Behavior Research Methods (2026) 58:7
https://doi.org/10.3758/s13428-025-02865-2

ORIGINAL MANUSCRIPT

Assessing the validity evidence for habit measures based on time pressure

Pablo Martínez-López[1,2] · **Antonio Vázquez-Millán**[1,2] · **Francisco Garre-Frutos**[3,4] · **David Luque**[1,2]

Received: 13 March 2025 / Accepted: 30 September 2025

Abstract

Animal research has shown that repeatedly performing a rewarded action leads to its transition into a habit—an inflexible response controlled by stimulus–response associations. Efforts to reproduce this principle in humans have yielded mixed results. Only two laboratory paradigms have demonstrated behavior habitualization following extensive instrumental training compared to minimal training: the forced-response task and the "aliens" outcome-devaluation task. These paradigms assess habitualization through distinct measures. The forced-response task focuses on the persistence of a trained response when a reversal is required, whereas the outcome-devaluation task measures reaction time switch costs—slowdowns in goal-directed responses conflicting with the trained habit. Although both measures have produced results consistent with the learning theory—showing stronger evidence of habits in overtrained conditions—their construct validity remains insufficiently established. In this study, participants completed 4 days of training in each paradigm. We replicated previous results in the forced-response task; in the outcome-devaluation task, a similar pattern emerged, observing the loss of a response speed advantage gained through training. We then examined the reliability of each measure and evaluated their convergent validity. Habitual responses in the forced-response task and reaction time switch costs in the outcome-devaluation task demonstrated good reliability, allowing us to assess whether individual differences remained stable. However, the two measures were not associated, providing no evidence of convergent validity. This suggests that these measures capture distinct aspects of the balance between habitual and goal-directed control. Our results highlight the need for further evaluation of the validity and reliability of current measures of habitual control in humans.

Keywords Habit formation · Learning · Replicability · Reliability · Convergent validity

Reward-based decision-making is critical for adaptive behavior. However, determining the best course of action—one that leads to the most favorable outcome—can be cognitively demanding, particularly in complex real-life situations. Within the associative dual-process framework, goal-directed processes evaluate different action plans and select the optimal option to achieve a desired outcome (de Wit & Dickinson, 2009). While goal-directed decisions are typically well aligned with an individual's actual needs, even in complex and uncertain environments, they come at the cost of high cognitive load and require time to execute. In contrast, habitual behavior is guided by stimulus–response (S–R) memory representations that develop through repeated experiences. Once a habit is established, perceiving the stimulus (S) directly activates the associated response (R), allowing for rapid and often effective decision-making, particularly in time-sensitive situations—such as braking when a traffic light suddenly turns red—or when cognitive resources are allocated elsewhere, like driving home while engaged in conversation with a passenger (Ashby et al., 2010; Graybiel & Grafton, 2015; for discussion, see Haith & Krakauer, 2018). Recent research has shown a renewed

✉ Pablo Martínez-López
pablomartinezlopez@uma.es

✉ David Luque
david.luque@gmail.com

1 Department of Basic Psychology, Faculty of Psychology and Speech Therapy, University of Málaga, Málaga, Spain

2 The Malaga Biomedical Research Institute and Nanomedicine Platform-IBIMA BIONAND Platform, Málaga, Spain

3 Department of Experimental Psychology, Faculty of Psychology, University of Granada, Granada, Spain

4 Mind, Brain and Behavior Research Center (CIMCYC), University of Granada, Granada, Spain

Published online: 24 November 2025

interest in studying these two complementary systems, as an imbalance between them is thought to contribute to the persistence of maladaptive behaviors, particularly in cases of excessive transdiagnostic impulsivity (Chen et al., 2024; Gillan et al., 2016; Sookud et al., 2025).

Habits result from extensive training and tend to govern behavior only in situations of low uncertainty and risk. Compelling evidence from studies with nonhuman animals shows that actions transition from goal-directed to habitual when repeated often enough in stable contexts (Adams, 1982; Dickinson et al., 1995; Moore et al., 2023; Thrailkill & Bouton, 2015). Animal laboratory research often employs instrumental overtraining to induce habit formation and the outcome devaluation paradigm to determine whether an action is goal-directed or habitual (Dickinson, 1985; Watson, 2024). In these experiments, animals are trained to form response–outcome associations that are valid in the presence of certain environmental stimuli (or discriminative stimuli)—for instance, when a light is turned on, pressing a lever to receive food. Importantly, there are varying levels of training on this instrumental relationship: overtraining versus minimal training. Habit formation is typically confirmed after overtraining when the previously valued reward is devalued—either through satiation or conditioned taste aversion (e.g., food poisoning)—and the animal continues to select the response linked to the now-devalued outcome in the presence of the discriminative stimulus. This is taken as evidence for the activation of a direct S–R memory representation that is controlling behavior.

Laboratory studies on human habits have adopted the general methodological approach taken in animal research, attempting to replicate key features of the outcome devaluation test. However, a direct translation has not been feasible, necessitating adaptations that have led to the development of several protocols (for a review, see Guida et al., 2022), some of which extend beyond the laboratory using smartphone-based tasks (Banca et al., 2024; Gera et al., 2024). Despite these variations, most of these protocols evaluate habitual responses by examining choices linked to outcomes that have been devalued (Gera et al., 2024; Gillan et al., 2014; Tricomi et al., 2009; Watson et al., 2014). Measuring habits in humans through the overt selection of devalued outcomes has been extensively employed to investigate the functioning of the habit system in healthy and special populations, as well as its neural substrates (for reviews, see Knowlton & Patterson, 2018; Wood & Rünger, 2016).

Despite the growing body of research on human habits, researchers have struggled to effectively examine the effect of the amount of training on behavior habitualization. As noted above, a central tenet of the dual-systems model of instrumental learning is that the transition from goal-directed to habitual behavior depends on experience, with the habit system prevailing after prolonged, stable training. However, human studies designed to parallel animal research have not consistently demonstrated this pattern (de Wit et al., 2018; Gera et al., 2023; Pool et al., 2022; Molinero et al., 2025). These findings highlight a critical issue in the field of habit learning: if existing protocols and measures fail to capture the expected transition from goal-directedness to habit—a fundamental prediction of the "habit" construct—how can we be certain that these protocols are tapping into the operation of the habit system? In other words, there is a validity issue in the field.

The limitations of human laboratory settings complicate the induction of new habitual behaviors. Real-life habits are formed over months (Lally et al., 2010), whereas laboratory protocols typically involve only a few days of training (at best), which might be insufficient to create strong habits. Additionally, the standard habit test requires the resolution of a response conflict between the outcomes of the goal-directed and habitual systems (Balleine & O'Doherty, 2010; Dolan & Dayan, 2013). Under these circumstances, isolating and detecting arguably weak habits can be challenging due to the dominance of the goal-directed system. Nevertheless, there is a potential approach to address this issue. The goal-directed and habitual systems are thought to operate at different speeds, which may provide an opportunity to detect the influence of the habit system. While S–R associations are triggered rapidly, goal-directed behavior is presumed to be slower (Keramati et al., 2011, 2016; Luque et al., 2017). Given sufficient time, goal-directed processes override habitual responses, allowing adaptation to new conditions. In this regard, most existing habit tests allow relatively slow responses (> 1 s), which may explain the lack of evidence for habit-driven behavior observed in human laboratory paradigms (Watson et al., 2022). In other words, while habits may form through training in these protocols, their expression is often masked by stronger, slower goal-directed decision-making processes. This suggests that one promising approach to making habits more detectable is to impose stricter time constraints on responses.

Supporting this idea, Hardwick et al. (2019) showed that habitual responding is detectable when participants are required to respond rapidly. In their experiments, participants learned four visuomotor associations under two training conditions: minimal (until achieving 20 consecutive correct responses) and overtraining (4 days of practice). Following training, two associations switched responses between them, requiring participants to learn the new associations until they again achieved 20 consecutive correct responses. Participants then completed the forced-response test, in which a stimulus was presented randomly within a sequence of four tones lasting 1.2 s. Participants were instructed to respond according to the updated associations and to synchronize their responses with the last tone, regardless of when the stimulus appeared. As a result, the interval

between the stimulus presentation and the required response varied across trials, manipulating the available preparation time to respond. When the imposed preparation time was too short to process the stimulus, participants were instructed to guess. The findings revealed that the proportion of habit errors—previously trained but now incorrect responses—significantly increased after four days of training, particularly when participants were required to respond rapidly (~300–600 ms). This result aligns closely with the notion that low-latency habitual control competes with slower goal-directed processes.

In this vein, Luque et al. (2020) further examined response interference between habits and goal-directed actions. Although goal-directed and habit systems operate at different speeds, both types of responses could be activated simultaneously within a specific time window—not too short, allowing the activation of the goal-directed response, but not too long, ensuring the habit remains active. Luque et al. hypothesized that during rapid but accurate goal-directed actions, any existing habit would still be active, albeit overridden by the goal-directed system. In such cases, if the habit and goal-directed systems produced conflicting responses (e.g., habit = responding left; goal-directed = responding right), the strength of the underlying habit could be inferred from its interfering effect on goal-directed action. To test this, Luque et al. (2020) designed a task where participants could still achieve a desired outcome during a devaluation test, but only if they switched from a previously trained, now-devalued response to a new, valued one. Consistent with previous findings (e.g., de Wit et al., 2018), the authors did not observe a significant difference in response selection between training conditions (i.e., 3 days vs. 1 day) during the test. However, they identified increased reaction time (RT) when participants successfully switched their responses in the overtraining condition. Notably, this RT switch cost effect was only observed under conditions where the time available to respond was reduced (i.e., up to 700 ms from stimulus onset). This finding suggests that the interfering effect of habits on goal-directed RT is detectable only under specific time constraints.

Luque et al. (2020) and Hardwick et al (2019) provide evidence of habit formation through overtraining under time-pressure conditions. It might be tempting to assume that these tasks will resolve the validity issues discussed earlier. However, while Hardwick et al. (2019) observed a higher proportion of habit errors (i.e., habitual response selection), Luque et al. (2020) identified a slowing of goal-directed responses as compared to baseline (i.e., an RT switch cost effect). Both measures have demonstrated sensitivity to training effects, suggesting they may serve as valid approaches to quantifying the strength of a previously trained response habit. Nonetheless, to further establish the link between these measures and the theoretical construct they aim to quantify, it is essential to test their convergent and discriminant validity (Cronbach & Meehl, 1955). Specifically, are habit errors and RT switch costs related as measures of habit formation?

The study by Nebe et al. (2024) is the only one to date that has attempted to address this question in the context of behavioral habit measures, reporting null correlations among the variables they tested. However, most of the behavioral measures included in their study were not sensitive to overtraining, potentially rendering them unsuitable for studying convergent validity. Moreover, Nebe et al. (2024) did not evaluate the reliability of their measures before examining their correlations. Low measurement reliability can undermine the validation process, as even if two measures of the same latent construct are strongly associated, measurement error can obscure their true relationship, making it difficult—if not impossible—to accurately assess their connection (Dang et al., 2020; Hedge et al., 2018).

Validation is a multifaceted concept often considered to follow a sequential assessment process (Flake et al., 2017) requiring the evaluation of multiple sources of validity (AERA et al., 2014). Our overarching objective was to provide evidence of construct validity for the measures purported to index habit activation in the tests developed by Hardwick et al. (2019) and Luque et al. (2020). More specifically, our preregistered aims were (1) to replicate previous results obtained with the forced-response task (Hardwick et al., 2019) and (2) to assess whether RTs for goal-directed responses are sensitive to overtraining in the forced-response task. Although participants in this paradigm are instructed to synchronize their responses with a tone, we hypothesized that they would exhibit a synchronization delay after overtraining, as they would need to override an existing S–R habit. This effect would be conceptually equivalent to the one observed by Luque et al. (2020). (3) We subsequently aimed to assess the construct validity of the measures identified as sensitive to overtraining. Specifically, we hypothesized that after overtraining, the proportion of habitual responses during the forced-response task and the magnitude of RT switch cost during the outcome-devaluation task would be positively correlated. Additionally, we expected the RT switch cost in the outcome-devaluation task to correlate with the equivalent effect observed in the forced-response task (as described in the second aim). (4) We also examined the reliability of the measures identified as sensitive to overtraining. Although this analysis was not preregistered, we included it because, as previously discussed, a proper assessment of a measure's validity must be accompanied by an evaluation of its reliability. Otherwise, the study would be incomplete and difficult to interpret. It is important to note that the reliability analysis presented here is purely descriptive and limited to the current dataset.

Method

This study was preregistered before data collection. Any non-preregistered analyses and hypotheses are explicitly stated as such in the text. The preregistration document is publicly available at https://osf.io/3bzx6.

Participants

To determine the target sample size, we used the {pwr} R package (Champely, 2006). We followed Simonsohn's small telescope approach (Simonsohn, 2015) to estimate the target effect size for the main analyses conducted on the forced-response test data. This approach is recommended for the replication of studies with large effect sizes and relatively small samples. The author suggests that the target effect size should correspond to the effect size detectable with 33% power in the original study. Hardwick et al. (2019) reported a larger proportion of habitual errors in the 4-day training condition compared to the minimal training condition, with 22 participants, $t(21) = 11.16$, $d_z = 2.38$. The effect size that could have been detected with 33% power in their study was $d_z = 0.34$. Based on this, we aimed to recruit at least 56 participants, ensuring 80% statistical power to detect this effect size with $\alpha = .05$. For the outcome-devaluation task, Luque et al. (2020) conducted a 2×2 repeated-measures analysis of variance (ANOVA) (training condition×stimulus value), revealing a main effect of training condition with $\eta_p^2 = .15$ based on a sample of 47 participants. Our planned sample size of 56 participants should provide adequate statistical power to detect this effect. Notably, when testing the degree of association between measures from both tasks, this planned sample size will enable us to detect a Pearson's r of .32 with 80% power (one-tailed).

Participants were required to visit the laboratory 12 times over 2 months, completing the forced-response task followed by the outcome-devaluation task. They completed the two tasks in this specific order. We did not counterbalance task order, because one of our preregistered objectives was to conduct a direct replication attempt of Experiment 1 of Hardwick et al. (2019); therefore, it was important that our participants start the forced-response task with no prior experience with other reinforcement learning tasks (as was the case in Hardwick et al., 2019).

To account for potential dropouts, we initially invited 126 students from the University of Málaga. Eighty students participated in the study, but six participants withdrew during the forced-response task, and two withdrew during the outcome-devaluation task. All participants had normal or corrected-to-normal vision. Participation was compensated with course credit, and the 10 students who achieved the highest scores in the outcome-devaluation task received a €30 reward. For the forced-response task analyses, we had to exclude 24 participants (see "Data analysis"). One participant was excluded from the outcome-devaluation task due to a technical error. Our final sample size included 50 participants for the forced-response task (43 women; $M_{age} = 19.7$; $SD_{age} = 1.56$), 71 participants (one participant missed a training session) for the outcome-devaluation task (60 women; $M_{age} = 19.87$; $SD_{age} = 1.38$), and 47 participants for the between-task correlational analyses (40 women; $M_{age} = 19.74$; $SD_{age} = 1.59$).

Apparatus

Participants were seated alone in front of a monitor and a computer, using a standard computer keyboard and, if necessary, a headset (Audio Technica ATH-M20x). The display measured 21.5 inches with a refresh rate of 60 Hz. Stimuli were presented using MATLAB 2022b (The MathWorks Inc., 2022) with Psychophysics Toolbox (Brainard, 1997; v3.0.19.4)

Stimuli, design, and procedure

Forced-response task

In this task, participants responded to the presentation of one of four stimuli (letters of the Phoenician alphabet) displayed in the center of the screen by pressing a specific key. To respond, they positioned one finger of their dominant hand over each of the keys "H," "U," "I," and "L" of the computer keyboard. After each response, a feedback box corresponding to the pressed key appeared on the screen, turning green for correct responses and red for incorrect ones. The assignment of specific keys to each stimulus was counterbalanced across participants.

All participants completed two training conditions: minimal and 4-day. The experiment followed a counterbalanced crossover design, with 10 days of separation between the two conditions. Thus, half of the participants completed the minimal condition before the 4-day condition, while for the other half, this order was reversed. Due to this within-subject training manipulation, the set of stimuli (Phoenician letters) used differed between the two conditions.

The forced-response task included three types of blocks: training, criterion, and forced-response test. In the 4-day condition, participants completed 10 daily training blocks for 4 consecutive days. On the fifth day, they completed the criterion blocks, followed by five forced-response test

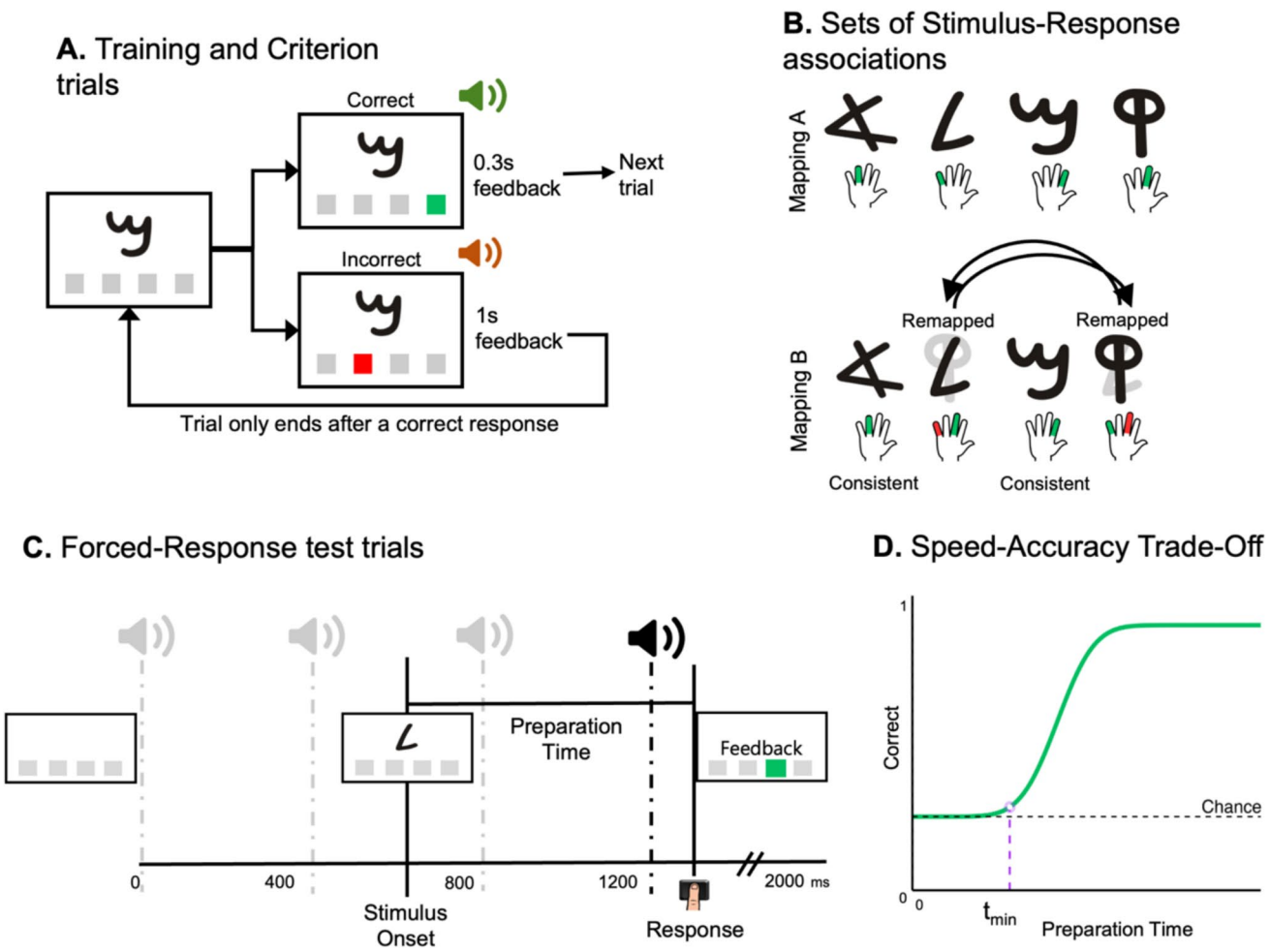

Fig. 1 The forced-response task. *Note*. **A** Schematic representation of the trial structure for training and criterion trials. In training trials, participants were instructed to respond as quickly as possible, while in criterion trials, they were required to focus on accuracy. **B** Example of the two sets of S–R associations: mapping A and mapping B. Each stimulus had one correct response (green-colored finger) associated. Mapping B was a modified version of mapping A in which two stimuli were remapped, switching their associated responses, while the other two stimuli remained consistent with mapping A. Responses to the remapped stimuli based on mapping A were classified as habitual errors (red-colored finger), whereas correct (adaptive) responses based on the updated mapping B were considered goal-directed (green-colored finger). **C** Schematic representation of the trial structure for forced-response test blocks. Stimulus onset varied randomly from trial onset (0 ms) to the onset of the fourth tone (1,200 ms). Participants were instructed to respond precisely at the onset of the fourth tone, regardless of when the stimulus appeared. Preparation time was defined as the RT of the actual response minus the stimulus onset. **D** Expected function for the probability of a correct response in forced-response trials. The chance level was set at 0.25, as there were four response options. t_{min} refers to the point of preparation time at which the function reaches 5% of its height

blocks. Participants in the minimal condition completed all the above blocks in a single day except the training phase (i.e., criterion and forced-response test blocks). In all blocks, the stimuli were presented pseudo-randomly,[1] ensuring that each stimulus appeared five times in subblocks of 20 trials, with no stimulus appearing more than twice consecutively.

Training blocks consisted of 100 free-RT trials, during which participants were instructed to respond as quickly as possible. Each trial began with a tone signaling the start of the trial, followed by the presentation of a stimulus. If the response was correct, a pleasant tone played, and after a 300-ms pause, the next trial began. If the response was incorrect, a buzzer sounded, and participants were unable to respond for 1 s. The trial continued until the correct response was given (see Fig. 1A). At the end of each block, participants received feedback on their completion time and were shown

[1] Due to a programming issue, the pseudo-random presentation order was the same for all participants during the first four sessions of the experiment. This issue was corrected before the session in which participants completed the forced-response test under overtrained conditions.

a comparison with their fastest block. They were encouraged to improve their speed and surpass their best performance in each subsequent block.

Criterion blocks were designed to assess participants' knowledge of specific S–R associations. Unlike training blocks, participants were informed that there were no time constraints, and their primary goal was to respond accurately to each stimulus. All other aspects of the procedure remained the same as in the training blocks. A criterion block ended once participants achieved five consecutive correct responses for each stimulus.

Participants first completed a criterion block with a set of four S–R associations, referred to as mapping A, which had been previously trained in the 4-day condition. After successfully meeting the criterion for mapping A, participants moved to a new criterion block. In this block, two of the four S–R associations from mapping A had their correct responses switched (remapped S–R associations), resulting in a (partially) new mapping, mapping B. Specifically, in mapping B, two associations remained unchanged (consistent condition), while the other two associations were remapped, meaning their correct responses differed from those in mapping A (remapped condition; Fig. 1B). These remapped stimuli were relevant for assessing whether the behavior was habitual or goal-directed.

The forced-response method aims to establish a continuum of allowable RTs by varying the stimulus onset on a trial-by-trial basis and instructing participants to always respond at a fixed time point (Haith et al., 2016). This task consisted of five blocks of 100 trials, during which participants heard four tones separated by 400 ms. They were instructed to respond at the onset of the fourth tone (1.2 s after trial start) based on the stimulus–response associations from mapping B. A key feature of this method was that the stimulus appeared at a random point within the tone sequence, meaning the imposed response preparation time—the interval between stimulus onset and the fourth tone—varied across trials. In trials where participants perceived there was insufficient time to process the stimuli (presumably, stimulus onsets 300 ms before the fourth tone), they were instructed to guess the response. As in the original study, each trial lasted 2 s, allowing for responses after the fourth tone. Unlike previous block types, in which trials lasted until participants responded correctly, only one response was permitted per trial during forced-response blocks. The actual preparation time for a response was defined as the RT minus the stimulus onset (see Fig. 1C). When participants' responses were poorly synchronized with the fourth tone, feedback was provided. If participants responded 100 ms before the fourth tone, the message "Too fast" was displayed in the center of the screen; if they responded 100 ms after the onset of the fourth tone, the message "Too slow" was displayed. Correct and incorrect responses were also visually indicated by coloring the corresponding feedback box during the trial.

During forced-response blocks, participants had to respond according to the (newly learned) mapping B. Thus, responses that followed the previous mapping A were classified as habitual errors, as they indicated reliance on prior learning. In contrast, correct responses to the remapped stimuli were considered goal-directed, reflecting successful adaptation to the new mapping. The two remaining types of responses were categorized as random errors.

Notably, in each training condition, before starting the first session, and also the last session of the overtrained condition, participants completed a familiarization task. This task included two training blocks and two forced-response test blocks using nonarbitrary stimuli (pictures of colored fingers).

Outcome-devaluation task

One month after completing the forced-response task, participants returned to the laboratory for the outcome-devaluation task. This task also included two training conditions, minimal and 4-day, presented in the reverse order of the forced-response task (except for three participants who were unable to attend the laboratory in this specific order) and spaced approximately 10 days apart. Different stimuli were used for each condition. In this task, participants took on the role of space traders, with the goal of earning as many points as possible by collecting diamonds. In each trial, they had to choose between two aliens to exchange a cookie for diamonds. Among the four possible cookies, two (S_{low}) were exchangeable for a low-value diamond (O^{10}, which is worth + 10 points), and the other two (S_{high}) could be traded for a high-value diamond (O^{100}, worth + 100 points). If participants correctly selected the alien that preferred the offered cookie, they received the corresponding optimal diamond reward (O^{10} or O^{100}). However, if they chose an alien that did not prefer that cookie, they received a less valuable diamond (O^{5}, worth only + 5 points). Thus, cookies functioned as discriminative cues that informed participants about the optimal response in each trial.

The task consisted of reward learning (training) and outcome-devaluation blocks (Fig. 2). Reward learning and outcome-devaluation blocks were composed of two types of trials: instrumental reward-learning trials and consumption trials.

Each trial began with a central fixation point (red) displayed for a random duration of 200 to 300 ms, followed by the presentation of a cookie stimulus at the center of the screen. Immediately afterward, two alien images appeared on either side of the screen, representing the two response options. These aliens were always displayed in the same position, with their placement counterbalanced across

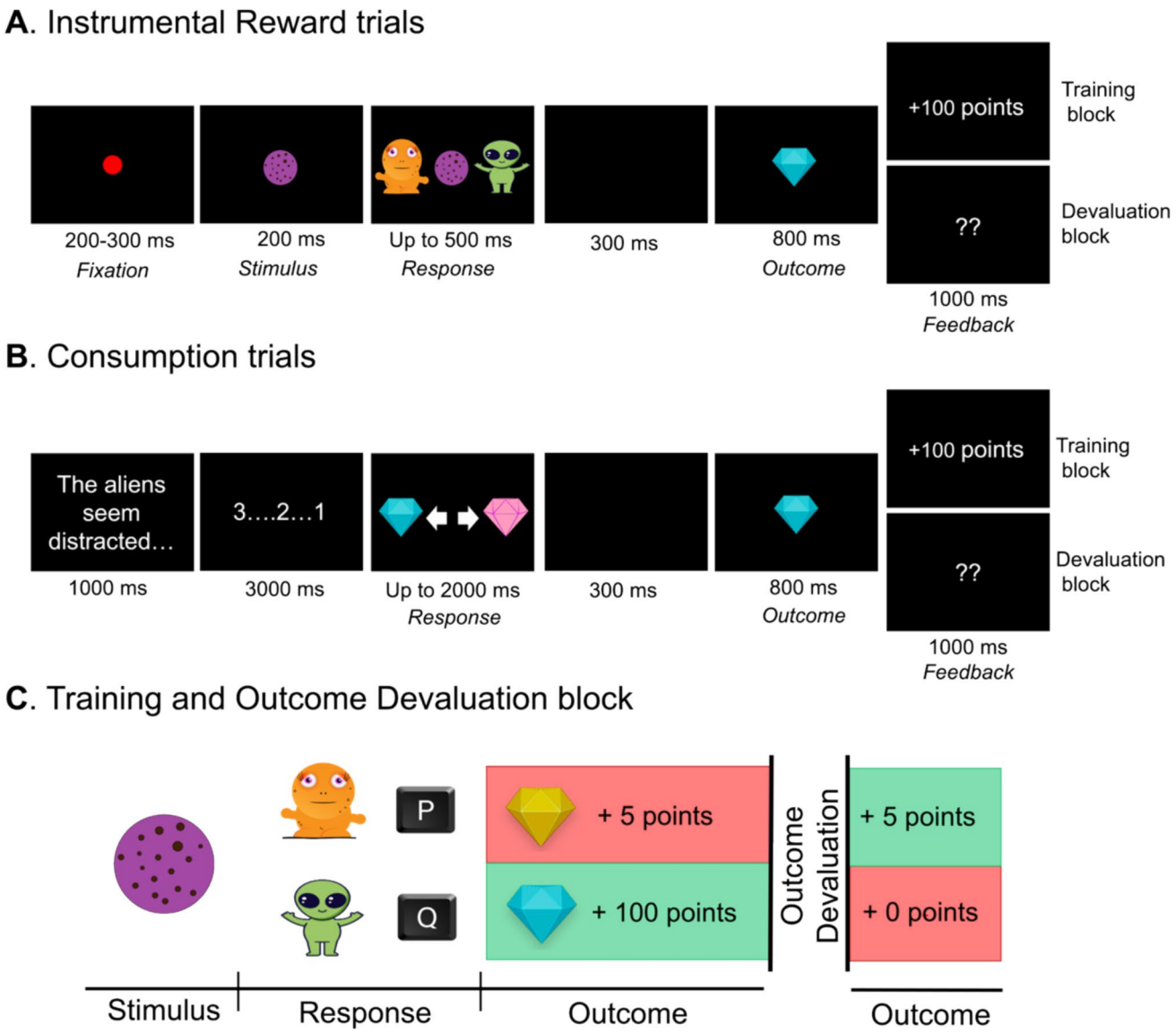

Fig. 2 The outcome-devaluation task. *Note*. **A** Example of an instrumental reward trial. Participants learned that each cookie (stimulus) indicated the optimal response for that trial (left or right, corresponding to the "Q" or "P" key). After responding, participants received the associated outcome, consisting of diamonds with different point values. During training blocks, the outcome included feedback on the number of points earned. However, during devaluation blocks, this information was hidden (nominal extinction). **B** Example of a consumption trial. Participants selected between two outcomes with different values (points), allowing assessment of their knowledge of the current outcome values. **C** Schema of the stimulus–response-outcome associations in each block. The example illustrates a high-value stimulus (+100), but the same logic applies to low-value stimuli (+10). The optimal response is shown in green, while the nonoptimal response is in red (+5 or +0). In contrast to training blocks, the goal-directed response in devaluation blocks is to switch responses and select the +5 diamond instead of the now-devalued outcome

participants. Participants pressed the "Q" key to give the cookie to the alien on the left or the "P" key to give it to the alien on the right. They had a 500-ms response window to make their choice. After responding, the screen went blank for 300 ms. If participants responded after the time limit, the message "Time out, please respond faster" was displayed, whereas if they responded before the aliens appeared, the message "Too early! No diamond for you!" was displayed. For on-time responses, one of the three diamond images appeared for 800 ms, followed by a feedback screen showing the points earned for that trial. In devaluation trials, the number of points was not displayed.

To assess participants' knowledge of the current outcome values, special consumption trials were included. Participants were informed that, occasionally, the aliens would be distracted, allowing them to obtain a diamond

without trading cookies. During these trials, the message "The aliens seem distracted…" was displayed for 1 s, followed by a 3-s countdown. After the countdown, O^{100} and O^{10} diamonds appeared on the left and right of the screen, with their positions randomly assigned for each trial. Participants pressed "Q" or "P" to select the left or right diamond. Responses made before the diamonds appeared or slower than 2 s resulted in "Too fast" or "Time out" messages. If the response was on time, the feedback was displayed as in regular trials.

At the beginning of each outcome-devaluation block, participants were informed that one diamond—O^{10} or O^{100}—had been devalued and would no longer earn points. The diamond worth five points was never devalued. In outcome-devaluation blocks, for half of the stimuli, the habitual response led to zero points, making the optimal response to switch and at least obtain the five-point diamond (see Fig. 2C). Although the feedback screen no longer displayed the number of points earned per trial, participants were explicitly informed at the start of the block that they would still receive the corresponding points.

Each block contained 52 trials, with each stimulus presented randomly 12 times. Consumption trials appeared at trials 13, 26, 39, and 52 in each block. In the 4-day condition, participants completed nine training blocks per day over four consecutive days. On the fifth day of the 4-day condition, participants completed five training blocks followed by two outcome-devaluation blocks, each one with a different devalued outcome (i.e., O^{10} or O^{100}), with the devaluation-block order randomly assigned. Following the devaluation blocks, participants completed two new training blocks identical to the first. The single session of the 1-day minimal condition followed the same structure. The cookie stimulus was presented for 800 ms in the first block of each session, 500 ms in the second block, and 200 ms in subsequent blocks to speed up the learning of S–R contingencies.

Finally, we illustrate the general design of the experiment in Fig. 3.

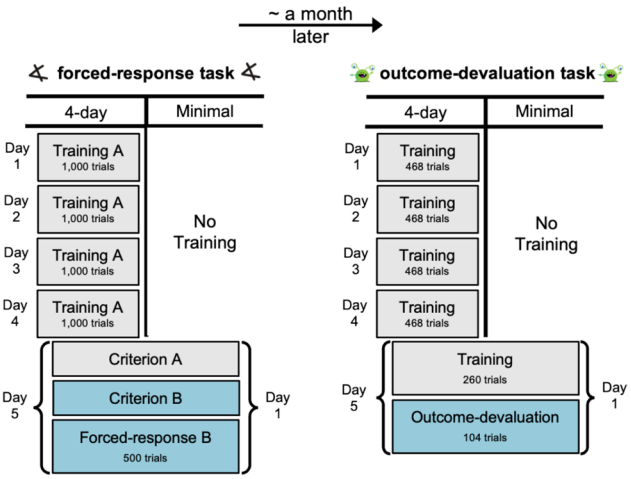

Fig. 3 General design of the experiment. *Note*. In each task, all participants completed both training conditions (i.e., 4-day and minimal). Training conditions were spaced by approximately 10 days. In the forced-response task, training conditions were presented counterbalanced across participants, whereas in the outcome-devaluation task, participants completed both conditions in the inverse order of the forced-response task. Colored boxes represent task blocks. In the forced-response task, letters "A" and "B" refer to the different stimulus–response mapping (i.e., mapping A or mapping B) that was used during the corresponding block

Data analysis

Statistical analyses and graphical visualizations were conducted using R (R Core Team, 2023; v4.3.0).

Forced-response task

Trials with timeouts were omitted (1.39%). No exclusion criteria were preregistered; however, 24 participants were excluded because they had no observations in at least one preparation time window of analysis. These preparation time windows of analysis were determined as in Hardwick et al.'s (2019) study: Like them, we fitted a cumulative Gaussian distribution to the speed–accuracy trade-off of consistent trials (trials that did not change in mapping B). This approach allowed us to determine the minimum preparation time that each participant needed to generate nonrandom responses (see Fig. 1D). This "t_{min}" was defined as the point at which the speed–accuracy trade-off function reached 5% of its height, hence responses started to be nonrandom—that is, the preparation time at which each participant began to respond above chance level. Note that 5% is an arbitrary threshold chosen by the original authors. The authors then used this t_{min} to split response data into three preparation time windows based on this reference point: t_{min} – 300 ms to t_{min}; t_{min} to t_{min} + 300 ms; t_{min} + 300 ms to t_{min} + 600 ms. The rationale for this non preregistered exclusion criterion is straightforward: We cannot conduct the preregistered repeated-measures (RM) ANOVA on the proportion of habitual errors across these three distinct preparation time windows while including participants with no observations in one of the time windows of analysis. We preregistered the same analysis using both RM ANOVA and mixed-model approaches. Values of t_{min} were computed for each participant using MATLAB 2022b and the original scripts provided in Hardwick et al. (2019). Although participants with no observations in at least one preparation time window of analysis could have been included in the mixed-model analysis (and other analysis), we excluded them from all the subsequent tests because they were clearly not following the

instructions of the forced-response task—i.e., they systematically avoided responding at specific times (usually very short preparation times).

Importantly, we have included in Supplemental Material A2 the analysis with these excluded participants (when their inclusion was possible).

Mixed models were implemented using the {lme4} R package (Bates et al., 2015) to handle unbalanced data and account for between-subject variability. For all the models fitted in this study, we coded categorical variables by deviation coding, and, following Barr et al. (2013), we specified the random-effects structure for participants and selected the model with the maximal feasible random-effects structure that did not lead to convergence issues (Matuschek et al., 2017). Regarding inference, we conducted Wald chi-squared tests to assess the significance of fixed effects using an alpha level of 5%. In addition, for hypothesis testing, we evaluated the significance of the fixed effects by comparing the maximal fixed-effects structure model against a reduced model (hierarchical test). This comparison was conducted using the likelihood ratio test (LRT) and Akaike information criterion (AIC). We report only the final selected models, while the full analysis workflow is available in the OSF project (https://osf.io/4s3c8/). After determining the final model, we used the {marginaleffects} R package (Arel-Bundock et al., 2024) to compute marginal means comparisons.

To facilitate comparisons between this study and Hardwick et al. (2019), we preregistered the RM ANOVA of the original study, and also included its translation to a mixed-model approach. We constructed a generalized linear mixed model (GLMM) to analyze habit errors (modeled with a binomial likelihood) as the dependent variable. The model included preparation time window, training condition, and their interaction as fixed effects. The random-effects structure included the main effects of the predictors. Given that the RM ANOVA and GLMM analyses produced similar outcomes and considering that the mixed-model approach provides a more robust framework for analyzing hierarchical data structures, we opted to present the RM ANOVA results in Supplemental Material A1. Following the notation of the {lme4} R package (Bates et al., 2015), our maximal model's formula was

$$\begin{aligned} \textit{habit error} \sim\ & \textit{training condition} * \textit{preparation time window} \\ & + (\textit{training condition} + \textit{preparation time window} | \textit{participant}) \end{aligned}$$

Following Hardwick et al. (2019), preparation time is expected to have a nonlinear effect on both habitual errors and correct responses. To examine this nonlinear relationship, we employed generalized additive mixed models (GAMM) using the {mgcv} R Package (Wood, 2023), which extends GLMMs by allowing predictors to be modeled as smooth functions rather than assuming linear relationships (Wood, 2017). To construct our target model, we selected from a range of models with different random-effects structures, starting with the most complex model (which included random smooths for each participant), to the simplest, which included only random intercepts for each participant. Model selection was based on the AIC. For habit errors, we fitted a GAMM with separate preparation time smooth functions for each training condition. Similarly, for correct responses, we constructed a GAMM incorporating separate smooth functions of preparation time for each training condition, along with a common smooth function for each stimulus type. Both selected models included the most complex random-effects structure.

We expected that overtrained (habitual) responses would be automatically activated in remapped trials even when participants correctly changed their responses. This activated S–R would slow down RTs for correct (goal-directed) responses, with the effect being most pronounced under high time pressure conditions (Luque et al., 2020). RTs were defined as the time elapsed between the start of the trial and the participant's response. Because participants were instructed to respond to the fourth tone, which occurred 1.2 s after the trial onset, RTs around this time point were expected. We filtered the data to include only correct responses and preparation times greater than each participant's t_{min}, thereby excluding very short responses that were most likely produced at random, and log-transformed RTs to normalize their distribution. We then fitted an LMM with log-transformed RTs as the dependent variable, including the three-way interaction of stimulus onset, stimulus type, and training condition as predictors. We included stimulus onset as a predictor because it indicates the time pressure in each trial. For example, since participants were instructed to always respond at the onset of the fourth tone (1.2 s after trial start), a trial where the stimulus appeared at 800 ms after trial start imposed greater time pressure than a trial where the stimulus appeared at 300 ms. To facilitate model convergence, the stimulus onset predictor was centered and scaled. The random-effects structure of this model included all main effects and interactions of the fixed-effects structure. The lme4 formula (Bates et al., 2015) of the maximal model was

$$\begin{aligned} \log(RT) \sim\ & \textit{training condition} * \textit{stimulus type} * \textit{stimulus onset} \\ & + (\textit{training condition} * \textit{stimulus type} * \textit{stimulus onset} | \textit{participant}) \end{aligned}$$

Outcome-devaluation task

We excluded trials with anticipations, that is, responses made before the response options appeared (0.18%) and timeouts (9.64%). Following Luque et al. (2020), we collapsed data from stimulus pairs associated with the same outcome value. As in the forced-response task, to facilitate comparisons between studies, we preregistered the analyses of the main dependent variables using both an RM ANOVA and a mixed-model approach. However, since both approaches produced

similar results, we reported RM ANOVAs in Supplemental Material A1. The mixed-model selection followed the same strategy as in the forced-response task.

We developed a GLMM to analyze response selection using binomial likelihood. The model included a three-way interaction between training condition, stimulus value, and block type (i.e., baseline and devalued blocks) as fixed effects. For the baseline, we used data from trials with the same stimulus (S_{high} or S_{low}) when participants made the "standard" response in the preceding training block (Luque et al., 2020). The maximal model included all main effects and two-way interactions of the fixed structure as random effects. The lme4 formula (Bates et al., 2015) for the maximal model was

$$\begin{aligned} &\textit{response selection} \sim \textit{training condition} \\ &\quad * \textit{stimulus value} * \textit{block type} \\ &\quad + \left(\begin{array}{l} \textit{training condition} * \textit{stimulus value} + \\ \quad \textit{training condition} * \textit{block type} + \\ \quad\quad \textit{stimulus value} * \textit{block type} \end{array} \middle| \textit{participant} \right) \end{aligned}$$

Similarly, we constructed an LMM for log-transformed RTs on trials where participants chose the highest available outcome. This model also included a three-way interaction between the training condition, stimulus value, and block type. The main effects and interactions were incorporated as random effects in the model. The lme4 formula (Bates et al., 2015) was as follows:

$$\begin{aligned} &\log(RT) \sim \textit{training condition} * \textit{stimulus value} * \textit{block type} \\ &\quad + (\textit{training condition} * \textit{stimulus value} * \textit{block type} | \textit{participant}) \end{aligned}$$

Reliability

We evaluated the internal consistency of each measure by calculating its split-half reliability using a permutation-based approach. In this method, all trials are randomly split into two halves, and a Pearson's r correlation is computed. We repeated this process 10,000 times, generating a distribution of split-half correlations that are then corrected using the Spearman–Brown formula to account for the splitting effect (Spearman, 1910). The mean of the coefficients was taken as the final estimate of reliability, and the 2.5th and 97.5th quantiles as the 95% bootstrapped CI. We used a modified version of the *split-half* R function (Garre-Frutos et al., 2024; Parsons, 2021).

Convergent validity

To evaluate convergent validity across measures from both tasks, we analyzed the correlations among the measures obtained from the 4-day training condition, where habitual behavior was expected to emerge. As preregistered, we only assessed convergent validity among measures that demonstrated sensitivity to the amount of training and were thus consistent with the expected characteristics of the "habit" construct. The preregistration plan included only "between-task" correlations; however, in an exploratory analysis, we also examined correlations between measures from the same task, provided they were sensitive to the amount of training. As noted earlier, we applied non-preregistered exclusion criteria in the forced-response task. However, for the sake of transparency, we report the convergent validity analyses including the maximum available sample in Supplemental Material A2.

Before conducting the analyses, we tested whether each measure followed a normal distribution using a quantile–quantile plot (Das & Imon, 2016). If the data were normally distributed, we used Pearson's r for correlation analyses; otherwise, we applied Kendall's tau (τ_b) as a non-parametric alternative. One-sided contrasts were used for preregistered tests. To complement the frequentist hypothesis testing, we reported the Bayes factor (BF) in cases where the absence of a relationship was key to our conclusions. This Bayesian analysis was performed using JASP (JASP Team, 2024; v0.19.2) with default prior settings. For each analysis, a Bayes factor robustness check is provided in Supplemental Material A4.

Results

Forced-response task

Training and criterion

Participants demonstrated improved performance, as RTs for correct responses ($t(49) = 9$, $p < .001$, $d_z = 1.29$, 95% CI [0.9, 1.66]) and errors ($t(49) = 7.89$, $p < .001$, $d_z = 1.13$, 95% CI [0.77, 1.48]) decreased significantly between the first and last training session (Fig. 4A). Practicing mapping A did not result in problems in learning the revised set of associations (mapping B), with no significant differences between training conditions in the number of trials needed to complete the corresponding criterion block ($t(49) = -0.2$, $p = .839$, $d_z = -0.03$, 95% CI [−0.31, 0.25]; Fig. 4B).

Forced-response test

Habit errors Our pattern of results replicates the one observed in Hardwick et al. (2019), observing a time-dependent pattern for habit expression. A GLMM on habit errors with preparation time window and training condition as predictors revealed a significant main effect of preparation time window, $\chi^2(2) = 252.57$, $p < .001$, a significant main effect of training condition, $\chi^2(1) = 94.5$, $p < .001$, and a significant interaction between preparation time window and

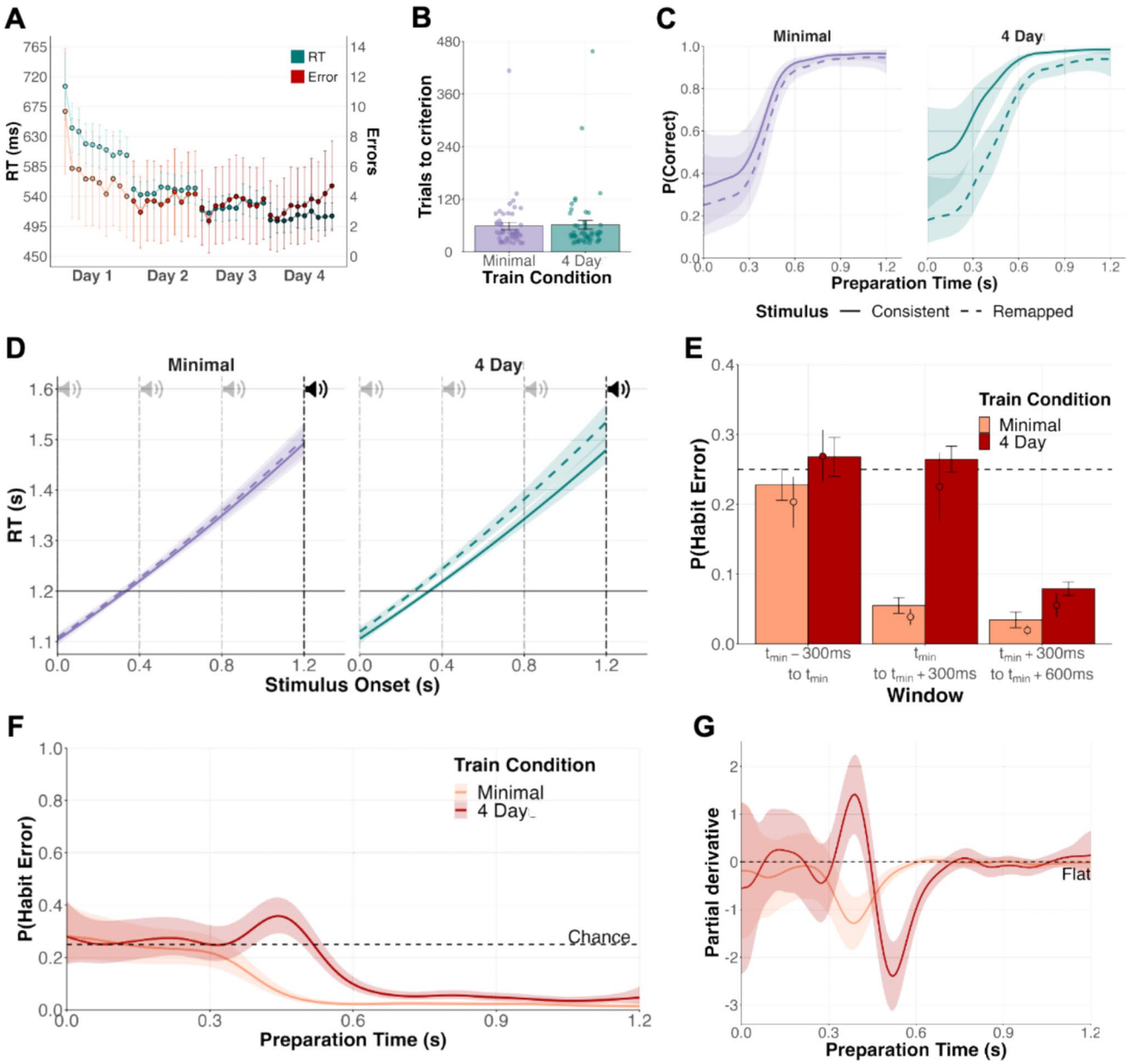

Fig. 4 Main results from the forced-response task. *Note*. **A** Averaged RT for correct responses and mean number of errors from the training blocks within the 4-day condition. Each dot corresponds to a training block, while error bars represent the standard error of the mean (SEM). **B** Number of trials required to complete the criterion block for mapping B (i.e., achieving five consecutive correct responses for each stimulus). Bars depict the mean number of trials across participants, with error bars indicating the SEM. Individual participant data are represented as dots. **C** GAMM predictions for correct responses across different types of stimuli, training conditions, and preparation times. Shaded areas indicate the 95% confidence interval (CI). **D** LMM predictions for goal-directed response RTs across stimulus types, training conditions, and stimulus onsets (i.e., excluding preparation times below t_{min}). Shaded areas represent the 95% CI. The dashed lines illustrate the temporal moment in which each tone was displayed during the trial, whereas the solid line represents the imposed RT. **E** Habit errors across preparation time windows for each training condition. Bars depict participants' mean proportion of habit errors, while error bars indicate the SEM. Dots show GLMM predictions, with error bars depicting the 95% CI. **F** GAMM predictions for habit errors across training conditions and preparation times. Shaded areas denote the 95% CI. **G** The partial derivative of the GAMM predictions for habit errors across training conditions and preparation times. Positive values indicate positive slopes (increasing likelihood of habit errors), negative values indicate negative slopes (decreasing likelihood of habit errors), and values near zero indicate a flattening of the function. Shaded areas represent the 95% CI

training condition, $\chi^2(2) = 138.72$, $p < .001$. Marginal means comparisons between training conditions showed no significant effect in the earliest preparation time window, $\beta = 0.04$, $z = 1.42$, $p = .155$, and a significant effect in the time window from t_{min} to $t_{min} + 300$ ms, $\beta = 0.19$, $z = 9.27$, $p < .001$, and $t_{min} + 300$ ms to $t_{min} + 600$ ms, $\beta = 0.05$, $z = 4.93$, $p < .001$ (see Fig. 4E).

In a non-preregistered analysis, we treated preparation time as a continuous variable rather than grouping it into time windows, and observed a nonlinear effect on habitual responses. Figure 4F illustrates GAMM predictions, which align closely with the findings of Hardwick et al. (2019), who observed that habitual errors appeared between 300 and 600 ms in the overtrained condition. Similarly, we found that habitual errors exceeded chance levels at preparation times ranging from 340 to 520 ms. Additionally, we observed an asymptotic increase in the probability of committing a habitual response in the 4-day condition compared to the minimal condition, particularly at high preparation times. To further explore the interaction between habitual and goal-directed processes, we conducted an additional, non-preregistered analysis, examining the change in slope (i.e., the partial derivative of the probability of committing a habitual error) as a function of preparation time. We observed that the likelihood of responding habitually increased at preparation times of 320 ms but decreased sharply at preparation times exceeding 450 ms, reaching an asymptote at 730 ms (Fig. 4G).

Correct responses We explored the nonlinear relationship between correct responses and preparation time. GAMM estimates show that after overtraining, participants required more preparation time to correctly respond to remapped stimuli relative to consistent stimuli. In contrast, in the minimal condition, where habit formation was presumably not established, there were no differences between stimulus types. This finding suggests that overtrained incompatible habits interfere with adaptation to novel conditions (i.e., remapped stimuli). However, it is important to note that the model exhibited high uncertainty at short preparation times, regardless of condition or stimulus type (Fig. 4C). This uncertainty is partly due to the small number of trials in which participants responded with short preparation times, highlighting a limitation of the task, which we will discuss further below.

Goal-directed reaction times We preregistered the hypothesis that RTs for goal-directed responses (i.e., nonrandom correct responses to remapped associations) would slow down after overtraining and with later stimulus onsets. Examining correct response RTs with an LMM, we found that, as expected, participants did not perfectly synchronize their responses with the tone that forced them to respond. While the model rendered the effect of training condition nonsignificant, $\chi^2(1) = 1.85$, $p = .174$, it revealed a significant effect of stimulus onset, $\chi^2(1) = 602.85$, $p < .001$ and stimulus type, $\chi^2(1) = 30.96$, $p < .001$, and a significant interaction between them, $\chi^2(1) = 5.22$, $p = .022$. More importantly, we found a significant training condition × stimulus type interaction, $\chi^2(1) = 31.18$, $p < .001$, and a significant three-way interaction with stimulus onset, $\chi^2(1) = 5.84$, $p = .016$. This pattern is explained by long synchronization delays for remapped stimuli, particularly after extensive training and when participants had less time to respond. These results highlight the difficulty of adapting to new associations when an incompatible habit has been formed and rapid responses are required, as illustrated in Fig. 4D. Marginal mean back-log-transformed RT comparisons between training conditions further supported this conclusion (consistent: $\beta = -3.7$, $z = -1.07$, $p = .284$; remapped: $\beta = 19.91$, $z = 4.35$, $p < .001$). These findings suggest that even when participants responded in a goal-directed manner, habitual responses were still activated and interfered with performance.

Outcome-devaluation task

Consumption trials

We conducted a non-preregistered analysis to examine whether accuracy (the proportion of optimal responses) differed across training conditions. A separate Friedman test revealed significant differences in accuracy between training conditions, $\chi^2(1) = 7.37$, $p = .007$, with more errors observed in the 4-day condition. Nine participants scored accuracy of 0.5 or less (i.e., they selected the devalued outcome in at least four of the eight consumption trials). Of these, two participants were from the minimal condition. Since it is crucial that participants in both groups have sufficient and comparable knowledge of the ongoing outcome values during devaluation, we conducted all main analyses again, this time excluding these nine participants. The results of these analyses are provided in Supplemental Material A3. However, since we found no differences in the results whether these participants were included or not, we report in the main text the preregistered analysis, in which consumption trials were not used to exclude participants, consistent with Luque et al. (2020).

Outcome devaluation

Response selection The GLMM yielded a significant effect of block type, $\chi^2(1) = 602.37$, $p < .001$, a significant block type × stimulus interaction, $\chi^2(1) = 26.67$, $p < .001$, and a significant block type × training condition interaction, $\chi^2(2) = 5.17$, $p = .023$. Marginal means comparisons

indicated that participants were less likely to select an outcome after its devaluation, $\beta = 0.6$, $z = 33.1$, $p < .001$. This effect was larger for S_{high}, $\beta = 0.66$, $z = 36.4$, $p < .001$ than for S_{low}, $\beta = 0.54$, $z = 23$, $p < .001$, explaining the block type × stimulus interaction. The block type × training condition interaction was driven by a larger devaluation (block type) effect in the 4-day condition, $\beta = 0.62$, $z = 32.1$, $p < .001$, compared to the minimal condition, $\beta = 0.57$, $z = 24.3$, $p < .001$.

As expected, and as preregistered, neither overtraining nor high-value outcomes resulted in greater persistence in response selection when the outcomes were devalued, meaning that they did not lead to stronger habitual behavior (Fig. 5A). This result is consistent with the findings from the original study (Luque et al., 2020, see also Nebe et al., 2024).

Goal-directed reaction times The LMM revealed a significant main effect of training condition, $\chi^2(1) = 12.12$, $p < .001$, block type, $\chi^2(1) = 83.61$, $p < .001$, and stimulus value, $\chi^2(1) = 21.69$, $p < .001$, a significant block type × stimulus value interaction, $\chi^2(1) = 40.4$, $p < .001$, and importantly, a significant training condition × block type interaction, $\chi^2(1) = 12.26$, $p < .001$. These interactions were further examined using marginal back-log-transformed means comparisons. These analyses revealed that the block type × stimulus value interaction was due to faster RTs at baseline compared to devaluation blocks, which was more pronounced for S_{high}, $\beta = -50.97$, $z = -10.32$, $p < .001$, than for S_{low}, $\beta = -13.43$, $z = -1.56$, $p = .118$. Reflecting an RT switch cost, the training condition × block type interaction indicated that RTs were faster in baseline after overtraining, $\beta = 28.61$, $z = 4.77$, $p < .001$, but not slower in devaluation blocks, $\beta = 1.54$, $z = 0.19$, $p = .849$ (see Fig. 5B). On the one hand, as preregistered, our pattern of results mirrors that found in Luque et al. (2020), who observed a greater RT switch cost after overtraining (4-day condition) compared to the minimal condition and for S_{high} (compared to the slow condition). On the other hand, contrary to the original study, we observed that the RT switch cost was driven by differences in baseline, while participants maintained similar RTs in devaluation blocks of both training conditions. This result is discussed in detail in the General discussion.

Convergent validity

The measures that demonstrated sensitivity to overtraining, as expected, were habit responses (forced-response task) and the RT switch cost (forced-response and outcome-devaluation tasks). As preregistered, for the following analyses, habit responses in the forced-response task were quantified as the proportion of habit errors in the preparation time window from t_{min} to $t_{min} + 300$ ms, as this window exhibited the strongest training effect. RT switch cost (forced-response task) was calculated as the RT for correct responses

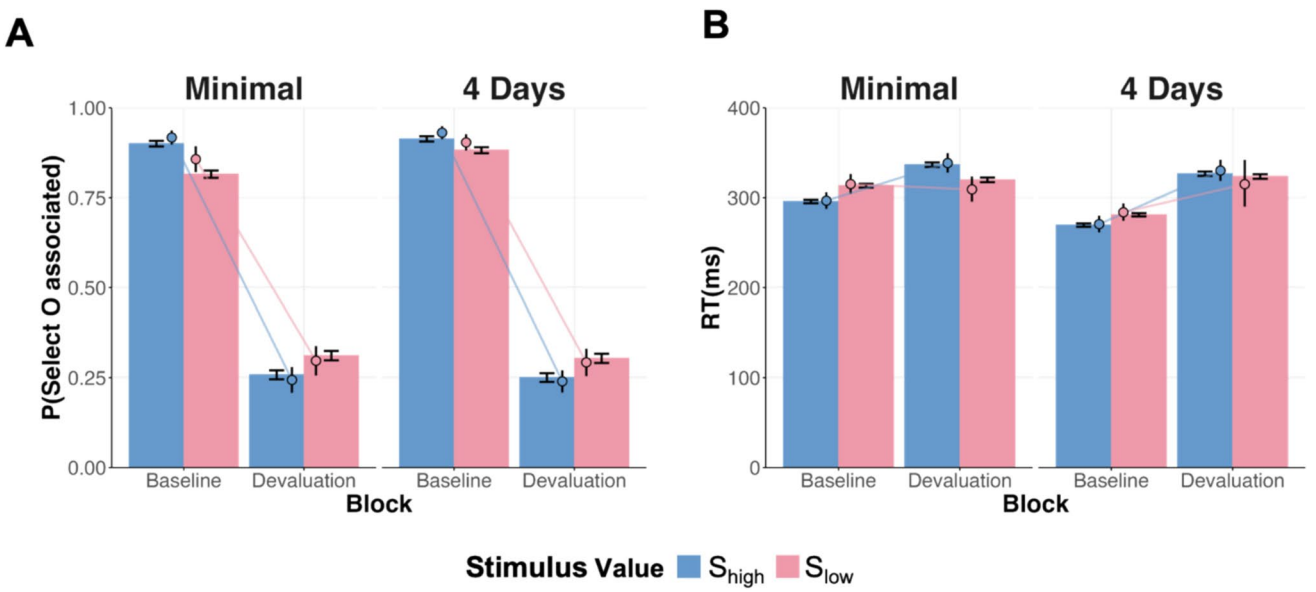

Fig. 5 Main results from the outcome-devaluation task. *Note*. **A** Response selection in each block type and training condition. Bars depict participants' mean proportion of responses associated with the stimulus presented, while error bars indicate the SEM. Dots show GLMM predictions, with error bars depicting the 95% CI. **B** Goal-directed response RTs in each block and training condition. Bars show mean RTs for responses associated with the most valuable outcome available, and error bars indicate the SEM. Dots show GLMM predictions, and error bars depict the 95% CI

to remapped minus RT for correct responses to consistent stimuli, filtering for preparation times above t_{min} to exclude responses with very short preparation times that are likely to be random. RT switch cost (outcome-devaluation task) was computed following Luque et al. (2020) by subtracting the RT of selecting an outcome when it was valued (baseline) from the RT of switching that response when the usual outcome was devalued, provided that participants correctly selected a still-valuable outcome; higher positive scores are indicative of a greater interference effect. In the current data, the stimulus value (whether the trained outcome was + 100 or + 10) was irrelevant for habit formation, as it did not interact with the training condition. Thus, we collapsed this factor in the current analyses.

Correlational analyses were conducted for between-task measures and within the forced-response task, using the maximum available sample for each analysis ($n = 47$ and $n = 50$, respectively). Descriptive, non-preregistered analyses were included to assess the reliability of these measures. For each score, we report the mean, standard deviation, Spearman–Brown reliability estimate (r_{sb}), and the 95% bootstrapped CI of the reliability. The results did not differ across samples. Habit responses in the forced-response task demonstrated high reliability, with a mean of 0.27 ($SD = 0.17$) and a Spearman–Brown reliability estimate of $r_{sb} = .83$, 95% CI [.66, .93]. Similarly, RT switch cost in the outcome-devaluation task showed strong reliability, with a mean of 0.05 ($SD = 0.04$) and $r_{sb} = .78$, 95% CI [.68, .85]. In contrast, RT switch cost in the forced-response task exhibited poor reliability, with a mean of 0.02 ($SD = 0.02$) and $r_{sb} = .49$, 95% CI [.23, .70].

Following our preregistration, we calculated Pearson correlations between measures from different tasks. The correlation between habit responses (forced-response task) and RT switch cost (outcome-devaluation task) was $r(45) = -.06$, $p = .648$, $BF_{0+} = 7.23$, while the correlation between RT switch cost in the outcome-devaluation task and the forced-response task was $r(45) = -.23$, $p = .943$, $BF_{0+} = 13.46$. Both tests were one-tailed, based on the preregistered expectation of a positive correlation. We also explored whether habit errors and RT switch cost—both from the forced-response task—were correlated. In this analysis, we did not find evidence for a relationship different from zero (two-tailed), $r(48) = .19$, $p = .179$, $BF_{01} = 2.36$ (see Fig. 6).

General discussion

Once a habit is acquired, it can interfere with our goals when achieving them requires actions that contradict the habit. To systematically study habit formation, experimental researchers attempt to induce habits in the laboratory by repeatedly pairing a specific response with a stimulus. These habits are then assessed using conflict tasks, which pit goal-directed behaviors against incompatible habitual responses. However, most current habit tests have failed to detect habit formation as a function of the amount of instrumental training (e.g., de Wit et al., 2018), despite this being a fundamental feature of the habit construct (Dickinson, 1985). Two promising protocols have demonstrated stronger evidence of habitualization in their overtrained conditions: the forced-response task (Hardwick et al., 2019) and the "aliens" outcome-devaluation task (Luque et al., 2020).

Although these protocols have produced results consistent with what is expected from a valid measure of habit strength, further validation is essential. Validation should be viewed as a continuous evidence accumulation process that ensures the credibility of score interpretations in research (Flake et al., 2022; Vazire et al., 2022). Measures designed to assess habit formation should not only be sensitive to overtraining effects but also reflect individual differences across participants, ensuring that they measure the same latent construct (convergent validity). Additionally, it is essential to assess reliability. Neglecting measurement reliability prevents researchers from determining the extent to which measurement error attenuates observed correlations and reduces statistical power (Parsons et al., 2019; Wiernik & Dahlke, 2020); it also limits the ability to compare effect sizes across populations and studies (e.g., Cooper et al., 2017). Thus, in this study, we aimed to evaluate the validity and reliability of laboratory protocols that are strong candidates for assessing the habit system, specifically those that have demonstrated sensitivity to overtraining manipulations (Hardwick et al., 2019; Luque et al., 2020).

We will summarize our preregistered objectives and the main results associated with them. Specifically, we preregistered the following aims. First, we aimed to replicate the main finding of Hardwick et al., (2019, Experiment 1), showing that habit errors are more frequent in an overtrained condition than in a minimal training condition. At the time of preregistration, this important study had not yet been replicated by any independent laboratory, and, to the best of our knowledge, the present study represents the first direct replication conducted by a laboratory other than the original one. It is worth noting that the effect has since been replicated by the original authors (Du & Haith, 2025). We have successfully replicated the main results from Hardwick et al. (2019, Experiment 1), supporting their conclusion that this task is suitable for studying habit commission under time pressure.

In the forced-response task, the second measure we investigated was the RT for goal-directed correct response switches, which has been shown to be sensitive to overtraining in the outcome-devaluation task (Luque et al., 2020; Nebe et al., 2024). Our second preregistered objective was to test whether RT for goal-directed responses is also sensitive

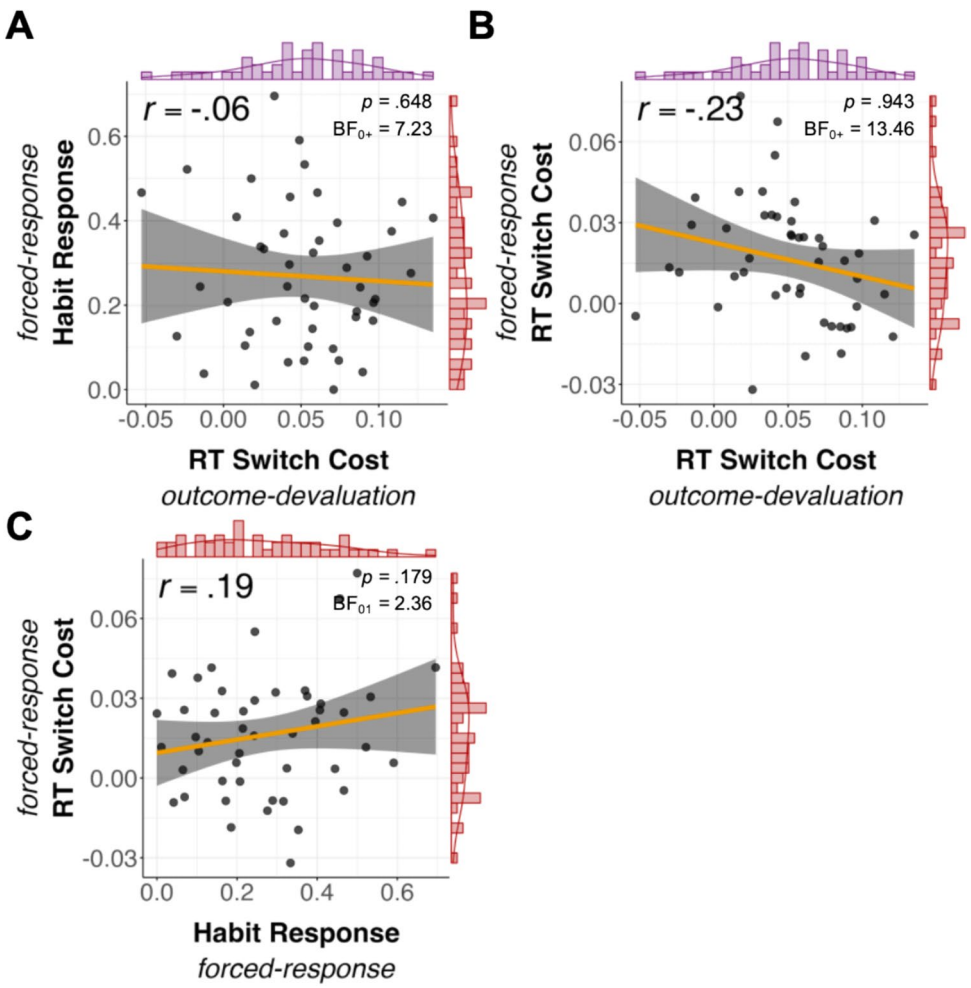

Fig. 6 Correlations across the habit measures. *Note*. **A** Scatterplot illustrating the correlation between the proportion of habit responses in the forced-response task and RT switch cost (in seconds) in the outcome-devaluation task ($n = 47$). **B** Scatterplot depicting the relationship between RT switch cost (in seconds) in the forced-response and outcome-devaluation tasks ($n = 47$). **C** Scatterplot showing the correlation between habit response and RT switch cost in the forced-response task ($n = 50$). In each figure, a marginal plot shows a histogram with a density plot overlapped for each measure

to the amount of training in this experimental paradigm. In other words, we sought to determine whether both relevant measures of habit formation could be obtained from a single paradigm, including habit errors (objective 1) and RT switch cost (objective 2). We successfully achieved this second objective, as our data revealed a significant slowdown in RTs for goal-directed responses in the 4-day condition compared to the minimal training condition. This finding confirms that the forced-response task, as implemented in the current study, provides evidence of progressive response habitualization through two dependent measures: habit errors and RT switch cost.

Our third preregistered objective was to advance our knowledge about the validity of measures obtained from the forced-response test (Hardwick et al., 2019) and outcome-devaluation test (Luque et al., 2020). However, as discussed previously, a proper validation assessment also requires evaluating measurement reliability. Consequently, we conducted non-preregistered, descriptive analyses to assess the reliability of measures that were found sensitive to overtraining. Specifically, we analyzed the number of habit responses produced with t_{min} to $t_{min} + 300$ ms preparation time in the forced-response test, the RT switch cost for correct goal-directed responses in the forced-response test, and the RT switch cost for correct goal-directed responses in the outcome-devaluation test. Our findings indicated that habit responses in the forced-response task and RT switch cost in the outcome-devaluation task demonstrated good reliability, whereas RT switch cost in the forced-response task showed poor reliability.

Continuing with the preregistered third objective, we analyzed the convergent validity of these measures. We initially

hypothesized that both RT-based measures would correlate; however, we did not find evidence that RT switch cost in the forced-response task positively correlates with any other measure. We also expected to observe a positive correlation between habit responses in the forced-response test and RT switch cost in the outcome-devaluation test, but our data supported the absence of a positive relationship.

Given these results, several questions arise that warrant discussion. One key question concerns why the habit response dependent variable was sensitive to the amount of training manipulation in the forced-response task but not in the outcome-devaluation task, despite both tasks imposing response time pressure during their tests. The forced-response test, which systematically varies the allowed response time on a trial-by-trial basis, is designed to identify which system predominates and controls responding at different preparation times. Consistent with Hardwick et al. (2019), our findings show that habitual responses emerge at short preparation times (~300 ms) but are rapidly overridden by goal-directed responses when sufficient time to respond is available (~500 ms). As a result, habit expression is overshadowed by goal-directed responses. In contrast, while our outcome-devaluation task also imposed time pressure for responding (which should encourage the expression of habitual responses), the available response time was fixed throughout the task (700 ms). Participants' actual RTs were around 500–550 ms after cue onset (Fig. 5; see also Luque et al., 2020, supplemental material). When we examined this same preparation time window in the forced-response test, we observed a steep decline in habitual errors (Fig. 4G), and by that point, habitual responses were nearly indistinguishable from chance (Fig. 4F). This suggests that participants in the outcome-devaluation test were adjusting their response times to the moment the goal-directed response became dominant over the habit. In other words, they took just enough time to ensure a correct goal-directed response, effectively minimizing habitual errors. Consequently, their responses in the outcome-devaluation test were primarily controlled by the goal-directed system, which explains why habit responses in this task did not significantly increase with overtraining.

This pattern of results is not entirely surprising, given that the outcome-devaluation task explicitly encouraged accurate response switching through its instructions. In such cases, using a fixed response time limit may not be ideal for obtaining a sensitive measure of habitual responses across participants, as they are likely to adjust their response times to optimize performance within the given constraints. In our opinion, shortening the response time limit even further does not appear to be a viable solution, as it would prevent participants from producing the instructed goal-directed response. This would likely lead to frustration and an increase in random responses, which could artificially inflate the number of detected "habitual" responses. Instead, a more valid approach to measuring habitual responses may involve de-emphasizing accuracy in task instructions while allowing participants to respond within a range of preparation times. This approach would reduce the frustration caused by frequent errors while still capturing habit-driven behavior. Additionally, it is advisable to include more than two response options, as this would help distinguish between random and habitual responses (Du & Haith, 2025).

Even when a goal-directed response is successfully produced, we can still study the functioning of the habit system through its interference with goal-directed processes, as demonstrated in Luque et al. (2020). For the outcome-devaluation task, the authors computed an RT switch cost score by subtracting the RT of correct responses during the training blocks (baseline) from those during the test blocks to control for confounding factors like fatigue, which could otherwise artificially affect RTs during the test. Using a baseline therefore seems necessary and it is common in the literature, especially when the amount of training is manipulated (e.g., Adams, 1982; Gera et al., 2024). Luque et al. (2020) showed that overtraining increased RTs for goal-directed switches, both with and without baseline correction (as reported in their supplemental material). Intriguingly, although we have replicated their data when the baseline was subtracted, the same pattern was not observed in our uncorrected data. Instead, we found no differences as a function of the amount of training in RT during the outcome devaluation test, and significant differences in the baselines (faster responses for the more trained associations). Arguably, the observed difference in the baseline RT precisely justifies the correction of the test data. However, the absence of difference in the uncorrected data is not ideal, because it allows us to explain the RT switch cost effect just because of the difference in the baseline. In general, manipulating the amount of experience leads to different baselines, and that produces interpretation problems. The solution to this issue is not straightforward; in any case, at least baseline data should be reported and analyzed, and the main results should be reported relative to the baseline (when needed) and without baseline correction, so the reader can have all the relevant information.

In the forced-response test, a latent interference effect was also evident when examining RTs for correct responses to remapped associations. We calculated the difference between the mean RT for correct responses to remapped associations and that for consistent associations (baseline). As predicted, participants in the overtrained condition took longer to respond correctly to novel associations compared to the minimal training condition. Furthermore, the more time pressure was applied (i.e., later stimulus onsets), the greater the RT switch cost. On this occasion, the effect was also observed when the baseline correction was not applied. These delays suggest that, at times, participants successfully

inhibited strong habitual responses, but doing so came at the cost of delaying the synchronization of their response with the imposed tone (see Fig. 4D).

The forced-response task proves to be a valuable tool for examining measures sensitive to overtraining effects at the group level: habit responses and RT switch cost. Currently, no other task offers the capability to study both habitual and goal-directed processes within a single paradigm. While these two measures likely capture different aspects of the complex interaction between habit and goal-directed systems, they should also share some variance related to the strength of a specific S–R habit. This raises the question of whether these measures are associated as indicators of habit system activation. Our results revealed a positive correlation pattern between habit responses and RT switch cost, though our data did not support the alternative hypothesis. One possible interpretation is that these measures reflect independent aspects of habit formation. However, there is an important caveat to consider: while the habit response measure demonstrated adequate internal consistency, the RT switch cost measure exhibited poor reliability. This low reliability may have attenuated the observed correlation, preventing us from drawing statistical inferences regarding convergent validity. In an exploratory analysis, we applied Spearman's (1904) attenuation-correction equation to estimate the correlation if both measures had perfect reliability. This correction suggested that the true correlation between them would be $r = .30$. To determine whether these measures truly capture the same cognitive construct, future studies should account for the expected attenuation when calculating sample sizes for correlational analyses. Additionally, efforts should be made to improve the reliability of these measures, as doing so would reduce measurement error and, hence, attenuation in observed correlations.

In this regard, as noted above, the RT switch measure was constructed using a "subtraction method." The reliability of a difference score is reduced when its components are correlated (Cronbach & Furby, 1970). Subtracting one from the other can remove systematic variability, decreasing the proportion of "true" variance relative to error variance and thus lowering reliability (Cronbach & Furby, 1970; see also Draheim et al., 2019). In our case, the components of the RT switch measure in the forced-response test were highly correlated ($r = 0.91$, see Supplemental Material A5), which likely reduced the true variance in our difference score. Several approaches have been proposed to enhance the reliability of difference scores (see Zorowitz & Niv, 2023). For our measure, we believe that the most straightforward approach would be to increase the ratio of variance between the component measures. The reliability of a difference score improves when the variances of its components differ substantially, as this reduces the proportion of shared variance (Chiou & Spreng, 1996; Zorowitz & Niv, 2023). This suggests that if we could reduce participants' synchronization delay, it is plausible that consistent trials would show minimal variance—with responses almost always perfectly synchronized—compared to remapped trials. In that case, using a difference score might not be necessary, and we could instead rely solely on the RT of remapped trials.

We preregistered the expectation that both measures of RT switch cost (one from each task) would correlate positively, reflecting the same underlying cognitive process (the response interference produced by activating an S–R habit). However, our data supported the absence of a positive correlation. Unlike the previous case, applying an attenuation correction to estimate the true correlation is not meaningful here, as the correlation was not even in the expected direction. If these measures capture, at least to some extent, the same cognitive construct, a positive and significant correlation should have been expected, as we preregistered. However, we did not find any evidence supporting this predicted association.

While our habit measures produced results consistent with those expected from the habit construct—as they were sensitive to the amount of training—we failed to find a positive association between them. This lack of correlation may be due to poor reliability in cases where the RT switch measure of the forced-response task was involved. However, measurement noise alone is unlikely to explain the absence of a correlation between habit responses from the forced-response task and RT switch cost from the outcome-devaluation task. This raises at least two possible explanations, each with important implications for research in this field.

One possibility is that there are no individual differences in how the habit and goal-directed systems interact to produce observable responses, or no differences stable enough to be captured across different tasks and/or at different time points. If this is the case, the extent to which a habit is formed or the ability of the goal-directed system to inhibit habits may be largely determined by external circumstances rather than by stable individual traits. A substantial body of research suggests that the functioning of these systems may be altered in psychopathologies (see Buabang et al., 2024). It remains possible that these systems do function differently in clinical populations but that in cognitively healthy individuals, individual differences are negligible. However, it is worth noting that much of the existing evidence for individual differences in habit formation comes from studies that do not use validated measures of habits. This leaves open the possibility that these studies were actually assessing other related cognitive constructs, such as impulsivity or cognitive control, rather than habit formation per se (e.g., Barzilay et al., 2022; Kalanthroff et al., 2016).

Another possibility is that RT switch cost in the outcome-devaluation task and habit responses in the forced-response

task are influenced, at least in part, by different cognitive processes. While both measures may capture the progressive increase in habitual control with increased training, it is also reasonable to think that they reflect distinct aspects of how habits and goal-directed processes interact. RT switch cost likely reflects not only habit strength but also the functioning of the goal-directed system, specifically its ability to inhibit the habit and produce the appropriate goal-directed response. In contrast, habit responses measured in the forced-response test may be less dependent on goal-directed control, as habitual errors occur primarily when the goal-directed system has not yet been activated.[2] The fact that the measures shared almost no variance suggests that captured processes might be qualitatively different. In this vein, however, it should also be considered that each task engages mechanisms beyond the balance between goal-directed and habit processes. Even though both tasks are intended to assess habitual behavior, these distinct influences could limit the extent of shared variance between the measures.

Future research should consider that, in the forced-response task, participants often struggled to synchronize their responses with the tone indicating when to respond. Poor synchronization led to an underrepresentation of particularly short preparation times, increasing uncertainty in these periods. In our study, it was not possible to fit a speed–accuracy trade-off function for 24 participants, as they were always artificially accurate by avoiding short preparation times that would likely result in random responses. Failing to adjust responses to the imposed tone represents strategic processes (Haith et al., 2016), which can interfere with measuring automatic processes and complicate the interpretability of computational models designed to isolate them. For instance, in an effort to improve computational model interpretability, Adkins et al. (2024) applied the forced-response paradigm to a different task and excluded between 30.6% and 40.3% of trials across four experiments because participants responded ± 100 ms off synchronization. Similarly, Adkins and Lee (2024) reported even higher exclusion rates for another task, omitting more than 58.1% of trials across three experiments—though not all exclusions were due to desynchronization issues. In our study, 37.87% of trials could have been considered desynchronized. However, since our focus was not only on overtrained responses but also on how habits interfere with goal-directed responses (as measured by RT switch cost), we chose not to exclude these trials.

The authors of the aforementioned studies argued that their results remained consistent without data exclusions. However, we believe that post hoc adjustments to address participants' synchronization issues (as described above) could expand the range of possible analysis decisions, a phenomenon referred to as "the garden of forking paths" (Gelman & Loken, 2013), where seemingly minor methodological choices may lead to substantially different statistical inferences. Accordingly, multiverse studies (Parsons, 2022; Steegen et al., 2016) using this paradigm may be necessary, where measures are computed using different a priori valid preprocessing pipelines to assess the robustness of statistical conclusions. More importantly, the difficulty participants had in synchronizing their responses suggests that the forced-response task may be challenging to perform. This difficulty raises concerns about the reliability of the paradigm in broader experimental settings and across diverse populations. If synchronization issues introduce significant measurement noise, they could ultimately restrict the paradigm's applicability and generalizability, making it suitable only for specific contexts. Future studies should explore ways to improve synchronization in the forced-response task. Possible solutions include refining instructions, providing extensive practice beforehand, or even developing a gamified version of the task designed to encourage precise synchronization.

Finally, there are limitations to our assessment of the convergent validity of habit measures that should be acknowledged. One of our preregistered objectives was to conduct a direct replication attempt, and accordingly, we did not counterbalance the order of the tasks. Additionally, in both tasks, participants completed two training conditions with a habit test, resulting in extensive experience with habit tests, which may not be ideal—although we found group-level evidence of habit expression in both tasks regardless of the training condition order, as reported in Supplemental Material A6. However, future studies should counterbalance task order and consider including only an overtrained condition per task to minimize potential cross-task influence. Lastly, correlations involving the RT cost measure in the forced response remained inconclusive due to its low reliability and the reduction in sample size after exclusions in this task, which should be considered when planning future studies.

In conclusion, the present study has contributed to habit research by examining the validity and reliability of two of its most promising paradigms. Our findings indicate that both tasks produced measures consistent with the expected characteristics of habits, demonstrating evidence of increased habitualization with extended training. The key measures examined were the number of habitual errors observed during testing and the RT cost of switching from a habit to a new goal-directed response. Our results

[2] It has also been argued that habitual behavior may be misconceptualized in the literature and be the result of outcome expectancy-driven processes rather than stimulus-driven processes (De Houwer et al., 2018). We do not enter into this ongoing debate, and anchor our study in the dual-process theory, which guided the design of the measures we assessed.

suggest that the forced-response task may be better suited to studying habitual errors, as errors in the outcome-devaluation task were not affected by the amount of training. By contrast, RT switch cost appears to be more adequately measured in the outcome-devaluation task, as only this task exhibited appropriate reliability for that measure. Importantly, these two measures were poorly correlated, likely because they capture different aspects of the balance between habitual and goal-directed processing. Habitual errors in the forced-response task arise from the (relatively) uncontested activation of the habit system, while RT switch cost in the outcome-devaluation task reflects the direct competition between these two systems when they are co-active.

The findings from this study provide important insights for future researchers examining individual differences in human decision-making and its relationship with psychopathology. Researchers should carefully consider which aspects of the balance between goal-directed (controlled) and habitual (automatic) processing are most relevant to their research. If the goal is to obtain a "purer" measurement of habit strength or automatic response activation, tasks similar to the forced-response task developed by Hardwick et al. (2019) would be more appropriate, particularly for populations that can manage a potentially demanding task. However, if the focus is on understanding how the goal-directed (control) system interacts with strong (automatic) habits, then RT switch cost, as measured in tasks like the one developed by Luque et al. (2017, 2020), could be a more suitable approach. Regardless of the approach chosen, we emphasize the importance of reporting reliability and gathering validity evidence for these measures. This study underscores the advantages of carefully evaluating the psychometric properties of tools designed to measure habit induction, ideally before applying them to more specific research questions, such as studying individual differences. Failing to do so risks conducting resource-intensive research that may ultimately yield null or difficult-to-interpret results.

Acknowledgements The present work forms part of the PhD dissertation of PML within the Psychology Doctoral Program at the University of Málaga, under the supervision of DL.

Authors' contributions Author roles were classified using the Contributor Role Taxonomy (CRediT; https://credit.niso.org/) as follows: PML: Conceptualization, Methodology, Software, Investigation, Data curation, Formal Analysis, Visualization, Validation, and Writing–original draft. AVM: Conceptualization, Methodology, Software, Investigation, Data curation, and Writing–review & editing. FGF: Data curation, Formal analysis, Visualization, Validation, and Writing–review & editing. DL: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, and Writing–review & editing.

Funding This article was part of the project PID2021-126767NB-I00 funded by MICIU/AEI/10.13039/501100011033 and FEDER, UE. PML was supported by the grant PRE2022-103151MICIU/ESF funded by MICIU/AEI/10.13039/501100011033 and ESF+. AVM was supported by the project ProyExcel_00287, funded by the Andalusian Autonomic Government. FGF was supported by an FPU predoctoral grant (ref. FPU20/00826).

Data availability All data used in the current study are publicly available at https://osf.io/4s3c8/.

Code availability Analysis scripts and experimental programs used in the current study are publicly available at https://osf.io/4s3c8/.

Declarations

Conflicts of interest We have no known conflicts of interest to disclose.

Ethics approval The procedure conformed to the ethical standards outlined in the 1975 Helsinki Declaration and was approved by the Human Research Ethics Advisory Committee (Psychology) of the University of Málaga (46–2020-H).

Consent to participate All participants provided informed consent before taking part in the study.

Consent to publish Not applicable.

References

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

Adkins, T. J., & Lee, T. G. (2024). Reward accelerates the preparation of goal-directed actions under conflict. *Journal of Cognitive Neuroscience, 36*(12), 2831–2846. https://doi.org/10.1162/jocn_a_02072

Adkins, T. J., Zhang, H., & Lee, T. G. (2024). People are more error-prone after committing an error. *Nature Communications, 15*(1), 6422. https://doi.org/10.1038/s41467-024-50547-y

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software, 111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences, 14*(5), 208–215. https://doi.org/10.1016/j.tics.2010.02.001

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

Banca, P., Ruiz, M. H., Gonzalez-Zalba, M. F., Biria, M., Marzuki, A. A., Piercy, T., Sule, A., Fineberg, N. A., & Robbins, T. W. (2024). Action-sequence learning, habits and automaticity in obsessive-compulsive disorder. *eLife*, *12*, RP87346. https://doi.org/10.7554/eLife.87346.3

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), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barzilay, S., Fradkin, I., & Huppert, J. D. (2022). Habitual or hyper-controlled behavior: OCD symptoms and explicit sequence

learning. *Journal of Behavior Therapy and Experimental Psychiatry, 75*, 101723. https://doi.org/10.1016/j.jbtep.2022.101723
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
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357
Buabang, E. K., Donegan, K. R., Rafei, P., & Gillan, C. M. (2024). Leveraging cognitive neuroscience for making and breaking real-world habits. *Trends in Cognitive Sciences*, S1364661324002663. https://doi.org/10.1016/j.tics.2024.10.006
Champely, S. (2006). *pwr: Basic Functions for Power Analysis* (p. 1.3–0) . https://doi.org/10.32614/CRAN.package.pwr
Chen, H., Xie, M., Ouyang, M., Yuan, F., Yu, J., Song, S., Liu, N., & Zhang, N. (2024). The impact of illness duration on brain activity in goal-directed and habit-learning systems in obsessive-compulsive disorder progression: A resting-state functional imaging study. *Neuroscience, 553*, 74–88. https://doi.org/10.1016/j.neuroscience.2024.06.018
Chiou, J., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 9*, 158–167.
Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology, 8*, 1482. https://doi.org/10.3389/fpsyg.2017.01482
Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80. https://doi.org/10.1037/h0029382
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957
Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences, 24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007
Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5. https://doi.org/10.11648/j.ajtas.20160501.12
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., & Dickinson, A. (2009). Associative theories of goal-directed behaviour: A case for animal–human translational models. *Psychological Research, 73*(4), 463–476. https://doi.org/10.1007/s00426-009-0230-6
de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A. C., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. M. (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
Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78. https://doi.org/10.1098/rstb.1985.0010
Dickinson, A., Balleine, B., 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
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
Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*(5), 508–535. https://doi.org/10.1037/bul0000192
Du, Y., & Haith, A. M. (2025). Dissociable habits of response preparation versus response initiation. *Nature Human Behaviour, 9*(9), 1941–1958. https://doi.org/10.1038/s41562-025-02215-4
Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063
Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist, 77*(4), 576–588. https://doi.org/10.1037/amp0001006
Garre-Frutos, F., Vadillo, M. A., González, F., & Lupiáñez, J. (2024). On the reliability of value-modulated attentional capture: An online replication and multiverse analysis. *Behavior Research Methods, 56*, 5986–6003. https://doi.org/10.3758/s13428-023-02329-5
Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*, 1–17.
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
Gera, R., Segev, B., & Schonberg, T. (2024). A novel free-operant framework enables experimental habit induction in humans. *Behavior Research Methods, 56*, 3937–3958. https://doi.org/10.3758/s13428-023-02263-6
Gillan, C. M., 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. M., Robbins, T. W., Sahakian, B. J., Van Den Heuvel, O. A., & Van Wingen, G. (2016). The role of habit in compulsivity. *European Neuropsychopharmacology, 26*(5), 828–840. https://doi.org/10.1016/j.euroneuro.2015.12.033
Graybiel, A. M., & Grafton, S. T. (2015). The striatum: Where skills and habits meet. *Cold Spring Harbor Perspectives in Biology, 7*(8), a021691. https://doi.org/10.1101/cshperspect.a021691
Guida, P., Michiels, M., Redgrave, P., Luque, D., & Obeso, I. (2022). An fMRI meta-analysis of the role of the striatum in everyday-life vs laboratory-developed habits. *Neuroscience and Biobehavioral Reviews, 141*, 104826. https://doi.org/10.1016/j.neubiorev.2022.104826
Haith, A. M., & Krakauer, J. W. (2018). The multiple effects of practice: Skill, habit and reduced cognitive load. *Current Opinion in Behavioral Sciences, 20*, 196–201. https://doi.org/10.1016/j.cobeha.2018.01.015
Haith, A. M., Pakpoor, J., & Krakauer, J. W. (2016). Independence of movement preparation and movement initiation. *The Journal of Neuroscience, 36*(10), 3007–3015. https://doi.org/10.1523/JNEUROSCI.3245-15.2016
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
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
JASP Team. (2024). *JASP (Version 0.19.2)* [Computer software]. https://jasp-stats.org/
Kalanthroff, E., Abramovitch, A., Steinman, S. A., Abramowitz, J. S., & Simpson, H. B. (2016). The chicken or the egg: What drives OCD? *Journal of Obsessive-Compulsive and Related Disorders, 11*, 9–12. https://doi.org/10.1016/j.jocrd.2016.07.005
Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS*

Computational Biology, 7(5), e1002055. https://doi.org/10.1371/journal.pcbi.1002055

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America, 113*(45), 12868–12873. https://doi.org/10.1073/pnas.1609094113

Knowlton, B. J., & Patterson, T. K. (2018). Habit Formation and the Striatum. In R. E. Clark & S. J. Martin (Eds.), *Behavioral Neuroscience of Learning and Memory* (pp. 275–295). Springer International Publishing. https://doi.org/10.1007/7854_2016_451

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

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, 37*(11), 3009–3017. https://doi.org/10.1523/JNEUROSCI.3205-16.2017

Luque, D., Molinero, S., Watson, P., López, F., J., & Le Pelley, M., E. (2020). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology: General*, *149*(8), 1449–1459. https://doi.org/10.1037/xge0000722

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Molinero, S., Martínez-López, P., Morís, J., Quintero, M. J., Cobos, P. L., López, F. J., & Luque, D. (2025). The degraded contingency test fails to detect habit induction in humans. *PLoS One, 20*(10), e0334087. https://doi.org/10.1371/journal.pone.0334087

Moore, S., Wang, Z., Zhu, Z., Sun, R., Lee, A., Charles, A., & Kuchibhotla, K. V. (2023). *Revealing abrupt transitions from goal-directed to habitual behavior*. https://doi.org/10.1101/2023.07.05.547783

Nebe, S., Kretzschmar, A., Brandt, M. C., & Tobler, P. N. (2024). Characterizing human habits in the lab. *Collabra: Psychology, 10*(1), 92949. https://doi.org/10.1525/collabra.92949

Parsons, S. (2021). splithalf: Robust estimates of split half reliability. *Journal of Open Source Software, 6*(60), 3041. https://doi.org/10.21105/joss.03041

Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2577

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive Behavioral Measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395. https://doi.org/10.1177/2515245919879696

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., 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, 29*(1), 16–28. https://doi.org/10.1101/lm.053413.121

R Core Team. (2023). *R: A Language and Environment for Statistical Computing* (Version 4.3.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Sookud, S., Martin, I., Gillan, C. M., & Wise, T. (2025). Impaired goal-directed planning in transdiagnostic compulsivity is explained by uncertainty about learned task structure. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. https://doi.org/10.1016/j.bpsc.2025.10.005

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72. https://doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

The MathWorks Inc. (2022). *MATLAB* (Version 2022b) [Computer software]. The MathWorks Inc. https://www.mathworks.com

Thrailkill, E. A., & Bouton, M. E. (2015). Contextual control of instrumental actions and habits. *Journal of Experimental Psychology: Animal Learning and Cognition, 41*(1), 69–80. https://doi.org/10.1037/xan0000045

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

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Watson, P. (2024). Defining and Measuring Habits Across Different Fields of Research. In Y. Vandaele (Ed.), *Habits: Their Definition, Neurobiology, and Role in Addiction* (pp. 3–22). Springer International Publishing.

Watson, P., Wiers, R. W., Hommel, B., & de Wit, S. (2014). Working for food you don't desire. Cues interfere with goal-directed food-seeking. *Appetite, 79*, 139–148. https://doi.org/10.1016/j.appet.2014.04.005

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

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science, 3*(1), 94–123. https://doi.org/10.1177/2515245919885611

Wood, S. N. (2017). *Generalized Additive Models: An introduction with R* (2nd Edition). Chapman and Hall/CRC.

Wood, S. N. (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (p. 1.9–1) . https://doi.org/10.32614/CRAN.package.mgcv

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology, 67*(1), 289–314. https://doi.org/10.1146/annurev-psych-122414-033417

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004