Supplementary Information for Temporal and state abstractions for efficient learning, transfer and composition in humans

Liyu Xia, Anne G. E. Collins

Potential asymmetry in Block 7 of Experiment 1

We checked whether the performance of circle and square in Block 7 was asymmetrically affected due to the interleaving of odd and even blocks (Fig. 2B). Specifically, participants might start Block 7 by using HO_1 in odd blocks; thus the negative transfer in the first stage of Block 7 would be primarily due to more key presses from the square, not the circle.

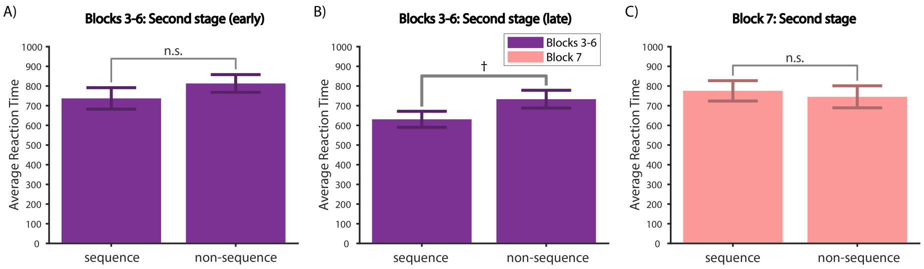
To test this possibility, we calculated average number of key presses in the first 5 trials for circle and square respectively in Block 7. However, we found no significant difference between the performance of circle and square in the first stage (paired t-test, $t(24) = 1.38, p = 0.18$); we also found no significant difference between the performance in the second stage following circle and square (paired t-test, $t(24) = 0.44, p = 0.66$).

Second stage reaction time and sequence learning effects

Sequence learning predicts that the reaction time of the “sequence” type to be faster than the “non-sequence” type. Therefore, we calculated the average reaction time (Supplementary Fig. S1, Supplementary Fig. S2) for both “sequence” and “non-sequence” error types in Experiment 1 and 2.

Experiment 1

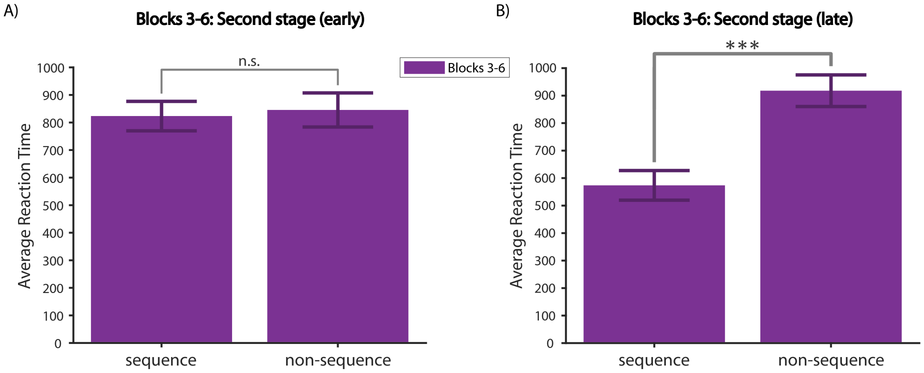
We broke down each block to 2 different time periods: early (trials 1-7 for each of the 4 branches in the second stage) and late (trials 8-15 for each of the 4 branches). Aggregating Blocks 3-6, we found a marginal effect of time period (2-way repeated measure ANOVA, $F(1, 21) = 3.0, p = 0.099$), which might be due to participants generally becoming faster as they progressed within a block. We also found a main effect of error type (2-way repeated measure ANOVA, $F(1, 21) = 4.5, p = 0.046$) on reaction time. Specifically, we found no significant difference ($t(23) = 1.3, p = 0.2$) between the reaction time of the “sequence”



Supplementary Figure S1: Experiment 1 reaction time. (A) Average reaction time for trials 1-7 for each of the 4 branches in the second stage for Blocks 3-6 for sequence (left) and non-sequence (right) error types. (B) Same as (A) for trials 8-15. (C) Average reaction time for sequence (left) and non-sequence (right) error types in the second stage of Block 7.

and "non-sequence" error types in the early time periods (Supplementary Fig. S1A). The "sequence" type was marginally faster (paired t-test, $t(22) = 1.9, p = 0.072$) than the "non-sequence" type in the late time period (Supplementary Fig. S1B). We also found no significant difference (paired t-test, $t(20) = 1.1, p = 0.3$) between the "sequence" and "non-sequence" types in the entire Block 7 (Supplementary Fig. S1C). These results suggest that the transfer effects we observed at the beginning of each block could not be due to pure sequence learning, which only start to take effect during learning saturation.

Experiment 2



Supplementary Figure S2: Experiment 2 reaction time. (A) Average reaction time for trials 1-4 for each of the 4 branches in the second stage for Blocks 3-6 for sequence (left) and non-sequence (right) error types. (B) Same as (A) for trials 5-8.

We also analyzed the reaction time (Supplementary Fig. S2) of the "sequence" and "non-sequence" error types in Blocks 5-6 in Experiment 2. As in Experiment 1, we broke down each block into 2 halved time periods: early (trials 1-4 for each of the 4 branches in the second stage) and late (trials 5-8 for each of the 4 branches). We found a main effect of time period and error type, and a significant interaction (2-way repeated measure ANOVA, time period: $F(1,16) = 8, p = 0.012$; error type: $F(1,16) = 16, p = 0.0009$; interaction: $F(1,16) = 15, p = 0.0013$). Specifically, there was no significant difference (Supplementary Fig. S2A) between the reaction time of the "sequence" and "non-sequence" types in the early time period (paired t-test, $t(21) = 0.61, p = 0.55$). However, the "sequence" type was significantly faster (Fig. S2B) than the "non-sequence" type in the late period (paired t-test, $t(17) = 4.8, p = 0.0002$). These results replicated the trend observed in the second stage of Experiment 1 (Supplementary Fig. S1): sequence learning might take effect during learning saturation, but not the beginning of blocks, where we typically expect to observe transfer effects.

Full description of the Option Model

The first stage of the Option Model is identical to the first stage of the Task-Set Model. The model tracks the probability P^1 of selecting each first stage task-set HO_i in different first stage contexts c^1_j, which encodes the current temporal (block) context (e.g. 8 contexts in the first stage of Experiment 1 due to 8 blocks). The model uses CRP to select HO: if contexts $\{c^1_{1:n}\}$ are clustered on $N^1 \leq n$ HOs, when the model encounters a new context c^1_{n+1}, the prior probability of selecting a new high-level option HO_{n+1} in this new context is set to:

$$P^1(HO_{n+1}|c^1_{n+1}) = \frac{\gamma^1}{Z^1}; \tag{1}$$

and the probability of reusing a previously created high-level option HO_i is set to:

$$P^1(HO_i|c^1_{n+1}) = \frac{N^1_i}{Z^1}, \tag{2}$$

where γ^1 is the clustering coefficient for the CRP, N^1_i is the number of first stage contexts clustered on HO_i, and $Z^1 = \gamma^1 + \sum_i N^1_i$ is the normalization constant. The new HO_{n+1} policy is initialized with uninformative Q-values $1/\#\{possible\,actions\} = \frac{1}{4}$. The model samples HO based on the conditional distribution over all HOs given the current temporal context. The model also tracks HO-specific policies via Q-learning. Once an HO is selected, a first stage policy is computed based on the HO's Q-values and the first stage stimulus F_i with softmax:

$$P(A^1_j|F_i, HO) = \frac{exp(\beta^1 * Q^1_{HO}(F_i, A^1_j))}{\sum_k exp(\beta^1 * Q^1_{HO}(F_i, A^1_k))}, \tag{3}$$

where β^1 is the inverse temperature. A first stage action A^1, ranging from A_1 to A_4, is then sampled from this softmax policy. After observing the outcome (moving on to the second stage or not), the model uses Bayes' Theorem to update P^1:

$$P^1(HO_k|c_j^1) = \frac{P(r|F_i, A^1, HO_k)P(HO_k|c_j^1)}{(\sum_l P(r|F_i, A^1, HO_l)P(HO_l|c_j^1))}, \tag{4}$$

where r is 1 if A^1 is correct and 0 otherwise, and $P(r|F_i, A^1, HO_l) = 1 - Q^1_{HO_l}(F_i, A^1)$ if $r = 0$, or $Q^1_{HO_l}(F_i, A^1)$ if $r = 1$. Then the Q-values of the HO with the highest posterior probability is updated:

$$Q^1_{HO}(F_i, A^1) = Q^1_{HO}(F_i, A^1) + \alpha^1 * (r - Q^1_{HO}(F_i, A^1)), \tag{5}$$

where α^1 is the learning rate. After each choice, the model decays the Q-values of each HO in the first stage based on f^1:

$$Q^1_{HO}(F_i, A^1_j) = (1 - f^1) * Q^1_{HO}(F_i, A^1_j) + f^1 * \frac{1}{4}. \tag{6}$$

Forgetting in the second stage is implemented similarly.

The second stage is similar to the second stage of the Task-Set Model. The only difference is that each MO has an MO-specific probability table P^2_{MO}. In the Task-Set Model, the CRP in the second stage using P^2 is independent of the first stage choices. In contrast, in the Option Model, the first stage choice determines which MO is activated (e.g. choosing A_1 for the circle in Experiment 1 is equivalent to choosing MO_1 as a whole, Fig. 2A), which then determines which probability table, P^2_{MO}, to use for running the CRP in the second stage and to select LOs. This implementation captures the essence of options in the HRL framework, in that selection of MO in the first stage constrains the policy chosen until the end of the second stage (where the option terminates). Specifically, for the P^2_{MO} activated by the MO chosen in the first stage, there are 16 contexts in the second stage of Experiment 1 (8 blocks and 2 first stage stimuli). If contexts $\{c^2_{1:n}\}$ are clustered on $N^2 \leq n$ LOs, when the model encounters a new context c^2_{n+1}, the prior probability of selecting a new low-level option LO_{n+1} in this new context is set to:

$$P^2_{MO}(LO_{n+1}|c^2_{n+1}) = \frac{\gamma^2}{Z^2}; \tag{7}$$

and the probability of reusing a previously created low-level option LO_i is set to:

$$P^2_{MO}(LO_i|c^2_{n+1}) = \frac{N^2_i}{Z^2}, \tag{8}$$

where γ^2 is the clustering coefficient for the CRP, N^2_i is the number of second stage contexts clustered on LO_i for the current MO, and $Z^2 = \gamma^2 + \sum_i N^2_i$ is the normalization constant. The new LO_{n+1} policy is initialized with uninformative Q-values $1/\#\{possible\, actions\} = \frac{1}{4}$. The model samples LO based on the conditional distribution over all LOs given the current context and MO. The model also tracks LO-specific policies via Q-learning. Once an LO is selected, a second stage policy is computed based on the LO's Q-values and the second stage stimulus S_i with softmax:

$$P(A_j^2|S_i, LO) = \frac{exp(\beta^2 * Q_{LO}^2(S_i, A_j^2))}{\sum_k exp(\beta^2 * Q_{LO}^2(S_i, A_k^2))}, \tag{9}$$

where β^2 is the inverse temperature. To account for the meta-learning heuristic, we add a free meta-learning parameter, m, that discourages selecting the same action in the second stage as in the first stage. Specifically, if π is the second stage policy as computed from softmax, we set $P(A^1) = m$, where A^1 is the action chosen in the first stage, and re-normalize:

$$P(A^{other}) = (1 - m) \times \pi(A^{other})/(1 - \pi(A^1)), \tag{10}$$

where A^{other} is any action other than A^1. A second stage action A^2, ranging from A_1 to A_4, is then sampled from this policy. After observing the outcome (moving on to the second stage or not), the model uses Bayes' Theorem to update P_{MO}^2:

$$P_{MO}^2(LO_k|c_j^2) = \frac{P(r|S_i, A^2, LO_k)P_{MO}^2(LO_k|c_j^2)}{(\sum_l P(r|S_i, A^2, LO_l)P_{MO}^2(LO_l|c_j^2))}, \tag{11}$$

where r is 1 if A^2 is correct and 0 otherwise, and $P(r|S_i, A^2, LO_l) = 1 - Q_{LO_l}^2(S_i, A^2)$ if $r = 0$, or $Q_{LO_l}^2(S_i, A^2)$ if $r = 1$. Then the Q-values of the LO with the highest posterior probability is updated:

$$Q_{LO}^2(S_i, A^2) = Q_{LO}^2(S_i, A^2) + \alpha^2 * (r - Q_{LO}^2(S_i, A^2)), \tag{12}$$

where α^2 is the learning rate.

Parameters for model simulations

Parameters used for main text

We used the set of parameters from Supplementary Table S1 in the main text to track participants' behavioral patterns both qualitatively and quantitatively.

A set of constrained parameters that capture behavior across all tasks qualitatively

In the main text, we selected parameters to try to trace participants' behavior patterns both quantitatively and qualitatively (Supplementary Table S1). Here we used another set of parameters (Supplementary Table S2) to (1) constrain parameters so that most experiments shared the same parameters while showing the qualitatively trends in participants' behavior and (2) show that the model can reproduce the same qualitative effects with a range of parameters.

Exp	Sample	Model	α^1	β^1	γ^1	f^1	α^2	β^2	γ^2	f^2	m
Exp 1	In-lab	Naive	0.5	4	NA	0.0025	0.7	10	NA	0.0001	0.01
		Flat	0.5	4	NA	0.0025	0.7	10	NA	0.0001	0.01
		Task-Set	1	2	14	0.0004	0.8	3	3	0.0002	0.01
		Option	1	2	14	0.0004	0.8	3	3	0.0002	0.01
	Mturk	Option	0.8	3	100	0.01	0.6	6	5	0.004	0.01
Exp 2	In-lab	Option	0.7	3	13	0.001	0.6	4	5	0.001	0.01
Exp 3	In-lab	Option	0.7	4	100	0.01	0.8	5	15	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.8	5	15	0.005	0.01
Exp 4	In-lab	Option	0.6	4	100	0.01	0.8	5	4	0.0002	0.01
	Mturk	Option	0.6	4	100	0.01	0.4	4	5	0.002	0.01
		Task-Set	0.6	4	100	0.01	0.4	4	5	0.002	0.01

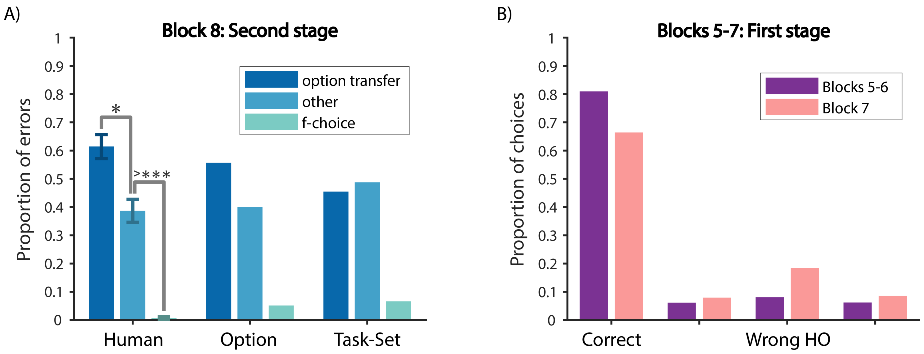
Supplementary Table S1: Parameters for the main text.

Exp	Sample	Model	α^1	β^1	γ^1	f^1	α^2	β^2	γ^2	f^2	m
Exp 1	In-lab	Naive	0.7	4	NA	0.001	0.7	4	NA	0.001	0.01
		Flat	0.7	4	NA	0.001	0.7	4	NA	0.001	0.01
		Task-Set	0.7	4	14	0.001	0.7	4	4	0.001	0.01
		Option	0.7	4	14	0.001	0.7	4	4	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	4	0.005	0.01
Exp 2	In-lab	Option	0.7	4	100	0.01	0.7	4	4	0.001	0.01
Exp 3	In-lab	Option	0.7	4	100	0.01	0.7	4	20	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	20	0.005	0.01
Exp 4	In-lab	Option	0.7	4	100	0.01	0.7	4	4	0.001	0.01
	Mturk	Option	0.7	4	100	0.01	0.5	4	4	0.005	0.01

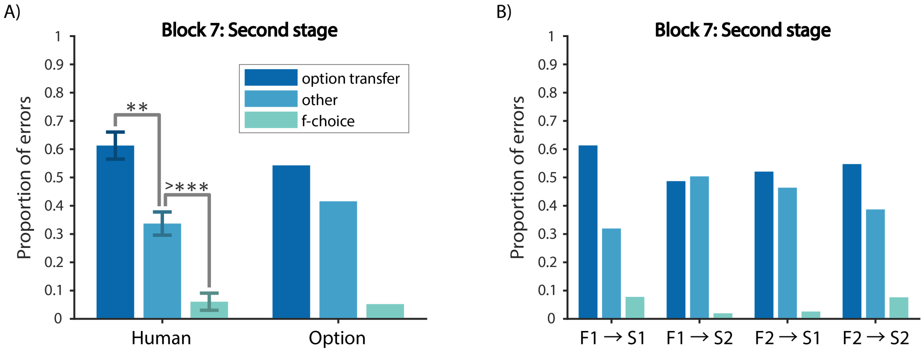
Supplementary Table S2: A second set of parameters that is constrained but still replicate transfer effects qualitatively.

In particular, we used $\alpha^1 = 0.7, \beta^1 = 4, \beta^2 = 4, m = 0.01$ for all experiments. For all in-lab experiments, we used $\alpha^2 = 0.7, f^2 = 0.001$; for all Mturk experiments, we used $\alpha^2 = 0.5, f^2 = 0.005$, which indicate slower learning rate and faster forgetting. For Experiment 1 in-lab, we used $\gamma^1 = 14, f^1 = 0.001$; for all other experiments, we used $\gamma^1 = 100, f^1 = 0.01$ to implement a lack of transfer effects in the first stage. We used $\gamma^2 = 20$ in Experiment 3 to model reduced option transfer in the second stage; for all other experiments, we used $\gamma^2 = 4$.

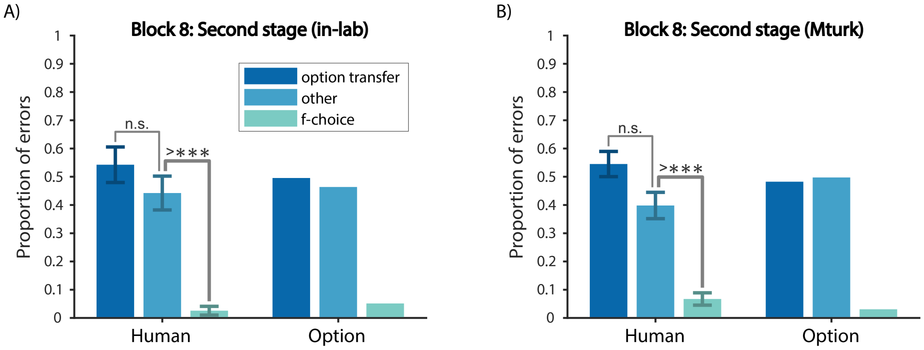
We recreated some of the representative analysis in the main text to demonstrate that this second set of parameters can replicate the transfer effects in human participants qualitatively well (Supplementary Fig. S3, Supplementary Fig. S4, Supplementary Fig. S5).



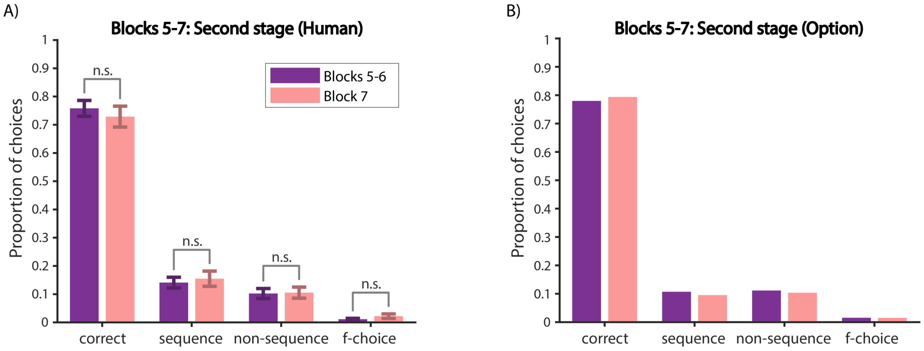
Supplementary Figure S3: Experiment 1 with parameters from Supplementary Table S2. (A) Error type analysis of the second stage in Block 8 for participants (left), the Option Model (middle) and the Task-Set Model (right). (B) Choice type analysis of the first stage in Blocks 5-7 for the Option Model.



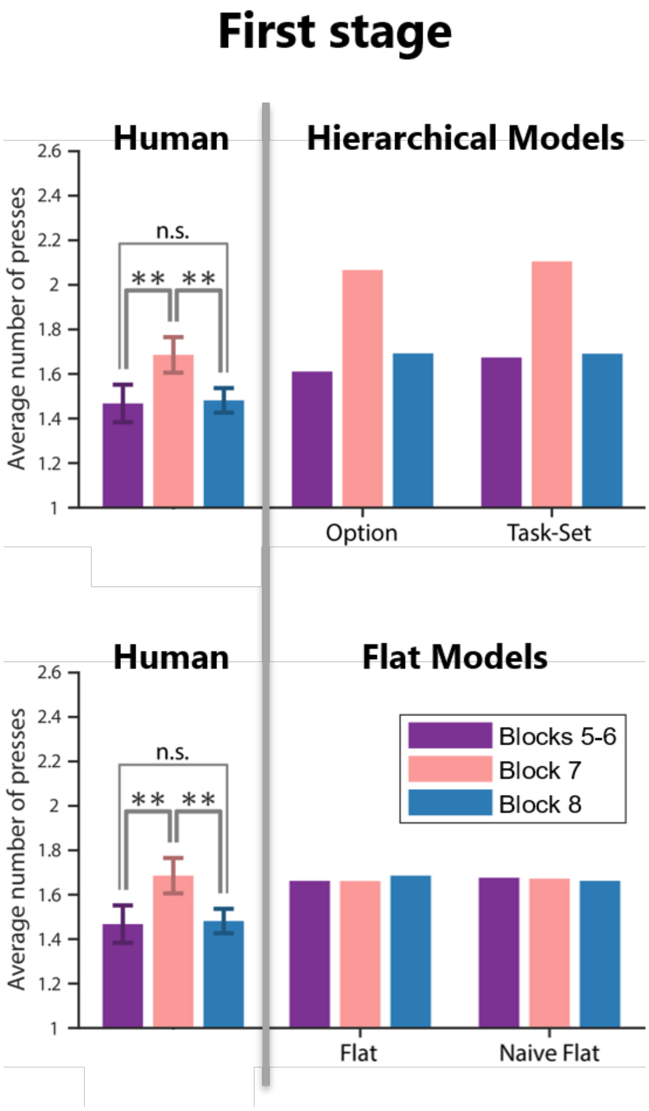
Supplementary Figure S4: Experiment 2 second stage choices with parameters from Supplementary Table S2 (A) Error type analysis of the second stage in Block 7 for participants (left) and the Option Model (right). (B) Error type analysis for each of the 4 branches in the second stage of Block 7 for the Option Model.



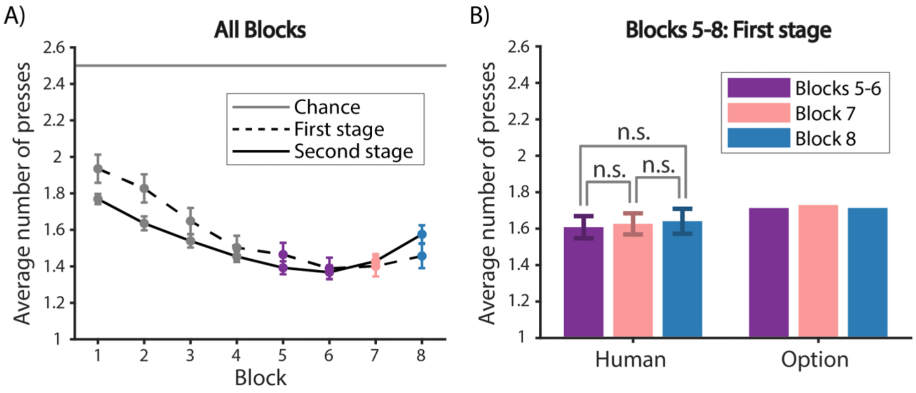
Supplementary Figure S5: Experiment 3 second stage choices with parameters from Supplementary Table S2. Error type analysis of the second stage in Block 8 for (A) in-lab participants (left) and the Option Model (right), and (B) Mturk participants (left) and the Option Model (right).



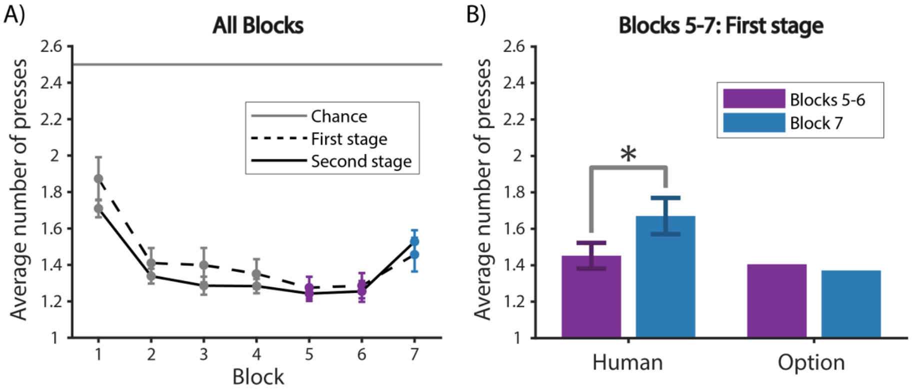
Supplementary Figure S6: Experiment 1 second stage choices. Choice type analysis of the second stage comparing Blocks 5-6 and Block 7 for (A) participants and (B) the Option Model. There was no significant difference across all choice types, indicating positive transfer in the second stage of Block 7.



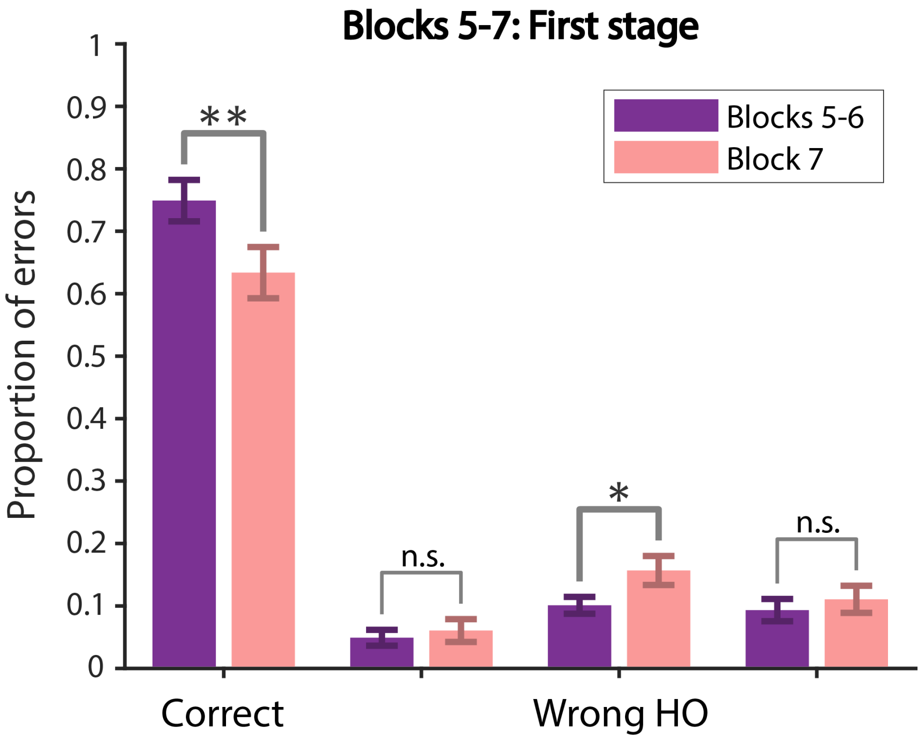
Supplementary Figure S7: Experiment 1 first stage transfer effects. Average number of first stage key presses in the first 10 trials of Block 5-8 for participants as well as model simulations. We ran 500 simulations of each hierarchical model (top) and flat model (bottom). See Supplementary Table S1 for model parameters. Behavioral results show patterns of positive and negative transfer predicted by hierarchical, but not flat RL models.



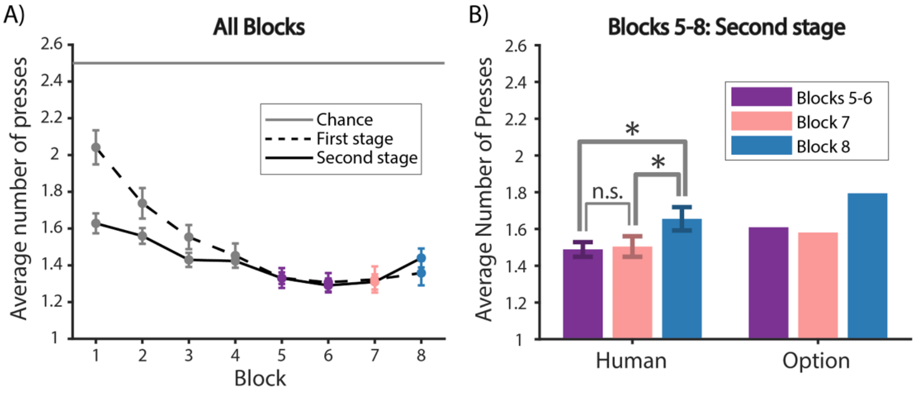
Supplementary Figure S8: Experiment 1 Mturk results. (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the first stage for participants (left) and the Option Model (right).



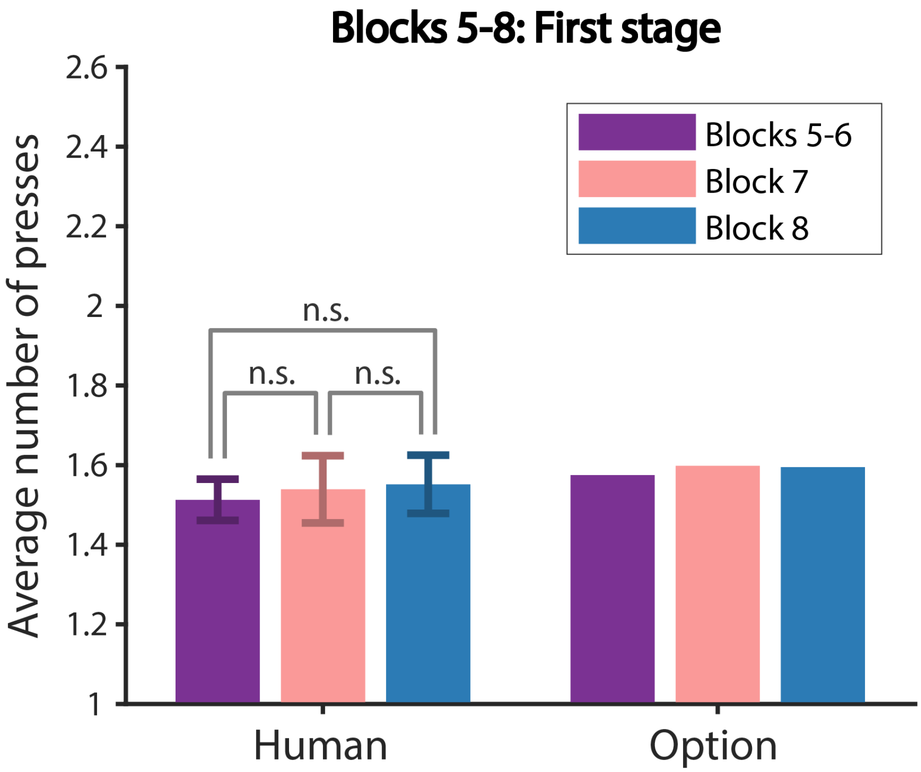
Supplementary Figure S9: Experiment 2 results. (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-7 for the first stage for participants (left) and the Option Model (right).



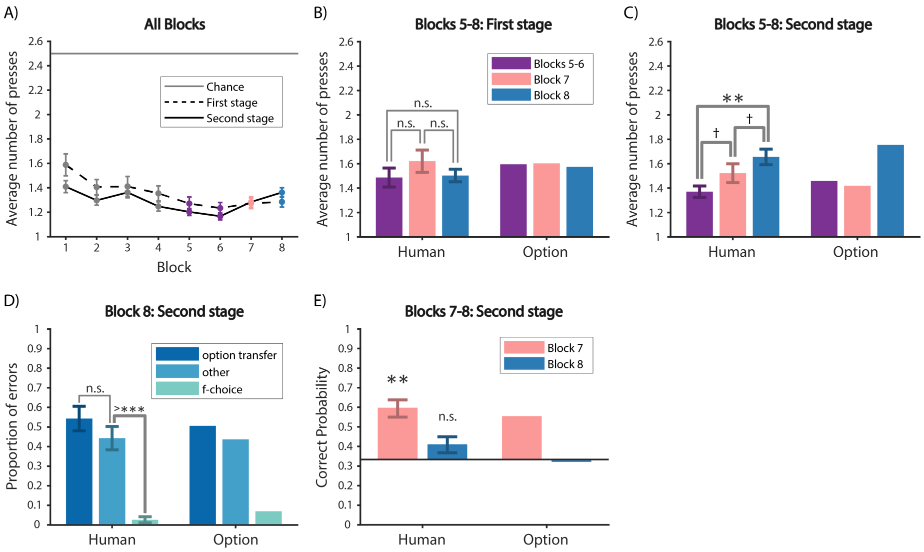
Supplementary Figure S10: Experiment 2 first stage choices. Choice type analysis of the first stage comparing Blocks 5-6 and Block 7. The only error type that significantly increased was the wrong *HO* error, suggesting that participants were perseverating in the first stage while learning the new mappings in the second stage of Block 7.



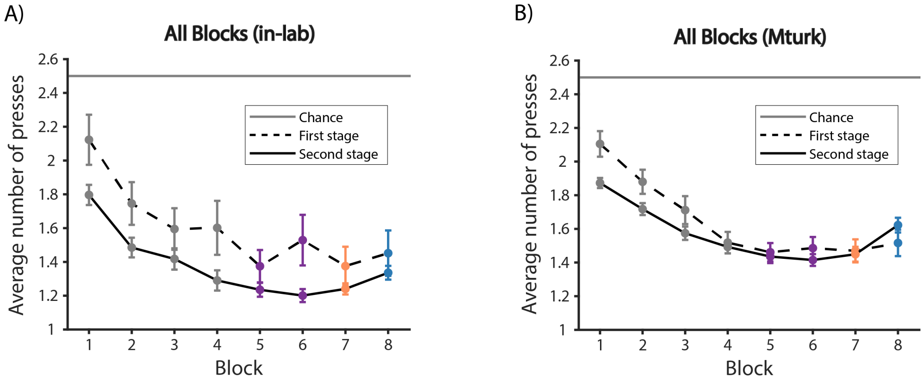
Supplementary Figure S11: Experiment 3 Mturk results. (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the second stage for participants (left) and the Option Model (right).



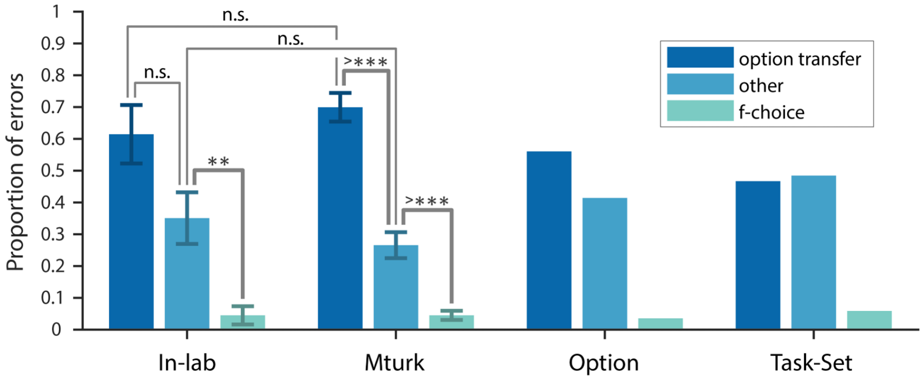
Supplementary Figure S12: Experiment 3 Mturk first stage choices. Average number of presses in the first 10 trials of Blocks 5-8 in the first stage for participants (left) and the Option Model (right). This shows a lack of transfer in the first stage, representative of Experiments 3-4 first stage for both in-lab and Mturk populations.



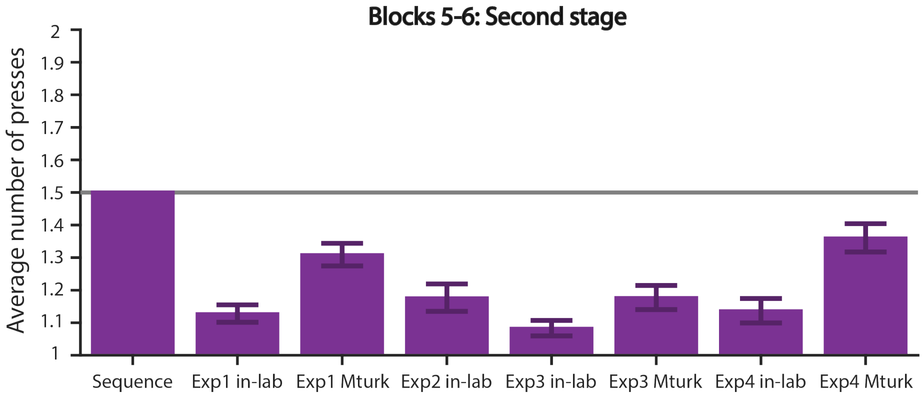
Supplementary Figure S13: Experiment 3 summary. (A) Average number of key presses in the first and the second stages per block. (B) Average number of key presses for the first 10 trials of Blocks 5-8 for the first stage for participants (left) and the Option Model (right). (C) Same as (B) for the second stage. (D) Error type analysis of the second stage in Block 8 for participants (left) and the Option Model (right). The proportion of option transfer error was not significantly different from other error, different from Experiment 1 and Experiment 2, suggesting reduced option transfer. (E) Probability of a correct first key press for the second stage of the first trial of each of the 4 branches in Blocks 7-8 for participants (left) and the Option Model (right).



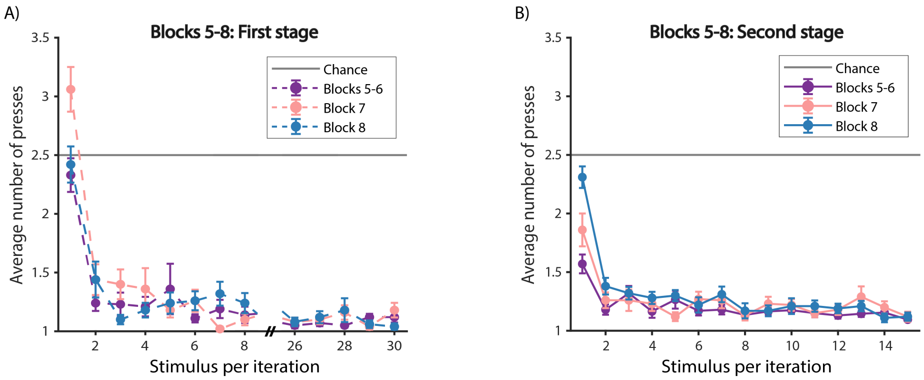
Supplementary Figure S14: Experiment 4 number of presses. Average number of key presses in the first and the second stages per block for (A) in-lab participants and (B) Mturk participants.



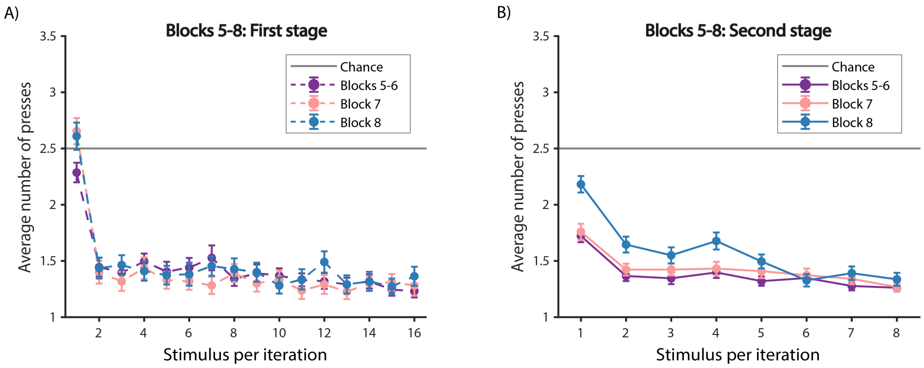
Supplementary Figure S15: Experiment 4 second stage errors reveal temporal options transfer and compositionality. Error type analysis of the second stage in Block 7 for the mismatch condition for in-lab participants, Mturk participants, the Option Model and the Task-Set Model.



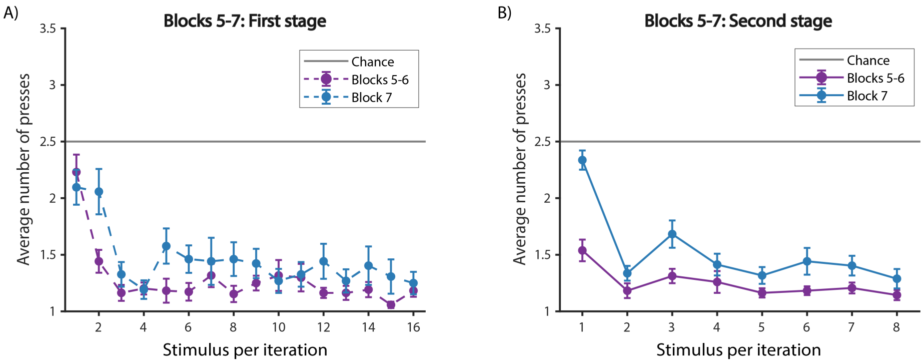
Supplementary Figure S16: Comparison of sequence learning model asymptotic performance with participants' performance in the last 10 trials of Blocks 5 and 6 across all 7 datasets (4 in-lab and 3 Mturk). While the sequence learning model is stuck at 1.5 presses/trial on average, participants performed significantly better.



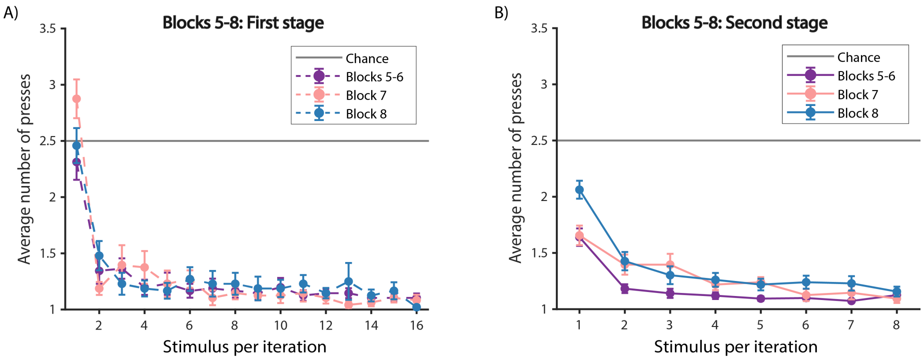
Supplementary Figure S17: Experiment 1 performance within Blocks 5-8 for in-lab participants. (A) First stage. (B) Second stage.



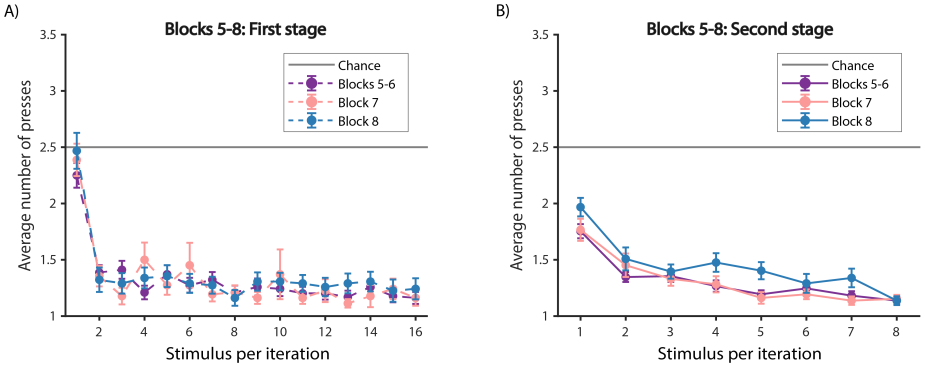
Supplementary Figure S18: Experiment 1 performance within Blocks 5-8 for Mturk participants. (A) First stage. (B) Second stage.



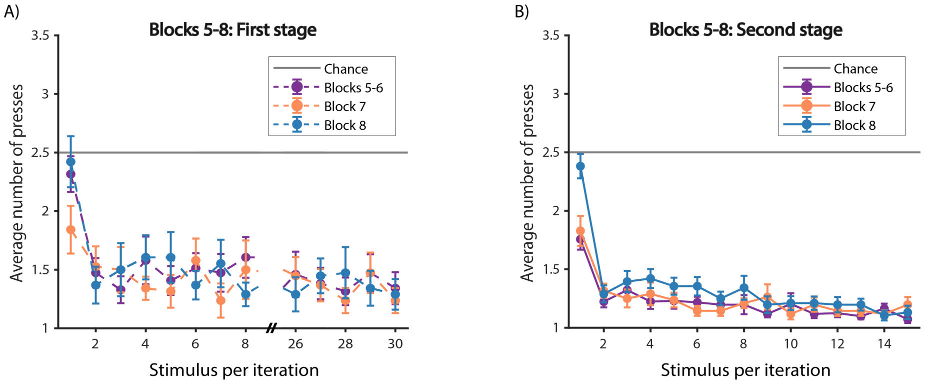
Supplementary Figure S19: Experiment 2 performance within Blocks 5-7. (A) First stage. (B) Second stage.



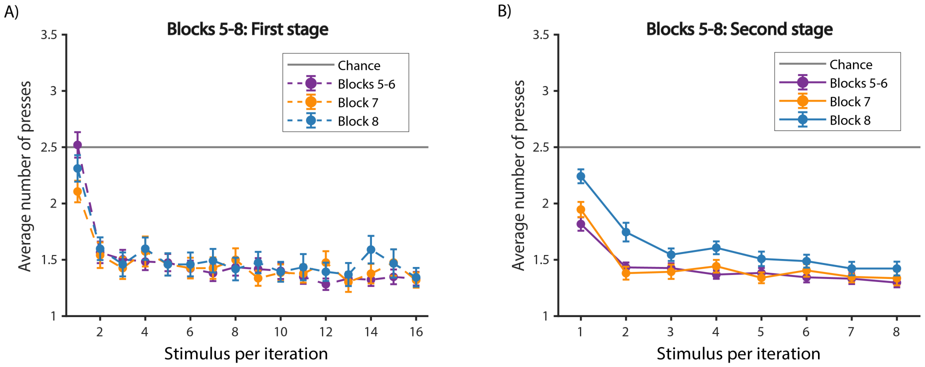
Supplementary Figure S20: Experiment 3 performance within Blocks 5-8 for in-lab participants. (A) First stage. (B) Second stage.



Supplementary Figure S21: Experiment 3 performance within Blocks 5-8 for Mturk participants. (A) First stage. (B) Second stage.



Supplementary Figure S22: Experiment 4 performance within Blocks 5-8 for in-lab participants. (A) First stage. (B) Second stage.



Supplementary Figure S23: Experiment 4 performance within Blocks 5-8 for Mturk participants. (A) First stage. (B) Second stage.

Exp	18-25	26-30	31-35	36-40	41+	Unknown	Total
Exp 1	14	18	26	23	33	2	116
Exp 3	4	9	18	9	25	0	65
Exp 4	14	17	24	15	40	0	110

Supplementary Table S3: Age range distribution for Mturk participants in Experiments 1, 3, and 4.