Appendix A: Summary of Experimental Conditions

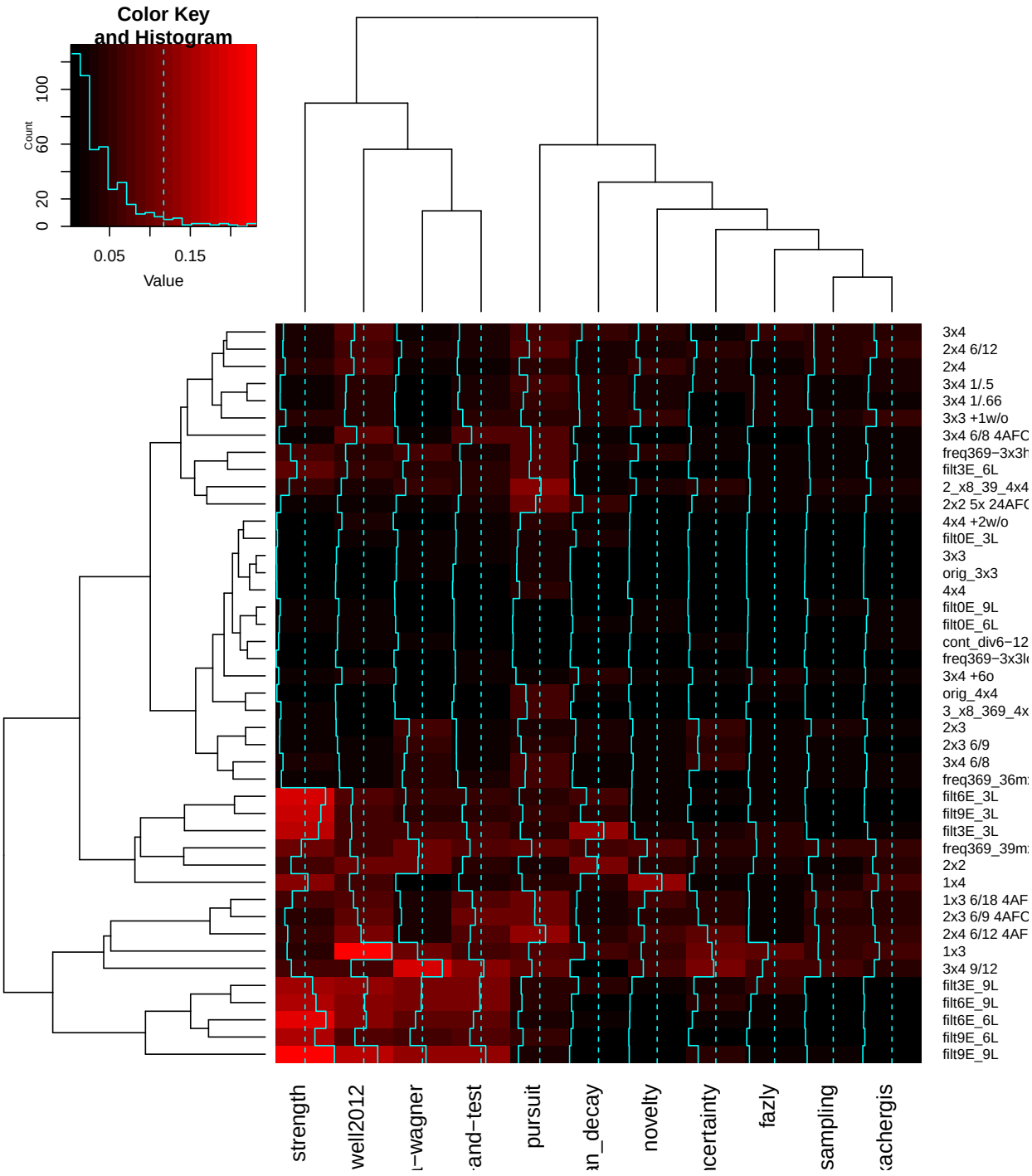
Table 1 describes the 44 experimental conditions included in the model comparison, including the number of trials, the number of words presented per trial (Words/Trial), the number of referents presented per trial (Objects/Trial), the number of to-be-learned word-referent pairs (Items), people's overall mean accuracy ($p(o|w)$ across all intended $w - o$ mappings) in each condition (Accuracy), the standard deviation of performance (SD), and the number of participants per condition (N).

Appendix B: Clustering experiments and models by misfit

The table below shows cross-validated model fit (SSE; sum of squared error) for each experimental condition.

Condition Name	Trials	Words/Trial	Objects/Trial	Items	Accuracy	SD	N
3x4	36	3	4	18	0.19	0.12	25
3x4 1/.5	36	3	4	18	0.22	0.11	25
3x4 1/.66	36	3	4	18	0.21	0.09	25
3x4 +6o	36	3	4	18	0.27	0.12	20
2x4	54	2	4	18	0.30	0.15	33
3x3 +1w/o	54	3	3	18	0.17	0.08	39
4x4 +2w/o	54	4	4	18	0.10	0.05	39
1x3 6/18 4AFC	108	1	3	18	0.67	0.10	43
2x3 6/9 4AFC	54	2	3	18	0.69	0.11	38
2x4 6/12 4AFC	54	2	4	18	0.62	0.12	31
3x4 6/8 4AFC	36	3	4	18	0.69	0.09	36
1x3	108	1	3	18	0.74	0.11	23
2x3	54	2	3	18	0.58	0.07	23
2x4 6/12	54	2	4	18	0.33	0.18	14
3x4 6/8	36	3	4	18	0.42	0.14	13
2x3 6/9	54	2	3	18	0.55	0.11	32
3x4 9/12	54	3	4	18	0.69	0.08	33
1x4	108	1	4	18	0.19	0.10	40
1x3	108	1	3	18	0.39	0.12	40
4x4	27	4	4	18	0.31	0.07	77
3x3	36	3	3	18	0.43	0.08	36
2x2	54	2	2	18	0.79	0.11	19
2x2 5x 24AFC	60	2	2	24	0.51	0.12	46
filt0E_3L	18	2	2	12	0.38	0.09	31
filt3E_3L	27	2	2	12	0.72	0.12	30
filt6E_3L	36	2	2	12	0.71	0.09	30
filt9E_3L	45	2	2	12	0.70	0.08	31
filt0E_6L	36	2	2	12	0.47	0.11	27
filt3E_6L	45	2	2	12	0.67	0.11	27
filt6E_6L	54	2	2	12	0.79	0.09	27
filt9E_6L	63	2	2	12	0.75	0.09	27
filt0E_9L	54	2	2	12	0.54	0.09	31
filt3E_9L	63	2	2	12	0.83	0.08	31
filt6E_9L	72	2	2	12	0.82	0.09	31
filt9E_9L	81	2	2	12	0.86	0.06	31
2_x8_39_4x4	27	4	4	18	0.41	0.16	30
3_x8_369_4x4	27	4	4	18	0.33	0.06	74
freq369-3x3loCD	36	3	3	18	0.33	0.08	102
freq369-3x3hiCD	36	3	3	18	0.56	0.12	26
freq369_36mx	36	3	3	18	0.45	0.16	62
freq369_39mx	36	3	3	18	0.62	0.14	66
orig_4x4	27	4	4	18	0.27	0.07	88
orig_3x3	36	3	3	18	0.43	0.08	104
cont_div6-12	36	3	3	18	0.43	0.11	40

Table 1: Summary of modeled datasets.



Takeaways: Some experimental conditions are hard for particular models to fit. For example, the strength-biased model has particular difficulty with the `filt` conditions[1] (Kachergis et al., 2012), which present a group of word-referent pairs early in training which then systematically co-occur with particular novel late-stage word-referent pairs, testing how strictly learners will maintain a mutual exclusivity (ME) constraint. The Trueswell2012 model also shows greater misfit in most of these conditions (except for the `filtXE_3L` conditions, which have only 3 repetitions of the late-stage pairings, and thus do not overwhelm learners' ME bias).

[1]Except for the `filtOE_` conditions, which consist only of the late-stage pairs, with no early stage.

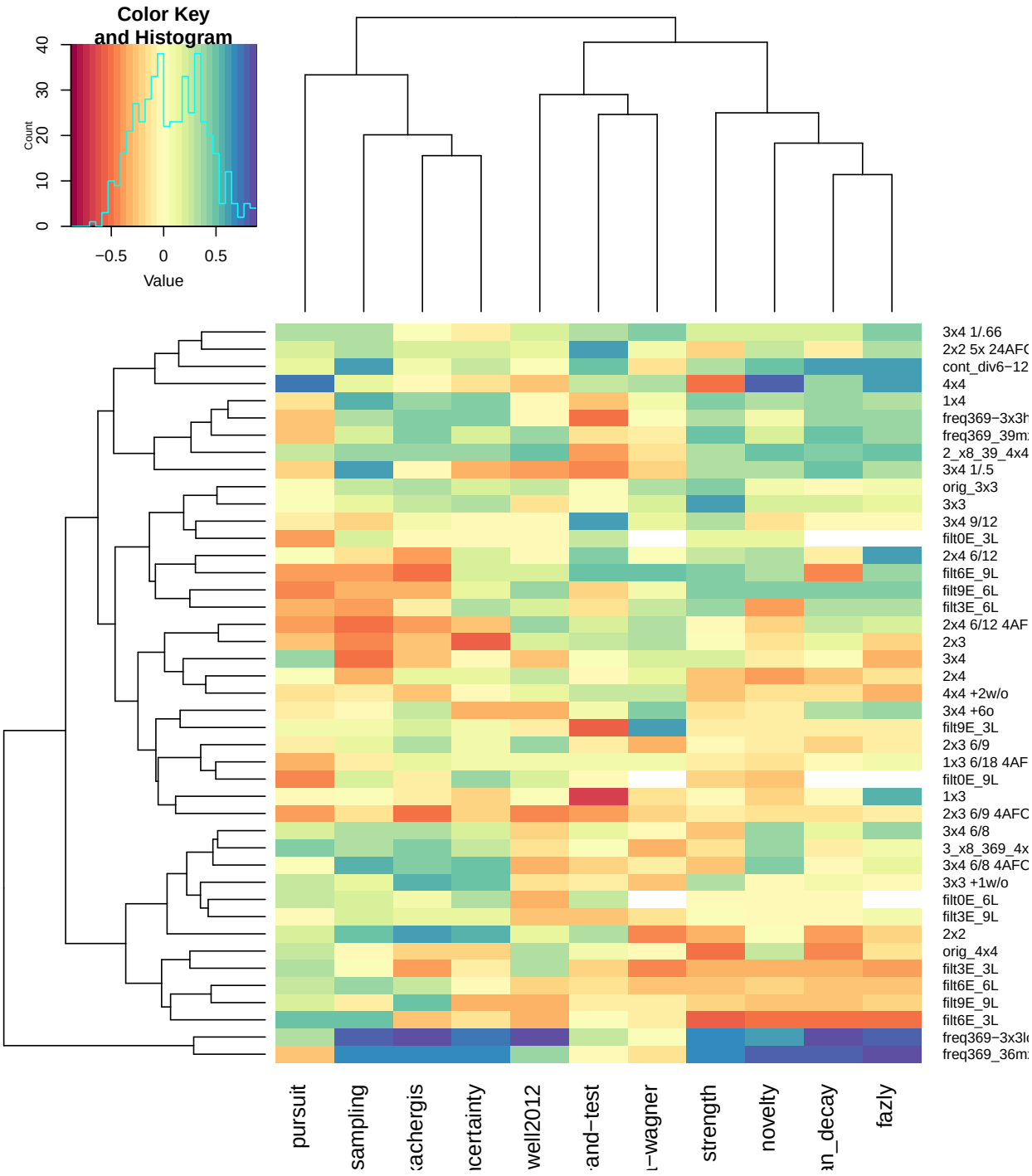
On the other hand, many experimental conditions are nearly equally well fit by all models, especially those that have a fixed number of repetitions per word-referent pair (e.g., 3x3, 4x4,)

Other ideas:

- Find experiment with maximal discrimination between the models? (highest SD of fit?)
- Find experiment that each model best predicts? and worst predicts?

Correlation of model predictions with item perforance per condition

Each cell displays the correlation coefficient of model vs. human item-level performance in a given experimental condition.



Each Model's Best- and Worst-Fitting Experiment

Which experiments are models best fitting?

Best and worst fitting models per experiment

(Could order conditions by how much they discriminate models?)

Model	best	worst	best_cond	worst_cond
Bayesian_decay	0.88	-0.53	freq369-3x3loCD	filt6E_3L
fazly	0.88	-0.52	freq369_36mx	filt6E_3L
kachergis	0.84	-0.52	freq369-3x3loCD	2x3 6/9 4AFC
trueswell2012	0.84	-0.43	freq369-3x3loCD	2x3 6/9 4AFC
novelty	0.78	-0.52	freq369_36mx	filt6E_3L
kachergis_sampling	0.78	-0.52	freq369-3x3loCD	3x4
uncertainty	0.76	-0.56	freq369-3x3loCD	2x3
pursuit	0.72	-0.44	4x4	filt0E_9L
strength	0.69	-0.53	freq369-3x3loCD	filt6E_3L
guess-and-test	0.62	-0.70	2x2 5x 24AFC	1x3
rescorla-wagner	0.61	-0.46	filt9E_3L	2x2

Table 2: Each model's best- and worst-fitting experiment.

Model	best_fits	worst_fits
Bayesian_decay	3	1
fazly	6	1
guess-and-test	4	7
kachergis	4	3
kachergis_sampling	5	3
novelty	1	2
pursuit	4	6
rescorla-wagner	4	8
strength	4	5
trueswell2012	5	4
uncertainty	3	3

Table 3: Number of experiments that each model fits best, and worst.