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

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

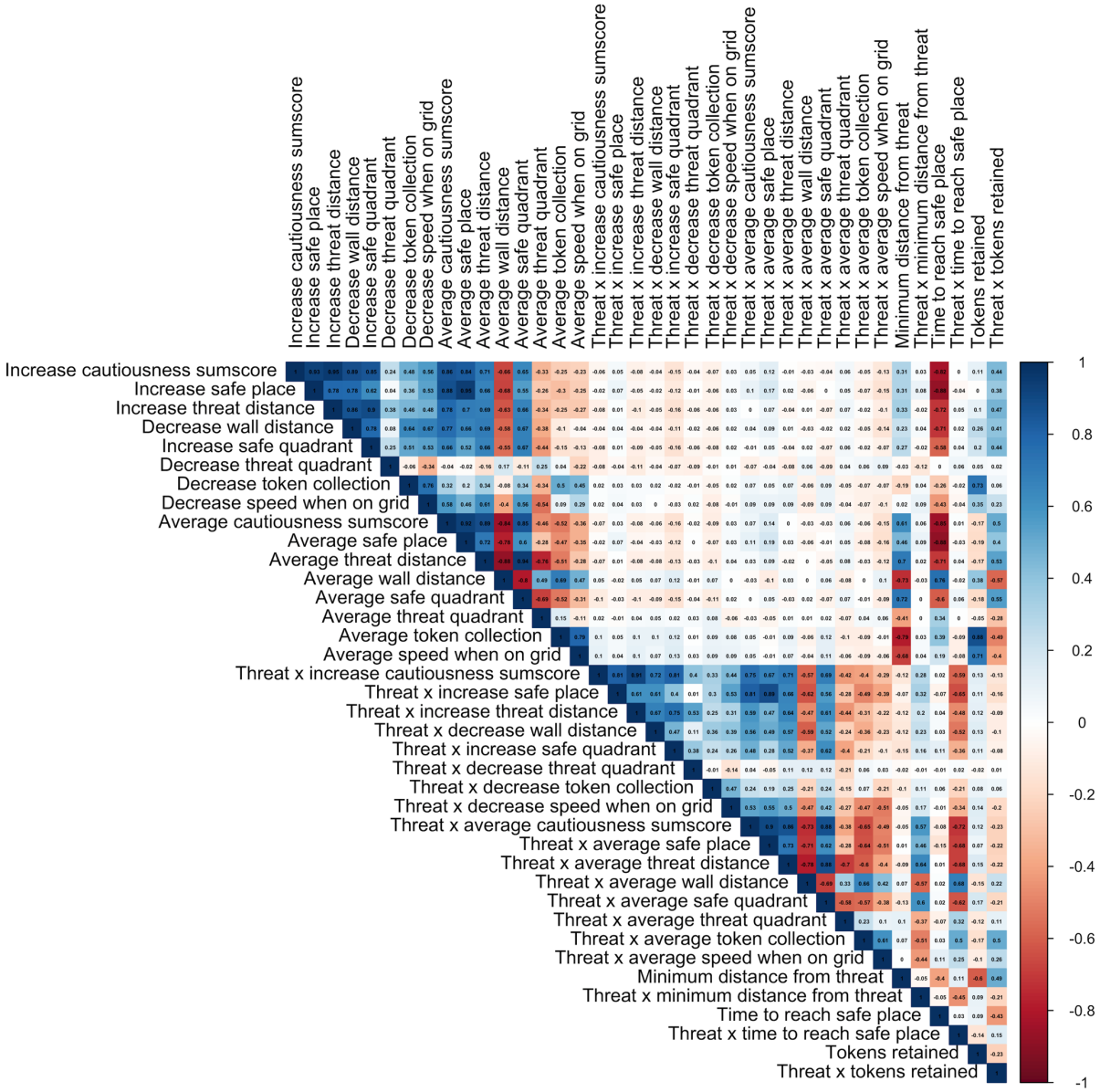
Supplementary Table 4. Extract from Barratt Impulsivity Scale extract, including the full Cognitive Complexity subscale, which is boxed in bold. The rest of the BIS is omitted for convenience.

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.

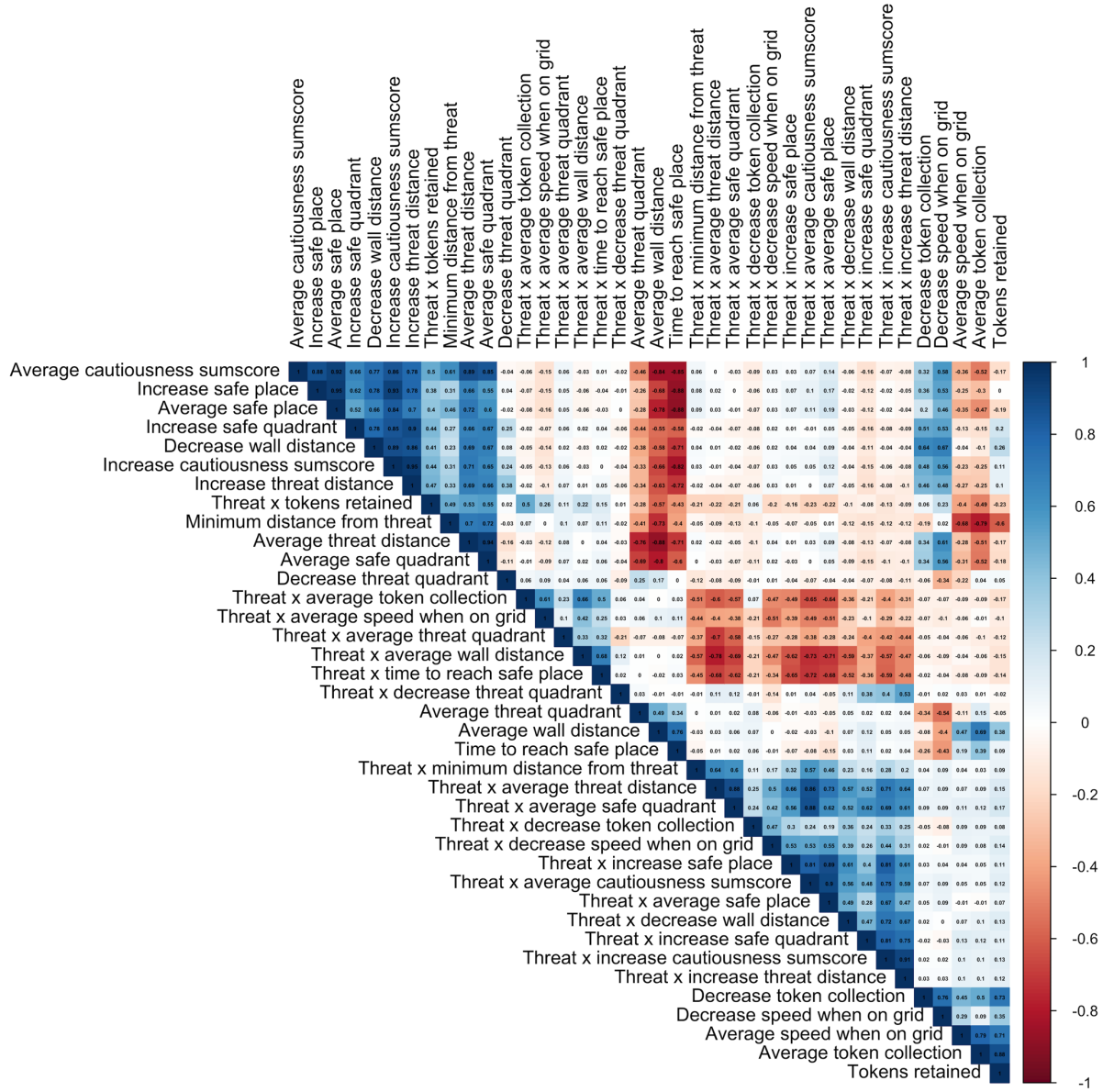
	Rarely	**Occasionally**	**Often**	**Always**
1. I plan tasks carefully.				
2. I do things without thinking.				
3. I make-up my mind quickly.				
4. I am happy go lucky.				
5. I don't "pay attention."				
6. I have "racing thoughts."				
7. I plan trips well ahead of time.				
8. I am self-controlled.				
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.

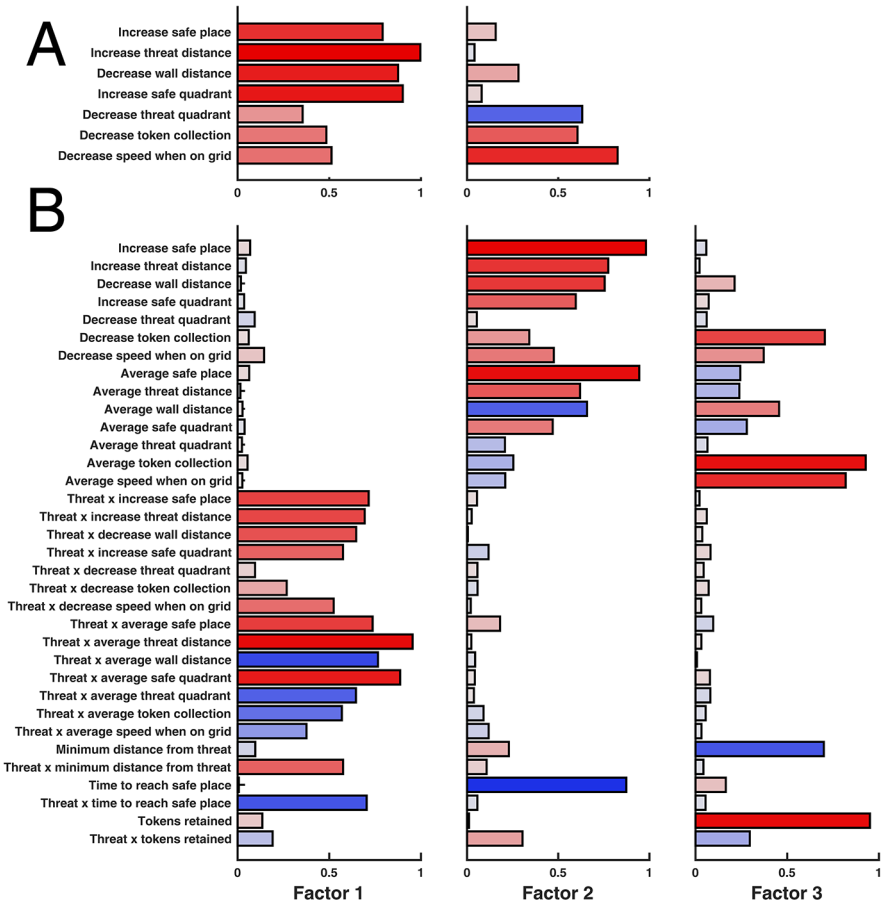
Supplementary Table 5: Consistency (Cronbach's alpha) at BSL, and test-retest reliability from baseline to FU-1 (over 11-32 months). Cronbach's alpha is for linear adaptation in the 7 time-dependent measures at baseline. Factor scores for discovery and confirmation samples are based on factor analysis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

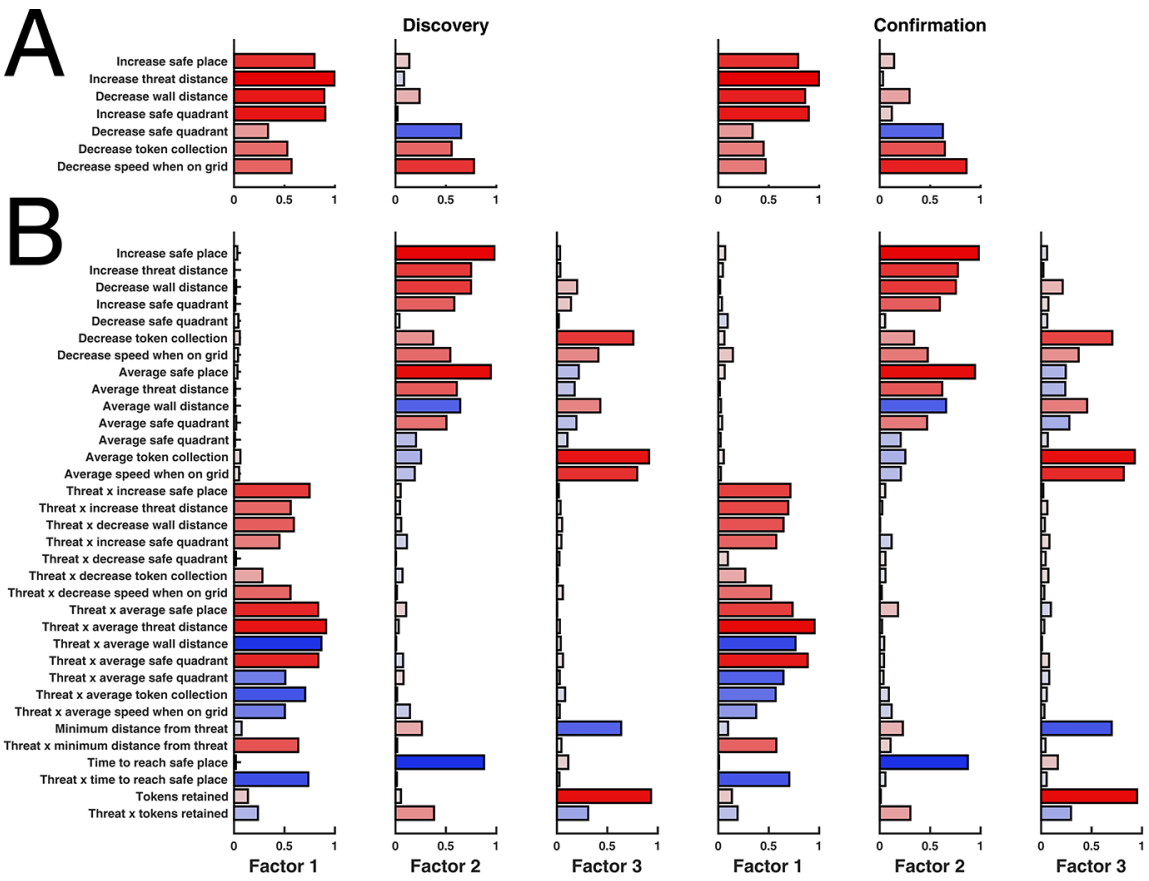
Tokens retained	0.696	0.677	0.689
Threat × tokens retained	0.17	0.172	0.173
Factor 1 7-msr FA	0.541	0.531	0.528
Factor 2 7-msr FA	0.640	0.619	0.639
Factor 1 <Sensitivity to threat probability> 34-msr FA	0.090	0.098	0.096
Factor 2 <Sensitivity to intra-epoch time> 34-msr FA	0.522	0.535	0.531
Factor 3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

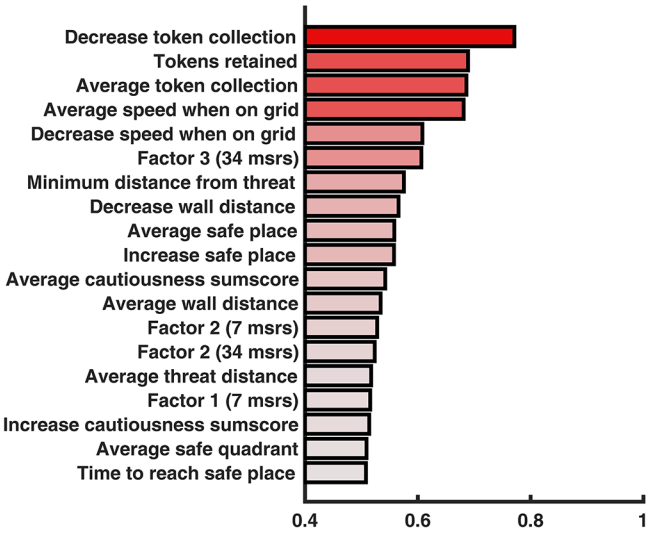
Test-retest reliability between BSL and FU-1 (e.g. over 1-3 years) was larger than $r_{tt} = .40$ for most measures unrelated to the predator differences, and was high as $r_{tt} = .70$ for some measures related to performance (see Supplementary Table 5, see Supplementary Figure 5 for measures with $r_{tt} > .50$ 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.

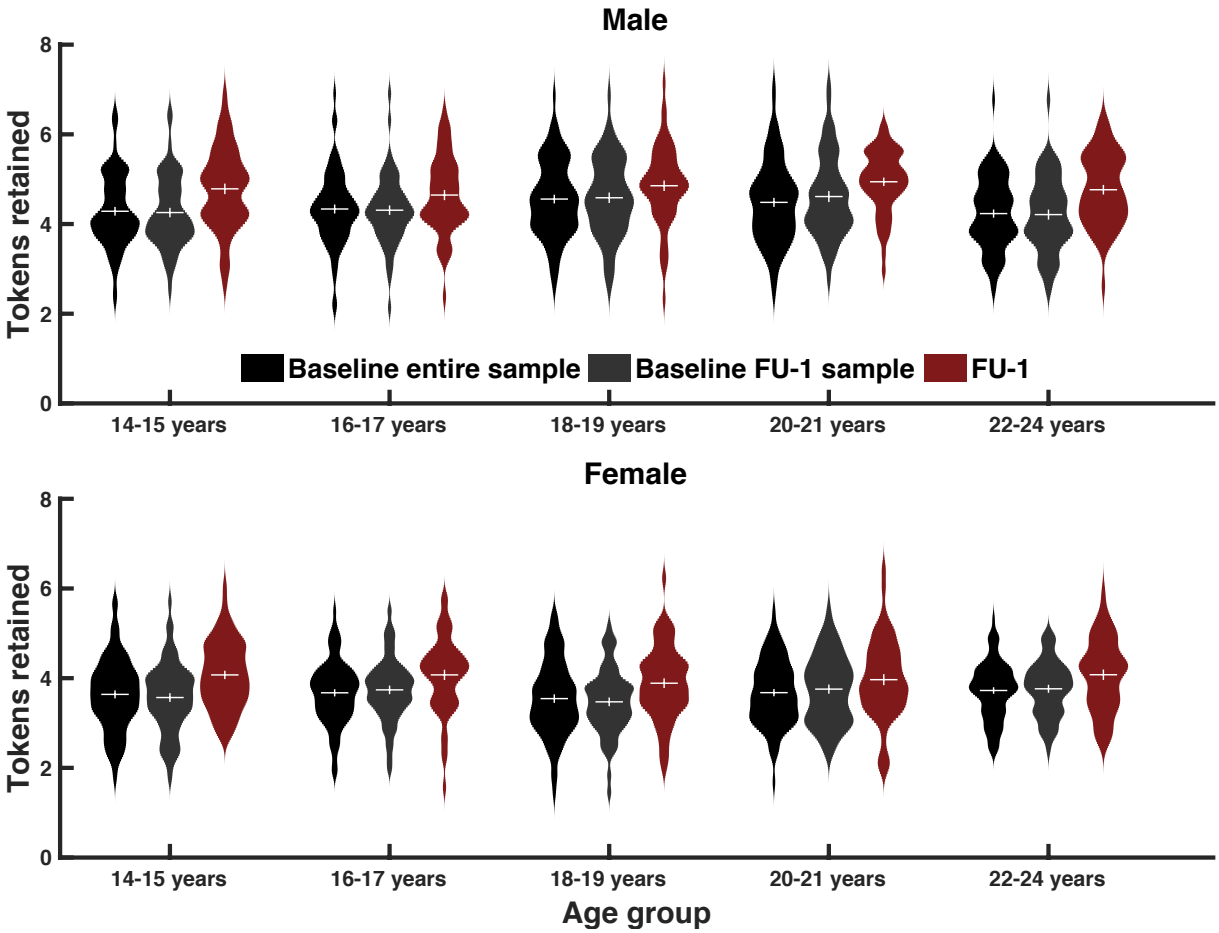
Finally, at the request of a reviewer, we computed, for each task variable, an age × (repetition + time interval) linear mixed effects model across the combined sample, and contrasted this with an age × repetition model. There were no significant age × time interval interactions at our alpha threshold 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 (95% CI)	t(df) p	Cohen's d (95% CI)	t(df) p
Increase cautiousness sumscore	1.35 (0.80-1.90)	t(62) = 5.38 p < .001	0.31 (-0.22-0.84)	t(54) = 1.16 p = .25
Increase safe place	1.34 (0.79-1.89)	t(62) = 5.33 p < .001	0.23 (-0.30-0.76)	t(54) = 0.87 p = .39
Increase threat distance	1.07 (0.53-1.59)	t(62) = 4.23 p < .001	0.28 (-0.25-0.81)	t(54) = 1.03 p = .31
Decrease wall distance	1.4 (0.84-1.95)	t(62) = 5.55 p < .001	-0.06 (-0.59-0.47)	t(54) = -0.22 p = .82
Increase safe quadrant	0.92 (0.40-1.44)	t(62) = 3.66 p < .001	0.51 (-0.03-1.05)	t(54) = 1.90 p = .062
Decrease threat quadrant	-0.2 (-0.69-0.30)	t(62) = -0.78 p = .44	0.18 (-0.35-0.71)	t(54) = 0.66 p = .51
Decrease token collection	0.95 (0.43-1.47)	t(62) = 3.77 p < .001	0.29 (-0.24-0.82)	t(54) = 1.08 p = .28
Decrease speed when on grid	1.14 (0.61-1.67)	t(62) = 4.54 p < .001	-0.11 (-0.63-0.42)	t(54) = -0.39 p = .70
Average cautiousness sumscore	1.32 (0.77-1.86)	t(62) = 5.25 p < .001	-0.28 (-0.81-0.25)	t(54) = -1.03 p = .31
Average safe place	1.37 (0.81-1.91)	t(62) = 5.42 p < .001	-0.07 (-0.60-0.46)	t(54) = -0.27 p = .79
Average threat distance	1.24 (0.69-1.78)	t(62) = 4.92 p < .001	-0.48 (-1.02-0.06)	t(54) = -1.79 p = .079
Average wall distance	-1.09 (-1.62--0.56)	t(62) = -4.33 p < .001	0.17 (-0.36-0.70)	t(54) = 0.64 p = .52
Average safe quadrant	0.83 (0.31-1.34)	t(62) = 3.29 p = .002	-0.59 (-1.13--0.05)	t(54) = -2.19 p = .033
Average threat quadrant	-0.72 (-1.23--0.21)	t(62) = -2.86 p = .006	0.22 (-0.31-0.75)	t(54) = 0.81 p = .42
Average token collection	0.04 (-0.45-0.54)	t(62) = 0.18 p = .86	0.69 (0.14-1.23)	t(54) = 2.54 p = .014
Average speed when on grid	-0.23 (-0.73-0.26)	t(62) = -0.92 p = .36	0.41 (-0.13-0.94)	t(54) = 1.51 p = .14
Threat x increase cautiousness sumscore	0.07 (-0.43-0.56)	t(62) = 0.27 p = .79	0.09 (-0.44-0.61)	t(54) = 0.32 p = .75
Threat x increase safe place	0.11 (-0.38-0.61)	t(62) = 0.45 p = .65	0.12 (-0.41-0.64)	t(54) = 0.43 p = .67
Threat x increase threat distance	-0.1 (-0.60-0.39)	t(62) = -0.42 p = .68	0.21 (-0.32-0.74)	t(54) = 0.76 p = .45
Threat x decrease wall distance	-0.11 (-0.60-0.39)	t(62) = -0.44 p = .67	0.37 (-0.16-0.91)	t(54) = 1.39 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 (95%-CI)	R2 (95%-CI)	R2 (95%-CI)
Decrease wall distance	0.000 (0.000-0.007)	0.001 (0.000-0.009)	0.002 (0.000-0.014)
Decrease token collection	0.004 (0.000-0.019)	0.001 (0.000-0.010)	0.012 (0.002-0.033)
Decrease speed when on grid	0.001 (0.000-0.012)	0.000 (0.000-0.006)	0.001 (0.000-0.009)
Average token collection	0.017 (0.004-0.040)	0.005 (0.000-0.020)	0.013 (0.002-0.034)
Average speed when on grid	0.011 (0.001-0.030)	0.004 (0.000-0.019)	0.003 (0.000-0.015)
Minimum distance from threat	0.014 (0.002-0.035)	0.006 (0.000-0.022)	0.003 (0.000-0.015)
Tokens retained	0.013 (0.002-0.034)	0.002 (0.000-0.013)	0.020 (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)