Supporting Text

Methods

Bayesian chunk learner (BCL)

Prior distributions An inventory I was defined as the set of chunks, and for each chunk the shapes it influenced (referred to as links). The prior probability of an inventory (used for computing the posterior over inventories in Eq. 7, and also used in Eq. 8) depended on the number of chunks H and the total number of links L:

$$\mathbb{P}(I) = \mathbb{P}(H)\,\mathbb{P}(L|H) \tag{S1}$$

The prior distribution of the parameters of an inventory θ_I (used in Eqs. 6 and 8) included appearance probability parameters $W = \left\{w_{ij}, w_{x_i}, w_{y_j}\right\}$, and spatial position parameters $C = \left\{c_{ij}, c_{x_i}, c_{y_j}, \sigma_{ij}, \sigma_{x_i}, \sigma_{y_j}\right\}$:

$$\mathbb{P}(\theta_I|I) = \mathbb{P}(W|I)\,\mathbb{P}(C|I) \tag{S2}$$

The prior distribution of appearance probability parameters W had a simple factorized form

$$\mathbb{P}(W|I) = \prod_i \mathbb{P}(w_{x_i}) \prod_j \mathbb{P}\big(w_{y_j}\big) \prod_{j,i\in \mathrm{par}_I(j)} \mathbb{P}(w_{ij}) \tag{S3}$$

where $\mathrm{par}_I(j)$ is the set of chunks that influence (have links to) shape j according to inventory I.

The prior distribution of spatial position parameters had a similar factorized form:

$$\mathbb{P}(C|I) = \prod_i \mathbb{P}(c_{x_i})\,\mathbb{P}(\sigma_{x_i}) \prod_j \mathbb{P}\big(c_{y_j}\big)\,\mathbb{P}\big(\sigma_{y_j}\big) \prod_{j,i\in \mathrm{par}_I(j)} \mathbb{P}(c_{ij})\,\mathbb{P}(\sigma_{ij}) \tag{S4}$$

Table 2 shows the specific distributions used in Eqs. S1-S4.

Sampling the posterior Computing the predictive probability of a test scene $\mathbb{P}(T)$ (Eq. 8) required integration over the joint posterior distribution of inventories and parameters.* This integral was analytically intractable and was thus approximated by a Monte Carlo integral by summing over samples from the posterior distribution. In order to ensure fair sampling, a reversible-jump Markov-chain Monte Carlo (MCMC) sampler (*1*) was constructed, which was suitable for sampling sub-spaces of different dimensionality, combined with an extended ensemble approach (*2*), in which Markov-chains were run in parallel at different 'temperatures' for better mixing. The integrals were approximated by sums over $500,000$ samples collected after having discarded the first $500,000$ samples to ensure that the Markov-chain had already attained its stationary distribution.

* Equation 8 formally assumes that participants do not select a single inventory, but rather keep track of the posterior probabilities of *all* possible inventories (Eq. 7) and integrate over this distribution to compute the familiarity (predictive probability) of a scene (model *averaging*). Another, computationally less extensive alternative would be to concentrate on a *single* inventory, the one with the maximum a posteriori (MAP) probability $\hat{I}_{\mathrm{MAP}} = \mathrm{argmax}_I\, \mathbb{P}(I|\mathcal{D})$, or the one with maximal marginal likelihood $\hat{I}_{\mathrm{ML}} = \mathrm{argmax}_I\, \mathbb{P}(\mathcal{D}|I)$ (model *selection*). We applied the model averaging variant because that gives strictly optimal performance. However, this issue had little practical relevance in our simulations because the posterior over inventories was overwhelmingly dominated by the marginal likelihood of a single inventory, in which case these approaches yield identical results (see also Supporting Text, and Figs. 7-9).

In each iteration of the MCMC sampler the likelihood of an inventory needed to be computed that required computing a sum over x and an integral over u for each scene (Eqs. 4-5). The sum over x was computed by extensively sweeping through all possible 2^H configurations of x, where H is the number of chunks in the inventory. The integral over u was calculated analytically, for each configuration of x, by making use of the fact that the joint distribution of u and v was Gaussian conditioned on x and y due to the model assuming a products of Gaussians form for the process generating spatial positions (Eqs. 2-3). This ensured that the *exact* value of the likelihood could be calculated.

The MATLAB code implementing the sampler is available from the authors upon request.

Associative learner (AL)

Prior distributions The prior distribution of model parameters θ (used for computing the posterior over parameters according to Eq. 7, with I being substituted by θ) was the product of the priors of individual parameters, including appearance and co-appearance parameters w_{y_j} and w_{jk}, absolute and relative spatial position parameters c_{y_j} and c_{jk}, and the corresponding spatial variance parameters σ_{y_j} and σ_{jk}:

$$\mathbb{P}(\theta) = \prod_j \mathbb{P}\big(w_{y_j}\big)\,\mathbb{P}\big(c_{y_j}\big)\,\mathbb{P}\big(\sigma_{y_j}\big) \prod_{j,k\geq j} \mathbb{P}(w_{jk})\,\mathbb{P}(c_{jk})\,\mathbb{P}(\sigma_{jk}) \qquad \text{(S5)}$$

Table 3 shows the specific distributions used in Eq. S5.

Sampling the posterior Just as the BCL, the AL also required sampling of the posterior (of model parameters θ). Since the number of parameters did not vary, a simpler Metropolis-Hastings MCMC sampler (*3*) was constructed, combined with extended ensemble techniques (*2*) for better mixing. The integrals were approximated by sums over $400,000$ samples collected after having discarded the first $700,000$ samples to ensure that the Markov-chain had already attained its stationary distribution.

The MATLAB code implementing the sampler is available from the authors upon request.

Statistically naïve models (FL, JFL, CPL)

Learning in the frequency learner (FL) amounted to storing the occurrence frequencies of individual shapes during the familiarization phase (see also Table 4). Choice probability of a test scene was calculated by summing the stored frequencies of shapes appearing in the scene and replacing the log predictive probabilities with these sums in Equation 9.[†] This model captured the patterns of experimental data extremely poorly; even the qualitative trends were opposite to those seen in human performance (Fig. 4). Therefore, fitting

[†] Although these models do not define an actual probability distribution over scenes (hence their name 'statistically naïve'), one can think of the (co)occurrence frequencies learned by the FL (JFL) as playing the role of the canonical parameter of an exponential-family model (such as that used by the AL) that were set to be equal to the sufficient statistic of the model (which is incorrect, in general). In an exponential family model, the probability of a data point (in our case, a scene) is directly related through an exponential function to the negative 'energy' of the data point, $\mathbb{P}(x) \propto e^{-E(x)}$, and the negative energy has the form $-E\,(x) = \sum_i \theta_i f_i\,(x)$, where $\{\theta_i\}$ is the canonical parameter, and $f_i\,(x)$ are some functions of x (in our case, each one is a binary indicator of a specific shape(pair) being present in the scene). The sufficient statistic of such a model is $\left\{\sum_n f_i\,(x_n)\right\}$ where x_n is the nth training data point, which in our case is just the counts of (co)occurrences of shapes in the familiarization scenes. Choice probabilities are related to the log probabilities of the test scenes (Eq. 9), which in this case are their negative energies (up to an additive constant that cancels out) that turn out to be just the summed frequencies we used to determine choice probabilities. This gives a formal motivation for the way choice probabilities were computed for these models. A further motivation was direct comparability to previously published results (e.g. see (*4*,*5*)) obtained by computing choice probabilities in this way.

parameter β (Eq. 9) would have resulted in uninformative plots ($\beta = 0$ was the trivial solution to minimize squared error) and thus it was fixed at $\beta = 1$ across all experiments.

In the case of the joint frequency learner (JFL) and conditional probability learner (CPL) histograms were built based on the familiarization scenes. One entry of the histogram represented the co-occurrence frequency (JFL) or conditional occurrence frequency (CPL) of a pair of shapes in a given relative spatial position (see also Table 4). Choice probability of a test scene was calculated by summing the entries of the histogram corresponding to all shape pairs present in the scene and replacing the log predictive probabilities with these sums in Equation 9.[†] (Note that the CPL and the AL differ in one crucial aspect: in the CPL, the familiarity of a test scene is only determined by the shapes that are present in it, while in the AL, the absence of shapes is also considered.) In each model, parameter β (see Eq. 9) was fitted for each experiment individually, except for the baseline and frequency-balanced experiments, which were fitted together (so that there were at least two data points to be fitted in each case). A summary of model descriptions can be found in Table 4.

Results

Model fits to human performance

The four experiments of Fig. 2 were fitted with all five models in order to quantitatively evaluate their performance. The summary of the fits of the three statistically naïve models together with the AL and the BCL is shown in Fig. 4. In the baseline experiment, when all combos appeared with equal frequency, all models, except the FL, could reproduce the results obtained with human participants (Fig. 4A). In the frequency-balanced experiment, humans learned the right inventory even though in the familiarization scenes some combos (pairs of shapes) were presented more frequently than others and particular juxtapositions of two frequent pairs appeared just as often as one of the rare pairs. These results ruled out the FL and JFL, which rely on frequency counting, while correlation-based models (CPL and AL) and the BCL showed the same pattern of performance as human learners. In the experiments with combos composed of more than two shapes, naïve statistical learners (FL, JFL, CPL) could only predict identical performance on all test-types, while both the AL and the BCL could well fit human performance (Figs. 4C and D).

Figures 5 and 6 show quantitatively how well the BCL and the AL fitted human performance when both were fit with the same procedure. In each model, parameter β was fitted to data from Experiments 1-4 (red, orange, yellow, and green symbols), and the β value thus obtained was then used to *predict* experimental percent correct values (using again Eq. 9) in the correlation-balanced experiment (blue symbols). (Note that for Figs. 2-3, the AL, but not the BCL, was fit individually for each experiment, see also Methods of the Main Text.) The advantage of the BCL was substantial (compare Figs. 5 and 6) and it was also robust to the particular choice of hyperparameters (compare Figs. 5A and B).

Maximum a posteriori inventories

Learning in the BCL amounted to inferring the posterior probability distribution of inventories (Eq. 7, see also footnote on p.1). However, our numerical simulations showed that familiarization in the experiments was extensive enough so that for each experiment more than 99% of this distribution was concentrated within one inventory (for a formal definition of what constitutes an inventory see Supporting Methods), the maximum a posteriori (MAP) inventory. Since the prior used for inventories was fairly broad (Table 2), this showed that the posterior was dominated by the marginal likelihoods (Eq. 6).

Analyzing the MAP inventories also gave intuitions as to why the observed patterns emerged in the choice probabilities in the different experiments (Figs. 7-9). In the first four experiments (Fig. 2) the chunks learned by the model corresponded directly to the combos used for constructing the familiarization scenes.

In the frequency-balanced experiment (Figs. 2B and 7B) the chunks' appearance parameters, w_{x_i} (red numbers in Fig. 7B), reflected the differences in the frequencies of different combos. Therefore, once two chunks for the two frequent pairs had been learned it was not parsimonious to assume an extra combo for their combination if, as in the experiments, each of them also occurred in the absence of the other a number of times. This resulted in low predictive probabilities for mixture pairs composed of shapes belonging to different frequent pairs.

In the triplet and quadruplet experiments (Figs. 2C-D and 8A-B), similar to human performance, the recognition of embedded pairs dropped to close to chance level from the significantly above-chance recognition of true triplets and quadruplets and embedded triplets. This happened because, in the model, the overall predictive probability of an embedded combo was effectively a sum of two terms (the other terms in the sum in Eq. 4 were negligible). One term in the sum expressed the embedded combo as mere 'noise', the case when no chunks are present and thus each individual element appears with independent low probabilities (according to its spontaneous appearance parameter, w_{y_j}, blue numbers in Fig. 8). This 'noise-based' explanation became increasingly unlikely as the size of the embedded combo grew: shapes appeared independently, the appearance probability of any single one was low, and thus many of them appearing had an exponentially diminishing chance. The other term in the sum expressed the embedded combo as a chunk failing to appear fully. This chunk-based explanation was relatively unlikely since during training combos had always appeared in full, with all of their constituent shapes present. (This almost deterministic relationship was well captured by the high chunk-dependent appearance parameters of the shapes learned in the model, w_{ij}, black numbers in Fig. 8). Because the limiting factor in the chunk-based explanation was the probability of a shape *not* appearing, the predictive probability of an embedded combo did not depend on the size of the embedded combo (the number of shapes that *did* appear), only on the number of shapes of the true combo *not* included in the embedded combo. Thus, in effect, the noise-based account, under which any two combos had the same probability, dominated over the chunk-based account, which could distinguish between a true embedded and a mixture embedded combo, for smaller but not for larger embedded combos.

In the last, correlation-balanced experiment (Figs. 3 and 9), all shapes in the second group-of-4, that were usually shown separately, were assigned to a single common chunk accounting for the scenes in which they were all shown together. Shapes appearing individually were easily accounted for by their spontaneous appearance probabilities (note relatively higher w_{y_j} values, blue numbers, in Fig. 9), therefore it was not parsimonious to introduce extra chunks for each of them separately.

References

1. Green, P.J. Reversible jump MCMC computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).

2. Iba, Y. Extended ensemble Monte Carlo. *Int J Mod Phys C* **12**, 623–656 (2001).

3. MacKay, D.J.C. *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, UK, 2003).

4. Fiser, J. & Aslin, R.N. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol Sci* **12**, 499–504 (2001).

5. Fiser, J. & Aslin, R.N. Encoding multielement scenes: statistical learning of visual feature hierarchies. *J Exp Psychol Gen* **134**, 521–37 (2005).

6. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

$\mathrm{Sigmoid}(x)$	$= 1/\left(1+\exp(-x)\right)$	$x \in \mathbb{R}$
$\mathrm{Bernoulli}(x;q)$	$= q^x \left(1-q\right)^{1-x}$	$x \in \{0,1\}$
$\mathrm{Geometric}(x;q)$	$= q\left(1-q\right)^x$	$x \in \mathbb{N}$
$\mathrm{TruncGeom}(x;q,N)$	$= \frac{1}{1-(1-q)^{N+1}} q \left(1-q\right)^x$	$x \in \{0,\ldots N\}$
$\mathrm{Exponential}(x;\lambda)$	$= \lambda e^{-\lambda x}$	$x \in \mathbb{R}, x \geq 0$
$\mathrm{Laplace}(x;\mu,b)$	$= \frac{1}{2b} e^{-\lvert x-\mu \rvert / b}$	$x \in \mathbb{R}$
$\mathrm{Normal}(x;\mu,\Sigma)$	$= \frac{1}{(2\pi)^{N/2} \lvert\Sigma\rvert^{1/2}} e^{-\frac{1}{2}(x-\mu)^\intercal \, \Sigma^{-1} \, (x-\mu)}$	$x \in \mathbb{R}^N$
$\text{Dirac-delta}(x;\mu)$	$= \prod_{i=1}^{N} \delta(x_i - \mu_i)$	$x \in \mathbb{R}^N$

Table 1: Functions and distributions used in the text.

Parameter	Prior distribution		Hyperparameters		
			Set 1[*]	Set 2[†]	
H	$\text{Geometric}(H; q)$	q	0.1	0.2	
L	$\text{TruncGeom}(L; q, H \cdot N)$	q	0.1	0.2	[‡]
w_{x_i}	$\text{Laplace}(w_{x_i}; \mu, b)$	μ	0	2	
		b	$\sqrt{2}$	1	
w_{y_j}	$\text{Laplace}\big(w_{y_j}; \mu, b\big)$	μ	-8	-6	
		b	1	$\frac{1}{\sqrt{2}}$	
w_{ij}	$\text{Exponential}(w_{ij}; \lambda)$	λ	$\frac{1}{6}$	$\frac{1}{3}$	[§]
c_{x_i}	$\text{Dirac-delta}(c_{x_i}; \mu)$	μ	$[0, 0]$	$[0, 0]$	[¶]
c_{y_j}	$\text{Dirac-delta}\big(c_{y_j}; \mu\big)$	μ	$[0, 0]$	$[0, 0]$	[¶]
c_{ij}	$\text{Normal}(c_{ij}; \mu, \Sigma)$	μ	$[0, 0]$	$[0, 0]$	[¶]
		Σ	$2 \cdot \mathbf{I}$	$2\sqrt{2} \cdot \mathbf{I}$	[‖][**]
σ_{x_i}	$\text{Dirac-delta}(\sigma_{x_i}; \mu)$	μ	4	$4\sqrt{2}$	[**]
σ_{y_j}	$\text{Dirac-delta}\big(\sigma_{y_j}; \mu\big)$	μ	4	$4\sqrt{2}$	[**]
σ_{ij}	$\text{Dirac-delta}(\sigma_{ij}; \mu)$	μ	$\sqrt{2}$	2	[**]

Table 2: Parameter priors for the ideal learner. For definitions of probability distributions see Table 1.

Notes:

[*] Values used for simulations plotted in Figures 2-3, and 5A.

[†] Values used for simulations plotted in Figure 5B.

[‡] N is the number of shapes.

[§] Non-negativity of link weights was motivated by Ref. *6*.

[¶] The origin $[0, 0]$ was defined as the geometrical center of the grid in which the shapes were presented.

[‖] $\mathbf{I}$ is the (2×2) identity matrix.

[**] The unit of spatial coordinates was the length of the edge of a cell in the grid.

Parameter	**Prior distribution**	
w_{jk}	$\mathrm{Normal}(w_{jk}; 0, 16)$	
w_{y_j}	$\mathrm{Normal}\left(w_{y_j}; 0, 4\right)$	
c_{jk}	$\mathrm{Normal}(c_{jk}; [0, 0]\,, 4\mathbf{I})$	[*][†][‡]
c_{y_j}	$\mathrm{Normal}\left(c_{y_j}; [0, 0]\,, 4\mathbf{I}\right)$	[*][†][‡]
σ_{ij}	$\mathrm{Exponential}(\sigma_{ij}; 5)$	[‡]
σ_{y_j}	$\mathrm{Exponential}\left(\sigma_{y_j}; 1/\sqrt{2}\right)$	[‡]

Table 3: Parameter priors for the associative learner. For definitions of probability distributions see Table 1.

Notes:

[*] The origin $[0, 0]$ was defined as the geometrical center of the grid in which the shapes were presented.

[†] **I** is the (2×2) identity matrix.

[‡] The unit of spatial coordinates was the length of the edge of a cell in the grid.

		FAMILIARIZATION	TEST
FL	single-shape frequency counting	freq(⧓), freq(⬇), freq(⚑) …	Σ for all shapes present freq(shape)
JFL	shape-pair frequency counting	freq(⧓⬇), freq(⬇⚑), freq(⚑⧓) …	Σ for all shape-pairs present freq(shape_1, shape_2)
CPL	shape-pair conditional probabilities	freq(⧓⬇) / freq(⧓), freq(⧓⬇) / freq(⬇) …	Σ for all shape-pairs present freq(shape_1, shape_2) / freq(shape_1)
AL	probabilistic associative learner	Prob(pair-wise correlations | familiarization scenes)	Prob(test scene | pair-wise correlations)
BCL	Bayesian model comparison	Prob(inventory of independent chunks | familiarization scenes)	Prob(test scene | inventory of independent chunks)

Table 4: Summary of alternative models used to explain human behavior in visual chunk learning experiments. For each model the quantities extracted from familiarization scenes and those used to judge the familiarity of a scene in the test phase of the experiments are shown. The frequency learner (FL) stores single shape occurrence frequencies during familiarization and adds up these frequencies for each shape present in a test scene. The joint frequency learner (JFL) stores co-occurrence frequencies of all possible shape-pairs and adds up these frequencies for each shape-pair present in a test scene. The conditional probability learner (CPL) stores conditional probabilities (co-occurrence frequencies normalized by individual shape frequencies) of all possible (ordered) shape-pairs and adds up these for each shape-pair present in a test scene. The associative learner (AL) infers the strength of associations between all possible pairs of shapes from the familiarization scenes, and based on these computes the probability of a test scene. The Bayesian chunk learner (BCL) infers the probabilities of possible chunk inventories from the familiarization scenes, and based on these computes the probability of a test scene.

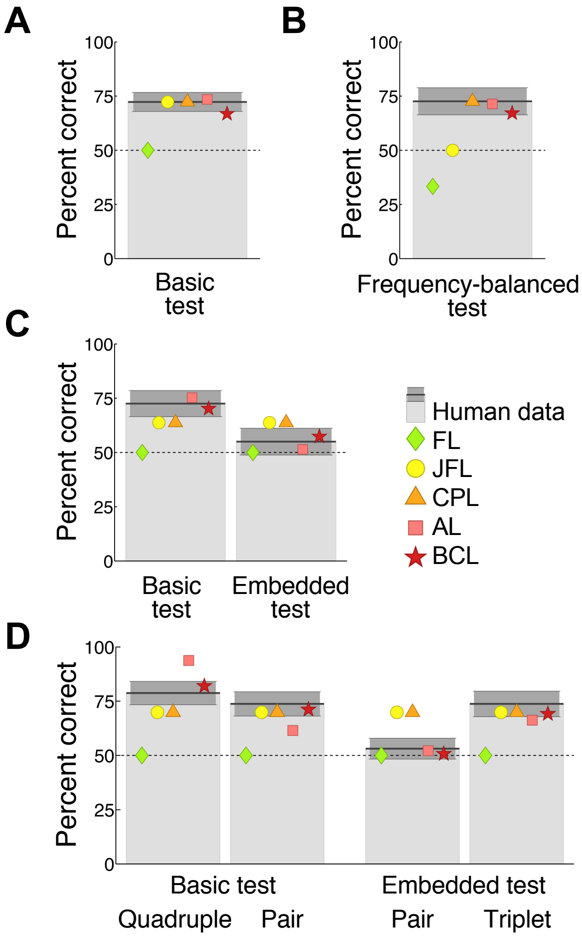

Figure 4: Summary of the performance of the five different models (colored symbols) compared to human performance (gray bars) for a series of experiments from Refs. (*4*) and (*5*) using increasingly complex inventories. Models shown are the frequency learner (FL), joint frequency learner (JFL), conditional probability learner (CPL), associative learner (AL), and the Bayesian chunk learner (BCL). Bars indicate average human performance (± 1 standard error of mean, s.e.m.). For a stringent comparison, parameters for the FL, JFL, CPL, and AL models were adjusted independently for each experiment to obtain best fits, whereas the BCL used a single parameter optimization across all experiments. **A,** Inventory containing six equal-frequency pairs. Only the FL failed to predict above-chance human performance on the basic test of true pairs vs. mixture pairs. **B,** Inventory containing six pairs of varying frequencies. The FL and JFL failed to predict above-chance human performance on the test of true rare pairs vs. frequency-balanced mixture pairs. **C,** Inventory containing four equal-frequency triplets. Human performance was above chance on the basic test of true triplets vs. mixture triplets and at chance on the test of embedded pairs vs. mixture pairs. All the statistically naïve models (FL, JFL, CPL) incorrectly predicted equivalent performance on the basic and embedded tests. **D,** Inventory containing two quadruples and two pairs, all with equal frequency. Human performance was above chance on the basic tests of true quadruples or pairs vs. mixture quadruples or pairs, and on the test of embedded triplets vs. mixture triplets, but it was at chance on the test of embedded pairs vs. mixture pairs. Again, the FL, JFL, and the CPL incorrectly predicted the same performance on all tests. Only the AL and the BCL captured the overall pattern of human performance in all these experiments.

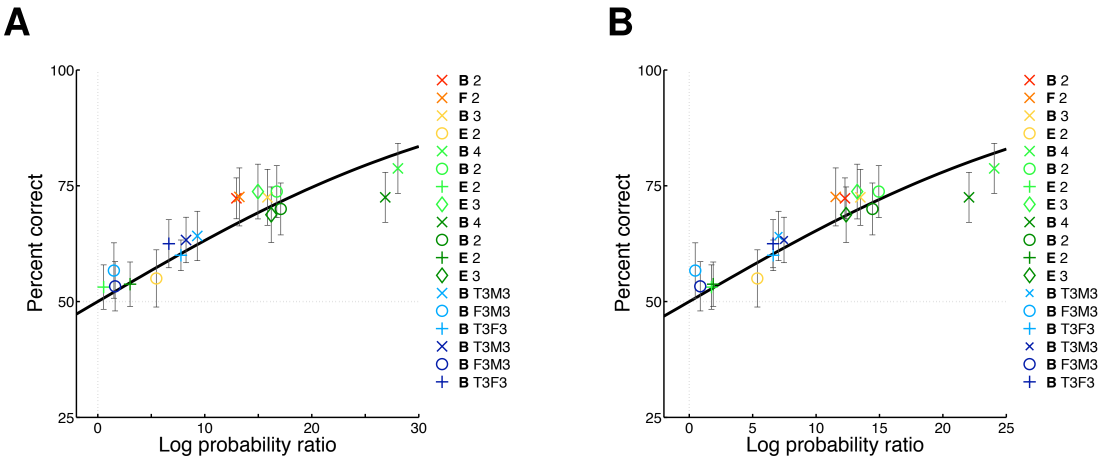

Figure 5: Summary fit of the Bayesian chunk learner (BCL) to human performance. **A,** Simulation results shown in the main text with hyperparameter set 1 (see Supporting Methods and Table 2). **B,** Results of control simulations with hyperparameter set 2 (see Supporting Methods and Table 2) to demonstrate robustness of model predictions to changes in hyperparameters. Each *colored symbol* shows the average percent correct value (±s.e.m.) from experimental data for one test type versus the log probability ratio used to determine simulated choice probabilities for the BCL (Eq. 9). *Colors* show experiment type: *red* – baseline experiment (Fig. 2A), *orange* – frequency-balanced experiment (Fig. 2B), *yellow* – triplet experiment (Fig. 2C), *green* – quadruplet experiment (Fig. 2D), *blue* – correlation-balanced experiment (Fig. 3). In the quadruplet and correlation-balanced experiments a group of participants were trained on the first half of the familiarization scenes and were tested after this limited training: *dark colors* – mid-familiarization test results, *light colors* – test results after completion of the familiarization phase (data plotted in Figs. 2D and 3B). *Letters* denote test type (Fig. 2, middle row): *B* – baseline test, *F* – frequency-balanced test, *E* – embedded test; in the correlation-balanced experiment: *T3M3* – true triplet versus mixture triplet test, *F3M3* – false triplet versus mixture triplet test, *T3F3* – true triplet versus false triplet test (see Fig. 3B). *Numbers* denote size of test combo: *2* – pair, *3* – triplet, *4* – quadruplet. *Thick black line* is the predicted choice probability of the BCL (Eq. 9) after fitting $\beta = 0.054$ (**A**) and $\beta = 0.063$ (**B**). The correlation (Pearson's correlation coefficient) between experimental data and model predictions is $r = 0.88$ ($p < 0.0002$) (**A**) and $r = 0.89$ ($p < 0.0001$) (**B**) for data used for fitting the model (red, orange, yellow, and green symbols), and $r = 0.92$ ($p < 0.006$) (**A**) and $r = 0.92$ ($p < 0.009$) (**B**) for predicted data (blue symbols). See also Methods of the Main Text and Supporting Text for details of the fitting procedure.

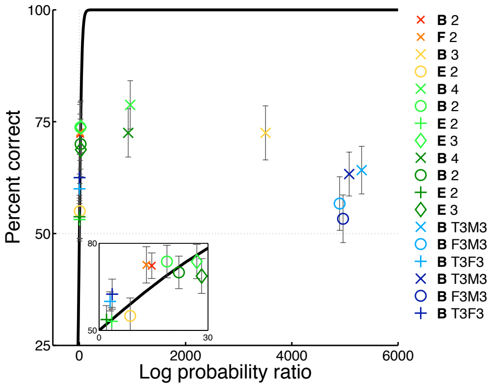

Figure 6: Summary fit of the associative learner (AL) to human performance. Each *colored symbol* shows the average percent correct value (±s.e.m.) from experimental data for one test type versus the log probability ratio used to determine simulated choice probabilities for the AL (Eq. 9). *Inset* shows magnified area of the graph. *Colors* show experiment type: *red* – baseline experiment (Fig. 2A), *orange* – frequency-balanced experiment (Fig. 2B), *yellow* – triplet experiment (Fig. 2C), *green* – quadruplet experiment (Fig. 2D), *blue* – correlation-balanced experiment (Fig. 3). In the quadruplet and correlation-balanced experiments participants were also tested after half of the familiarization scenes had been shown: *dark colors* – mid-familiarization test results, *light colors* – test results after completion of the familiarization phase (data plotted in Figs. 2D and 3B). *Letters* denote test type (Fig. 2, middle row): *B* – baseline test, *F* – frequency balanced test, *E* – embedded test; in the correlation-balanced experiment: *T3M3* – true triplet versus mixture triplet test, *F3M3* – false triplet versus mixture triplet test, *T3F3* – true triplet versus false triplet test (see Fig. 3B). *Numbers* denote size of test combo: *2* – pair, *3* – triplet, *4* – quadruplet. *Thick black line* is the predicted choice probability of the AL (Eq. 9) after fitting $\beta = 0.042$. The correlation (Pearson's correlation coefficient) between experimental data and model predictions is $r = 0.71$ ($p < 0.01$) for data used for fitting the model (red, orange, yellow, and green symbols), and $r = -0.23$ ($p > 0.65$) for predicted data (blue symbols). See also Methods of the Main Text and Supporting Text for details of the fitting procedure.

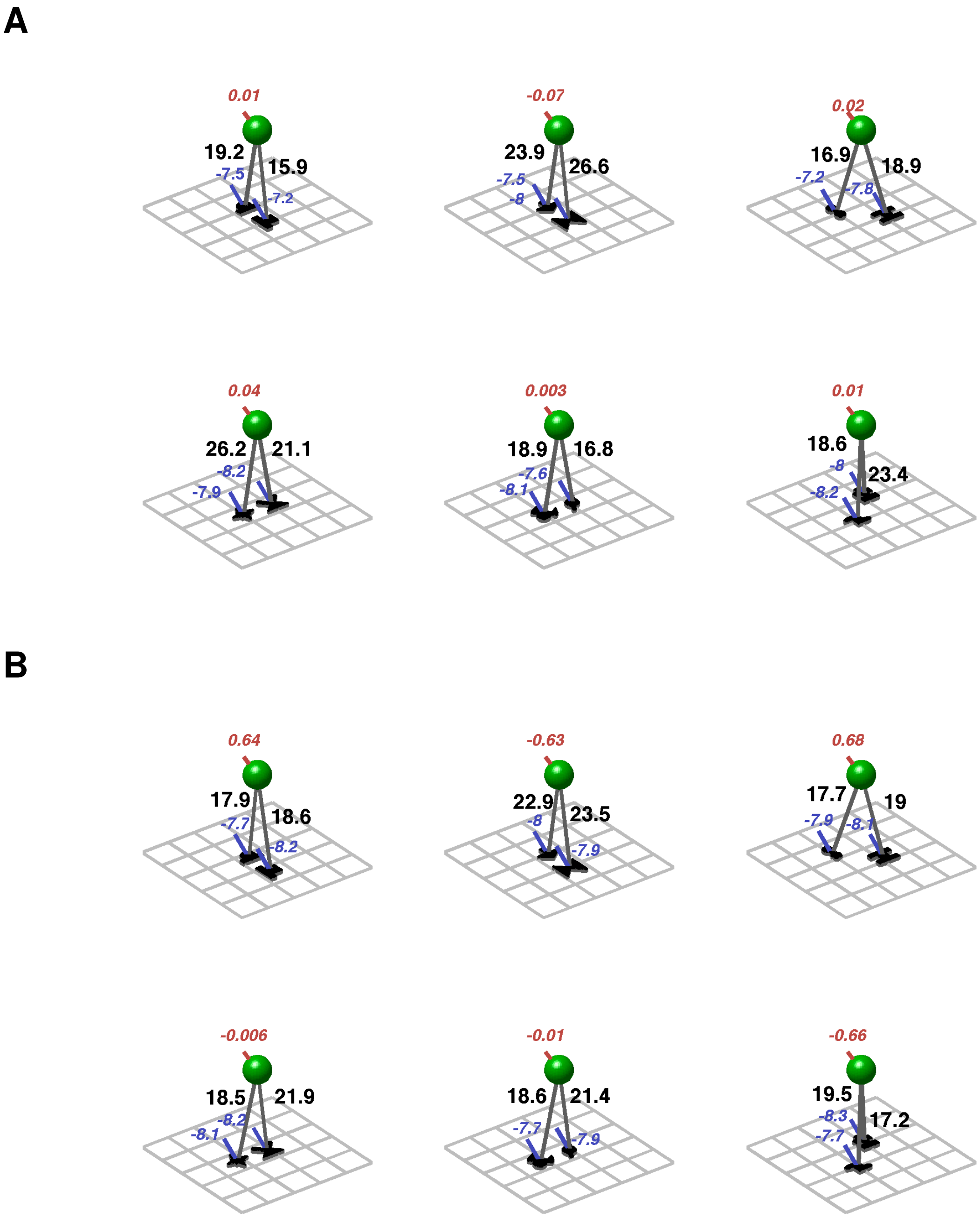

Figure 7: Maximum a posteriori (MAP) inventories, and posterior means of the parameters within the MAP inventories, learned by the BCL in the baseline (**A**, Fig. 2A) and frequency-balanced experiments (**B**, Fig. 2B). Each panel shows the parameters belonging to a chunk (*green sphere*) and to the *shapes* it influences. The tilted grid shows the grid used in the visual scenes for reference. The position of the green sphere (along the grid) shows the spatial position parameter $c_{x_i} = [0, 0]$ of the chunk (fixed without learning), and the *red number* shows its appearance parameter, w_{x_i}. The positions of the *shapes* in the grid show their relative spatial position parameters, c_{ij}, and the *blue numbers* and *black numbers* show their 'spontaneous' and chunk-dependent appearance parameters, w_{y_j} and w_{ij}, respectively.

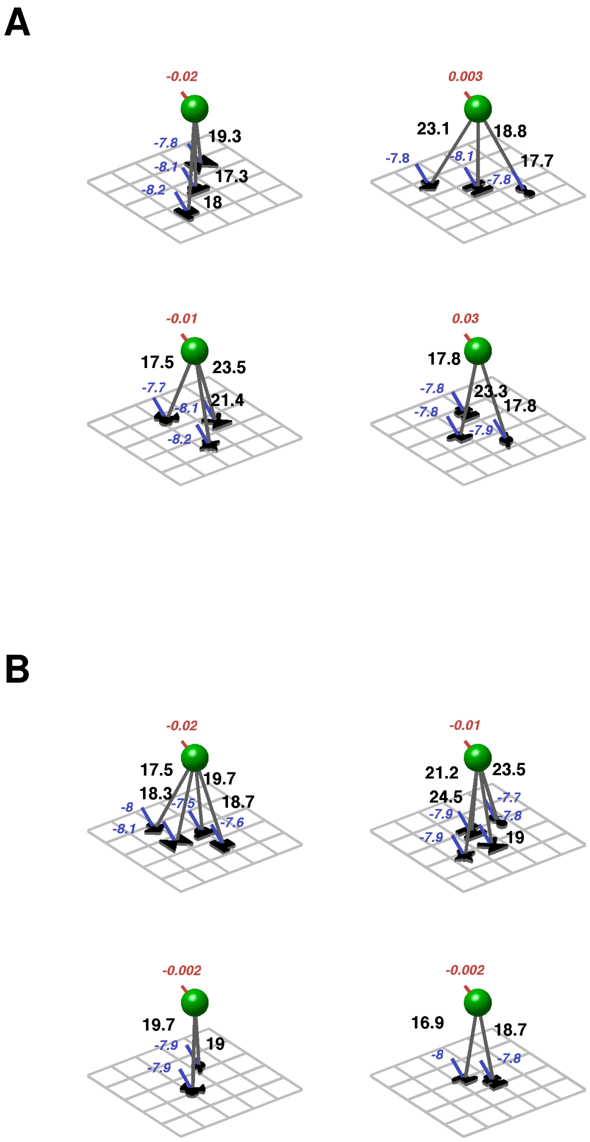

Figure 8: Maximum a posteriori (MAP) inventories, and posterior means of the parameters within the MAP inventories, learned by the BCL in the triplet (**A**, Fig. 2C) and quadruplet experiments (**B**, Fig. 2D). Each panel shows the parameters belonging to a chunk (*green sphere*) and to the *shapes* it influences. The tilted grid shows the grid used in the visual scenes for reference. The position of the green sphere (along the grid) shows the spatial position parameter $c_{x_i} = [0, 0]$ of the chunk (fixed without learning), and the *red number* shows its appearance parameter, w_{x_i}. The positions of the *shapes* in the grid show their relative spatial position parameters, c_{ij}, and the *blue numbers* and *black numbers* show their ‘spontaneous’ and chunk-dependent appearance parameters, w_{y_j} and w_{ij}, respectively.

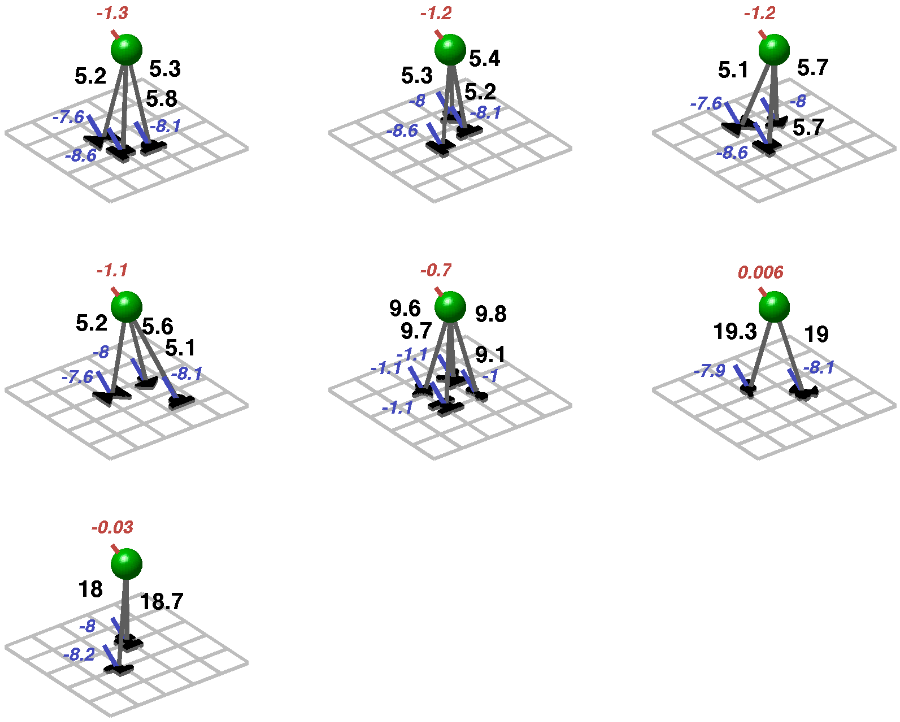

Figure 9: Maximum a posteriori (MAP) inventories, and posterior means of the parameters within the MAP inventories, learned by the BCL in the correlation-balanced experiment (Fig. 3). Each panel shows the parameters belonging to a chunk (*green sphere*) and to the *shapes* it influences. The tilted grid shows the grid used in the visual scenes for reference. The position of the green sphere (along the grid) shows the spatial position parameter $c_{x_i} = [0, 0]$ of the chunk (fixed without learning), and the *red number* shows its appearance parameter, w_{x_i}. The positions of the *shapes* in the grid show their relative spatial position parameters, c_{ij}, and the *blue numbers* and *black numbers* show their 'spontaneous' and chunk-dependent appearance parameters, w_{y_j} and w_{ij}, respectively.