How should we measure exploration?

Kristin Witte[1,+,*], **Mirko Thalmann**[1,+,*], **and Eric Schulz**[1]

[1]Helmholtz Munich, Center for Computational Health, Neuherberg, 85764, Germany
[*]kristin.witte@helmholtz-munich.de; mirko.thalmann@helmholtz-munich.de
[+]these authors contributed equally to this work

ABSTRACT

An increasing number of studies have used multi-armed bandit tasks to investigate individual differences in exploration behavior. However, the psychometric properties of exploration measures remain unexplored. We examine the test-retest reliability, convergent, divergent, and external validity of model-based estimates of exploration strategies using three canonical paradigms. Our results revealed poor to moderate reliability, with minimal correlations for the same strategy across tasks. By refining the computational models, we identified two convergently valid latent factors representing value-guided and directed exploration. However, these factors showed no significant correlation with self-reported exploration tendencies and only weak associations with mood fluctuations and symptoms of anxiety and depression. The exploration factors were, however, highly correlated with working memory capacity, questioning whether they provide additional information beyond performance-related constructs. To improve future research, we suggest simplifying common computational models and using multiple tasks to more accurately measure exploration strategies and mitigate spurious correlations arising from task-specific factors.

Keywords: exploration-exploitation; exploration strategies; few-armed bandits; reliability; validity

Introduction

To explore or not to explore has significant implications. For many organisms, exploration is key to discovering new options. In humans, this might be akin to joining a new social group or trying out a new sport. The unfamiliar option could turn out worse than the one already known, but it might also prove better. Exploitation, on the other hand, involves selecting the option that one knows already to be good. The balance between exploration and exploitation is essential, and has been examined in a wide range of domains (for a review see Mehlhorn et al.[1]). With regards to human behavior, one question has been of chief interest in the past two decades: What are the cognitive strategies people apply to explore and exploit their environment?

Exploration strategies have been typically studied in so-called few-armed bandit tasks. In these tasks, participants are instructed to maximize their rewards by sequentially selecting between a few response options (i.e., the arms of the bandit). At the onset of each round, participants lack prior information about the arms and must learn from experience which arm yields the highest average reward, thereby creating an exploration-exploitation dilemma. Paid out rewards on the arms are usually randomly sampled from a distribution, for example, a normal distribution. Three of the most commonly used few-armed bandit tasks to study exploratory behavior are the Two-armed bandit[2], the Restless bandit[3], and the Horizon task[4]. In the Two-armed bandit paradigm, each arm is associated with a mean reward, with trial-by-trial payouts perturbed by Gaussian noise. Usually, a round lasts no more than ten trials, but participants complete several rounds. The Restless bandit typically uses four arms, with mean rewards evolving according to a Gaussian random walk over only one round of 200 trials, and observed rewards generated by adding noise to these means. Randomly-walking arms are assumed to motivate continued exploratory behavior over the course of a round. The Horizon task, a modified Two-armed bandit, addresses the confound between an agent's estimate of an arm's value and its associated uncertainty, which typically become correlated over trials due to preferential sampling of the higher-value arm. The Horizon task incorporates an initial phase of four forced choices, followed by six free choices, effectively de-correlating value from uncertainty on the first free choice across all rounds. Consequently, the first free choice provides a statistically cleaner measure of how individuals integrate information about value and uncertainty in their decision-making process.

These three tasks have been used to answer a wide variety of questions: First, the Two-armed bandit has been used to study how people explore in general[2,5], how exploration is reflected in brain activity[6] and in pupil dilation[7], how groups differ in exploratory behavior (genetically[8] and developmentally[9]), and how individual differences in exploration strategies are related to psychiatric traits[10]. Second, the Restless bandit has typically only been used to study average exploratory behavior[3,11–13]. Third, the Horizon task has been used to study the average participants' exploration profile behaviorally[4] and in the brain[14], group differences in exploration strategies (drugs[15], development[16,17], humans vs. large language models[18]), and again individual differences and their relation to psychological traits[19]. One main conclusion of this body of research is that humans use a

mixture of value-guided and uncertainty-guided exploration strategies[20,21]. In other words, individuals tend to explore arms despite lower expected rewards (i.e., value) and arms about which they possess less information (i.e., uncertainty).

While initial research focused on exploration strategies on the population level, recent work has used few-armed bandit tasks as measurement instruments, i.e. to look at individual differences of exploratory behavior. For example, Fan et al.[10] showed that people who describe themselves as anxious make less use of uncertainty in their decisions and tend to respond more deterministically. These results and similar ones[15,19,22–24] provide the groundwork for understanding cognition in people with psychiatric conditions, and for providing instruments to improve these conditions therapeutically. However, we only know very little about the psychometric properties of exploration strategies as inferred using computational model parameters. Specifically, it is yet to be established whether (a) individual differences in model-based estimates of exploration strategies are stable over time (i.e., test-retest reliability), which is a prerequisite for their use as measurement tools, (b) the same exploration strategy is correlated across tasks (i.e., convergent validity), which is a prerequisite for considering the strategy to be a general phenomenon, rather than a task-specific artifact, (c) the strategies are sufficiently distinct from a general, performance-related construct (i.e., divergent validity with respect to working memory), and (d) the strategies are predictive of self-report measures of exploration (i.e., criterion validity).

While some studies showed mediocre to acceptable test-retest reliabilities of model parameters (Four-armed Restless bandit with discrete rewards[25], probabilistic reversal learning task[26], Two-armed bandit[27], predictive inference task[28,29]), to the best of our knowledge, so far no study has evaluated the validity of exploration strategies as inferred by computational model parameters (but see Anvari et al.[30] for a model-agnostic analysis and Jach et al.[31] for an investigation of the convergent and external validity of model parameters extracted from information seeking tasks). This is problematic, because a lack of reliability and convergent validity can lead to spurious correlations between variables and inconsistent research results such as the inconsistent findings on the relationship between exploration and anxiety[10,19,23,24,32]. It is currently unclear whether the reliability of these measures is sufficient, and whether they indeed measure a general exploration construct, which would allow us to use them as stable traits in individual difference research.

In general, correlating model parameters derived from a single task, for example with psychometrically developed scales, is problematic for several reasons. First, according to classical test theory, any measure contains error variance. Second, the resulting construct variance can be contaminated by task-specific variance[33]. Third, parameter trade-offs, induced during model fitting[34], may lead to systematic bias in the estimated model parameters. Individual differences in these parameters, in effect, cannot be interpreted to reflect variability in the construct in question. In the past, using a broad range of tasks, representing a cognitive construct through a confirmatory factor analysis has been successfully applied to extract general factors of working-memory capacity[33] and components of reaction-time distributions[35], and to circumvent the three just mentioned problems. In particular, the confirmatory approach allows us (a) to remove correlations between parameters within a single task induced during the model fitting by modeling correlated residuals, which is, for example, not possible with principal component analysis (i.e., PCA), and (b) to extract generalizable aspects of exploration strategies by modeling the correlations between the same strategy in several tasks, and therefore effectively ignoring task-specific artifacts (e.g., task strategies). Furthermore, similar work in the domain of category learning[36,37] has shown that a general factor of category learning was highly correlated with working memory capacity, pointing to a lack of divergent validity, and possibly to conceptual overlap between the constructs[38,39].

The present work therefore aims to systematically examine the recoverability, test-retest reliability, convergent, divergent and external validity of three exploration tasks. Additionally, we present specific recommendations of how to improve the measurement of exploration strategies. Task-based exploration parameters had mediocre test-retest reliability, at best. Theoretically-informed latent constructs of the exploration strategies could not be computed when we used the previously suggested models in these three tasks. Unifying the computational models to estimate the strategies, and removing a highly-correlated strategy from the model in the Two-armed bandit allowed us to derive a general construct of value-guided exploration, which was temporally highly stable. We were also able to derive a general factor of directed exploration, however only for the second task session. The two strategies were highly correlated with each other, and value-guided exploration was highly correlated with working-memory capacity. Moreover, the strategies were unrelated to self-reported exploration behavior.

Results

General approach

In each of two sessions separated by six weeks, participants completed a series of three working memory paradigms, three few-armed bandit paradigms, and five questionnaires (see Fig.1). This allowed us to assess the test-retest reliability of all used measures. We used three canonical few-armed bandit paradigms to test whether it is possible to extract generalizable aspects of the proposed exploration strategies (i.e., convergent validity). In particular, we used the Horizon task introduced by Wilson et al.[4], the Two-armed bandit introduced by Gershman[2], and the Restless bandit introduced by Daw et al.[3] In all three

bandit paradigms, participants were required to collect as many rewards as possible in a limited number of trials, creating an exploration-exploitation dilemma.

We used the three working-memory paradigms to assess whether exploration differs from a general, performance-based ability (i.e., divergent validity). The three paradigms consisted of two complex-span tasks (an operation span task, OS, and a symmetry span task, SS)[40] taken from the short working-memory capacity battery by Oswald et al.[41] and a working-memory updating task (modeled after von Bastian et al.[42]). The questionnaires included two scales to assess the external validity of exploration, i.e. the Curiosity and Exploration Inventory (CEI)[43], the Openness subscale from the short form of the Big Five Inventory-2 (BFI-2)[44], as well as three scales to assess the criterion validity of the exploration strategies, i.e. the State Trait Inventory of Cognitive and Somatic Anxiety (STICSA)[45], the Patient Health Questionnaire Mood Subscale (PHQ-9)[46], and the Positive Affect Negative Affect Scale (PANAS)[47].

To estimate the extent to which participants' choices were guided by different exploration strategies, we fit those computational models to the choices in the three bandit paradigms that have previously been proposed by the respective authors.[2,4,11]. Although the implementation differs slightly across models, they all assume that people sequentially learn the expected values of the arms of the bandit and their uncertainties, and subsequently integrate this information to make a choice. The models in the Horizon task and in the Restless bandit propose that it is only the learned values (value-guided exploration) and their associated uncertainties (directed exploration), which guide people's choices. The model in the Two-armed bandit additionally assumes that the relation between the learned values and the total uncertainty across both arms (random exploration or Thompson sampling) also affects choices. We implemented all three models as hierarchical Bayesian models, which made parameter estimation more robust, improved out-of-sample prediction and the parameters' reliabilities, and did not change anything about the individual estimates and group-level patterns (see Tables **??** and **??** and Figures **??** and **??** in the SI). Note that our main analyses and hypotheses as well as the exclusion criteria were preregistered with the Open Science Framework (https://osf.io/cavj3). All task and analysis code as well as all raw data is available under https://osf.io/ra7su/.

Recoverability

We tested the parameter recovery for the commonly used models in all three bandit tasks by fitting the observed choices from the first session with the respective model, simulating data from the fitted parameters, re-fitting that data and estimating the correlation between the fitted and recovered parameters. All parameters recovered reasonably well (see Figure 2). Recovery in the Horizon task, however, was notably worse than in the other tasks. We additionally observed a strong off-diagonal correlation in the Two-armed bandit (Figure 2B) between value-guided exploration and random exploration (-0.79). This points towards redundancy in the parameters and multicollinearity in the predictors. This in turn can reduce the interpretability of the parameter estimates. Lastly, the parameter recovery for the Restless bandit task was reliable with correlations as high as 0.95 and low off-diagnonal correlations.

Replicability

We then tested whether the group-level exploration patterns were replicable across time points (see Figure 3). For the Horizon task, the group-level pattern changed qualitatively, with value-guided exploration disappearing completely in the second session. This was due to the fact that participants used value-guided exploration in the long and the short horizon to the same degree in Session 2, but not in Session 1. Notably, because of the difference calculation of the strategies, variance in the group-level posterior distributions was inflated (see top panel of Fig. 3 [48–50]). For the Two-armed bandit and the Restless bandit, the group-level patterns replicated well across time points. Note however, that for the Two-armed bandit there was virtually no signature of value-guided exploration as all variance was taken up by the highly correlated random exploration parameter. Also note that for the Restless bandit, the group-level parameter for directed exploration was negative, indicating that participants actively avoided uncertain options instead of exploring them.

Reliability

Next, we evaluated whether the model parameters were consistent over the six-week period. Specifically, we calculated the Intraclass Correlation coefficient (ICC) between the model parameters estimated at the first and at the second session. In line with previous work on the Two-armed bandit[27], the model parameters demonstrated mediocre reliabilities at best, with some parameters having a poor reliability[51]. The shared variance between measurement time points (R^2) was less than 0.36 for all parameters, and for most even less than 0.27. The low reliability observed in the Horizon task may be partially attributed to the fact that exploration is defined by the difference between two experimental conditions, and difference scores are notoriously known for providing low reliabilities[48–50]. As the stable proportion of the variance in task-based model parameters is mediocre, the parameters are affected to a large degree by theoretically less interesting, unwanted sources of variance. Besides the model parameters, we also analyzed a few model-free task-based measures. The task-based measures, including performance and switch probability, exhibited moderate to good reliability. Remarkably, the switch probability (i.e. the proportion of trials on

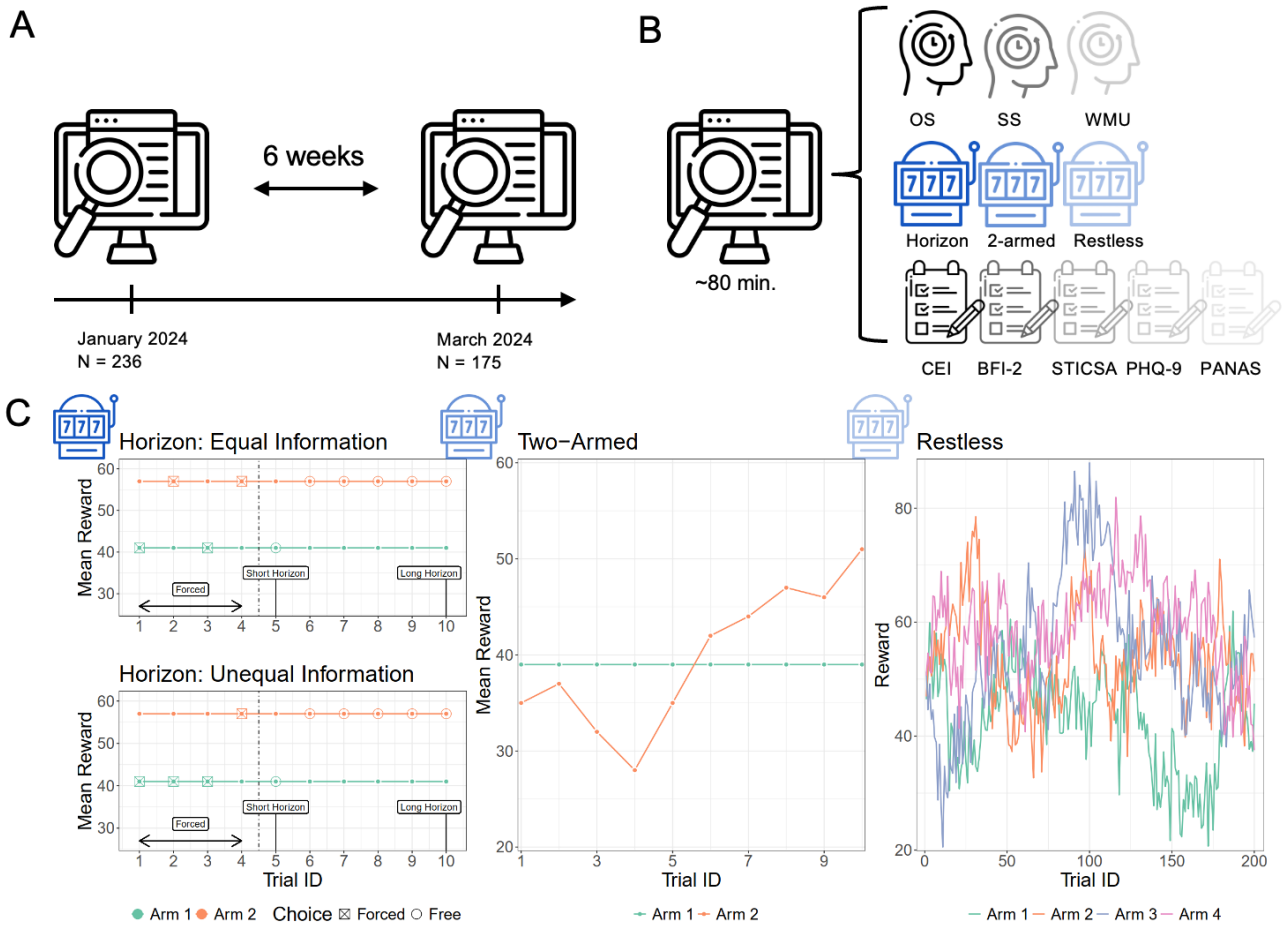

Figure 1. Study overview. **A:** Our study consisted of two study sessions six weeks apart. In the first session, 236 participants passed the inclusion checks and were re-invited for the second session. Out of these 236 participants, 175 participated in the second session and passed the inclusion checks again. **B:** Each study session lasted for around 80 minutes and consisted of three working memory tasks, three bandit tasks and five questionnaires. **C:** Reward structures of the three bandit tasks. In the Horizon task, the two mean rewards were stable within each block. In the Two-armed bandit, either one or both or none of the arms had drifting mean rewards throughout a round. In the Restless bandit, all arms had mean rewards that were continuously drifting throughout the 200 trials. See Methods for details.

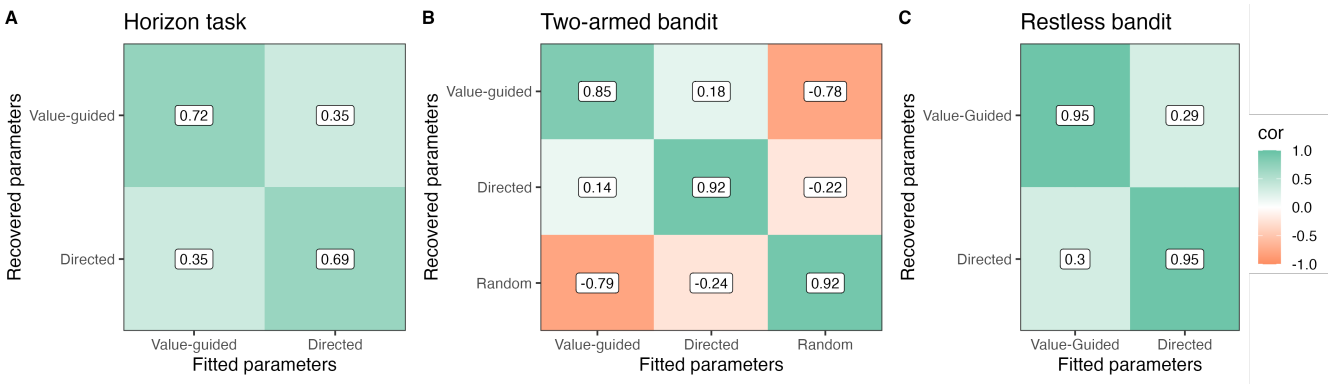

Figure 2. Parameter recovery using the default models for each task. Numbers indicate the correlation between fitted and recovered parameter estimates when using the data from the first session.

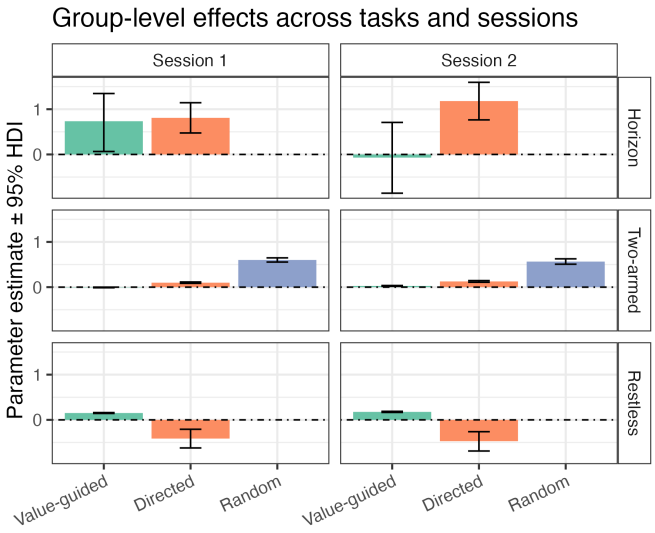

Figure 3. Replicability of the group-level effects. The height of the bars indicates the mean of the posterior distribution for that parameter at that time point. The error bars indicate the 95% Highest Density Interval of the posterior distribution.

which participants selected a different arm than on the previous trial) was consistently more reliable than any of the model-based parameters (see Figure 4A).

Convergent Validity

The then tested whether the model parameters of the same exploration strategy correlated across tasks. The parameters showed poor convergent validity (Figure 4D). Specifically, the value-guided exploration parameters showed very low correlations across tasks with some correlations even being negative ($r = -0.17$ between Restless bandit and Two-armed bandit). For the directed exploration parameters, the correlation between parameters from the Two-armed bandit and the Restless bandit was 0.39. Parameters from the Horizon task were however very weakly correlated with the parameters from the other two tasks (0.12 and 0.17). This poor convergent validity in the model parameters can be explained by two factors: Firstly, in the Two-armed bandit task, the model parameters for value-guided and random exploration are strongly negatively correlated and traded off with each other. Secondly, in the Horizon task, the parameters are calculated as the difference between the long and the short horizon, making them conceptually different from the other two tasks. In addition to the model parameters, we also tested the convergent validity of a model-free approximation of exploration, namely the switch probability (P(switch), Figure 4B). This measure of exploration seemed reasonably consistent across tasks with correlations between 0.36 and 0.49. These correlations are also comparable in strength to the correlations in performance between the three different working memory tasks (Figure 4C).

External & Criterion Validity

We only observed relatively small correlations between the model parameters of exploration and questionnaire-based measures of exploration (strongest correlation was .21 between value-guided exploration on the Restless bandit and the questionnaire measure of Openness to new experiences; see Figure 4E). However, we observed strong correlations between task-based measures and model parameters on the one hand and working memory capacity on the other (highest r = 0.4). This suggests that measures from the bandit paradigms are affected by individual differences in cognitive capacity (see Collins & Frank[52] and Collins et al.[53]) .

Improving the Measurement of Exploration

The low replicability of the model parameters in the Horizon task, the overall mediocre reliabilities of the exploration strategies in all tasks, and the lack of convergent validity motivated us to perform three changes to the measurement of exploration.

First, we removed random exploration from the model in the Two-armed bandit. This change was motivated by the redundancy in the two strategies as indicated by the high correlation between their respective predictors (on average between .83 to .94 across trials 2-10 in the four conditions; see Fig. **??** in the SI), between the recovered parameters (-.79 and -.78; see Fig. 2), and between the eventual estimates of the two strategies (-.88; see Fig. 4). In addition, random exploration in the Two-armed bandit task likely captured value-guided variance given its positive correlation with value-guided exploration in the Restless bandit. One way to avoid these problems is to omit one of the two highly-correlated predictors[56].

Second, we only used the data from the long horizon condition to estimate value-guided exploration and directed exploration in the Horizon task. Exploration is defined as the change in the respective regression parameter from the short to the long horizon condition. This is problematic because it differs from the definition as the absolute value of the respective parameters in the other two bandit tasks. Difference measures are moreover known to have a stronger contamination by error variance[48,49,57], which reduces their reliability. As it has been hypothesized and observed that people explore more in the long horizon condition than in the short horizon condition[4,15,58], we exclusively focused on the first free choice from the former.

Third, we computed latent factors of the two remaining exploration strategies to remove unwanted sources of variance (i.e., task-specific variance, error variance, and variance due to parameter trade-offs induced during model fitting[34]) from the strategy estimates. For that we used a confirmatory factor analysis including correlations between residuals of parameters estimated from the same task to capture parameter trade-offs separately in the model[35]. That is, we explicitly tested whether the model parameters in the three tasks measured generalizable aspects of the two strategies.

These changes resulted in three desirable effects for the measurement of exploration (see Fig. 5). First, they improved parameter recovery in the Two-armed bandit by reducing the off-diagonal correlations to an acceptable level. Parameter recovery for the Horizon task, however, was on average unchanged. Second, the group-level exploration pattern in the Horizon task was qualitatively replicable across sessions. It is visible in Fig. 5 though, that participants barely used directed exploration on average. A similar pattern was visible in the Two-armed bandit when analyzing choices for each trial separately (see Fig. **??** in the SI). Hence, the increase in directed exploration from the short to the long horizon condition, as reported before, reflects rather a decreased tendency to avoid uncertainty than active exploration. Third, while the convergent validity of directed exploration stayed about the same, the one of value-guided exploration improved (see Figure 6B). Finally, the changes, on average, did not affect the reliabilities of the model parameters, as their positive and negative effects were roughly balanced. That is, the reliability of value-guided exploration in the Horizon task slightly increased, but the reliability of both strategies in the Two-armed bandit slightly decreased (see Figure 6A).

Although the correlations between directed exploration across bandits stayed about the same, those for value-guided exploration increased (see panel B in Fig. 6). This allowed us to evaluate *construct validity*. Therefore, we first created separate measurement models for working-memory capacity, value-guided exploration, and directed exploration on the data from session 1 using the lavaan package[59] in R[60]. These measurement models were just identifiable with df = 0. They also allowed us to quantify the contribution from each individual model parameter to the construct via the standardized factor loadings, which are shown in Table 1. Even though all three tasks contributed approximately equally to working-memory capacity and value-guided exploration, it is visible that directed exploration was mostly accounted for by the Restless bandit, to a much smaller degree by the Two-armed bandit, and only negligibly by the Horizon task. We then compared two structural models to test whether value-guided and directed exploration differ on the latent level. The first model included separate latent factors of value-guided exploration and directed exploration. The second model included only one general factor of exploration, with all six model parameters loading on this general factor. Because the first model converged with a negative variance on the session 1 data, but not on the session 2 data, we used the latter for the model comparison. The fit indices of both models are shown in Table **??** in the SI. We approximated the Bayes factor with the method described in[61]. The evidence was clearly in favor of the two-factor model, with a Bayes factor > 1000, supporting the theoretical claim that value-guided exploration and directed exploration are two separate strategies.

We also created latent factors based on participants' compound scores on each questionnaire. Here, we determined the

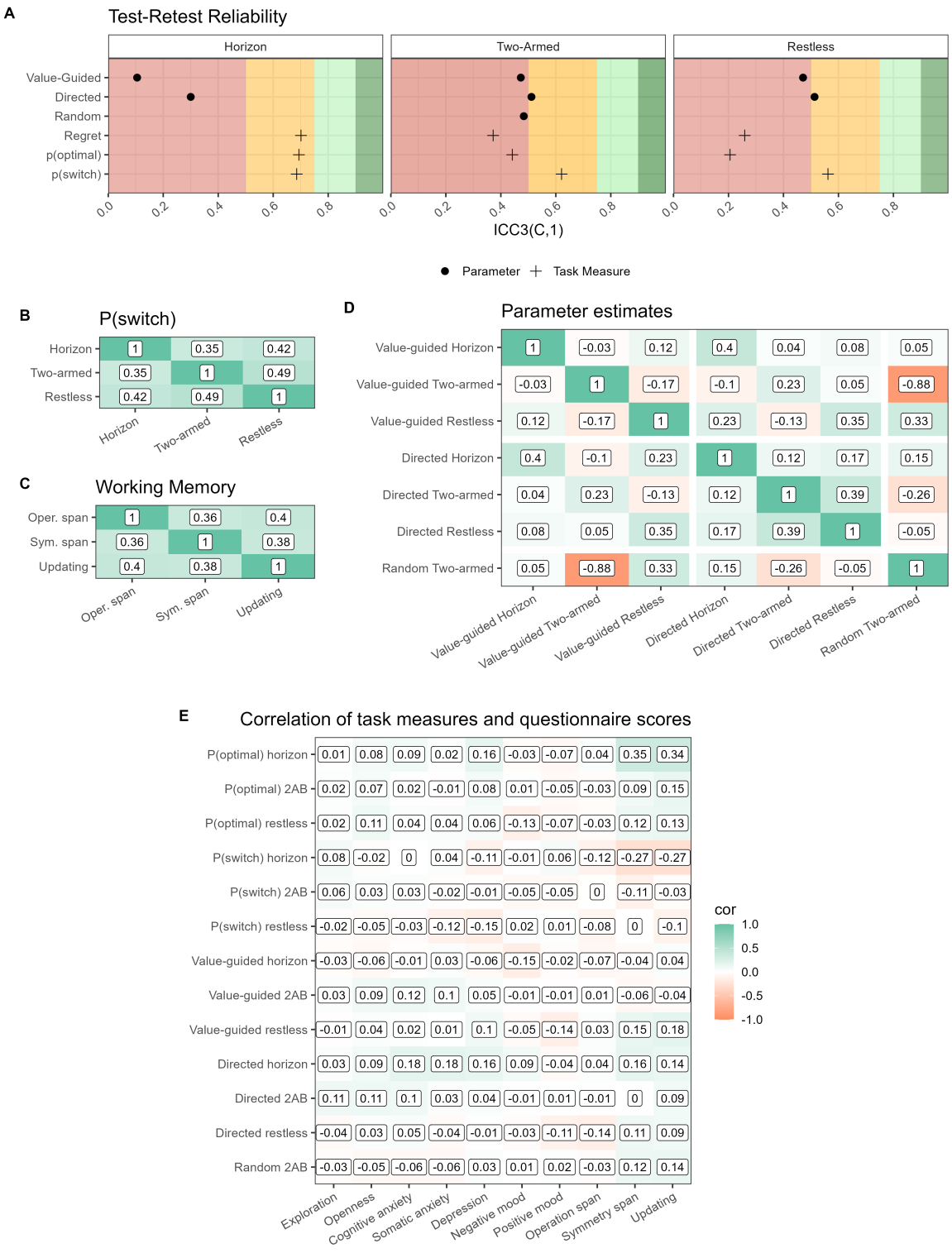

Figure 4. **A:** Test-retest reliability of model parameters and task measures. We calculated the reliabilities in the form of the respective intra-class correlation coefficient ICC(3,1)[54] as a measure of consistency[55]. The red, orange, light green and dark green background indicates bad, acceptable, good and very good reliability, respectively. Circles indicate the reliabilities of model-based parameters. Crosses indicate the reliabilities of model-free task measures. **B:** Convergent validity of the trial-by-trial switch probability. Correlations between participants' average proportion of trials where they selected a different option than on the previous trial across tasks. **C:** Convergent validity of the recall accuracy in the working memory tasks. **D:** Convergent validity of the model parameters in the bandit tasks. Correlations between subject-level model parameters across tasks. **E:** External validity of task measures and model parameters. Correlations of subject-level task measures as well as model parameters with all questionnaire scores and performance in all three working memory tasks.

factor structure by correlations between questionnaire scores. The final model was the only model that reached convergence with an acceptable model fit. The fit indices of that model can be found in Table **??** in the SI.

Model	Indicator	Standardized Loading
Working-Memory Capacity	Operation Span	0.625
Working-Memory Capacity	Symmetry Span	0.583
Working-Memory Capacity	Updating	0.647
Value-Guided	Horizon	0.597
Value-Guided	Two-armed	0.535
Value-Guided	Restless	0.696
Directed	Horizon	0.137
Directed	Two-armed	0.444
Directed	Restless	0.891

Table 1. Standardized factor loadings for the three measurement models of working-memory capacity, value-guided exploration, and directed exploration.

To evaluate *divergent validity* and *external validity*, we initially created a structural model including the two latent exploration factors and a working-memory capacity factor. We then correlated these factors with each other as well as with the two questionnaire measures of exploration. All correlations between the exploration factors and the questionnaire measures of exploration were small and statistically not significant (all ps > .05). Specifically, directed exploration correlated to 0.15 with the CEI scale and to -0.01 with the BIG5 openness subscale. Value-guided exploration correlated to 0.05 with the CEI scale and to -0.13 with BIG5 openness subscale. The two latent factors of exploration were however highly correlated with each other ($r = 0.64$, $p < .001$) and with the latent factor for working memory capacity ($r = 0.26$, $p < .001$ and $r = 0.47$, $p < .001$ for directed and value-guided exploration, respectively).

To further assess external and criterion validity, we created a structural model including a latent factor for self-reported exploration, two latent factors for positive and negative mood, respectively, and a latent factor for anxiety and depression based on participants' compound scores for each questionnaire subscale. This choice of factor structure was motivated by the similarity of the theoretical constructs as well as their strong observed correlation ($r = 0.57$, $p < .001$). The factors for positive and negative mood only consisted of the corresponding PANAS subscale [1] Lastly, combining depression, somatic anxiety and cognitive anxiety in one factor has both a strong theoretical basis in the computational psychiatry literature[62,63] and is motivated by the large correlations between these three measures (all $r > 0.64$, all $p < .001$). When investigating the correlations between the latent factors of task behavior and the latent factors based on the questionnaire scores, we found all correlations to be very small and again not statistically significant (all ps > .05), underscoring the absence of a link between the self-report measures and task behavior (see Figure 6C).

Discussion

In this study, we tested the reliability and validity of exploration strategies as inferred via computational model parameters in a large-scale individual differences study. Using existing models from the literature, we observed mixed results: While we achieved good parameter recovery, there were significant parameter trade-offs and weak signatures of directed exploration. Additionally, test-retest reliability was mediocre at best for all three tasks, and convergent and external validity were poor. To address these issues, we refined the measurements of exploration strategies by estimating only two strategies in the Two-Armed Bandit and only utilizing data from the long horizon in the Horizon task. Furthermore, we calculated latent factors to eliminate unwanted sources of variance, including task-specific, error, and model-fitting induced variances.

Following these improvements, test-retest reliability of single model parameters remained mediocre but we successfully extracted a temporally stable latent trait of value-guided exploration, with a reliability of 0.91. This finding suggests that value-guided exploration can be reliably measured over time. Moreover, we only identified a latent factor of directed exploration in the second session, but not in the first session, which may indicate that participants converged upon a strategy after sufficient exposure to the tasks. However, it also points towards difficulties in measuring a stable trait of directed exploration, and in using approximations of directed exploration using task-based parameters with single time point measurements. Compared to a recent study by Anvari et al.[30], which did not find converging evidence of exploration tendencies using model-free task measures, our study portrays a more positive picture. One possible reason for the discrepancy is that our selection of tasks was

[1] Note: No other confirmatory model combining these subscales with other scales in a latent factor converged.

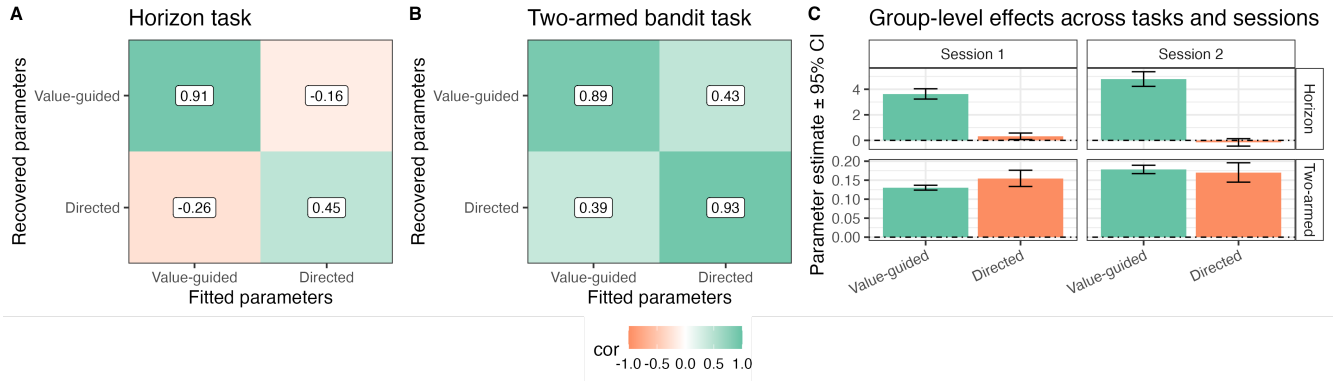

Figure 5. **A-B:** Parameter recovery following the improvements to the computational modeling, again using data from the first session. Numbers indicate the correlations between fitted and recovered subject-level parameters. **C:** Replicability of group-level effects from one session to the next following the improvements in the computational modeling. Note that the y axis is on a different scale for the different tasks.

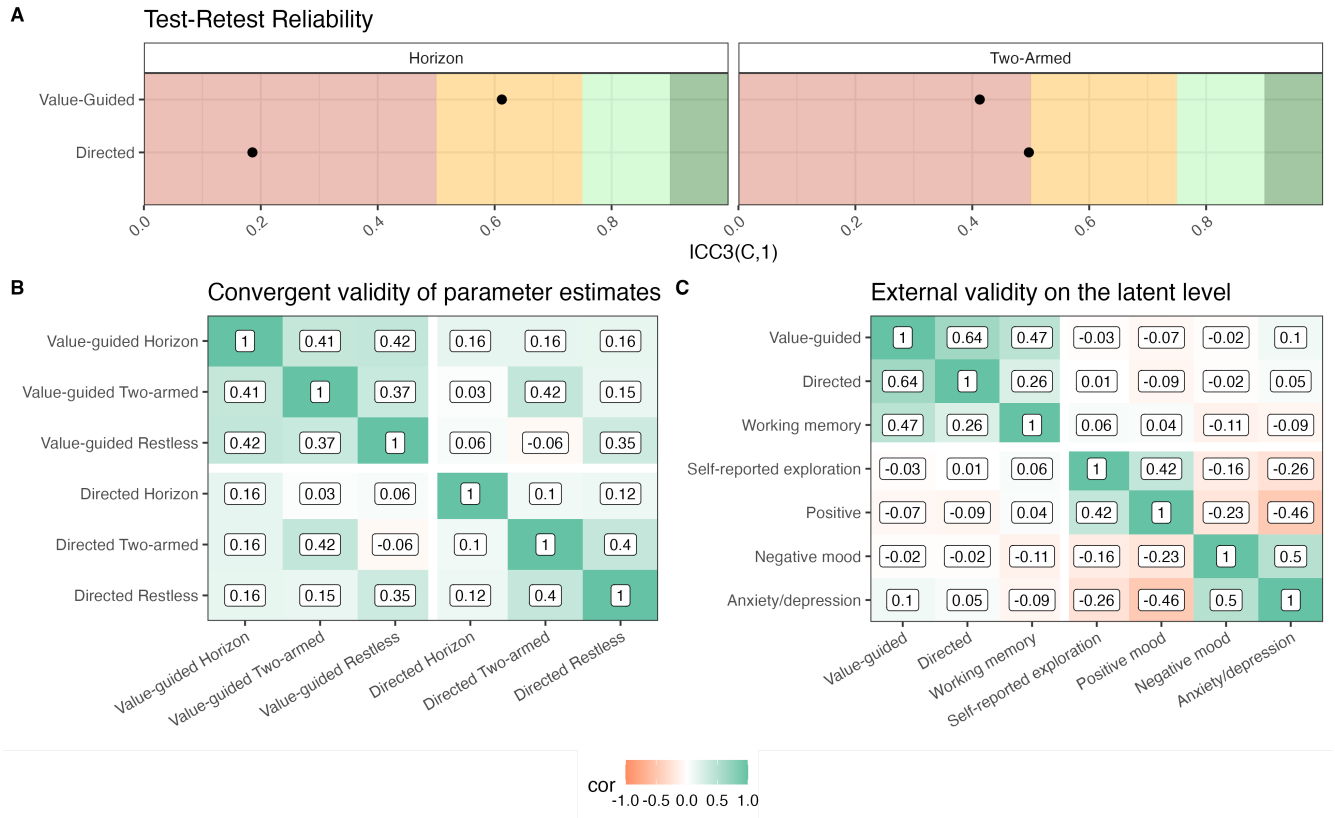

Figure 6. **A:** Test-retest reliability of model parameters from the Horizon task and the Two-armed bandit following our improvements to the modeling process. Red, orange, light green and dark green backgrounds indicate poor, acceptable, good and very good reliability, respectively. We again calculated the reliabilities in the form the respective intra-class correlation coefficient ICC(3,1)[54] as a measure of consistency[55] **B:** Convergent validity of the model parameters following improvements on the model fitting process. **C:** Correlations of latent factors from bandit tasks and questionnaires.

more constrained, focusing on the exploration-exploitation dilemma in few-armed bandits, while Anvari et al. also included other paradigms (e.g., optional stopping, decision from experience).

Notably, value-guided exploration was highly correlated with working memory capacity (r = 0.47). This is in line with the observation that participants whose working memory was taxed with a secondary task (i.e., lower capacity), showed decreased value-guided exploration on the horizon task[64]. This strong correlation with a general, performance-based cognitive construct, points towards general problems in the measurement of individual differences using behavioral tasks (for similar issues in the measurement of individual differences in category learning, see Lewandowsky[36] and Lewandowsky[37]). How much meaningful variability between people can we extract from existing behavioral cognitive paradigms, and few-armed bandits in particular? Moreover, none of the latent factors derived from the behavioral tasks were associated with the questionnaire scales or latent factors calculated from these scales. This lack of association suggests that the constructs measured by the behavioral tasks and questionnaires measure distinct aspects of individual differences between people. This observation adds to a growing body of research showing that behavioral tasks and questionnaires capture qualitatively different aspects of behavior[65], as has previously been shown in the domains of self-control[66,67], thought control[68] and risk[69].

Furthermore, the two exploration factors were highly correlated with each other (r = 0.64), highlighting potential issues with using these types of bandit tasks to distinguish between the two strategies. The strategies were however sufficiently distinct for the structural model with separate latent factors to be a better fit to the data than a model with just one exploration factor, providing evidence that the two are indeed distinct exploration strategies. A grain of salt on top of that interpretation in our study, however, is that people rarely actively engaged in directed exploration. In fact, the respective model parameters were negative in the Restless bandit, they turned negative after five choices in the Two-armed bandit (see Fig. **??** in the SI), and they were indistinguishable from zero on the first free choice in the Horizon task. This pattern is more in line with the active avoidance of uncertainty on most trials.

Given the consistently observed negative sign of directed exploration parameters, a critical question arises: Can directed exploration truly be estimated in these types of bandit tasks? One potential way forward is to use these tasks but limit the number of trials per round (e.g., stopping after five choices). This approach might help to isolate the effects of directed exploration. Another promising direction could be to focus more on tasks that de-correlate exploration from exploitation, such as the observe-or-bet task[70,71]. This type of task design could provide clearer insights into the distinct mechanisms of exploration. Additionally, more complex or gamified tasks might offer a better framework for studying exploration strategies[72,73]. However, as Allen et al.[72] have pointed out, the reliability and validity of these new tasks must be rigorously tested and cannot be assumed a priori.

Finally, it is worth considering whether humans necessarily use these principled algorithms. The history of these algorithms stems from artificial agents[74,75] that require exploration strategies to function effectively, for example, to avoid getting stuck in a local minimum. However, just because artificial agents need these kinds of strategies does not mean that humans need to use these exact strategies. Instead, they could rely on more heuristic approaches rather than explicitly computing their uncertainty about each option[13,32,76]. In fact, the difference between the rewards earned using the optimal strategy and a heuristic strategy is usually no more than a few cents. It therefore remains an open question whether participants in these experimental studies are simply not incentivized enough to seek maximum rewards[77], or whether these complex strategies are simply not a good representation of how people make (real-world) choices.

Conclusion

We measured value-guided and directed exploration using three distinct bandit tasks. Our findings suggest that using all three tasks provides a comprehensive assessment of value-guided exploration. However, if constraints limit the use of all tasks, the Restless bandit task is recommended due to its highest factor loadings observed in both session 1 and session 2. Despite our ability to identify a factor of directed exploration, the generally negative parameter values indicate that these bandit tasks may not be well-suited for studying directed exploration. The data suggest that participants tend to converge quickly on a preferred option, exhibiting minimal exploration behavior. This rapid convergence highlights a potential limitation in using these tasks to investigate directed exploration comprehensively.

Methods

Participants

We invited participants on the Prolific platform (prolific.com) who had completed at least 5 studies prior to participation, had an acceptance rate of at least 95%, were between the ages of 18 and 50, reported English as their primary & first language, and did not report any language-related disorders. Prior to participation, they all gave informed consent and were informed about the privacy policy and their payment. Payment consisted of a fixed portion (GBP 9 for each session) and a bonus portion (up to GBP 6 for each session) that rewarded performance on the behavioral tasks but not the questionnaires. The study was

performed in accordance with the relevant guidelines and regulations approved by the ethics committee of the University of Tuebingen (project nr. 202/2023B02, study title: Psychopathology and cognitive processes). Experiments were presented to participants using a combination of HTML, javaScript, CSS with custom code, and jsPsych[78]. For the working-memory tasks, we adapted code made available by Luthra & Todd[79].

In session 1, we collected data from 357 participants. We subsequently excluded participants who failed any of the following inclusion criteria: A full dataset, no use of external aids during the working memory or bandit tasks (as judged by self-report), bandit task performance better than the 95th percentile of chance performance as given by a binomial distribution, less comprehension attempts during the bandit task portion than the mean + 2SD (such as to exclude participants who were merely guessing the correct answers instead of reading the instructions), correctly answering both of the two attention check questions in the questionnaire portion and performing better than the 95th percentile of chance performance in the processing portion of the working memory tasks. On the basis of these exclusion criteria, we selected a total of 236 participants, whom we invited back for a second session. We subsequently applied the same criteria to the data we gathered from the second session, resulting in a final sample of 175 participants (84 females, mean age = 34.23, SD = 7.77).

Behavioral Paradigms and Questionnaires

General procedure

Both sessions followed the same sequence of tasks: participants started with the three working-memory tasks, which were presented in the same order for every participant, followed by the three bandit tasks, followed by the five questionnaires all presented on the same page. We presented the bandit tasks in randomized order across participants but fixed the order across sessions for every individual participant. The reasoning was the following: Order effects can impact the reliability and precision of our results[80] (for example, if the order effect is reflected in an increase of inattentive responding over time leading to a more noisy distribution in later tasks). Additionally, completely randomizing tasks within a single participant complicates the interpretation of factor loadings across sessions. To mitigate these issues, we randomized the order of bandit tasks across participants but kept the order fixed across sessions for each individual participant.

In the Horizon task and the Two-armed bandit, participants selected a slot machine by pressing S or K on the keyboard, in the Restless bandit, they selected a slot machine with S, D, K, or L. In the working-memory updating task and in the operation span task, participants recalled the memoranda by typing on the keyboard. In the symmetry span task, they recalled the locations by sequentially clicking with the mouse on the respective grid cells; in both processing parts, participants responded with the up (correct) and down arrow keys (incorrect) on the keyboard. The stimulus sets (memoranda and processing items in the working-memory tasks and rewards in the bandit tasks) were pre-sampled and the same for all participants. We decided upon the final set sizes in the working-memory tasks and all reward sets in the bandit tasks in an initial pilot study to avoid ceiling and floor effects.

Bandit tasks

Horizon task. We adapted the same general procedure as in the original publication[4] with the sole difference that the history of rewards was not available to participants. Participants played one practice round of the task which was followed by 80 task rounds. In each round, there was a message above the slot machines indicating whether they were playing a long or a short round. During the first four trials, participants had to select the slot machine that was highlighted for them. After these initial four forced choices, the slot machines briefly disappeared from the screen and there was a message reminding the participants whether they were about to make one or six free choices, depending on whether they were in a long or a short round.

When generating the reward set for the Horizon task, we fully crossed Horizon length and available information and made sure that each of the four resulting conditions had the same reward on average. We also ensured that both arms were on average, across the entire task, equally rewarding. Finally, we ensured that each combination of the Horizons and information conditions was equally difficult by assigning the same average reward differences to each combinations. Those differences were 30, 20, 12, 8, and 4, as these were the differences that avoided both ceiling and floor effects during piloting.

Two-armed bandit. We followed the procedure outlined by Fan et al.[10], which adapts the task originally proposed by Gershman[2]. Participants chose freely between two slot machines for 30 rounds, each consisting of 10 choices. The reward conditions varied: sometimes both arms had stable mean rewards, sometimes only one arm was stable, and sometimes both arms had mean rewards that drifted according to a random walk (for details, see Fan et al.[10]). For the drifting arms, we ensured that the average reward differences between the two arms did not exceed 15 over an entire round. Additionally, we biased the means of the generating reward distributions towards an 8-point difference, which proved effective in distinguishing between participants during piloting.

Restless bandit. In the Restless bandit task[3], participants could freely choose between four slot machines. They played one round that lasted for 200 choices. The rewards were sampled from four randomly walking mean rewards with standard deviation of 4. The mean rewards were sampled the same as in Daw et al.[3] according to $\mu_{i,t+1} = \lambda * \mu_{i,t} + (1-\lambda) * \theta + v_{innov}$, with lambda set to .9836, the decay center set to 50, with the only exception that the innovation variance / diffusion noise was set to

the lower value of 7.84 (sd = 2.8). We initially sampled several random walks and then finally selected two random walks such that there was variability in the which arm was the best arm across trials, and we made sure that each arm was never best for an extended period to motivate exploration.

Working Memory Tasks

Operation span and symmetry span. In both tasks, the presentation of memoranda was interleaved with the presentation of distracting processing problems. Memoranda were presented for 1000 ms, processing problems for maximally 6000 ms, but they were replaced immediately with the next memorandum after participants had responded to a given problem. Participants were instructed to recall the memoranda in order of presentation and to respond to the processing problems as accurately and as quickly as possible. In the operation span task, set size varied from 4 to 8, in the symmetry span task from 3 to 6. Every set size was tested twice in each session. We used a strict scoring scheme for calculating the proportion of correct responses. That is, only memoranda recalled in the serial position, in which they were presented, were scored as correct.

The memoranda in the operation span task were selected from all consonants (with the exception of J and Y), those in the symmetry span task were selected from the possible 16 locations in a 4x4 grid. Processing problems consisted of the validation of equations (additions and subtractions, the result always being a one-digit number) in the operation span task and of the judgment of symmetry of patterns along the vertical axis in an 8x8 grid in the symmetry task. The problems were constructed in the following way: half of the problems were correct, half incorrect. Incorrect equations in the operation span task were created by randomly adding or subtracting 1 or 2 to a correct equation. Asymmetric patterns were created by randomly changing 3 or 4 grid cells in one half of the grid.

Working-Memory Updating. In every trial, five digits (from 0 to 9) were presented in five adjacent, differently colored boxes in the middle of the screen for 5000 ms. In each of the following seven updating steps one of the five digits was replaced by a randomly sampled digit from the same set. The digit to be updated was presented for 1250 ms within the respectively colored box. We added an inter-stimulus interval of 250 ms in between presentation of the initial set and the first updating step as well as between the remaining updating steps. Participants were instructed to recall the five final digits in order of presentation. As for the other two working memory tasks, we used strict scoring. There were 20 updating trials in total. We added five trials without updating, in which participants were instructed to recall the initial set immediately to ensure correct encoding of the initial set (see[42]).

Questionnaires

We used two different questionnaires to investigate real-world exploration behavior: The Curiosity and Exploration Inventory (CEI)[43] and the Openness subscale from the Big Five Personality Inventory-2 (BFI-2)[44]. The CEI comprises two subscales, namely exploration and absorption. Here, we only used the former, which consists of four items to be rated on a 7-point Likert scale. The Openness subscale from the BFI-2 comprises six items, which are rated on a 5-point Likert scale.

In addition, we administered two psychiatric questionnaires to gain insight into the levels of anxiety and depression experienced by the participants. With regard to the former, we opted to utilize the trait version of the State Trait Inventory of Cognitive and Somatic Anxiety (STICSA)[45]. It comprises 21 items in total, 10 of which pertain to cognitive anxiety and 11 to somatic anxiety. Participants were invited to indicate their level of agreement with each item on a 4-point scale, ranging from "almost never" to "almost always". We assessed depressivity using the Patient Health Questionnaire Mood Subscale (PHQ-9)[46]. It consists of nine items, which are rated on a four-point scale ranging from "not at all" to "nearly every day".

Finally, we were interested in exploring whether behavior on the bandit tasks is influenced by transient fluctuations in mood. We therefore assessed positive and negative mood using the Positive Affect Negative Affect Scale (PANAS)[47]. This scale consists of 20 adjectives, 10 describing positive affect and 10 describing negative affect. Participants were invited to rate how much each adjective described the way they were feeling at that moment on a 5-point scale ranging from "very slightly or not at all" to "extremely".

Computational Modeling

We used the same computational models as proposed in the source studies. For the Horizon task, we used the model proposed by Wilson and colleagues[4] modeling only the first free choice in every round after the four forced choices.

As compared to the other two bandit tasks, learning was not implemented in an incremental fashion but only once for the fifth choice (i.e. the first free choice). The expected value of arm j, $E_j(t=5)$, was calculated as the average of the observed rewards during the forced choices. $I_j(t=5)$ was an indicator variable tracking the accumulated information about each arm j (i.e. how often it has been selected):

$$E_j(t=5) = \frac{\Sigma_{t=1}^{4} \delta_j(t) * r(t)}{\Sigma_{t=1}^{4} \delta_j(t)} \quad (1)$$

$$I_j(t=5) = \Sigma_{t=1}^{4} \delta_j(t) \quad (2)$$

where $r(t)$ is the reward received on trial t and $\delta_j(t)$ = 1 if arm j was chosen on trial t, and 0 otherwise.

In this approach, exploration is defined as the difference in the model parameters between the long horizon and the short horizon conditions. For value-guided exploration, that is reflected in the difference between the parameters relating $E_1(t=5)-E_2(t=5)$ to choice probability. For directed exploration, it is reflected in the difference between the parameters relating $I_1(t=5)-I_2(t=5)$ to choice probability. For the Two-armed bandit, we used the model proposed by Gershman[2], and for the Restless bandit, we used the model introduced by Daw and colleagues[3].

Accordingly, in both of those tasks, we used a Bayesian updating rule, i.e., the Kalman filter[81,82], to learn the expected value $E_j(t)$ and the variance $V_j(t)$ of the average reward on a given arm j on trial t

$$E_j(t) = E_j(t-1) + \delta_j(t) * K_j(t) * [r(t) - E_j(t-1)] \tag{3}$$

$$V_j(t) = [1 - \delta_j(t) * K_j(t)] * [V_j(t-1) + \sigma^2_{innov}] \tag{4}$$

where $r(t)$ is the reward received on trial t, σ^2_{innov} is the innovation variance, and $\delta_j(t)$ = 1 if arm j was chosen on trial t, and 0 otherwise. The learning rate $K_j(t)$, i.e., the Kalman gain, is calculated as follows:

$$K_j(t) = \frac{V_j(t-1) + \sigma^2_{innov}}{V_j(t-1) + \sigma^2_{innov} + \sigma^2_{noise}} \tag{5}$$

Note that σ_{innov}, σ_{noise}, $E_j(0)$, and $V_j(0)$ were set to the generating values and not fitted (see Danwitz et al.[83], Gershman[2], and Speekenbrink[84]). Then, considering the diffusion process, the priors $E_j(t+1)$ and $V_j(t+1)$ are updated in the following way before making the next decision:

$$E_j(t+1) = \lambda * E_j(t) + (1-\lambda) * C_{decay} \tag{6}$$

with λ and C_{decay} referring to the decay rate and the decay center of the diffusion process, respectively. We updated the priors according to the diffusion process only for the restless bandit (see Daw et al.[3]), but not for the Two-armed bandit (see Gershman[2]).

The choice model for the restless bandit was an upper-confidence bound policy:

$$P(C(t)=j) = \frac{exp(\tau * E_j(t) + \beta * sqrt(V_j(t) + \sigma^2_{innov}))}{\sum_{k=1}^{4} * exp(\tau * E_k(t) + \beta * sqrt(V_k(t) + \sigma^2_{innov}))} \tag{7}$$

The choice models in the Horizon task and in the Two-armed bandit were implemented via a logistic regression:

$$P(C(t)=j) = \frac{1}{1 + exp(-(\beta_0 + \boldsymbol{\beta} * \mathbf{x}))} \tag{8}$$

with $\mathbf{x}$ containing the independent variables $E_1(t)-E_2(t)$, $V_1(t)-V_2(t)$ (or $I_1(t)-I_2(t)$ in the case of the Horizon task), and $\boldsymbol{\beta}$ containing the respective parameters of the logistic regression. Note that β_0, the intercept of the regression, refers to a side bias, i.e., if someone preferentially chooses the left or the right arm without any further knowledge.

We implemented all the originally proposed models as hierarchical Bayesian models. We changed the softmax decision rule, which has shown to be the best account for human choices in the Restless bandit task[3,11], to an upper-confidence bound decision rule. These two adaptations neither changed the pattern of group-level effects nor the absolute parameter values of the individual effects as compared to the originally proposed models (see Figure **??** in the SI). The new implementation, however, led to three desired effects. First, it reduced the problem of parameter outliers, likely due to parameter trade-offs when fitting individual data. Second, the model parameters had a higher reliability. Third, it improved the models' predictive accuracy on held-out data(see Figure **??** in the SI).

To fit the logistic regression choice models in the Horizon task and the Two-armed bandit, we used the `brms` package[85] in R[86]. A comparison showed that neither the group-level nor the subject-level parameter estimates were changed in our hierarchical Bayesian implementation compared to the originally proposed subject-level implementation (see Wilson et al.[4], Figure **??** in the SI). However, both the reliability and the out-of-sample prediction improved (see Table **??** and **??**, respectively).

References

1. Mehlhorn, K. *et al.* Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* **2**, 191–215, DOI: 10.1037/dec0000033 (2015).

2. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42, DOI: 10.1016/j.cognition.2017.12.014 (2018).

3. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879, DOI: 10.1038/nature04766 (2006). Number: 7095 Publisher: Nature Publishing Group.

4. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074, DOI: 10.1037/a0038199 (2014). Publisher: US: American Psychological Association.

5. Banks, J., Olson, M. & Porter, D. An experimental analysis of the bandit problem. *Econ. Theory* **10**, 55–77, DOI: 10.1007/s001990050146 (1997).

6. Tomov, M. S., Truong, V. Q., Hundia, R. A. & Gershman, S. J. Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nat. communications* **11**, 2371 (2020).

7. Fan, H. *et al.* Pupil size encodes uncertainty during exploration. *J. cognitive neuroscience* **35**, 1508–1520 (2023).

8. Gershman, S. J. & Tzovaras, B. G. Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia* **120**, 97–104, DOI: 10.1016/j.neuropsychologia.2018.10.009 (2018).

9. Chierchia, G. *et al.* Confirmatory reinforcement learning changes with age during adolescence. *Dev. Sci.* **26**, e13330, DOI: 10.1111/desc.13330 (2023). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/desc.13330.

10. Fan, H., Gershman, S. J. & Phelps, E. A. Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. *Nat. Hum. Behav.* **7**, 102–113, DOI: 10.1038/s41562-022-01455-y (2023). Number: 1 Publisher: Nature Publishing Group.

11. Speekenbrink, M. & Konstantinidis, E. Uncertainty and Exploration in a Restless Bandit Problem. *Top. Cogn. Sci.* **7**, 351–367, DOI: 10.1111/tops.12145 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12145.

12. Navarro, D. J., Tran, P. & Baz, N. Aversion to Option Loss in a Restless Bandit Task. *Comput. Brain & Behav.* **1**, 151–164, DOI: 10.1007/s42113-018-0010-8 (2018).

13. Ferguson, T. D., Fyshe, A., White, A. & Krigolson, O. E. Humans Adopt Different Exploration Strategies Depending on the Environment. *Comput. Brain & Behav.* **6**, 671–696, DOI: 10.1007/s42113-023-00178-1 (2023).

14. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* **6**, e27430, DOI: 10.7554/eLife.27430 (2017). Publisher: eLife Sciences Publications, Ltd.

15. Dubois, M. *et al.* Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *eLife* **10**, e59907, DOI: 10.7554/eLife.59907 (2021).

16. Harms, M. B. *et al.* The structure and development of explore-exploit decision making. *Cogn. Psychol.* **150**, 101650 (2024).

17. Somerville, L. H. *et al.* Charting the expansion of strategic exploratory behavior during adolescence. *J. Exp. Psychol. Gen.* **146**, 155–164, DOI: 10.1037/xge0000250 (2017).

18. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120, DOI: 10.1073/pnas.2218523120 (2023). Publisher: Proceedings of the National Academy of Sciences.

19. Dubois, M. & Hauser, T. U. Value-free random exploration is linked to impulsivity. *Nat. Commun.* **13**, 4542, DOI: 10.1038/s41467-022-31918-9 (2022).

20. Speekenbrink, M. Chasing Unknown Bandits: Uncertainty Guidance in Learning and Decision Making. *Curr. Dir. Psychol. Sci.* **31**, 419–427, DOI: 10.1177/09637214221105051 (2022).

21. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14, DOI: 10.1016/j.conb.2018.11.003 (2019).

22. Blanco, N. J., Otto, A. R., Maddox, W. T., Beevers, C. G. & Love, B. C. The influence of depression symptoms on exploratory decision-making. *Cognition* **129**, 563–568, DOI: 10.1016/j.cognition.2013.08.018 (2013).

23. Aberg, K. C., Toren, I. & Paz, R. A neural and behavioral trade-off between value and uncertainty underlies exploratory decisions in normative anxiety. *Mol. Psychiatry* **27**, 1573–1587, DOI: 10.1038/s41380-021-01363-z (2022).

24. Smith, R. *et al.* Lower Levels of Directed Exploration and Reflective Thinking Are Associated With Greater Anxiety and Depression. *Front. Psychiatry* **12** (2022).

25. Mkrtchian, A., Valton, V. & Roiser, J. P. Reliability of Decision-Making and Reinforcement Learning Computational Parameters. *Comput. Psychiatry* **7**, DOI: 10.5334/cpsy.86 (2023).

26. Waltmann, M., Schlagenhauf, F. & Deserno, L. Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. *Behav. Res. Methods* DOI: 10.3758/s13428-021-01739-7 (2022).

27. Schurr, R., Reznik, D., Hillman, H., Bhui, R. & Gershman, S. J. Dynamic computational phenotyping of human cognition. *Nat. Hum. Behav.* **8**, 917–931, DOI: 10.1038/s41562-024-01814-x (2024). Publisher: Nature Publishing Group.

28. Loosen, A. M., Seow, T. X. & Hauser, T. U. Consistency within change: Evaluating the psychometric properties of a widely used predictive-inference task. *Behav. Res. Methods* 1–17 (2024).

29. Vrizzi, S., Najar, A., Lemogne, C., Palminteri, S. & Lebreton, M. Comparing the test-retest reliability of behavioral, computational and self-reported individual measures of reward and punishment sensitivity in relation to mental health symptoms. (2023).

30. Anvari, F., Billinger, S., Analytis, P. P., Franco, V. R. & Marchiori, D. Testing the convergent validity, domain generality, and temporal stability of selected measures of people's tendency to explore. DOI: 10.31234/osf.io/jnuwz (2023). Publisher: OSF.

31. Jach, H. K. *et al.* Individual differences in information demand have a low dimensional structure predicted by some curiosity traits. *Proc. Natl. Acad. Sci.* **121**, e2415236121 (2024).

32. Witte, K., Wise, T., Huys, Q. J. & Schulz, E. Exploring the unexplored: Worry as a catalyst for exploratory behavior in anxiety and depression. (2024).

33. Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O. & Schulze, R. Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence* **30**, 261–288, DOI: 10.1016/S0160-2896(01)00100-3 (2002).

34. Ratcliff, R. & Tuerlinckx, F. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychon. Bull. & Rev.* **9**, 438–481, DOI: 10.3758/BF03196302 (2002).

35. Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M. & Wittmann, W. W. Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *J. Exp. Psychol. Gen.* **136**, 414, DOI: 10.1037/0096-3445.136.3.414 (2007). Publisher: US: American Psychological Association.

36. Lewandowsky, S. Working memory capacity and categorization: Individual differences and modeling. *J. Exp. Psychol. Learn. Mem. Cogn.* **37**, 720–738, DOI: 10.1037/a0022639 (2011). Place: US Publisher: American Psychological Association.

37. Lewandowsky, S., Yang, L.-X., Newell, B. R. & Kalish, M. L. Working memory does not dissociate between different perceptual categorization tasks. *J. Exp. Psychol. Learn. Mem. Cogn.* **38**, 881–904, DOI: 10.1037/a0027298 (2012).

38. Bringmann, L. F., Elmer, T. & Eronen, M. I. Back to Basics: The Importance of Conceptual Clarification in Psychological Science. *Curr. Dir. Psychol. Sci.* **31**, 340–346, DOI: 10.1177/09637214221096485 (2022). Publisher: SAGE Publications Inc.

39. Wulff, D. U. & Mata, R. Using embeddings to automate jingle–jangle detection and tackle taxonomic incommensurability, DOI: 10.31234/osf.io/9h7aw (2023).

40. Turner, M. L. & Engle, R. W. Is working memory capacity task dependent? *J. Mem. Lang.* **28**, 127–154, DOI: 10.1016/0749-596X(89)90040-5 (1989).

41. Oswald, F. L., McAbee, S. T., Redick, T. S. & Hambrick, D. Z. The development of a short domain-general measure of working memory capacity. *Behav. Res. Methods* **47**, 1343–1355, DOI: 10.3758/s13428-014-0543-2 (2015).

42. von Bastian, C. C., Souza, A. S. & Gade, M. No evidence for bilingual cognitive advantages: A test of four hypotheses. *J. Exp. Psychol. Gen.* **145**, 246–258, DOI: 10.1037/xge0000120 (2016). Place: US Publisher: American Psychological Association.

43. Kashdan, T. B., Rose, P. & Fincham, F. D. Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *J. personality assessment* **82**, 291–305 (2004).

44. Soto, C. J. & John, O. P. Short and extra-short forms of the big five inventory–2: The bfi-2-s and bfi-2-xs. *J. Res. Pers.* **68**, 69–81 (2017).

45. Ree, M. J., French, D., MacLeod, C. & Locke, V. Distinguishing cognitive and somatic dimensions of state and trait anxiety: Development and validation of the state-trait inventory for cognitive and somatic anxiety (sticsa). *Behav. Cogn. Psychother.* **36**, 313–332 (2008).

46. Kroenke, K., Spitzer, R. L. & Williams, J. B. The phq-9: validity of a brief depression severity measure. *J. general internal medicine* **16**, 606–613 (2001).

47. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: the panas scales. *J. personality social psychology* **54**, 1063 (1988).

48. Salthouse, T. A. & Hedden, T. Interpreting Reaction Time Measures in Between-Group Comparisons. *J. Clin. Exp. Neuropsychol.* **24**, 858–872, DOI: 10.1076/jcen.24.7.858.8392 (2002). Publisher: Routledge _eprint: https://doi.org/10.1076/jcen.24.7.858.8392.

49. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186, DOI: 10.3758/s13428-017-0935-1 (2018).

50. Zorowitz, S. & Niv, Y. Improving the reliability of cognitive task measures: A narrative review. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging* **8**, 789–797 (2023).

51. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Medicine* **15**, 155–163, DOI: 10.1016/j.jcm.2016.02.012 (2016).

52. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035, DOI: 10.1111/j.1460-9568.2011.07980.x (2012). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-9568.2011.07980.x.

53. Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A. & Frank, M. J. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *The J. Neurosci.* **34**, 13747–13756, DOI: 10.1523/JNEUROSCI.0989-14.2014 (2014).

54. Shrout, P. E. & Fleiss, J. L. Intraclass Correlations : Uses in Assessing Rater Reliability. *Psychol. Bull.* **86**, 420–428 (1979).

55. McGraw, K. O. & Wong, S. P. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychol. Methods* **1**, 30–46 (1996).

56. Daoud, J. I. Multicollinearity and Regression Analysis. *J. Physics: Conf. Ser.* **949**, 012009, DOI: 10.1088/1742-6596/949/1/012009 (2017).

57. Enkavi, A. Z. *et al.* Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proc. Natl. Acad. Sci.* **116**, 5472–5477, DOI: 10.1073/pnas.1818430116 (2019).

58. Dubois, M. & Hauser, T. U. Value-free random exploration is linked to impulsivity. *Nat. Commun.* **13**, 4542, DOI: 10.1038/s41467-022-31918-9 (2022). Number: 1 Publisher: Nature Publishing Group.

59. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **48**, 1–36, DOI: 10.18637/jss.v048.i02 (2012).

60. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2024).

61. Wagenmakers, E.-J. A practical solution to the pervasive problems ofp values. *Psychon. Bull. & Rev.* **14**, 779–804, DOI: 10.3758/BF03194105 (2007). Number: 5.

62. Wise, T., Robinson, O. J. & Gillan, C. M. Identifying transdiagnostic mechanisms in mental health using computational factor modeling. *Biol. Psychiatry* **93**, 690–703 (2023).

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

64. Cogliati Dezza, I., Cleeremans, A. & Alexander, W. Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. Exp. Psychol. Gen.* **148**, 977, DOI: 10.1037/xge0000546 (2019). Publisher: US: American Psychological Association.

65. Dang, J., King, K. M. & Inzlicht, M. Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends cognitive sciences* **24**, 267, DOI: 10.1016/j.tics.2020.01.007 (2020).

66. Duckworth, A. L. & Kern, M. L. A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* **45**, 259–268, DOI: 10.1016/j.jrp.2011.02.004 (2011).

67. Moffitt, T. E. *et al.* A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* **108**, 2693–2698, DOI: 10.1073/pnas.1010076108 (2011).

68. Göbel, K., Hensel, L., Schultheiss, O. C. & Niessen, C. Meta-analytic evidence shows no relationship between task-based and self-report measures of thought control. *Appl. Cogn. Psychol.* **36**, 659–672, DOI: 10.1002/acp.3952 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3952.

69. Frey, R., Pedroni, A., Mata, R., Rieskamp, J. & Hertwig, R. Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* **3**, e1701381, DOI: 10.1126/sciadv.1701381 (2017).

70. Tversky, A. & Edwards, W. Information versus reward in binary choices. *J. Exp. Psychol.* **71**, 680, DOI: 10.1037/h0023123 (1966). Publisher: US: American Psychological Association.

71. Navarro, D. J., Newell, B. R. & Schulze, C. Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cogn. Psychol.* **85**, 43–77, DOI: 10.1016/j.cogpsych.2016.01.001 (2016).

72. Allen, K. *et al.* Using games to understand the mind. *Nat. Hum. Behav.* **8**, 1035–1043, DOI: 10.1038/s41562-024-01878-9 (2024). Publisher: Nature Publishing Group.

73. Donegan, K. R. *et al.* Using smartphones to optimise and scale-up the assessment of model-based planning. *Commun. Psychol.* **1**, 31 (2023).

74. Kaelbling, L. P. *Learning in embedded systems* (MIT press, 1993).

75. Lai, T. L. & Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. applied mathematics* **6**, 4–22 (1985).

76. Thalmann, M. & Schulz, E. Simple, Idiosyncratic Decision Heuristics in a Two-Armed Bandit Task. In *2023 Conference on Cognitive Computational Neuroscience*, DOI: 10.32470/CCN.2023.1240-0 (Cognitive Computational Neuroscience, Oxford, UK, 2023).

77. Zorowitz, S., Solis, J., Niv, Y. & Bennett, D. Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nat. human behaviour* **7**, 1667–1681 (2023).

78. De Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* **47**, 1–12, DOI: 10.3758/s13428-014-0458-y (2015).

79. Luthra, M. & Todd, P. M. Role of Working Memory on Strategy Use in the Probability Learning Task. *Proc. Annu. Meet. Cogn. Sci. Soc.* **41** (2019).

80. Goodhew, S. C. & Edwards, M. Translating experimental paradigms into individual-differences research: Contributions, challenges, and practical recommendations. *Conscious. Cogn.* **69**, 14–25 (2019).

81. Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **82**, 35–45, DOI: 10.1115/1.3662552 (1960).

82. Kalman, R. E. & Bucy, R. S. New Results in Linear Filtering and Prediction Theory. *J. Basic Eng.* **83**, 95–108, DOI: 10.1115/1.3658902 (1961).

83. Danwitz, L., Mathar, D., Smith, E., Tuzsus, D. & Peters, J. Parameter and Model Recovery of Reinforcement Learning Models for Restless Bandit Problems. *Comput. Brain & Behav.* DOI: 10.1007/s42113-022-00139-0 (2022).

84. Speekenbrink, M. Identifiability of Gaussian Bayesian bandit models. In *2019 Conference on Cognitive Computational Neuroscience*, DOI: 10.32470/CCN.2019.1335-0 (Cognitive Computational Neuroscience, Berlin, Germany, 2019).

85. Bürkner, P.-C. Brms: An R package for bayesian multilevel models using stan. *J. Stat. Softw.* **80**, DOI: 10.18637/jss.v080.i01 (2017).

86. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2020).

Acknowledgments

This work was supported by the Max Planck Society, Helmholtz Munich, a Jacobs Research Fellowship, and an ERC Starting Grant to E.S. We thank Alexander D. Kipnis for helpful discussions, and Marcel Binz for valuable feedback on a previous version of this manuscript.

Author contributions statement

All authors conceived the experiments, K.W. and M.T. conducted the experiments, K.W. and M.T. analyzed the results. All authors reviewed the manuscript.