Predictors of risky foraging behaviour in healthy young people

Dominik R. Bach[1,2¶*], Michael Moutoussis[1¶], Aislinn Bowler[1], Neuroscience in Psychiatry Network consortium[3], Raymond J. Dolan[1]

[1]Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, and Wellcome Centre for Human Neuroimaging, University College London, London WC1 3BG, UK

[2]Computational Psychiatry Research; Department of Psychiatry, Psychotherapy, and Psychosomatics; Psychiatric Hospital; University of Zurich, 8032 Zurich, Switzerland

[3] please see Supplementary Table 1.

¶ Joint first authors

* Corresponding author: d.bach@ucl.ac.uk; Postal address: 10-12 Russell Square, London WC1N 5BH, UK,

Abstract

During adolescence and early adulthood, learning when to avoid threats and when to pursue rewards, becomes crucial. Using a risky foraging task, we investigated individual differences in this dynamic across 781 individuals aged 14 to 24, which were split into a hypothesis-generating discovery sample and a hold-out. Sex was the most important predictor of cautious behaviour and performance. Males earned one standard deviation, or 20%, more reward than females, collected more reward when there was little to lose, and reduced foraging to the same level as females when potential losses became high. Other independent predictors of cautiousness and performance were self-reported daringness, IQ, and self-reported cognitive complexity. We found no evidence for an impact of age or maturation. Thus maleness, a high IQ or self-reported cognitive complexity, and self- reported daringness, predicted greater success in risky foraging, possibly due to better exploitation of low-risk opportunities in high-risk environments.

Introduction

Arbitrating risk and benefits is a challenge for all animals, including humans. Animals foraging for nutrition are often faced with fertile open spaces that expose them to potentially fatal predation [1]. Many human situations echo this scenario. For example, driving entails an exposure to mortal dangers, but risk-taking taxi drivers earn more money on average [2]. Adolescence and early adulthood is a critical period during which an ability to balance cautiousness versus daring emerges as a character trait [3], but is also associated with miscalculations that lead to harms such as traffic accidents [4], unwanted pregnancy, and substance-related morbidity [5]. However, this does not affect all adolescents to the same extent. For example, sex differences in adolescent risk-taking are important [5].

Yet, there is a debate in relation to the cognitive and neurobiological mechanisms of youth risk taking and its predictors [6, 7, 8]. This is likely, at least in part, to reflect a relative lack of suitable laboratory tasks to measure behavioural risk-taking. Adolescent risk miscalculations typically involve emotionally arousing and extended sequences of events with real consequences [3], where the latter are often not fully known [9]. By contrast, laboratory tasks assessing risk taking in adolescents typically do not involve these features and fall into two broad classes [10]: hypothetical or real monetary decisions (e.g. [11]), usually involving economic lotteries (e.g. [12]), and game-like tasks with intuitive cover stories such as the "balloon task" [13]. In economic revealed preference tasks, all stakes and outcomes are fully described, unlike what pertains in real-life scenarios [9]. Risk-seeking in these tasks appears to monotonically decrease from childhood into adulthood, and this fails to capture a reported characteristic mid-adolescence peak in real-life risk taking [14]. When lotteries involve ambiguity or uncertainty [15, 16], adolescents avoid outcomes of unknown probability to a lesser extent, and invest less in information searching, compared to children or adults [9]. A further discrepancy with real-life risk taking is the relative insignificance of potential (petty financial) outcomes in these tasks, which tends to obviate an element of emotional arousal [3]. This also applies to the often-used, game-like, virtual Balloon task where the worst outcome in the cover story is bursting a party balloon [13].

Here, we sought to address these shortcomings in order to identify antecedents of individual differences in adolescent risk taking. To this end, we drew on ethological ideas of risky foraging for reward under threat of predation [17, 18, 19]. Risky foraging is a common biological scenario, which is extended over time, and affords detection of relatively low-risk opportunities in high-risk environments [20, 21]. We previously described a human risky foraging task, resembling rodent approach/avoidance conflict tests, which necessitates multiple sequential decisions and exploration of stakes and probabilities in real time. Within an ethological framing, which includes virtual 'death', we showed that cautious behaviour relies on similar neural circuits to that of rodent tasks where harm is looming [22, 23, 24, 25, 26]. This task allows, in principle, a comparison to real-life scenarios, as well as an assessment of the "dynamic flow of decision-making" [27].

In this task, players weigh potential gains against two potential threat features. The first is the presence of one or other task scenario associated with a different overall threat probability, where these are learned by experience and signalled by frame colour. The second threat feature relates to foraging being extended over time. Early on, there is little to lose by getting caught, and participants typically forage vigorously, but reduce such foraging with the passage of time, as they accumulate gains. We have previously shown that this latter behavioural change over time is impacted by a range of anxiolytics as well as by hippocampal lesions, with no consistent effect reported in the case of threat probability [23, 24, 25]. In a related task, we have provided neuroimaging and lesion evidence suggesting that distinct neural substrates account for a representation of threat probability and loss magnitude respectively [22, 26].

Here, we exploited a large dataset, involving an accelerated longitudinal design that spanned a limited age range (14-24 years), to probe how individual differences between adolescents and young adults, including sex, age, IQ, and a large variety of mental health measures, account for success in risky foraging. More specifically, we were interested in examining how these factors shape sensitivity to the presence of potential threat, differences in threat probability, and the passage of intra-epoch time. Importantly, the task provides a rich set of behavioural measures. By splitting our data into discovery and confirmation samples, we exploit these measures in an entirely data-driven way and form hypotheses that were first pre-registered before confirmation in the hold-out sample.

Results

Analysis strategy

N = 781 adolescents, drawn from the Neuroscience in Psychiatry Network (NSPN) 2400 cohort [28], completed 81 epochs (i.e. trials) of a pac-man style computer game in which they foraged for tokens under threat of virtual predation (see figure 1). The data curator (MM) randomised these data into a sex- and age-balanced discovery sample (N = 492) and a hold-out confirmation sample (N = 289, table 1). The data analyst (DRB) had access to the discovery sample alone. This sample was used in an exploratory analysis to derive 9 distinct pre-registered hypotheses (https://osf.io/hrce6/, registered on 28.07.2018, table 2) prior to accessing data for the confirmation sample. We then tested these hypotheses in the hold-out sample without parameter re-fitting. We report out-of-sample predictive performance (i.e. with parameters fixed from the discovery sample) from these tests, as well as in-sample predictive performance (i.e. with parameters fitted in-sample) for the combined sample [29]. To put our findings into context, we report effect sizes (but no inference statistics) from post-hoc exploratory analyses.

During data exploration in the discovery sample, we considered 38 task variables (see Supplementary Results 1 for a detailed list). These included 4 summary statistics (see Extended Data Figure 1 for their derivation) for each of 7 previously validated, (intra-epoch) time-dependent variables and their weighted "cautiousness" sum score [25], and 6 additional time-independent measures including overall performance (tokens retained after the epoch had finished), all averaged over the 81 epochs of the game. We analysed how these 38 task variables related to 32 predictor variables. The latter included sex, age, IQ, and self-report questionnaire data that covered a broad range of mental health symptoms and dispositions (see methods). All bivariate linear or quadratic relationships discovered in these 2432 tests formed part of our hypotheses summarised in table 2; any relationship not mentioned therein was not significant at an alpha level of $p < .001$ in the discovery sample.

Task properties

Test-retest reliability over two years for those task variables included in the confirmatory analysis exceeded $r_{tt} > .5$ (see Supplementary Results 1 for test-retest variability of all task variables). These 38 variables were correlated (see Supplementary Results 1) and can be conceptually characterised as contributing to at least three interpretable and replicable factors; namely sensitivity to threat probability, sensitivity to intra-epoch time, and performance achieved (see Supplementary Results 1 for factor analysis). However, the factor scores explained less variance in predictor variables than some of the individual task variables. Therefore, in what follows we focused on the latter.

Task performance, ie. tokens retained after the predator wakes up, stands out among task variables. In a post-hoc analysis across the combined sample, average token collection during the epoch was the best predictor of performance (76.9% explained variance), decrease in wall distance over the epoch was the best predictor of remaining variance (12.4% explained variance), and average wall distance was the best predictor of still remaining variance (1.1% explained variance). At the suggestion of a reviewer, we investigated average survival rates. None of the 32 predictors explained more than 1.6% variance in survival rates in the discovery sample. Among task variables, survival was best explained by the average cautiousness sum score (67.1% explained variance in combined sample); increase in safe quadrant presence as the epoch progressed explained most of the remaining variance (3.8% explained variance); and average wall distance explained most of still remaining variance

(1.7% explained variance). Another variable considered during the revision was average latency to initiate escape (defined as first movement away from the threat) once the predator woke up, which took on values between 100 ms (minimum latency defined by the maximum attainable speed) and 500 ms, explaining 1.4% variance in raw survival rate. Notably, the predator moved at 40 grid movements per second, while the maximum attainable speed of the human player was 10 grid movements per second. Thus, the player's sensorimotor performance during escape was by design less decisive for survival than the player's location on the grid when the predator woke up.

Sex was the strongest predictor of behavioural differences in the task

Amongst the 32 predictor variables, sex was the single best predictor of performance (tokens retained after the predator woke up), explaining 17.0% variance in this metric in the combined sample. Males earned 19.9%, or approximately 1 standard deviation, more tokens than females (see figure 2B, table 2, Extended Data Figure 2, Supplementary Table 2). In the discovery sample, sex was associated with 7 task variables (including task performance) at an alpha level of $p < .001$. Males collected more tokens per time unit, moved faster (see figure 2A), and approached closer to predator location (see figure 2B). Over time, they also decreased their distance to walls, token collection rate and overall speed, more rapidly than females. To collectively confirm these associations, we computed a multiple logistic regression model that predicted sex from these 7 task variables (H1, table 2). This model was confirmed without refitting, using a random permutation test in the hold-out sample.

Descriptively, male behaviour for most measures converged to that of females towards the end of an epoch, when they had collected a number of tokens, and the cost of predation was getting high. Consequently, post-hoc analyses across the entire sample revealed no evidence that males were less successful in avoiding a predator (sex difference in average survival rate 0.52%; -1.81%-2.84% 95% parametric confidence interval; LBF = 0.90 in favour of a model without sex difference). Figure 2C shows occupancy heat maps for the peak token collection period and illustrates how females, on average, maintained a closer proximity to the safe place and to the walls than did males.

Next, we asked if any of the other 6 task variables relating to sex mediated the observed performance difference between males and females (see figure 2E). In a post-hoc mediation analysis [30], over the combined sample, average token collection (82%) and decrease in token collection over the epoch (13%) mediated the largest proportion of the performance difference. For detailed results and further mediation analyses, see Supplementary Results 2.

Sex differences in the time people had spent with computer games in their daily lives could, in principle, affect performance in the present task. If, for example, males spent more play time and already performed better than females, then this could explain our results. In this case, one would expect a steeper performance increase over trials in female participants. To test this hypothesis in a pre-registered analysis, we analysed the epoch-by-epoch performance trajectory (H2, table 2). In our discovery sample, we found that males and females started at the same performance level, but males increased their performance more steeply over repeated epochs than females (see Supplementary Results 3). This is the opposite of what one would expect if females' overall worse performance were explained by less experience with computer games. However, this sex difference was not confirmed in the hold-out sample (see table 2 and Supplementary Results 3). Nevertheless, we did not find any evidence for a steeper epoch-by-epoch performance trajectory in females in the confirmation sample. Also, the individual slope of the epoch-by-epoch performance trajectory mediated only a negligible proportion (0%) of the sex effect on performance (see Supplementary Results 2).

Because of a strong effect of sex, we controlled for sex in all further analyses. All significant relationships reported in what follows remained significant when sex was modelled as covariate.

Self-reported daringness, IQ, and cognitive complexity predict better performance

We next investigated the relationship between 29 self-report measures, IQ, and age, and task variables. Self-reported daringness, measured with the CADS questionnaire, explained 3.9% of performance variance in the combined sample. In the discovery sample, participants with higher self-reported daringness collected, and retained, more tokens, and decreased their token collection more rapidly with the progression of time over an epoch (H4, see figure 3AC, table 2, Supplementary Table 2, Extended Data Figure 3). These associations were collectively confirmed in the hold-out.

In a post-hoc mediation analysis, over the combined sample, average token collection mediated 87% of the CADS effect on performance. For detailed results and further mediation analyses, see Supplementary Results 2.

Because daringness related to 3 summary-statistics of time-dependent measures, we assessed in a pre-registered analysis (H7, see table 2) whether daringness predicted continuous intra-epoch trajectories of these, or other, time-dependent measures. Thus, we computed the average intra-epoch trajectory for the 20 individuals with highest self-reported daringness in the discovery sample, for each of the 7 time-dependent measures (see figure 3B). We then projected each individual's trajectory onto this trajectory. This quantifies the extent to which an individual pursues the strategy that high CADS daringness scorers use (note that in this approach, an individual can use this strategy to a higher extent that the highest CADS daringness scorers). In the discovery sample, this metric was predicted by self-reported daringness for distance from walls, token collection rate, and speed on grid (see figure 3BC, table 2, Supplementary Table 2). These associations were collectively confirmed in the hold-out.

The next best predictor of performance was IQ, as measured with WASI-I, which explained 3.8% performance variance across the combined sample. In the discovery sample, participants with higher IQ decreased their token collection more as time progressed during an epoch, and retained more tokens (H3, see figure 4AB, table 2, Supplementary Table 2, Extended Data Figure 4). These associations were collectively confirmed in the hold-out.

In a post-hoc mediation analysis, over the combined sample, decrease in token collection mediated 80% of the IQ effect on performance. For detailed results and further mediation analyses, see Supplementary Results 2.

Finally, self-reported cognitive complexity as measured by the BIS explained 2.2% of the variance in performance across the combined sample. In the discovery sample, individuals with higher questionnaire scores (corresponding to lower cognitive complexity) reduced their token collection rate less over time, and retained fewer tokens (H6, see figure 4CD, table 2, Supplementary Table 2, Extended Data Figure 4). These associations were collectively confirmed in the hold-out.

In a post-hoc mediation analysis, over the combined sample, a decrease in token collection mediated 94% of the cognitive complexity effect on performance. For detailed results and further mediation analyses, see Supplementary Results 2.

Three associations found in the discovery sample were not confirmed in the hold-out (H5/8/9, see table 2). Furthermore, none of the underlying bivariate relationships replicated in the confirmation sample, even when we did not correct for multiple comparison (see Supplementary Table 2). In particular, the quadratic bivariate relationship of the anxiety questionnaire RCMAS with task variables was not confirmed (H8). Notably, there was also no linear relationship of self-reported anxiety with any task measure in the discovery sample at our alpha level. However, Bayes Factors were not strongly in favour of a model without RCMAS for any task measure (|LBF| < 3), such that we cannot firmly rule out a true effect of self-reported anxiety on task variables.

Absence of evidence for an impact of age or maturation on behaviour

Surprisingly in our cross-sectional analysis, age did not predict any task variable, including survival rates, either in a linear or quadratic manner, or when splitting the discovery sample as a function of sex. There was no age by sex interaction for any of the task variables. However, Bayes Factors were not strongly in favour of a model without age for any analysis (|LBF| < 3), such that we cannot rule out a population effect of age. All of these results replicated in the confirmation sample.

Our accelerated longitudinal design also enabled us to ask whether passage of time, as index of maturation at this age, had an impact on task measures. In a subsample of n = 63 participants, distributed across discovery and confirmation sample, who returned 6 months after the first visit (BSL) and played the game again (visit FU-R), the pattern of changes between BSL/FU-R and FU-R/FU-1 suggested an absence, or insignificant impact, of maturation (see Supplementary Results 4). Furthermore, in the larger number of participants (N = 567) who took part in BSL and FU-1 (after 11-32 months), but not necessarily in FU-R, any behavioural change between the two assessments was best explained by repetition and not by the time elapsed, for discovery, confirmation and combined sample, as evidenced by Bayesian model comparison (all LBF > 3 in favour of the simpler model without time, see Supplementary Results 4). The impact of age at BSL on the impact of repetition of the task is reported in Supplementary Results 4.

Relation of risk aversion and related economic preferences to risky foraging task

In a final post-hoc analysis, we examined how a pure economic risk aversion measure related to behavioural indices of cautiousness in our task, harnessing an economic risk preference paradigm in which participants made a choice between a sure amount and a lottery (see Extended Data Figure 5). A propensity to choosing the lottery was parameterised in a variant of an economic risk-return model [31], specifically the mean-variance-skewness (MVS) model [32, 33]. This conceptualises choice as logistic function of the difference between the certain amount, and a weighted sum of the lottery's expected value, variance, and skewness. The relation between economic task parameters on the one hand, and the 7 task variables that related significantly to predictor variables on the other, is reported in Supplementary Results 5. None of these relations exceeded 1.7% explained variance. In keeping with previous research, we found that males were less averse to increasingly variable gambles than females (Cohen's d = 0.28; 0.13-0.42, 95% parametric confidence interval) without a pronounced difference in skewness preference (Cohen's d = -0.08; -0.20-0.07). Aversion to variable gambles explained 1.9% of the sex effect on performance (proportion of mediation 2%; -0.1%-5.0%, 95% bootstrap confidence interval), preference for skewed gambles explained 0.1% (-0.4%-1.0%), and choice temperature mediated 2.1% (0.3%-5.0%).

Discussion

In this paper, we investigated antecedents of individual differences for risky foraging in adolescents and early adulthood. We explored a large number of relationships in a discovery sample, pre-registered 9 selected hypotheses, and then tested these in an independent hold-out sample. Our main finding is that sex was the best predictor of cautiousness as well as performance, with a 20% payment gap between the sexes. Independent of sex, self-reported daringness, cognitive complexity, and measured IQ, also predicted better task performance. At the same time, there was no evidence that internalizing measures such as clinical anxiety were associated with behaviour in the task. Neither did we find evidence that developmental time related to improved performance or reduced risk-taking, both in cross-sectional and accelerated longitudinal analyses.

An extensive literature suggests greater risk-taking for males than females, including many self-report and experimental broadly defined 'risk' measures [34]. Our study reveals a more nuanced picture. Male adolescents took calculated risks that made them more successful in the long term. Male adolescents were less cautious (ie. collected on average more tokens and moved closer to the predator's location) when their potential losses were small early in an epoch, but adapted their behaviour to the same level as females (ie. decreased token collection to a greater extent over time) when a potential loss increased towards the end of the epoch. Our mediation analysis showed that higher initial, and more steeply decreasing, token collection explained most of the payment gap in our task. Therefore, male adolescents behaved daringly but not recklessly, in line with studies showing greater adjustment-to-risk in males in the Cambridge Gambling Task [35, 36]. Males are often reported to prefer economic risk (higher variability in outcomes) more than females [34, 37]. We replicated this finding here in an economic risk preference task. However, this preference did not mediate a large proportion of the sex effect on performance rendering it more likely to represent a separate propensity. Interestingly, our findings support a field study [2], which suggested a provision of economic bonuses for taking real-life risk in the 'gig economy' may disadvantage women, even in the absence of employer and customer discrimination.

What might account for sexual dimorphism in risky foraging behaviour? One explanation that we examined is that males are more practiced in computer games. While there is evidence for equivalent exposure to video games in both sexes [38], male adolescents engage more in the most violent-action-like video games [39], which provide for more intense sensorimotor training [40, 41]. However, sensorimotor practice alone is unlikely to explain our results. First, as per design of the task, sensorimotor practice has only a very small impact on escape success, which is mostly determined by a player's location on the grid when the predator wakes up. We found that males moved closer to the sleeping predator but did not get caught more often, an observation not explained by better sensorimotor performance and which we speculate instead depends on meta-cognitive abilities. Furthermore, we found no evidence that females increased their performance more over repeated epochs (i.e. with increased sensorimotor training). On the contrary, performance increased more steeply for males with experience, at least under high threat probability, supporting an argument of greater metacognitive ability to learn based on observing one's own performance. It is possible that training in violent-action-like games improves such ability, including habituation in perceiving 'apparent death' as an affordable outcome (in virtual reality), such that it can more easily be included into utility calculations [17].

A complementary explanation for females' overall higher cautiousness might be that signals of potential threat presence weigh more negatively into their subjective perception of reward itself, possibly based on their life experience of potential threats. On the other hand, males may be more sensitive to quantitative threat features, particularly related to the way that losses vary over time. In our study, potential threat is signalled by a predator shadow looming in a corner, which stays constant over time. In a study of human avoidance learning, females engaged in avoidance behaviour more quickly and for longer during signalled threat periods than was the case for males [42, 43]. Thus, in our study, threat signals may motivate females more against vigorous foraging despite a lack of actual hazard early in each epoch. Males may take into account the actual loss magnitude to a larger extent, which is variable over the course of an epoch. There was no difference between males and females in sensitivity to threat probability. Notably, the interpretation that decreased foraging over the epoch corresponds to increased cautiousness is plausible but there are alternative explanations, including a decreasing marginal utility of collecting additional tokens, or subjectively increasing hazard rate (which is objectively decreasing over time).

Two aspects of externalizing disposition were associated with task performance, but in opposite directions. Dispositional decision impulsivity, measured by ‘BIS cognitive complexity’, was anticorrelated with performance, while 'CADS daringness' correlated with performance and other task measures. These results provide an external validation for our in-task findings, albeit with small effect sizes. Furthermore, participants with high IQ performed better. Participants who did best were those that saw themselves as daring (high CADS daringness) and engaged in foraging to a greater extent, but were thoughtful as opposed to reckless (high IQ and low cognitive impulsivity), allowing them to decrease their foraging more steeply as potential loss increased. Mediation analysis suggested that although IQ and self-reported cognitive complexity were to some extent related, they represented for the most part separate influences on task behaviour. How these influences play out, and in particular how IQ leads to better performance, remains to be determined. We suggest a mediation is likely to reflect multiple influences, from faster reaction times through to reduced Pavlovian bias for losses [44].

We did not find evidence that self-reported anxiety predicted behaviour. Notably, approach/avoidance conflict paradigms such as the one we use here are designed to temporarily elicit cautious behaviour, not to distinguish among individuals based on self-reported anxiety. GABAergic anxiolytics consistently decrease cautiousness in rodent approach/avoidance conflict tasks [45, 46, 47], and in their human analogues [24, 25, 48], whereas other anxiolytic manipulations (such as chronic SSRI treatment) do not (or only inconsistently) reduce cautiousness in these tasks (see [47] for review). This suggests that cautiousness in this category of tasks (and possibly real-life cautiousness) is not directly related to, or determined by, feelings of anxiety, or their representation in questionnaire measures. More generally, while some models of human emotion make an implicit assumption that behaviour relates to concurrent subjective feeling, there is relatively little evidence for such a direct link [49], see for reviews e.g. [50, 51]. This has motivated a view that regards reported feelings as representations inferred from both interoception and from mechanisms that generate behaviour [50, 52], presumably with considerable interindividual variability in this inference, as is the case for other meta-cognitive and interoceptive processes [53]. In our view, this calls into question the viability of any straightforward mapping between cautious behaviour in approach/avoidance conflict tasks and self-reported anxiety, a mapping that has received limited empirical support to date. Interestingly, there was only a modest relation of anxiety and daringness in our sample (r = -.07), with some individuals expressing high values on both metrics. Thus, one may speculate that anxiety and daringness/cautiousness represent partly separate propensities that relate to different aspects of every-day behaviour on the one hand, and clinical symptoms on the other.

How reduced cautiousness in our task relates to catastrophic risk miscalculations, which can characterise the behaviour of male adolescents with a typical peak in mid-adolescence, remains to be determined. In our task, we did not detect any sex difference in terms of virtual survival. Furthermore, across the age range investigated here (14-24 years), we found no evidence for an effect of age or maturation on any task measure, including virtual survival. This raises a question as to which type of risk-taking our task measures. Reduced cautiousness in our task was linked to increased performance and thus may conceptually relate to adaptive risk-taking [7] where the latter has recently been suggested to be indexed by self-reported sensation-seeking [8]. Sensation seeking is viewed to peak around 16 years of age [8], which is not the pattern we observe for cautious behaviour in our task. Impulsive risk-taking, thought to be maladaptive [7], might relate to task survival rates, but again we found no impact of age. To conclusively rule out a relation with age it will be desirable to investigate a wider age range of subjects, including children (see e.g. [9]). However, we note that in a recent post-hoc analysis of adults between 18-57 years playing a similar game as ours, we observed a negative association of cautiousness with age. In other words, older participants were more risk-taking than younger ones by this metric [25]. This implies our task might measure a type of risk-taking that is a relatively stable trait (as indexed by moderate test-retest reliability over 2 years) but one that is unrelated to the specific type of risk-taking that leads to adolescents' increased vulnerability to catastrophic outcomes.

In conclusion, we found that young people do not fully maximize returns in a setting where sensory features – but not actual consequences – approximate a prey situation. Attributes of male sex, a self-assessment as ‘daring’, high ‘cognitive complexity’ (i.e. low reflection impulsivity) and having higher IQ help maximize monetary returns. We found no evidence that age or maturation played a role in the 14-24 years age range. We speculate that specific subcategories of externalizing disposition, rather than internalizing features such as anxiety, dominate in determining better or worse performance under risky foraging in young people.

Methods

Ethics statement

This research complied will all relevant ethical regulations. Written informed consent was obtained for all participants over the age of 16, and written consent from a parent/legal guardian was obtained for younger participants together with their assent. Ethical approval was granted by the National Health Service Research Ethics Committee (project ID 97546). Participant compensation for this task was between £0.00 and £5.00 with an average of £2.50.

Participants and Design

Our sample consisted of individuals recruited from the Neuroscience in Psychiatry Network (NSPN) 2400 Cohort. This was a community-based sample of young people living in either Cambridgeshire or Greater London, UK (Kiddle et al., 2018). The sampled age range was chosen to capture a high-risk period for onset of a range of common mental health problems, including a period of peak incidence of adolescent risk taking. Using purposive sampling, we recruited approximately equally in 5 age and sex groups (14–15, 16–17, 18–19, 20–21, 22–24 years old), until 785 participants were tested as a 'baseline cognition cohort' between 2013 and 2016. All participants also filled several batteries of questionnaires, described below. Participants performed multiple tasks with different analysis methods and anticipated effect sizes. Thus, overall sample size was heuristically determined. Below we provide a post-hoc power analysis for the task reported here.

Study data were collected and managed using REDCap electronic data capture tools hosted at University College London. REDCap (Research Electronic Data Capture) is a secure, webbased application designed to support data capture for research studies [54]. N = 781 participants provided complete data for the task reported here at baseline. Incomplete data were not analysed [28]. An accelerated longitudinal design was then used, in which the baseline sample was invited for follow up testing. N = 568 (N = 567 complete data) participants attended after 11-32 months, the FU-1 sample. A small subset of participants was additionally tested 6 months after baseline testing (FU-R, N = 68 participants, N = 64 complete data) to test for task stability and help interpret the FU-1 results.

Human risky foraging task

This task was developed to reflect established rodent approach/avoidance conflict tests in a pac-man style computer game [23, 24, 25]. The current study used a shortened version with reduced number of threat probability levels (2 instead of 3) and fewer trials per condition (20 instead of 40), and was presented using the MATLAB toolbox Cogent (www.vislab.ucl.ac.uk). The task included 80 'epochs' (and one bonus epoch, see below), that is, time periods in which participants had the opportunity to accumulate monetary tokens. In each epoch, they collected tokens on a 24 × 16 grid while under threat of being chased by a predator. Being caught resulted in the loss of all tokens collected in that epoch (see Figure 1A). One corner of the grid was a location safe from predator attack. The safe place was either the player's starting place or the opposite corner, randomly balanced over epochs. The 80 epochs were divided into five blocks of 16 epochs and approximately 5 minutes duration, with short self-paced breaks.

Tokens: At all times, ten tokens were uniformly distributed on the grid, and every 2 s one of the tokens changed its position randomly, in order to encourage uniform foraging across the grid. Collected tokens were replaced in a random position on the grid, and the number of collected tokens was displayed above the grid.

Predator: The predator was initially inactive in the corner diagonal to the safe place. Participants were instructed that the predator could become active and chase participants any time. Colour of the frame around the grid indicated two distinct predator wake-up probabilities (0.25 or 0.75), which participants learned to distinguish. Participants started either in the same place as the predator ('active') or in the safe place, ie. opposite the predator ('passive'). Notably, all epochs entailed going out onto the grid to collect tokens, and we have previously shown that over the course of an epoch, behaviour becomes comparable for the two starting positions [23, 24, 25], such that we averaged data over this factor, except for single-trial analysis of performance where such averaging was not possible.

Movements on the grid: Participants coordinated their movements by pressing the four computer keyboard arrow keys. No diagonal movements were possible. Participants could move at a maximum speed of 10 grid blocks per second if they held a key pressed. Both predators had the same speed of 40 grid blocks per second.

Epoch duration: Duration of the foraging phase was randomly drawn from 3 s, 6.5 s, 10 s, or 13.5 s. After the pre-determined foraging phase duration, the predator either woke up for a 5-second chase phase, or the next epoch started. Only the foraging phase was analysed. Before each epoch started, there was a 3 s countdown with a preview of the grid layout, during which the player could not move, to facilitate orientation on the grid.

Post-task questions: Participants rated on a visual analogue scale (ranging from 0% to 100%) the wake-up probability of the two predators. Finally, participants were given the choice to select the predator that they would like to face in a final bonus round. The majority of participants preferred the low-threat predator (discovery sample: 68%, confirmation sample 69%, both $p < .001$ in binomial test). They rated the wake-up rate of the low-threat predator as smaller than of the high-threat predator (mean ± standard deviation: discovery sample

47.5% ± 19.02% vs. 61.7% ± 17.84%; $t_{491} = 10.7$; $p < .001$; confirmation sample 46.1% ± 17.62% vs. 64.3% ± 17.69%; $t_{324} = 11.9$; $p < .001$).

Payment: At the end of the game, the average number of retained tokens over the whole task was transformed into a monetary reimbursement that was added to a constant fee for the whole testing day. Participants were truthfully told that average earnings from this task were expected to be £2.50 (and a maximum of £5.00) depending on performance.

Task measures

We took advantage of the substantial sample size to carry out extensive exploration in a 'discovery' sub-sample, and relied on an independent out-of-sample testing to validate our key hypotheses (see Analysis strategy below) Thus we analysed in total 38 task measures in the discovery sample.

First, we extracted seven previously reported continuous behavioural variables for each 1-second time bin within each epoch: (1) proportion of presence in safe place (the only grid block which the predator could not enter), (2) distance (as the crow flies) from threat (i.e., from the predator), (3) distance from nearest wall, (4) presence in safe quadrant (i.e. the quarter of the grid surrounding the safe place), (5) presence in threat quadrant (i.e. quarter of the grid surrounding the predator position), (6) token collection, and (7) speed when outside the safe place. Furthermore, we combined them into (8) a summary measure by weighting each measure by its theoretically possible range within the task, as reported previously [25]. We then averaged these measures across trials for each task condition. In doing so, we had to account for the different duration of epochs. For analysis of trajectory similarity, we used mean imputation, which is the strategy we had used for trajectory analysis in previous publications [23, 24, 25]. For the computation of epoch-summary scores (see below), due to a coding error that was detected after pre-registration, we imputed missing values with zeros. In supplementary results 6, we show that the resulting epoch-summary scores span the same space as scores computed with mean imputation and weighted least square regression. We further analyse the impact of this method on results, and show that all our key findings are replicated when using mean imputation before computing epoch-summary scores.

Normatively, as the number of collected tokens increases over an epoch within our task, so do potential losses, and participants should become more cautious by retreating to the safe place. We previously observed that the linear component of this intra-epoch adaptation of behaviour is reduced by the anxiolytics lorazepam, pregabalin and valproate, and by hippocampus and amygdala lesions [23, 24, 25]. We have also observed lesion-induced overall changes in the average behaviour, and in the impact of threat probability [23]. This motivated computing, for each of the 8 measures, the following 4 epoch-summary scores: (1) slope of a linear ordinary least squares regression on intra-epoch time, across both predators; (2) average over time and both predators; (3) difference in regression slope between the predators; (4) average difference over time, between the predators (see Extended Data Figure 1). We note that (2) corresponds to overall threat sensitivity, (1) to sensitivity to passage of time during an epoch, (4) to sensitivity to threat probability, and (3) to the interaction between time and threat probability. All measures were then recoded such that higher values mean higher cautiousness overall, or more cautiousness later (as opposed to earlier) during the epoch. Overall, this yielded 8 x 4 = 32 task measures. Furthermore, we analysed three summary statistics that did not depend on time: the number of tokens retained after the epoch ended (including predator chase phase) as main performance measure, the time until first (re-)entry into the safe place, and the minimum distance to the predator during foraging. For each of these three measures, we computed the average across conditions, and the difference between high and low threat probability, thus yielding another 6 task measures. Overall, 38 task measures were analysed.

Demographic and self-report measures

As demographic measures, we included sex and age on the day of the cognitive task. An estimate of total IQ was obtained using the 2-subtest version of the Wechsler Abbreviated Scale of Intelligence – First Edition (WASI-I) [55]. Self-report questionnaires were sent out in three waves not synchronised with the cognitive task battery. If more than one questionnaire pack was returned, we linearly interpolated questionnaire values from the two closest time points to the time point of the cognitive task. We measured symptoms with the Mood and Feelings Questionnaire (MFQ) [56], the Revised Children's Manifest Anxiety Scale (RCMAS) [57], the Rosenberg Self-Esteem Scale (RSE) [58], the Kessler Psychological Distress Scale (K10) [59], and the 'Behaviour Checklist'. The latter was a new brief self-report instrument based on the DSM-IV criteria for conduct disorder [28]. Dispositions were assessed by the sum scores for the 3 subscales of the Antisocial Process Screening Device (APD) [60], the 3 measures of the Child and Adolescent Disposition Scale (CADS, see Supplementary Table 3) [61], the 9 subscales of the Schizotypal Personality Questionnaire (SPQ) [62], the 3 subscales of the Inventory of Callous-Unemotional Traits (ICU) [63], and the 6 first-order factors of the Barratt Impulsive Scale (BIS, see Supplementary Table 4) [64]. Overall, 32 demographic and self-report measures covering 'internalizing' characteristics that might be related to over-cautiousness and anxiety, and 'externalizing' measures that might relate to inability or unwillingness to exercise thoughtful caution, were included in the analysis.

Statistical analysis: exploration-confirmation analysis

The high dimensionality of the data set and the many possible ways of analysing it posed a formidable multiple comparison problem. This is why we opted for a rigorous out-of-sample validation approach. All analyses were performed in a 'discovery' sub-sample and selected hypotheses from this analysis were pre-registered. The hold-out sample was then used for confirmation. The data analyst (DRB), who was located outside the NSPN centres, had no access to the primary NSPN database. He was provided the two samples via the data curator (MM) sequentially, so as not to have access to the hold-out sample until after the confirmation analysis was pre-registered at the Open Science Framework on 28.07.2018 (https://osf.io/hrce6/). All models to be confirmed were included in this pre-registration as RData files. The discovery sample comprised around 2/3 of the data, randomly drawn for each sex and age group from the entire sample (N = 492), while the remaining cases constituted the confirmation sample (N = 289). In addition, the data analyst had access to all data from the N = 64 FU-R cases, which were distributed between discovery and confirmation samples. The size of discovery and confirmation sample was determined heuristically. At our chosen alpha level of $\alpha = .001$, the discovery sample was sufficiently large to detect a bivariate correlation of $R^2 = .10$ with > 99% power, a correlation of $R^2 = .05$ with 96% power, and a correlation of $R^2 = .02$ with 44% power.

All statistical tests (with the exception of random permutation tests) were two-tailed. All statistical models (with the exception of random permutation tests) assumed normality of the residuals, but this assumption was not formally tested.

First, we constructed quantitative hypotheses of how task measures related to demographic/psychometric variables. We computed 1216 bivariate regressions between each task measure on the one hand, and each demographic/self-report variable on the other. Only findings at an alpha level of $p < .001$ were retained. To minimise the number of confirmation tests, we then built multiple regression models to predict each psychometric/demographic measure simultaneously from all those task measures that had a significant bivariate relation. These models were then applied to the confirmation data set without refitting, and tested by randomly permuting the dependent variable 10^5 times. We present the ratio of explained variance (R^2) in the hold-out sample as best estimate of the out-of-sample predictive performance, and the explained variance for a multiple regression model with the same predictors, fit on the combined sample, as best estimate of in-sample predictive performance [29]. For all effect sizes, we computed 95% confidence intervals from the sampling distribution of 10^5 bootstrapped samples, taking into account the asymmetric distributions. Notably, confidence intervals are included due to journal requirements and do not reflect the posterior plausibility of true parameter values [65]. We then estimated the relative contribution of each task measure in predicting the demographic/self-report measure in sequential procedure, by residualising on each step each task measure with respect to the measure that shared most variance with the predictor in the previous step.

Next, we were interested in quadratic relationships between demographic/self-report variables, and task variables. We computed 1216 multiple regression models to predict each task measure from a second order polynomial of each demographic/self-report variable (ie. from the variable and its square). Findings at an alpha-level of $p < .001$ were retained. We then build multiple regression models to relate the questionnaire variable to the several task measures with significant bivariate quadratic relations. To do so, we z-scored each relating task variable across participants, and then averaged over all relating task variables. We then build a multiple regression model to predict this task sum score from a second order polynomial of the questionnaire measure. This model was then applied to the confirmation data set without refitting and with the normalisation parameters established in the discovery sample, and subject to the aforementioned random permutation test. Notably, neither of the two hypotheses derived in this way was confirmed in the hold-out sample.

CADS daringness was the self-report variable with the highest number of relationships to task variables. We were thus interested whether CADS daringness did not only predict summary statistics from the task, but also the intra-epoch trajectories. To address this, we averaged each of the 7 time-dependent measures across the predator factor. We then created the average intra-epoch trajectory for the 20 individuals scoring highest on CADS daringness. For the remaining individuals, we calculated the scalar product between individual trajectory and high-daringness trajectory, separately for each measure. We then computed the bivariate relation between each trajectory similarity measure and CADS daringness. Findings at an alpha-level of $p < .001$ were retained. We then built a multiple regression model to predict CADS daringness from all trajectory similarity measures that significantly related to CADS daringness, similar to the approach described above. In the confirmation sample, we computed, for all participants, trajectory similarity with the high-daringness trajectories from the discovery sample. We then applied this model to the confirmation data set without refitting, in a random permutation test.

All bivariate relations and multiple regression models in the discovery sample were replicated after controlling for sex as a covariate. All linear and trajectory models that were significant in the confirmation sample were followed up to examine sex as a possible confound, by re-fitting the multiple regression model with sex as covariate, and all confirmed hypotheses were replicated in this procedure.

Some individual bivariate relationships were not replicated in the confirmation sample, although the joint predictive model was confirmed. To follow up on this discrepancy, all confirmed multiple regression models were refitted to the entire data set, to test for a dataset × predictor interaction (indicating different weights in discovery and confirmation sample); and no significant differences between the regression weights for discovery and confirmation sample were found with this approach.

In all cases where we interpret null results, we computed evidence for a linear model with the predictor in question, and a null model without the predictor. We approximated model evidence by extracting Akaike Information Criterion [66] with the R function "AIC". We computed log Bayes Factors (LBF) as LBF = 0.5 (AIC_{pred} - AIC_{null}), where positive values are evidence in favour of the simpler null model.

We reasoned that performance differences between the sexes could reflect a differential prior experience with computer games, and this could lead to a different performance trajectory over trials. Thus, we analysed sex effects on the trajectory of task performance over trials, as indexed by the number of tokens retained (after catch phase). We only analysed the first 64 trials, as the duration (which strongly influences performance) of the remaining 16 trials was unbalanced. First we used Bayesian model selection to pinpoint the best model for the trial-by-trial trajectory across categories. We compared linear, quadratic, logarithmic and square root models by Akaike Information Criterion. We then tested the effect of sex in a 2 (sex) x 2 (predator) x 2 (task) ANCOVA with log(trl) as continuous covariate.

Developmental and practice effects on the task were tested by analysing data from the baseline and follow-up measures. We first tested, for each task measure, the change between BSL and FU-R, and between FU-R and FU-1. To disentangle the effect of maturation and practice in the FU-1 sample, we computed, for each task measure, a linear mixed-effects model with repetition as within-subject factor, and time between BSL and FU-1 as continuous predictor. We compared this with a model not including time, and extracted Akaike Information Criterion to approximate model evidence.

In our analysis, we combined weighted summary statistics of the seven time-dependent measures into cautiousness sum scores. We have previously suggested that linear change in cautiousness reflects a "loss adaptation" sum score [25]. To assess the internal consistency of this metric, we report Cronbach's alpha of the 7 contributing statistics. We assessed test-retest reliability of all 38 task measures by computing the bivariate correlation between BSL and FU-R, or FU-1, respectively. Finally, we investigated the internal structure of 7 linear change coefficients, or all 34 task measures (excluding the collinear sum scores) by computing an exploratory factor analysis using maximum likelihood factorisation with varimax rotation. Parallel analysis suggested 2 or 6 factors, respectively. We computed exploratory factor analysis with these numbers of factors. We then split the discovery data set into random partitions and found that for the factor analysis of 34 task measures, only 3 factors robustly replicated between different partitions. These 3 factors had a cumulative factor loading of .71. This is why we chose to report and interpret the first 2, or 3, factors of the exploratory factor analysis. Because factor analysis was computed with a higher number of factors, this choice does not impact upon the factor solution. We then defined a confirmatory factor analysis by retaining all factor loadings above an absolute threshold of .2. The exploratory factor analysis was computed in the confirmation sample in the same way, using 2 or 6 factors as determined on the discovery sample, of which we retained 2 or 3 factors. The confirmatory factor analysis did not converge in the confirmation sample and is not reported here. To nevertheless confirm the exploratory factor analysis in the confirmation sample, we computed factor values in the confirmation sample using either the loadings derived in the discovery sample, or factor loadings from an exploratory factor analysis on the confirmation sample. We then assessed the correlation between the two ways of computing factor values.

Further exploratory analyses that did not yield additional insights within the discovery sample included predicting the change in task measures between BSL and FU-1 from demographic/self-report measures at BSL or from change in these measures between BSL and FU-1, and canonical correlation analysis between task and demographic/self-report measures.

Statistical analysis: post-hoc analysis in the combined sample

After gaining access to the complete data set, the following exploratory and non-planned analyses were performed. We only report parameter estimates and, where appropriate, confidence intervals, but provide no inference statistics.

First, we analysed how task performance related to all other task measures. Furthermore, at the suggestion of a reviewer, we analysed virtual survival rates, ie. the rate with which participants were caught when the predator woke up. We analysed how survival rate related to the 38 task measures in the entire sample, and to the 32 predictor measures in the discovery sample.

Second, we analysed mediation of the impact of predictors on task variables using the R package 'mediation', which uses a quasi-Bayesian Monte Carlo approximation to estimate causally mediated, and non-mediated direct, effects [30].

Third, we analysed a separate ‘lottery task’, administered to the same cohort (N = 781 available data sets, baseline only), in order to assess canonical economic risk preferences. This followed very closely the methodology of Symmonds et al [33], but the lotteries were simplified to four ‘slices’ (see Extended Data Figure 5), to reduce cognitive load and facilitate large scale testing. Please see [33] for further details.

The probability of choosing the roulette was modelled as:

$$\pi_r = zexp\left(\frac{E_r + w_{var}VAR_r + w_{skew}SKEW_R - E_{sure}}{\tau}\right)$$

Where the index *r* refers to the roulette and *sure* to sure amount, E is expectation (mean), VAR the variance, SKEW the skewness, τ the decision temperature parameter, w_{var}, w_{skew} the parameters quantifying the taste of the individual for variable and skewed distributions of outcome respectively and *z* is a normalizing factor ensuring choice probabilities add to 1. A lower w_{var} was therefore hypothesized to statistically explain less avoidance to risk-related states (distance from predator, etc) in our main task. The model was fitted by finding, for each individual, the set of parameters that maximized the likelihood of the data.

Data availability

Anonymised data are available on the Open Science Framework (https://osf.io/mnbfy/) [67]. Full data are available upon reasonable request from the corresponding authors or from OpenNSPN@medschl.cam.ac.uk.

Code availability

All custom code used for the analysis is available on the Open Science Framework (https://osf.io/mnbfy/) [67]. After extracting task measures using Matlab 2017b, all discovery analyses were performed in R 3.4.1 (www.r-project.org), using the following toolboxes: R.matlab v 3.6.1, abind 1.4-5, reshape2 1.4.3, nlme 3.1-131.1, lme4 1.1-15, lmerTest 2.0-36, nFactors 2.3.3, sem 3.1-9. Confirmation and post-hoc analyses were performed in R 3.5.2, using the following toolboxes: R.matlab v 3.6.2, abind 1.4-5, reshape2 1.4.3, psych 1.8.12, lme4 1.1-21, lmerTest 3.1-0, sem 3.1-9, pracma 2.2.5, mediation 4.5.0, gvlma 1.0.0.3, DescTools 0.99.30, corrplot 0.84.

References

1. Lima SL, Dill LM. Behavioral decisions made under the risk of predation: a review and prospectus. *Canadian journal of zoology* **68**, 619-640 (1990).

2. Cook C, Diamond R, Hall J, List JA, Oyer P. The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers.). National Bureau of Economic Research (2018).

3. Steinberg L. Risk taking in adolescence: what changes, and why? *Annals of the New York Academy of Sciences* **1021**, 51-58 (2004).

4. Schwebel DC, Severson J, Ball KK, Rizzo M. Individual difference factors in risky driving: the roles of anger/hostility, conscientiousness, and sensation-seeking. *Accid Anal Prev* **38**, 801-810 (2006).

5. Eaton DK*, et al.* Youth risk behavior surveillance--United States, 2007. *MMWR Surveill Summ* **57**, 1-131 (2008).

6. Somerville LH, Casey BJ. Developmental neurobiology of cognitive control and motivational systems. *Current opinion in neurobiology* **20**, 236-241 (2010).

7. Romer D, Reyna VF, Satterthwaite TD. Beyond stereotypes of adolescent risk taking: Placing the adolescent brain in developmental context. *Dev Cogn Neurosci* **27**, 19-34 (2017).

8. Khurana A, Romer D, Betancourt LM, Hurt H. Modeling Trajectories of Sensation Seeking and Impulsivity Dimensions from Early to Late Adolescence: Universal Trends or Distinct Sub-groups? *J Youth Adolesc* **47**, 1992-2005 (2018).

9. van den Bos W, Hertwig R. Adolescents display distinctive tolerance to ambiguity and to uncertainty during risky decision making. *Scientific reports* **7**, 40962 (2017).

10. Frey R, Pedroni A, Mata R, Rieskamp J, Hertwig R. Risk preference shares the psychometric structure of major psychological traits. *Sci Adv* **3**, e1701381 (2017).

11. Overman WH, Frassrand K, Ansel S, Trawalter S, Bies B, Redmond A. Performance on the IOWA card task by adolescents and adults. *Neuropsychologia* **42**, 1838-1851 (2004).

12. Deakin J, Aitken M, Robbins T, Sahakian BJ. Risk taking during decision-making in normal volunteers changes with age. *J Int Neuropsychol Soc* **10**, 590-598 (2004).

13. Lauriola M, Panno A, Levin IP, Lejuez CW. Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task. *J Behav Decis Making* **27**, 20-36 (2014).

14. Defoe IN, Dubas JS, Figner B, van Aken MA. A meta-analysis on age differences in risky decision making: adolescents versus children and adults. *Psychol Bull* **141**, 48-84 (2015).

15. Bach DR, Hulme O, Penny WD, Dolan RJ. The known unknowns: neural representation of second-order uncertainty, and ambiguity. *Journal of Neuroscience* **31**, 4811-4820 (2011).

16. Bach DR, Dolan RJ. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat Rev Neurosci* **13**, 572-586 (2012).

17. Korn CW, Bach DR. Maintaining homeostasis by decision-making. *PLoS computational biology* **11**, e1004301 (2015).

18. Korn CW, Bach DR. Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications* **9**, 325 (2018).

19. Korn CW, Bach DR. Minimizing threat via heuristic and optimal policies recruits hippocampus and medial prefrontal cortex. *Nat Hum Behav*, (2019).

20. Caraco T. Energy Budgets, Risk and Foraging Preferences in Dark-Eyed Juncos (Junco-Hyemalis). *Behavioral Ecology and Sociobiology* **8**, 213-217 (1981).

21. Kolling N, Behrens TE, Mars RB, Rushworth MF. Neural mechanisms of foraging. *Science* **336**, 95-98 (2012).

22. Khemka S, Barnes G, Dolan RJ, Bach DR. Dissecting the Function of Hippocampal Oscillations in a Human Anxiety Model. *J Neurosci* **37**, 6869-6876 (2017).

23. Bach DR*, et al.* Human Hippocampus Arbitrates Approach-Avoidance Conflict. *Current Biology* **24**, 541-547 (2014).

24. Bach DR, Korn CW, Vunder J, Bantel A. Effect of valproate and pregabalin on human anxiety-like behaviour in a randomised controlled trial. *Transl Psychiatry* **8**, 157 (2018).

25. Korn CW, Vunder J, Miro J, Fuentemilla L, Hurlemann R, Bach DR. Amygdala Lesions Reduce Anxiety-like Behavior in a Human Benzodiazepine-Sensitive Approach-Avoidance Conflict Test. *Biological psychiatry* **82**, 522-531 (2017).

26. Bach DR, Hoffmann M, Finke C, Hurlemann R, Ploner CJ. Disentangling Hippocampal and Amygdala Contribution to Human Anxiety-Like Behavior. *J Neurosci* **39**, 8517-8526 (2019).

27. Mobbs D, Kim JJ. Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Curr Opin Behav Sci* **5**, 8-15 (2015).

28. Kiddle B*, et al.* Cohort Profile: The NSPN 2400 Cohort: a developmental sample supporting the Wellcome Trust NeuroScience in Psychiatry Network. *Int J Epidemiol* **47**, 18-19g (2018).

29. Yarkoni T, Westfall J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci* **12**, 1100-1122 (2017).

30. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. *J Stat Softw* **59**, (2014).

31. Markowitz H. Portfolio selection. *Journal of Finance* **7**, 77-91 (1952).

32. Symmonds M, Bossaerts P, Dolan RJ. A behavioral and neural evaluation of prospective decision-making under risk. *Journal of Neuroscience* **30**, 14380-14389 (2010).

33. Symmonds M, Wright ND, Bach DR, Dolan RJ. Deconstructing risk: Separable encoding of variance and skewness in the brain. *Neuroimage*, (2011).

34. Byrnes JP, Miller DC, Schafer WD. Gender differences in risk taking: A meta-analysis. *Psychological Bulletin* **125**, 367-383 (1999).

35. Lewis G, Srinivasam R, Roiser JP, Blakemore SJ, Flouri E, Lewis G. *Risk taking to obtain reward: gender differences and associations with emotional and depressive symptoms in a nationally representative cohort of UK adolescents*. https://doi.org/10.1101/644450 (2019).

36. van den Bos R, Taris R, Scheppink B, de Haan L, Verster JC. Salivary cortisol and alpha-amylase levels during an assessment procedure correlate differently

with risk-taking measures in male and female police recruits. *Front Behav Neurosci* **7**, 219 (2013).

37. Fisher PJ, Yao R. Gender differences in financial risk tolerance. *J Econ Psychol* **61**, 191-202 (2017).

38. Stuart K. UK gamers: more women play games than men, report finds. In: *The Guardian*) (2014).

39. DeCamp W. Who plays violent video games? An exploratory analysis of predictors of playing violent games. *Pers Indiv Differ* **117**, 260-266 (2017).

40. Green CS, Bavelier D. Action video game modifies visual selective attention. *Nature* **423**, 534-537 (2003).

41. Dye MW, Green CS, Bavelier D. Increasing Speed of Processing With Action Video Games. *Curr Dir Psychol Sci* **18**, 321-326 (2009).

42. Sheynin J*, et al.* Behaviourally inhibited temperament and female sex, two vulnerability factors for anxiety disorders, facilitate conditioned avoidance (also) in humans. *Behav Processes* **103**, 228-235 (2014).

43. Sheynin J, Moustafa AA, Beck KD, Servatius RJ, Myers CE. Testing the role of reward and punishment sensitivity in avoidance behavior: a computational modeling approach. *Behav Brain Res* **283**, 121-138 (2015).

44. Moutoussis M*, et al.* Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS computational biology* **14**, e1006679 (2018).

45. Calhoon GG, Tye KM. Resolving the neural circuits of anxiety. *Nature Neuroscience* **18**, 1394-1404 (2015).

46. Gray JA, McNaughton N. *The neuropsychology of anxiety: An enquiry into the functions of the septohippocampal system*. Oxford University Press (2000).

47. Kirlic N, Young J, Aupperle RL. Animal to human translational paradigms relevant for approach avoidance conflict decision making. *Behav Res Ther* **96**, 14-29 (2017).

48. Biedermann SV*, et al.* An elevated plus-maze in mixed reality for studying human anxiety-related behavior. *BMC Biol* **15**, 125 (2017).

49. DeWall CN, Baumeister RF, Chester DS, Bushman BJ. How Often Does Currently Felt Emotion Predict Social Behavior and Judgment? A Meta-Analytic Test of Two Theories. *Emot Rev*, (2015).

50. Bach DR, Dayan P. Algorithms for survival: a comparative perspective on emotions. *Nat Rev Neurosci* **18**, 311-319 (2017).

51. LeDoux JE. Semantics, Surplus Meaning, and the Science of Fear. *Trends in cognitive sciences* **21**, 303-306 (2017).

52. Barrett LF. The theory of constructed emotion: An active inference account of interoception and categorization. *Soc Cogn Affect Neurosci*, (2016).

53. Rouault M, Seow T, Gillan CM, Fleming SM. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological psychiatry* **84**, 443-451 (2018).

54. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**, 377-381 (2009).

55. Wechsler D. *Wechsler Abbreviated Scale of Intelligence*. The Psychological Corporation: Harcourt Brace & Company (1999).

56. Costello EJ, Angold A. Scales to assess child and adolescent depression: checklists, screens, and nets. *J Am Acad Child Adolesc Psychiatry* **27**, 726-737 (1988).

57. Reynolds CR, Richmond BO. What I Think and Feel: a revised measure of Children's Manifest Anxiety. *J Abnorm Child Psychol* **25**, 15-20 (1997).

58. Rosenberg M. *Conceiving the self*. Basic Books (1979).

59. Kessler R, Mroczek D. Final versions of our non-specific psychological distress scale. *Memo dated March* **10**, 1994 (1994).

60. Frick PH, Hare RD. *The Antisocial Process Screening Device*. Multi-Health Systems (2001).

61. Lahey BB, Rathouz PJ, Applegate B, Tackett JL, Waldman ID. Psychometrics of a self-report version of the Child and Adolescent Dispositions Scale. *J Clin Child Adolesc Psychol* **39**, 351-361 (2010).

62. Raine A. The SPQ: a scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophr Bull* **17**, 555-564 (1991).

63. Kimonis ER*, et al.* Assessing callous-unemotional traits in adolescent offenders: validation of the Inventory of Callous-Unemotional Traits. *Int J Law Psychiatry* **31**, 241-252 (2008).

64. Patton JH, Stanford MS, Barratt ES. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol* **51**, 768-774 (1995).

65. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. *Psychon Bull Rev* **23**, 103-123 (2016).

66. Burnham KP, Anderson DR. Multimodel inference - understanding AIC and BIC in model selection. *Sociol Method Res* **33**, 261-304 (2004).

67. Bach DR, Moutoussis MM, NSPN consortium, Bowler A, Dolan RJ. NSPN_AAC. (2020). DOI 10.17605/OSF.IO/MNBFY

Acknowledgements

We thank the the NSPN management and research assistant teams. Cognitive experiments were realized using Cogent 2000 developed by the Cogent 2000 team at the FIL and the ICN, and Cogent Graphics developed by John Romaya at the Wellcome Department of Imaging Neuroscience. The Wellcome Trust funded the 'Neuroscience in Psychiatry Project' (NSPN). All NSPN members (Supplementary Table 1) are supported by Wellcome Strategic Award (095844/7/11/Z). DRB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2018 CoG-816564 ActionContraThreat). MM and DRB receive support from the National Institute for Health Research (NIHR) UCLH Biomedical Research Centre. RJD is supported by a Wellcome Investigator Award (098362/Z/12/Z). Peter Fonagy (NSPN consortium, Supplementary Table 1) is in receipt of a NIHR Senior Investigator Award (NF-SI-0514-10157) and was in part supported by the NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames at Barts Health NHS Trust. The Max Planck – UCL Centre for Computational Psychiatry and Ageing is a joint initiative of the Max Planck Society and UCL. The Wellcome Centre for Human Neuroimaging is funded by core funding from the Wellcome Trust (203147/Z/16/Z). The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

DRB, MM, NSPN consortium, and RJD contributed to conception and design of this work. MM, NSPN consortium, and RJD contributed to acquisition of the data. DRB and MM analysed the data. DRB, MM, AB and RJD contributed to interpretation of the data and to drafting and revising the manuscript.

Competing interests

The authors declare no competing interests.

Neuroscience in Psychiatry Network consortium

Michael Moutoussis, Aislinn Bowler, Raymond J. Dolan: Max Planck-UCL Centre for Computational Psychiatry and Ageing Research, and Wellcome Centre for Human Neuroimaging, University College London, London WC1 3BG, UK

A full list of members and their affiliations appears in the Supplementary Information

Figures

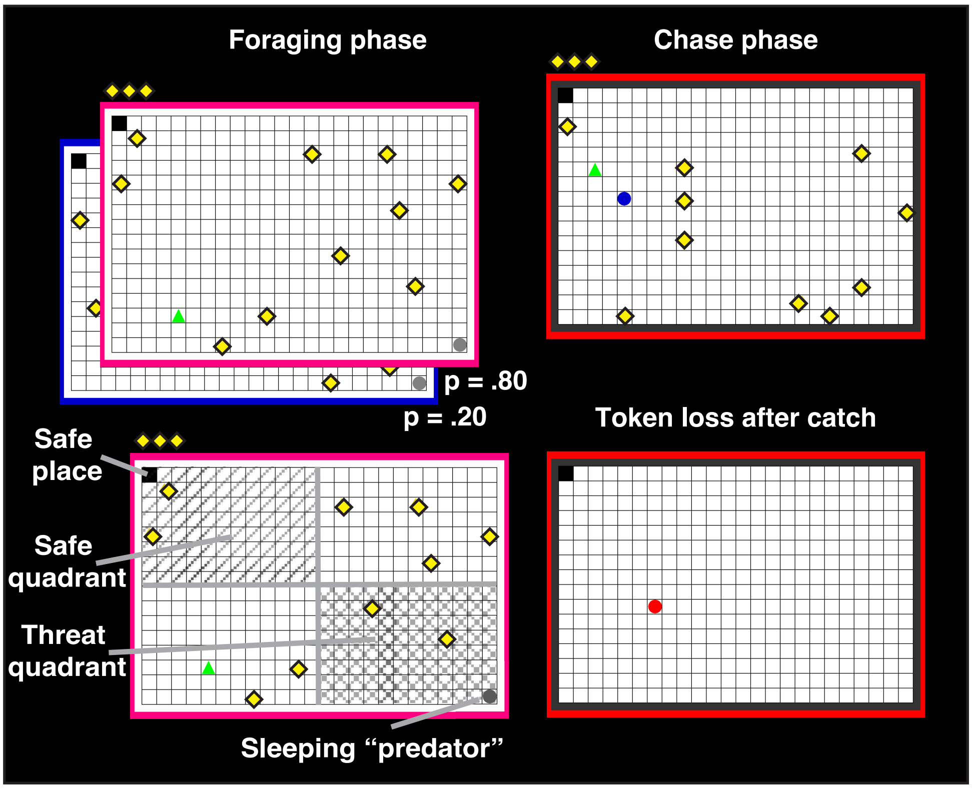


Figure 1. Risky foraging task, building on rodent approach/avoidance conflict tests. In each of 81 game 'epochs', the participant forages for monetary tokens on a grid, where a virtual predator can wake-up and give chase at any time. If caught, the player loses all tokens. Each epoch starts with a fresh 'life' and zero tokens. The result from randomly selected epochs is paid out in money at the end. Thus, players are incentivized to retain as many token as possible on each epoch.

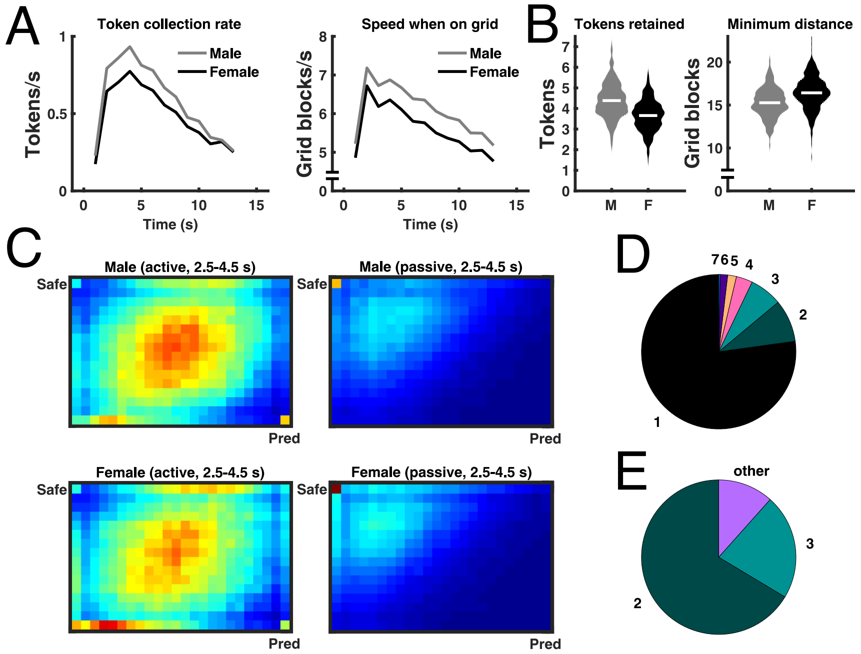


Figure 2. Relation between sex and task measures. A: Intra-epoch trajectories of token collection rate and speed when on grid, illustrating the sex differences in derived summary statistics (corresponding to measures 2-3, 5, 7 in panel D). B: Distribution of time-independent statistics for males and females: tokens retained (i.e. performance), and minimum distance from threat (corresponding to measures 1 and 4 in panel D). White lines: mean. Standard errors are smaller than line width and not displayed. C: Heat maps illustrating the probability of being in each position on the grid during 2.5-4.5 s after epoch start, for epochs in which the player starts in the predator position ('active') or in the safe place ('passive'). Females stay closer to the safe place and to the walls than males. D: Proportion of additionally explained variance by each task measure, after residualising already explained variance, and normalized for the overall explained variance. Labels for pie segments: 1. Tokens retained, 2. Average token collection rate, 3. Decrease in token collection rate, 4. Minimum distance from threat, 5. Decrease in speed when on grid, 6. Decrease in distance from walls, 7. Average speed when on grid. In terms of bivariate relations, sex explained in the combined sample 17.0% (12.4%-22.0%), 15.9% (11.5%-20.8%), 14.7% (10.4%-19.5%), 9.0% (5.5%-13.2%), 3.0% (1.1%-5.8%), 1.0% (0.1%-2.8%), and 9.1% (5.6%-13.3%)

variance (parametric 95%-CI) of these task measures. E: Proportion of mediation of the sex effect on performance. Numeric pie segment labels are the same as in D; other: remaining proportion in the sex effect on performance, explained by variables that were not part of the mediation analysis and not included in 2-7. Supplementary Table 2 and Extended Data Figure 2 show results separately for discovery and confirmation sample.

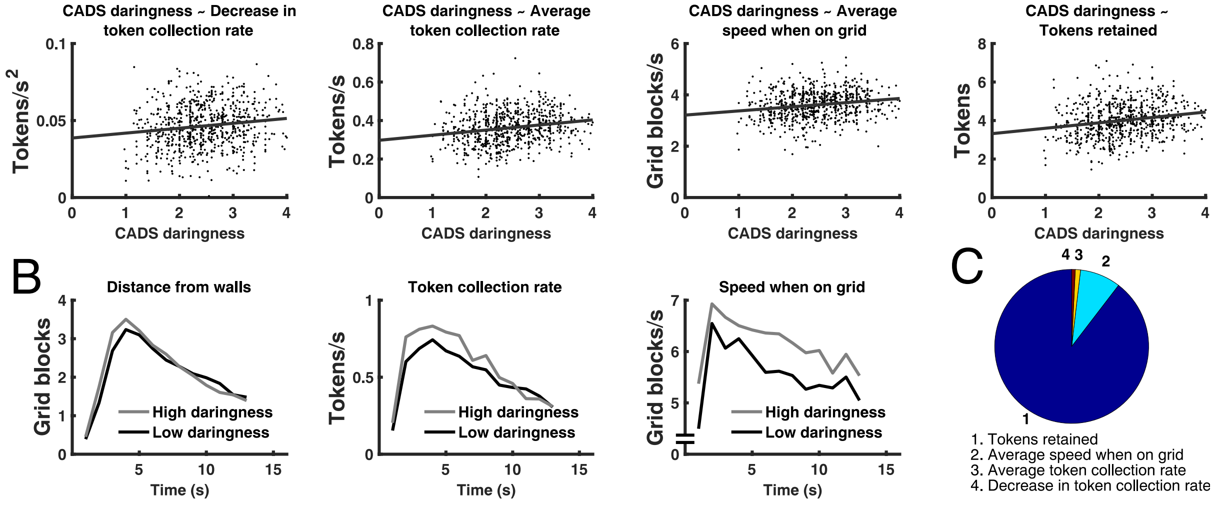


Figure 3. Relation of self-reported daringness (CADS questionnaire) with task measures. A: Individual task measures that relate to daringness. B: Average trajectories of the 20 highest-scoring and the 20 lowest-scoring individuals in the discovery sample for those measures in which daringness predicted trajectory similarity. C: Proportion of additionally explained variance by each task measure, after residualizing already explained variance, and normalized for the overall explained variance. In terms of bivariate relations, daringness explained, across the entire sample 3.9% (1.6%-7.1%), 3.4% (1.3%-6.4%), 3.9% (1.6%-7.0%), and 2.0% (0.5%-4.5%) variance (parametric 95%-CI) in the task measures as listed in C. Supplementary Table 2 and Extended Data Figure 3 show results separately for discovery and confirmation sample.

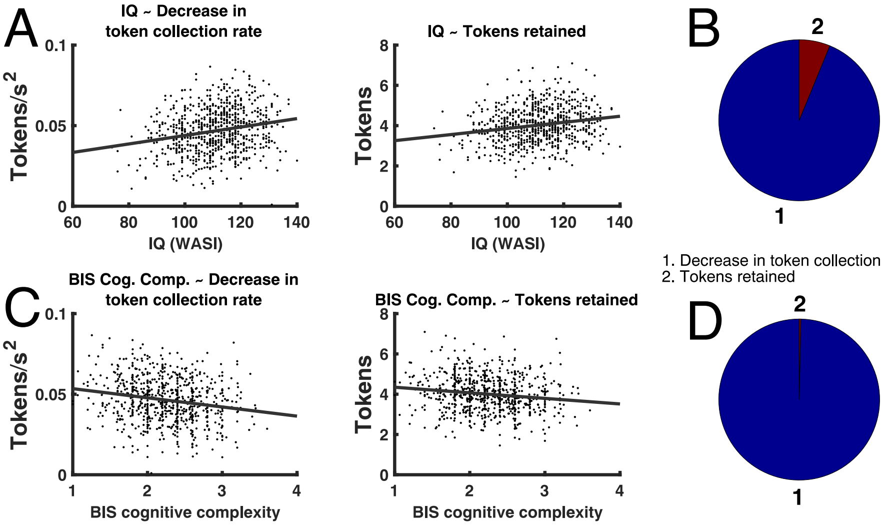


Figure 4. Relation of IQ (measured with WASI) and self-reported cognitive complexity (BIS questionnaire) with task measures. A: Individual task measures that relate to IQ. B: Proportion of additionally explained variance by each task measure, after residualizing already explained variance, and normalized for the overall explained variance. In terms of bivariate relations, IQ explained, across the entire sample, 4.6% (2.1%-7.9%) and 3.8% (1.6%-6.9%) variance (parameteric 95%-CI) in the task measures as listed in B. C: Individual task measures that relate to self-reported cognitive complexity. D: Proportion of additionally explained variance by each task measure, after residualizing already explained variance, and normalized for the overall explained variance. In terms of bivariate relations, self-reported cognitive complexity explained, across the entire sample, 3,8% (1.5%-6.9%) and 2.2% (0.6%-4.8%) variance (95%-CI) in the task measures as listed in B. Supplementary Table 2 and Extended Data Figure 4 show results separately for discovery and confirmation sample.

Tables

Table 1: Age distribution, and psychometric measures predictive of behaviour, for the discovery and confirmation samples.

	Discovery: Males			Discovery: Females			Confirmation: Males			Confirmation: Females		
14-15	47			54			24			27		
16-17	48			51			21			27		
18-19	47			50			19			27		
20-21	48			49			28			33		
22-24	47			51			28			32		
	Mean	±	SD	Mean	±	SD	Mean	±	SD	Mean	±	SD
IQ (WASI)	111.95	±	11.41	109.45	±	11.18	111.89	±	11.6	109.99	±	11.05
CADS daringness	2.67	±	0.55	2.24	±	0.61	2.61	±	0.59	2.28	±	0.61
BIS cogn. compl.	2.18	±	0.47	2.32	±	0.45	2.22	±	0.49	2.27	±	0.48

See **Supplementary Results 4** for performance (tokens retained) of the different age groups.

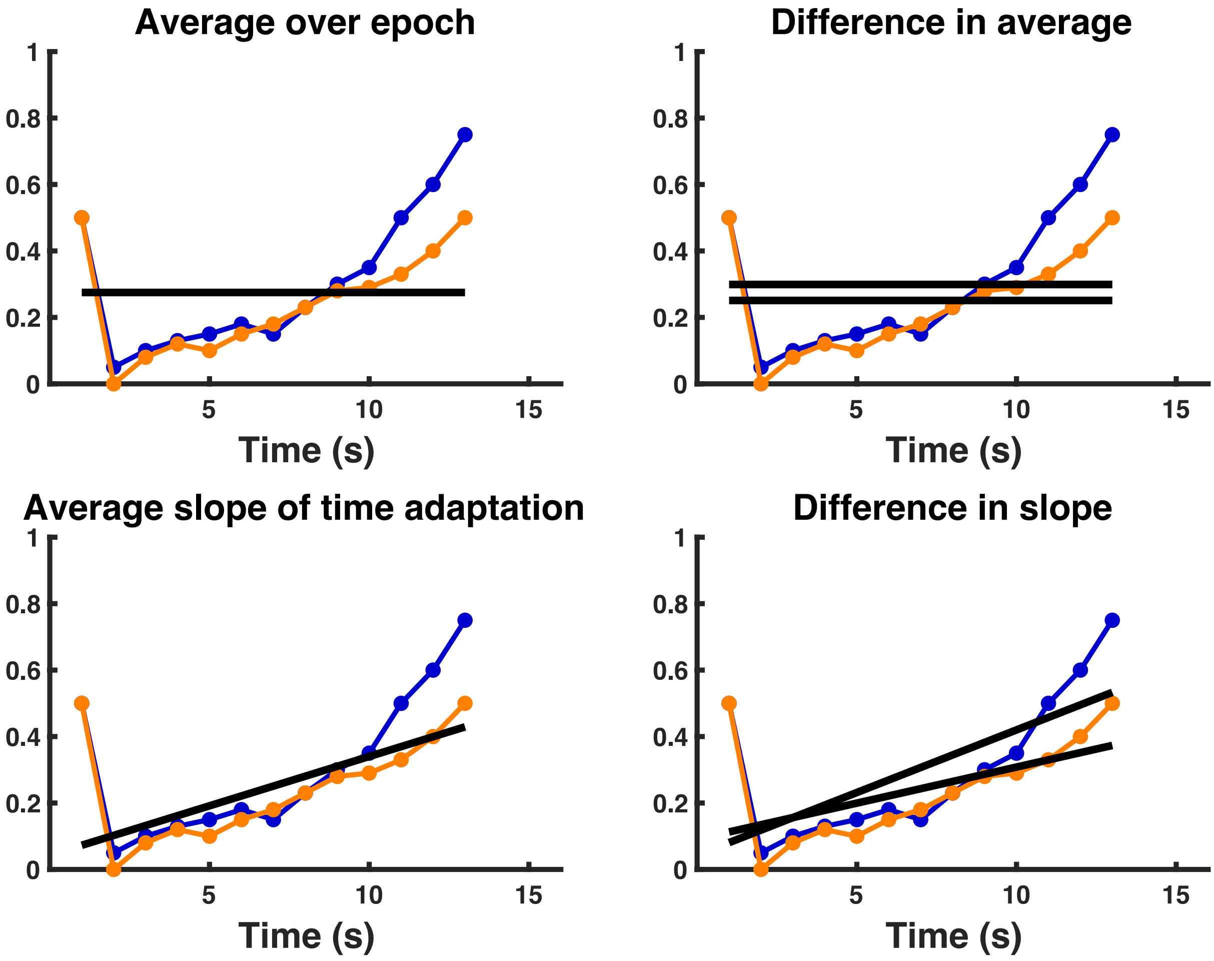
Table 2: Pre-registered hypotheses, and results of the confirmation analysis.

Hypothesis	**Task variables in predictive model**	**Discovery sample: accuracy (bootstrapped 95%-CI); test statistic and significance level (uncorrected) from multiple (H1: logistic) regression; H2: LME results**	**Confirmation sample: accuracy (bootstrapped 95%-CI) of the discovery model; significance level (uncorrected) from non-parametric random permutation test; H2: LME results**	**Combined sample: accuracy (bootstrapped 95% CI) of the refitted joint predictive model**	**Non-confirmed bivariate relations included in the predictive model**
H1: Male sex is associated with less cautious behavior and higher performance, as indexed by a weighted combination of 7 task measures	Decrease in distance from walls, Decrease in token collection rate, Decrease in speed when on grid, Average token collection rate, Average speed when on grid, Minimum distance from threat, Tokens retained	69.3% (63.8%-72.6%) χ2(7) = 126.1 p < .001	69.2% (64.0%-74.7%) p < .001	69.7% (65.9%-73.0%)	Decrease in distance from walls
H2: Male participants increase their performance (tokens retained) more over repeated epochs than females	N/A	F(1, 30982) = 9.66 p < .001 (see Supplementary Results 3)	F(1, 1819) = 0.2 p = .65 (see Supplementary Results 3)	N/A	N/A
H3: Higher IQ is associated with less cautious behaviour and higher performance, as indexed by a weighted combination of 2 task measures	Decrease in token collection rate, Tokens retained	4.0% (-0.2%-6.5%) F(2, 487) = 10.1 p < .001	6.7% (2.4%-12.2%) p < .001	4.9% (1.9%-7.7%)	All confirmed
H4: Higher CADS subscale 'Daringness' is associated with less cautious behaviour and higher performance, as indexed by a weighted combination of 4 task measures	Decrease in token collection rate, Average token collection rate, Average speed, Tokens retained	4.0% (-1.0%-6.2%) F(4, 456) = 4.8 p < .001	4.3% (0%-9.7%) p < .001	4.3% (4.0%-8.2%)	Decrease in token collection rate
H5: Higher BIS factor 'Self-control' is associated with decrease in threat quadrant presence	Decrease in threat quadrant presence	2.5% (-0.9%-4.5%) F(1, 453) = 11.7 p < .001	Not confirmed p > .99	Not confirmed	None confirmed
H6: Higher BIS factor 'Cognitive complexity' is associated with less cautious behavior and higher performance, as indexed by a weighted combination of 2 task measures	Decrease in token collection rate, Tokens retained	4.5% (-0.4%-7.4%) F(2, 459) = 10.8 p < .001	3.0% (-1.6%-8.5%) p = .002	3.8% (0%-6.1%)	All confirmed
H7: CADS subscale 'Daringness' is associated with similarity of intra-epoch behavioral trajectory to trajectory of 20 highest CADS 'Daring' scorers in discovery sample (behavioral trajectory indexed by weighted combination of 3 time-dependent task variables)	Wall distance, Token collection rate, Speed	5.4% (0.0%-8.3%) F(3, 435) = 8.3 p < .001	6.3% (1.6%-11.9%) p < .001	6.0 % (5.2%-10.7%)	All confirmed
H8: RCMAS 'Anxiety' sum score quadratically predicts cautious behavior, as indexed by an average of 4 normalized and recoded task measures	Average cautiousness score, Average safe place presence, Average distance from walls, Time to reach safe place	3.2% (-0.9%-5.4%) F(2, 460) = 7.5 p < .001	Not confirmed p > .99	Not confirmed	None confirmed
H9: SPQ subscale 'Odd or eccentric behaviour' quadratically predicts cautious behavior, as indexed by an	Average distance from threat, Average distance from walls	3.2% (-0.9%-5.4%) F(2, 454) = 7.4 p < .001	Not confirmed p > .99	Not confirmed	None confirmed

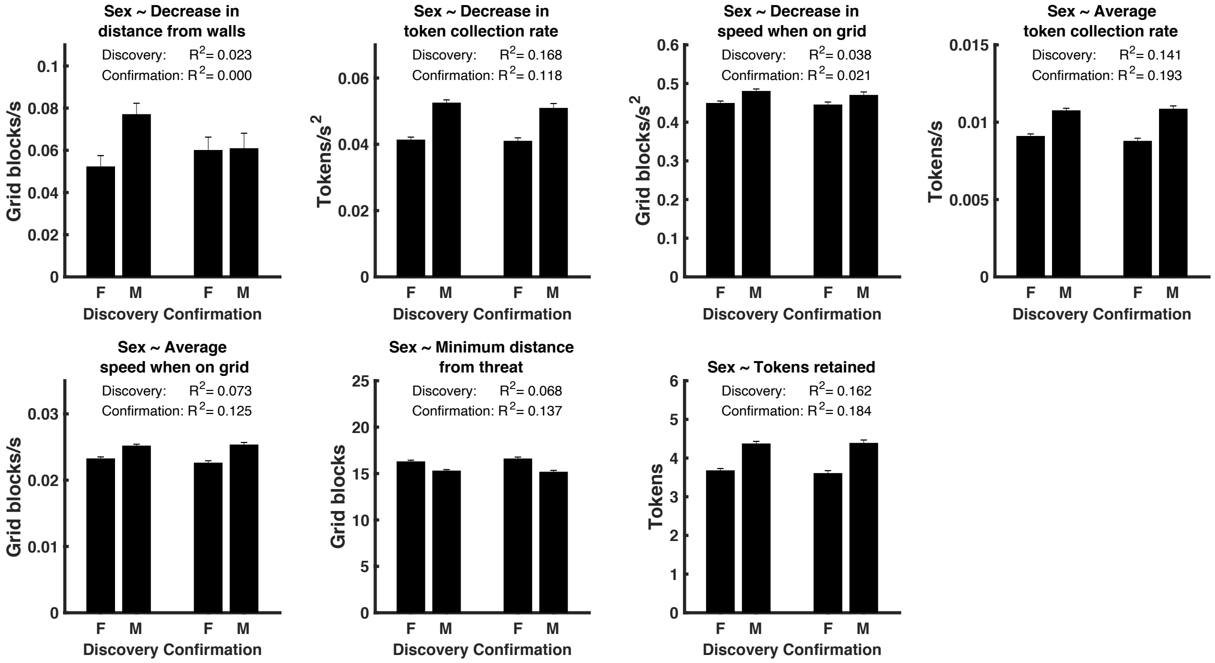
average of 2 normalized and recoded task measures					

All predictive models and their coefficients were part of the pre-registration (https://osf.io/hrce6/, registered on 28.07.2018). Confirmatory analysis was based on a random permutation test of model predictions in the hold-out sample, without re-fitting. The table shows accuracy and 95% bootstrap confidence intervals of the discovery model together with test statistics, on which the pre-registration was based. It also shows accuracy of the discovery model and 95% bootstrap confidence intervals in the confirmation sample as estimate of out-of-sample predictive performance [29], and significance level of a random permutation test. As best estimate of in-sample predictive performance and to compare with the out-of-sample predictive performance from the confirmation sample, we also show accuracy and 95% bootstrap confidence intervals of a model fitted in the combined sample. Notably, the procedure of deriving confidence intervals is unrelated to the permutation test; they are included due to journal requirements and do not reflect the posterior plausibility of true parameter values [65]. Accuracy is shown as percent correct for sex and percent variance explained for other variables (N/A for unconfirmed models). The last column highlights task variables contained in confirmed predictive models, for which the underlying bivariate relationship was not confirmed at $p < .001$ in the confirmation sample, even when not correcting for multiple comparison. See Supplementary Table 2 for a full list of bivariate relations in discovery and confirmation sample. See Extended Data Figures 2-4 for analysis of the predictive models separately in confirmation and discovery sample.

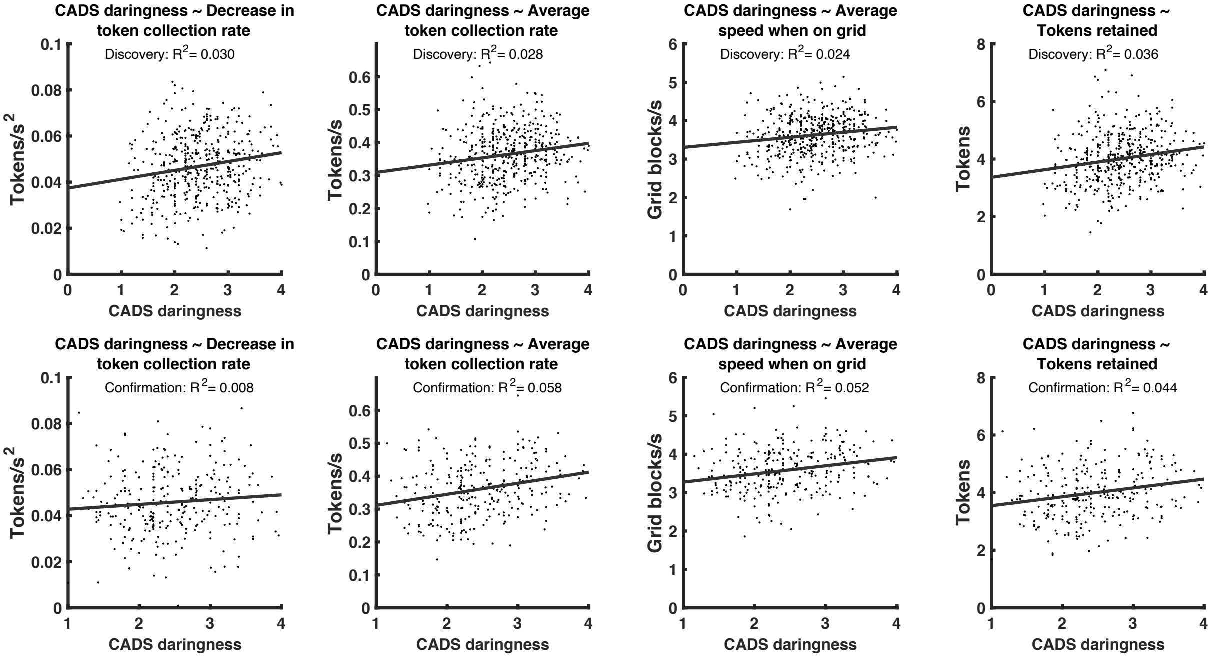
Extended Data Figures



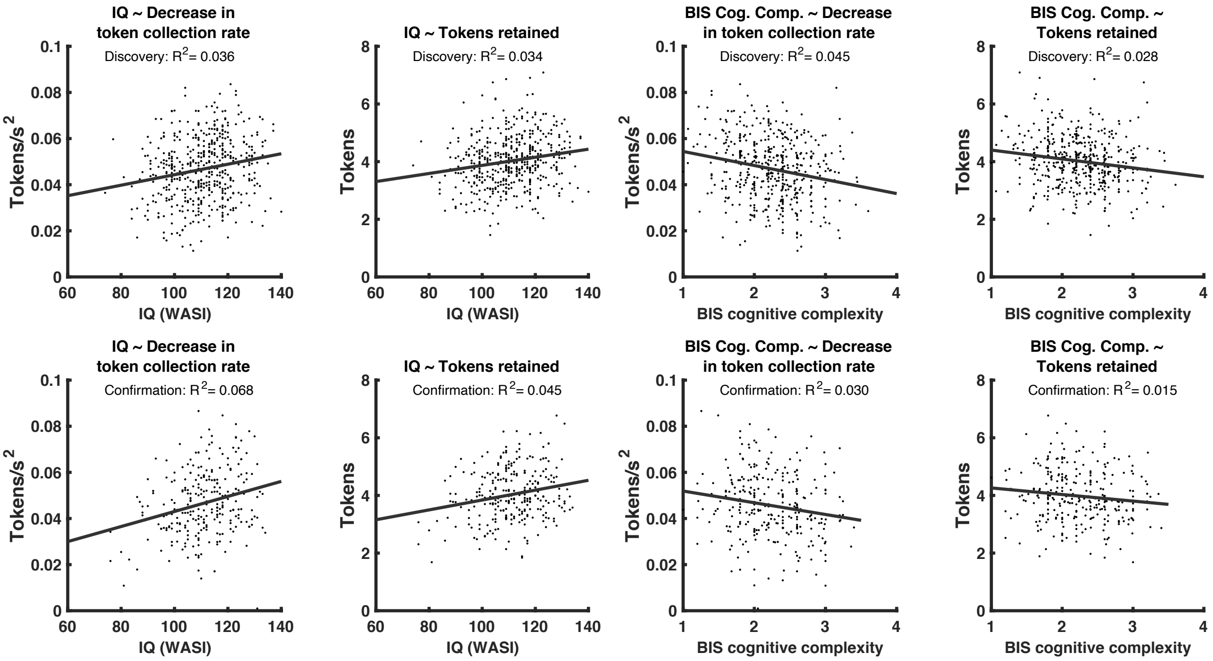
Extended Data Figure 1. Extraction of summary statistics from time-dependent variables. Four summary statistics are extracted for each of 7 time-dependent task measures, and for their time-dependent weighted sum (example data). Blue: low threat probability; orange: high threat probability. Example data are averaged over the active/passive (ie. starting position) factor.



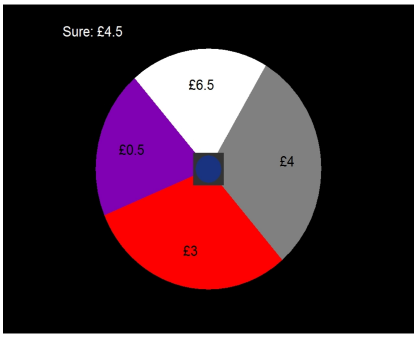
Extended Data Figure 2. Association of individual task variables with sex. Results from linear regressions fitted separately on discovery and confirmation sample. See supplementary table 2 for statistical tests of the individual relations. To confirm these associations collectively, we fitted a multiple logistic regression on the discovery data (registered hypothesis H1), which was confirmed. See Table 2 in main text for hypothesis summary and discovery/ confirmation results. A multiple logistic regression across the entire sample weakly favoured a model with common regression weights over one with separate weights for discovery and confirmation sample (LBF = 2.8).



Extended Data Figure 3. Association of individual task variables with CADS daringness. Results from linear regressions fitted separately on discovery and confirmation sample. See supplementary table 2 for statistical tests of the individual relations. To confirm these associations collectively, we computed a multiple regression model on the discovery data (registered hypothesis H4), which was confirmed. See Table 2 in main text for hypothesis summary and discovery/confirmation results. A multiple logistic regression across the entire sample favoured a model with common regression weights over one with separate weights for discovery and confirmation sample (LBF = 3.2). For the association of CADS with intra-epoch trajectories shown in Fig. 3 and Supplementary Table 2, we computed a multiple regression model with these three measures on the discovery data (registered hypothesis H7), which was confirmed (see Table 2). A multiple logistic regression across the entire sample weakly favoured a model with common regression weights over one with separate weights for discovery and confirmation sample (LBF = 2.3).



Extended Data Figure 4. Association of individual task variables with IQ and BIS cognitive complexity. Results from linear regressions fitted separately on discovery and confirmation sample. See supplementary table 2 for statistical tests of the individual relations. To confirm the associations with IQ collectively, we computed a multiple regression model on the discovery data (registered hypothesis H3), which was confirmed. See Table 2 in main text for hypothesis summary and discovery/confirmation results. A multiple logistic regression across the entire sample weakly favoured a model with common regression weights over one with separate weights for discovery and confirmation sample (LBF = 2.5). For BIS cognitive complexity, the multiple regression model (registered hypothesis H6) was confirmed as well (see Table 2). A multiple logistic regression across the entire sample weakly favoured a model with common regression weights over one with separate weights for discovery and confirmation sample (LBF = 2.7).



Extended Data Figure 5. Lottery (revealed economic preference) task. The roulette task involved a choice between the sure amount (upper left) and a four-sector roulette, just complex enough to define an Expectation, Variance and Skewness over roulette outcomes. The square in the middle of the roulette indicated a timer to maintain a reasonable pace of trials.

Supplementary tables

Supplementary Table 1: Neuroscience in Psychiatry consortium author list and affiliations.

Neuroscience in Psychiatry Network Study & Consortium Author list	
Principal Investigators	Edward Bullmore (CI from 01/01/2017) [1,2,3]
	Raymond Dolan [4,5]
	Ian Goodyer (CI until 01/01/2017) [1]
	Peter Fonagy [6]
	Peter Jones [1]
NSPN funded staff	Michael Moutoussis [4,5]
	Tobias Hauser [4,5]
	Sharon Neufeld [1]
	Rafael Romero-Garcia [1,2]
	Michelle St Clair [1]
	Petra Vértes [1,2]
	Kirstie Whitaker [1,2]
	Becky Inkster [1]
	Gita Prabhu [4,5]
	Cinly Ooi [1]
	Umar Toseeb [1]
	Barry Widmer [1]
	Junaid Bhatti [1]
	Laura Villis [1]
	Ayesha Alrumaithi [1]
	Sarah Birt [1]
	Aislinn Bowler [5]
	Kalia Cleridou [5]
	Hina Dadabhoy [5]
	Emma Davies [1]
	Ashlyn Firkins [1]
	Sian Granville [5]
	Elizabeth Harding [5]
	Alexandra Hopkins [4,5]
	Daniel Isaacs [5]
	Janchai King [5]
	Danae Kokorikou [5,6]
	Christina Maurice [1]
	Cleo McIntosh [1]
	Jessica Memarzia [1]
	Harriet Mills [5]
	Ciara O'Donnell [1]
	Sara Pantaleone [5]
	Jenny Scott [1]
Affiliated Scientists	Pasco Fearon [6]
	John Suckling [1]
	Anne-Laura van Harmelen [1]
	Rogier Kievit [4,7]
Affiliations	
1 Department of Psychiatry, University of Cambridge, United Kingdom	
2 Behavioural and Clinical Neuroscience Institute, University of Cambridge, United Kingdom	
3 ImmunoPsychiatry, GlaxoSmithKline Research and Development, United Kingdom	
4 Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, UK	
5 Wellcome Centre for Human Neuroimaging, University College London, United Kingdom	
6 Research Department of Clinical, Educational and Health Psychology, University College London, United Kingdom	
7 Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, United Kingdom	

Supplementary Table 2: Bivariate relationships of task variables with confirmed predictors. Bivariate relationships were selected in the discovery sample at an alpha level of p < .001. A joint model including all bivariately significant task measures per predictor variable was then confirmed without re-fitting in the hold-out sample (permutation test p < .05 after Holmes-Bonferroni correction for 9 comparisons). See figures S2-4 for the confirmation tests. In this table, p-values for the bivariate relationships serve for illustration purposes and are not corrected for multiple comparison. 95%-CI: 95% parametric confidence interval. For confidence intervals on correlation coefficients that include zero, an upper limit for R2 is given.

	R2	*t(df)*	*R2*	*t(df)*	*R2*
	(95%-CI)	*p*	*(95%-CI)*	*p*	*(95%-CI)*
	Discovery	*Discovery*	*Confirmation*	*Confirmation*	*Combined*
Sex					
Decrease wall distance	0.023	t(490)= 3.40	0	t(287)= 0.09	0.01
	(0.004-0.056)	p < .001	(< 0.015)	p = .93	(0.001-0.028)
Decrease token collection	0.168	t(490)= 9.94	0.118	t(287)= 6.21	0.147
	(0.111-0.231)	p < .001	(0.057-0.195)	p < .001	(0.104-0.195)
Decrease speed when on grid	0.038	t(490)= 4.37	0.021	t(287)= 2.49	0.03
	(0.012-0.077)	p < .001	(0.001-0.066)	p = .013	(0.011-0.058)
Average token collection	0.141	t(490)= 8.95	0.193	t(287)= 8.28	0.159
	(0.088-0.201)	p < .001	(0.116-0.278)	p < .001	(0.115-0.208)
Average speed when on grid	0.073	t(490)= 6.21	0.125	t(287)= 6.40	0.091
	(0.035-0.123)	p < .001	(0.062-0.203)	p < .001	(0.056-0.133)
Minimum distance from threat	0.068	t(490)= -5.96	0.137	t(287)= -6.76	0.09
	(0.031-0.116)	p < .001	(0.071-0.217)	p < .001	(0.055-0.132)
Tokens retained	0.162	t(490)= 9.74	0.184	t(287)= 8.03	0.17
	(0.106-0.225)	p < .001	(0.109-0.269)	p < .001	(0.124-0.220)
CADS daringness					
Decrease token collection	0.03	t(459)= 3.80	0.008	t(272)= 1.50	0.02
	(0.007-0.069)	p < .001	(< 0.043)	p = .14	(0.005-0.045)
Average token collection	0.028	t(459)= 3.66	0.058	t(272)= 4.09	0.039
	(0.006-0.065)	p < .001	(0.016-0.122)	p < .001	(0.016-0.070)
Average speed when on grid	0.024	t(459)= 3.32	0.052	t(272)= 3.87	0.034
	(0.004-0.058)	p < .001	(0.013-0.114)	p < .001	(0.013-0.064)
Tokens retained	0.036	t(459)= 4.12	0.044	t(272)= 3.54	0.039
	(0.010-0.076)	p < .001	(0.009-0.103)	p < .001	(0.016-0.071)
CADS daringess and trajectory similarity					
Wall distance	0.027	t(437)= 3.47	0.053	t(272)= 3.91	0.037
	(0.005-0.064)	p < .001	(0.013-0.116)	p < .001	(0.014-0.068)
Token collection	0.053	t(437)= 4.94	0.069	t(272)= 4.48	0.058
	(0.019-0.100)	p < .001	(0.022-0.136)	p < .001	(0.029-0.096)
Speed on grid	0.032	t(437)= 3.82	0.051	t(272)= 3.83	0.039
	(0.008-0.072)	p < .001	(0.012-0.113)	p < .001	(0.016-0.071)
IQ					
Decrease token collection	0.036	t(488)= 4.26	0.068	t(268)= 4.41	0.046
	(0.010-0.075)	p < .001	(0.021-0.135)	p < .001	(0.021-0.079)

Tokens retained	0.034 (0.010-0.072)	t(488)= 4.15 p < .001	0.045 (0.009-0.105)	t(268)= 3.57 p < .001	0.038 (0.016-0.069)
BIS cognitive complexity					
Decrease token collection	0.045 (0.015-0.088)	t(460)= -4.63 p < .001	0.03 (0.003-0.081)	t(274)= -2.89 p = .004	0.038 (0.015-0.069)
Tokens retained	0.028 (0.006-0.065)	t(460)= -3.65 p < .001	0.015 (0.000-0.055)	t(274)= -2.01 p = .046	0.022 (0.006-0.048)

Supplementary Table 3. Child and Adolescent Disposition Scale 'Daringness' (the other CADS items are omitted for convenience).

These questions are of your personality. When you answer these questions, please think about the last 12 months and tick the box that you feel best describes you.

	Not at all	**Just a little**	**Pretty much / pretty often**	**Very much / very often**
3. Are you daring and adventurous?				
6. Do you like rough games and sports?				
9. Do you enjoy doing things that are risky or dangerous?				
11. Do you like things that are exciting and loud?				
50. Are you brave?				

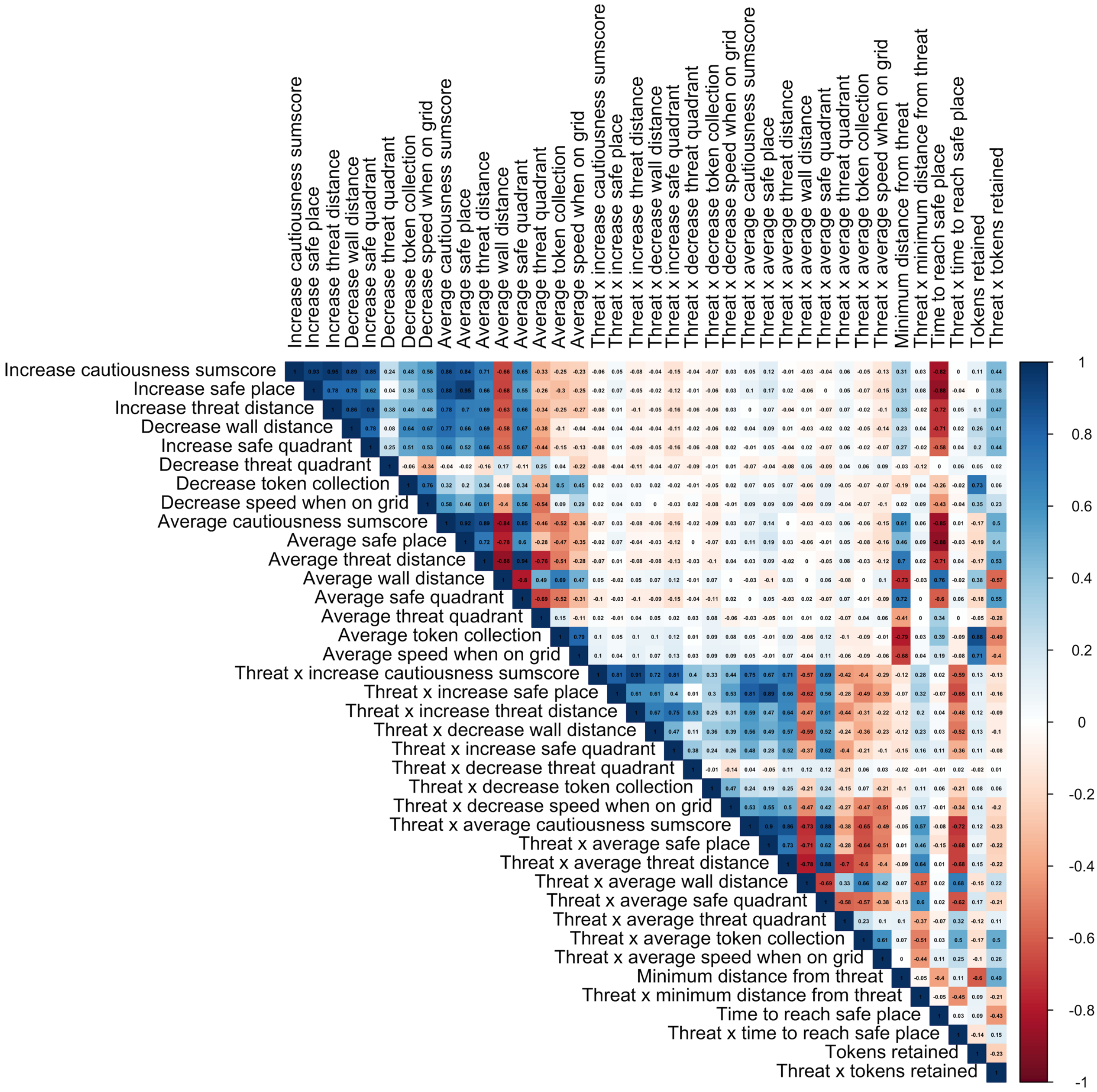
Supplementary Table 4. Extract from Barratt Impulsivity Scale extract, including the full Cognitive Complexity subscale, which is boxed in bold. The rest of the BIS is omitted for convenience.

People differ in the ways they act and think in different situations. This is a test to measure some of the ways in which you act and think. Read each statement and put and tick in the appropriate box. Do not spend too much time on any statement. Answer quickly and honestly.

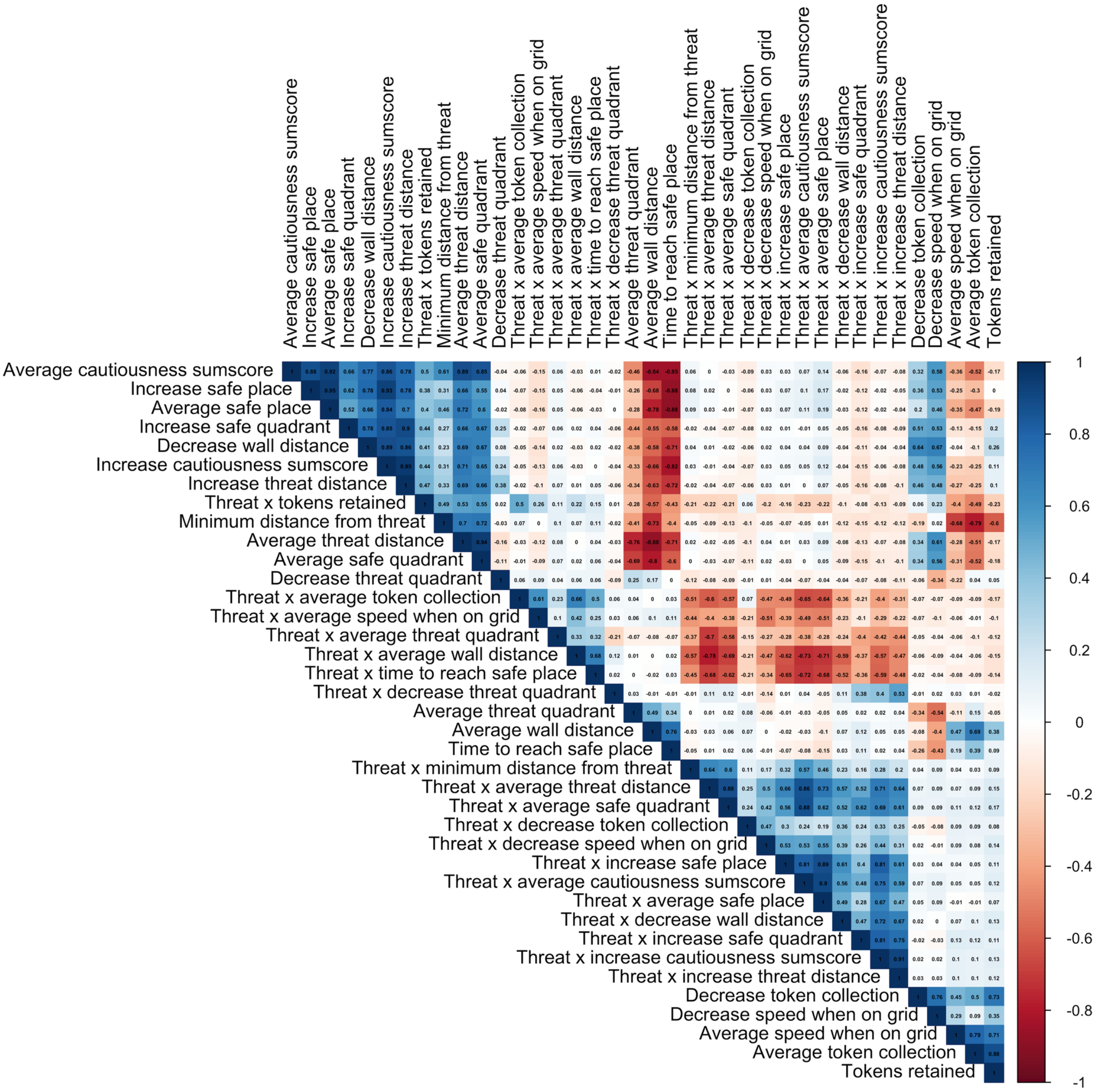
	Rarely	**Occasionally**	**Often**	**Always**
1. I plan tasks carefully.				
2. I do things without thinking.				
3. I make-up my mind quickly.				
4. I am happy go lucky.				
5. I don't "pay attention."				
6. I have "racing thoughts."				
7. I plan trips well ahead of time.				
8. I am self-controlled.				
9. I concentrate easily.				
10. I save regularly.				
11. I "squirm" at plays and lectures.				
12. I am a careful thinker.				
…				

Supplementary results

1. Psychometric properties of the task



Supplementary figure 1. Correlation matrix of the 38 task variables at BSL, for the combined sample. Variables appear in systematic order.



Supplementary Figure 2. Correlation matrix of the 38 task variables at BSL, for the combined sample. This is the same as Supplementary Figure 1 but with variables ordered by hierarchical clustering for better visualisation.

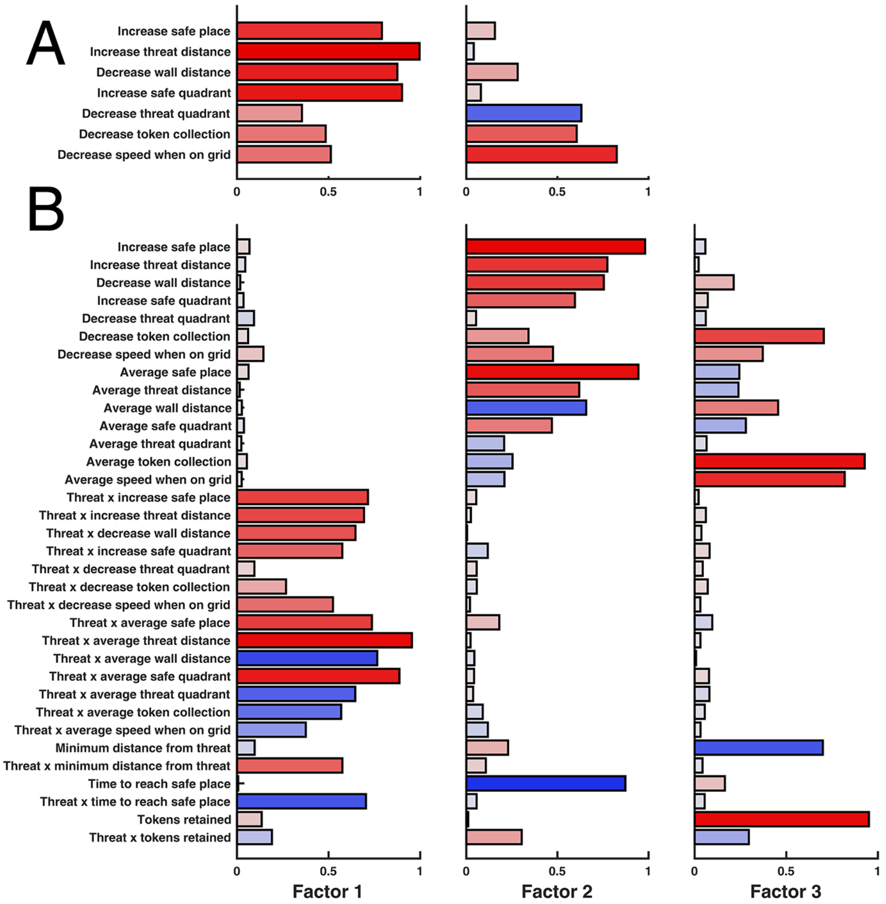
Supplementary Table 5: Consistency (Cronbach's alpha) at BSL, and test-retest reliability from baseline to FU-1 (over 11-32 months). Cronbach's alpha is for linear adaptation in the 7 time-dependent measures at baseline. Factor scores for discovery and confirmation samples are based on factor analysis of discovery sample. Factor scores for the combined sample are based on a factor analysis of the combined sample.

	Discovery	**Confirmation**	**Combined**
Cronbach's alpha	0.87	0.86	0.86
Increase cautiousness sum score	0.521	0.503	0.514
Increase safe place	0.584	0.516	0.558
Increase threat distance	0.434	0.430	0.431
Decrease wall distance	0.562	0.573	0.566
Increase safe quadrant	0.356	0.381	0.365
Decrease threat quadrant	0.403	0.361	0.378
Decrease token collection	0.795	0.735	0.771
Decrease speed when on grid	0.600	0.623	0.608
Average cautiousness sum score	0.561	0.515	0.542
Average safe place	0.591	0.507	0.558
Average threat distance	0.510	0.533	0.517
Average wall distance	0.518	0.556	0.534
Average safe quadrant	0.518	0.502	0.509
Average threat quadrant	0.388	0.450	0.407
Average token collection	0.704	0.659	0.686
Average speed when on grid	0.679	0.684	0.681
Threat × increase cautiousness sum score	0.041	0.068	0.049
Threat × increase safe place	0.049	-0.017	0.025
Threat × increase threat distance	0.047	0.098	0.063
Threat × decrease wall distance	-0.029	0.008	-0.014
Threat × increase safe quadrant	0.081	0.108	0.090
Threat × decrease threat quadrant	-0.019	-0.034	-0.025
Threat × decrease token collection	0.040	0.093	0.063
Threat × decrease speed when on grid	-0.016	0.007	-0.007
Threat × average cautiousness sum score	0.150	0.084	0.124
Threat × average safe place	0.138	0.047	0.103
Threat × average threat distance	0.086	0.183	0.125
Threat × average wall distance	0.055	0.088	0.069
Threat × average safe quadrant	0.127	0.134	0.130
Threat × average threat quadrant	0.083	0.162	0.113
Threat × average token collection	0.127	-0.016	0.070
Threat × average speed when on grid	0.049	-0.042	0.014
Minimum distance from threat	0.590	0.551	0.575
Threat × minimum distance from threat	0.007	0.080	0.035
Time to reach safe place	0.536	0.465	0.508
Threat × time to reach safe place	0.114	0.074	0.099

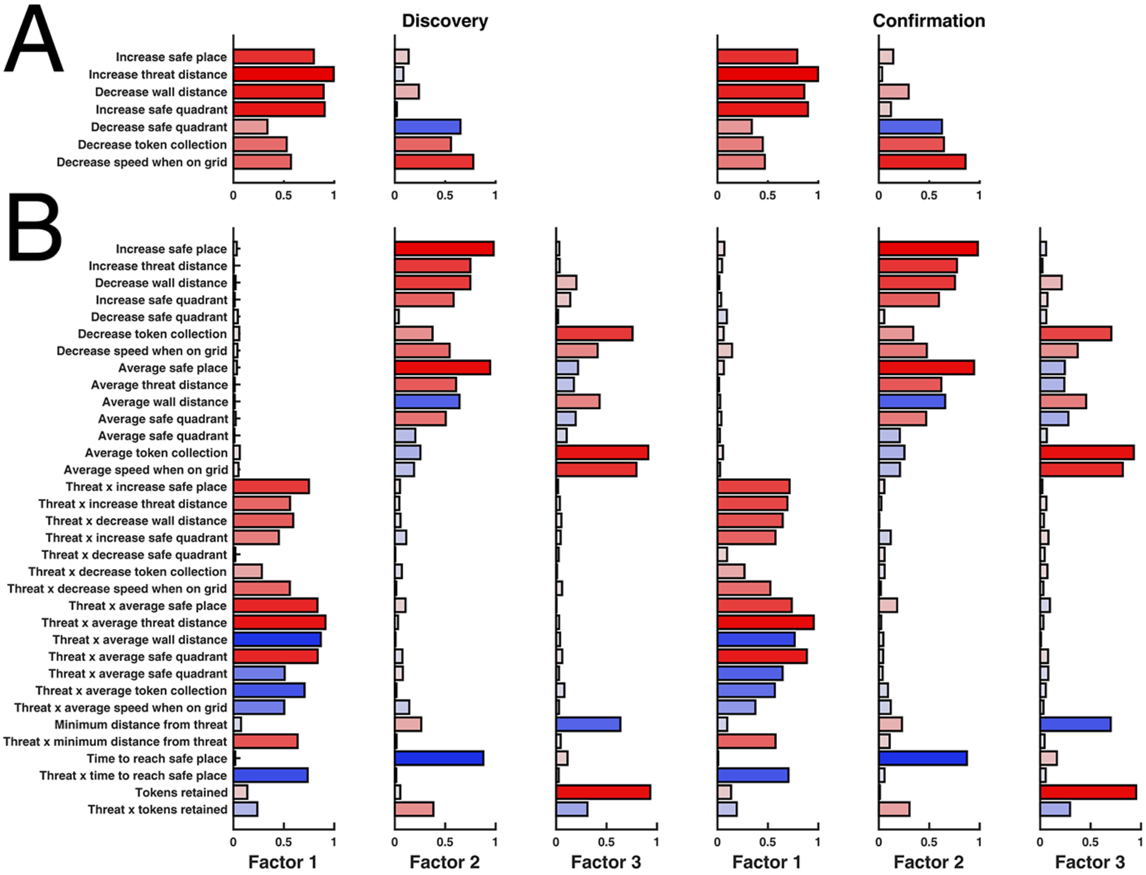
Tokens retained	0.696	0.677	0.689
Threat × tokens retained	0.17	0.172	0.173
Factor 1 7-msr FA	0.541	0.531	0.528
Factor 2 7-msr FA	0.640	0.619	0.639
Factor 1 <Sensitivity to threat probability> 34-msr FA	0.090	0.098	0.096
Factor 2 <Sensitivity to intra-epoch time> 34-msr FA	0.522	0.535	0.531
Factor 3 <Performance> 34-msr FA	0.708	0.694	0.69

Linear adaptation in the 7 previously reported task measures showed a high internal consistency as indexed by Cronbach's alpha = .86 in the combined sample (see Supplementary Table 5). Nevertheless, parallel analysis suggested a 2-factor solution for these 7 measures. After varimax rotation, 4 measures loaded dominantly onto one factor, and three on the other (see Supplementary Figure 3A). This factor analysis replicated between discovery and confirmation sample (see Supplementary Figure 4A). In parallel analysis using 34 task measures (excluding the 4 collinear sum scores), a 6-factor solution was preferred. The first 3 factors could meaningfully be interpreted and replicated over partitions of the discovery sample (see Supplementary Figure 3B). These three factors replicated between the discovery and confirmation sample (see Supplementary Figure 4B).

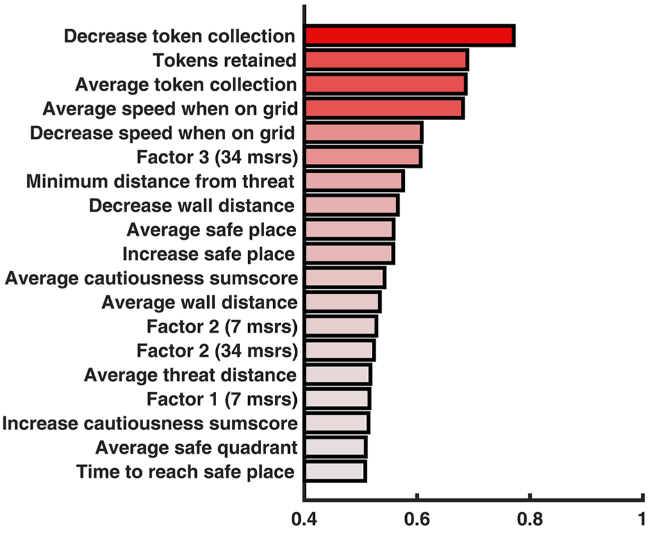
Test-retest reliability between BSL and FU-1 (e.g. over 1-3 years) was larger than r_{tt} = .40 for most measures unrelated to the predator differences, and was high as r_{tt} = .70 for some measures related to performance (see Supplementary Table 5, see Supplementary Figure 5 for measures with r_{tt} > .50 in the combined sample). Overall, although not designed to do so, it appears that the task quantifies traits that are stable over time, particularly the case for those measures predicted by sex, IQ, and self-report variables.



Supplementary Figure 3: Exploratory factor analysis (EFA) with varimax rotation in the combined sample. A: EFA on linear adaptation in 7 previously reported time-sensitive measures. B: EFA on all 34 independent task measures (excluding 4 linearly dependent sum scores). See Supplementary Figure 4 for comparison between discovery and confirmation sample.



Supplementary Figure 4: Factor loadings from two exploratory factor analyses (EFA) on discovery and confirmation samples with varimax rotation. Positive loadings are in red shades, negative in blue. The factor loadings almost perfectly replicated between discovery and confirmation data set. Specifically, we computed factor scores in the confirmation data set, either using loadings derived from the discovery data set or loadings derived from the confirmation data set. For both EFAs, the two factor scores were highly correlated. EFA with 7 measures: Factor 1, r > 0.99 (0.99-1.00, 95% parametric confidence interval; t(287) = 364.8; p < .001); Factor 2, r = 0.99 (0.99-1.00, 95% parametric confidence interval; t(287) = 162.7; p < .001). EFA with 34 measures: Factor 1, r > 0.99 (0.99-1.00, 95% parametric confidence interval; t(287) = 211.0; p < .001); Factor 2, r > 0.99 (0.99-1.00, 95% parametric confidence interval; t(287) = 234.6.8; p < .001), Factor 3, r > 0.99 (0.99-1.00, 95% parametric confidence interval; t(287) = 139.6; p < .001). To further confirm the factor structure of the 34-measure EFA, we defined a confirmatory factor model as structural equation model that included only factor loadings above 0.2. However, this model did not converge.



Supplementary Figure 5: Test-retest-reliability from BSL to FU-1 for the combined sample, showing all measures for which rtt > .50. See Supplementary Table 5 for a full list split into discovery and confirmation sample.

2. Mediation analysis

Here we report effect sizes (but, following journal policy, no inference statistics) from post-hoc mediation analysis across the combined sample. For each predictor, we constrained this mediation analysis to task variables that were related to this predictor, and other predictor variables that were themselves related to task variables, namely sex, IQ, CADS daringness, and BIS cognitive complexity.

Sex difference in performance - mediation by task variables

Average token collection mediated the largest proportion of the performance difference between sexes (82%; 75%-90%; 95% bootstrap confidence interval). After accounting for this variable, the rate of decrease in token collection (as the epoch progressed) carried the next highest proportion of mediation (13%; 8%-19%). After accounting for both these variables, estimated proportion of mediation was 0% (-2%-0%) for minimum distance, 0% (-2%-1%) for decrease in speed, 0% (-2%-1%) for average speed, and -2% (-5%-1%) for decrease in wall distance. At the request of a reviewer, we analysed whether the slope of the trial-by-trial performance trajectory mediated the sex effect on performance. This variable mediated 3% (1%-6%). After accounting for average token collection, which mediated a much higher proportion of variance, the individual performance slope mediated 0% (0%-1%).

Sex difference in performance - mediation by other variables related to performance

Among the variables considered, self-reported daringness mediated the highest proportion of the sex effect (5%, 0%-11%), while IQ mediated 4% (1%-7%), and BIS cognitive complexity 3% (1%-6%). Because of the relatively small proportion of mediation, we did not investigate the unique contribution of these covariates.

CADS daringness effect on performance - mediation by task variables and by other variables related to performance

Average token collection mediated the greatest proportion of a daringness effect on performance (87%; 70%-107%), similar to the sex effect on performance. After accounting for this variable, the rate of decrease in token collection (as the epoch progressed) carried the next highest proportion of mediation (2%; -1%-6%), while average speed mediated 0% (-1%-1%). Regarding other predictor variables, IQ mediated -1% (-1%-5%) and BIS cognitive complexity mediated 3% (-2%-10%) of the daringness effect on performance.

IQ effect on performance - mediation by task variables and by other variables related to performance

Decrease in token collection mediated 80% of an IQ effect on performance (80%; 62%-104). No other task variables were related to IQ. Self-reported daringness mediated -2% (-12%-6%) and self-reported cognitive complexity mediated 13% (3%-30%)   of the IQ effect on performance

BIS cognitive complexity effect on performance - mediation by task variables and by other variables related to performance

Decrease in token collection mediated 94% of a BIS cognitive complexity effect on performance (68%-151%). No other task variables were related to cognitive complexity. Self-reported daringness mediated 6% (-4%-19%) and IQ mediated 33% (17%-73%) of a BIS cognitive complexity on performance. Thus, both IQ and BIS cognitive complexity may have a separate impact on task performance. The proportion of cognitive complexity variance mediated by IQ was descriptively higher than vice versa (13%, see above).

3. Analysis of performance trajectory over trials (H2)

Under a hypothesis that males performed better because they had more experience with computer game, we expected that females may improve their performance more over trials than male. We first determined the curvature of the performance trajectory across both sexes and all conditions. Model evidence indicated that a logarithmic trajectory fitted the data decisively better than a linear, quadratic, or square root trajectory (LBF -44 with respect to the linear model). We then computed a threat level × task × trial × sex linear effects model with trial as logarithmic predictor across all conditions. This model revealed a significant trial × sex interaction (F(1, 30982) = 9.66; p < .001), but in the opposite direction than expected, i.e. male participant increased their performance over time more than females. However, this exploratory finding was not replicated in the confirmation sample (H2, F(1, 18193) = 0.2; p = .65). Instead, we found a significant threat level × trial × sex interaction (F(1, 18193.1) = 10.46; p < .001), implying that males increase their performance more than females only in the high threat probability condition. Across the entire sample, both the trial × sex (F(1, 49189) = 9.79; p = .002 uncorrected) and the threat level × trial × sex (F(1, 49541) = 4.15; p < .042 uncorrected) interactions were significant.

4. Analysis of maturation effects

A subsample of n = 63 participants, distributed across discovery and confirmation sample, returned 6 months after the first visit (BSL) and played the game again (visit FU-R). N = 55 of these participants also came back 11-32 months after visit 1 for another session (FU-1). Many task measures changed in the 6 months between BSL and FU-R, while over a much longer (5-26-month) interval between FU-R and FU-1, only four measures changed systematically, indicating practice as opposed to maturation effects. A full list of changes is found in Supplementary Table 6 below.

To address further whether observed systematic changes between visits were due to maturation or practice, we capitalised on the variable interval between BSL and FU-1, availing of the larger number of participants who took part in BSL and FU-1 (after 11-32 months), but not necessarily in FU-R. Thus, in the discovery sample (N = 357) we computed a model that included task measures at visits BSL and FU-1 as dependent variables, and repetition and time interval between BSL and FU-1 as predictors. We observed an effect of repetition in many measures, but for no variable was this effect better explained by the time elapsed between the two visits. For each of the 38 task measures, Bayesian model comparison favoured a model with repetition but without time effects (LBF > 3 in favour of the simpler model). This finding was corroborated in both the confirmation sample (N = 210) and the combined sample (N = 567). Overall, we found no evidence for any effects of maturation on task behaviour.

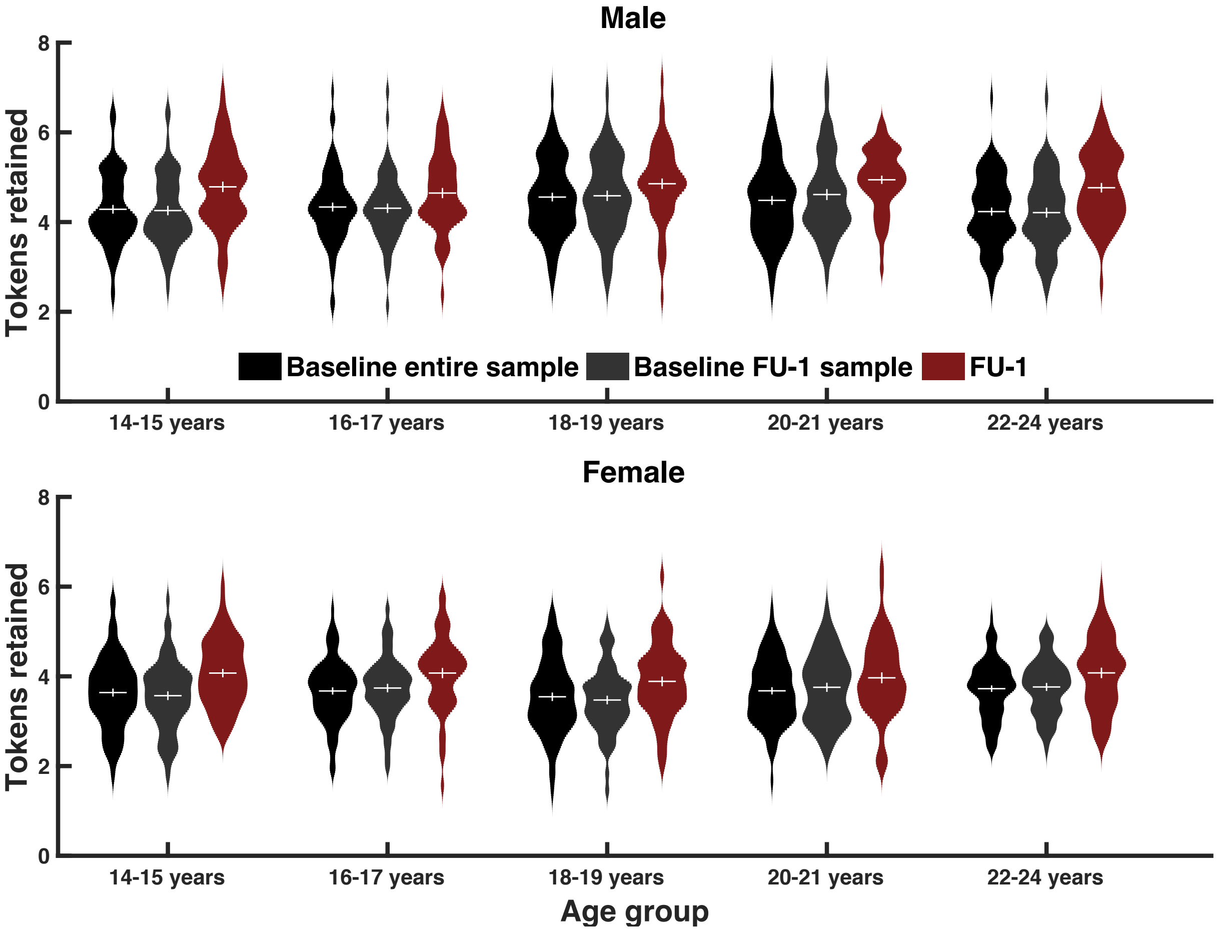
Finally, at the request of a reviewer, we computed, for each task variable, an age × (repetition + time interval) linear mixed effects model across the combined sample, and contrasted this with an age × repetition model. There were no significant age × time interval interactions at our alpha threshold of $p < .001$. Even at $p < .05$, there was only one significant finding, but here the simpler model explained the data better than the more complex one (BF difference > 700 in favour of the simpler model). Thus, we did not find any evidence that age moderates the impact of maturation in this sample. The impact of age on repetition of the task is illustrated in Supplementary Figure 11 below.

Supplementary Table 6. Effect of maturation: change in 38 task measures between BSL and FU-R, or FU-R and FU-1.

	BSL-FU-R		FU-R - FU-1	
	Cohen's d	t(df)	Cohen's d	t(df)
	(95% CI)	p	(95% CI)	p
Increase cautiousness sumscore	1.35	t(62) = 5.38	0.31	t(54) = 1.16
	(0.80-1.90)	p < .001	(-0.22-0.84)	p = .25
Increase safe place	1.34	t(62) = 5.33	0.23	t(54) = 0.87
	(0.79-1.89)	p < .001	(-0.30-0.76)	p = .39
Increase threat distance	1.07	t(62) = 4.23	0.28	t(54) = 1.03
	(0.53-1.59)	p < .001	(-0.25-0.81)	p = .31
Decrease wall distance	1.4	t(62) = 5.55	-0.06	t(54) = -0.22
	(0.84-1.95)	p < .001	(-0.59-0.47)	p = .82
Increase safe quadrant	0.92	t(62) = 3.66	0.51	t(54) = 1.90
	(0.40-1.44)	p < .001	(-0.03-1.05)	p = .062
Decrease threat quadrant	-0.2	t(62) = -0.78	0.18	t(54) = 0.66
	(-0.69-0.30)	p = .44	(-0.35-0.71)	p = .51
Decrease token collection	0.95	t(62) = 3.77	0.29	t(54) = 1.08
	(0.43-1.47)	p < .001	(-0.24-0.82)	p = .28
Decrease speed when on grid	1.14	t(62) = 4.54	-0.11	t(54) = -0.39
	(0.61-1.67)	p < .001	(-0.63-0.42)	p = .70
Average cautiousness sumscore	1.32	t(62) = 5.25	-0.28	t(54) = -1.03
	(0.77-1.86)	p < .001	(-0.81-0.25)	p = .31
Average safe place	1.37	t(62) = 5.42	-0.07	t(54) = -0.27
	(0.81-1.91)	p < .001	(-0.60-0.46)	p = .79
Average threat distance	1.24	t(62) = 4.92	-0.48	t(54) = -1.79
	(0.69-1.78)	p < .001	(-1.02-0.06)	p = .079
Average wall distance	-1.09	t(62) = -4.33	0.17	t(54) = 0.64
	(-1.62--0.56)	p < .001	(-0.36-0.70)	p = .52
Average safe quadrant	0.83	t(62) = 3.29	-0.59	t(54) = -2.19
	(0.31-1.34)	p = .002	(-1.13--0.05)	p = .033
Average threat quadrant	-0.72	t(62) = -2.86	0.22	t(54) = 0.81
	(-1.23--0.21)	p = .006	(-0.31-0.75)	p = .42
Average token collection	0.04	t(62) = 0.18	0.69	t(54) = 2.54
	(-0.45-0.54)	p = .86	(0.14-1.23)	p = .014
Average speed when on grid	-0.23	t(62) = -0.92	0.41	t(54) = 1.51
	(-0.73-0.26)	p = .36	(-0.13-0.94)	p = .14
Threat x increase cautiousness sumscore	0.07	t(62) = 0.27	0.09	t(54) = 0.32
	(-0.43-0.56)	p = .79	(-0.44-0.61)	p = .75
Threat x increase safe place	0.11	t(62) = 0.45	0.12	t(54) = 0.43
	(-0.38-0.61)	p = .65	(-0.41-0.64)	p = .67
Threat x increase threat distance	-0.1	t(62) = -0.42	0.21	t(54) = 0.76
	(-0.60-0.39)	p = .68	(-0.32-0.74)	p = .45
Threat x decrease wall distance	-0.11	t(62) = -0.44	0.37	t(54) = 1.39
	(-0.60-0.39)	p = .67	(-0.16-0.91)	p = .17
Threat x increase safe quadrant	0.08	t(62) = 0.33	-0.25	t(54) = -0.94

	(-0.41-0.58)	p = .74	(-0.78-0.28)	p = .35
Threat x decrease threat quadrant	0.08	t(62) = 0.31	0.13	t(54) = 0.48
	(-0.42-0.57)	p = .75	(-0.40-0.66)	p = .63
Threat x decrease token collection	0.17	t(62) = 0.66	0.19	t(54) = 0.69
	(-0.33-0.66)	p = .51	(-0.34-0.72)	p = .49
Threat x decrease speed when on grid	0.15	t(62) = 0.61	-0.03	t(54) = -0.12
	(-0.34-0.65)	p = .54	(-0.56-0.50)	p = .91
Threat x average cautiousness sumscore	0.32	t(62) = 1.25	0.09	t(54) = 0.33
	(-0.18-0.81)	p = .21	(-0.44-0.62)	p = .74
Threat x average safe place	0.28	t(62) = 1.09	0.09	t(54) = 0.34
	(-0.22-0.77)	p = .28	(-0.44-0.62)	p = .73
Threat x average threat distance	0.35	t(62) = 1.38	0.09	t(54) = 0.34
	(-0.15-0.84)	p = .17	(-0.44-0.62)	p = .73
Threat x average wall distance	-0.25	t(62) = -0.98	-0.2	t(54) = -0.73
	(-0.74-0.25)	p = .33	(-0.72-0.33)	p = .47
Threat x average safe quadrant	0.39	t(62) = 1.53	0.04	t(54) = 0.15
	(-0.11-0.88)	p = .13	(-0.49-0.57)	p = .88
Threat x average threat quadrant	-0.52	t(62) = -2.05	0.15	t(54) = 0.54
	(-1.02--0.01)	p = .045	(-0.38-0.67)	p = .59
Threat x average token collection	-0.08	t(62) = -0.33	-0.02	t(54) = -0.08
	(-0.58-0.41)	p = .74	(-0.55-0.51)	p = .94
Threat x average speed when on grid	0.06	t(62) = 0.23	-0.19	t(54) = -0.70
	(-0.44-0.55)	p = .82	(-0.72-0.34)	p = .49
Minimum distance from threat	0.73	t(62) = 2.90	-1.17	t(54) = -4.33
	(0.22-1.24)	p = .005	(-1.74--0.59)	p < .001
Threat x minimum distance from threat	0.61	t(62) = 2.43	0.09	t(54) = 0.32
	(0.11-1.12)	p = .018	(-0.44-0.62)	p = .75
Time to reach safe place	-1.72	t(62) = -6.81	-0.22	t(54) = -0.83
	(-2.29--1.13)	p < .001	(-0.75-0.31)	p = .41
Threat x time to reach safe place	-0.2	t(62) = -0.80	-0.2	t(54) = -0.76
	(-0.70-0.29)	p = .43	(-0.73-0.33)	p = .45
Tokens retained	0.98	t(62) = 3.89	0.62	t(54) = 2.31
	(0.45-1.50)	p < .001	(0.08-1.16)	p = .025
Threat x tokens retained	1.02	t(62) = 4.04	-0.34	t(54) = -1.27
	(0.49-1.54)	p < .001	(-0.87-0.19)	p = .21

95%-CI: 95% parametric confidence interval.



Supplementary Figure 6. Distribution of performance (tokens retained) split by sex and age group, for BSL and FU-1. White lines show mean and standard error of the mean.

5. Relation of task variables and parameters from an economic lottery

Supplementary table 7. Post-hoc analysis of bivariate relationships of the task variables that related to predictors to paramters from the economic lottery task. In line with journal policy, we state effect sizes but no inference statistics. 95%-CI: 95% parametric confidence interval. For confidence intervals on correlation coefficients that include zero, an upper limit for R2 is given.

	Preference for variable gambles	Preference for skewed gambles	Choice temperature
	R2	R2	R2
	(95%-CI)	(95%-CI)	(95%-CI)
Decrease wall distance	0.000	0.001	0.002
	(0.000-0.007)	(0.000-0.009)	(0.000-0.014)
Decrease token collection	0.004	0.001	0.012
	(0.000-0.019)	(0.000-0.010)	(0.002-0.033)
Decrease speed when on grid	0.001	0.000	0.001
	(0.000-0.012)	(0.000-0.006)	(0.000-0.009)
Average token collection	0.017	0.005	0.013
	(0.004-0.040)	(0.000-0.020)	(0.002-0.034)
Average speed when on grid	0.011	0.004	0.003
	(0.001-0.030)	(0.000-0.019)	(0.000-0.015)
Minimum distance from threat	0.014	0.006	0.003
	(0.002-0.035)	(0.000-0.022)	(0.000-0.015)
Tokens retained	0.013	0.002	0.020
	(0.002-0.034)	(0.000-0.013)	(0.005-0.044)

6. Analysis of the task parameter extraction method

Because the epochs had variable duration, fewer data points were available later in the epoch compared to early in the epoch. In previous publications we had used mean imputation, i.e. we ignored missing data points when computing the average over trials for each time bin. The same strategy was used in the current analysis of trajectory similarity. To compute regression coefficients for these averaged trajectories however, it may be appropriate to take into account the different number of available data points per time bin, for example using weighted linear squares regressions with estimated coefficients

$$\hat{\beta} = (X^T W X)^{-1} X^T W y;$$

where y is the vector of data per time bin (averaged over trials), X the design matrix, and W a diagonal matrix that contains, for each time bin, the proportion of available observations out of all trials.

Due to a coding error that was detected after pre-registration, we had instead used zero imputation before averaging trajectories over trials, followed by ordinary least squares regression. Using this method, the resulting coefficients are of the form

$$\hat{\beta} = (X^T X)^{-1} X^T W y;$$

where y is the averaged trajectory one would have obtained using mean imputation. It is easy to see that these coefficients span the same space as the ones obtained in the aforementioned WLS approach, such that they do not contain different information.

However, we note that the discrete criterion for inclusion of any task measure into our predictive models may results in different predictive accuracy, and that the interpretation of the coefficients in these models may be different. This is why we replicated the original analysis, using mean imputation and WLS regression. First, we refitted the preregistered predictive models in the discovery sample, using the same nominal task variables, and tested their predictive performance in the confirmation sample (Supplementary Table 8). All models were confirmed, such that our key findings can be seen as independent from the method of task variable extraction. Second, we used the same approach of including task variables into the predictive models at an alpha threshold of $p < .001$ for the bivariate relationship. The confirmed models from this set related to the same predictor variables as in our original analysis. We note that fewer task variables were included and that the predictive performance was descriptively lower than in the original analysis. Third, we summarise all bivariate relationships relating to our original models in Supplementary Table 9.

Supplementary Table 8: Joint predictive models and their performance when using WLS-derived task parameters instead of OLS parameters with zero imputation. Further models predicting CADS subscale "prosocial behaviour", RCMAS, and SPQ subscale "odd or eccentric behaviour" in the discovery sample were not confirmed and are not shown here. Notably, the procedure of deriving confidence intervals is unrelated to the permutation test; they are included due to journal requirements and do not reflect the posterior plausibility of true parameter values [65].

Predictive model	**Task variables in predictive model**	**Discovery sample: accuracy (95% bootstrap confidence interval) of the joint predictive model and parametric test**	**Confirmation sample: accuracy (95% bootstrap confidence interval) of the discovery model; significance level (uncorrected) from non-parametric random permutation test**
Sex: model with pre-registered task variables	Decrease in distance from walls, Decrease in token collection rate, Decrease in speed when on grid, Average token collection rate, Average speed when on grid, Minimum distance from threat, Tokens retained	71.3% (66.4%-75.2%) $\chi^2(7) = 126.4$, $p < .001$	68.2% (63.0%-73.7%) $p < .001$
Sex: model with new inclusion of task variables	Decrease in token collection rate, Average token collection rate, Average speed when on grid, Minimum distance from threat, Tokens retained	65.9% (67.8%-74.8%) $\chi^2(5) = 120.8$ $p < .001$	68.2% (63.0%-73.7%) $p < .001$
IQ: model with pre-registered task variables	Decrease in token collection rate, Tokens retained	4.2% (-0.1%-6.8%) F(2, 487) = 10.7 $p < .001$	7.5% (3.2%-12.9%) $p < .001$
IQ: model with new inclusion of task variables	Tokens retained	3.4% (-0.3%-5.8%) F(1, 488) = 17.24 $p < .001$	4.8% (0.7%-10.2%) $p < .001$
CADS daringness: model with pre-registered task variables	Decrease in token collection rate, Average token collection rate, Average speed, Tokens retained	3.9% (-0.1%-6.8%) F(4, 456) = 4.7 $p = .001$	4.3% (-0.2%-9.6%) $p < .001$
CADS daringness: model with new inclusion of task variables	Tokens retained	3.6% (-0.5%-6.1%) F(1, 459) = 17.0 $p < .001$	4.7% (0.5%-9.9%) $p < .001$
BIS cognitive complexity: model with pre-registered task variables (identical: with new inclusion of task variables)	Decrease in token collection rate, Tokens retained	4.2% (-0.5%-6.9%) F(2, 459) = 9.9 $p < .001$	3.5% (-0.6%-8.7%) $p = .001$

Supplementary Table 9: Bivariate relationships of task variables extracted with WLS regression with confirmed predictors. Bivariate relationships are shown if they were included in the preregistered analysis; no new task variables were included in the discovery analysis with WLS regression. In this table, p-values for the bivariate relationships serve for illustration purposes and are not corrected for multiple comparison. 95%-CI: 95% parametric confidence interval. For confidence intervals on correlation coefficients that include zero, an upper limit for R2 is given.

	R2 *(95%-CI)* *Discovery*	*t(df)* *p* *Discovery*	*R2* *(95%-CI)* *Confirmation*	*t(df)* *p* *Confirmation*	*R2* *(95%-CI)* *Combined*
Sex					
Decrease wall distance	0.016 (0.001-0.045)	t(490)= 2.82 p = .005	0.001 (< 0.019)	t(287)= -0.39 p = .70	0.005 (0.000-0.020)
Decrease token collection	0.043 (0.015-0.085)	t(490)= 4.72 p < .001	0.011 (< 0.048)	t(287)= 1.81 p = .071	0.029 (0.010-0.057)
Decrease speed when on grid	0 (< 0.012)	t(490)= -0.49 p = .63	0.004 (< 0.031)	t(287)= -1.04 p = .30	0.001 (< 0.011)
Average token collection	0.094 (0.050-0.148)	t(490)= 7.13 p < .001	0.162 (0.091-0.245)	t(287)= 7.46 p < .001	0.117 (0.078-0.162)
Average speed when on grid	0.07 (0.032-0.119)	t(490)= 6.07 p < .001	0.109 (0.050-0.185)	t(287)= 5.94 p < .001	0.083 (0.050-0.124)
Minimum distance from threat	0.068 (0.031-0.116)	t(490)= -5.96 p < .001	0.137 (0.071-0.217)	t(287)= -6.76 p < .001	0.09 (0.055-0.132)
Tokens retained	0.162 (0.106-0.225)	t(490)= 9.74 p < .001	0.184 (0.109-0.269)	t(287)= 8.03 p < .001	0.17 (0.124-0.220)
CADS					
Decrease token collection	0.006 (< 0.028)	t(459)= 1.65 p = .100	0.009 (< 0.044)	t(272)= -1.56 p = .12	0 (< 0.007)
Average token collection	0.021 (0.003-0.054)	t(459)= 3.13 p = .002	0.072 (0.024-0.141)	t(272)= 4.60 p < .001	0.037 (0.015-0.068)
Average speed when on grid	0.019 (0.002-0.051)	t(459)= 2.98 p = .003	0.053 (0.013-0.115)	t(272)= 3.91 p < .001	0.03 (0.011-0.059)
Tokens retained	0.036 (0.010-0.076)	t(459)= 4.12 p < .001	0.044 (0.009-0.103)	t(272)= 3.54 p < .001	0.039 (0.016-0.071)
IQ					
Decrease token collection	0.018 (0.002-0.049)	t(488)= 3.03 p = .003	0.057 (0.015-0.121)	t(268)= 4.02 p < .001	0.03 (0.010-0.058)
Tokens retained	0.034 (0.010-0.072)	t(488)= 4.15 p < .001	0.045 (0.009-0.105)	t(268)= 3.57 p < .001	0.038 (0.016-0.069)
BIS cognitive complexity					
Decrease token collection	0.024 (0.004-0.058)	t(460)= -3.34 p < .001	0.029 (0.003-0.080)	t(274)= -2.88 p = .004	0.026 (0.008-0.053)
Tokens retained	0.028 (0.006-0.065)	t(460)= -3.65 p < .001	0.015 (0.000-0.055)	t(274)= -2.01 p = .046	0.022 (0.006-0.048)