Appendix S8: Clustering analyses

We performed clustering analyses using the MClust package in R (Scrucca, Fop, Murphy, & Raftery, 2017).

Fig. S15 BIC plot for model comparison among a large variety of clustering solutions to describe the joint distribution of baseline and long follow-up model-fit (integrated likelihood) measures. The spherical, equal-volume Gaussian model resulting in the four-cluster solution discussed in the main text is the grey triangle ‘EII’ at 4 components. Key: "EII" : spherical, equal volume; "VII" : spherical, unequal volume; "EEI" : diagonal, equal volume and shape; "VEI": diagonal, varying volume, equal shape ; "EVI" : diagonal, equal volume, varying shape; "VVI" : diagonal, varying volume and shape; "EEE" : ellipsoidal, equal volume, shape, and orientation; "EEV" : ellipsoidal, equal volume and equal shape "VEV": ellipsoidal, equal shape "VVV": ellipsoidal, varying volume, shape, and orientation.
Fig S16. Simplified plots showing Mood, IQ and parameter values for the four clusters identified. Y axes are transformed / rescaled to facilitate visual comparisons. There are no significant differences in Mood (A.). The rest of the plots (B.-H.) show very similar values for the two smaller clusters, 1 and 2 but the blue cluster, 4, who had the best fits in both baseline and long follow-up, consistently has parameters more advantageous to better performance, and higher IQ, whereas the green cluster, 3, shows the exact opposite pattern. See table below for formal statistical comparisons.
The table below shows post-hoc contrast for differences between means of parameters, Mood and IQ of the four different clusters.

<table>
<thead>
<tr>
<th>Mood (MFQ to power 1/3)</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>z value</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>2 - 1 == 0</td>
<td>-1.717</td>
</tr>
<tr>
<td>3 - 1 == 0</td>
<td>-1.421</td>
</tr>
<tr>
<td>4 - 1 == 0</td>
<td>0.483</td>
</tr>
<tr>
<td>3 - 2 == 0</td>
<td>0.480</td>
</tr>
<tr>
<td>4 - 2 == 0</td>
<td>2.492</td>
</tr>
<tr>
<td>4 - 3 == 0</td>
<td>2.269</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pavlovian Bias (log transf.)</th>
<th>Return Sensitivity (log transf.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>z value</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>2 - 1 == 0</td>
<td>-0.968</td>
</tr>
<tr>
<td>3 - 1 == 0</td>
<td>-0.762</td>
</tr>
<tr>
<td>4 - 1 == 0</td>
<td>2.682</td>
</tr>
<tr>
<td>3 - 2 == 0</td>
<td>0.313</td>
</tr>
<tr>
<td>4 - 2 == 0</td>
<td>3.970</td>
</tr>
<tr>
<td>4 - 3 == 0</td>
<td>4.119</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>appet. Lrn. rate (log transf.)</th>
<th>Aver. Lrn. rate (raised to 1/4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>z value</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>2 - 1 == 0</td>
<td>0.537</td>
</tr>
<tr>
<td>3 - 1 == 0</td>
<td>0.19995</td>
</tr>
<tr>
<td>4 - 1 == 0</td>
<td>-3.270</td>
</tr>
<tr>
<td>3 - 2 == 0</td>
<td>3.638</td>
</tr>
<tr>
<td>4 - 2 == 0</td>
<td>-2.973</td>
</tr>
<tr>
<td>4 - 3 == 0</td>
<td>4.191</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Irreducible noise (log transf.)</th>
<th>Go bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>z value</td>
<td>Pr(&gt;</td>
</tr>
<tr>
<td>2 - 1 == 0</td>
<td>-1.343</td>
</tr>
<tr>
<td>3 - 1 == 0</td>
<td>-5.276</td>
</tr>
<tr>
<td>4 - 1 == 0</td>
<td>2.989</td>
</tr>
<tr>
<td>3 - 2 == 0</td>
<td>-4.060</td>
</tr>
<tr>
<td>4 - 2 == 0</td>
<td>4.729</td>
</tr>
<tr>
<td>4 - 3 == 0</td>
<td>9.858</td>
</tr>
</tbody>
</table>

Table S17. Tukey contrasts for multiple comparisons of means of the four clusters. Note how group 4, that fitted poorly on at both baseline and long follow-up stands out as different to the others and starkly contrasts with group 3, which fitted best on both occasions. Tukey tests were done separately for each measure, following linear mixed effects analyses. All entire analyses are statistically significant, at least p < 0.0001, except for Mood which is non-significant. A rigorous criterion would be to threshold the values above for 0.05/8 = 0.00625, and 0.01/8=0.00125, which means that in the table ‘*’ = trend, ‘**’ = stat. sig. at 0.05, ‘***’ = stat. sig. at least 0.01. Note that the significant results regarding return sensitivity and irreducible noise look strong but are not independent from the measure that was used to cluster participants, and hence offer little new information.
If a clustering solution is forced to utilize two clusters only, corresponding to the two clear peaks of the distribution, two equipopulous clusters form, the first, C1, approximately corresponding to the union of the green and red (2 and 3) clusters above, and C2 corresponding to the union of the black and blue (1 and 4) clusters. The parameter and likelihood comparison between these two larger clusters is shown in Table S18.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Long Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean C1</td>
<td>mean C2</td>
</tr>
<tr>
<td></td>
<td>6.64</td>
<td>9.26</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>0.46</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>0.087</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>-74.2 (59.7%)</td>
<td>-59.5 (66.1%)</td>
</tr>
<tr>
<td></td>
<td>Pcorr.</td>
<td>Pcorr.</td>
</tr>
<tr>
<td></td>
<td>&lt;1e-10</td>
<td>&lt;1e-10</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>9.1e-6</td>
<td>9.1e-6</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>int.likelihood (predict. per trial)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-74.2 (59.7%)</td>
<td>-59.5 (66.1%)</td>
</tr>
<tr>
<td></td>
<td>Pcorr.</td>
<td>Pcorr.</td>
</tr>
<tr>
<td></td>
<td>&lt;1e-10</td>
<td>&lt;1e-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S18. Means of the parameters if participants are classified two clusters, C1 and C2, of the joint distribution over integrated likelihood. P values regard Wilcoxon comparisons between clusters at the same time point, corrected for 14 comparisons. Note that the model accounts significantly better for the behaviour of the second cluster at both timepoints (better int. likelihood / predictability per trial).