stress may, speculatively, have increased during the test phase, leading to the more pronounced habit indicator at the end of the test phase. Continuous tracking of stress levels during the test phase of outcome devaluation paradigms may be needed to test this assumption (see, e.g., Heller et al., 2018).

Heart rate variability as indexed by the SDNN did not predict habitual response strength. However, higher heart rate variability predicted less accurate and slower responses during both training phases and slower responses in compatible, incompatible, and neutral trials in the test phases of both tasks. Potentially, these results reflect lower perceived time pressure in individuals with higher heart rate variability during the tasks. In free trials where individuals could choose any response, however, higher heart rate variability predicted faster responses. Speculatively, this may reflect that individuals with higher heart rate variability perceived the decision-making process in free trials as less complex than individuals with lower heart rate variability. However, these two speculative explanations cannot be justified by the data since we did not collect self-reports on perceived time pressure or the perceived complexity of responding. Although the findings suggest that baseline heart rate and heart rate variability may be able to contribute to the explanation of responses in learning tasks (e.g., Howell & Hamilton, 2022, there is no evidence of an impact of SDNN and heart rate on habitual control. As mentioned in the limitations section, this may also be related to a potential lack of reliability of the outcome devaluation task used in this study. Of note, unexpectedly, trait anxiety, as measured with the STAI-T anxiety factor score, was uncorrelated with heart rate variability as indexed by the SDNN, potentially reflecting that SDNN is an unspecific indicator of overall autonomic functioning and arousal rather than a specific marker of trait anxiety (Kim et al., 2018; Riganello et al., 2012).

Several limitations need to be considered when interpreting the results of this study. First, in both tasks, the retrospective self-reported avoidance motivation was higher than the approach motivation, which may reflect that the subjective values of the outcomes may not have been entirely comparable. Of note, the outcomes' delivery differed in temporal proximity to the responses. While the aversive stimulations were administered immediately after each incorrect response, the rewards were not paid until the end of the study (i.e., with a temporal delay). Speculatively, some participants may not have believed that the financial rewards would be paid at the end of the study, which may have compromised the motivational valence of the reward. Additionally, while the intensity of the aversive outcome was individually calibrated, this was not the case for the financial rewards. Second, the straightforward interpretation of the experimental results was impeded by uncertainty about the direct comparability of the two task versions. The outcomes in the test phases were intended to limit potential undetected goals and thereby elevate the internal validity. However, the outcomes also introduced systematic differences in the task design since the aversive outcomes in the test phase of the approach task were more aversive and produced stronger stress than the reward outcomes in the test phase of the avoidance task. Third, the trait anxiety measure used in this study may have lacked interindividual variance and validity. Despite the pre-screening, the range of variance of trait anxiety may have been too restricted to detect specific trait anxiety effects. Additionally, the validity of the STAI-T anxiety subscale to measure trait anxiety is uncertain and may also, similarly to the STAI-T general score, measure unspecific psychological vulnerability (Bados et al., 2010; Balsamo et al., 2013; Knowles & Olatunji, 2020). Since most studies on the impact of trait anxiety on habitual responding have used the STAI-T sum score (e.g., Flores et al., 2018; Gillan et al., 2016; Gillan et al., 2014; Patterson et al., 2019), implementing the STAI-T anxiety subscale score may have enabled more direct compatibility with earlier studies. However, current trait anxiety measures may be too strongly affected by low validity to support detecting specific associations with behavior in outcome devaluation tasks (see Knowles & Olatunji, 2020). The inclusion of more clearly delineated measures of anxiety-related measures, such as, for example, intolerance of uncertainty (e.g., Boswell et al., 2013), may be beneficial in future clinically oriented action control studies. Fourth, the low correlations between the different habit indices in the tasks suggest a critical view of the reliability of these measures. Low reliability of behavioral tasks complicates the detection of interindividual differences in task performance (Enkavi et al., 2019). Systematically investigating the temporal stability of

individual outcome devaluation effects may advance research on interindividual differences in habitual control.

To summarize, we observed more pronounced indicators for habitual approach than habitual avoidance after an identical amount of training. However, the more pronounced habit indicator in the approach task may also reflect that the participants, on average, experienced higher stress levels in this task's test phase that involved aversive stimulations after incorrect goal-directed responses. Unexpectedly, the approach and avoidance task versions were, thus, not entirely parallel, which impedes direct comparisons. Concerning the effect of trait anxiety, higher trait anxiety predicted more pronounced low-cost habitual responses, as indicated by the reaction time compatibility effect. This effect of trait anxiety was independent of the task version. However, trait anxiety tended to additionally predict more pronounced costly habitual approach than habitual costly avoidance indicated by the accuracy compatibility effect. Of note, trait anxiety did not affect the performance in neutral control trials that were independent from the previous extensive training. Thus, trait anxiety only predicted goal-directed responses that were related to the previously extensively trained responses and did not predict impaired performance in the test phase in general. These results suggest an unspecific aberration of action control in more highly trait-anxious individuals after extensive training, which may be strengthened in threatening contexts. However, the study also emphasizes that details of outcome devaluation paradigms, such as the specific implementation of the outcome devaluation procedure, may decisively affect the habitual response measures. Therefore, the current results need to be interpreted cautiously. Future studies may compare the acquisition of habitual approach and avoidance with parallel tasks that do not differ regarding elicited emotions, associated cognitions, and stress.

5. General Discussion

This thesis investigated whether trait anxiety and anxiety disorders predict a stronger acquisition of habitual avoidance. Evidence of such potential associations may inform models on the maintenance of persistent avoidance behaviors in individuals with anxiety disorders. A potentially stronger tendency of individuals with anxiety disorders to develop habitual avoidance was proposed in the research literature repeatedly (e.g., Arnaudova et al., 2017; LeDoux & Daw, 2018; LeDoux et al., 2017; Pittig et al., 2020) but the evidence for such claims was not clear. Our studies added to the null findings concerning an amplification of avoidance habit acquisition in individuals with higher trait anxiety or anxiety disorders.

The outcome devaluation paradigm used in this thesis to measure habitual avoidance is a well-established experimental action control task but has received criticism, especially concerning internal validity (de Houwer et al., 2018; Moors et al., 2017; Watson & Wit, 2018). We aimed to address this critique by explicitly addressing the internal validity of the outcome devaluation tasks used in the thesis. Therefore, the tasks in this thesis differed from outcome devaluation studies commonly used in the research literature in several ways. As most significant difference, behavioral outcomes were presented in the test phases of the current studies to create a valid conflict between habitual and goal-directed responses. This conflict allowed us a) to reduce goal-directed strategies in the test phase, b) to test habits without relying on null differences between responses to valued and devalued stimuli, and c) to specifically operationalize costly habitual avoidance that was associated with a monetary cost. In two additional indicators for low-cost habitual avoidance, incompatible responses were not associated with a loss of reward. I will describe the potential impact of the behavioral outcomes in the test phase later in this discussion. Of note, we further adapted details of the test phases and the instructions (i.e., in Experiment 2 of Study 1 and, lastly in Study 3) to account for shortcomings that we identified during the research process, such as the trial-and-error learning in Experiment 1 of Study 1 and in Experiment 2. However, all studies in this thesis, except for the approach training task in the third study, featured identical training phases and outcome devaluation procedures. Therefore, the test phase result differences between the studies were, arguably, not caused by training differences, which allowed us to identify task characteristics that affected the habit indicators in the test phase.

In Study 1, we developed an outcome devaluation paradigm variation to investigate costly and non-costly indicators of habitual avoidance. In Experiment 1 of Study 1, costly, persistent

outcome devaluation effects were observed, which we tentatively interpreted as indicating costly habitual avoidance. In this experiment, the stimulus-response-outcome (i.e., S-R-O) contingencies after the outcome devaluation procedure were uninstructed, introducing a potential bias by a better-safe-than-sorry strategy. In Experiment 2 of Study 1, the S-R-O contingencies in the test phase were explicitly instructed, which, arguably, eliminated trial-and-error learning, but also reduced the task difficulty. In Experiment 2, we observed a small non-costly outcome devaluation effect. However, the costly outcome devaluation effect observed in Experiment 1 disappeared. The explicit instructions before the test phase of the outcome devaluation paradigm thus decisively reduced costly habitual avoidance, potentially due to a facilitation of competing goal-directed responses.

Using the experimental design without instructions, Study 2 demonstrated that participants with and without anxiety disorders showed a comparable acquisition of costly and low-costly habitual avoidance. Thus, we observed no stronger habitual avoidance acquisition in participants with anxiety disorders compared to healthy control participants. In an exploratory subgroup comparison, indicators of costly and non-costly habitual avoidance were more pronounced in participants with panic disorder than in participants with social anxiety disorder. Speculatively, this subgroup difference may have resulted from biases resulting from task features, since individuals with panic disorders may be more sensitive to the bodily effects of aversive electrotactile stimulations than individuals with social anxiety disorders. Generally, the results in Study 2 were subject to the same limitations concerning the interpretation of the results than the results of Experiment 1 in Study 1.

In Study 3, trait anxiety predicted slightly stronger indicators for low-cost habitual avoidance and approach. Trait anxiety also tended to predict a stronger indicator for costly habits after approach training than after extensive avoidance training. Of note, this effect was only approaching significance. Generally, independently from trait anxiety, we observed stronger indicators for acquired approach habits than for avoidance habits in this study. This task difference might have resulted from higher stress levels and higher perceived task difficulty in the approach task's test phase where incorrect responses were followed by aversive outcomes. Therefore, the two task versions for approach and avoidance habits were not entirely parallel versions, which was a limitation of the study. Heart rate variability and heart rate did not contribute to the prediction of habitual responses in this study. In sum, Study 3 did not indicate a specific tendency to acquire habitual avoidance in individuals with higher trait anxiety.

To summarize, although we observed significant indicators for habit acquisition in all three studies in the thesis, we did not find evidence for an increased acquisition of habitual avoidance in individuals with high trait anxiety or with anxiety disorders. The current studies, in this regard, concur with other studies reporting no elevated acquisition of habitual avoidance in individuals with elevated trait anxiety (Gillan et al., 2021; Patterson et al., 2019, but see Flores et al., 2018 for more ambiguous results) or anxiety disorders (Roberts et al., 2022). The observed positive association between trait anxiety and approach habit acquisition adds to the literature that includes null findings concerning a faster acquisition of habitual approach in higher trait anxiety (Gillan et al., 2016; Gillan et al., 2021) but also evidence for increased approach habits in individuals with social anxiety disorder (Alvares et al., 2014; Alvares et al., 2016).

The current studies suggest not only that trait anxiety and anxiety disorders were not associated with increased habitual avoidance, but also indicate that task features can have a pronounced effect on habitual response indicators. Specifically, the variations of the contingency instructions and the complexity of the S-R-O contingencies impacted the strength of the habit indicators. The studies highlight that outcome devaluation studies are sensitive to variations of the experimental design features, reflecting already existing criticisms of the validity and reliability of outcome devaluation study results (e.g., Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023; de Houwer et al., 2022; de Houwer et al., 2018; Moors et al., 2017). Clinically oriented action control research and basic action control research rely on valid measures of action control processes. Therefore, in the following, I will discuss the validity issues we encountered in the thesis and propose potential ways forward.

5.1 What did we measure? Issues of internal validity

A recurrent challenge in all studies of this thesis was the difficulty of interpreting the observed responses as indicators of habits. These difficulties were apparent even though the thesis addressed existing threats to internal validity of the paradigm, such as null results testing, the potential bias of outcome devaluation effects by goal-directed strategies, and incomplete outcome devaluation strategies (see Buabang, Boddez, et al., 2023; de Houwer, 2019; de Houwer et al., 2022; Moors et al., 2017). Thus, despite the adjustments we made, we encountered unclarities concerning how task details may have elicited or facilitated goal-directed strategies, suggesting a continued need for considering internal validity issues in action control research.

5.1.1 Task difficulty and the facilitation of goal-directed responses

The difficulty of goal-directed responses in the test phase was not equal in all experiments in this thesis. In Experiment 1 of Study 1 and Study 2, the participants needed to infer the S-R-O contingencies at the beginning of the test phase, which gave rise to trial-and-error learning. Therefore, unadjusted responses at the beginning of the test phase could have reflected habitual control or trial-and-error learning and unexpected strategies, such as a better-safe-than-sorry strategy, potentially biased the results. The large observed habit indicators immediately after the outcome devaluation procedure could, thus, not be interpreted as indicators for habitual avoidance. However, the habit indicators in these two uninstructed tasks were apparent until the end of the task when the participants had acquired knowledge of the new S-R-O associations. These habit indicators at the end of the test phase were interpreted as indicators for acquired habitual avoidance. In Experiment 2 of Study 1 and Study 3, the instructions before the test phase removed the need to infer the S-R-O contingencies, but also reduced the task difficulty in the test phase. There were also task difficulty differences between these two instructed tasks that may have impacted the results: in Experiment 2 of Study 1, which was a relatively easy task (i.e., nine stimuli needed to be categorized into three categories), only a slight reaction time compatibility effect emerged, indicating little habitual avoidance. In Study 3, where the test phase was more difficult (i.e., 180 different stimuli needed to be categorized), we observed no acquired habitual avoidance.

One explanation for an association between higher task difficulty and reduced habit effects may be that goal-directed responses may less effectively inhibit habitual tendencies in more complex tasks (see Hardwick et al., 2019). Under low task difficulty (i.e., when demands on cognitive resources are low), goal-directed responses may efficiently inhibit habitual responses (e.g., Strack & Deutsch, 2004). However, even though participants may be able to inhibit habitual responses to perform an incompatible goal-directed response in easy tasks, the participants may be unable to do so when performing the goal-directed responses is demanding. Thus, one might expect more pronounced outcome devaluation effects in tasks with more complex test phases. This hypothesis is backed up by the finding that task difficulty moderated the effect of low working memory capacity on habitual responses in one study (Otto, Raio, et al., 2013). Working memory capacity is needed for planning complex goal-directed responses. Thus, if goal-directed planning is difficult, goal-directed processes may be unable to inhibit competing habitual responses (see Hardwick et al., 2019). Of note, if goal-directed planning is too difficult, seemingly habitual responses may also result from error-prone or faulty goal-

directed processes (see Feher da Silva & Hare, 2020). Future studies may, therefore, aim to ensure that the participants understand the task instructions, for example, by including practice trials or post-experimental questionnaires on the participant's understanding of the task.

Systematically accounting for task difficulty effects may also benefit studies on the impact of anxiety on habitual control. Task-dependent effects of trait anxiety have already been proposed in the attentional control theory (Berggren & Derakshan, 2013) but have not been transferred to action control research systematically. Potentially, high trait anxiety, similar to low working memory capacity, may be associated with more pronounced habitual control only under high task difficulties that preclude the compensation of efficiency deficits in goal-directed control. If deficits in goal-directed control cannot be compensated, goal-directed processes may less effectively inhibit habitual tendencies, and increased habitual responses may be expected. Future action control studies on the effects of trait anxiety effects may, for example, vary the difficulty of goal-directed responses in the test phase while keeping the training duration and the outcome devaluation procedure constant.

5.1.2 Variations of outcome devaluation task as norm rather than exception

One unanswered question in action control research is why several studies in humans reported evidence for outcome devaluation effects (e.g., Gillan et al., 2015; Gillan et al., 2014; Gillan et al., 2011; Schwabe & Wolf, 2009, 2010; Tricomi et al., 2009) while other studies were unable to detect such effects (e.g., de Houwer et al., 2022; de Houwer et al., 2018; de Wit et al., 2018). The different habit indicators in the current studies suggest that the experimental variations decisively impacted the strength of habitual responses. Since the training duration was constant in all experiments, the differences in outcome devaluation effects between the studies cannot be attributed to differences in training. Of note, the amount of 100 repetitions per response in the training phase in the studies in this thesis was relatively high in comparison with other human studies that included training phases ranging from 21 repetitions per response (Alvares et al., 2014) to 30 repetitions (Gillan et al., 2014), 40 repetitions (Gillan et al., 2015; Roberts et al., 2022), and up to 98 repetitions (Zwosta et al., 2018). In animal studies, the training phase can consist of several hundred repetitions per response (Adams & Dickinson, 1981). The high and constant number of repetitions may have supported the internal validity and the comparability between the studies in this thesis.

In general, the observed impact of task difficulty as a result of different task instructions and S-R-O contingencies with different complexity, as well as the impact of different stress levels during the task as a result of different presented outcomes may imply that the

heterogeneity of results between different outcome devaluation studies can be partly explained by the different utilized outcome devaluation paradigm variations in the literature. In rodents, differences in the schedules of reinforcement, the number and type of the trained instrumental responses and the reinforcers, the training duration, and the exact outcome devaluation procedure have been suggested to impact the results in outcome devaluation paradigms (Perez & Dickinson, 2019; Watson et al., 2022). Although these features vary substantially within human action control studies, systematic investigations of their effects in humans are rare (but see the studies on different training durations, e.g., de Houwer et al., 2022; de Wit et al., 2018; Gera et al., 2023; Pool et al., 2022). The variety of paradigm variations without systematic investigations of their effects can be problematic since these variations may address slightly different underlying action regulation and learning mechanisms (Schreiner et al., 2020). Outcome devaluation studies, therefore, have been compared to a black box (Hommel, 2019). The term outcome devaluation paradigm may then be understood as an umbrella term for several tasks measuring different action-related learning mechanisms, cognitions, or strategies (Schreiner et al., 2020; Vandaele & Janak, 2018).

5.1.3 The complexity of goals in human action control

The internal validity of outcome devaluation tasks to measure goal-directed and habitual control in humans may be threatened when the task is translated from animal research without considering that humans can pursue more complex goals and strategies in the task. Humans can pursue multiple and conflicting goals with various complexities simultaneously (Moors et al., 2017). In contrast, memory, reasoning, and prospective planning are arguably less complex in rodents than humans (e.g., Azkona & Sanchez-Pernaute, 2022). Therefore, results in studies with rodents may be less biased by elaborate goals and goal-directed strategies than outcome devaluation tasks in humans. Relatedly, outcome devaluation procedures in animals have been described as implementing more robust devaluation procedures than outcome devaluation procedures in human studies (Buabang, Boddez, et al., 2023). Since it has become increasingly evident that undetected goals or strategies that result from ambiguous or incomplete outcome devaluation procedures pose a central threat to the internal validity of outcome devaluation tasks, less robust outcome devaluation procedures in humans may also impede direct translations of results between rodents and human studies (Buabang, Boddez, et al., 2023; de Houwer et al., 2018; Moors et al., 2017; Watson & Wit, 2018). Developing experimental designs that rule out potential biases by goals and strategies in the test phase of outcome devaluation studies may be the central aim of future action control research. One way to rule

out that goal-directed strategies bias responses in a devaluation paradigm may be the implementation of extensive pilot studies with post-experimental retrospective interviews to elucidate first-person experiences during the task, including the experienced cognitions and emotions (e.g., Petitmengin, 2006; Tewes, 2018).

5.1.4 Costly or low-cost habitual avoidance: a beneficial distinction?

The most salient difference between the current studies and commonly used outcome devaluation tasks was the introduction of costs for non-adjusted responses in the test phase of the current studies, which served three main purposes: First, the costs should alleviate the problem of non-adjusting responses as a beneficial strategy. By introducing costs for non-adjusted responses, the impact of cognitive strategies threatening the internal validity of the task, such as strategies to reduce cognitive effort or a better-safe-than-sorry strategy, should be reduced. Second, incorporating the concept of costly habitual avoidance aimed at a more externally valid operationalization of persistent maladaptive avoidance (see Krypotos et al., 2015). Third, presenting outcomes during the test phase allowed us to measure habitual responses with compatibility effects and avoided null hypothesis testing (see de Houwer et al., 2018; Watson & Wit, 2018), which is impossible when test phases are carried out in extinction. Fourth, the involvement of costs enabled us to investigate more than one single habit indicator simultaneously. Earlier outcome devaluation studies usually only analyzed the frequency but not the speed of responses that were compatible with the previously trained responses. If participants showed no differences between the frequencies of devalued and non-devalued responses, habitual responses were inferred in these studies. The analysis of reaction time and accuracy data in this study was uncommon, since most outcome devaluation studies focused solely on accuracy (but see Luque et al., 2019). However, there is evidence for the benefits of the analysis of two-step tasks with reaction times and accuracy as compared with accuracy alone (Shahar et al., 2019).

The use of three habit indicators in the current studies may have facilitated the detection of habitual behavior, which may be considered a benefit. In each study, we operationalized avoidance habits with the accuracy compatibility effect, the reaction time compatibility effect, and the compatibility effect in free trials. Repeatedly, only one or two of the three indicators were significantly affected by the extensive training (i.e., the accuracy compatibility effect and the free trial compatibility effect in Experiment 1 of Study 1 and Study 2; the significant reaction time compatibility effect in Experiment 2 of Study 1; the significant accuracy compatibility effect in the approach task in Study 3). The costly habit indicator (i.e., the

accuracy compatibility effect) was significantly pronounced in three of the four experiments, and may, therefore, considered to be relatively robust. However, if goal-directed behavior is very undemanding, overt habitual responses may be effectively inhibited, and reaction time effects may then be more sensitive habit indicators (Luque et al., 2019), as was tentatively suggested considering the results of Experiment 2 in Study 1. Therefore, the current studies may indicate that using costly and non-costly outcome devaluation effects allow for a more sensitive detection of habitual processes than using only one, non-costly, indicator that had been used in earlier outcome devaluation studies. Further studies may go beyond the limits of the current studies by analyzing reaction times and accuracy concurrently, for example, using drift-diffusion models (Johnson & Ratcliff, 2014).

The three habit indicators correlated differently within each study, indicating that they do not measure one shared construct or underlying process. In Experiment 1 of Study 1 and Study 2, which were very similar tasks, the costly habitual avoidance indicator correlated with both non-costly indicators of habitual avoidance. In Experiment 2 of Study 1, the costly habit indicator correlated with one of the non-costly indicators. In Study 3, which featured a relatively similar design, the indicators were not correlated at all. Therefore, the thesis could not resolve the ambiguity about the phenomena that are operationalized in outcome devaluation studies (e.g., de Houwer et al., 2018; Hommel, 2019). Of note, it has been proposed that goal-directed actions (Gillan et al., 2016) and habitual actions (Evans & Stanovich, 2013; Moors & de Houwer, 2006) are guided not by one but several underlying processes and would be better analyzed in terms of different contributing components than as presumably coherent, overarching constructs. The distinction between costly and low-cost habitual avoidance may support such a more precise conceptualization. Future research may aim to systematically describe which cognitive processes contribute to costly and low-cost habit indicators (see Schreiner et al., 2020).

Interestingly, a systematic association between trait anxiety and anxiety disorders with costly and low-cost habitual avoidance acquisition did not emerge in the current studies. This is a difference from goal-directed avoidance where individuals with higher trait anxiety (Pittig & Scherbaum, 2020) and anxiety disorders (Pittig, Boschet, et al., 2021) seem to present elevated costly but not low-cost avoidance. The specific tendency toward stronger goal-directed costly avoidance in individuals with high trait anxiety (i.e., Pittig & Scherbaum, 2020) or anxiety disorders (Pittig, Boschet, et al., 2021) has tentatively been explained by a reduced impact of competing rewards in these individuals. Since habitual responses are assumed to be

independent of the associated response outcomes, it is unsurprising that habitual avoidance, different from goal-directed avoidance, is not specifically affected by competing rewards.

Despite the benefits of including costs for habitual responses, the costs did not resolve all threats to the internal validity of the outcome devaluation paradigm. The costs may have affected the test phase results at least in three unintended ways. First, the different amount of costs accompanying non-adjusted behaviors may have influenced the degree of goal-directedness in the test phase. In the experiments that examined avoidance habits, the costs for habitual responses in incompatible trials were a loss of monetary rewards (i.e., Study 1, Study 2, and avoidance task of Study 3). In the task on approach habits in Study 3, non-adjusted responses in incompatible trials were, however, accompanied by aversive electrotactile stimulations. Interindividual differences concerning the estimations of the values of these costs may have affected the goal-directed responses and, therefore, the strength of competing habitual responses. Future studies may investigate whether higher costs for non-adjusted responses are associated with a higher motivation for goal-directed responses, and consequently, with smaller effects on costly habit indicators. This may being investigated by varying the magnitude of rewards (e.g., one Cent per correct answer vs. one Euro per correct answer) or aversive outcomes (e.g., slight electrical stimulation vs. aversive electrical stimulation) that are being used as costs.

The introduction of costs had several trickle-down effects concerning other features of the experimental design. First, a side effect of introducing costly and low-cost avoidance was the need for different conditions in the test phase. To create these conditions, more complex stimuli were presented in the test phase (i.e., stimuli combining background color and object picture) than in studies without costs, where identical stimuli are presented in training and test. This, arguably, increased the task difficulty of the current studies in comparison to studies with identical stimuli in training and test. As already explained, more difficult tasks may facilitate the detection of habitual responses via a less effective inhibition by goal-directed responses (Hardwick et al., 2019) which may explain why we found indicators for habitual responses in all current studies. Introducing new stimuli in the test phase also entailed that the background colors did not predict any of the outcomes in the test phase. Instead, the background colors that had been presented in the training phase can be seen as a distractor stimulus in the test phase. In this regard, the adapted outcome devaluation task resembles, for example, the Simon task or Stroop tasks that measure whether irrelevant task information interferes with intended behaviors (e.g., Proctor, 2011; Williams et al., 1996). Additionally, due to introducing four

conditions, the test phases in the outcome devaluation tasks in this thesis were considerably longer than those of earlier outcome devaluation paradigms, which ranged from, for example, four (Gillan et al., 2014) or ten (Gillan et al., 2016) to 50 trials per stimulus (Valentin et al., 2007) to a fixed duration of three minutes (Tricomi et al., 2009). However, although the more extensive test phase and three different habit indicators in the current thesis may have supported the detection of small outcome devaluation effects, they may also have introduced effects of boredom that may have reduced goal-directed responses throughout the test phase (see Meier et al., 2023). Boredom may decrease the motivation for goal-directed responses and, thereby, bias outcome devaluation task results. Future studies may collect data on the subjective level of boredom during or after the experiment to control for such potential effects.

5.1.5 Unintended stress effects as potential confounders

Although investigating stress effects on habitual responses was not an explicit aim of the current studies, stress may still have influenced their results. Potential unintended effects of stress were most apparent in Study 3, where aversive outcomes were presented in the test phase of the approach, but not the avoidance training task version. Higher stress in the test phase of the approach task version than in the avoidance task version as indicated by the retrospective self-reports in Study 3 may have partly caused the more pronounced habit indicators and the more pronounced effect of trait anxiety in the approach than in the avoidance task in this study, indicating that stress can influence outcome devaluation task results even without explicit stress manipulations.

One explanation for stress effects in outcome devaluation tasks is the stress-related release of stress hormones and subsequent changes in memory-related processes (Schwabe & Wolf, 2013; Wirz et al., 2018). Stress may also cause non-adjusted responses or a reduced adjustment of responses in experimental tasks due to giving rise to undetected goals and strategies. For example, stress may amplify preferences for undemanding behaviors (i.e., enhanced demand avoidance; Picciotto & Fabio, 2023). Stress may also activate goal-directed stress reduction strategies such as pursuing rewards to self-soothe (Buabang, Boddez, et al., 2023). Stress may also enhance the motivation to respond fast rather than accurately, which may impede goal-directed control and amplify habitual control (Raio et al., 2020). Participants with increased reactivity to experimentally induced stress (e.g., participants with psychiatric disorders) may use such stress-induced, goal-directed strategies more strongly than less sensitive participants, which may lead to a systematic, biased overestimation of habitual responses in stress-sensitive groups (Buabang, Boddez, et al., 2023). Since experiments on avoidance habits often involve

aversive stimuli that may induce stress, investigating such potential unintended stress effects may be one aim of future research on avoidance habits. To control for potential stress differences between different conditions or tasks, future studies may, for example, collect data on the subjective experiences of stress during the experiment (for this approach, see Heller et al., 2018), or in post-experimental retrospective interviews (see Petitmengin, 2006).

5.1.6 Self-efficacy and anhedonia as potential explanatory variables

Using delineated, clinically relevant individual characteristics as explanatory variables may potentially benefit action control studies on habitual avoidance in general. The construct of trait anxiety, but also current diagnostic classification systems have been criticised to suffer from unclear construct validity and reliability (e.g., Balsamo et al., 2013; Knowles & Olatunji, 2020; Roefs et al., 2022), which arguably impedes identifying clear associations in experimental studies. Gillan et al. (2016), relatedly, proposed that the transdiagnostic symptom dimension *compulsivity* may be a more promising candidate mechanism for explaining increased habitual approach tendencies than the broad diagnostic category of obsessive-compulsive disorder. Similarly, future studies on maladaptive persistent avoidance in anxiety disorders may include symptoms that are not directly related to fear and anxiety but frequently co-occur with anxiety disorders, such as reduced self-efficacy or anhedonia (Muris, 2002; Winer et al., 2017).

Elevated anhedonia and reduced self-efficacy have been discussed to contribute to persistent passive avoidance via a general reduction of active goal pursuit (e.g., Winer et al., 2017; Winer et al., 2019). Anhedonia, a symptom described as reduced liking and wanting, may specifically reduce the goal-directed approach of potentially rewarding outcomes, but may also decrease active avoidance of threats due to impaired relief learning (Heller et al., 2018; Leng et al., 2022). Such reduced activity levels may resemble elevated passive avoidance without being directly associated with threat expectations, fear, or trait anxiety. Self-efficacy describes an individual's expectations about their ability to act constructively and successfully in the face of demands or aversive situations (e.g., Raeder et al., 2019). Reduced self-efficacy may, therefore, reduce the approach to potentially threatening stimuli (Bandura, 1986). In depressive disorders, a cycle between depressive symptoms and reduced goal-directed activities has been extensively discussed as explanation for the maintenance of symptoms (e.g., Ferster, 1973; Grahek et al., 2019). Potentially, a general reduction of goal-directed activities may similarly contribute to maintaining persistent maladaptive passive avoidance in anxiety disorders (e.g., Winer et al., 2017; Winer et al., 2019). Nonetheless, how depression-related symptoms such as anhedonia

or reduced self-efficacy may specifically maintain maladaptive active and passive avoidance has yet to be systematically investigated in future studies (see Kalin, 2020).

Additionally, in outcome devaluation tasks, higher levels of anhedonia may be associated with a less positive valuation of rewards associated with goal-directed responses in the test phase, which may decrease motivation for goal-directed behavioral adjustment. Individuals with reduced self-efficacy may underestimate their capacity to control the outcomes in a task, which can even impair successful fear extinction (Raeder et al., 2019; Zlomuzica et al., 2015). In outcome devaluation tasks, reduced self-efficacy may negatively affect the self-perceived capability to effectively adjust responses in the test phase. Such reduced estimation of the efficacy of one's effort may lead to less successful initiation of goal-directed responses and potentially cause an overestimation of habitual responses in individuals with reduced self-efficacy. A similar potential mechanism concerning the effect of elevated anhedonia seems plausible. However, these hypotheses have not been investigated yet.

5.2 Where can we apply the results? Issues of external validity

Do response tendencies in outcome devaluation tasks contribute to our understanding of complex naturalistic maladaptive and persistent avoidance? Thus, can results derived from outcome devaluation tasks currently be applied diagnostically or therapeutically in clinical practice? The generalizability to contexts outside the laboratory (i.e., external validity) of experimental paradigms can be seen as a prerequisite for applying research findings in clinical interventions (Krypotos et al., 2018). Ensuring appropriate external validity of controlled laboratory experiments for naturalistic, biopsychosocial action control processes is not trivial (Field & Kersbergen, 2020; Watson et al., 2022). The generalizability of outcome devaluation task results has not been empirically investigated yet. One study investigated whether the performance in a laboratory action control task, the slips-of-action task, was associated with real-life habit formation as operationalized with the efficiency to adapting to differently colored keys to one's home, but did not find pronounced associations between experimental and real-life indicators for habit formation (Linnebank et al., 2018). Although attempts to operationalize habit formations in daily life with diary studies have been reported (Lally et al., 2010; Linnebank et al., 2018), such data have not yet been associated with habit tendencies as derived from outcome devaluation tasks

Of note, the lack of data on the generalizability of the laboratory evidence was not taken into account when a habit component of maladaptive avoidance was proposed based on the

available experimental data from outcome devaluation paradigms and two-step sequential decision-making tasks (e.g., Arnaudova et al., 2017; LeDoux & Daw, 2018; LeDoux et al., 2017; Pittig & Scherbaum, 2020; Pittig et al., 2020). The theoretical proposals were, thus, grounded on findings from outcome devaluation tasks without considering the current lack of evidence on the generalizability of these findings to avoidance behaviors in settings outside of the laboratory. In the following, threats to the external validity of outcome devaluations and potential solutions will be discussed.

5.2.1 Active and passive avoidance profiles

Habits have mostly been brought up to explain maladaptive behaviors characterized by a loss of control, some degree of stereotypicality, or a feeling of urge to perform the behavior. These characteristics are present, for example, in obsessive-compulsive disorder (Gillan et al., 2015; Gillan et al., 2014)or substance use disorders (Everitt & Robbins, 2016), which have been centered in clinically oriented habit research. In contrast, the degree of activity in avoidance can considerably vary from the performance of an avoidance response (i.e., active avoidance) to the inhibition of an approach response (i.e., passive avoidance; LeDoux & Daw, 2018). Differentiating between passive and active avoidance may be interesting since habitual responses may be more pronounced in individuals with stronger tendencies to avoid actively than in individuals with stronger passive avoidance tendencies. It has already been proposed that results from outcome devaluation paradigms may be more externally valid for explaining active than passive maladaptive avoidance responses (Roberts et al., 2022). A tendency to acquire habitual responses may, then, be a stronger risk factor for the development of persistent active avoidance than for the development of persistent passive avoidance. Persistent passive avoidance, in contrast, may, speculatively, be more associated with learned helplessness (Maier & Seligman, 2016; Seligman & Johnston, 1973), reduced self-efficacy, or elevated anhedonia, as has already been discussed earlier in this section. However, research on these hypotheses is scarce. Similarly, a comparison of the external validity of outcome devaluation paradigms to explain naturalistic active and passive avoidance has yet to be conducted. To do so, results from laboratory outcome devaluation studies might be associated with active and passive avoidance measures as obtained from experiential sampling techniques, observational, or interview studies.

5.2.2 Persistent avoidance in social contexts

Biopsychosocial models of human behavior assume that behavior is influenced by social, psychological, and biological processes (e.g., Lehman et al., 2017). Social influences on

maladaptive persistent avoidance may be conceptualized by taking into account that avoidance can be embedded and extended into external structures (e.g., Rowlands, 2010). The estimation of goal values, for example, may be influenced by social and cultural contexts (Oettingen et al., 2008). Experimental situations can also be understood as social contexts associated with socially influenced goals and social norms, as illustrated by social desirability effects in experiments (Larson, 2019; Nederhof, 1985). In this regard, one tentative hypothesis may state that, especially in outcome devaluation studies without costs for habitual responses, some individuals may adjust behaviors in the test phase because they consider this to be the expected behavior from attentive study participants, and they aim at adhering to this expectation. Future empirical studies on potential social desirability effects in outcome devaluation studies may shed light on such potential biases.

In naturalistic settings, action control may interact with or depend on the social structures surrounding the individual. For example, daily routines that facilitate or impair approach or avoidance may depend on the daily routines of others, such as coworkers, friends, or family members. The social context may, therefore, stabilize the maladaptive avoidance of an individual (Hunger-Schoppe et al., 2022). Attempts to modify maladaptive persistent avoidance may thus benefit from targeting not only the cognitions and behaviors of a patient but also their social contexts (Hunger-Schoppe et al., 2022). Systemic and family-based interventions have been demonstrated to be effective treatment options for anxiety disorders (Carr, 2016), but their specific effectiveness for reducing avoidance has not been demonstrated yet. Future models and studies on persistent maladaptive avoidance may integrate the current evidence on learning mechanisms on an individual level with specific social context effects.

The embeddedness of behavioral control in social contexts may also be apparent when individuals actively shape and adjust their environment (Rowlands, 2010). Certain decisions, such as deciding for a specific job or a place to live, may create structures that open up or restrict opportunities for avoidance and approach. Structures of daily life may, thus, reduce the necessity to actively avoid feared stimuli because the structure restricts encounters with potentially threatening stimuli or situations (Rowlands, 2010). For example, if an individual with social anxiety chooses a job with few social interactions, they may experience fewer social interactions in daily life without actively avoiding social situations. Potentially, actively adjusting environmental structures of life to facilitate opportunities for encountering feared situations may be relevant when individuals have created structures tailored to reduce such encounters with feared stimuli (Boyle, 2019). For example, a person with social anxiety may

become encouraged to intentionally add tasks to their job that are associated with meeting clients daily to facilitate opportunities for fear extinction. Such active creation of structures that facilitate fear extinction may tentatively be framed as promoting meta-learning but also as a strategy to increase self-efficacy, since individuals may become more aware that they can actively create environments with ample or poor opportunities for fear extinction learning. Similarly, it has already been proposed that changes in the structure of daily life, such as relocating, can be accompanied by a facilitation of the development of new goal-directed behaviors (Verplanken & Roy, 2016). Lastly, groups of individuals may be empowered to create lower levels of objective threats in their communities or neighborhoods to reduce the necessity to actively or passively avoid realistic threats in their daily lives (e.g., Fitzsimons & Fuller, 2002).

5.3 Clinical implications

The main finding of the current studies – i.e., that higher trait anxiety or anxiety disorders were not associated with a more pronounced acquisition of habitual avoidance – concurs with previous reports on null effects of trait anxiety and anxiety disorders on the acquisition of habitual avoidance (Flores et al., 2018; Gillan et al., 2014; Patterson et al., 2019; Roberts et al., 2022) and approach (Gillan et al., 2016; Gillan et al., 2021). Of note, this thesis did not address the potential development of habitual avoidance over the course of anxiety disorders as a simple result of repetition (see LeDoux et al., 2017; Arnaudova et al., 2017), which is, therefore, still a potential mechanism leading to persistent avoidance in anxiety disorders. To investigate this question, future studies may analyze longitudinal data on habitual avoidance and associate these data with the progression of anxiety disorders.

In the current studies, we were repeatedly confronted with the problem that several mechanisms besides habitual tendencies may have contributed to non-adjusted responses in our outcome devaluation tasks, which aligns with earlier critical accounts (Buabang, Boddez, et al., 2023; de Houwer et al., 2018; Moors et al., 2017). Arguably, the internal validity of outcome devaluation paradigms for measuring habitual control needs to be ascertained prior to applying the results in clinical settings. Future experimental studies on habitual avoidance may focus specifically on the internal and external validity of outcome devaluation paradigms (see Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023; de Houwer et al., 2018; Moors et al., 2017). Studies to ascertain the validity and reliability of operationalizations of habitual control may precede clinical applications of outcome devaluation task results. The already mentioned inclusion of habitual responses in the Research Domain Criteria Framework

(National Institute of Mental Health, 2023), for example, implies that habitual responses may be used as behavioral markers in diagnostic contexts. However, given the current lack of reliable and valid habit measures, such a diagnostic application of the habit concept seems to lack empirical justification (e.g., Buabang, Köster, et al., 2023; Das, 2015).

Given the current data and methodological challenges, a goal-directed perspective on maladaptive avoidance can be considered to constitute a more evidence-based path than a habit perspective for developing interventions to reduce maladaptive persistent avoidance. A similar proposal has been discussed for the role of habits in addiction disorders, given the lack of reliable and valid research, but also due to the complex biopsychosocial influences on human behavior (Buabang, Boddez, et al., 2023; Field & Kersbergen, 2020). Potential interventions may, for example, target the gap between wanting to act adaptively and but finding oneself again and again to behave maladaptively (Gollwitzer, 1999). One way to support the reduction of maladaptive avoidance may be to support the implementation of approach motivations. Implementation intentions, also termed "if-then planning", aim to support the development of close associations between environmental cues and goal-directed behaviors (Gollwitzer, 1999). One example of an implementation intention that may facilitate goal-directed social approach is "If I see my colleague this morning, I will say Hello" or "If my colleagues are going for dinner today, I will join them". Implementation intentions are goal-directed since the association between actions and environmental cues is explicitly planned. However, implementation intentions are assumed to reduce the cognitive effort needed for goal-directed actions via direct associations between environmental stimuli and actions. Meta-analytic evidence indicates the effectiveness of implementation intentions to support behavioral changes in a variety of mental disorders, including anxiety disorders (Toli et al., 2016). Experimental evidence suggests that approach implementation intentions can mitigate attentional threat biases in socially anxious individuals (Webb et al., 2010) and reduce avoidance frequency in healthy individuals (Karsdorp et al., 2016). However, the effectiveness of implementation intentions for mitigating maladaptive avoidance in anxiety disorders has not been investigated yet. Studies may, for example, address whether implementation intentions can facilitate the translation of approach intentions from therapeutic settings to daily life.

5.4 Outlook

This thesis investigated the development of habitual avoidance in trait anxiety and anxiety disorders while simultaneously addressing validity concerns regarding the operationalization of action control processes. We did not find evidence for enhanced acquisition of avoidance habits

in individuals with anxiety disorders and trait anxiety. Of note, we did not investigate whether avoidance habits may result from the mere repetition of avoidance in anxiety disorders. Therefore, the studies do not allow inferences on this specific mechanism. Additionally, the current studies did not systematically address whether trait anxiety or anxiety disorders are associated with habitual avoidance only under boundary conditions, such as tasks with high cognitive demands. Exploring the internal and external validity of the experimental paradigms in the associative dual-process framework can be considered a central topic to ensure appropriate and meaningful interpretations of experimental results. I would like to conclude by suggesting three potential routes for further investigations.

First, this thesis did not investigate whether habitual, outcome-insensitive avoidance may develop as a result of avoidance repetition over the course of anxiety disorders (e.g., LeDoux & Daw, 2018). Instead, the thesis investigated a potentially more pronounced transition between goal-directed and habitual avoidance in individuals with higher trait anxiety or with anxiety disorders. There is currently, to our best knowledge, no evidence of a longitudinal shift from goal-directed to habitual avoidance behaviors in anxiety disorders. Obtaining such evidence would, however, be interesting to assess the potential development of habitual avoidance as a function of repetition only. Longitudinal studies may allow investigating a potential increase of habitual avoidance during the development of anxiety disorders.

Second, future studies may intensify existing efforts to ascertain an adequate internal validity of action research tasks. One central question, in this regard, is to ensure whether non-adjusted responses in outcome devaluation tasks result from goals or goal-directed strategies (see Buabang, Boddez, et al., 2023; Buabang, Köster, et al., 2023). Incorporating detailed qualitative interviews following experimental tasks or in pilot studies may allow to capture task-related goals or strategies of participants (Petitmengin, 2006). Similarly, collecting data on participants' subjective experiences during experimental tasks may support the development of experimental paradigms that are less biased by, for instance, boredom, reduced self-efficacy, or elevated anhedonia that may influence the results of outcome devaluation tasks due to a systematic reduction of goal-directedness (e.g., Meier et al., 2023).

To elucidate the impact of task variations, future studies may systematically compare how differences in action control task designs influence the obtained action control parameters. In this regard, studies may compare response-outcome contingency degradation procedures and outcome devaluation procedures, but also different outcome devaluation procedures, such as selective satiation, pairings of the to-be-devalued outcome with aversive outcomes, or the

complete removal of the outcomes. Such systematic comparisons may allow to develop more precise concepts of the phenomena that are being measured. Another theoretical approach to increase internal validity may be to focus on the specific effects of attention and memory processes in outcome devaluation studies. Albeit challenging, this may support delineating potential boundary conditions of trait anxiety effects and identify attention- and memory-associated subcomponents of habitual and goal-directed action control. For example, future studies on trait anxiety effects on avoidance control may associate interindividual differences of attentional control efficiency and action control processes (e.g., Berggren & Derakshan, 2013; Cisler & Koster, 2010).

Third, future research may understand persistent maladaptive avoidance as behavior situated within social contexts. A change of perspective in action control research may be to ask how social structures in daily life, neighborhoods, communities, or families influence approach and avoidance behaviors. The studies in this thesis suggest that trait anxiety and anxiety disorders do not strongly impact whether we act habitually or goal-directedly in laboratory tasks. The thesis also showed how laborious designing internally and externally valid experimental tasks in the realm of action control is. It may be worthwhile to consider using research strategies that bridge individual behaviors and the surrounding social contexts. Adopting a person-centered, socio-ecological research perspective in addition to laboratory research (Department of Community Health, 2022) would be theoretically and methodologically challenging but may generate new information on the maintaining conditions of maladaptive, persistent avoidance.

6. References

Ach, N. (1910). *Über den Willensakt und das Temperament: Eine experimentelle Untersuchung*. Quelle & Meyer.

Adams, C. D. (1982). Variations in the Sensitivity of Instrumental Responding to Reinforcer Devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *34*(2b), 77–98. https://doi.org/10.1080/14640748208400878

Adams, C. D., & Dickinson, A. (1981). Instrumental Responding following Reinforcer Devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*(2b), 109–121. https://doi.org/10.1080/14640748108400816

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Computational Biology*, *11*(12), e1004648. https://doi.org/10.1371/journal.pcbi.1004648

Akerstedt, T., Anund, A., Axelsson, J., & Kecklund, G. (2014). Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of Sleep Research*, *23*(3), 240–252. https://doi.org/10.1111/jsr.12158

Akerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *The International Journal of Neuroscience*, *52*(1-2), 29–37. https://doi.org/10.3109/00207459008994241

Allan, N. P., Raines, A. M., Capron, D. W., Norr, A. M., Zvolensky, M. J., & Schmidt, N. B. (2014). Identification of anxiety sensitivity classes and clinical cut-scores in a sample of adult smokers: Results from a factor mixture model. *Journal of Anxiety Disorders*, *28*(7), 696–703. https://doi.org/10.1016/j.janxdis.2014.07.006

Alvares, G. A., Balleine, B. W., & Guastella, A. J. (2014). Impairments in goal-directed actions predict treatment response to cognitive-behavioral therapy in social anxiety disorder. *PloS One*, *9*(4), e94778. https://doi.org/10.1371/journal.pone.0094778

Alvares, G. A., Balleine, B. W., Whittle, L., & Guastella, A. J. (2016). Reduced goal-directed action control in autism spectrum disorder. *Autism Research : Official Journal of the International Society for Autism Research*, *9*(12), 1285–1293. https://doi.org/10.1002/aur.1613

Alvares, G. A., Quintana, D. S., Kemp, A. H., van Zwieten, A., Balleine, B. W., Hickie, I. B., & Guastella, A. J. (2013). Reduced heart rate variability in social anxiety disorder: Associations with gender and symptom severity. *PloS One*, *8*(7), e70468. https://doi.org/10.1371/journal.pone.0070468

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Andlin-Sobocki, P., Jönsson, B., Wittchen, H.-U., & Olesen, J. (2005). Cost of disorders of the brain in Europe. *European Journal of Neurology, 12 Suppl 1*, 1–27. https://doi.org/10.1111/j.1468-1331.2005.01202.x

Arnaudova, I., Kindt, M., Fanselow, M., & Beckers, T. (2017). Pathways towards the proliferation of avoidance in anxiety and implications for treatment. *Behaviour Research and Therapy*, *96*, 3–13. https://doi.org/10.1016/j.brat.2017.04.004

Aupperle, R. L., McDermott, T. J., White, E., & Kirlic, N. (2023). The neuropsychology of anxiety: An approach–avoidance decision-making framework. In G. G. Brown, T. Z. King, K. Y. Haaland, & B. Crosson (Eds.), *APA handbook of neuropsychology, Volume 1: Neurobehavioral disorders and conditions: Accepted science and open questions (Vol. 1)* (pp. 767–787). American Psychological Association. https://doi.org/10.1037/0000307-036

Aupperle, R. L., & Paulus, M. P. (2010). Neural systems underlying approach and avoidance in anxiety disorders. *Dialogues in Clinical Neuroscience*, *12*(4), 517–531. https://doi.org/10.31887/DCNS.2010.12.4/raupperle

Aylward, J., Valton, V., Ahn, W.-Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature Human Behaviour*, *3*(10), 1116–1123. https://doi.org/10.1038/s41562-019-0628-0

Azkona, G., & Sanchez-Pernaute, R. (2022). Mice in translational neuroscience: What R we doing? *Progress in Neurobiology*, *217*, 102330. https://doi.org/10.1016/j.pneurobio.2022.102330

Bados, A., Gómez-Benito, J., & Balaguer, G. (2010). The state-trait anxiety inventory, trait version: Does it really measure anxiety? *Journal of Personality Assessment*, *92*(6), 560–567. https://doi.org/10.1080/00223891.2010.513295

Ball, T. M., & Gunaydin, L. A. (2022). Measuring maladaptive avoidance: From animal models to clinical anxiety. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *47*(5), 978–986. https://doi.org/10.1038/s41386-021-01263-4

Balleine, B. W., & Dezfouli, A. (2019). Hierarchical Action Control: Adaptive Collaboration Between Actions and Habits. *Frontiers in Psychology*, *10*, 2735. https://doi.org/10.3389/fpsyg.2019.02735

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407–419. https://doi.org/10.1016/S0028-3908(98)00033-1

Balleine, B. W., & Dickinson, A. (2005). Effects of outcome devaluation on the performance of a heterogeneous instrumental chain. *International Journal of Comparative Psychology*, *18*(4).

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Balsamo, M., Romanelli, R., Innamorati, M., Ciccarese, G., Carlucci, L., & Saggino, A. (2013). The State-Trait Anxiety Inventory: Shadows and Lights on its Construct Validity. *Journal of Psychopathology and Behavioral Assessment*, *35*(4), 475–486. https://doi.org/10.1007/s10862-013-9354-5

Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience*, *17*(3), 327–335. https://doi.org/10.31887/DCNS.2015.17.3/bbandelow

Bandura, A. (1986). Fearful expectations and avoidant actions as coeffects of perceived self-inefficacy. *The American Psychologist*, *41*(12), 1389–1391. https://doi.org/10.1037/0003-066X.41.12.1389

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, *133*(1), 1–24. https://doi.org/10.1037/0033-2909.133.1.1

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Belin, D., Belin-Rauscent, A., Murray, J. E., & Everitt, B. J. (2013). Addiction: Failure of control over maladaptive incentive habits. *Current Opinion in Neurobiology*, *23*(4), 564–572. https://doi.org/10.1016/j.conb.2013.01.025

Berggren, N., & Derakshan, N. (2013). Attentional control deficits in trait anxiety: Why you see them and why you don't. *Biological Psychology*, *92*(3), 440–446. https://doi.org/10.1016/j.biopsycho.2012.03.007

Berggren, N., & Eimer, M. (2021). The role of trait anxiety in attention and memory-related biases to threat: An event-related potential study. *Psychophysiology*, *58*(3), e13742. https://doi.org/10.1111/psyp.13742

Berner, L. A., Fiore, V. G., Chen, J. Y., Krueger, A., Kaye, W. H., Viranda, T., & Wit, S. de (2023). Impaired belief updating and devaluation in adult women with bulimia nervosa. *Translational Psychiatry*, *13*(1), 2. https://doi.org/10.1038/s41398-022-02257-6

Bieling, P. J., Antony, M. M., & Swinson, R. P. (1998). The State-Trait Anxiety Inventory, Trait version: Structure and content re-examined. *Behaviour Research and Therapy*, *36*(7-8), 777–788. https://doi.org/10.1016/S0005-7967(98)00023-0

Boswell, J. F., Thompson-Hollands, J., Farchione, T. J., & Barlow, D. H. (2013). Intolerance of uncertainty: A common factor in the treatment of emotional disorders. *Journal of Clinical Psychology*, *69*(6), 630–645. https://doi.org/10.1002/jclp.21965

Bouton, M. E. (2018). *Learning and behavior: A contemporary synthesis* (2nd ed.). Sinauer Associates.

Boyle, L. E. (2019). The (un)habitual geographies of Social Anxiety Disorder. *Social Science & Medicine (1982)*, *231*, 31–37. https://doi.org/10.1016/j.socscimed.2018.03.002

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. https://doi.org/10.1016/0005-7916(94)90063-9

Brain Products GmbH. (2018a). *BrainVision Analyzer (Version 2.1)* [Computer software].

Brain Products GmbH. (2018b). *BrainVision Recorder (Version 1.21.0402)* [Computer software].

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Brown, T. A., Campbell, L. A., Lehman, C. L., Grisham, J. R., & Mancill, R. B. (2001). Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *Journal of Abnormal Psychology*, *110*(4), 585–599. https://doi.org/10.1037//0021-843x.110.4.585

Bruce, S. E., Yonkers, K. A., Otto, M. W., Eisen, J. L., Weisberg, R. B., Pagano, M., Shea, M. T., & Keller, M. B. (2005). Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: A 12-year prospective study. *The American Journal of Psychiatry*, *162*(6), 1179–1187. https://doi.org/10.1176/appi.ajp.162.6.1179

Buabang, E. K., Boddez, Y., Wolf, O. T., & Moors, A. (2023). The role of goal-directed and habitual processes in food consumption under stress after outcome devaluation with taste aversion. *Behavioral Neuroscience*, *137*(1), 1–14. https://doi.org/10.1037/bne0000439

Buabang, E. K., Köster, M., Hogarth, L., & Moors, A. (2023). *Poor Reliability and Validity of Habit Effects in Substance Use and Novel Insights From a Goal-Directed Perspective.* https://doi.org/10.31234/osf.io/79ykb

Bublatzky, F., Alpers, G. W., & Pittig, A. (2017). From avoidance to approach: The influence of threat-of-shock on reward-based decision making. *Behaviour Research and Therapy*, *96*, 47–56. https://doi.org/10.1016/j.brat.2017.01.003

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Cain, C. K. (2019). Avoidance Problems Reconsidered. *Current Opinion in Behavioral Sciences*, *26*, 9–17. https://doi.org/10.1016/j.cobeha.2018.09.002

Carr, A. (2016). How and Why Do Family and Systemic Therapies Work? *Australian and New Zealand Journal of Family Therapy*, *37*(1), 37–55. https://doi.org/10.1002/anzf.1135

Carvalho, J. P., & Hopko, D. R. (2011). Behavioral theory of depression: Reinforcement as a mediating variable between avoidance and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *42*(2), 154–162. https://doi.org/10.1016/j.jbtep.2010.10.001

Chalmers, J. A., Quintana, D. S., Abbott, M. J.-A., & Kemp, A. H. (2014). Anxiety Disorders are Associated with Reduced Heart Rate Variability: A Meta-Analysis. *Frontiers in Psychiatry*, *5*, 80. https://doi.org/10.3389/fpsyt.2014.00080

Christiana, J. M., Gilman, S. E., Guardino, M., Mickelson, K., Morselli, P. L., Olfson, M., & Kessler, R. C. (2000). Duration between onset and time of obtaining initial treatment among people with anxiety and mood disorders: An international survey of members of mental health patient advocate groups. *Psychological Medicine*, *30*(3), 693–703. https://doi.org/10.1017/S0033291799002093

Cisler, J. M., & Koster, E. H. W. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clinical Psychology Review*, *30*(2), 203–216. https://doi.org/10.1016/j.cpr.2009.11.003

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews. Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Colwill, R. M., Delamater, A. R., & Lattal, K. M. (2022). Developments in associative theory: A tribute to the contributions of Robert A. Rescorla. *Journal of Experimental Psychology. Animal Learning and Cognition*, *48*(4), 245–264. https://doi.org/10.1037/xan0000344

Colwill, R. M., & Rescorla, R. A. (1985). Instrumental responding remains sensitive to reinforcer devaluation after extensive training. *Journal of Experimental Psychology. Animal Behavior Processes*, *11*(4), 520–536. https://doi.org/10.1037/0097-7403.11.4.520

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Corr, P. J. (2013). Approach and Avoidance Behaviour: Multiple Systems and their Interactions. *Emotion Review*, *5*(3), 285–290. https://doi.org/10.1177/1754073913477507

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

Craske, M. G., Stein, M. B., Eley, T. C., Milad, M. R., Holmes, A., Rapee, R. M., & Wittchen, H.-U. (2017). Anxiety disorders. *Nature Reviews. Disease Primers*, *3*, 17024. https://doi.org/10.1038/nrdp.2017.24

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13817–13822. https://doi.org/10.1073/pnas.1506367112

Da Feher Silva, C., & Hare, T. A. (2018). A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict

about the effects of reward and transition on the stay probability. *PloS One*, *13*(4), e0195328. https://doi.org/10.1371/journal.pone.0195328

Das, J. P. (2015). Three Faces of Cognitive Processes. In *Cognition, Intelligence, and Achievement* (pp. 19–47). Elsevier. https://doi.org/10.1016/B978-0-12-410388-7.00003-8

Daw, N. D. (2015). Of goals and habits. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13749–13750. https://doi.org/10.1073/pnas.1518488112

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks : The Official Journal of the International Neural Network Society*, *22*(3), 213–219. https://doi.org/10.1016/j.neunet.2009.03.004

de Houwer, J. (2019). On How Definitions of Habits Can Complicate Habit Research. *Frontiers in Psychology*, *10*, 2642. https://doi.org/10.3389/fpsyg.2019.02642

de Houwer, J., Buabang, E. K., Boddez, Y., Köster, M., & Moors, A. (2022). Reasons to Remain Critical About the Literature on Habits: A Commentary on Wood et al. (2022). *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 17456916221131508. https://doi.org/10.1177/17456916221131508

de Houwer, J., Tanaka, A., Moors, A., & Tibboel, H. (2018). Kicking the habit: Why evidence for habits in humans might be overestimated. *Motivation Science*, *4*(1), 50–59. https://doi.org/10.1037/mot0000065

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology. General*, *147*(7), 1043–1065. https://doi.org/10.1037/xge0000402

Delorme, C., Salvador, A., Valabrègue, R., Roze, E., Palminteri, S., Vidailhet, M., Wit, S. de, Robbins, T. W., Hartmann, A., & Worbe, Y. (2016). Enhanced habit formation in Gilles de la Tourette syndrome. *Brain : A Journal of Neurology*, *139*(Pt 2), 605–615. https://doi.org/10.1093/brain/awv307

Department of Community Health (Ed.). (2022). *Community Health: Grundlagen, Methoden, Praxis* (1. Auflage). Beltz Juventa.

Dickinson, A. (1985). Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010

Dickinson, A., Balleine, B. W., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, *23*(2), 197–206. https://doi.org/10.3758/BF03199935

Digitimer Ltd. (n.d.). *DS7A High Voltage Constant Current Stimulator* [Apparatus].

Dinsmoor, J. A. (1954). Punishment. I. The avoidance hypothesis. *Psychological Review*, *61*(1), 34–46. https://doi.org/10.1037/h0062725

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology : CB*, *30*(15), R860-R865. https://doi.org/10.1016/j.cub.2020.06.051

Du, Y., Krakauer, J. W., & Haith, A. M. (2022). The relationship between habits and motor skills in humans. *Trends in Cognitive Sciences*, *26*(5), 371–387. https://doi.org/10.1016/j.tics.2022.02.002

Dymond, S. (2019). Overcoming avoidance in anxiety disorders: The contributions of Pavlovian and operant avoidance extinction methods. *Neuroscience and Biobehavioral Reviews*, *98*, 61–70. https://doi.org/10.1016/j.neubiorev.2019.01.007

Elwood, L. S., Wolitzky-Taylor, K., & Olatunji, B. O. (2012). Measurement of anxious traits: A contemporary review and synthesis. *Anxiety, Stress, and Coping*, *25*(6), 647–666. https://doi.org/10.1080/10615806.2011.582949

Endler, N. S., & Kocovski, N. L. (2001). State and trait anxiety revisited. *Journal of Anxiety Disorders*, *15*(3), 231–245. https://doi.org/10.1016/S0887-6185(01)00060-3

Engelhard, I. M., van Uijen, S. L., van Seters, N., & Velu, N. (2015). The Effects of Safety Behavior Directed Towards a Safety Cue on Perceptions of Threat. *Behavior Therapy*, *46*(5), 604–610. https://doi.org/10.1016/j.beth.2014.12.006

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Enkavi, A. Z., & Poldrack, R. A. (2021). Implications of the Lacking Relationship Between Cognitive Task and Self-report Measures for Psychiatry. *Biological Psychiatry.*

Cognitive Neuroscience and Neuroimaging, *6*(7), 670–672. https://doi.org/10.1016/j.bpsc.2020.06.010

Ernst, M., & Paulus, M. P. (2005). Neurobiology of decision making: A selective review from a neurocognitive and clinical perspective. *Biological Psychiatry*, *58*(8), 597–604. https://doi.org/10.1016/j.biopsych.2005.06.004

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*, *67*, 23–50. https://doi.org/10.1146/annurev-psych-122414-033457

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and Performance: The Processing Efficiency Theory. *Cognition & Emotion*, *6*(6), 409–434. https://doi.org/10.1080/02699939208409696

Eysenck, M. W., & Derakshan, N. (2011). New perspectives in attentional control theory. *Personality and Individual Differences*, *50*(7), 955–960. https://doi.org/10.1016/j.paid.2010.08.019

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion (Washington, D.C.)*, *7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*. Advance online publication. https://doi.org/10.1038/s41562-020-0905-y

Ferster, C. B. (1973). A functional analysis of depression. *The American Psychologist*, *28*(10), 857–870. https://doi.org/10.1037/h0035605

Field, M., & Kersbergen, I. (2020). Are animal models of addiction useful? *Addiction (Abingdon, England)*, *115*(1), 6–12. https://doi.org/10.1111/add.14764

Fitzsimons, S., & Fuller, R. (2002). Empowerment and its implications for clinical practice in mental health: A review. *Journal of Mental Health*, *11*(5), 481–499. https://doi.org/10.1080/09638230020023

Flores, A., López, F. J., Vervliet, B., & Cobos, P. L. (2018). Intolerance of uncertainty as a vulnerability factor for excessive and inflexible avoidance behavior. *Behaviour Research and Therapy*, *104*, 34–43. https://doi.org/10.1016/j.brat.2018.02.008

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third edition). SAGE.

Friedel, E., Koch, S. P., Wendt, J [Jean], Heinz, A., Deserno, L., & Schlagenhauf, F. (2014). Devaluation and sequential decisions: Linking goal-directed and model-based behavior. *Frontiers in Human Neuroscience*, *8*, 587. https://doi.org/10.3389/fnhum.2014.00587

Friedman, B. H. (2007). An autonomic flexibility-neurovisceral integration model of anxiety and cardiac vagal tone. *Biological Psychology*, *74*(2), 185–199. https://doi.org/10.1016/j.biopsycho.2005.08.009

Garr, E., Padovan-Hernandez, Y., Janak, P. H., & Delamater, A. R. (2021). Maintained goal-directed control with overtraining on ratio schedules. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *28*(12), 435–439. https://doi.org/10.1101/lm.053472.121

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. *EBL-Schweitzer*. Cambridge University Press. http://swb.eblib.com/patron/FullRecord.aspx?p=288457

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., Zheng, T., & Dorie, V. (2022). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models* (Version 1.12-2) [Computer software]. https://CRAN.R-project.org/package=arm

Gera, R., Bar Or, M., Tavor, I., Roll, D., Cockburn, J., Barak, S., Tricomi, E., O'Doherty, J. P., & Schonberg, T. (2023). Characterizing habit learning in the human brain at the individual and group levels: A multi-modal MRI study. *NeuroImage*, *272*, 120002. https://doi.org/10.1016/j.neuroimage.2023.120002

Gillan, C., Apergis-Schoute, A. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2015). Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. *The American Journal of Psychiatry*, *172*(3), 284–293. https://doi.org/10.1176/appi.ajp.2014.14040525

Gillan, C., 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*. https://doi.org/10.7554/eLife.11305

Gillan, C., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M., Fineberg, N. A., Sahakian, B. J., & Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological Psychiatry*, *75*(8), 631–638. https://doi.org/10.1016/j.biopsych.2013.02.002

Gillan, C., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *The American Journal of Psychiatry*, *168*(7), 718–726. https://doi.org/10.1176/appi.ajp.2011.10071062

Gillan, C., & Robbins, T. W. (2014). Goal-directed learning and obsessive-compulsive disorder. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1655). https://doi.org/10.1098/rstb.2013.0475

Gillan, C., Vaghi, M. M., Hezemans, F. H., van Ghesel Grothe, S., Dafflon, J., Brühl, A. B., Savulich, G., & Robbins, T. W. (2021). Experimentally induced and real-world anxiety have no demonstrable effect on goal-directed behaviour. *Psychological Medicine*, *51*(9), 1467–1478. https://doi.org/10.1017/S0033291720000203

Glück, V. M., Boschet-Lange, J. M., Pittig, R., & Pittig, A. (2023). *Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm: Manuscript submitted for publication.*

Glück, V. M., Zwosta, K., Wolfensteller, U., Ruge, H., & Pittig, A. (2021). Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm. *Behaviour Research and Therapy*, *146*, 103964. https://doi.org/10.1016/j.brat.2021.103964

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *The American Psychologist*, *54*(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493

Graf, A. (2019). *Boxenstopp für eine A380 - Putzen, checken, tanken: Eine Reportage von Andreas Graf* [Video]. Hessischer Rundfunk.

Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neuroscience and Biobehavioral Reviews*, *102*, 371–381. https://doi.org/10.1016/j.neubiorev.2019.04.011

Hardwick, R. M., Forrence, A. D., Krakauer, J. W., & Haith, A. M. (2019). Time-dependent competition between goal-directed and habitual response preparation. *Nature Human Behaviour*, *3*(12), 1252–1262. https://doi.org/10.1038/s41562-019-0725-0

Hartley, C. A., & Phelps, E. A. (2012). Anxiety and decision-making. *Biological Psychiatry*, *72*(2), 113–118. https://doi.org/10.1016/j.biopsych.2011.12.027

HealthMeasures. (2019). *PROMIS Anxiety Scoring Manual*. https://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Anxiety_Scoring_Manual.pdf

Heeren, A., & McNally, R. J. (2018). Social Anxiety Disorder as a Densely Interconnected Network of Fear and Avoidance for Social Situations. *Cognitive Therapy and Research*, *42*(1), 103–113. https://doi.org/10.1007/s10608-017-9876-3

Heinzen, E., Sinnwell, J., Atkinson, E., Gunderson, T., & Dougherty, G. (2021). *arsenal: An Arsenal of 'R' Functions for Large-Scale Statistical Summaries. R package version 3.6.3 [*Computer software]. Ethan Heinzen, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson and Gregory Dougherty. https://CRAN.R-project.org/package=arsenal

Heller, A. S., Ezie, C. E. C., Otto, A. R., & Timpano, K. R. (2018). Model-based learning and individual differences in depression: The moderating role of stress. *Behaviour Research and Therapy*, *111*, 19–26. https://doi.org/10.1016/j.brat.2018.09.007

Hendriks, S. M., Spijker, J., Licht, C. M. M., Hardeveld, F., Graaf, R. de, Batelaan, N. M., Penninx, B. W. J. H., & Beekman, A. T. F. (2016). Long-term disability in anxiety disorders. *BMC Psychiatry*, *16*, 248. https://doi.org/10.1186/s12888-016-0946-y

Hofmann, S. G., & Hay, A. C. (2018). Rethinking avoidance: Toward a balanced approach to avoidance in treating anxiety disorders. *Journal of Anxiety Disorders*, *55*, 14–21. https://doi.org/10.1016/j.janxdis.2018.03.004

Hommel, B. (2019). Binary Theorizing Does Not Account for Action Control. *Frontiers in Psychology*, *10*, 2542. https://doi.org/10.3389/fpsyg.2019.02542

Howell, B. C., & Hamilton, D. A. (2022). Baseline heart rate variability (HRV) and performance during a set-shifting visuospatial learning task: The moderating effect of trait negative affectivity (NA) on behavioral flexibility☆. *Physiology & Behavior*, *243*, 113647. https://doi.org/10.1016/j.physbeh.2021.113647

Hox, J. J., Moerbeek, M., & van Schoot, R. de. (2018). *Multilevel analysis: Techniques and applications* (Third edition). *ProQuest Ebook Central*. Routledge Taylor & Francis. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5046900

Hull, C. L. (1943). *Principles of Behavior*. Appleton-Century-Crofts.

Hunger-Schoppe, C., Schweitzer, J., Hilzinger, R., Krempel, L., Deußer, L., Sander, A., Bents, H., Mander, J., & Lieb, H. (2022). Integrative systemic and family therapy for social anxiety disorder: Manual and practice in a pilot randomized controlled trial (SOPHO-CBT/ST). *Frontiers in Psychology*, *13*, 867246. https://doi.org/10.3389/fpsyg.2022.867246

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., & Dayan, P. (2015). Decision-Theoretic Psychiatry. *Clinical Psychological Science*, *3*(3), 400–421. https://doi.org/10.1177/2167702614562040

Ilango, A., Shumake, J., Wetzel, W., & Ohl, F. W. (2014). Contribution of emotional and motivational neurocircuitry to cue-signaled active avoidance learning. *Frontiers in Behavioral Neuroscience*, *8*, 372. https://doi.org/10.3389/fnbeh.2014.00372

Ishikawa, K., Oyama, T., & Okubo, M. (2021). The malfunction of domain-specific attentional process in social anxiety: Attentional process of social and non-social stimuli. *Cognition & Emotion*, *35*(6), 1163–1174. https://doi.org/10.1080/02699931.2021.1935217

Jacobson, N. C., & Newman, M. G. (2014). Avoidance mediates the relationship between anxiety and depression over a decade later. *Journal of Anxiety Disorders*, *28*(5), 437–445. https://doi.org/10.1016/j.janxdis.2014.03.007

James, W. (2021). *The Principles of Psychology: Volumes 1 and 2. Complete works.* Amazon Fulfillment. (Original work published 1890)

Johnson, E. J., & Ratcliff, R. (2014). Computational and Process Models of Decision Making in Psychology and Behavioral Economics. In *Neuroeconomics* (pp. 35–47). Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00003-6

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.

Kalin, N. H. (2020). The Critical Relationship Between Anxiety and Depression. *The American Journal of Psychiatry*, *177*(5), 365–367. https://doi.org/10.1176/appi.ajp.2020.20030305

Karsdorp, P. A., Geenen, R., Kroese, F. M., & Vlaeyen, J. W. S. (2016). Turning Pain Into Cues for Goal-Directed Behavior: Implementation Intentions Reduce Escape-Avoidance Behavior on a Painful Task. *The Journal of Pain*, *17*(4), 499–507. https://doi.org/10.1016/j.jpain.2015.12.014

Kasper, S. (2006). Anxiety disorders: Under-diagnosed and insufficiently treated. *International Journal of Psychiatry in Clinical Practice*, *10 Suppl 1*, 3–9. https://doi.org/10.1080/13651500600552297

Kaufman, J., & Charney, D. (2000). Comorbidity of Mood and Anxiety Disorders. *Depression and Anxiety*(12), 69–76.

Keefe, J. R., McCarthy, K. S., Dinger, U., Zilcha-Mano, S., & Barber, J. P. (2014). A meta-analytic review of psychodynamic therapies for anxiety disorders. *Clinical Psychology Review*, *34*(4), 309–323. https://doi.org/10.1016/j.cpr.2014.03.004

Kemper, C. J., Ziegler, M., & Taylor, S. (2011). *ASI-3 - Angstsensitivitätsindex-3.* https://doi.org/10.23668/psycharchives.4526

Keren, G., & Schul, Y. (2009). Two Is Not Always Better Than One: A Critical Evaluation of Two-System Theories. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x

Kessler, R. C., Olfson, M., & Berglund, P. A. (1998). Patterns and predictors of treatment contact after first onset of psychiatric disorders. *American Journal of Psychiatry*(155), 62–69.

Kim, H.-G., Cheon, E.-J., Bai, D.-S., Lee, Y. H., & Koo, B.-H. (2018). Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investigation*, *15*(3), 235–245. https://doi.org/10.30773/pi.2017.08.17

Kirlic, N., Young, J., & Aupperle, R. L. (2017). Animal to human translational paradigms relevant for approach avoidance conflict decision making. *Behaviour Research and Therapy*, *96*, 14–29. https://doi.org/10.1016/j.brat.2017.04.010

Knappe, S., Klotsche, J., Heyde, F., Hiob, S., Siegert, J., Hoyer, J., Strobel, A., LeBeau, R. T., Craske, M. G., Wittchen, H.-U., & Beesdo-Baum, K. (2014). Test-retest reliability and sensitivity to change of the dimensional anxiety scales for DSM-5. *CNS Spectrums*, *19*(3), 256–267. https://doi.org/10.1017/S1092852913000710

Knowles, K. A., & Olatunji, B. O. (2020). Specificity of trait anxiety in anxiety and depression: Meta-analysis of the State-Trait Anxiety Inventory. *Clinical Psychology Review*, *82*, 101928. https://doi.org/10.1016/j.cpr.2020.101928

Kruglanski, A. W., Erbs, H.-P., Pierro, A., Mannetti, L., & Chun, W. Y. (2006). On Parametric Continuities in the World of Binary Either Ors. *Psychological Inquiry*, *17*(3), 153–165. https://doi.org/10.1207/s15327965pli1703_1

Krypotos, A.-M., Blanken, T. F., Arnaudova, I., Matzke, D., & Beckers, T. (2017). A Primer on Bayesian Analysis for Experimental Psychopathologists. *Journal of Experimental Psychopathology*, *8*(2), 140–157. https://doi.org/10.5127/jep.057316.

Krypotos, A.-M., Effting, M., Kindt, M., & Beckers, T. (2015). Avoidance learning: A review of theoretical models and recent developments. *Frontiers in Behavioral Neuroscience*, *9*, 189. https://doi.org/10.3389/fnbeh.2015.00189

Krypotos, A.-M., Vervliet, B., & Engelhard, I. M. (2018). The validity of human avoidance paradigms. *Behaviour Research and Therapy*, *111*, 99–105. https://doi.org/10.1016/j.brat.2018.10.011

Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit formation in the real world. *European Journal of Social Psychology*, *40*(6), 998–1009. https://doi.org/10.1002/ejsp.674

Larson, R. B. (2019). Controlling social desirability bias. *International Journal of Market Research*, *61*(5), 534–547. https://doi.org/10.1177/1470785318805305

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (Eds.). (1981). *Das State-Trait-Angstinventar (STAI): theoretische Grundlagen und Handanweisung*. Beltz.

LeBeau, R. T., Glenn, D. E., Hanover, L. N., Beesdo-Baum, K., Wittchen, H.-U., & Craske, M. G. (2012). A dimensional approach to measuring anxiety for DSM-5. *International Journal of Methods in Psychiatric Research*, *21*(4), 258–272. https://doi.org/10.1002/mpr.1369

LeDoux, J. E., & 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

LeDoux, J. E., Moscarello, J., Sears, R., & Campese, V. (2017). The birth, death, and resurrection of avoidance: A reconceptualization of a troubled paradigm. *Molecular Psychiatry*, *22*(1), 24–36. https://doi.org/10.1038/mp.2016.166

Lehman, B. J., David, D. M., & Gruber, J. A. (2017). Rethinking the biopsychosocial model of health: Understanding health as a dynamic system. *Social and Personality Psychology Compass*, *11*(8), Article e12328. https://doi.org/10.1111/spc3.12328

Leng, L., Beckers, T., & Vervliet, B. (2022). No joy - why bother? Higher anhedonia relates to reduced pleasure from and motivation for threat avoidance. *Behaviour Research and Therapy*, *159*, 104227. https://doi.org/10.1016/j.brat.2022.104227

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.7.2) [Computer software]. https://CRAN.R-project.org/package=emmeans

Levis, D. J., & Boyd, T. L. (1979). Symptom maintenance: An infrahuman analysis and extension of the conservation of anxiety principle. *Journal of Abnormal Psychology*, *88*(2), 107–120. https://doi.org/10.1037/0021-843X.88.2.107

Lewin, K. (1922a). Das Problem der Willensmessung und das Grundgesetz der Assoziation. I. *Psychological Research*, *1*(1), 191–302. https://doi.org/10.1007/BF00410391

Lewin, K. (1922b). Das Problem der Willensmessung und das Grundgesetz der Assoziation. II. *Psychological Research*, *2*(1), 65–140. https://doi.org/10.1007/BF02412947

Linnebank, F. E., Kindt, M., & de Wit, S. (2018). Investigating the balance between goal-directed and habitual control in experimental and real-life settings. *Learning & Behavior*, *46*(3), 306–319. https://doi.org/10.3758/s13420-018-0313-6

Loijen, A., Vrijsen, J. N., Egger, J. I. M., Becker, E. S., & Rinck, M. (2020). Biased approach-avoidance tendencies in psychopathology: A systematic review of their assessment and modification. *Clinical Psychology Review*, *77*, 101825. https://doi.org/10.1016/j.cpr.2020.101825

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shiban, Y., Schmitz, A., Straube, B., . . . 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

Lorimer, B., Kellett, S., Nye, A., & Delgadillo, J. (2021). Predictors of relapse and recurrence following cognitive behavioural therapy for anxiety-related disorders: A systematic review. *Cognitive Behaviour Therapy*, *50*(1), 1–18. https://doi.org/10.1080/16506073.2020.1812709

Lovibond, P. F., Davis, N. R., & O'Flaherty, A. S. (2000). Protection from extinction in human fear conditioning. *Behaviour Research and Therapy*, *38*(10), 967–983. https://doi.org/10.1016/s0005-7967(99)00121-7

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Lovibond, P. F., Mitchell, C. J., Minard, E., Brady, A., & Menzies, R. G. (2009). Safety behaviours preserve threat beliefs: Protection from extinction of human fear conditioning by an avoidance response. *Behaviour Research and Therapy*, *47*(8), 716–720. https://doi.org/10.1016/j.brat.2009.04.013

Lovibond, P. F., Saunders, J. C., Weidemann, G., & Mitchell, C. J. (2008). Evidence for expectancy as a mediator of avoidance and anxiety in a laboratory model of human avoidance learning. *Quarterly Journal of Experimental Psychology (2006)*, *61*(8), 1199–1216. https://doi.org/10.1080/17470210701503229

Lovibond, S. H., & Lovibond, P. F. (1996). *Manual for the depression anxiety stress scales*. Psychology Foundation of Australia.

Luck, C. C., & Lipp, O. V. (2016). Instructed extinction in human fear conditioning: History, recent developments, and future directions. *Australian Journal of Psychology*, *68*(3), 209–227. https://doi.org/10.1111/ajpy.12135

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Lundh, L.-G. (2019). The Crisis in Psychological Science and the Need for a Person-Oriented Approach. In J. Valsiner (Ed.), *Theory and History in the Human and Social Sciences. Social Philosophy of Science for the Social Sciences* (pp. 203–223). Springer International Publishing. https://doi.org/10.1007/978-3-030-33099-6_12

Luque, D., Beesley, T., Morris, R. W., Jack, B. N., Griffiths, O., Whitford, T. J., & Le Pelley, M. E. (2017). Goal-Directed and Habit-Like Modulations of Stimulus Processing during Reinforcement Learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *37*(11), 3009–3017. https://doi.org/10.1523/JNEUROSCI.3205-16.2017

Luque, D., & Molinero, S. (2020). *A critical assessment of the goal replacement hypothesis for habitual behaviour*. https://doi.org/10.31234/osf.io/rw26a

Luque, D., Molinero, S., Watson, P., López, F. J., & Le Pelley, M. E. (2019). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology. General*. Advance online publication. https://doi.org/10.1037/xge0000722

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, *38*(1), 50–67. https://doi.org/10.3758/LB.38.1.50

Maier, S. F., & Seligman, M. E. P. (2016). Learned helplessness at fifty: Insights from neuroscience. *Psychological Review*, *123*(4), 349–367. https://doi.org/10.1037/rev0000033

Malloy, P., & Levis, D. J. (1988). A laboratory demonstration of persistent human avoidance. *Behavior Therapy*, *19*(2), 229–241. https://doi.org/10.1016/S0005-7894(88)80045-5

Margraf, J., & Cwik, J. C. (2017). *Mini-DIPS Open Access: Diagnostisches Kurzinterview bei psychischen Störungen*. https://doi.org/10.13154/rub.102.91

Marmorstein, N. R. (2012). Anxiety disorders and substance use disorders: Different associations by anxiety disorder. *Journal of Anxiety Disorders*, *26*(1), 88–94. https://doi.org/10.1016/j.janxdis.2011.09.005

McCurdy, B. H., Scozzafava, M. D., Bradley, T., Matlow, R., Weems, C. F., & Carrion, V. G. (2022). Impact of anxiety and depression on academic achievement among underserved school children: Evidence of suppressor effects. *Current Psychology*, 1–9. https://doi.org/10.1007/s12144-022-03801-9

McNaughton, N. (2018). What do you mean 'anxiety'? Developing the first anxiety syndrome biomarker. *Journal of the Royal Society of New Zealand*, *48*(2-3), 177–190. https://doi.org/10.1080/03036758.2017.1358184

Meier, M., Martarelli, C., & Wolff, W. (2023). *Bored participants, biased data? How boredom can influence behavioral science research and what we can do about it.* https://doi.org/10.31234/osf.io/hzfqr

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001

Mendlowicz, M. V., & Stein, M. B. (2000). Quality of life in individuals with anxiety disorders. *The American Journal of Psychiatry*, *157*(5), 669–682. https://doi.org/10.1176/appi.ajp.157.5.669

Mesri, B., Niles, A. N., Pittig, A., LeBeau, R. T., Haik, E., & Craske, M. G. (2017). Public speaking avoidance as a treatment moderator for social anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *55*, 66–72. https://doi.org/10.1016/j.jbtep.2016.11.010

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Mineka, S., & Oehlberg, K. (2008). The relevance of recent developments in classical conditioning to understanding the etiology and maintenance of anxiety disorders. *Acta Psychologica*, *127*(3), 567–580. https://doi.org/10.1016/j.actpsy.2007.11.007

Mobbs, D., Marchant, J. L., Hassabis, D., Seymour, B., Tan, G., Gray, M., Petrovic, P., Dolan, R. J., & Frith, C. D. (2009). From threat to fear: The neural organization of defensive fear systems in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(39), 12236–12243. https://doi.org/10.1523/JNEUROSCI.2378-09.2009.

Mogg, K., & Bradley, B. P. (1998). A cognitive-motivational analysis of anxiety. *Behaviour Research and Therapy*, *36*(9), 809–848. https://doi.org/10.1016/s0005-7967(98)00063-1

Moitra, E., Herbert, J. D., & Forman, E. M. (2008). Behavioral avoidance mediates the relationship between anxiety and depressive symptoms among social anxiety disorder patients. *Journal of Anxiety Disorders*, *22*(7), 1205–1213. https://doi.org/10.1016/j.janxdis.2008.01.002

Moors, A., Boddez, Y., & de Houwer, J. (2017). The Power of Goal-Directed Processes in the Causation of Emotional and Other Actions. *Emotion Review*, *9*(4), 310–318. https://doi.org/10.1177/1754073916669595

Moors, A., & de Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*(8), 831–864. https://doi.org/10.1037/bul0000051

Mowrer, O. H. (1951). Two-factor learning theory: Summary and comment. *Psychological Review*, *58*(5), 350–354. https://doi.org/10.1037/h0058956

Mowrer, O. H., & Lamoreaux, R. R. (1946). Fear as an intervening variable in avoidance conditioning. *Journal of Comparative Psychology*, *39*, 29–50. https://doi.org/10.1037/h0060150

Muris, P. (2002). Relationships between self-efficacy and symptoms of anxiety disorders and depression in a normal adolescent sample. *Personality and Individual Differences*, *32*(2), 337–348. https://doi.org/10.1016/S0191-8869(01)00027-7

National Institute of Mental Health. (2023). *Definitions of the RDoC Domains and Constructs*. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs

Neal, D. T., Wood, W., Wu, M., & Kurlander, D. (2011). The pull of the past: When do habits persist despite conflict with motives? *Personality & Social Psychology Bulletin*, *37*(11), 1428–1437. https://doi.org/10.1177/0146167211419863

Nebe, S., Reutter, M., Baker, D. H., Bölte, J., Domes, G., Gamer, M., Gärtner, A., Gießing, C., Gurr, C., Hilger, K., Jawinski, P., Kulke, L., Lischke, A., Markett, S., Meier, M., Merz, C. J., Popov, T., Puhlmann, L. M. C., Quintana, D. S., . . . Feld, G. B. (2023). Enhancing precision in human neuroscience. *ELife*, *12*. https://doi.org/10.7554/eLife.85980

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Nilges, P., & Essau, C. (2015). Die Depressions-Angst-Stress-Skalen: Der DASS--ein Screeningverfahren nicht nur für Schmerzpatienten [Depression, anxiety and stress scales: DASS--A screening procedure not only for pain patients]. *Schmerz*, *29*(6), 649–657. https://doi.org/10.1007/s00482-015-0019-z

OECD. (2018). *Health at a Glance: Europe 2018.* https://doi.org/10.1787/health_glance_eur-2018-en

Oettingen, G., Sevincer, A. T., & Gollwitzer, P. M. (2008). Goal Pursuit in the Context of Culture. In *Handbook of Motivation and Cognition Across Cultures* (pp. 191–211). Elsevier. https://doi.org/10.1016/B978-0-12-373694-9.00009-X

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. *The Psychiatric Clinics of North America*, *33*(3), 557–577. https://doi.org/10.1016/j.psc.2010.04.002

Olatunji, B. O., Cisler, J. M., & Tolin, D. F. (2007). Quality of life in the anxiety disorders: A meta-analytic review. *Clinical Psychology Review*, *27*(5), 572–581. https://doi.org/10.1016/j.cpr.2007.01.015

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae: NEO-PI-R; Manual: Revidierte Fassung*. Hogrefe.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751–761. https://doi.org/10.1177/0956797612463080

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 20941–20946. https://doi.org/10.1073/pnas.1312011110

Patterson, T. K., Craske, M. G., & Knowlton, B. J. (2019). Enhanced Avoidance Habits in Relation to History of Early-Life Stress. *Frontiers in Psychology*, *10*, 1876. https://doi.org/10.3389/fpsyg.2019.01876

Paulus, M. P., & Yu, A. J. (2012). Emotion and decision-making: Affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, *16*(9), 476–483. https://doi.org/10.1016/j.tics.2012.07.009

Peak, J., Hart, G., & Balleine, B. W. (2019). From learning to action: The integration of dorsal striatal input and output pathways in instrumental conditioning. *The European Journal of Neuroscience*, *49*(5), 658–671. https://doi.org/10.1111/ejn.13964

Perez, O. D., & Dickinson, A. (2019). *A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior*. https://doi.org/10.1101/807800

Perez, O. D., & Dickinson, A. (2023). *Dual-system avoidance: extension of a theory*. https://doi.org/10.1101/2023.05.24.542134

Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences*, *5*(3-4), 229–269. https://doi.org/10.1007/s11097-006-9022-2

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Sciences*, *20*(6), 414–424. https://doi.org/10.1016/j.tics.2016.03.013

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology*, *4*, 92. https://doi.org/10.3389/fpsyg.2013.00092

Picciotto, G., & Fabio, R. A. (2023). Does stress induction affect cognitive performance or avoidance of cognitive effort? *Stress and Health : Journal of the International Society for the Investigation of Stress*. Advance online publication. https://doi.org/10.1002/smi.3280

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*(3), 263–283. https://doi.org/10.1177/1073191111411667

Pittig, A. (2019). Incentive-based extinction of safety behaviors: Positive outcomes competing with aversive outcomes trigger fear-opposite action to prevent protection from fear extinction. *Behaviour Research and Therapy*, 103463. https://doi.org/10.1016/j.brat.2019.103463

Pittig, A., Alpers, G. W., Niles, A. N., & Craske, M. G. (2015). Avoidant decision-making in social anxiety disorder: A laboratory task linked to in vivo anxiety and treatment outcome. *Behaviour Research and Therapy*, *73*, 96–103. https://doi.org/10.1016/j.brat.2015.08.003

Pittig, A., Arch, J. J., Lam, C. W. R., & Craske, M. G. (2013). Heart rate and heart rate variability in panic, social anxiety, obsessive-compulsive, and generalized anxiety disorders at baseline and in response to relaxation and hyperventilation. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *87*(1), 19–27. https://doi.org/10.1016/j.ijpsycho.2012.10.012

Pittig, A., Boschet, J. M., Glück, V. M., & Schneider, K. (2021). Elevated costly avoidance in anxiety disorders: Patients show little downregulation of acquired avoidance in face of competing rewards for approach. *Depression and Anxiety*, *38*(3), 361–371. https://doi.org/10.1002/da.23119

Pittig, A., Brand, M., Pawlikowski, M., & Alpers, G. W. (2014). The cost of fear: Avoidant decision making in a spider gambling task. *Journal of Anxiety Disorders*, *28*(3), 326–334. https://doi.org/10.1016/j.janxdis.2014.03.001

Pittig, A., & Dehler, J. (2019). Same fear responses, less avoidance: Rewards competing with aversive outcomes do not buffer fear acquisition, but attenuate avoidance to accelerate subsequent fear extinction. *Behaviour Research and Therapy*, *112*, 1–11. https://doi.org/10.1016/j.brat.2018.11.003

Pittig, A., Glück, V. M., Boschet, J. M., Wong, A. H. K., & Engelke, P. (2021). Increased Anxiety of Public Situations During the COVID-19 Pandemic: Evidence From a Community and a Patient Sample. *Clinical Psychology in Europe*, *3*(2), e4221. https://doi.org/10.32872/cpe.4221

Pittig, A., Hengen, K., Bublatzky, F., & Alpers, G. W. (2018). Social and monetary incentives counteract fear-driven avoidance: Evidence from approach-avoidance decisions. *Journal of Behavior Therapy and Experimental Psychiatry*, *60*, 69–77. https://doi.org/10.1016/j.jbtep.2018.04.002

Pittig, A., & Scherbaum, S. (2020). Costly avoidance in anxious individuals: Elevated threat avoidance in anxious individuals under high, but not low competing rewards. *Journal of Behavior Therapy and Experimental Psychiatry*, *66*, 101524. https://doi.org/10.1016/j.jbtep.2019.101524

Pittig, A., Schulz, A. R., Craske, M. G., & Alpers, G. W. (2014). Acquisition of behavioral avoidance: Task-irrelevant conditioned stimuli trigger costly decisions. *Journal of Abnormal Psychology*, *123*(2), 314–329. https://doi.org/10.1037/a0036136

Pittig, A., Treanor, M., LeBeau, R. T., & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research.

Neuroscience and Biobehavioral Reviews, 88, 117–140. https://doi.org/10.1016/j.neubiorev.2018.03.015

Pittig, A., & Wong, A. H. K. (2021). Incentive-based, instructed, and social observational extinction of avoidance: Fear-opposite actions and their influence on fear extinction. *Behaviour Research and Therapy, 137,* 103797. https://doi.org/10.1016/j.brat.2020.103797.

Pittig, A., & Wong, A. H. K. (2022). Reducing the return of avoidance and fear by directly targeting avoidance: Comparing incentive-based and instructed extinction of avoidance to passive fear extinction. *Journal of Experimental Psychopathology*, *13*(4), 204380872211364. https://doi.org/10.1177/20438087221136424

Pittig, A., Wong, A. H. K., Glück, V. M., & Boschet, J. M. (2020). Avoidance and its bi-directional relationship with conditioned fear: Mechanisms, moderators, and clinical implications. *Behaviour Research and Therapy, 126,* 103550. https://doi.org/10.1016/j.brat.2020.103550

Pool, E. R., Gera, R., Fransen, A., Perez, O. D., Cremer, A., Aleksic, M., Tanwisuth, S., Quail, S., Ceceli, A. O., Manfredi, D. A., Nave, G., Tricomi, E., Balleine, B. W., Schonberg, T., Schwabe, L., & O'Doherty, J. P. (2022). Determining the effects of training duration on the behavioral expression of habitual control in humans: A multilaboratory investigation. *Learning & Memory (Cold Spring Harbor, N.Y.), 29*(1), 16–28. https://doi.org/10.1101/lm.053413.121

Popp, R., Fulda, S., & Schwarz, J. F. (2011). *Karolinska-Schläfrigkeits-Skala (Karolinska Sleepiness Scale). Kompendium Schlafmedizin für Ausbildung, Klinik und Praxis*. Deutsche Gesellschaft für Schlafforschung und Schlafmedizin. https://doi.org/Hrsg

Popp, R., Fulda, S., Schwarz, J., Åkerstedt, T., & Deutsche Gesellschaft für Schlafforschung und Schlafmedizin (2011). Karolinska-Schläfrigkeits-Skala (Karolinska Sleepiness Scale). *Deutsche Gesellschaft Für Schlafforschung Und Schlafmedizin (Hrsg) Kompendium Schlafmedizin Für Ausbildung, Klinik Und Praxis (17. Erg. Lfg. 5/11). Eco-Med, Landsberg*.

Porter, E., & Chambless, D. L. (2015). A systematic review of predictors and moderators of improvement in cognitive-behavioral therapy for panic disorder and agoraphobia. *Clinical Psychology Review, 42,* 179–192. https://doi.org/10.1016/j.cpr.2015.09.004

Proctor, R. W. (2011). Playing the Simon game: Use of the Simon task for investigating human information processing. *Acta Psychologica, 136*(2), 182–188. https://doi.org/10.1016/j.actpsy.2010.06.010

Quaedflieg, C. W. E. M., Stoffregen, H., Sebalo, I., & Smeets, T. (2019). Stress-induced impairment in goal-directed instrumental behaviour is moderated by baseline working memory. *Neurobiology of Learning and Memory*, *158*, 42–49. https://doi.org/10.1016/j.nlm.2019.01.010

R Core Team. (2022). *R: A language and environment for statistical computing*. http://www.R-project.org/

Rachman, S. (1976). The passing of the two-stage theory of fear and avoidance: Fresh possibilities. *Behaviour Research and Therapy*, *14*(2), 125–131. https://doi.org/10.1016/0005-7967(76)90066-8

Radenbach, C., Reiter, A. M. F., Engert, V., Sjoerds, Z., Villringer, A., Heinze, H.-J., Deserno, L., & Schlagenhauf, F. (2015). The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology*, *53*, 268–280. https://doi.org/10.1016/j.psyneuen.2014.12.017

Raeder, F., Karbach, L., Struwe, H., Margraf, J., & Zlomuzica, A. (2019). Low Perceived Self-Efficacy Impedes Discriminative Fear Learning. *Frontiers in Psychology*, *10*, 1191. https://doi.org/10.3389/fpsyg.2019.01191

Raio, C. M., Konova, A. B., & Otto, A. R. (2020). Trait impulsivity and acute stress interact to influence choice and decision speed during multi-stage decision-making. *Scientific Reports*, *10*(1), 7754. https://doi.org/10.1038/s41598-020-64540-0

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Rebar, A. L., Gardner, B., Rhodes, R. E., & Verplanken, B. (2018). The Measurement of Habit. In B. Verplanken (Ed.), *The Psychology of Habit* (pp. 31–49). Springer International Publishing. https://doi.org/10.1007/978-3-319-97529-0_3

Reiss, S. (1991). Expectancy model of fear, anxiety, and panic. *Clinical Psychology Review*, *11*(2), 141–153. https://doi.org/10.1016/0272-7358(91)90092-9

Remmers, C., & Zander, T. (2018). Why You Don't See the Forest for the Trees When You Are Anxious: Anxiety Impairs Intuitive Decision Making. *Clinical Psychological Science*, *6*(1), 48–62. https://doi.org/10.1177/2167702617728705

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*(3), 151–182. https://doi.org/10.1037/h0024475

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 (Ed.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton- Century-Crofts.

Riganello, F., Garbarino, S., & Sannita, W. G. (2012). Heart Rate Variability, Homeostasis, and Brain Function. *Journal of Psychophysiology*, *26*(4), 178–203. https://doi.org/10.1027/0269-8803/a000080

Roberts, C., Apergis-Schoute, A. M., Bruhl, A., Nowak, M., Baldwin, D. S., Sahakian, B. J., & Robbins, T. W. (2022). Threat reversal learning and avoidance habits in generalised anxiety disorder. *Translational Psychiatry*, *12*(1), 216. https://doi.org/10.1038/s41398-022-01981-3

Robinson, O. J., Vytal, K., Cornwell, B. R., & Grillon, C. (2013). The impact of anxiety upon cognition: Perspectives from human threat of shock studies. *Frontiers in Human Neuroscience*, *7*, 203. https://doi.org/10.3389/fnhum.2013.00203

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W. M., Wiers, R. W., Borsboom, D., & Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy*, *153*, 104096. https://doi.org/10.1016/j.brat.2022.104096

Roelofs, K. (2017). Freeze for action: Neurobiological mechanisms in animal and human freezing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1718). https://doi.org/10.1098/rstb.2016.0206

Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press. https://doi.org/10.7551/mitpress/9780262014557.001.0001?locatt=mode:legacy

Rudaz, M., Craske, M. G., Becker, E. S., Ledermann, T., & Margraf, J. (2010). Health anxiety and fear of fear in panic disorder and agoraphobia vs. Social phobia: A prospective longitudinal study. *Depression and Anxiety*, *27*(4), 404–411. https://doi.org/10.1002/da.20645

Sacha, J. (2014). Interplay between heart rate and its variability: A prognostic game. *Frontiers in Physiology*, *5*, 347. https://doi.org/10.3389/fphys.2014.00347

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Scholl, C., Baladron, J., Vitay, J., & Hamker, F. H. (2022). Enhanced habit formation in Tourette patients explained by shortcut modulation in a hierarchical cortico-basal ganglia model. *Brain Structure & Function*, *227*(3), 1031–1050. https://doi.org/10.1007/s00429-021-02446-x

Scholten, W. D., Batelaan, N. M., van Balkom, A. J., Wjh Penninx, B., Smit, J. H., & van Oppen, P. (2013). Recurrence of anxiety disorders and its predictors. *Journal of Affective Disorders*, *147*(1-3), 180–185. https://doi.org/10.1016/j.jad.2012.10.031

Schreiner, D. C., Renteria, R., & Gremel, C. M. (2020). Fractionating the all-or-nothing definition of goal-directed and habitual decision-making. *Journal of Neuroscience Research*, *98*(6), 998–1006. https://doi.org/10.1002/jnr.24545

Schulz, E., Huys, Q. J. M., Bach, D. R., Speekenbrink, M., & Krause, A. (2016, February 2). *Better safe than sorry: Risky function exploitation through safe optimization*. http://arxiv.org/pdf/1602.01052v2

Schwabe, L., & Wolf, O. T. (2009). Stress prompts habit behavior in humans. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(22), 7191–7198. https://doi.org/10.1523/JNEUROSCI.0979-09.2009

Schwabe, L., & Wolf, O. T. (2010). Socially evaluated cold pressor stress after instrumental learning favors habits over goal-directed action. *Psychoneuroendocrinology*, *35*(7), 977–986. https://doi.org/10.1016/j.psyneuen.2009.12.010

Schwabe, L., & Wolf, O. T. (2013). Stress and multiple memory systems: From 'thinking' to 'doing'. *Trends in Cognitive Sciences*, *17*(2), 60–68. https://doi.org/10.1016/j.tics.2012.12.001

Seligman, M. E., & Campbell, B. A. (1965). Effect of intensity and duration of punishment on extinction of an avoidance response. *Journal of Comparative and Physiological Psychology*, *59*, 295–297. https://doi.org/10.1037/h0021845

Seligman, M. E., & Johnston, J. (1973). A cognitive theory of avoidance learning. In F. J. McGuigan & D. B. Lumsden (Eds.), *Contemporary approaches to conditioning and learning.* V. H. Winston & Sons.

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, *5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*, 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Shin, Y. K., Proctor, R. W., & Capaldi, E. J. (2010). A review of contemporary ideomotor theory. *Psychological Bulletin*, *136*(6), 943–974. https://doi.org/10.1037/a0020541

Sjouwerman, R., Niehaus, J., & Lonsdorf, T. B. (2015). Contextual Change After Fear Acquisition Affects Conditioned Responding and the Time Course of Extinction Learning-Implications for Renewal Research. *Frontiers in Behavioral Neuroscience*, *9*, 337. https://doi.org/10.3389/fnbeh.2015.00337

Smeets, T., Ashton, S. M., Roelands, S. J. A. A., & Quaedflieg, C. W. E. M. (2023). Does stress consistently favor habits over goal-directed behaviors? Data from two preregistered exact replication studies. *Neurobiology of Stress*, *23*, 100528. https://doi.org/10.1016/j.ynstr.2023.100528

Smith, A. R., Ebert, E. E., & Broman-Fulks, J. J. (2016). The relationship between anxiety and risk taking is moderated by ambiguity. *Personality and Individual Differences*, *95*, 40–44. https://doi.org/10.1016/j.paid.2016.02.018

Solomon, R. L., Kamin L. J., & Wynne, L. C. (1953). Traumatic avoidance learning: The outcomes of several extinction procedures with dogs. *Journal of Abnormal Psychology*, *48*(2), 291–302. https://doi.org/10.1037/h0058943

Somers, J. M., Goldner, E. M., Waraich, P., & Hsu, L. (2006). Prevalence and incidence studies of anxiety disorders: A systematic review of the literature. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *51*(2), 100–113. https://doi.org/10.1177/070674370605100206

Spielberg, J. M., Heller, W., Silton, R. L., Stewart, J. L., & Miller, G. A. (2011). Approach and Avoidance Profiles Distinguish Dimensions of Anxiety and Depression. *Cognitive Therapy and Research*, *35*(4), 359–371. https://doi.org/10.1007/s10608-011-9364-0

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.

Stock, A., & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research*, *68*(2-3), 176–188. https://doi.org/10.1007/s00426-003-0154-5

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Struijs, S. Y., Lamers, F., Rinck, M., Roelofs, K., Spinhoven, P., & Penninx, B. W. J. H. (2018). The predictive value of Approach and Avoidance tendencies on the onset and course of depression and anxiety disorders. *Depression and Anxiety.* Advance online publication. https://doi.org/10.1002/da.22760

Sussman, T. J., Jin, J., & Mohanty, A. (2016). Top-down and bottom-up factors in threat-related perception and attention in anxiety. *Biological Psychology*, *121*(Pt B), 160–172. https://doi.org/10.1016/j.biopsycho.2016.08.006

Talmi, D., & Pine, A. (2012). How costs influence decision values for mixed outcomes. *Frontiers in Neuroscience*, *6*, 146. https://doi.org/10.3389/fnins.2012.00146

Taylor, S. (1995). Anxiety sensitivity: Theoretical perspectives and recent findings. *Behaviour Research and Therapy*, *33*(3), 243–258. https://doi.org/10.1016/0005-7967(94)00063-p

Taylor, S., Zvolensky, M. J., Cox, B. J., Deacon, B., Heimberg, R. G., Ledley, D. R., Abramowitz, J. S., Holaway, R. M., Sandin, B., Stewart, S. H., Coles, M., Eng, W., Daly, E. S., Arrindell, W. A., Bouvard, M., & Cardenas, S. J. (2007). Robust dimensions of anxiety sensitivity: Development and initial validation of the Anxiety Sensitivity Index-3. *Psychological Assessment*, *19*(2), 176–188. https://doi.org/10.1037/1040-3590.19.2.176

Telch, M. J., Schmidt, N. B., Jaimez, T. L., Jacquin, K. M., & Harrington, P. J. (1995). Impact of cognitive-behavioral treatment on quality of life in panic disorder patients. *Journal of Consulting and Clinical Psychology*, *63*(5), 823–830. https://doi.org/10.1037//0022-006x.63.5.823

Tewes, C. (2018). The Phenomenology of Habits: Integrating First-Person and Neuropsychological Studies of Memory. *Frontiers in Psychology*, *9*, 550. https://doi.org/10.3389/fpsyg.2018.01176

Thorndike, E. L. (1913). Ideo-motor action. *Psychological Review*, *20*(2), 91–106. https://doi.org/10.1037/h0072027

Toli, A., Webb, T. L., & Hardy, G. E. (2016). Does forming implementation intentions help people with mental health problems to achieve goals? A meta-analysis of experimental studies with clinical and analogue samples. *The British Journal of Clinical Psychology*, *55*(1), 69–90. https://doi.org/10.1111/bjc.12086

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tolman, R. M., Himle, J., Bybee, D., Abelson, J. L., Hoffman, J., & van Etten-Lee, M. (2009). Impact of social anxiety disorder on employment among women receiving welfare benefits. *Psychiatric Services (Washington, D.C.)*, *60*(1), 61–66. https://doi.org/10.1176/ps.2009.60.1.61

Toyama, A., Katahira, K., & Ohira, H. (2019). Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *Journal of Mathematical Psychology*, *91*, 88–102. https://doi.org/10.1016/j.jmp.2019.03.007

Tran, T. D., Tran, T., & Fisher, J. (2013). Validation of the depression anxiety stress scales (DASS) 21 as a screening instrument for depression and anxiety in a rural community-based cohort of northern Vietnamese women. *BMC Psychiatry*, *13*, 24. https://doi.org/10.1186/1471-244X-13-24

Treanor, M., & Barry, T. J. (2017). Treatment of avoidance behavior as an adjunct to exposure therapy: Insights from modern learning theory. *Behaviour Research and Therapy*, *96*, 30–36. https://doi.org/10.1016/j.brat.2017.04.009

Trew, J. L. (2011). Exploring the roles of approach and avoidance in depression: An integrative model. *Clinical Psychology Review*, *31*(7), 1156–1168. https://doi.org/10.1016/j.cpr.2011.07.007

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The European Journal of Neuroscience*, *29*(11), 2225–2232. https://doi.org/10.1111/j.1460-9568.2009.06796.x

Uniacke, B., Timothy Walsh, B., Foerde, K., & Steinglass, J. (2018). The Role of Habits in Anorexia Nervosa: Where We Are and Where to Go From Here? *Current Psychiatry Reports*, *20*(8), 61. https://doi.org/10.1007/s11920-018-0928-5

Vaghi, M. M., Cardinal, R. N., Apergis-Schoute, A. M., Fineberg, N. A., Sule, A., & Robbins, T. W. (2019). Action-Outcome Knowledge Dissociates From Behavior in Obsessive-Compulsive Disorder Following Contingency Degradation. *Biological*

Psychiatry. Cognitive Neuroscience and Neuroimaging, *4*(2), 200–209. https://doi.org/10.1016/j.bpsc.2018.09.014

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(15), 4019–4026. https://doi.org/10.1523/JNEUROSCI.0564-07.2007

van Dis, E. A. M., van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L. H., van den Heuvel, R. M., Cuijpers, P., & Engelhard, I. M. (2020). Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *77*(3), 265–273. https://doi.org/10.1001/jamapsychiatry.2019.3986

van Uijen, S. L., Leer, A., & Engelhard, I. M. (2018). Safety Behavior After Extinction Triggers a Return of Threat Expectancy. *Behavior Therapy*, *49*(3), 450–458. https://doi.org/10.1016/j.beth.2017.08.005

Vandaele, Y., & Janak, P. H. (2018). Defining the place of habit in substance use disorders. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *87*(Pt A), 22–32. https://doi.org/10.1016/j.pnpbp.2017.06.029

Verhoeven, A., & Wit, S. de. (2018). The Role of Habits in Maladaptive Behaviour and Therapeutic Interventions. In B. Verplanken (Ed.), *The Psychology of Habit* (pp. 285–303). Springer International Publishing. https://doi.org/10.1007/978-3-319-97529-0_16

Verplanken, B., & Roy, D. (2016). Empowering interventions to promote sustainable lifestyles: Testing the habit discontinuity hypothesis in a field experiment. *Journal of Environmental Psychology*, *45*, 127–134. https://doi.org/10.1016/j.jenvp.2015.11.008

Vervliet, B., & Indekeu, E. (2015). Low-Cost Avoidance Behaviors are Resistant to Fear Extinction in Humans. *Frontiers in Behavioral Neuroscience*, *9*, 351. https://doi.org/10.3389/fnbeh.2015.00351

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R. N., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: A common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–352. https://doi.org/10.1038/mp.2014.44

Voon, V., Reiter, A., Sebold, M., & Groman, S. (2017). Model-Based Control in Dimensional Psychiatry. *Biological Psychiatry*, *82*(6), 391–400. https://doi.org/10.1016/j.biopsych.2017.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). Ez does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*(6), 1229–1235. https://doi.org/10.3758/PBR.15.6.1229

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*(1), 3–22. https://doi.org/10.3758/BF03194023

Wahl, I., Löwe, B., & Rose, M. (2011). Das Patient-Reported Outcomes Measurement Information System (PROMIS®): Übersetzung der Item-Banken für Depressivität und Angst ins Deutsche. *Klinische Diagnostik Und Evaluation, 4*, 236–261.

Wake, S., van Reekum, C. M., & Dodd, H. (2021). The effect of social anxiety on the acquisition and extinction of low-cost avoidance. *Behaviour Research and Therapy, 146*, 103967. https://doi.org/10.1016/j.brat.2021.103967

Wang, T., Li, M., Xu, S., Liu, B., Wu, T., Lu, F., Xie, J., Peng, L., & Wang, J. (2019). Relations between trait anxiety and depression: A mediated moderation model. *Journal of Affective Disorders*, *244*, 217–222. https://doi.org/10.1016/j.jad.2018.09.074

Ward, R. T., Lotfi, S., Sallmann, H., Lee, H.-J., & Larson, C. L. (2020). State anxiety reduces working memory capacity but does not impact filtering cost for neutral distracters. *Psychophysiology*, *57*(10), e13625. https://doi.org/10.1111/psyp.13625

Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L., & Turner, K. (2022). Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neuroscience and Biobehavioral Reviews*, *142*, 104869. https://doi.org/10.1016/j.neubiorev.2022.104869

Watson, P., & Wit, S. de (2018). Current limits of experimental research into habits and future directions. *Current Opinion in Behavioral Sciences*, *20*, 33–39. https://doi.org/10.1016/j.cobeha.2017.09.012

Webb, T. L., Ononaiye, M. S. P., Sheeran, P., Reidy, J. G., & Lavda, A. (2010). Using implementation intentions to overcome the effects of social anxiety on attention and appraisals of performance. *Personality & Social Psychology Bulletin*, *36*(5), 612–627. https://doi.org/10.1177/0146167210367785

Williams, J. M., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*(1), 3–24. https://doi.org/10.1037/0033-2909.120.1.3

Wilmer, M. T., Anderson, K., & Reynolds, M. (2021). Correlates of Quality of Life in Anxiety Disorders: Review of Recent Research. *Current Psychiatry Reports*, *23*(11), 77. https://doi.org/10.1007/s11920-021-01290-4

Winer, E. S., Bryant, J., Bartoszek, G., Rojas, E., Nadorff, M. R., & Kilgore, J. (2017). Mapping the relationship between anxiety, anhedonia, and depression. *Journal of Affective Disorders*, *221*, 289–296. https://doi.org/10.1016/j.jad.2017.06.006

Winer, E. S., Jordan, D. G., & Collins, A. C. (2019). Conceptualizing anhedonias and implications for depression treatments. *Psychology Research and Behavior Management*, *12*, 325–335. https://doi.org/10.2147/PRBM.S159260

Wirz, L., Bogdanov, M., & Schwabe, L. (2018). Habits under stress: Mechanistic insights across different types of learning. *Current Opinion in Behavioral Sciences*, *20*, 9–16. https://doi.org/10.1016/j.cobeha.2017.08.009

Wise, R. A., & Koob, G. F. (2014). The development and maintenance of drug addiction. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *39*(2), 254–262. https://doi.org/10.1038/npp.2013.261

Wit, S. de, Niry, D., Wariyar, R., Aitken, M. R. F., & Dickinson, A. (2007). Stimulus-outcome interactions during instrumental discrimination learning by rats and humans. *Journal of Experimental Psychology. Animal Behavior Processes*, *33*(1), 1–11. https://doi.org/10.1037/0097-7403.33.1.1

Wittchen, H.-U., Kessler, R. C., Pfister, H., Höfler, M., & Lieb, R. (2000). Why do people with anxiety disorders become depressed? A prospective-longitudinal community study. *Acta Psychiatrica Scandinavica*, *102*, 14–23. https://doi.org/10.1111/j.0065-1591.2000.acp29-03.x

Wong, A. H. K., & Lovibond, P. F. (2021). Breakfast or bakery? The role of categorical ambiguity in overgeneralization of learned fear in trait anxiety. *Emotion (Washington, D.C.)*, *21*(4), 856–870. https://doi.org/10.1037/emo0000739

Wood, W., Mazar, A., & Neal, D. T. (2022). Habits and Goals in Human Behavior: Separate but Interacting Systems. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *17*(2), 590–605. https://doi.org/10.1177/1745691621994226

Wood, W., & Rünger, D. (2016). Psychology of Habit. *Annual Review of Psychology*, *67*, 289–314. https://doi.org/10.1146/annurev-psych-122414-033417

World Medical Association (2013). Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*.

Xia, W., Dymond, S., Lloyd, K., & Vervliet, B. (2017). Partial reinforcement of avoidance and resistance to extinction in humans. *Behaviour Research and Therapy*, *96*, 79–89. https://doi.org/10.1016/j.brat.2017.04.002

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews. Neuroscience*, *7*(6), 464–476. https://doi.org/10.1038/nrn1919

Zlomuzica, A., Preusser, F., Schneider, S., & Margraf, J. (2015). Increased perceived self-efficacy facilitates the extinction of fear in healthy participants. *Frontiers in Behavioral Neuroscience*, *9*, 270. https://doi.org/10.3389/fnbeh.2015.00270

Zwosta, K., Ruge, H., Goschke, T., & Wolfensteller, U. (2018). Habit strength is predicted by activity dynamics in goal-directed brain systems during training. *NeuroImage*, *165*, 125–137. https://doi.org/10.1016/j.neuroimage.2017.09.062

Appendix

Supplementary Material for Study 1

Additional experimental phases

Exploratorily, we added a *revaluation* phase and a *reinstatement* phase after the competition phase in both experiments. Subsequent to the competition phase, the US electrode was re-attached to the participants forearms, which was emphasized both verbally and by on-screen instructions. Therefore, participants were aware about the possibility of the occurrence of aversive stimuli. We did not instruct participants other than emphasizing that the electrodes were now re-attached and that they should continue with the task. The aim of the revaluation phase was to evaluate whether the revaluation of the aversive US outcome would impact overtraining-compatible vs. overtraining-incompatible responding. The revaluation phase consisted of one block of randomized trials (i.e., 54 trials) identical to the trials in the competition phase (i.e., with reward outcomes but without aversive stimulation outcomes). After this block of trials, the reinstatement phase started seamlessly. At the beginning of the reinstatement phase, one single unannounced US was delivered during an ITI (see Lonsdorf et al., 2017). This reinstatement test was conducted to examine whether the unwarned occurrence of an aversive stimulus would impact overtraining-compatible vs. overtraining-incompatible responding, potentially due to return of fear. The reinstatement phase consisted of one block of randomized trials (i.e., 54 trials) which were again identical to the trials in the competition phase. Importantly, data from these explorative phases do not address habitual responding, as the aversive US was a valuable outcome in both phases and outcome insensitivity of responding therefore cannot be assessed. The aim of the two phases was to generate further hypotheses about the degree of adjustment of overtrained avoidance responses when aversive outcomes are revalued or reinforced.

Statistical analyses

Costly avoidance was defined as higher accuracy in overtraining-compatible as compared with overtraining-incompatible trials. Low cost avoidance was operationalized as a) higher reaction time in overtraining-compatible vs. overtraining-incompatible trials, and b) as the proportion of overtrained-compatible responses in free trials with the colors that had been overtrained (i.e., low cost avoidance when overtrained-compatible responses > 50%). Changes in the dependent variables costly (i.e., accuracy differences) and low-cost avoidance (i.e.,

reaction time differences) between the last block of the competition phase, the revaluation phase and the reinstatement phase were tested with repeated-measures ANOVAs with factor Trial Type (overtraining-compatible and overtraining-incompatible) and factor Phase (last block of the competition phase vs. revaluation phase vs. reinstatement phase). Changes in low-cost avoidance as indicated by overtraining-compatible responding in free trials were analyzed with a univariate repeated-measures ANOVA with the factor Phase (last block of the competition phase vs. revaluation phase vs. reinstatement phase). Statistical analyses were identical for both experiments.

Results Experiment 1

Accuracy. Accuracy differed significantly between overtraining-compatible and overtraining-incompatible trials, $F(1,54) = 8.574$, $p = .05$, $\eta^2 = .065$. A post-hoc comparison showed that accuracy in overtraining-compatible trials was higher than in overtraining-incompatible trials, $t(54) = 2.928$, $p_{holm} = .005$, $d = 0.395$. There was a significant effect of Phase, $F(2,108) = 3.495$, $p = .034$, $\eta^2 = .015$. A post-hoc comparison showed a tendency towards lower accuracy in the revaluation phase as compared with the other two phases, $ps_{holm} = .062$, $ds \geq 0.300$. The interaction between Trial Type and Phase was not significant, $F(2,108) = 1.841$, $p = .164$, $\eta^2 = .009$, indicating that the strength of costly avoidance did not differ between the last block of the competition phase, the revaluation phase and the Reinstatement phase.

The proportion of overtraining-compatible responding in free trials did not differ between the three phases, $F(2,108) = 2.699$, $p = .072$, $\eta^2 = .048$.

Reaction time. Reaction time did not differ between overtraining-compatible and overtraining-incompatible trials, $F(1,54) = 1.018$, $p = .318$, $\eta^2 = .009$. There was a significant effect of Phase, $F(2,108) = 14.142$ $p < .001$, $\eta^2 = .075$. Post-hoc tests indicated that overall reaction time in the revaluation phase was lower than in the last block of the competition phase, $t(54) = 5.081$, $p_{holm} < .001$, $d = 0.685$, and lower in the reinstatement phase than in the last three blocks of the reinstatement phase, $t(54) = 3.902$, $p_{holm} < .001$, $d = 0.526$. There was no significant interaction between Trial Type and Phase, $F(2,108) = 1.570$, $p = .213$, $\eta^2 = .004$, indicating that non-costly avoidance as indicated by reaction times did not differ between the competition phase, the revaluation phase and the reinstatement phase.

Results Experiment 2

Accuracy. Accuracy did not differ between habit-compatible and habit-incompatible trials, $F(1,72) = 0.519$, $p = .474$, $\eta^2 = .002$, indicating no costly avoidance. There was no effect of

Phase, $F(2,144) = 1.902$, $p = .153$, $\eta^2 = .011$, and no interaction between Trial Type and Phase, $F(2,144) = 0.239$, $p = .787$, $\eta^2 = .001$, indicating that costly avoidance did not differ between the reinstatement phase and the revaluation phase.

The proportion of overtraining-compatible responding in free trials did not differ between the three phases, $F(2,144) = 0.398$, $p = .672$, $\eta^2 = .005$.

Reaction time. Responding in overtraining-compatible trials was faster than in overtraining-incompatible trials as indicated by a significant effect of trial type, $F(1,72) = 4.726$, $p = .033$, $\eta^2 = .013$ and a follow-up t test, $t(72) = 2.174$, $p_{holm} = .033$, $d = 0.254$. There was a significant effect of Phase, $F(2,144) = 8.525$, $p < .001$, $\eta^2 = .059$. Follow-up t tests indicated lower overall reaction time in the revaluation phase as compared with both the last block of the competition phase, $t(72) = 4.094$, $p_{holm} < .001$, $d = 0.479$, and with the reinstatement phase, $t(72) = 2.514$, $p_{holm} = .026$, $d = 0.294$. There was, however, no significant interaction between Trial Type and Phase, $F(2,144) = 0.689$, $p = .504$, $\eta^2 = .002$, indicating that the strength of low-cost avoidance as indicated by a higher reaction time in overtraining-incompatible trials did not differ between phases.

EZ Drift Diffusion Model

We applied the EZ model exploratorily to investigate whether the EZ DDM parameters may add new facets to the interpretation of the data in the light of the speed-accuracy trade-off affecting the differentiation between costly (i.e., accuracy difference) and low-cost habitual avoidance (i.e., reaction time difference). Drift diffusion models for choice data combine response accuracy and response speed to estimate psychologically meaningful parameters. We used this simplified drift-diffusion model, which, to our knowledge, is currently the simplest available drift-diffusion model, to calculate the parameters drift rate v, boundary separation a, and nondecision time T_{er} separately for habit-compatible, habit-incompatible and neutral control trials (Wagenmakers et al., 2007). However, the task was not designed to yield data for drift-diffusion model analyses. Therefore, the number of 60 observations per condition in our experiments may not have produced sufficient data points to reliably estimate the EZ model parameters. In two papers, Wagenmakers and colleagues (Wagenmakers et al., 2008; Wagenmakers et al., 2007) reported that $N \geq 100$ observations per condition was sufficient for the EZ parameter estimation.

Results Experiment 1

Mean drift rate *v*, an indicator of the ease of responding, was highest in habit-compatible trials (v = 0.284), slightly lower in neutral control trials (v = 0.249) and lowest in habit-incompatible trials (v = 0.203). In line with the frequentist accuracy data analysis, this indicates more ease in processing habit-compatible trials than neutral and habit-incompatible trials. Boundary separation *a* was highest in habit-compatible trials (a = 0.063), followed by habit-incompatible trials (a = 0.061) and neutral control trials (a = 0.057). Thus, boundary separation differed slightly between conditions, potentially indicating that participants responded most conservatively in habit-compatible trials, intermediately conservative in habit-incompatible trials, and least conservatively in neutral control trials. Nondecision time T_{er} was similar across trial types, with T_{er} = 502 ms in habit-compatible trials and T_{er} = 506 ms in both habit-incompatible trials and neutral control trials.

Results Experiment 2

Mean drift rate *v* was highest in habit-compatible trials (v = 0.490), slightly lower in neutral control trials (v = 0.0481) and lowest in habit-incompatible trials (v = 0.469). In line with the frequentist accuracy data analysis, this indicates that habit-compatible trials were processed more easily than neutral control trials and habit-incompatible trials. Boundary separation *a* was highest in habit-incompatible trials (a = 0.071), followed by habit-compatible trials (a = 0.070) and neutral control trials (a = 0.065), indicating more conservative responding in habit-incompatible and habit-compatible trials compared with neutral control trials. Nondecision time T_{er} was similar across trial types, with T_{er} = 446.25 ms in habit-compatible trials, T_{er} = 447.54 in habit-incompatible trials and T_{er} = 451.87 ms in neutral control trials.

Supplementary Material for Study 2

Supplement A: Results of the complete sample analysis (*N* = 124)

All models in Supplement A include quadratic and linear random slopes for each participant.

Table A.1

Model predicting accuracy in the training phase

	df	X^2	*p*
Intercept	1	1478.96	<.001
Block	2	640.62	<.001
Group	1	2.15	.142
Group x Block	2	7.59	.023

Table A.2

Model predicting reaction times in the training phase

	df	X^2	*p*
Intercept	1	5045.94	<.001
Block	2	99.23	<.001
Group	1	1.07	.300
Block x Group	2	2.31	.315

Table A.3

Model predicting accuracy in compatible and incompatible trials

	df	X^2	*p*
Intercept	1	271.85	<.001
Condition	1	78.46	<.001
Group	1	16.14	<.001
Block	2	432.26	<.001
Condition x Group	1	1.94	.163
Condition x Block	2	6.75	.034
Group x Block	2	10.79	.005
Condition x Group x Block	2	17.32	<.001

Table A.4

Model predicting accuracy within compatible trials (post hoc analysis)

	df	X^2	*p*
Intercept	1	224.80	<.001
Group	1	9.54	.002
Block	2	194.06	<.001
Group x Block	2	13.37	.001

Table A.5

Model predicting accuracy within incompatible trials (post hoc analysis)

	df	X^2	*p*
Intercept	1	128.40	<.001
Group	1	8.61	.003
Block	2	277.56	<.001
Group x Block	2	0.58	.749

Table A.6

Model predicting accuracy in compatible and incompatible trials within the anxiety group (post hoc analysis)

	df	X^2	*p*
Intercept	1	107.17	<.001
Condition	1	33.09	<.001
Block	2	154.30	<.001
Condition x Block	2	19.50	<.001

Table A.7

Model predicting accuracy in compatible and incompatible trials within the control group (post hoc analysis)

	df	X^2	*p*
Intercept	1	126.40	<.001
Condition	1	44.99	<.001

Table A.7 *(continued)*

	df	X^2	*p*
Block	2	181.10	<.001
Condition x Block	2	3.99	.136

Table A.8

Model predicting reaction times in compatible and incompatible trials

	df	X^2	*p*
Intercept	1	5519.43	<.001
Condition	1	0.60	.440
Group	1	0.06	.815
Block	2	3.28	.194
Condition x Group	1	3.33	.068
Condition x Block	2	0.62	.735
Group x Block	2	2.65	.266
Condition x Group x Block	2	0.14	.932

Table A.9

Model predicting responses in free trials

	df	X^2	*p*
Intercept	1	5.14	.023
Block	2	0.51	.774
Group	1	0.17	.681
Group x Block	2	3.59	.166

Table A.10

Model predicting accuracy in neutral trials

	df	X^2	*p*
Intercept	1	200.43	<.001
Block	2	207.63	<.001
Group	1	9.48	.002

Table A.10 (continued)

	df	X^2	*p*
Group x Block	2	7.90	.019

Supplement B: Subsample analysis with patients without psychopharmacological medication and matched controls

Summary

As a considerable proportion of the anxiety patient group used psychopharmacological medication at the time of their participation in the experiment, we performed a subsample analysis to assess whether medication systematically impacted the comparison between patients and healthy controls. We first estimated the models analogously to the main analysis (see section 2.8 in the main text), with the data of the subgroup of unmedicated patients (n = 40) and their matched healthy controls (n = 40). As several models did not converge due to the relatively small sample size, we simplified the random effects structure by removing the random intercept from all models and computed linear models (i.e., instead of LMMs) and logistic regression models (i.e., instead of GLMMs) with the R package *stats*. As several of the models featured non-normally distributed residuals, we additionally estimated robust models with the R package *robustbase*. As the robust models did not produce substantially different results (i.e., significance of predictors), only the results of the linear and logistic regression models are reported here. All analyses are openly available (https://osf.io/nr28s/). In sum, this analysis did not reveal any substantiative differences as compared with the main analysis, indicating no substantial bias by medication.

Table B.1

Predictor contributions in the model to predict accuracy in the training phase

	df	X^2	*p*
Block	2	626.00	<.001
Group	1	10.27	.001
Group x Block	2	2.13	.141

Table B.2

Predictor contributions in the model to predict reaction times in the training phase

	df	*F*	*p*
Block	2	168.47	<.001
Group	1	140.59	<.001
Block x Group	2	1.62	.198

Table B.3

Estimated means to predict accuracy in compatible and incompatible trials

	df	X^2	*p*
Condition	1	29.44	<.001
Group	1	54.31	<.001
Block	2	197.60	<.001
Condition x Group	1	0.90	.343
Condition x Block	2	1.68	.431
Group x Block	2	21.62	<.001
Condition x Group x Block	2	17.48	<.001

Table B.4

Predictor contributions in the model to predict reaction times in compatible and incompatible trials

	df	*F*	*p*
Condition	1	<.01	.977
Group	1	0.51	.476
Block	2	0.50	.606
Condition x Group	1	1.41	.235
Condition x Block	2	0.02	.983
Group x Block	2	11.68	<.001
Condition x Group x Block	2	0.54	.582

Table B.5

Predictor contributions in the model for responses in free trials

	df	X^2	*p*
Block	2	2.04	.362
Group	1	0.23	.629
Group x Block	2	7.52	.023

Table B.6

Predictor contributions in the model for accuracy in neutral trials

	df	X^2	*p*
Block	2	177.10	<.001
Group	1	60.01	<.001
Group x Block	2	9.49	.009

Supplement C: Effect of the dimensional anxiety score

To explore the effects of the current strength of anxiety symptoms, we conducted a dimensional analysis using the whole range of the anxiety symptom strength score. The anxiety symptom strength score was obtained by averaging the individual *z* standardized PROMIS-A-SF scores, the DSM-Cross D scores, and the DASS anxiety scores. The resulting anxiety symptom strength score correlated highly with each of the single anxiety measures (see Table C.1) and discriminated clearly between the groups with only a small overlap (see Figure C.1).

The statistical modeling was performed in parallel to the group comparison analysis. Thus, all descriptions about the group comparison modeling apply to the dimensional analysis, with the exception that in the dimensional analysis, the group variable was replaced with the individual anxiety symptom strength score in all models. As several models did not converge, the random slopes for block were removed and the random effects structure was simplified to a random intercept for participant ID for all dimensional models.

Table C.1

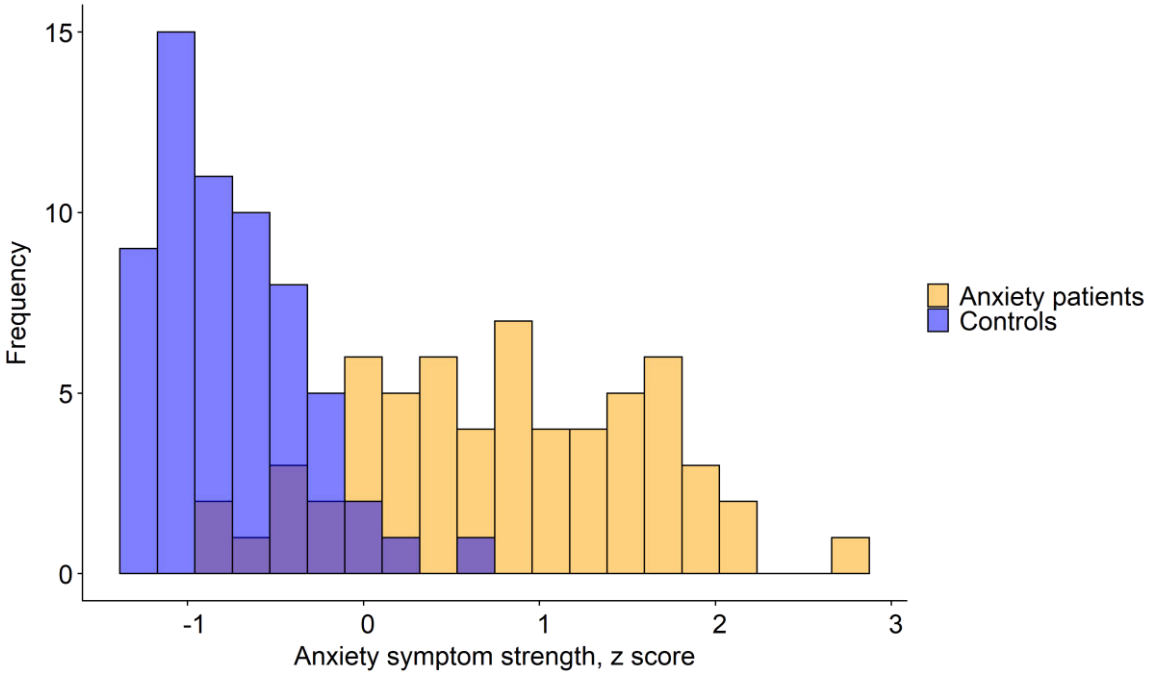
Correlations between the anxiety measures

Variable	Anxiety symptom score	DASS anxiety subscale	DSM-Cross D	PROMIS
Anxiety symptom score	—			
DASS anxiety subscale	.94***	—		
DSM-Cross D	.96***	.85***	—	
PROMIS	.97***	.86***	.92***	—

Note. Pearson correlations with Bonferroni-Holm correction. *$p < .05$, **$p < .01$, ***$p < .001$.

Figure C.1

Distribution of the anxiety symptom strength variable by group



Results

Training phase

Accuracy. The GLMM with the predictors *anxiety symptom strength*, *block*, and their interaction (see Table C.2) showed a significant effect of the predictor *block*, χ^2 (2) = 1000.15, $p < .001$, and non-significant contributions of a*nxiety symptom strength*, χ^2 (1) = 0.84, $p = .360$, and the interaction between *anxiety symptoms strength* and *block*, χ^2 (2) = 4.68, $p = .097$. In sum, accuracy increased quickly after the beginning of the task and constantly remained high afterward, independently of anxiety symptom strength.

Table C.2

Model predicting accuracy in the training phase

	df	X^2	*p*
Intercept	1	1245.57	<.001
Block	2	1000.15	<.001
Anxiety symptom strength	1	0.84	.360
Anxiety symptom strength x Block	2	4.68	.097

Reaction time. The LMM with the predictors *anxiety symptom strength, block,* and their interaction (see Table C.3) showed that the interaction between the *anxiety symptom score* and *block* was a significant predictor, χ^2 (24) = 17.37, $p < .001$, indicating that the anxiety symptom strength had a stronger decelerating impact in earlier blocks. This result diverges from the group comparison analysis, where the interaction between *group* and *block* did not significantly predict reaction times.

Table C.3

Model predicting reaction times in the training phase

	df	X^2	*p*
Intercept	1	5035.35	<.001
Block	2	811.26	<.001
Anxiety symptom strength	1	0.13	.719
Anxiety symptom strength x Block	2	17.37	<.001

Test phase

Accuracy. The GLMM yielded in a non-significant prediction of the three-way interaction between *anxiety score*, *condition*, and *block*, and non-significant interactions between *anxiety score* and *condition*, between *anxiety symptom score* and *block*, and between *block* and *condition* (see Table C.4). The interaction between the a*nxiety symptom score* and *condition* was a significant predictor, $\chi^2(1) = 4.15$, $p = 0.042$, indicating that the impact of the anxiety symptom score was associated with a larger accuracy compatibility effect (see Figure C.3). *Block* was also a significant predictor, $\chi^2(9) = 882.92$, $p < 0.001$, indicating higher accuracy in

later blocks (linear trend estimate: 67.67, SE = 2.63, p < .001) and a stronger increase in acuracy in earlier blocks than in later blocks (quadratic trend estimate: -29.66, SE = 2.59, $p < .001$).

Table C.4

Model predicting accuracy in compatible and incompatible trials

	df	X^2	*p*
Intercept	1	256.82	<.001
Condition	1	69.82	<.001
Anxiety score	1	3.21	.069
Block	2	882.92	<.001
Condition x Anxiety score	1	4.15	.042
Condition x Block	2	9.64	.008
Anxiety score x Block	2	2.23	.328
Condition x Anxiety score x Block	2	6.61	.037

Reaction times. The LMM with the continuous *anxiety score* as predictor resulted in a significant effect of *block*, $\chi^2(1) = 7.43$, $p = .024$, while the other predictors did not significantly explain response times (see Table C.5).

Table C.5

Model predicting reaction times in compatible and incompatible trials

	df	X^2	*p*
Intercept	1	6036.53	<.001
Condition	1	0.83	.363
Anxiety score	1	0.31	.579
Block	2	7.43	.024
Condition x Anxiety score	1	0.27	.601
Condition x Block	2	0.02	.974
Anxiety score x Block	2	5.95	.051
Condition x Anxiety score x Block	2	0.89	.640

Free trials

The GLM with the *anxiety symptom score* and *block* as fixed factors indicated that none of these variables or their interaction significantly predicted training-compatible responding (see Table C.6). Thus, the exploratory hypothesis that the *anxiety symptom score* would be associated with a longer lasting preference for training-compatible responding was not confirmed.

Table C.6

Model predicting responses in free trials

	df	X^2	*p*
Intercept	1	5.82	.016
Block	2	5.63	.060
Anxiety score	1	0.79	.373
Anxiety score x Block	2	8.28	.016

Learning control trials

In the GLM with the *anxiety symptom score* and *block* as fixed factors (see Table C.7 and Figure C.4), the *anxiety score* did not predict accuracy in learning control trials, $\chi^2(1) = 0.53$, $p = .466$, while *block* was a significant predictor, $\chi^2(9) = 392.83$, $p < .001$, with an increase in accuracy over the blocks (linear trend: 1.99, $SE = 0.08$, $p < .001$), which slowed down during the test phase (quadratic trend: -0.78, $SE = 0.12$, $p < .001$). The interaction between *block* and *anxiety score* was, again, non-significant, $\chi^2(9) = 7.64$, $p = 571$.

Table C.7

Model predicting responses in neutral trials

	df	X^2	*p*
Intercept	1	213.90	<.001
Block	2	387.36	<.001
Anxiety score	1	0.53	.468
Anxiety score x Block	2	4.51	.105

Supplement D: Analysis of the exploratory reinstatement and revaluation phases

An exploratory revaluation phase and an exploratory reinstatement phase after the test phase were presented to explore the adjustment of extensively trained avoidance when aversive outcomes were again revalued. After the test phase as described in the main paper, the experiment was paused, and the electrode for the deliverance of the aversive electrotactile stimulations was re-attached to the participant's arm (i.e., revaluation of the aversive outcome). This revaluation procedure was emphasized both verbally and by on-screen instructions. The subsequent *revaluation phase* consisted of three blocks of randomized trials (i.e., 54 trials) which were identical to the trials in the test phase (i.e., no electrotactile stimulations were delivered). After these revaluation phase trials, one single unannounced US was delivered during an ITI (i.e., reinstatement, see Lonsdorf et al., 2017) to explore whether the unwarned occurrence of the aversive stimulus would lead to a further change in goal-directed control and therefore to a rise in compatibility effects, potentially due to a return of fear. The three blocks of randomized trials (i.e., 54 trials, *reinstatement phase*) after this unannounced electrotactile stimulation were again identical to the trials in the test phase. The data analysis was conducted in analogy to the data analysis for the complete data (see section 3.9 in the main text body), but, as several models did not converge, the random slopes for block were removed and the random effects structure was simplified to a random intercept for participant ID for all models analyzing the revaluation and reinstatement phase.

Results

Revaluation phase

Accuracy compatibility effect. In the GLM with accuracy in the revaluation phase as criterion and with *group*, *condition* (i.e., compatible and incompatible), *block*, and all their interactions as fixed factors (see Table D.1), a significant interaction between *group* and *condition* emerged, indicating a larger accuracy compatibility effect in the control group which was independent from the blocks (see Figure D.1).

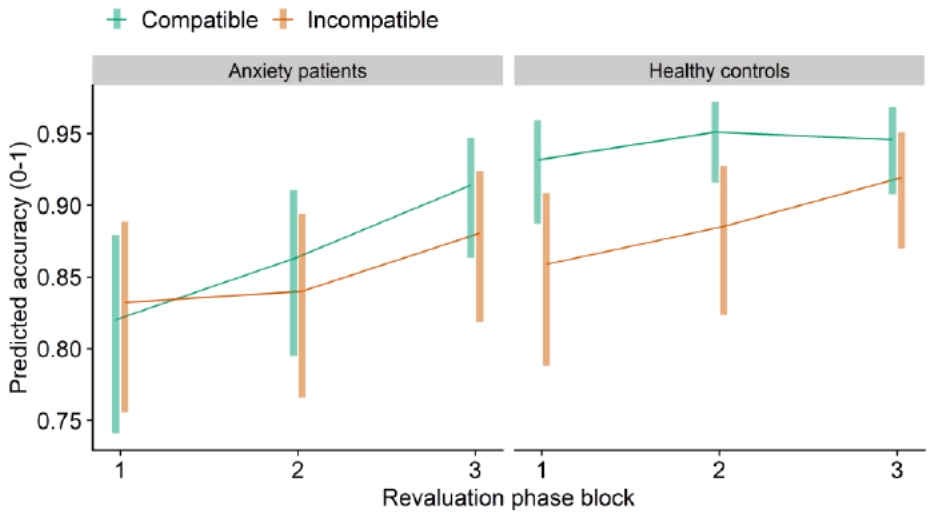
Table D.1

Estimated means to predict accuracy in compatible and incompatible trials in the revaluation phase

	df	X^2	*p*
Intercept	1	227.88	<.001
Condition	1	15.63	<.001
Group	1	5.17	.023
Block	2	15.62	<.001
Condition x Group	1	6.55	.011
Condition x Block	2	0.61	.789
Group x Block	2	1.18	.556
Condition x Group x Block	2	2.63	.269

Figure D.1

Estimated means for accuracy during the revaluation phase



Note: Error bars depict 95 % confidence intervals.

Reaction time compatibility effect. In the LMM with reaction times in the revaluation phase as criterion and with *group*, *condition* (i.e., compatible and incompatible), *block*, and all their interactions as fixed factors (see Table D.2), a tendency towards a significant interaction between condition and block emerged, indicating a slight tendency towards a stronger

compatibility effect (i.e., slower reaction times in incompatible vs. compatible trials) in the beginning of the revaluation phase.

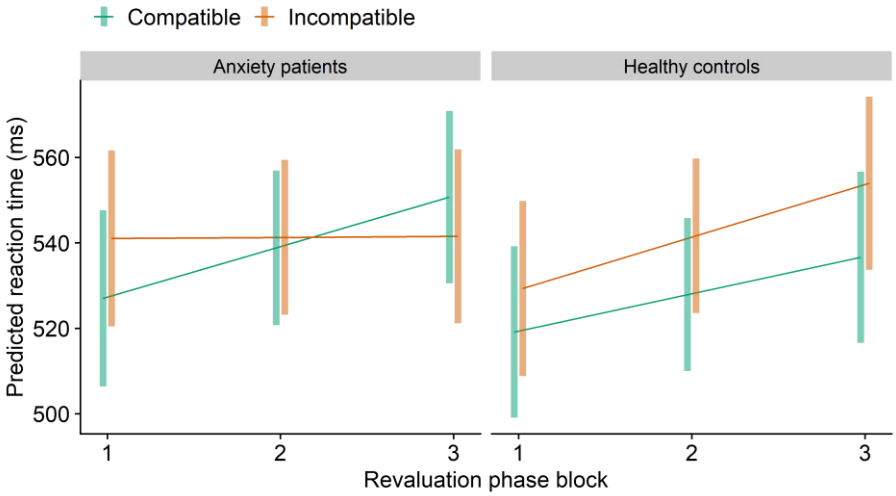
Table D.2

Estimated means to predict reaction times in compatible and incompatible trials in the revaluation phase

	df	X^2	*p*
Intercept	1	7619.24	<.001
Condition	1	4.15	.042
Group	1	0.18	.668
Block	2	12.80	.002
Condition x Group	1	2.09	.148
Condition x Block	2	5.85	.054
Group x Block	2	1.09	.581
Condition x Group x Block	2	2.62	.270

Figure D.2

Estimated means for reaction times in the revaluation phase



Note: Error bars depict 95 % confidence intervals.

Compatibility effect in free trials. In the GLM to model the compatibility of responses in free trials with block, group and all their interactions as predictors, all included predictors were insignificant (see Table D.3).

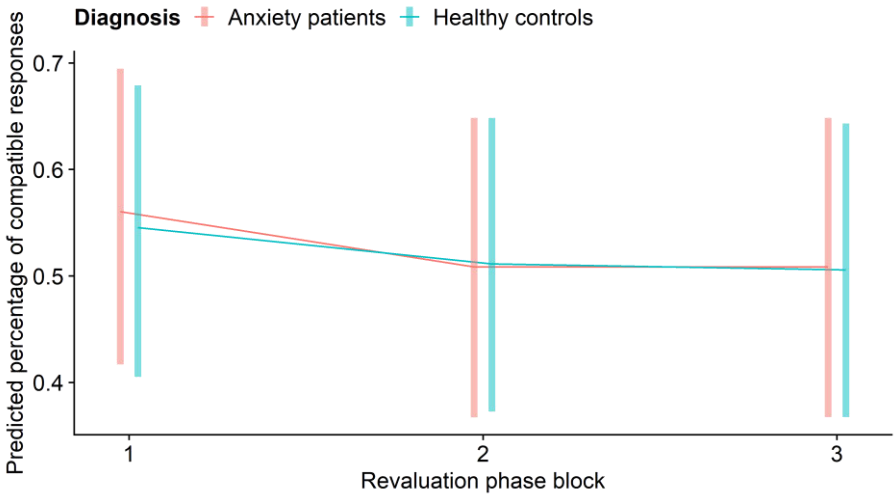
Table D.3

Estimated means to predict compatible responses in free trials in the revaluation phase

	df	X^2	*p*
Intercept	1	0.25	.617
Group	1	<0.01	.957
Block	2	1.82	.402
Group x Block	2	0.06	.971

Figure D.3

Estimated means for the compatibility effect in free trials in the revaluation phase



Note: Error bars depict 95 % confidence intervals.

Results: Reinstatement phase

Accuracy compatibility effect. In the GLM with accuracy in the reinstatement phase as criterion and with *group*, *condition* (i.e., compatible and incompatible), *block*, and all their interactions as fixed factors, *group* and *condition* (see Table D.4) were significant single predictors, indicating lower accuracy in anxiety patients (OR = 0.70 [0.51, 0.96]) and higher accuracy in compatible trials (OR = 1.21 [1.07, 1.37]).

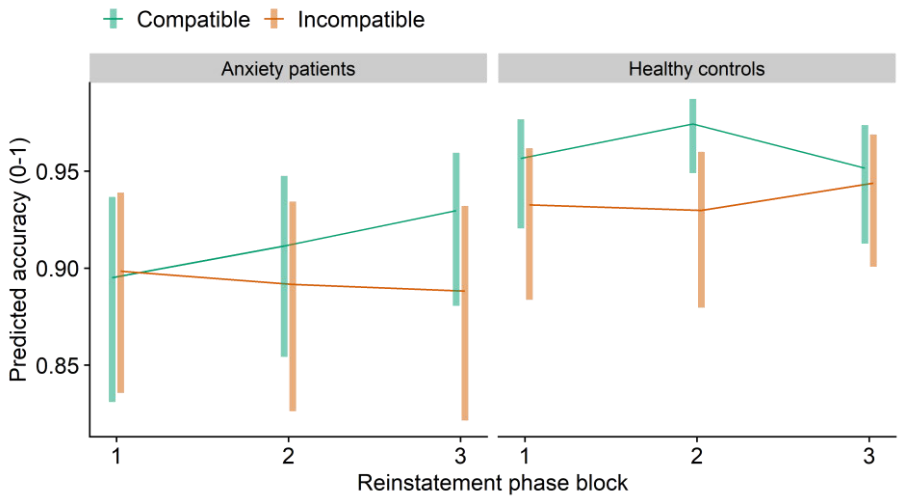
Table D.4

Estimated means to predict accuracy in compatible and incompatible trials in the reinstatement phase

	df	X^2	*p*
Intercept	1	229.93	<.001
Condition	1	9.94	.002
Group	1	5.11	.024
Block	2	1.08	.584
Condition x Group	1	1.70	.192
Condition x Block	2	2.08	.353
Group x Block	2	1.08	.583
Condition x Group x Block	2	4.04	.132

Figure D.4

Estimated means for accuracy during the reinstatement phase



Note: Error bars depict 95 % confidence intervals.

Reaction time compatibility effect. In the LMM with reaction times in the revaluation phase as criterion and with *group*, *condition* (i.e., compatible and incompatible), *block*, and all their interactions as fixed factors, none of the included variables was significantly predictive of reaction time (see Table D.5).

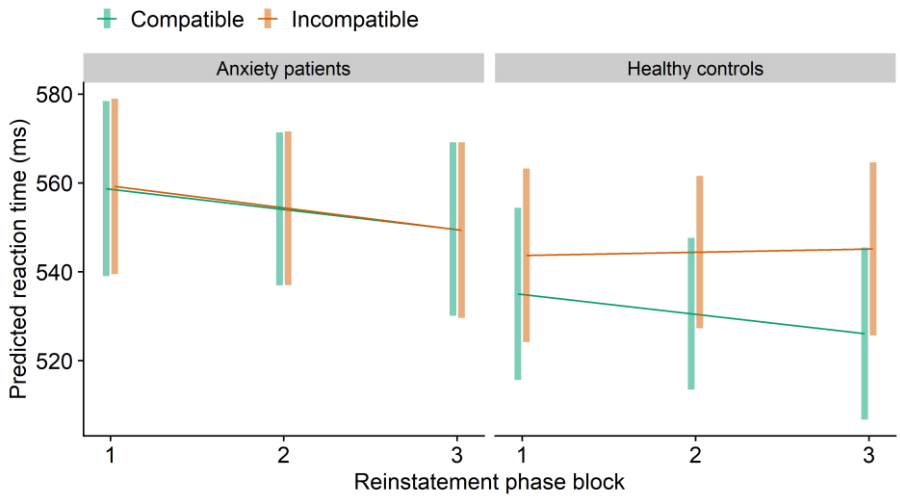
Table D.5

Estimated means to predict reaction times in compatible and incompatible trials in the reinstatement phase

	df	X^2	*p*
Intercept	1	8766.72	<.001
Condition	1	3.21	.073
Group	1	2.08	.150
Block	2	1.99	.371
Condition x Group	1	3.09	.079
Condition x Block	2	1.28	.527
Group x Block	2	2.10	.350
Condition x Group x Block	2	0.46	.793

Figure D.5

Estimated means for reaction times in the reinstatement phase



Note: Error bars depict 95 % confidence intervals.

Compatibility effect in free trials. In the GLM to model the compatibility of responses in free trials with block, group and all their interactions as predictors, none of the included variables was significantly predictive (see Table D.6)

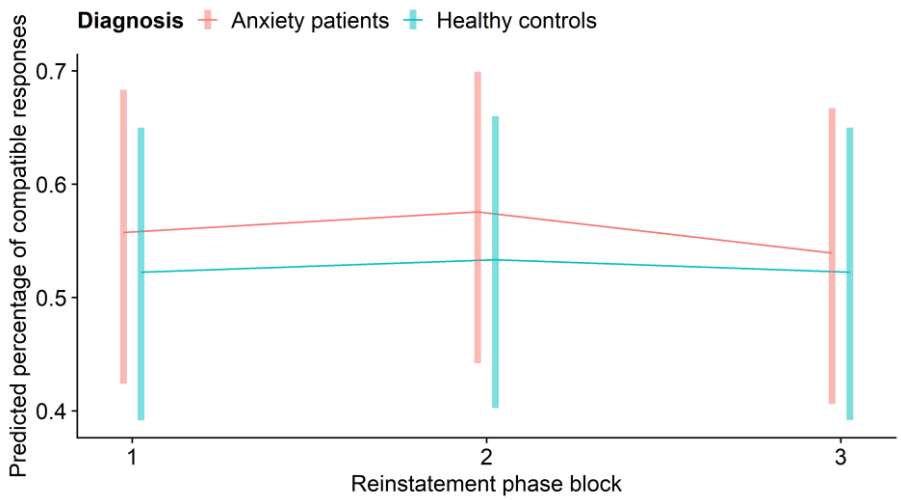
Table D.6

Estimated means to predict compatible responses in free trials in the reinstatement phase

	df	X^2	*p*
Intercept	1	0.95	.330
Group	1	0.14	.712
Block	2	0.42	.812
Group x Block	2	0.13	.939

Figure D.6

Estimated means for compatible responses in free trials in the reinstatement phase



Note: Error bars depict 95 % confidence intervals.

Supplement E: Comparison between patients with social anxiety disorder and patients with panic disorder with or without agoraphobia

Summary

In this exploratory analysis, patients with social anxiety disorder diagnosis (SAD, n = 21) were compared with patients with panic disorder diagnosis with or without agoraphobia (PD, n = 21). Patients diagnosed with only agoraphobia or with both SAD and PD were not included in this analysis. We first estimated the models as in the main analysis (see section 3.9 in the main text). As several models did not converge due to the relatively small sample size, we further simplified the random effects structure by removing the random slopes and the random intercept from all models and instead computed linear models (i.e., instead of LMMs) and logistic regression models (i.e., instead of GLMMs) with the R package *stats*. The impact of block was modeled by including a linear and a quadratic trend for *block* in all regression models. As several of the models featured non-normally distributed residuals, we additionally estimated robust models with the R package *robustbase*. These robust models did not produce divergent results. Thus, we report only the results of the linear and logistic regression models.

The analysis revealed slower responses in the PD subgroup both in the training and the test phase (see Supplementary Tables E.3, E.5). Additionally, the PD subgroup displayed a larger accuracy compatibility effect (see Table E.4) and larger compatibility effect in free trials (see Table E.6) than the SAD subgroup. Neither the accuracy in the training phase nor the accuracy in neutral trials differed between the PD and the SAD subgroup. Also, the subgroups did not differ in any of the psychological symptom measures or in their motivation to avoid and to approach; however, the average age in the PD subgroup (M = 31.33 years, SD = 11.72) was significantly higher than in the SAD subgroup (M = 23.38 years, SD = 5.32).

Table E.1

Comparison of sample characteristics in the diagnostic subgroups

	Social anxiety disorder (*N* = 21)	Panic disorder (*N* = 21)	Test statistic	*P*[3]	ES[4]
Age	23.38 (5.32)	31.33 (11.72)	88[1]	**.013**	.51[4]
Gender (% women)	13 (62%)	14 (67%)	0.1[2]	>.999	<0.01[5]
Subjective aversiveness (post) (VAS, 0-100)	55.24 (24.52)	54.76 (24.21)	221[1]	>.999	0.04[4]
Anxiety symptoms (DASS, 0-21)	8.90 (4.73)	10.44 (4.82)	178[1]	>.999	0.17[4]
Anxiety symptoms (PROMIS, 0-72)	18.05 (6.14)	18.38 (6.55)	217[1]	>.999	0.02[4]
Anxiety symptoms (DSM Cross-D, 0-48)	19.43 (10.05)	20.38 (8.76)	200[1]	>.999	0.08[4]
Depression symptoms (DASS, 0 - 21)	5.67 (3.79)	4.88 (4.78)	270[1]	>.999	0.19[4]
Arousal (pre) (SAM, 0 - 9)	4.95 (1.36)	4.71 (1.74)	250[1]	>.999	0.12[4]
Arousal (post) (SAM, 0 - 9)	4.29 (1.65)	4.38 (1.88)	220	>.999	<0.01[4]
Sleepiness (pre) (KSS, 1 - 10)	5.19 (1.75)	5.19 (1.81)	224[1]	>.999	0.02
Sleepiness (post) (KSS, 1 - 10)	5.38 (1.72)	5.19 (1.97)	237[1]	>.999	0.07[4]
General reward sensitivity (pre) (VAS, 1 - 10)	6.18 (2.07)	5.61 (1.87)	257	>.999	0.14[4]
Avoidance motivation (post) (VAS, 0 - 100)	62.62 (27.28)	70.00 (29.83)	164[1]	>.999	0.31[4]
Approach motivation (post) (VAS, 0 - 100)	70.48 (26.17)	50.00 (34.24)	300[1]	.596	0.22[4]

Note. [1]Wilcoxon's *W*. [2]Pearson's Chi-squared test. [3]Bonferroni-Holm corrected. [4]Wilcoxon effect size for independent samples. [5]Phi effect size.

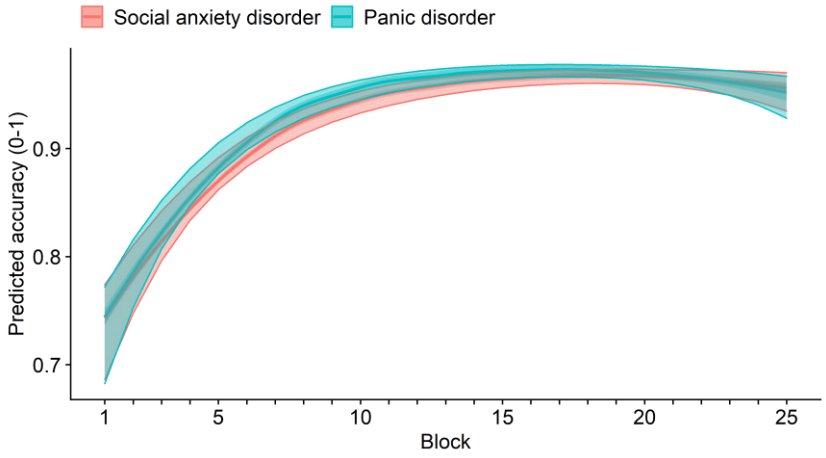
Table E.2

Predictor contributions in the model to predict accuracy in the training phase

	df	X^2	*p*
Block	2	390.26	<.001
Subgroup	1	1.78	.182
Subgroup x Block	2	1.28	.526

Figure E.1

Predicted accuracy during training



Note. Error bands display display 95 % confidence intervals.

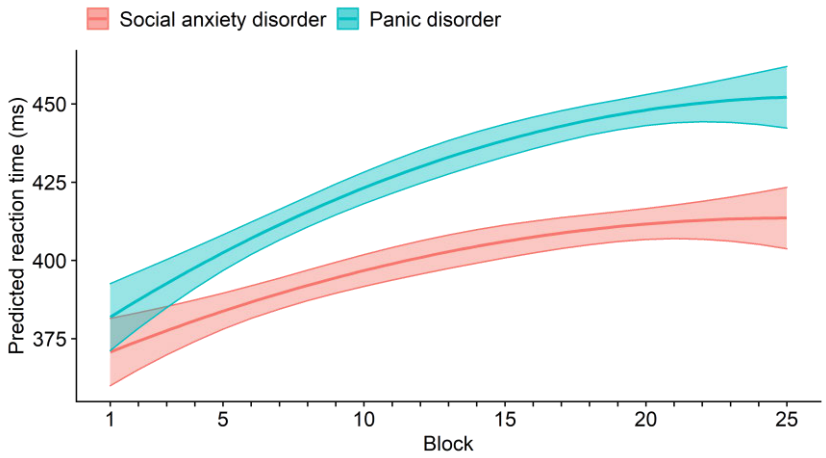
Table E.3

Predictor contributions in the model to predict reaction times in the training phase

	df	*F*	*p*
Block	2	84.87	<.001
Subgroup	1	118.69	<.001
Subgroup x Block	2	4.94	.007

Figure E. 2

Predicted reaction times during training



Note. Error bands display display 95 % confidence intervals.

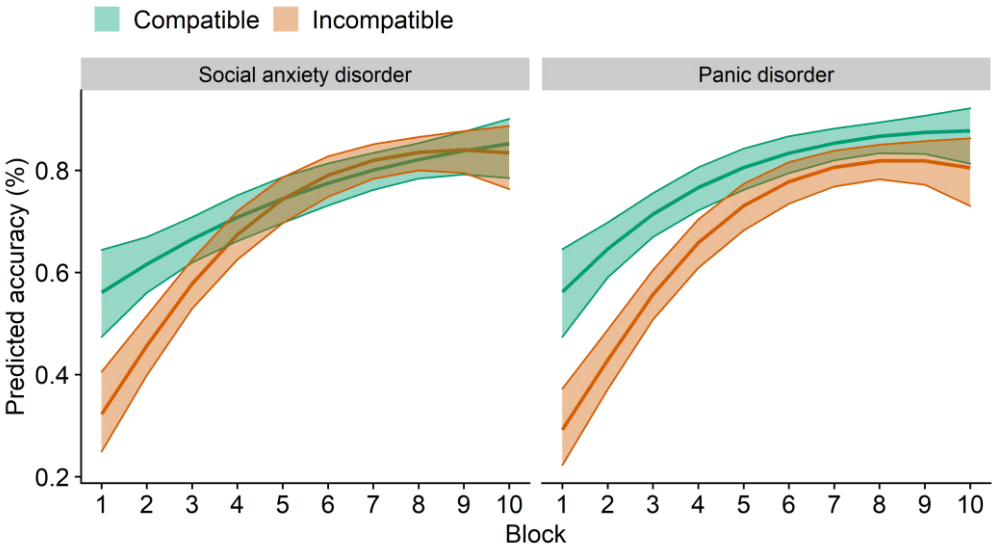
Table E.4

Estimated means to predict accuracy in compatible and incompatible trials

	df	X^2	*p*
Condition	1	21.23	<.001
Subgroup	1	0.80	.370
Block	2	288.38	<.001
Condition x Subgroup	1	4.97	.026
Condition x Block	2	15.26	<.001
Group x Block	2	0.78	.677
Condition x Subgroup x Block	2	0.43	.806

Figure E.3

Predicted accuracy during test



Note. Error bands display display 95 % confidence intervals.

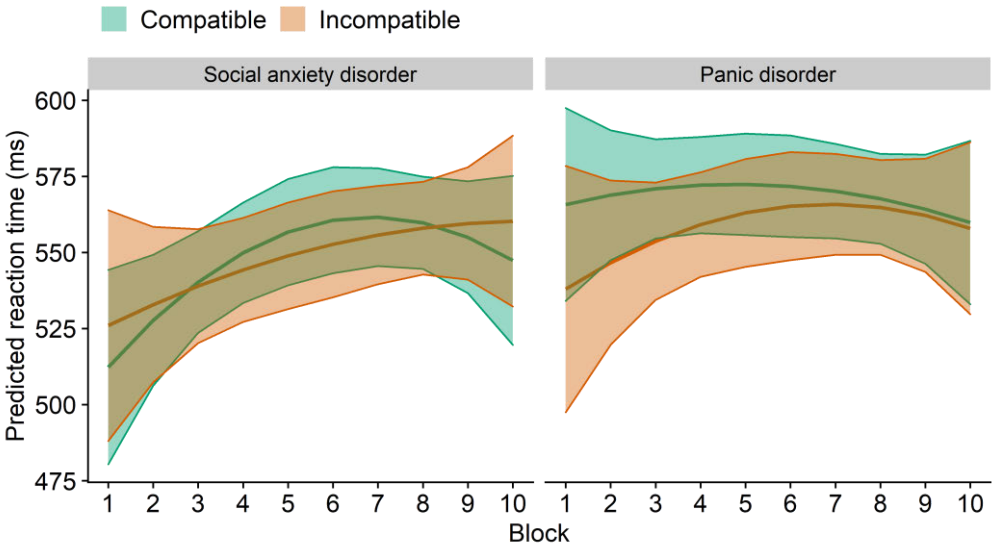
Table E.5

Predictor contributions in the model to predict reaction times in compatible and incompatible trials

	df	*F*	*p*
Condition	1	0.66	.416
Subgroup	1	6.29	.012
Block	2	3.12	.045
Condition x Subgroup	1	0.83	.363
Condition x Block	2	0.38	.946
Subgroup x Block	2	0.92	.398
Condition x Subgroup x Block	2	0.46	.632

Figure E.4

Predicted reaction times during test



Note. Error bands display display 95 % confidence intervals.

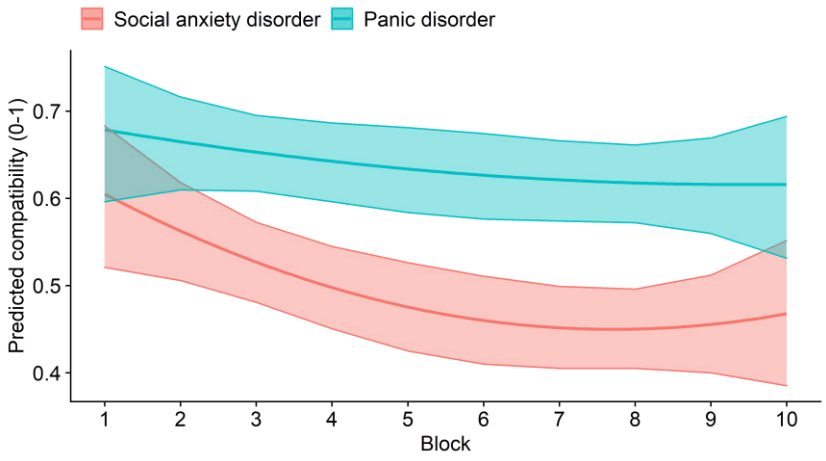
Table E.6

Predictor contributions in the model for responses in free trials

	df	X^2	*p*
Block	2	8.63	.013
Subgroup	1	34.30	<.001
Subgroup x Block	2	1.30	.522

Figure E.5

Predicted percentage of compatible responses in free trials



Note. Error bands display 95 % confidence intervals.

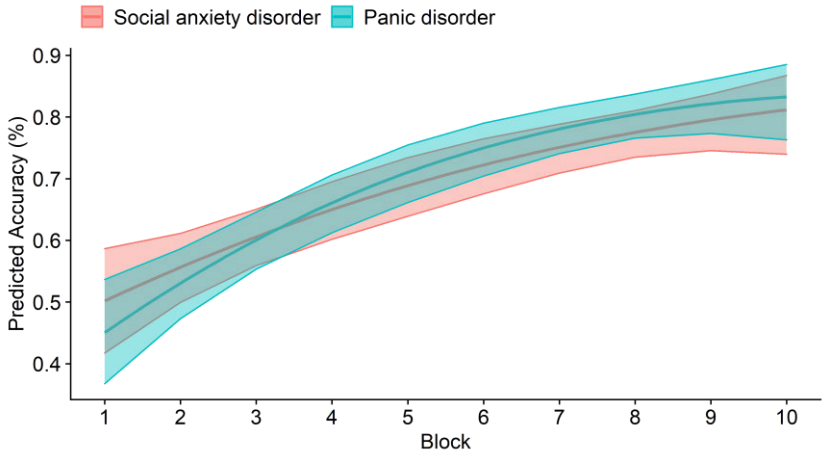
Table E.7

Significance of variable contributions in the model for accuracy in neutral trials

Variable	*df*	X^2	*p*
Block	2	97.60	<.001
Subgroup	1	0.31	.580
Subgroup x Block	2	1.43	.489

Figure E.6

Predicted accuracy in neutral trials



Note. Error bands display display 95 % confidence intervals.

Supplement F: Replication analysis

We aimed to replicate the results from an earlier study with a very similar experimental task, which, however, included a slightly more extensive test phase (i.e., 15 blocks with 18 trials each instead of 10 blocks with 18 trials each in the current experiment; see Experiment 1 in Glück et al., 2021). We reproduced the statistical analysis precisely as reported in this earlier paper. This replication analysis did not test the potential effects of group or anxiety symptom strength but only the effects of condition and block. The Pearson correlations with accuracy in neutral trials needed to be recalculated because they had not been computed in the previous paper. The main question of the replication analyses was, first, whether the extensive training would induce lasting compatibility effects in the test phase of this experiment (i.e., accuracy compatibility effect, reaction time compatibility effect, and compatibility effect in free trials) and, second, whether the different compatibility effects would correlate with each other, supporting the interpretation that the different compatibility effects indicate a general tendency towards repeating extensively trained responses.

The ANOVA for *accuracy* with the factors *block* (first vs. second half of blocks) and c*ondition* (compatible vs. incompatible trials) yielded a significant interaction, $F(1, 123) = 4.55$, $p = .035$, $\eta^2 = .003$, indicating that the accuracy compatibility effect was larger in the first than in the second half of blocks. Replicating the previous study's results, pairwise post-hoc comparisons within the blocks revealed significant accuracy compatibility effects in the first and second half of trials, ps $\leq .002$, ds $\geq .291$. Also, in line with the previous study's results, the ANOVA for *reaction times* with the factors *block* (first vs. second half of blocks) and *condition* (compatible vs. incompatible) yielded no significant interaction, $F(1,123) = 0.11$, $p = .740$, $\eta^2 < .001$, and no significant main effects, all Fs ≤ 0.514, all ps $\geq .475$, all ds $< .001$. The *free trial compatibility effect* was analyzed with a Wilcoxon test. On average, the previously reinforced response was performed in 55.20 % of responses in free trials (SD = 25.54) with color stimuli that had already been presented during the extensive avoidance training, which was remarkably similar to the previous experiment (M = 55.70%, SD = 29.44). This percentage was significantly above chance level, $W = 3906$, $p < .019$, $r_{bs} = .26$ and did not significantly differ between the blocks, $F(1, 123) = 2.19$, $p = .141$, $\eta^2 = .002$. Post-hoc pairwise comparisons revealed that the accuracy compatibility effect in free trials was significantly larger than chance in the first half of trials, $W = 3655$, $p = .001$, $r_{bs} = .264$, and tended to be significantly larger than chance in the second half of blocks, $W = 2550$, $p = .070$, $r_{bs} = .137$, again indicating that the extensive training influenced the post-devaluation responses.

Supplementary Material for Study 3

Supplement A: Descriptive information

Table A.1

Correlations of the anxiety and depression questionnaire scores

	Variable	*n*	*M*	*SD*	1	2	3	4	5
1.	ASI-3	93	21.17	12.51	—				
2.	DASS-21, Anxiety subscale	95	4.25	5.95	.58***	—			
3.	NEO-PI-R, N1 scale	95	16.25	6.81	.54***	.52***	—		
4.	STAI-T	95	38.68	11.14	.66***	.66***	.76***	—	
5.	STAI-T, Anxiety factor	95	14.29	4.85	.66***	.66***	.78***	.94***	—
6.	DASS-21, Depression subscale	95	6.06	6.60	.47***	.57***	.50***	.68***	.67***

Note. Pearson correlations. *p < .05, **p < .01, ***p < .001 (Bonferroni-Holm corrected).

Table A.2

Correlations of the average compatibility effects between tasks

			Avoidance training task			
		Variable	1	2	3	4
Approach training task	1.	Accuracy compatibility effect	-.03	-	-	-
	2.	RT compatibility effect	-	.01	-	-
	3.	Free trials compatibility effect	-	-	-.05	-
	4.	Accuracy in neutral trials	-	-	-	.23

Note. Pearson correlations. *p < .05 (Bonferroni-Holm corrected).

Table A.3

Correlations of the compatibility effects in the avoidance training task

	Variable	*n*	*M*	*SD*	1	2	3	4
1.	Accuracy compatibility effect	95	-0.2%	5.2	1.00	-.18	.16	.04
2.	Reaction time compatibility effect	95	<0.1 ms	21.4	-	1.00	.10	<.01
3.	Free trials compatibility effect	95	49.5%	3.6	-	-	1.00	.20
4.	Accuracy in neutral trials	95	92.1%	6.4	-	-	-	1.00

Note. Pearson correlations. *p < .05 (Bonferroni-Holm corrected).

Table A.4

Correlations of the compatibility effects within the approach training task

	Variable	*n*	*M*	*SD*	1	2	3	4
1.	Accuracy compatibility effect	95	1.6%	5.9	1.00	.11	.01	-.20
2.	Reaction time compatibility effect	95	3.4 ms	23.6	-	1.00	-.12	-.04
3.	Free trial habit effect	95	50.3%	4.4		-	1.00	.03
4.	Accuracy in neutral trials	95	91.7%	6.9	-	-	-	1.00

Note. Pearson correlations. *p < .05 (Bonferroni-Holm corrected).

Figure A.1

Distributions of the anxiety measures in the complete sample (N = 95)

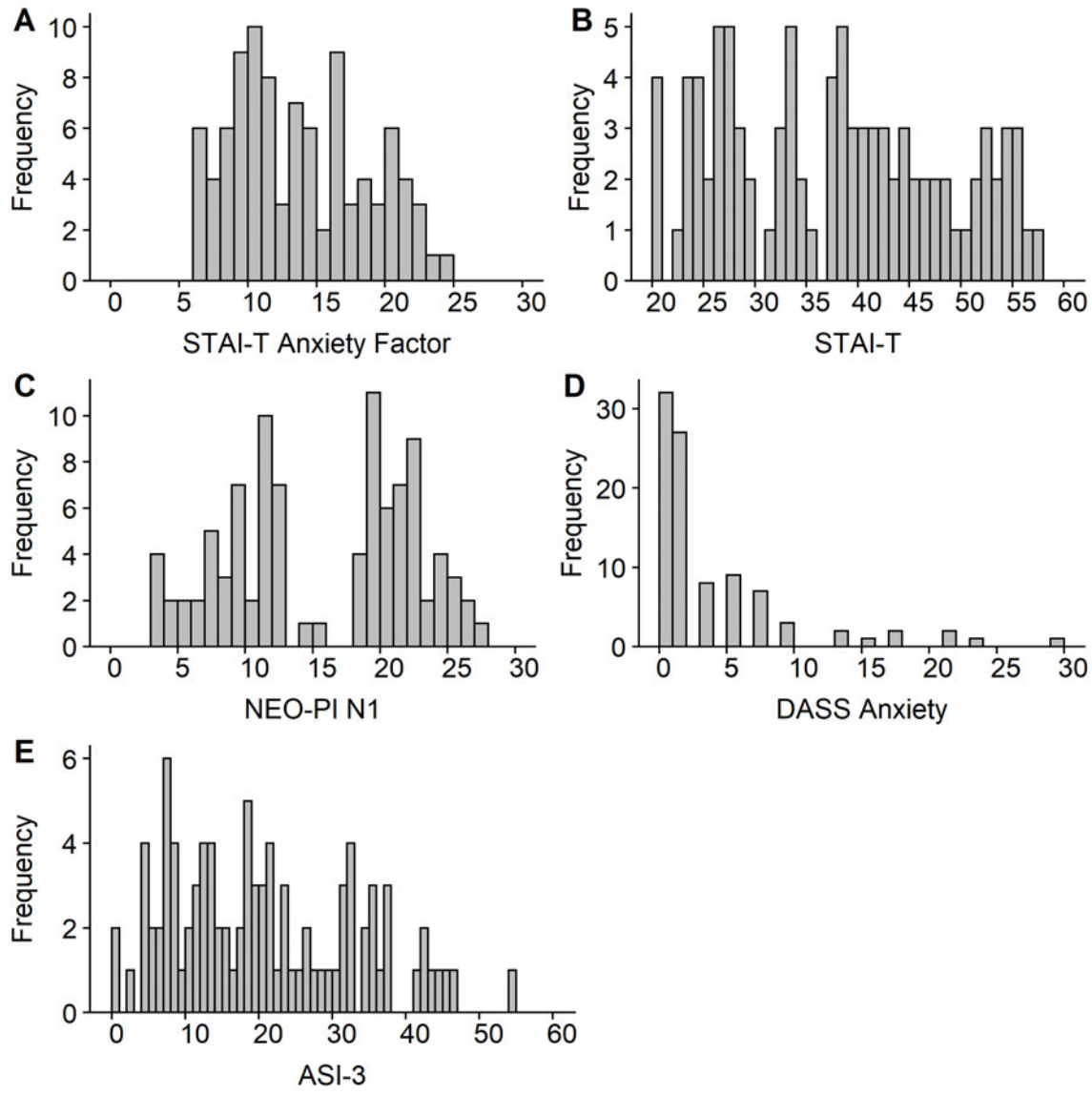
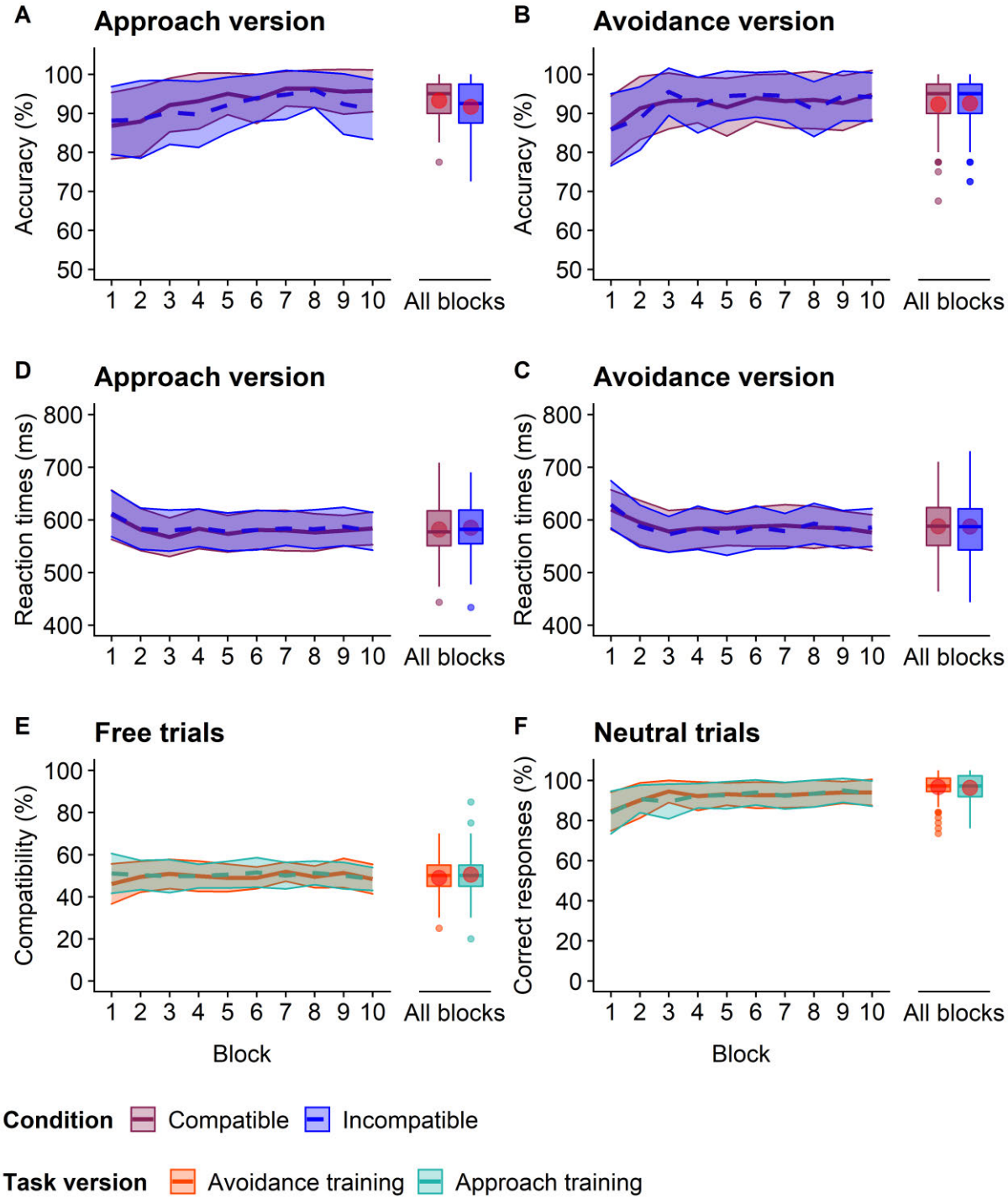


Figure A.2

Descriptive behavioral data



Note. Significance bands display standard errors of the mean.

Supplement B: Detailed GLMM and LMM results

Table B.1

GLMM for the accuracy in training

	df	X^2	*p*
Intercept	1	909.36	<.001***
Task	1	7.45	.006**
Trait anxiety	1	0.62	.431
Block	2	1281.35	<.001***
Session number	1	197.28	<.001***
Trait anxiety x Task	1	0.33	.565
Task x Block	2	12.92	.002**
Trait anxiety x Block	2	1.79	.409
Task x Trait anxiety x Block	2	11.05	.004**

Note. $N = 95$.

Table B.2

Exploratory GLMM with the DASS anxiety subscale sum as anxiety indicator to predict accuracy in training

	df	X^2	*p*
Intercept	1	877.99	<.001***
Task	1	7.28	.007**
DASS anxiety	1	0.06	.799
Block	2	1374.90	<.001***
Session number	1	188.86	<.001***
DASS anxiety x Task	1	0.82	.365
Task x Block	2	11.48	.003**
DASS anxiety x Block	2	1.60	.450
Task x DASS anxiety x Block	2	23.07	<.001***

Table B.3

Exploratory GLMM with the trait ASI-3 sum score as anxiety indicator to predict the accuracy in training

	df	X^2	*p*
Intercept	1	942.30	<.001***
Task	1	10.46	.001**
ASI-3	1	4.06	.044*
Block	2	1219.03	<.001***
Session number	1	221.21	<.001***
ASI-3 x Task	1	5.45	.020*
Task x Block	2	17.43	<.001***
ASI-3 x Block	2	5.18	.075
Task x ASI-3 x Block	2	3.42	.181

Table B.4

Results of the LMM to predict reaction times during training

	df	X^2	*p*
Intercept	1	6214.95	<.001***
Task	1	189.24	<.001***
Trait anxiety	1	2.16	.142
Block	2	903.13	<.001***
Session number	1	0.31	.579
Trait anxiety x Task version	1	0.02	.895
Task x Block	2	31.57	<.001***
Trait anxiety x Block	2	3.04	.219
Task x Trait anxiety x Block	2	2.48	.289

Note. Only data from compatible and incompatible trials.

Table B.5

Results of the GLMM predicting accuracy during the test phase

	df	X^2	*p*
Intercept	1	354.35	< .001***
Condition	1	3.50	.061
Task	1	2.24	.135
Trait anxiety	1	0.13	.716
Block	2	133.27	< .001***
Session number	1	1.72	.190
Condition x Task	1	7.53	.006**
Condition x Trait anxiety	1	0.17	.677
Task x Trait anxiety	1	1.04	.308
Condition x Block	2	3.16	.206
Task x Block	2	2.67	.263
Trait anxiety x Block	2	1.12	.570
Condition x Task x Trait anxiety	1	3.75	.053
Condition x Task x Block	2	4.68	.096
Block x Condition x Trait anxiety	2	2.36	.307
Task x Trait anxiety x Block	2	0.40	.818
Block x Condition x Task x Trait anxiety	2	1.17	.557

Note. Only data from compatible and incompatible trials.

Table B.6

Results of the LMM predicting reaction times during the test phase

	df	X^2	*p*
Intercept	1	4411.20	< .001***
Condition	1	0.87	.352
Task	1	19.89	< .001***
Trait anxiety	1	1.05	.306
Block	2	91.94	< .001***
Session number	1	2.17	.141
Condition x Task	1	1.04	.310

Table B.6 (continued)

	df	X^2	*p*
Condition x Trait anxiety	1	6.64	.010*
Task x Trait anxiety	1	10.19	.001**
Condition x Block	2	1.25	.536
Task x Block	2	1.24	.539
Trait anxiety x Block	2	11.14	.004**
Condition x Task x Trait anxiety	1	0.10	.754
Condition x Task x Block	2	3.83	.147
Block x Condition x Trait anxiety	2	2.08	.354
Task x Trait anxiety x Block	2	5.04	.080
Block x Condition x Task x Trait anxiety	2	2.46	.293

Note. Only data from compatible and incompatible trials.

Table B.7

GLMM to predict compatible responses in free trials

	df	X^2	*p*
Intercept	1	0.08	.782
Block	2	1.00	.605
Session number	1	0.35	.556
Task	1	0.44	.475
Trait anxiety	1	0.51	.475
Trait anxiety x Task version	1	0.13	.723
Task x Block	2	0.75	.687
Block x Trait anxiety	2	0.62	.735
Block x Task x Trait anxiety	2	3.16	.206

Table B.8

LMM to predict reaction times in free trials

	df	X^2	*p*
Intercept	1	5891.33	<.001***
Block	2	113.91	<.001***
Session number	1	4.92	.027*
Task	1	5.43	.020*
Trait anxiety	1	0.20	.656
Trait anxiety x Task	1	4.57	.033*
Task x Block	2	0.75	.686
Block x Trait anxiety	2	1.07	.585
Block x Task x Trait anxiety	2	4.24	.120

Table B.9

Exploratory LMM to predict reaction times in free trials

	df	X^2	*p*
Intercept	1	5886.34	<.001***
Block	2	112.11	<.001***
Session number	1	4.94	.019*
Task	1	5.49	.027*
Compatibility	1	0.98	.322
Trait anxiety	1	0.20	.658
Trait anxiety x Task	1	4.59	.032*
Task x Block	2	0.75	.686
Block x Trait anxiety	2	1.13	.569
Block x Task x Trait anxiety	2	4.24	.120
Compatibility x Task	1	1.03	.310
Compatibility x Block	2	0.01	.995
Compatibility x Trait anxiety	1	0.59	.444
Compatibility x Trait anxiety x Task	1	< 0.01	.921
Compatibility x Task x Block	1	1.61	.448
Compatibility x Trait anxiety x Task	1	3.76	.153
Block x Task x Trait anxiety x Compatibility	2	3.53	.171

Table B.10

GLMM to predict accuracy in neutral control trials

	df	X^2	*p*
Intercept	1	344.76	<.001***
Block	2	65.71	<.001***
Session number	1	0.02	.879
Task	1	0.12	.728
Trait anxiety	1	0.02	.885
Trait anxiety x Task	1	0.22	.639
Task x Block	2	0.47	.789
Block x Trait anxiety	2	1.03	.599
Block x Task x Trait anxiety	2	0.26	.880

Table B.11

LMM to predict reaction times in neutral control trials

	df	X^2	*p*
Intercept	1	4564.85	<.001***
Block	2	52.52	<.001***
Session number	1	0.30	.586
Task	1	9.35	.002**
Trait anxiety	1	0.62	.429
Trait anxiety x Task	1	0.03	.873
Task x Block	2	1.21	.546
Block x Trait anxiety	2	4.96	.084
Block x Task x Trait anxiety	2	1.38	.502

Supplement C: Exploratory analyses including SDNN and heart rate

Table C.1

GLMM to predict accuracy in training

Variable	*df*	X^2	*p*
Intercept	1	811.70	<.001***
Task	1	8.50	.004**
Trait anxiety	1	1.00	.318
Block	2	859.69	<.001***
SDNN	1	8.49	.004**
Heart rate	1	0.06	.812
Session number	1	149.73	<.001***
Trait anxiety x Task	1	0.15	.694
Task x Block	2	4.03	.133
Trait anxiety x Block	2	0.86	.651
Task x SDNN	1	2.11	.146
Task x Trait anxiety x Block	2	5.23	.073

Table C.2

LMM to predict reaction times during training

	df	X^2	*p*
Intercept	1	6349.26	<.001***
Task	1	194.36	<.001***
Trait anxiety	1	2.21	.137
Block	2	903.46	<.001***
Session number	1	0.28	.598
SDNN	1	9.06	.003**
Heart rate	1	21.43	<.001***
Trait anxiety x Task version	1	0.08	.775
Task x Block	2	31.96	<.001***
Trait anxiety x Block	2	3.10	.212
Task x SDNN	1	1.82	.177
Task x Trait anxiety x Block	2	2.49	.288

Note. Only data from compatible and incompatible trials.

Table C.3

GLMM predicting accuracy during the test phase

	df	X^2	*p*
Intercept	1	363.21	< .001***
Condition	1	3.73	.053
Task	1	3.06	.080
Trait anxiety	1	0.09	.764
Block	2	120.25	< .001***
SDNN	1	2.48	.115
Heart rate	1	0.68	.410
Session number	1	0.92	.337
Condition x Task	1	7.22	.007**
Condition x Trait anxiety	1	0.16	.689
Task x Trait anxiety	1	1.25	.264
Condition x Block	2	2.76	.251
Task x Block	2	2.41	.300
Trait anxiety x Block	2	1.15	.563
Condition x SDNN	1	0.08	.783
Task x SDNN	1	0.59	.445
Condition x Task x Trait anxiety	1	3.52	.061
Condition x Task x Block	2	4.46	.108
Block x Condition x Trait anxiety	2	2.07	.355
Task x Trait anxiety x Block	2	0.38	.825
Condition x Task x SDNN	1	0.71	.399
Block x Condition x Task x Trait anxiety	2	1.08	.584

Note. Only data from compatible and incompatible trials.

Table C.4

LMM predicting reaction times during the test phase

	df	X^2	*p*
Intercept	1	4228.01	< .001***
Condition	1	0.65	.420
Task	1	23.24	< .001***
Trait anxiety	1	0.93	.336
Block	2	91.92	< .001***
SDNN	1	5.21	.023*
Heart rate	1	1.14	.285
Session number	1	1.81	.178
Condition x Task	1	1.22	.269
Condition x Trait anxiety	1	6.17	.013*
Task x Trait anxiety	1	9.41	.002**
Condition x Block	2	1.25	.536
Task x Block	2	0.93	.627
Trait anxiety x Block	2	11.15	.004**
Condition x SDNN	1	1.75	.187
Task x SDNN	1	0.13	.716
Condition x Task x Trait anxiety	1	0.11	.746
Condition x Task x Block	2	3.83	.147
Block x Condition x Trait anxiety	2	2.08	.353
Task x Trait anxiety x Block	2	5.06	.080
Condition x Task x SDNN	1	<0.01	.978
Block x Condition x Task x Trait anxiety	2	2.40	.301

Note. Only data from compatible and incompatible trials.

Table C.5

GLMM to predict compatible responses in free trials

	df	X^2	*p*
Intercept	1	0.08	.783
Block	2	0.95	.622
Session number	1	0.39	.530
SDNN	1	0.04	.842
Heart rate	1	0.09	.762
Task	1	0.51	.475
Trait anxiety	1	0.51	.475
Trait anxiety x Task version	1	0.12	.731
Task x Block	2	0.73	.695
Block x Trait anxiety	2	0.60	.740
Task x SDNN	1	0.27	.301
Block x Task x Trait anxiety	2	3.16	.207

Table C.6

LMM to predict reaction times in free trials

Variable	*df*	X^2	*p*
Intercept	1	5572.98	<.001***
Block	2	114.67	<.001***
Session number	1	1.49	.223
SDNN	1	9.99	.002**
Heart rate	1	5.77	.016*
Task	1	3.66	.056
Trait anxiety	1	0.20	.656
Trait anxiety x Task	1	5.88	.015*
Task x Block	2	0.72	.697
Block x Trait anxiety	2	1.08	.584
Task x SDNN	1	7.08	.008**
Block x Task x Trait anxiety	2	4.17	.124

Table C.7

GLMM to predict accuracy in neutral control trials

Variable	*df*	X^2	*p*
Intercept	1	339.50	<.001***
Block	2	67.09	<.001***
Session number	1	0.02	.900
SDNN	1	0.21	.648
Heart rate	1	0.03	.870
Task	1	0.12	.728
Trait anxiety	1	0.03	.870
Trait anxiety x Task	1	0.19	.663
Task x Block	2	0.47	.792
Block x Trait anxiety	2	1.03	.599
Task x SDNN	1	<0.01	.958
Block x Task x Trait anxiety	2	0.26	.878

Table C.8

LMM to predict reaction times in neutral control trials

Variable	*df*	X^2	*p*
Intercept	1	4307.26	<.001***
Block	2	52.92	<.001***
Session number	1	0.80	.372
SDNN	1	8.97	.003**
Heart rate	1	0.02	.880
Task	1	12.78	<.001***
Trait anxiety	1	0.44	.508
Trait anxiety x Task	1	0.03	.873
Task x Block	2	1.20	.549
Block x Trait anxiety	2	5.01	.082
Task x SDNN	1	0.01	.907
Block x Task x Trait anxiety	2	1.37	.504

Supplement D: Exploratory analysis with ASI-3 score as trait anxiety indicator

Table D.1

Results of the GLMM predicting accuracy in the test phase

Variable	*df*	X^2	*p*
Intercept	1	363.21	< .001***
Condition	1	3.70	.054
Task	1	1.99	.159
Trait anxiety (ASI)	1	3.69	.055
Block	2	128.44	< .001***
Session number	1	1.56	.212
Condition x Task	1	7.15	.008**
Condition x Trait anxiety (ASI)	1	0.12	.731
Task x Trait anxiety (ASI)	1	0.18	.674
Condition x Block	2	3.69	.158
Task x Block	2	3.22	.200
Trait anxiety x Block	2	1.37	.503
Condition x Task x Trait anxiety	1	1.90	1.68
Condition x Task x Block	2	4.13	.127
Block x Condition x Trait anxiety	2	4.53	.104
Task x Trait anxiety x Block	2	0.96	.618
Block x Condition x Task x Trait anxiety	2	0.72	.698

Note. Only data from compatible and incompatible trials.

Table D.2

Results of the LMM predicting reaction times during the test phase

Variable	*df*	X^2	*p*
Intercept	1	4284.57	< .001***
Condition	1	1.52	.218
Task	1	16.88	< .001***
Trait anxiety (ASI)	1	3.76	.053
Block	2	91.94	< .001***
Session number	1	1.63	.202

Table D.2 (continued)

	df	X^2	*p*
Condition x Task	1	0.91	.340
Condition x Trait anxiety	1	0.48	.489
Task x Trait anxiety	1	4.90	.027
Condition x Block	2	1.26	.532
Task x Block	2	1.66	.437
Trait anxiety x Block	2	1.27	.531
Condition x Task x Trait anxiety	1	0.18	.674
Condition x Task x Block	2	3.68	.159
Block x Condition x Trait anxiety	2	0.06	.972
Task x Trait anxiety x Block	2	1.06	.588
Block x Condition x Task x Trait anxiety	2	1.82	.402

Note. Only data from compatible and incompatible trials.

Table D.3

GLMM to predict compatible responses in free trials

Variable	*df*	X^2	*p*
Intercept	1	0.14	.705
Block	2	1.18	.555
Session number	1	0.37	.545
Task	1	0.28	.598
Trait anxiety	1	0.28	.599
Trait anxiety x Task version	1	0.12	.729
Task x Block	2	0.89	.641
Block x Trait anxiety	2	0.66	.720
Block x Task x Trait anxiety	2	2.29	.318

Table D.4

LMM to predict response times in free trials

Variable	*df*	X^2	*p*
Intercept	1	6204.57	<.001
Block	2	110.28	<.001
Session number	1	3.80	.051
Task	1	4.70	.030
Trait anxiety	1	0.50	.480
Trait anxiety x Task version	1	0.19	.667
Task x Block	2	0.75	.688
Block x Trait anxiety	2	3.61	.165
Block x Task x Trait anxiety	2	0.17	.920

Table D.5

GLMM to predict correct responses (i.e., accuracy) in neutral trials

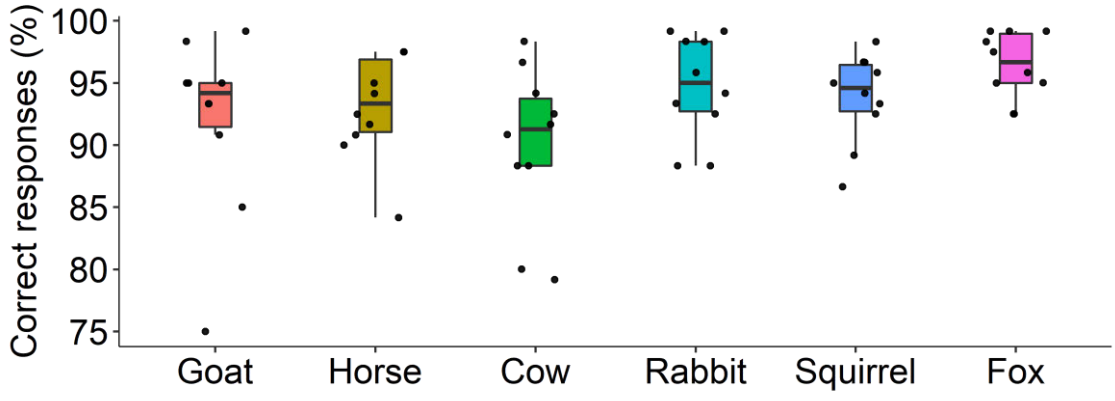
Variable	*df*	X^2	*p*
Intercept	1	344.09	<.001
Block	2	58.59	<.001
Session number	1	<0.01	.986
Task	1	0.05	.826
Trait anxiety	1	1.79	.182
Trait anxiety x Task version	1	0.74	.389
Task x Block	2	0.65	.723
Block x Trait anxiety	2	0.60	.741
Block x Task x Trait anxiety	2	0.53	.769

Supplement E: Pilot study results

Ten healthy participants without self-reported cardiovascular, respiratory or neurological diseases, acute physical illness, bipolar or psychotic disorders, current psychopharmacological medication, or pregnancy, were paid 9 € or received one hour course credit. The participants took part in the pilot study individually. During the pilot experiment, the participants wore headphones which were used to signal correct responses with a soft tone. The trial structure was identical to the structure of trials in the test phase of the avoidance training phase as described in the main paper. The participants were instructed that they would see pictures from two categories of animals on the computer screen, and that for one category, the left of two target buttons on the computer keyboard would be correct, while for the other category, the right target button would be correct. For each correct response, they would receive a small monetary rewards, which was signaled by a realistic picture of a 50 Cent coin. The participants were presented with three sets of pictures. Each set contained pictures from two animal categories with 60 pictures per animal category in 15 blocks with 8 pictures each. The pictures were black-and-white pictures of animals with a size of 2.1 x 2.1 cm, presented centered on the screen on a white background. The picture sequence within each block were pseudo-randomized. We descriptively analyzed the percentage of correct responses to each picture and each category (see Figure E.1). The pictures from categories which had turned out to be more difficult were re-examined and ambiguous pictures which had exceptionally low accuracy were exchanged for more easily identifiable pictures. The adapted picture sets were then used for the experiment.

Figure E.1

Average accuracy of correct responses per animal category in the pilot study



Note. $N = 10$. Black dots display single individuals' average values.

Statement of individual author contributions

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Graduate School
Life Sciences

Statement of individual author contributions and of legal second publication rights to manuscripts included in the dissertation

Manuscript 1 (complete reference): Glück, V. M., Zwosta, K., Wolfensteller, U., Ruge, H., & Pittig, A. (2021). Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm. *Behaviour Research and Therapy, 146,* 103964. https://doi.org/10.1016/j.brat.2021.103964

Participated in	**Author Initials**, Responsibility decreasing from left to right				
Study Design Methods Development	AP, VG AP, VG	UW, KZ UW, KZ	HR HR		
Data Collection	VG				
Data Analysis and Interpretation	VG, AP	KZ, UW	HR		
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 VG VG VG VG	 AP AP AP AP	 KZ, UW, HR KZ, UW, HR KZ, UW, HR		

Manuscript 2 (complete reference): Glück, V. M., Boschet-Lange, J. M., Pittig, R., & Pittig, A. (2023). Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm. *Behaviour Research and Therapy, 170,* 104417. https://doi.org/10.1016/j.brat.2023.104417

Participated in	**Author Initials**, Responsibility decreasing from left to right				
Study Design Methods Development	VG, AP VG, AP	JB JB			
Data Collection	VG	JB	AP	RP	
Data Analysis and Interpretation	VG	AP	JB		
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 VG VG VG VG	 AP AP AP	 JB JB JB	 RP RP RP	

Manuscript 3 (complete reference): The impact of trait anxiety on approach and avoidance habits in an outcome devaluation paradigm. [Manuscript in preparation]					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	VG VG	AP AP			
Data Collection	VG				
Data Analysis and Interpretation	VG	AP			
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 VG VG VG VG	 AP AP AP			

If applicable, the doctoral researcher confirms that she/he has obtained permission from both the publishers (copyright) and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Valentina Glück

Doctoral Researcher's Name	Date	Place	Signature
Prof. Dr. Andre Pittig	11.12.2023	Göttingen	
Primary Supervisor's Name	Date	Place	Signature

Statement of individual author contributions to figures/tables/chapters

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Graduate School
Life Sciences

Statement of individual author contributions to figures/tables of manuscripts included in the dissertation

Manuscript 1 (complete reference):

Glück, V. M., Zwosta, K., Wolfensteller, U., Ruge, H., & Pittig, A. (2021). Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm. *Behaviour Research and Therapy, 146,* 103964. https://doi.org/10.1016/j.brat.2021.103964

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	VG	AP			
2	VG	AP			
3	VG	AP			
4	VG	AP			
5	VG	AP			

Manuscript 2 (complete reference):

Glück, V. M., Boschet-Lange, J. M., Pittig, R., & Pittig, A. (2023). Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm. *Behaviour Research and Therapy, 170,* 104417. https://doi.org/10.1016/j.brat.2023.104417

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	VG	AP	JB		
2	VG	AP	JB		
3	VG	AP	JB		
Table					
1	VG	AP	JB	RP	

Manuscript 3 (complete reference):

Glück, V. M. & Pittig, A. (2023). The impact of trait anxiety on approach and avoidance habits in an outcome devaluation paradigm. [Manuscript in preparation].

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	VG	AP			
2	VG	AP			
3	VG	AP			
4	VG	AP			
Table	**Author Initials,** Responsibility decreasing from left to right				
1	VG	AP			

I also confirm my primary supervisor's acceptance.

Valentina Glück	Würzburg	19.12.2023	
Doctoral Researcher's Name	Date	Place	Signature

List of publications

Glück, V. M., Boschet-Lange, J. M., Pittig, R., & Pittig, A. (2023). Persistence of extensively trained avoidance is not elevated in anxiety disorders in an outcome devaluation paradigm. *Behaviour Research and Therapy, 170*, 104417. https://doi.org/10.1016/j.brat.2023.104417

Glück, V. M., Engelke, P., Hilger, K., Wong, A. H. K., Boschet, J. M., & Pittig, A. (2023). A network perspective on real-life threat, anxiety, and avoidance. *Journal of Clinical Psychology.* Advance online publication. https://doi.org/10.1002/jclp.23575

Glück, V. M., Zwosta, K., Wolfensteller, U., Ruge, H., & Pittig, A. (2021). Costly habitual avoidance is reduced by concurrent goal-directed approach in a modified devaluation paradigm. *Behaviour Research and Therapy*, *146*, 103964. https://doi.org/10.1016/j.brat.2021.103964

Pittig, A., **Glück, V. M.,** Boschet, J. M., Wong, A. H. K., & Engelke, P. (2021). Increased Anxiety of Public Situations During the COVID-19 Pandemic: Evidence From a Community and a Patient Sample. *Clinical Psychology in Europe*, *3*(2), e4221. https://doi.org/10.32872/cpe.4221

Pittig, A., Boschet, J. M., **Glück, V. M.**, & Schneider, K. (2021). Elevated costly avoidance in anxiety disorders: Patients show little downregulation of acquired avoidance in face of competing rewards for approach. *Depression and Anxiety*, *38*(3), 361–371. https://doi.org/10.1002/da.23119

Wong, A. H. K., **Glück, V. M**., Boschet, J. M., & Engelke, P. (2020). Generalization of extinction with a generalization stimulus is determined by learnt threat beliefs. *Behaviour Research and Therapy*, *135*, 103755. https://doi.org/10.1016/j.brat.2020.103755

Pittig, A., Wong, A. H. K., **Glück, V. M.**, & Boschet, J. M. (2020). Avoidance and its bi-directional relationship with conditioned fear: Mechanisms, moderators, and clinical implications. *Behaviour Research and Therapy*, *126*, 103550. https://doi.org/10.1016/j.brat.2020.103550

Affidavit

I hereby confirm that my thesis entitled *Habitual avoidance in trait anxiety and anxiety disorders* is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

_____ _____

Place, Date Signature

Eidesstaatliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation *Habituelles Vermeidungsverhalten bei Ängstlichkeit und Angststörungen* eigenständig, d.h. insbesondere selbstständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

_____ _____

Ort, Datum Unterschrift