None of the above: A Bayesian account of the detection of novel categories

Dan Navarro
School of Psychology
University of New South Wales

Charles Kemp
School of Psychology
Carnegie Mellon University

compcogscisydney.com/projects.html#noneoftheabove

Dragon

Unicorn

Dragon

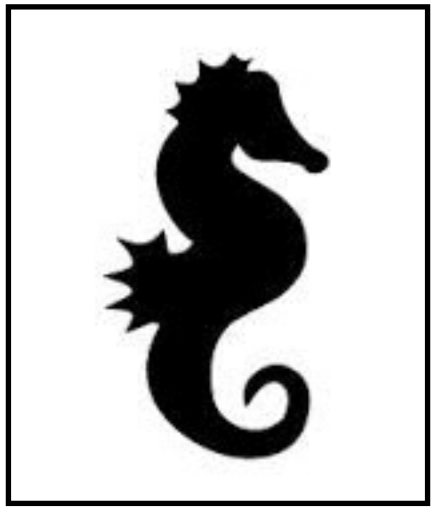
Dragon

Unicorn

Dragon

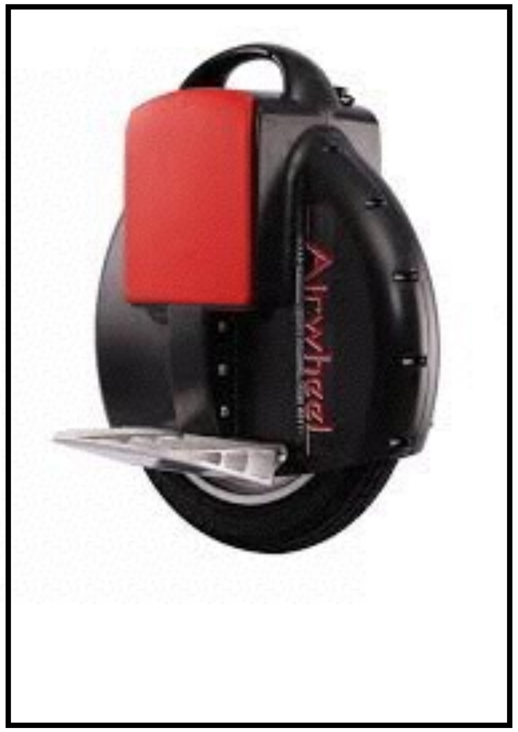
Unicorn

Unicorn



Is this a dragon or a unicorn?

Unicycle?
Segway?
Roomba?



Unicycle?
Segway?
Roomba?

None of the above… this is the first item from a novel category

The "mental dictionary" of categories is extensible… how do we know when to extend it?

air wheels

dabs

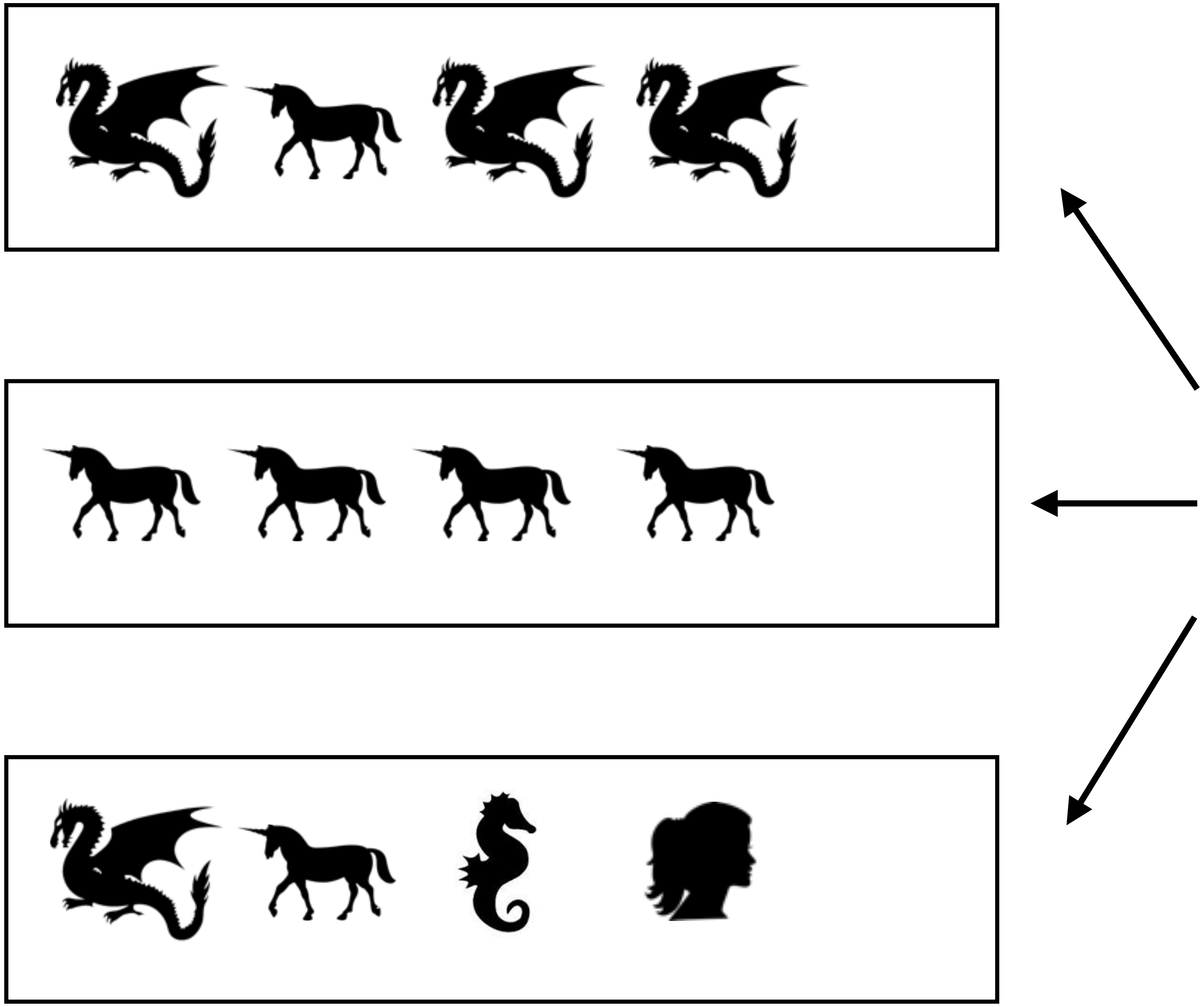
dwarf planets

Structure of the talk

- Qualitative desiderata, models, a priori predictions
- Experiments with minimal cues
    - Exp. 1: people satisfy the desiderata
    - Exp. 2: no they don’t
- An absurd number of computational models
- Experiments with similarity structure
    - Exp. 3: people integrate similarity & distribution
    - Exp. 4: a better version of Exp. 3
- Conclusions

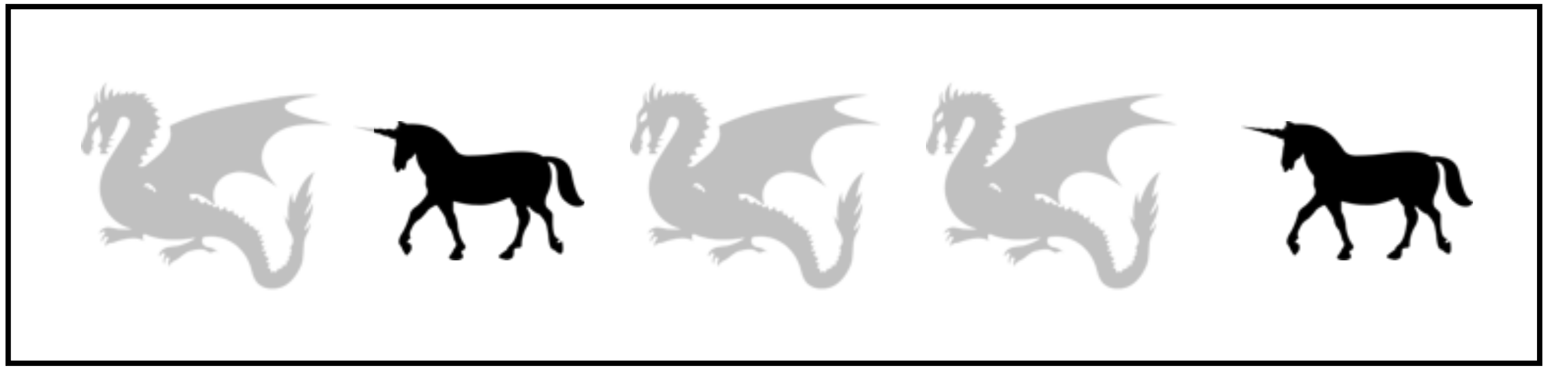
Qualitative desiderata for the discovery of new categories…

(Zabell 2011)

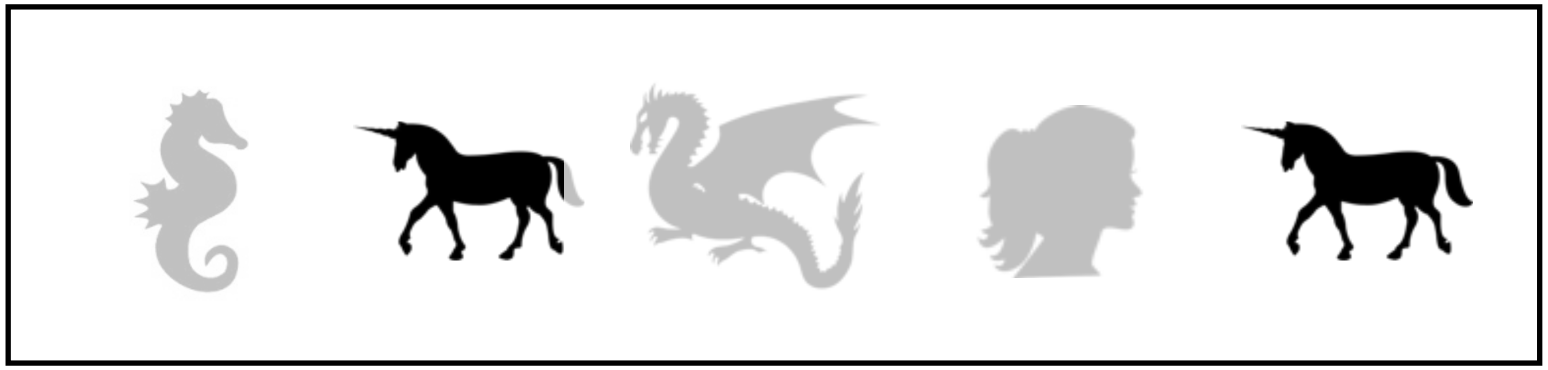


Any sequence of observations is *possible*, so I must (a priori) assign non-zero probability to them

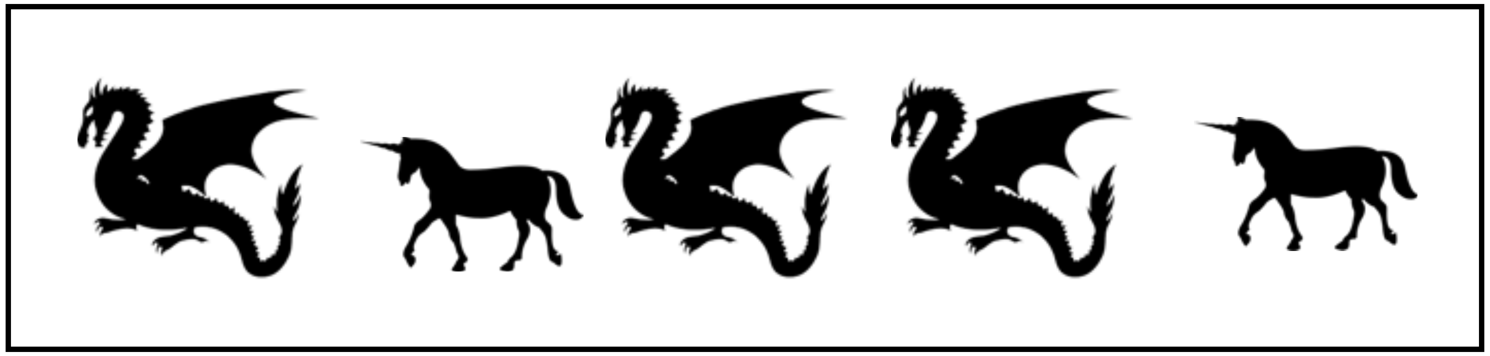
Same number of unicorns… so my beliefs about P(unicorn) should be the same



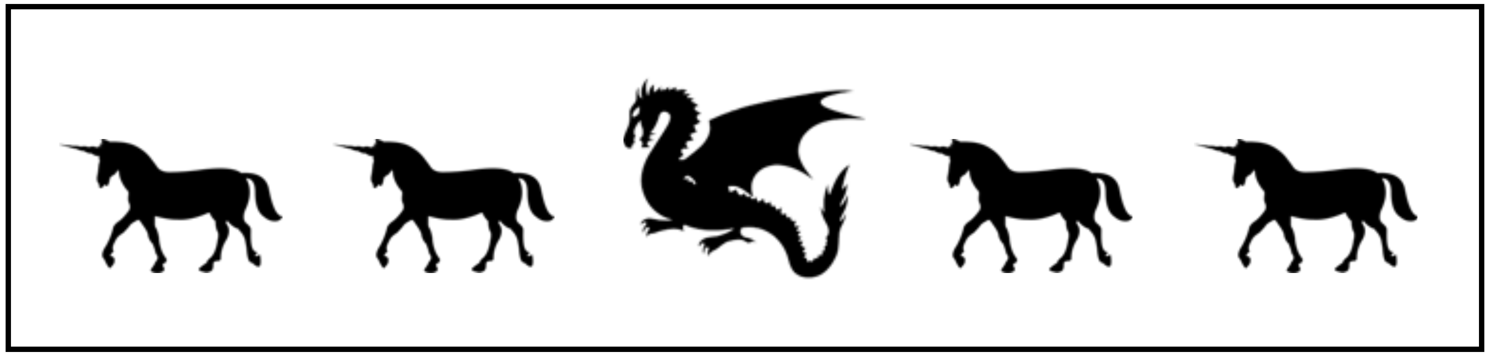
= 2/5 unicorns



= 2/5 unicorns



= 2 categories



= 2 categories

Same number of familiar categories so the probability of a new category is the same

What prior beliefs must a learner have in order to satisfy those desiderata?

Bayesian category learning models use the “Chinese restaurant process” (CRP)…

(Anderson 1990, Sanborn, Griffiths & Navarro 2010, etc)

$$\boxed{P(\text{old } k) \propto n_k}$$

← “Strength” associated with an existing category is proportional to its frequency

Bayesian category learning models use the “Chinese restaurant process” (CRP)…

(Anderson 1990, Sanborn, Griffiths & Navarro 2010, etc)

$$\boxed{\begin{array}{c} P(\text{old } k) \propto n_k \\ P(\text{new}) \propto \theta \end{array}}$$

↑

There is a ***fixed*** strength
associated with novelty

… but it's a special case: the full* solution to the problem is the *generalised* CRP

(Zabell, 2011)

$$\boxed{\begin{aligned} P(\text{old } k) &\propto n_k - \alpha \\ P(\text{new}) &\propto \theta + K\alpha \end{aligned}}$$

* sort of

… but it’s a special case: the full* solution to the problem is the *generalised* CRP

(Zabell, 2011)

$$\boxed{P(\text{old } k) \propto n_k - \alpha}$$

← “Strength” of old categories is slightly attenuated relative to the CRP…

* sort of

… but it’s a special case: the full* solution to the problem is the *generalised* CRP

(Zabell, 2011)

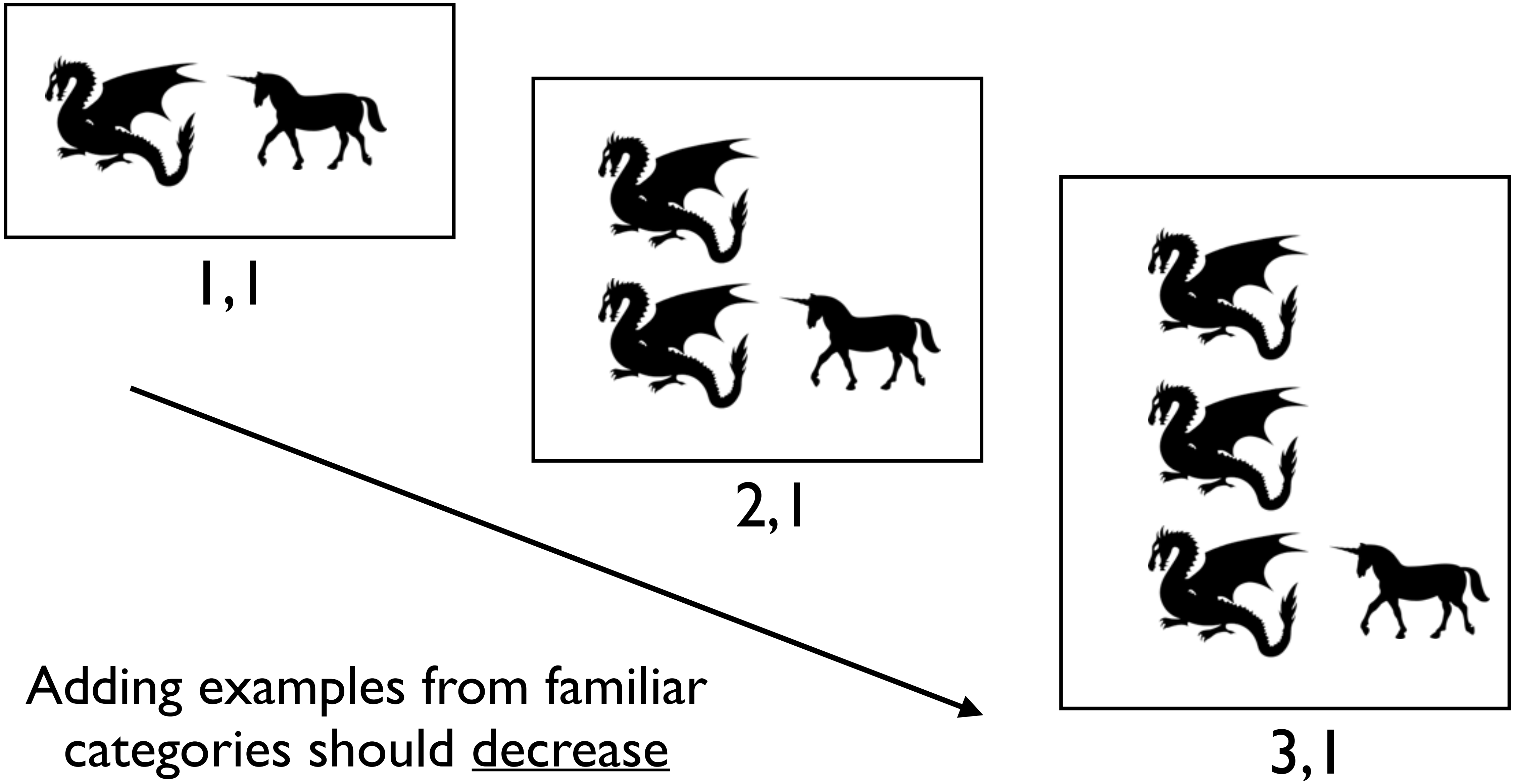
$$\boxed{\begin{aligned} P(\text{old } k) &\propto n_k - \alpha \\ P(\text{new}) &\propto \theta + K\alpha \end{aligned}}$$

… because every time a new category appears, P(new) goes up

* sort of

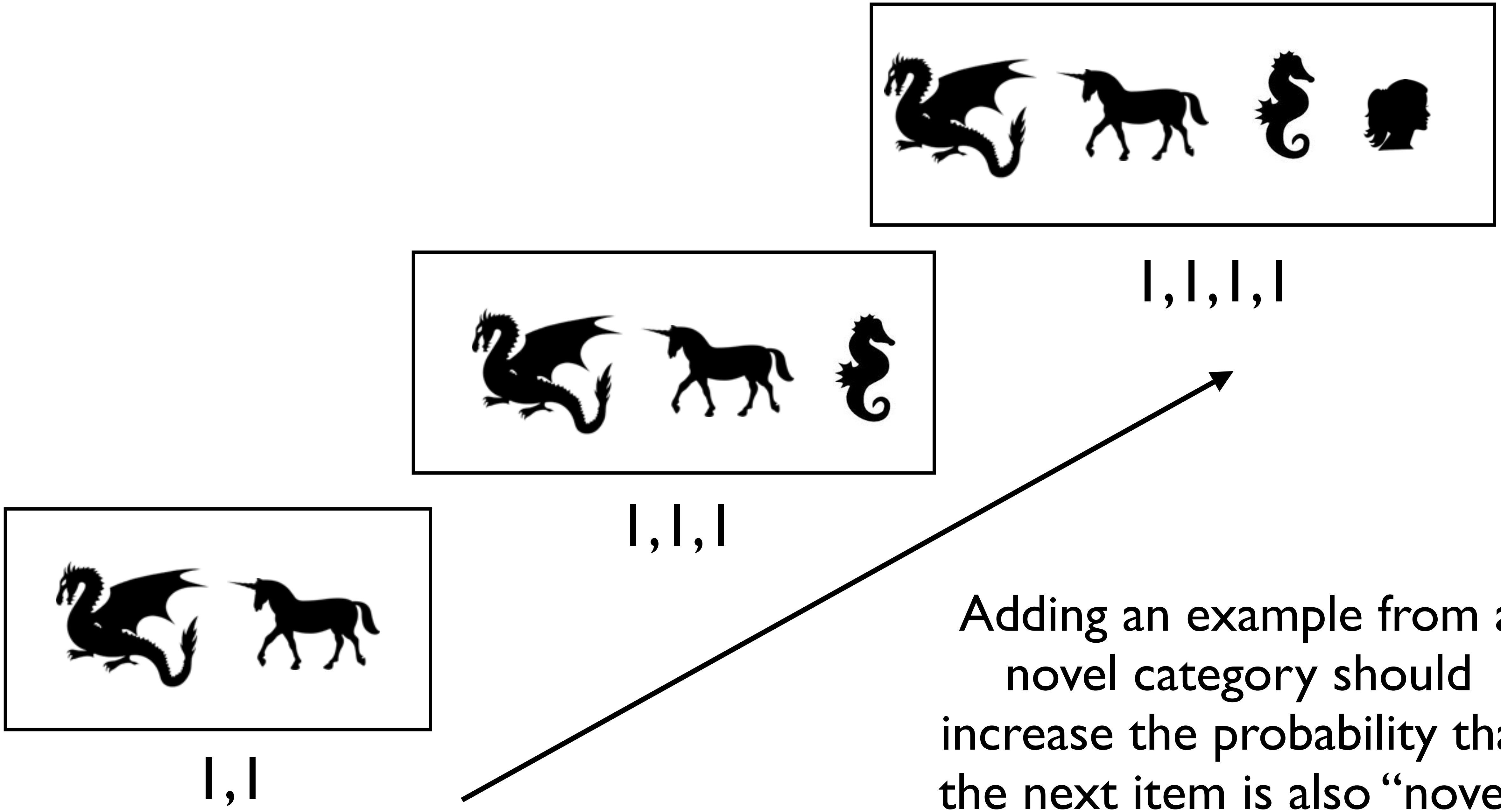
What empirical predictions does the G-CRP make for human novelty detection?

1: The familiar addition effect



Adding examples from familiar categories should decrease the probability of labelling the next thing as "novel"

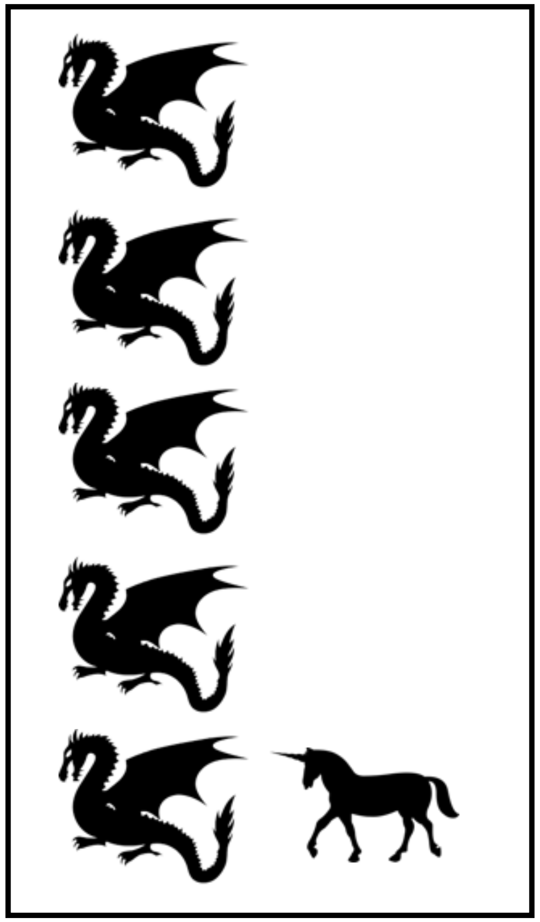
2: The novel addition effect



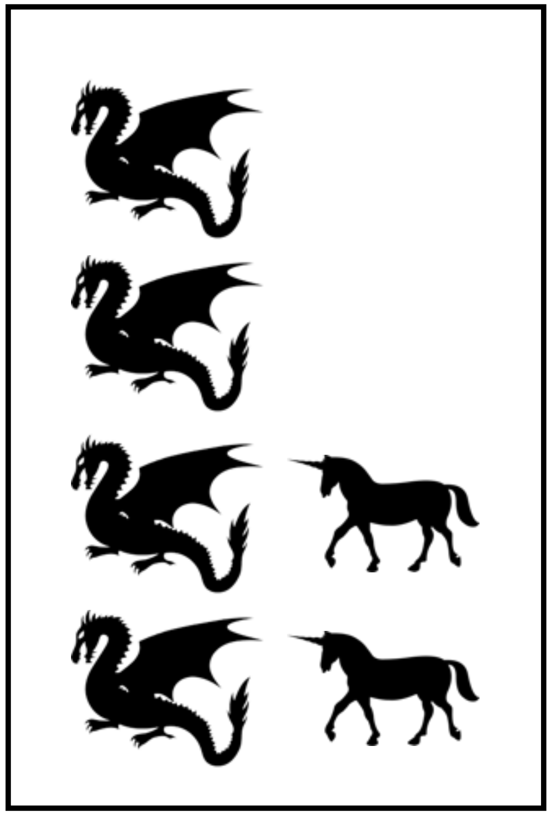
Adding an example from a novel category should increase the probability that the next item is also "novel"

3: No effect of transfer

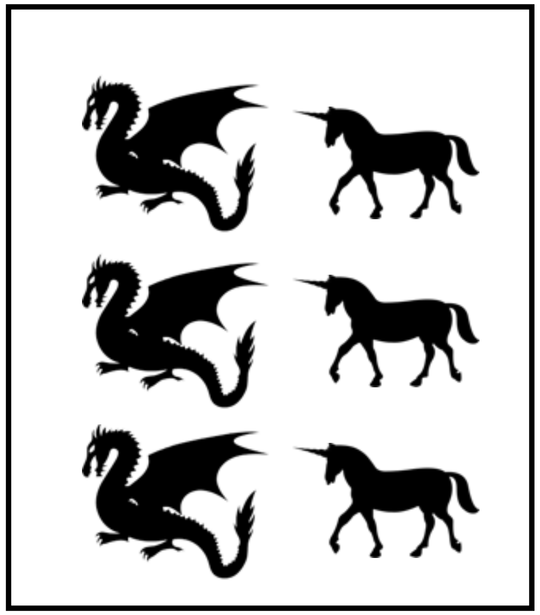
Nothing else about the frequency table matters except the number of exemplars N and the number of categories K



5,1

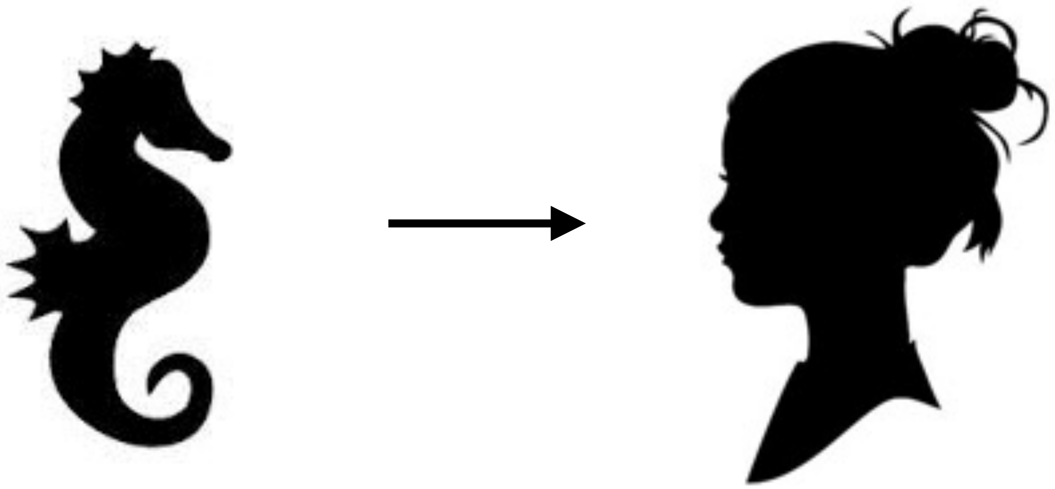


4,2

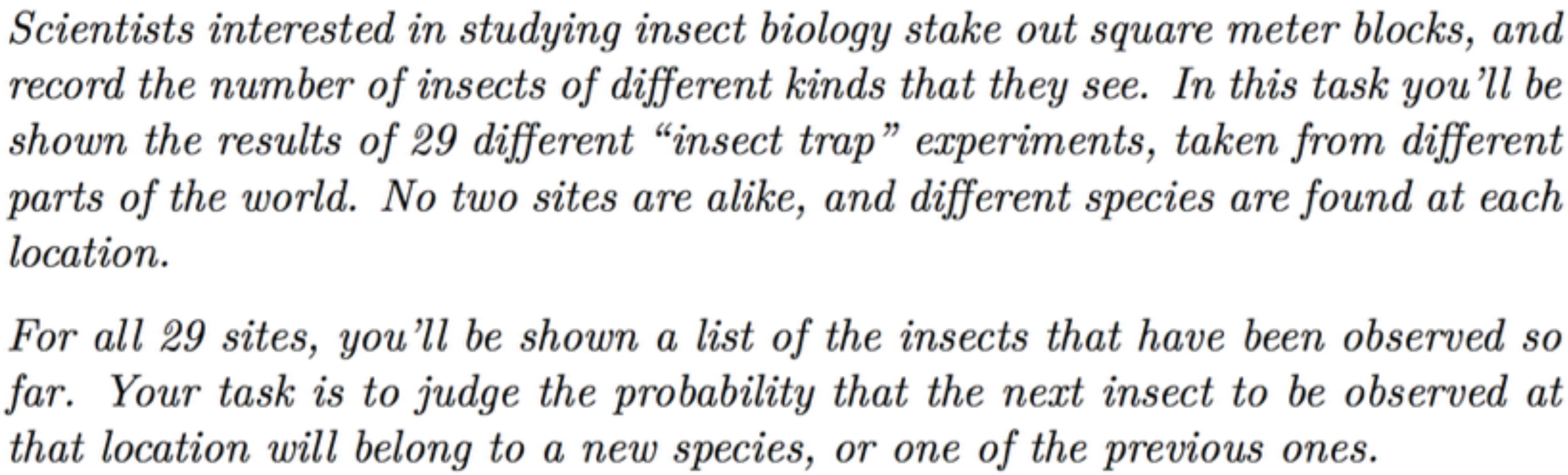


3,3

Experiments…



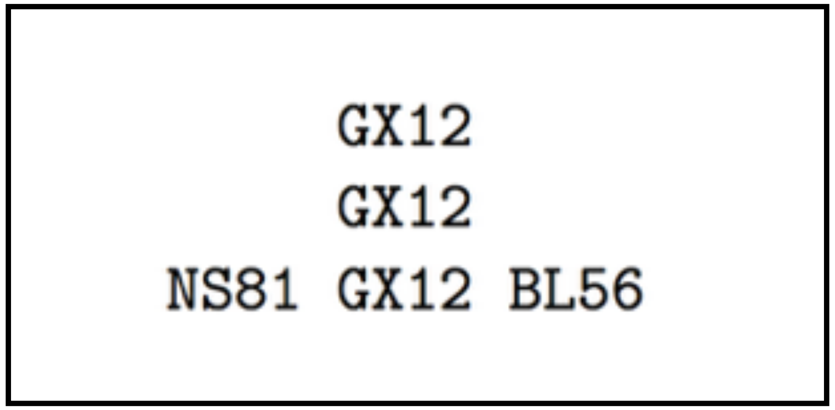
Experiment 1



Scientists interested in studying insect biology stake out square meter blocks, and record the number of insects of different kinds that they see. In this task you'll be shown the results of 29 different "insect trap" experiments, taken from different parts of the world. No two sites are alike, and different species are found at each location.

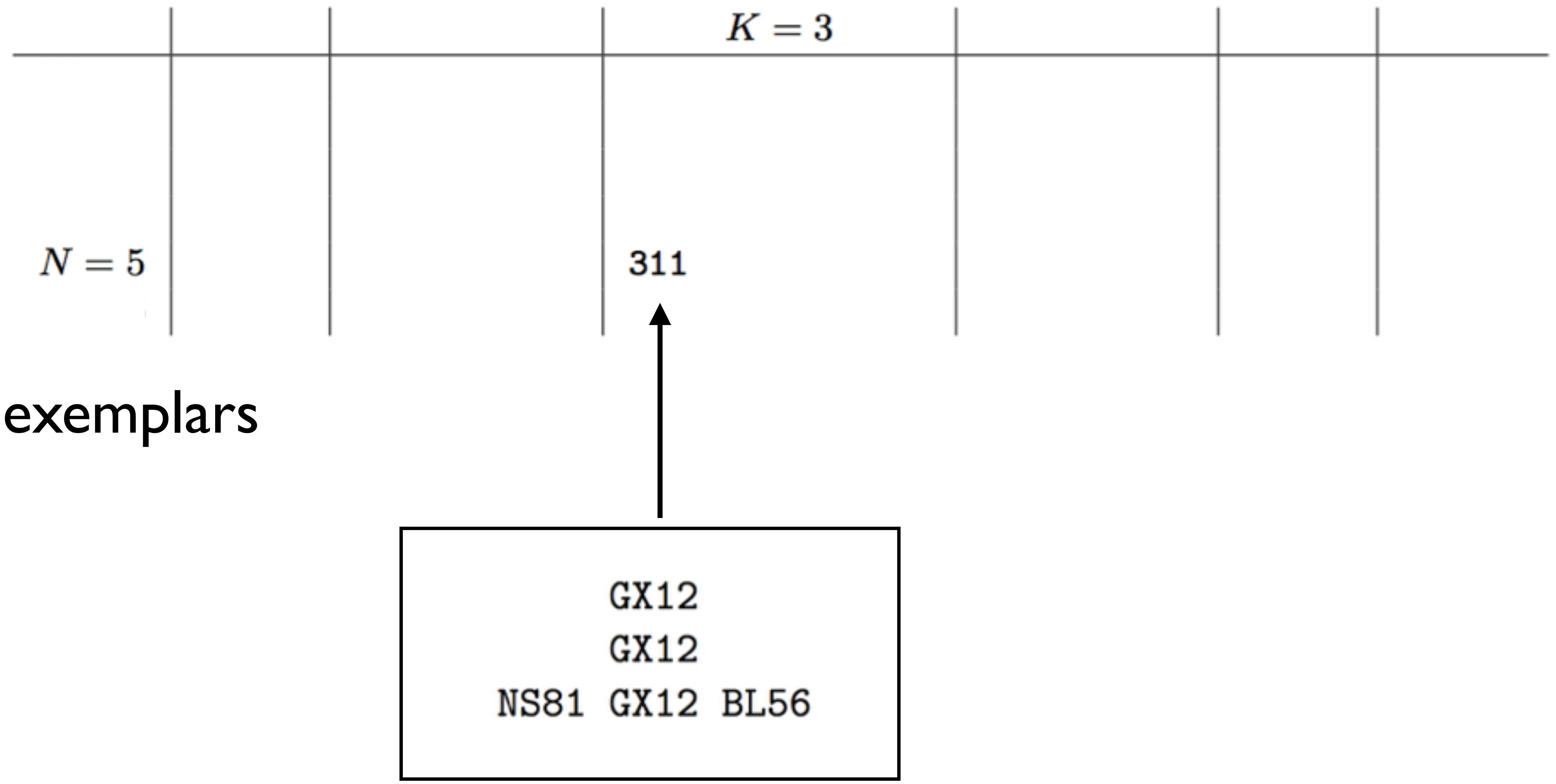
For all 29 sites, you'll be shown a list of the insects that have been observed so far. Your task is to judge the probability that the next insect to be observed at that location will belong to a new species, or one of the previous ones.

Stimuli were just arbitrary alphanumeric labels, to prevent similarity effects



```
     GX12
     GX12
NS81 GX12 BL56
```

categories



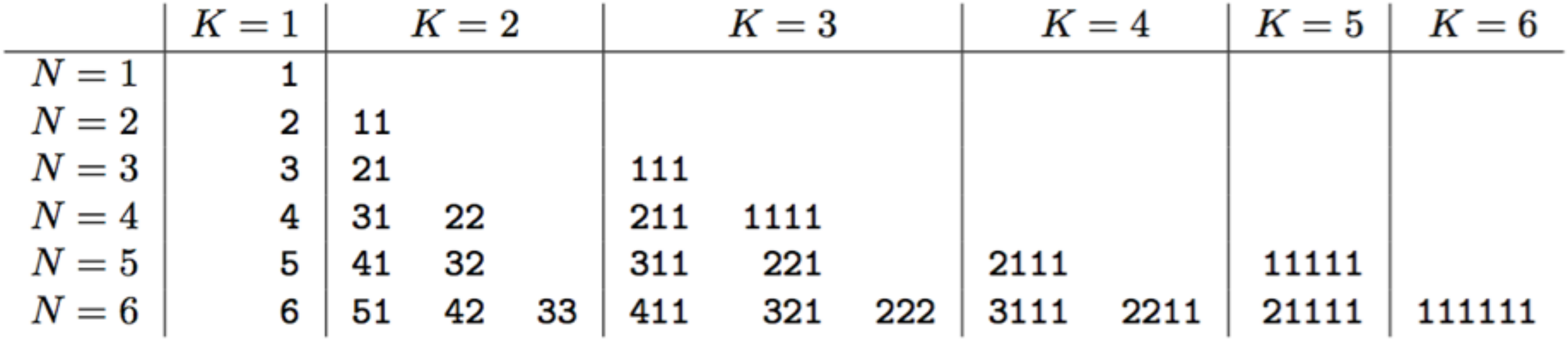
$K = 3$

$N = 5$

311

exemplars

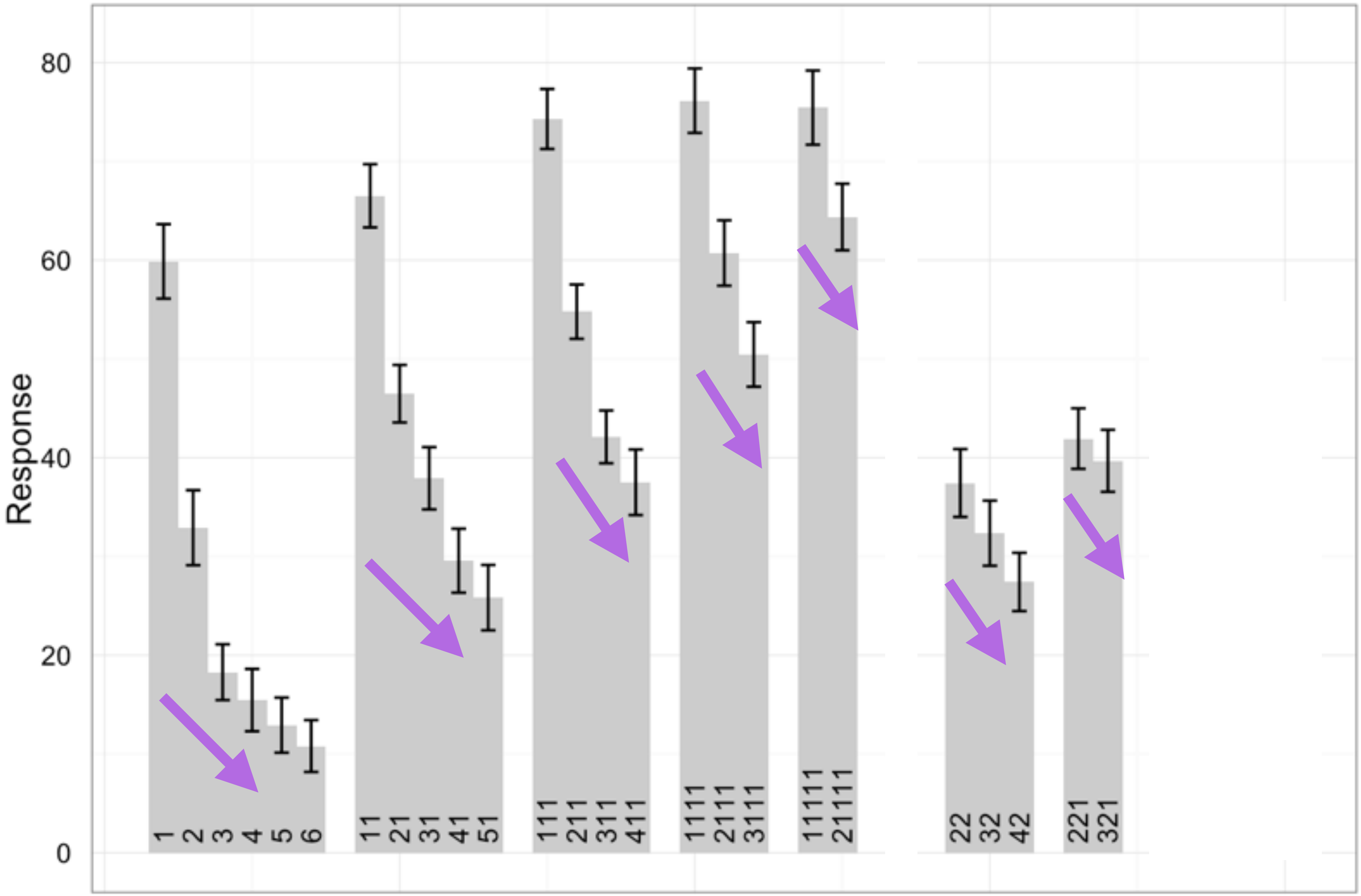
GX12
GX12
NS81 GX12 BL56



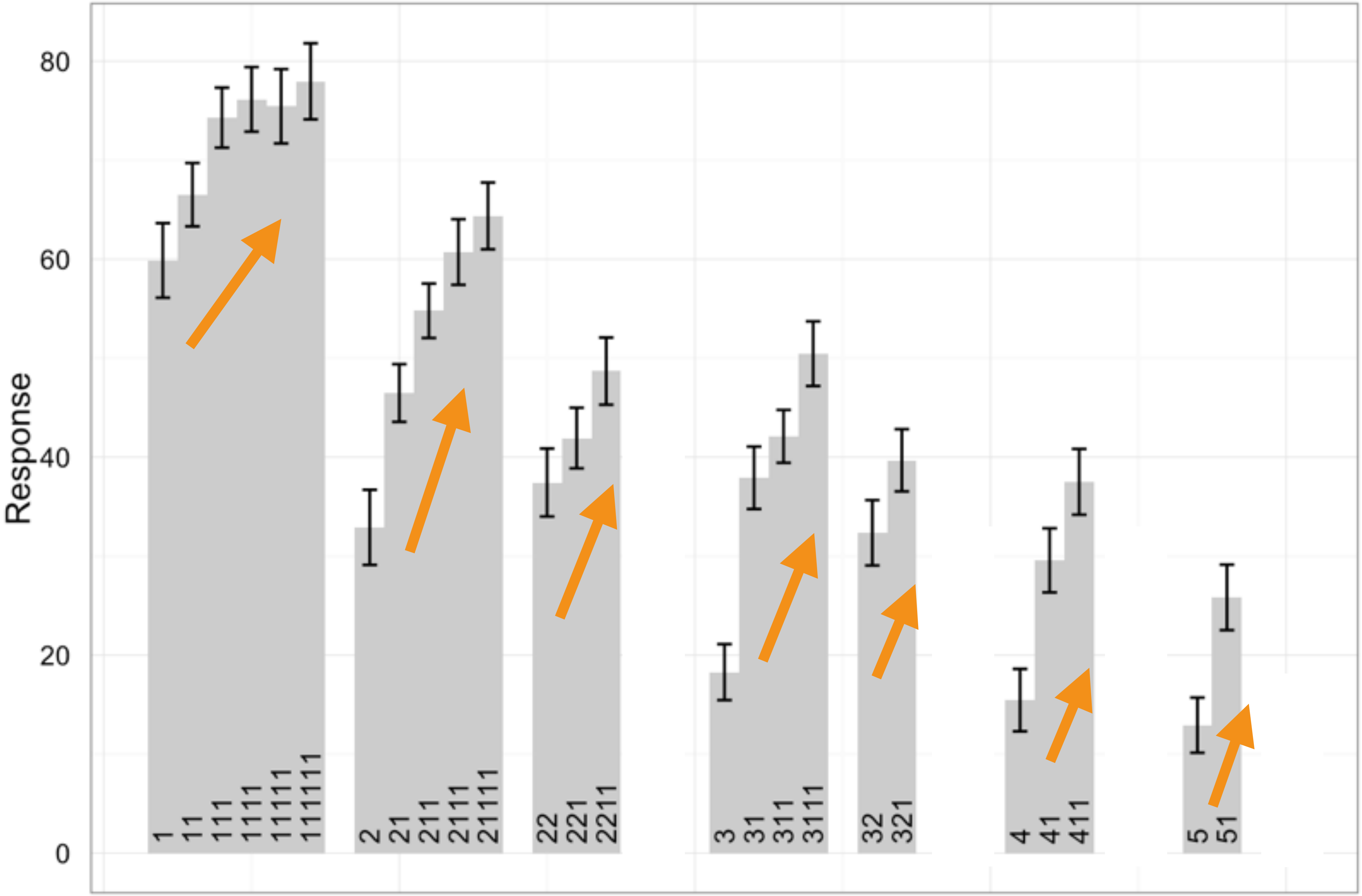
	$K=1$	$K=2$			$K=3$			$K=4$		$K=5$	$K=6$
$N=1$	1										
$N=2$	2	11									
$N=3$	3	21			111						
$N=4$	4	31	22		211	1111					
$N=5$	5	41	32		311	221		2111		11111	
$N=6$	6	51	42	33	411	321	222	3111	2211	21111	111111

Judge the probability that the next item will come from a new category, for every possible frequency table with 6 or fewer exemplars

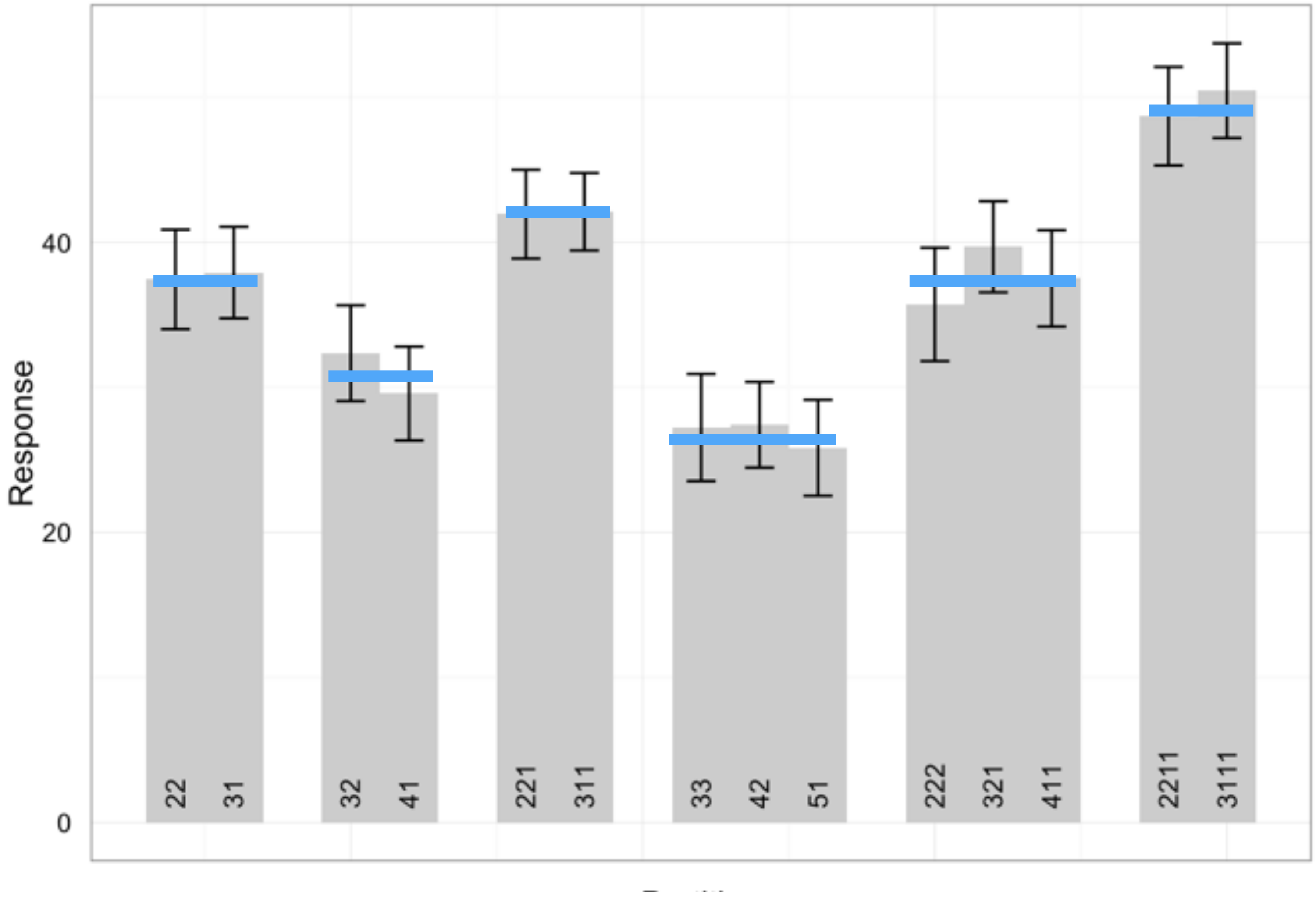
Familiar addition effect… ✓



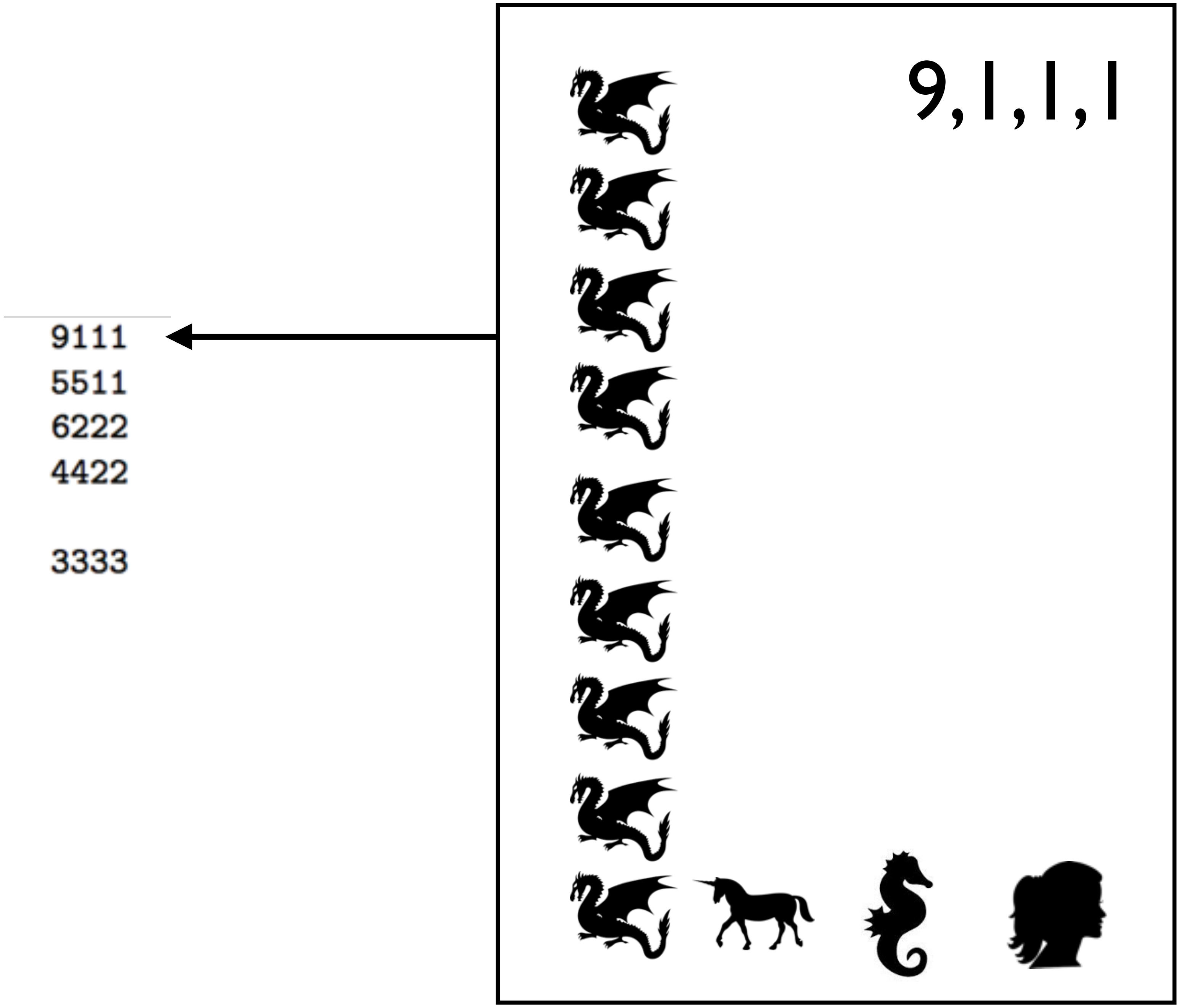
Novel addition effect… ✓



No transfer effect!



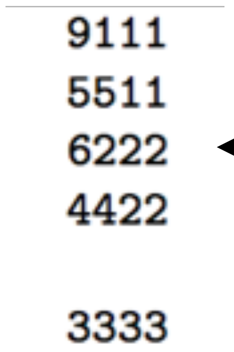
Experiment 2



9111
5511
6222
4422

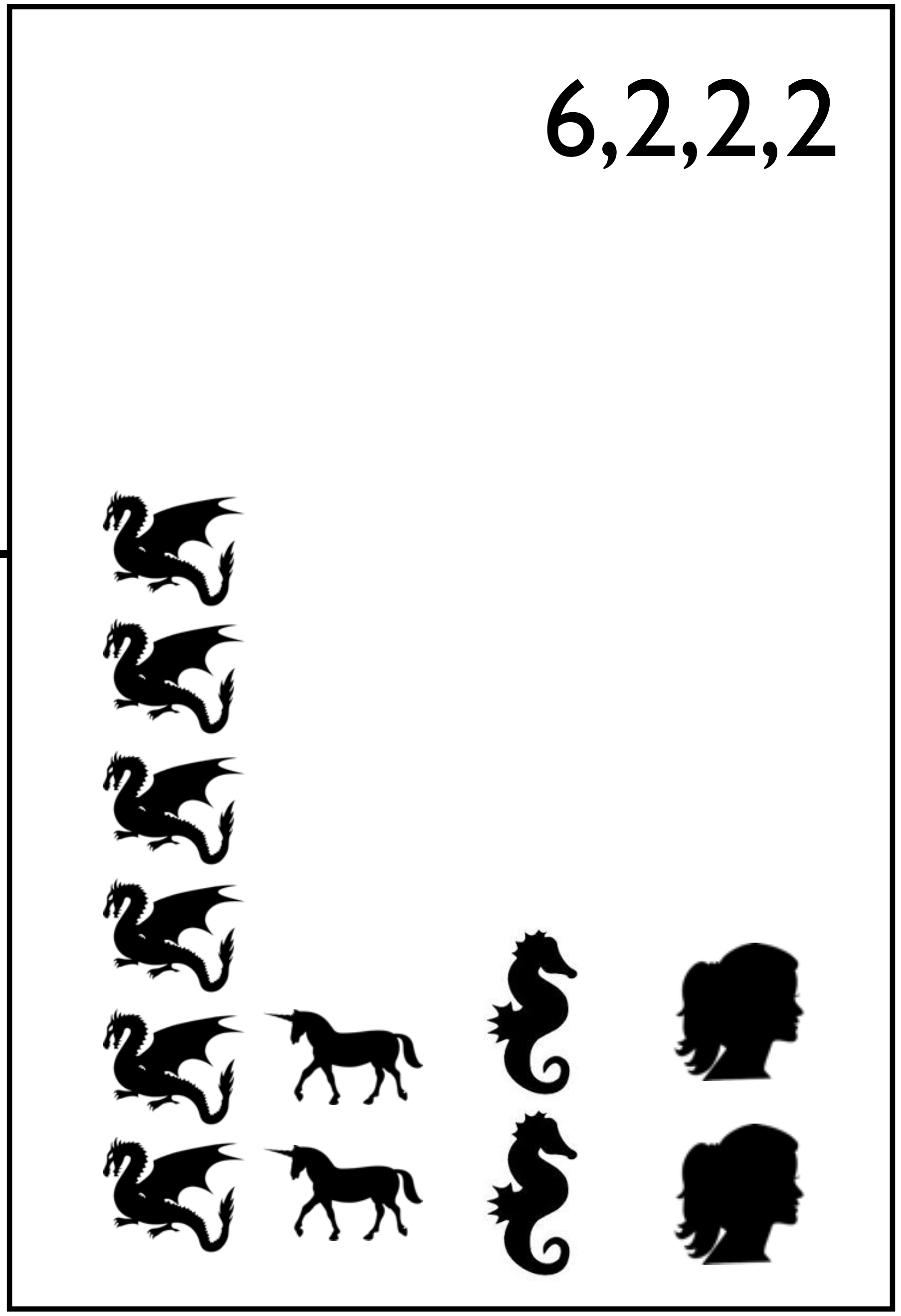
3333

Experiment 2



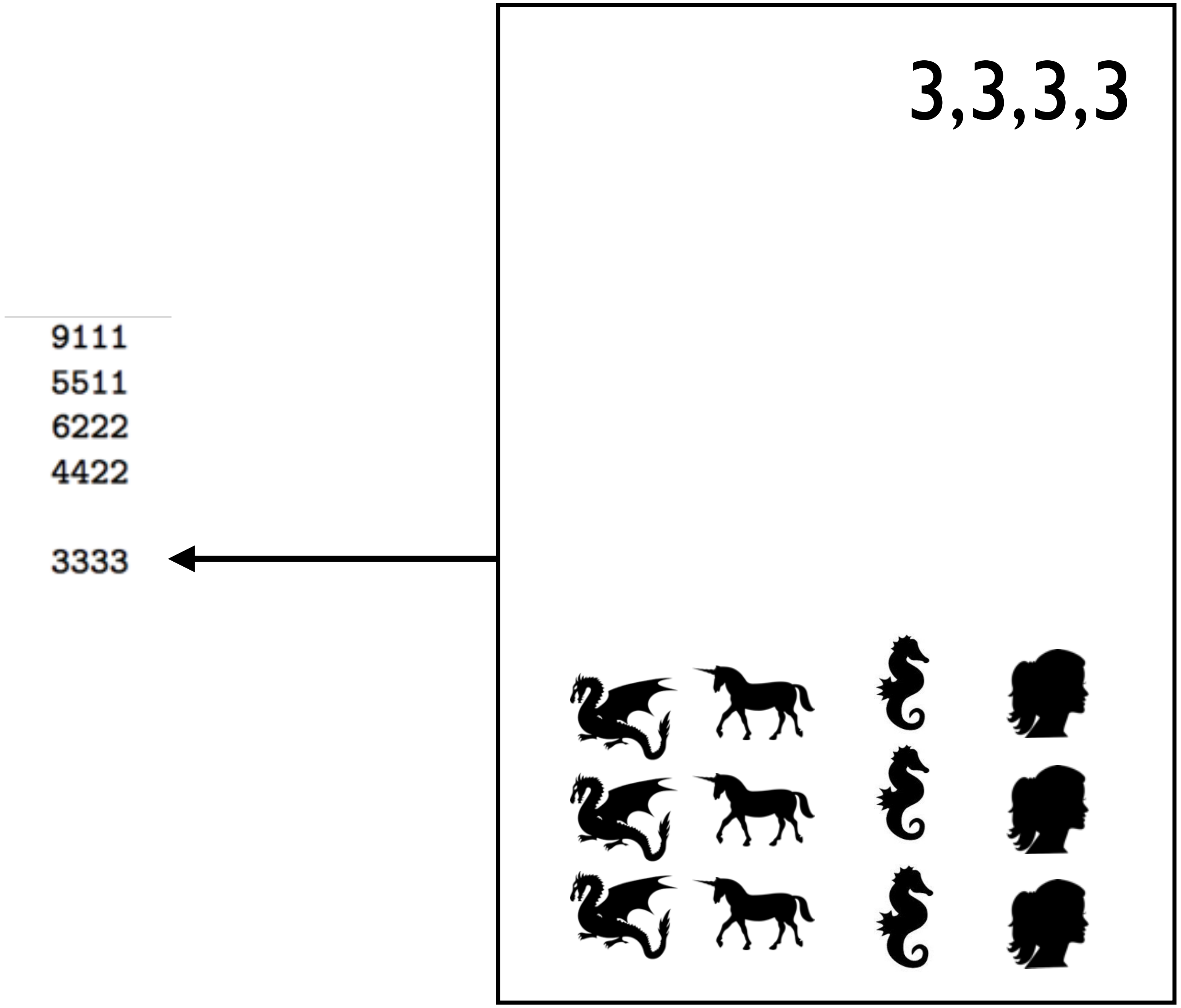
9111
5511
6222
4422

3333



6,2,2,2

Experiment 2

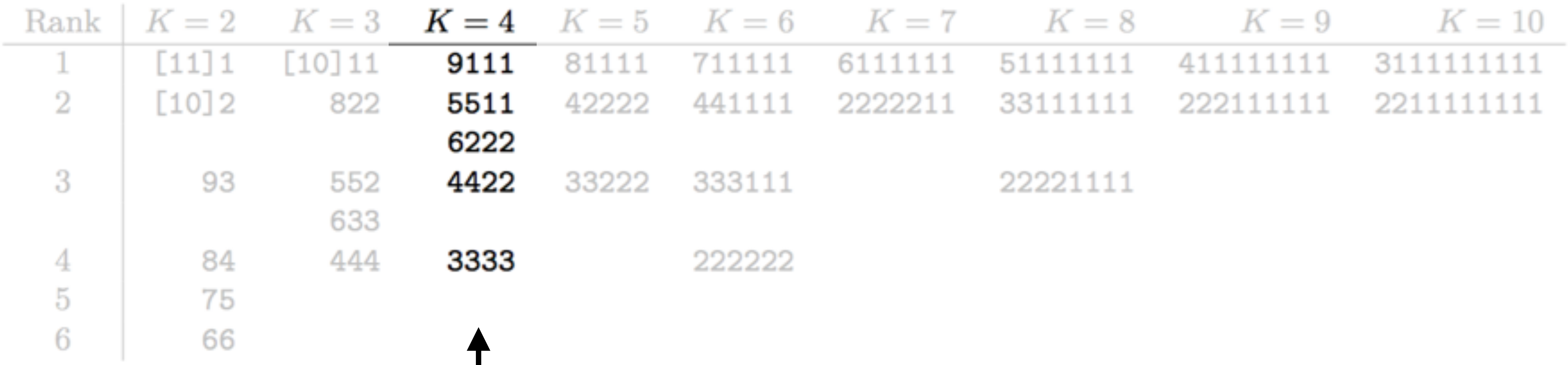


9111
5511
6222
4422

3333

3,3,3,3

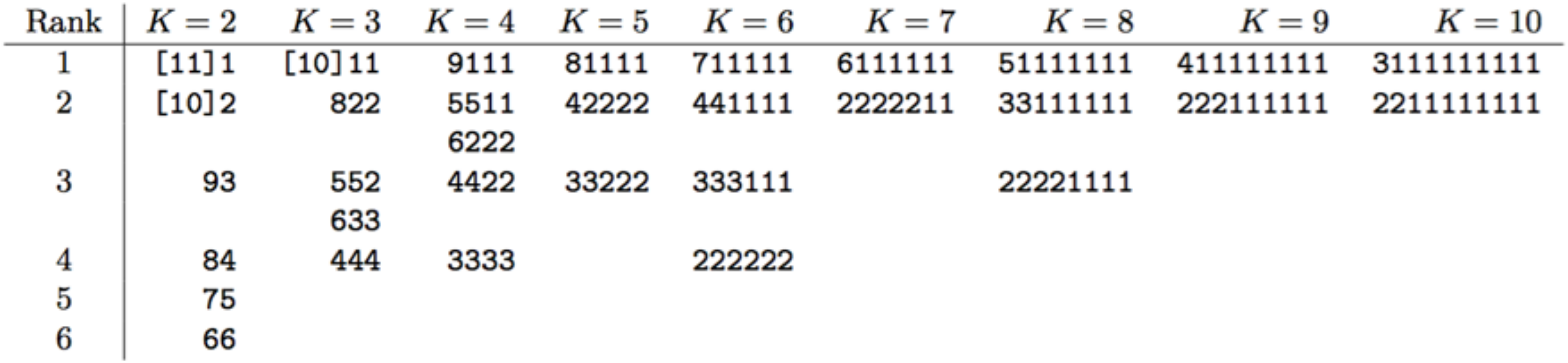
Experiment 2



Rank	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
1	[11]1	[10]11	**9111**	81111	711111	6111111	51111111	411111111	3111111111
2	[10]2	822	**5511** **6222**	42222	441111	2222211	33111111	222111111	2211111111
3	93	552 633	**4422**	33222	333111		22221111		
4	84	444	**3333**		222222				
5	75								
6	66								

12 objects in 4 categories

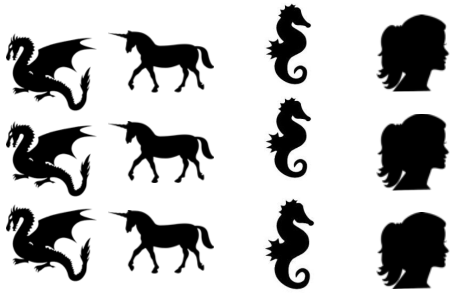
Experiment 2



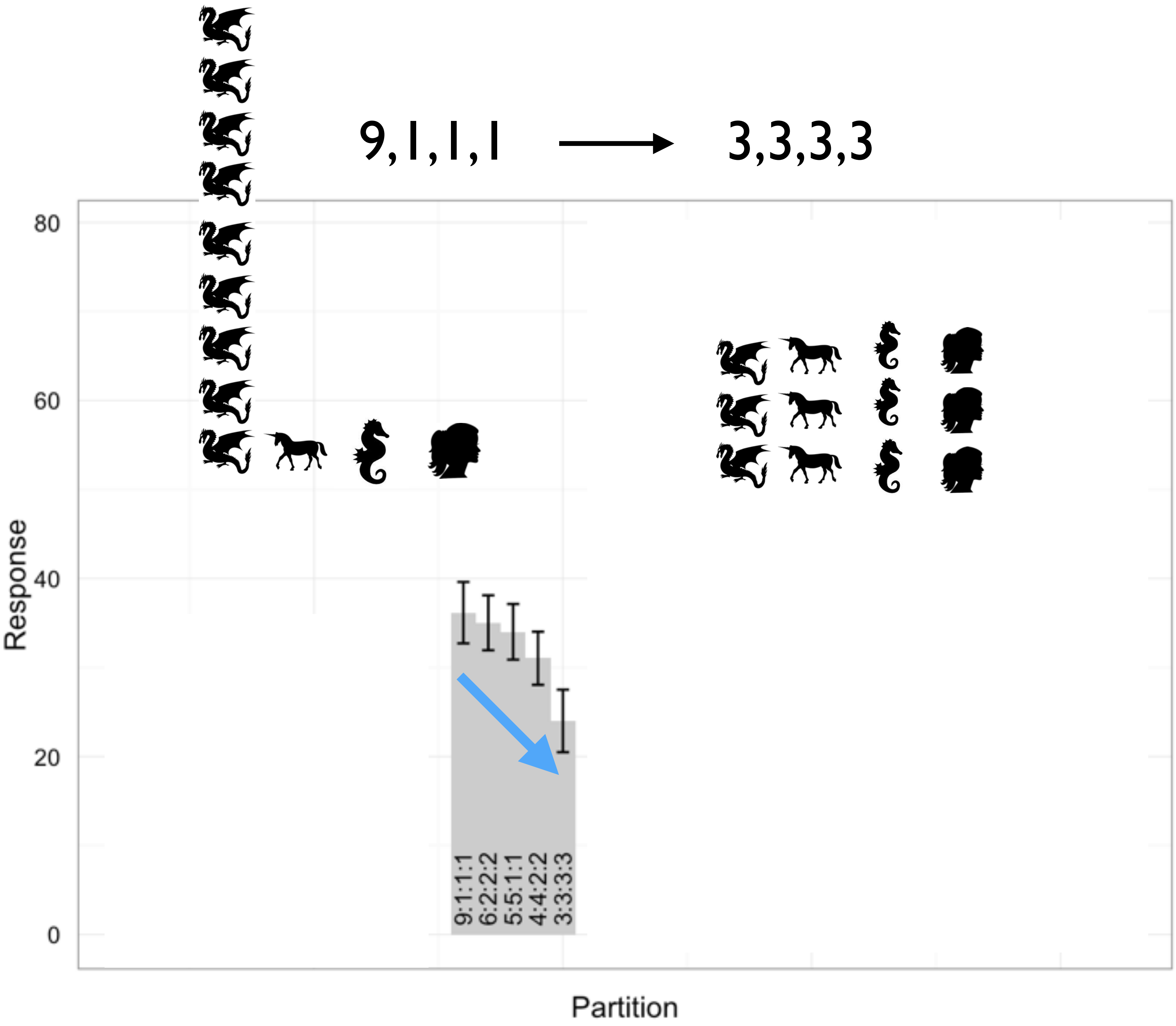
Rank	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
1	[11]1	[10]11	9111	81111	711111	6111111	51111111	411111111	3111111111
2	[10]2	822	5511 6222	42222	441111	2222211	33111111	222111111	2211111111
3	93	552 633	4422	33222	333111		22221111		
4	84	444	3333		222222				
5	75								
6	66								

Vary the number of categories from 2 to 10

9,1,1,1 → 3,3,3,3



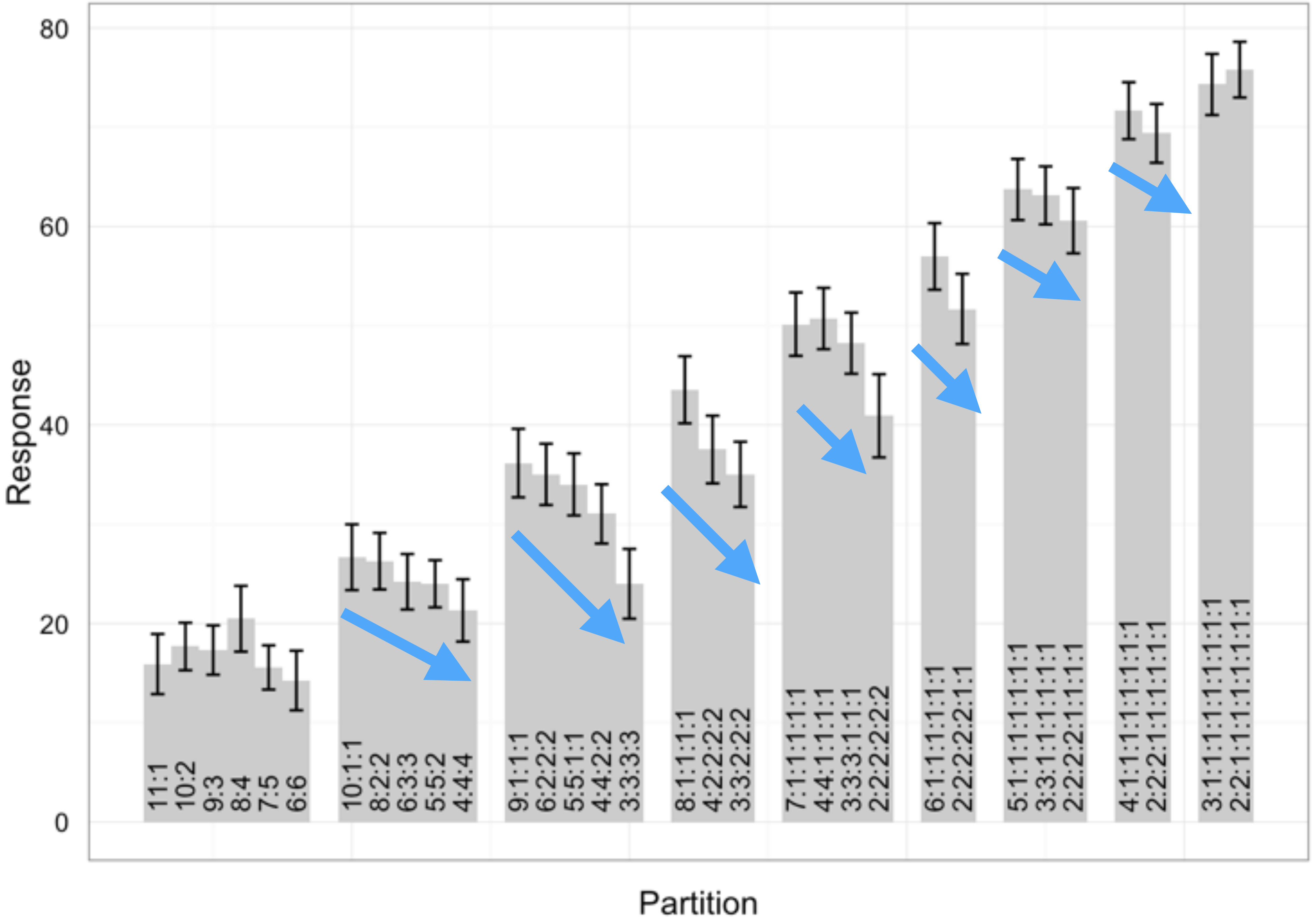
Does this transfer have an effect?



9,1,1,1 → 3,3,3,3

The transfer effect exists

(it's small, so you need bigger frequency tables)



Do these results pose a serious theoretical challenge to categorisation models?

A list of heuristic methods for estimating the probability that the next object will be novel

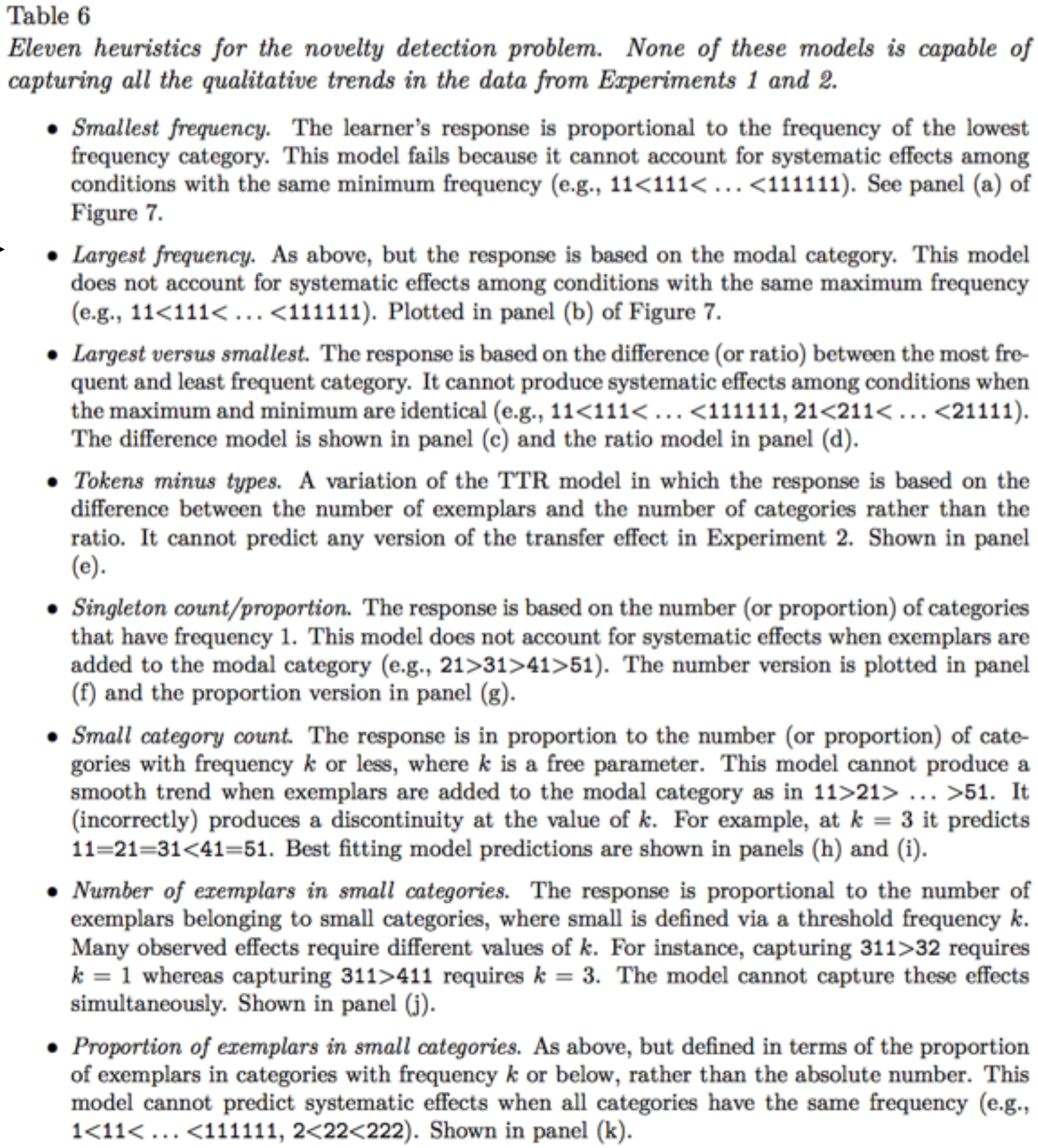
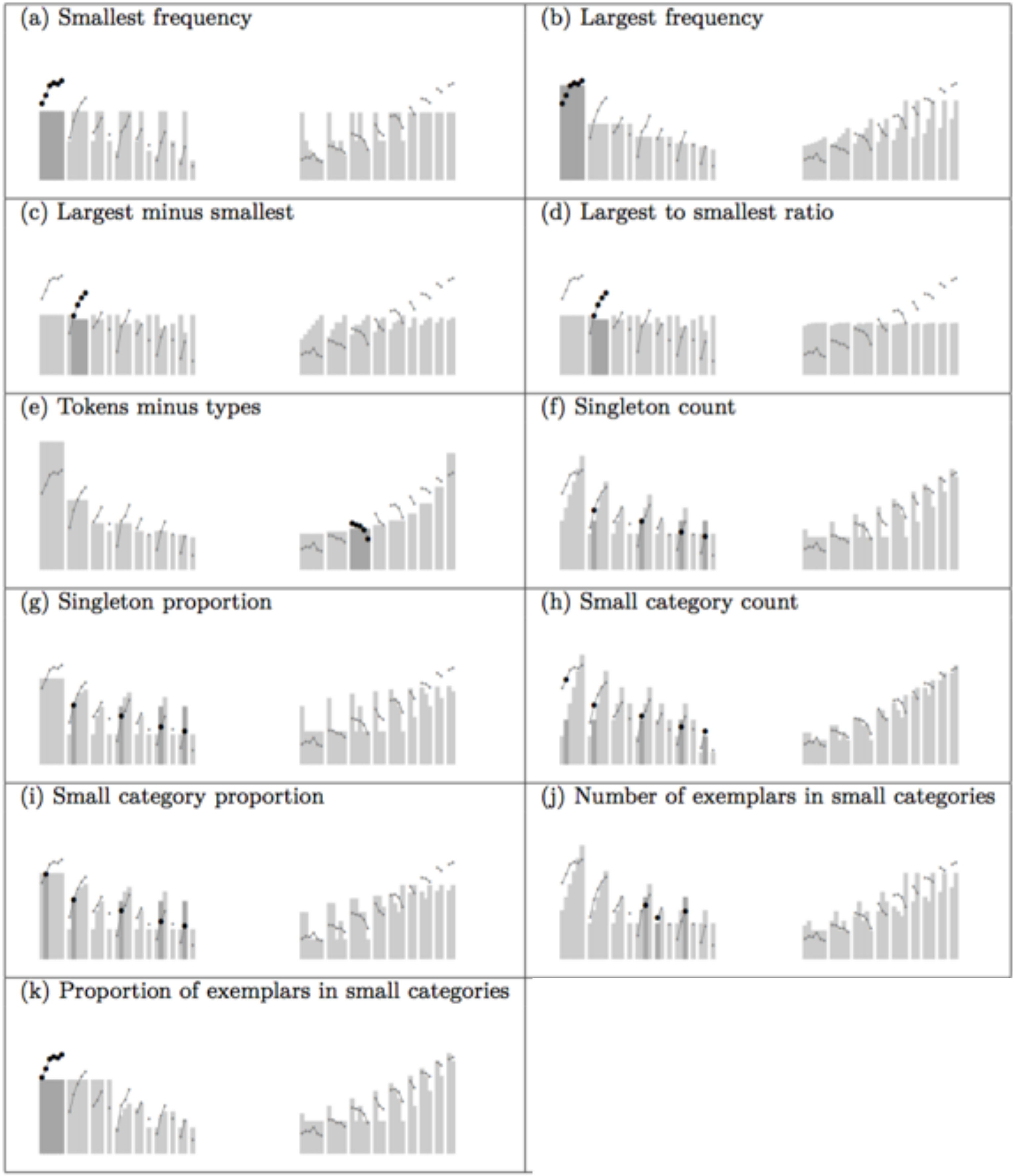


Table 6

Eleven heuristics for the novelty detection problem. None of these models is capable of capturing all the qualitative trends in the data from Experiments 1 and 2.

- *Smallest frequency.* The learner's response is proportional to the frequency of the lowest frequency category. This model fails because it cannot account for systematic effects among conditions with the same minimum frequency (e.g., 11<111< ... <111111). See panel (a) of Figure 7.
- *Largest frequency.* As above, but the response is based on the modal category. This model does not account for systematic effects among conditions with the same maximum frequency (e.g., 11<111< ... <111111). Plotted in panel (b) of Figure 7.
- *Largest versus smallest.* The response is based on the difference (or ratio) between the most frequent and least frequent category. It cannot produce systematic effects among conditions when the maximum and minimum are identical (e.g., 11<111< ... <111111, 21<211< ... <21111). The difference model is shown in panel (c) and the ratio model in panel (d).
- *Tokens minus types.* A variation of the TTR model in which the response is based on the difference between the number of exemplars and the number of categories rather than the ratio. It cannot predict any version of the transfer effect in Experiment 2. Shown in panel (e).
- *Singleton count/proportion.* The response is based on the number (or proportion) of categories that have frequency 1. This model does not account for systematic effects when exemplars are added to the modal category (e.g., 21>31>41>51). The number version is plotted in panel (f) and the proportion version in panel (g).
- *Small category count.* The response is in proportion to the number (or proportion) of categories with frequency k or less, where k is a free parameter. This model cannot produce a smooth trend when exemplars are added to the modal category as in 11>21> ... >51. It (incorrectly) produces a discontinuity at the value of k. For example, at $k = 3$ it predicts 11=21=31<41=51. Best fitting model predictions are shown in panels (h) and (i).
- *Number of exemplars in small categories.* The response is proportional to the number of exemplars belonging to small categories, where small is defined via a threshold frequency k. Many observed effects require different values of k. For instance, capturing 311>32 requires $k = 1$ whereas capturing 311>411 requires $k = 3$. The model cannot capture these effects simultaneously. Shown in panel (j).
- *Proportion of exemplars in small categories.* As above, but defined in terms of the proportion of exemplars in categories with frequency k or below, rather than the absolute number. This model cannot predict systematic effects when all categories have the same frequency (e.g., 1<11< ... <111111, 2<22<222). Shown in panel (k).

They don’t work ⟶



(a) Smallest frequency

(b) Largest frequency

(c) Largest minus smallest

(d) Largest to smallest ratio

(e) Tokens minus types

(f) Singleton count

(g) Singleton proportion

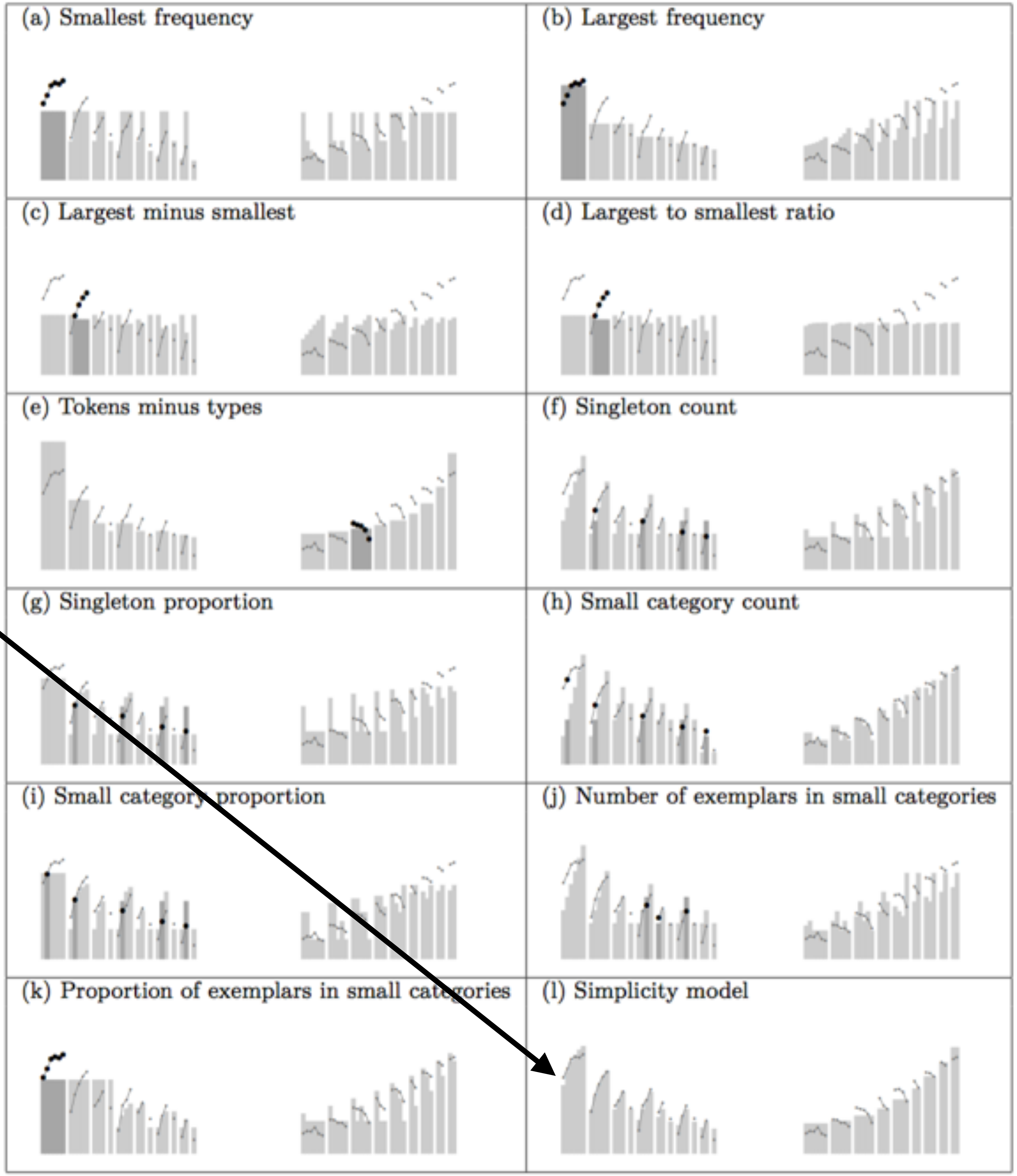
(h) Small category count

(i) Small category proportion

(j) Number of exemplars in small categories

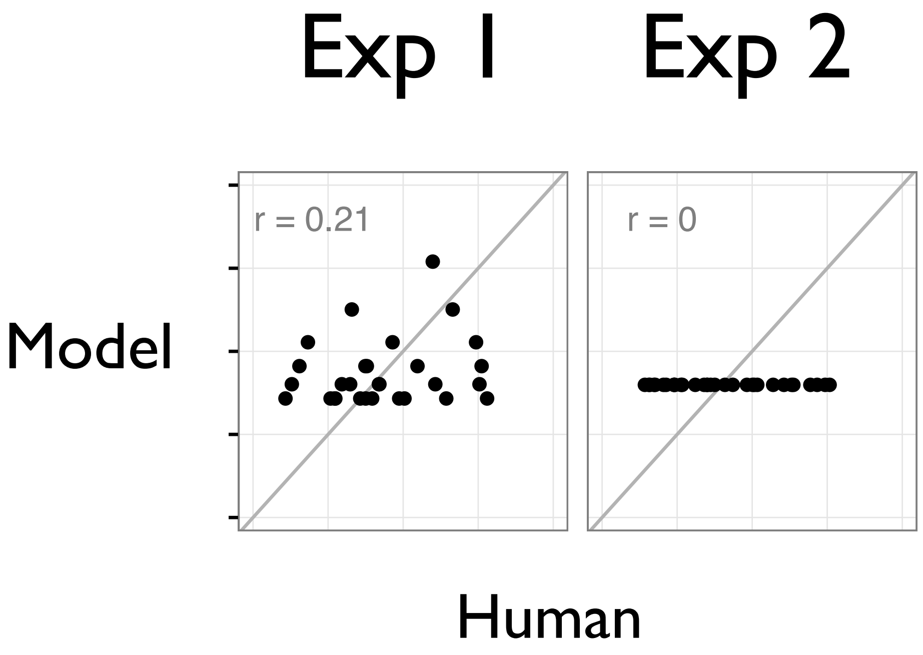
(k) Proportion of exemplars in small categories

Most existing category learning models (SUSTAIN, simplicity, etc) also fail



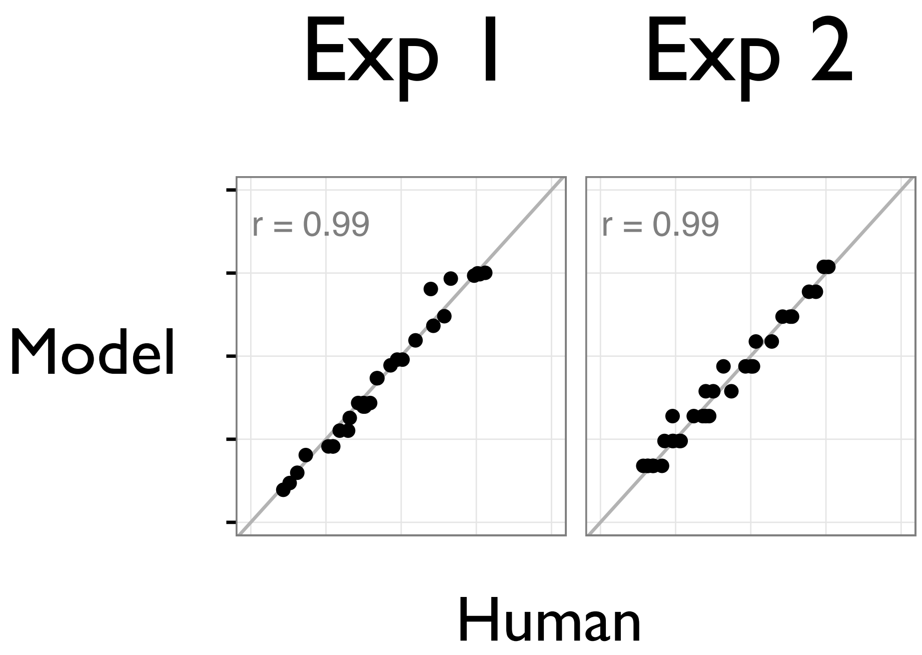
What about the Bayesian models?

Despite being near-universal among Bayesian models of categorisation, the CRP is terrible



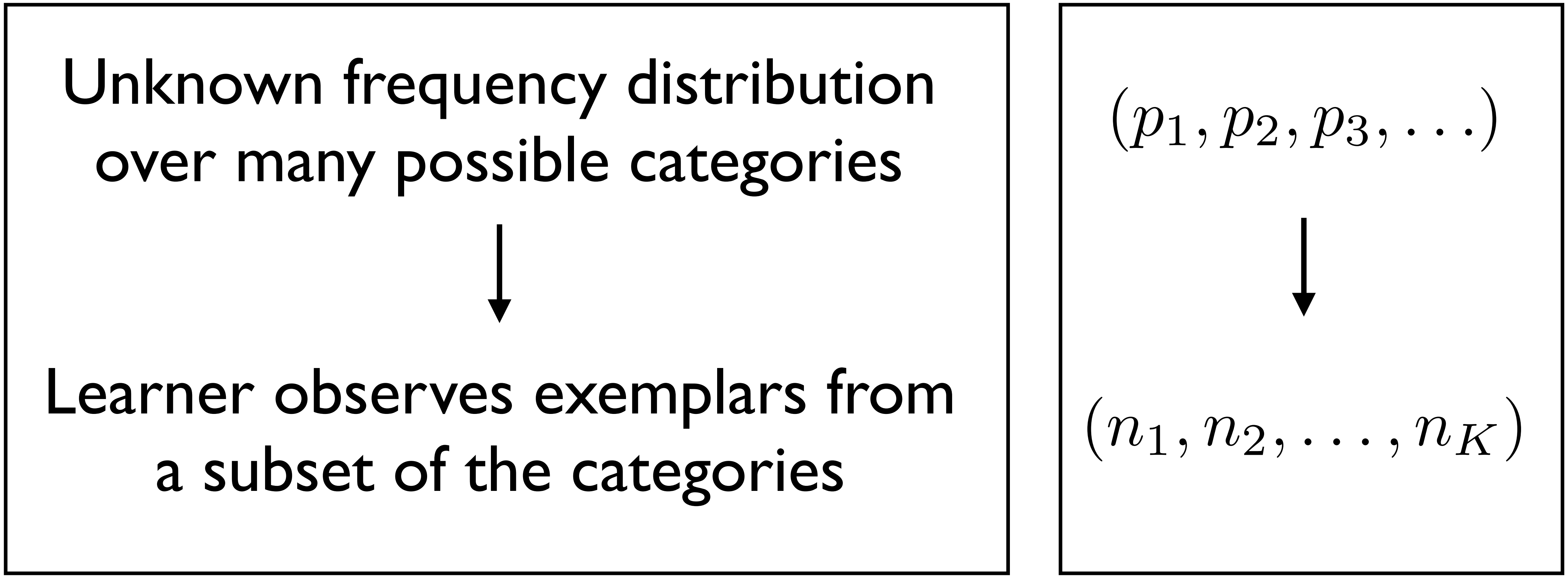
Good quantitative fit?	Predicts transfer effect?
✗	✗

The generalised CRP does better, but misses the transfer effect



Good quantitative fit?	Predicts transfer effect?
✓	✗

Generalised CRP



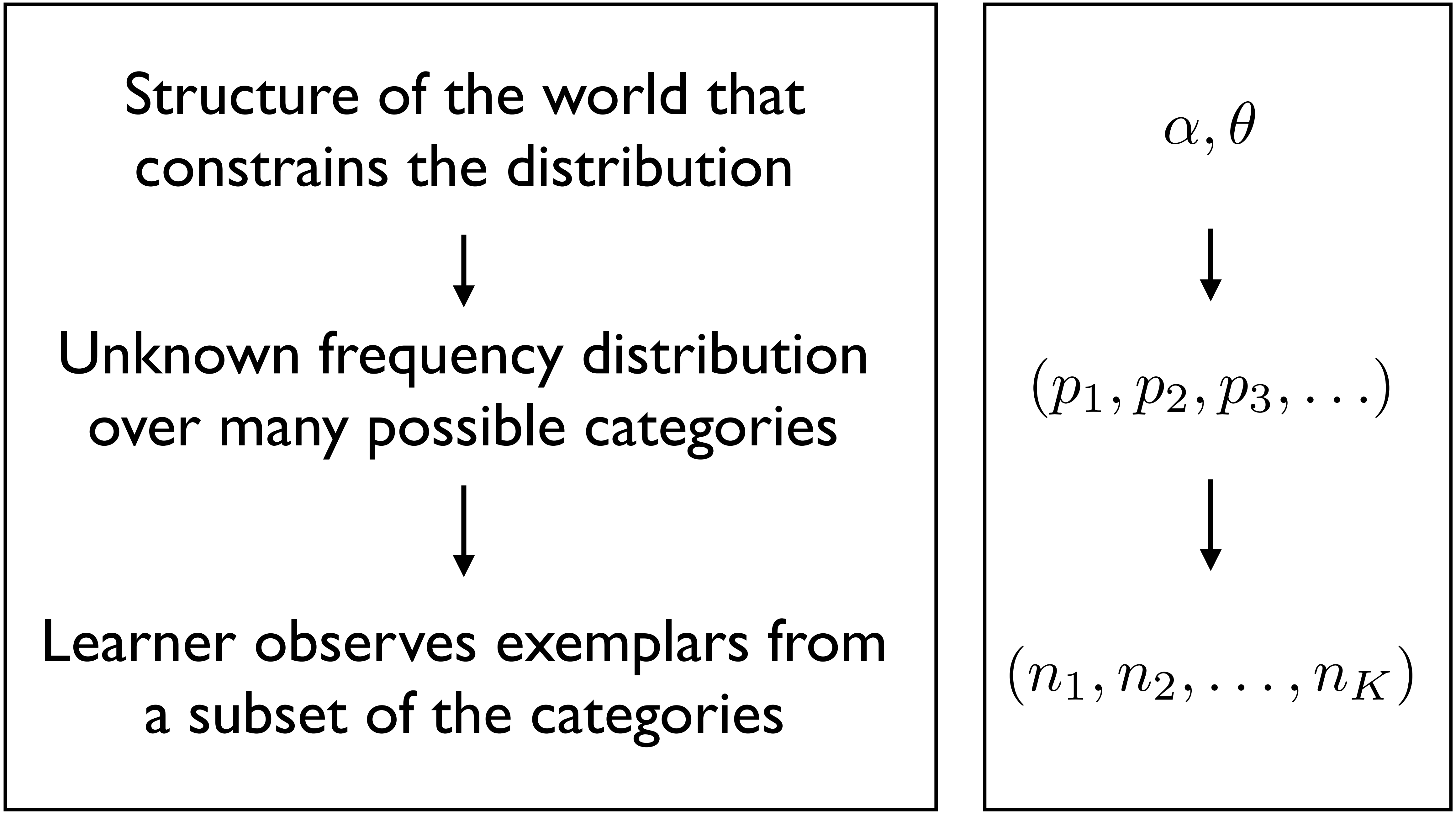
Unknown frequency distribution over many possible categories

Learner observes exemplars from a subset of the categories

$(p_1, p_2, p_3, \ldots)$

$(n_1, n_2, \ldots, n_K)$

Hierarchical generalised CRP



Structure of the world that constrains the distribution

↓

Unknown frequency distribution over many possible categories

↓

Learner observes exemplars from a subset of the categories

α, θ

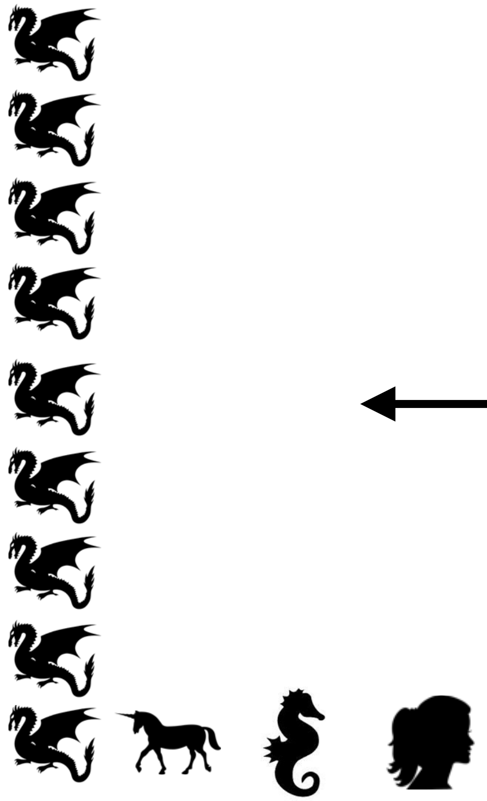
↓

$(p_1, p_2, p_3, \ldots)$

↓

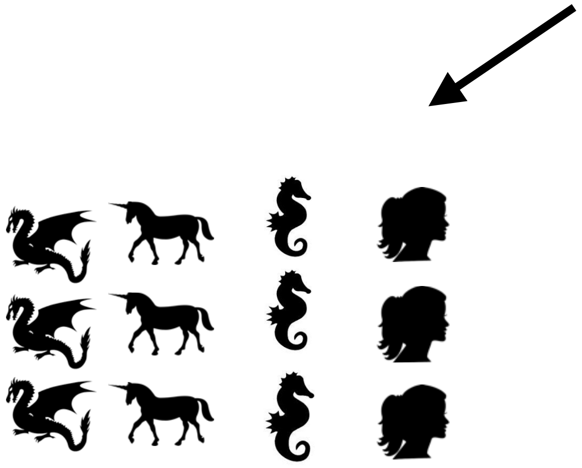
$(n_1, n_2, \ldots, n_K)$

Structure of the world that constrains the distribution



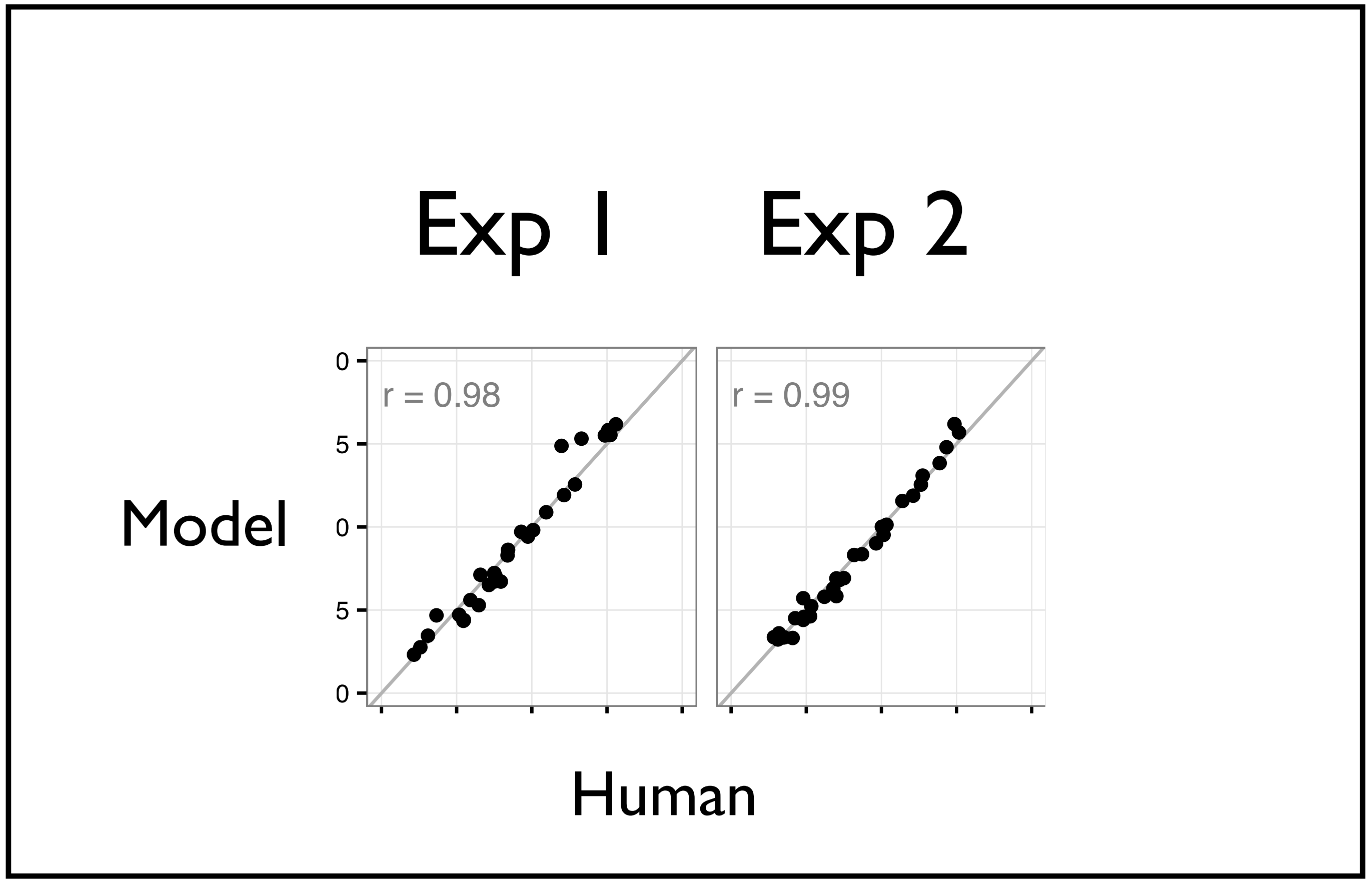
The HG-CRP model learns that this is a world with many low-frequency categories (infers a high α) and expects to see even more low-frequency categories

Structure of the world that constrains the distribution



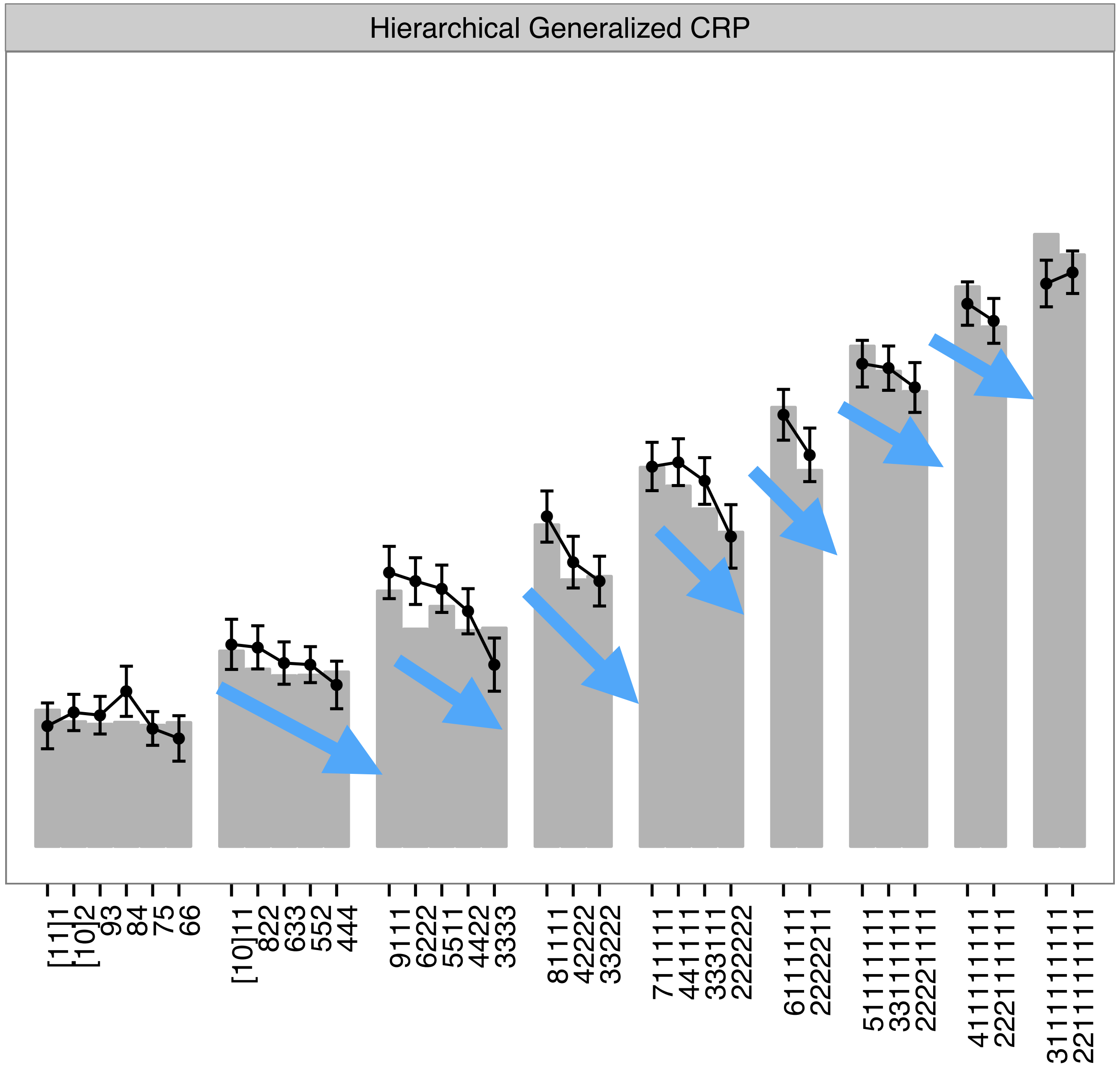
The HG-CRP model learns that this is a world with very few low-frequency categories (infers a lowα) and does not expect to see more LF categories

HG-CRP provides a good quantitative fit



✓

It also captures the transfer effect



✓

Is this actually a categorisation problem?

(a.k.a. Do people still do this in a standard task when similarity information exists?)

In most categorisation tasks we have similarity information

Dragon

Unicorn

Unicorn

Dragon

Dragon

Unicorn

Unicorn

Dragon

High similarity target is *less* likely to be novel

???

Dragon

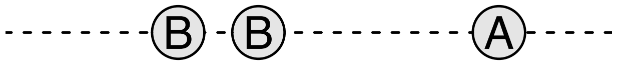
Unicorn

Unicorn

Dragon

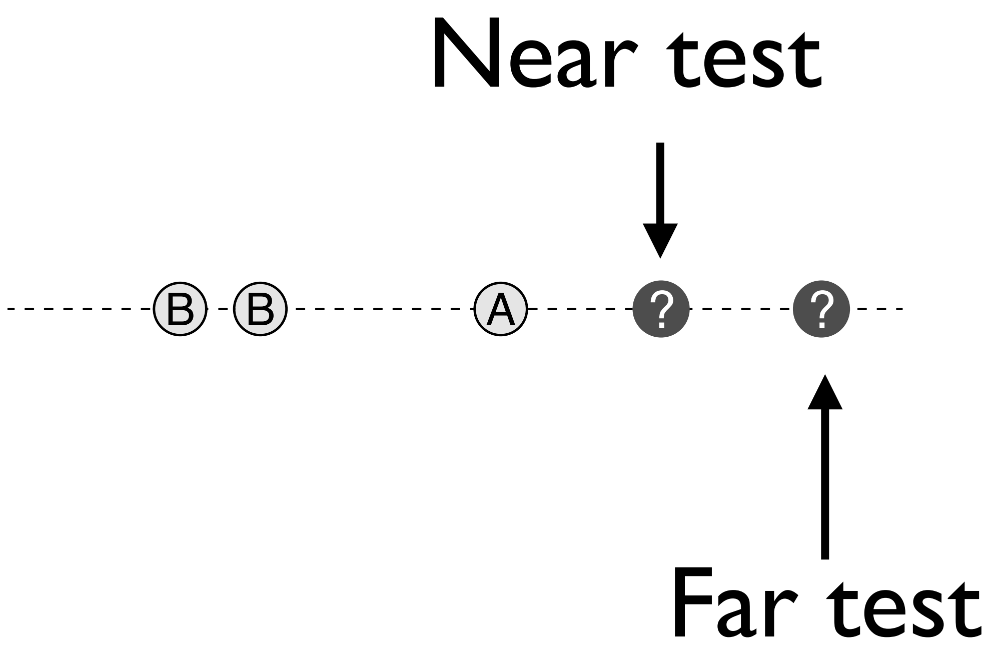
Low similarity target is *more* likely to be novel

???

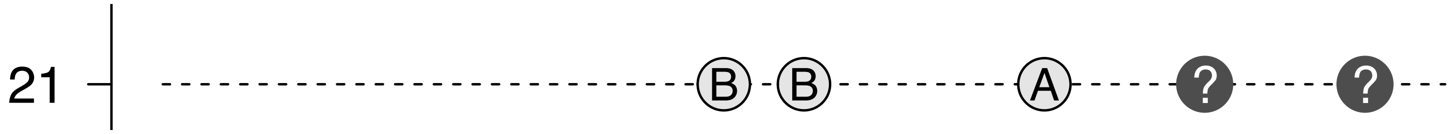


Training items vary on a single continuous stimulus dimension

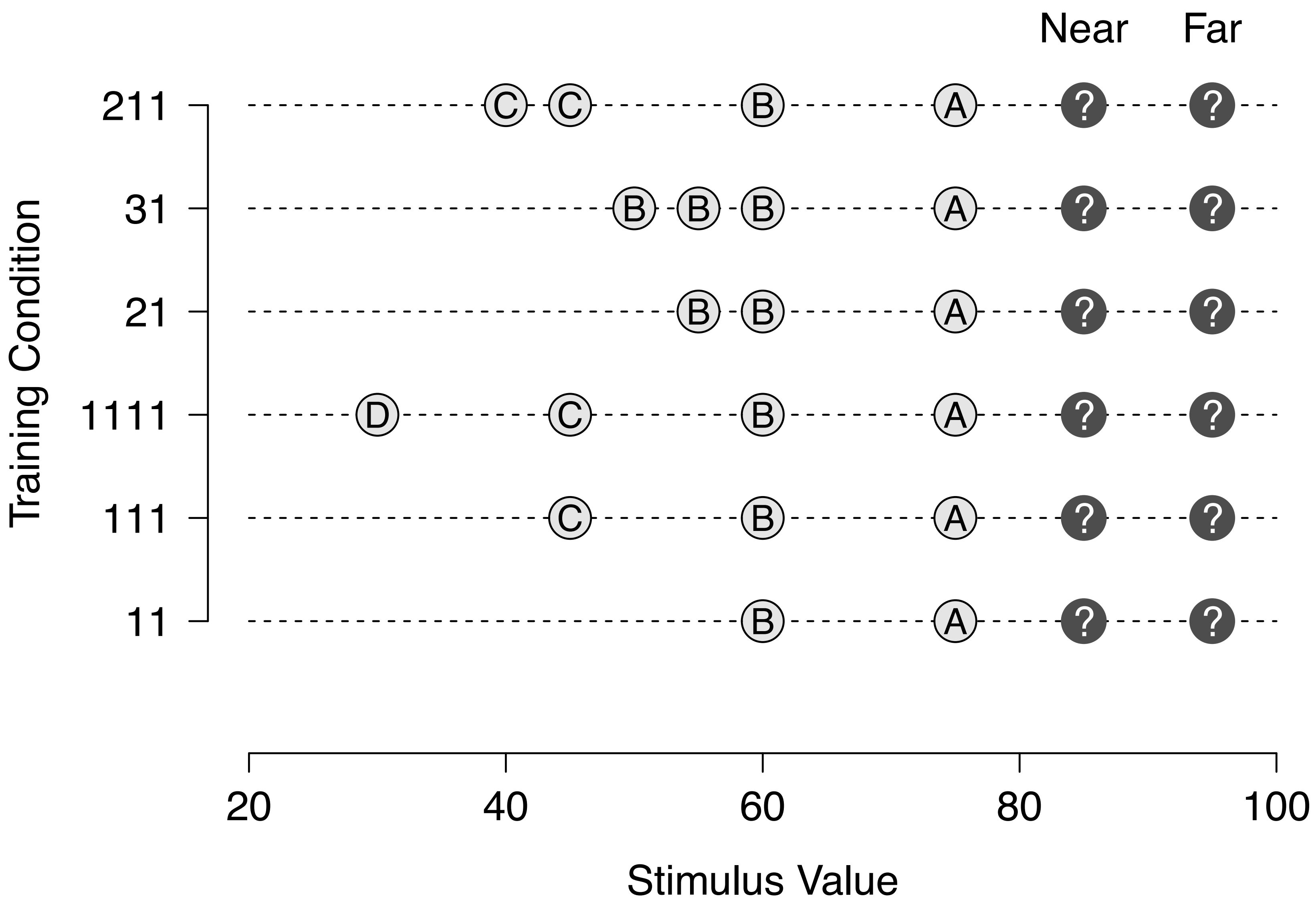
Similarity manipulation:



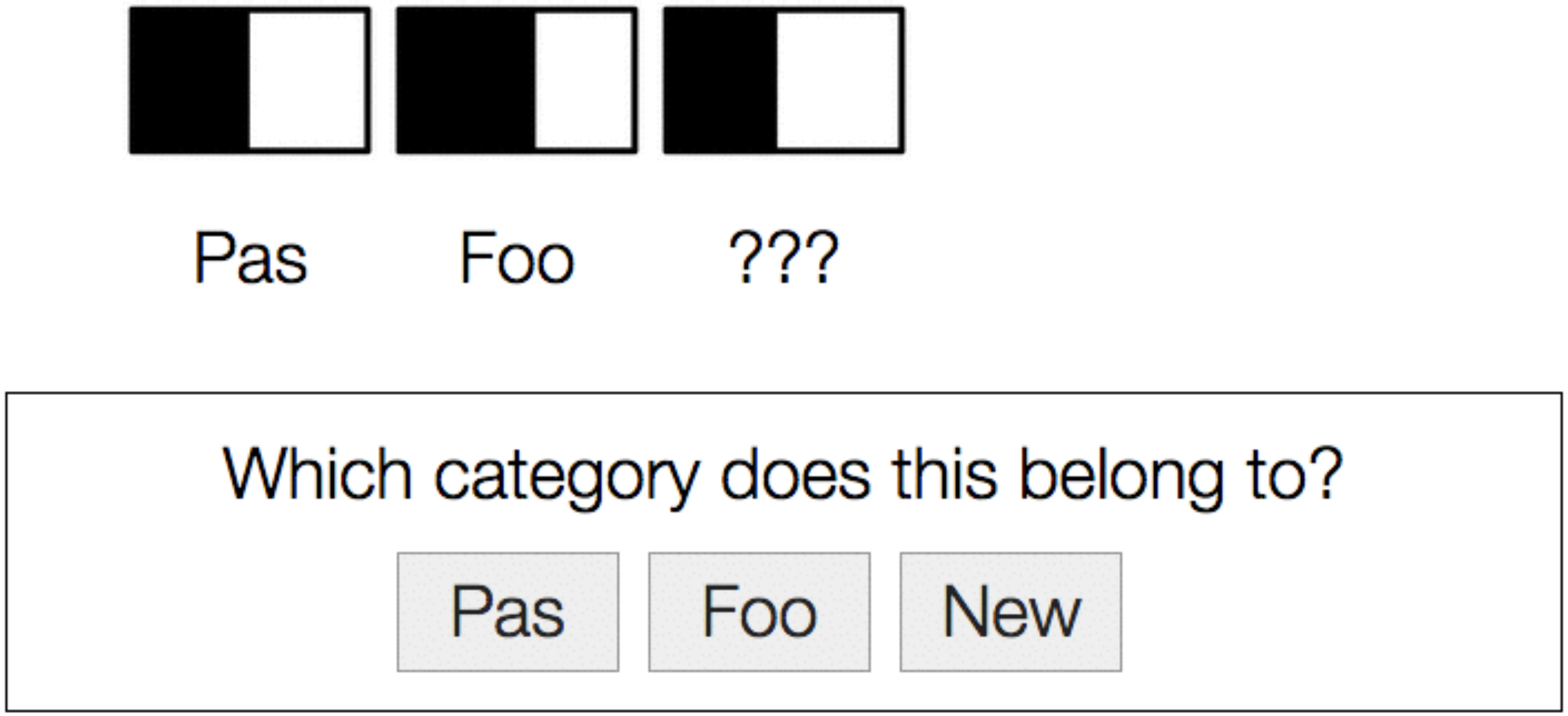
The training items form a 2,1 frequency table



6 tables x 2 tests = 12 categorisation tasks

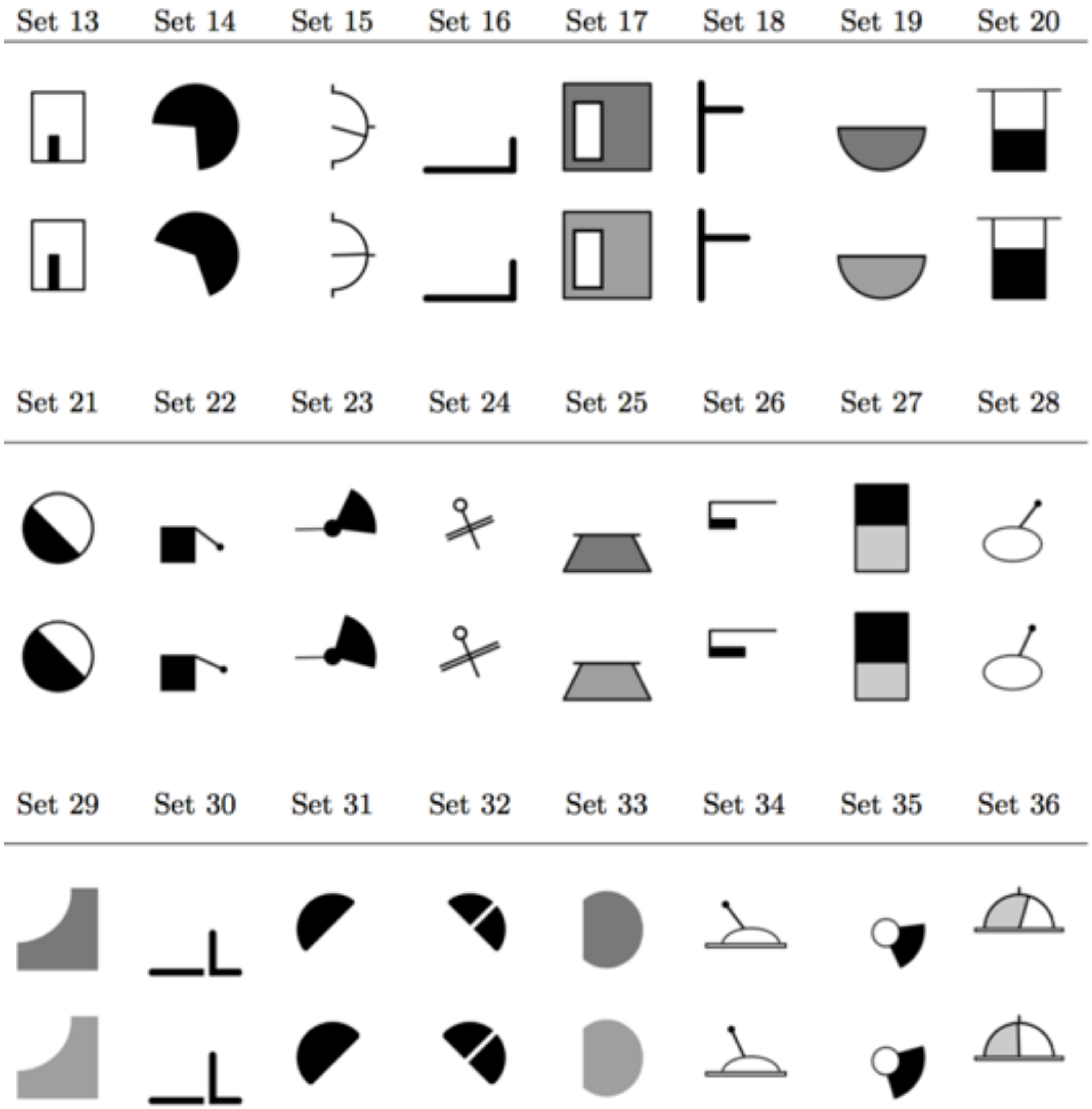


Experiment 3

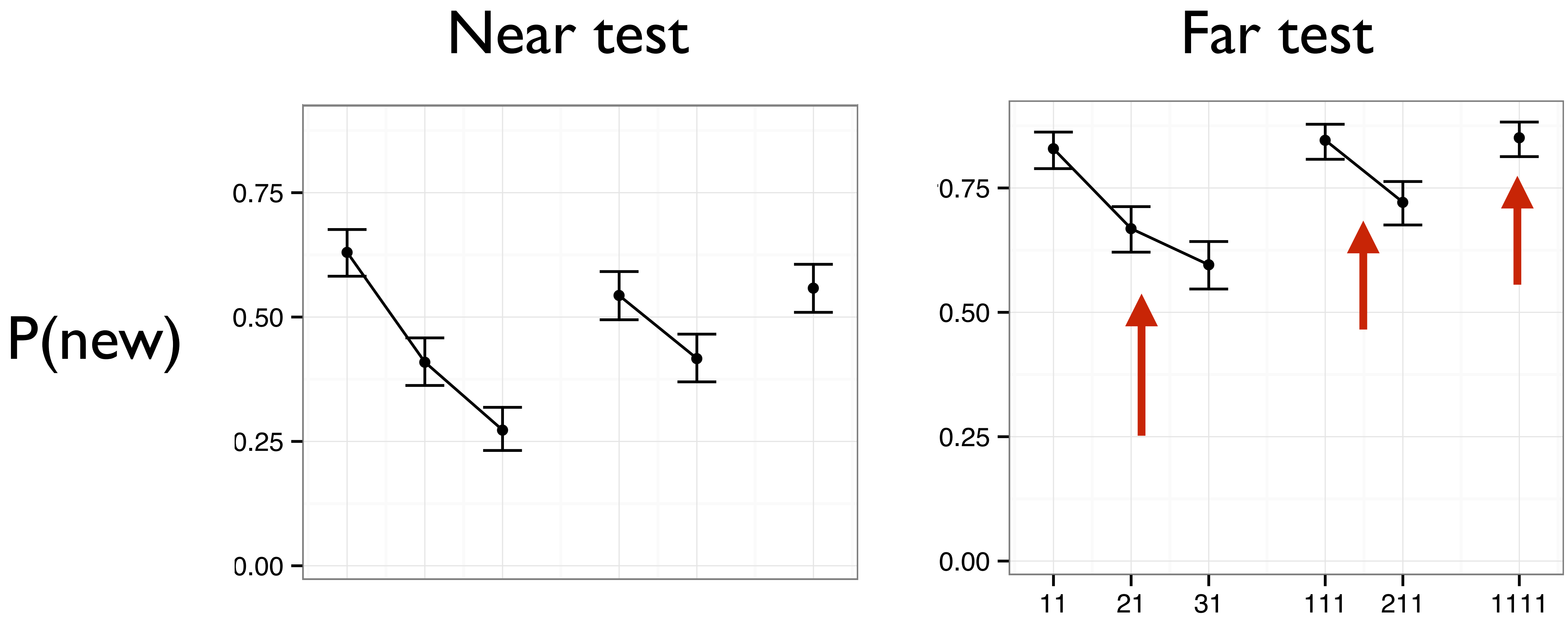


Which category does this belong to?

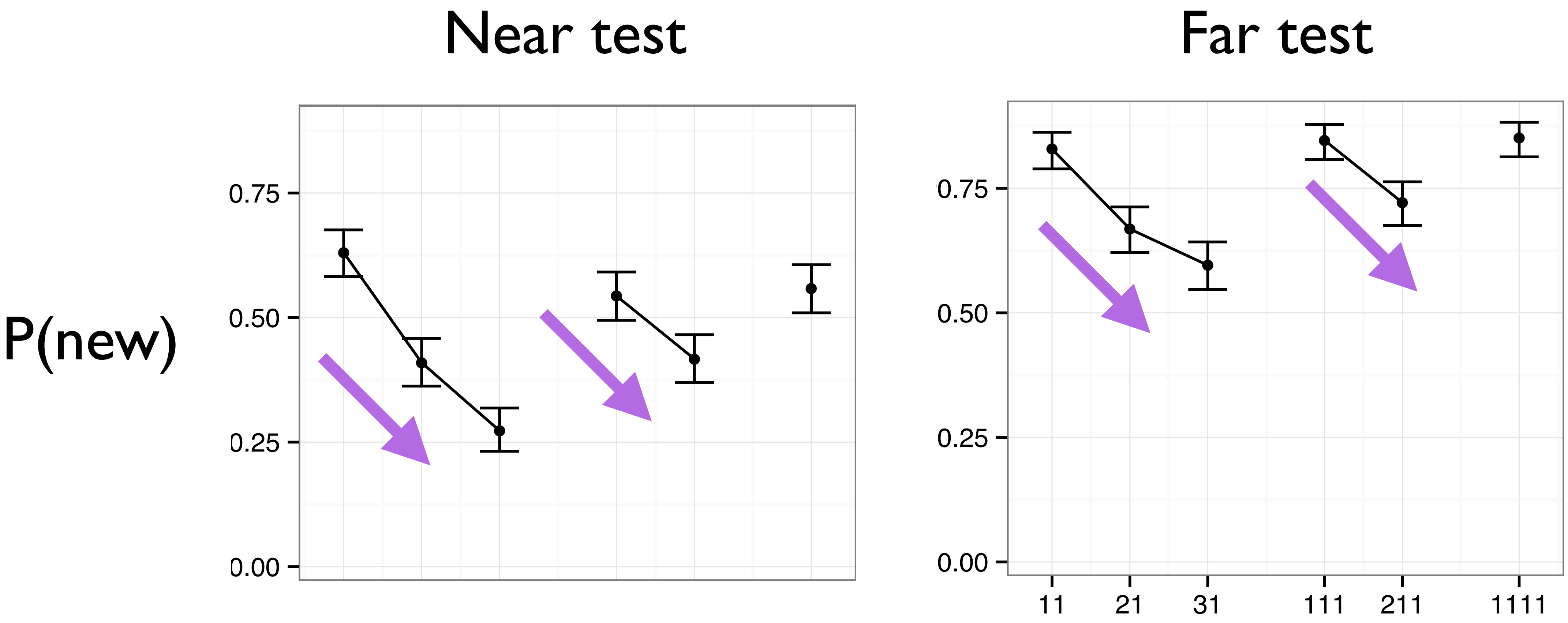
Lots of stimulus sets used:



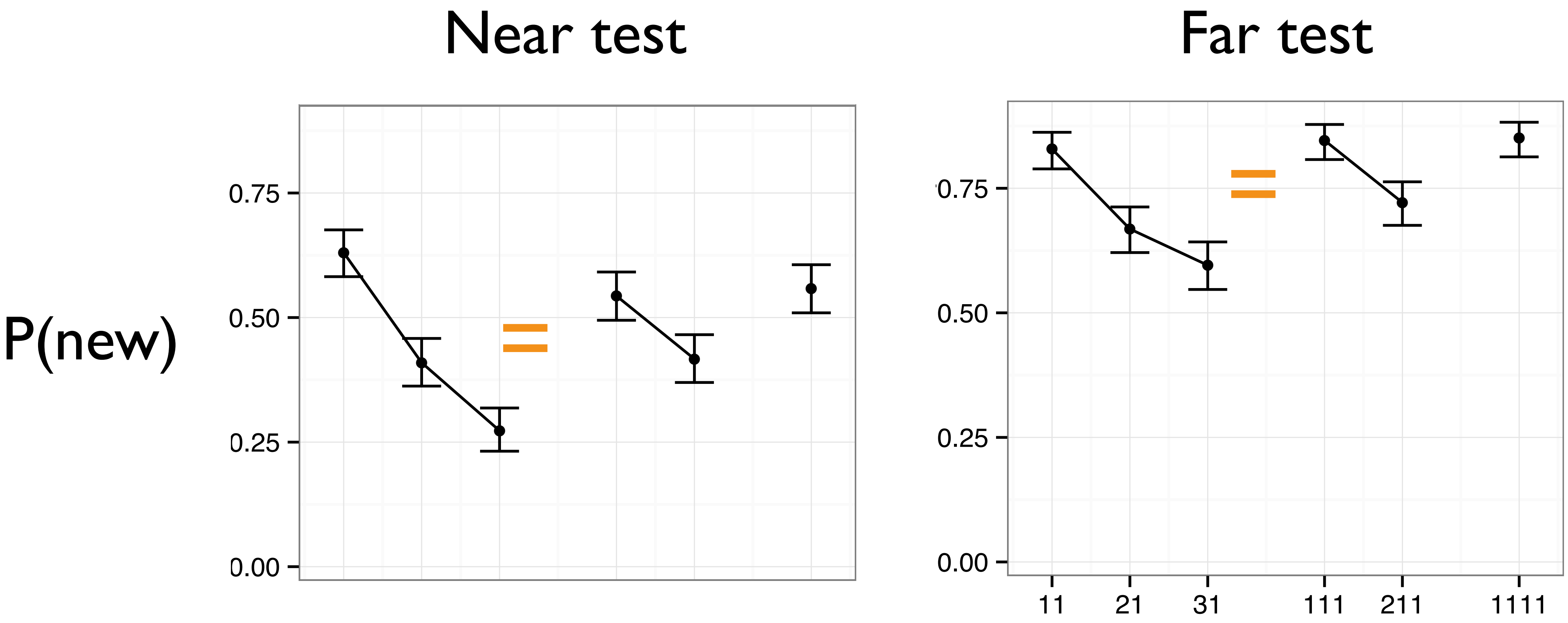
Similarity effect



Familiar addition



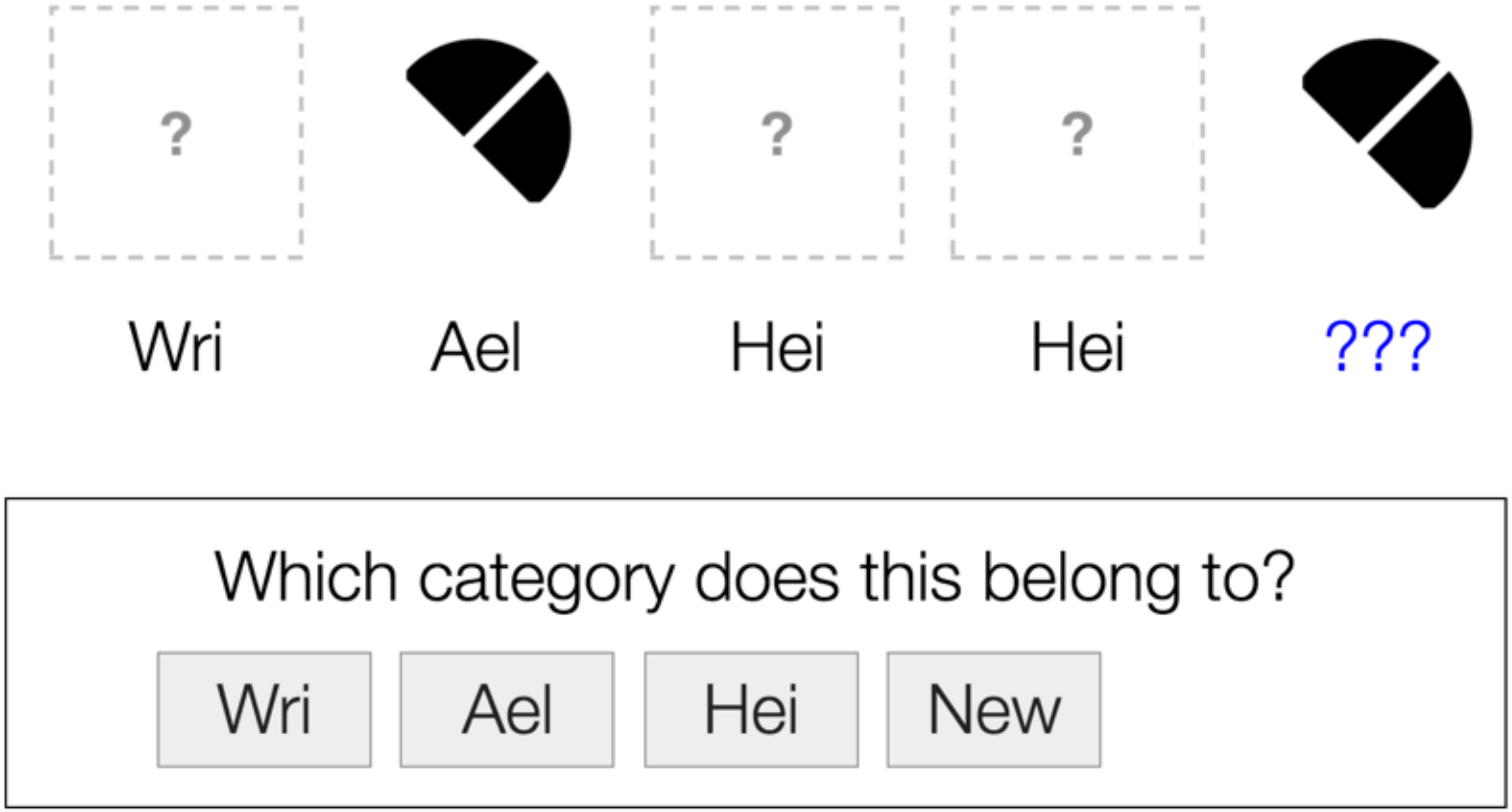
Novel addition?



Near test

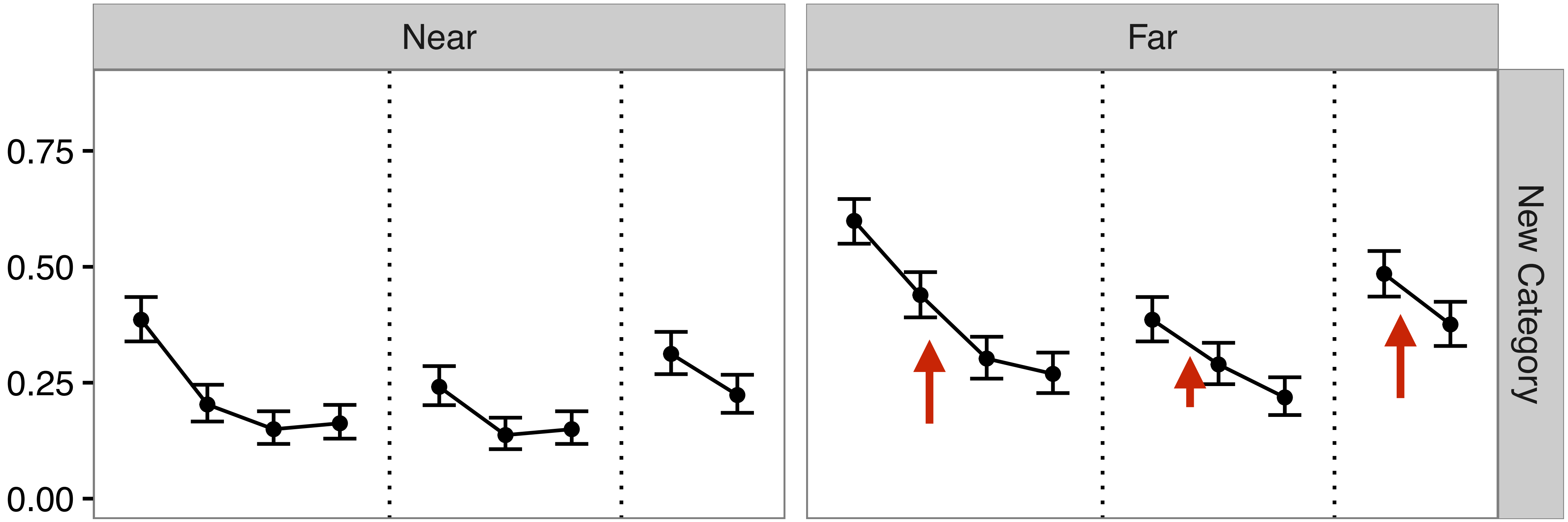
Far test

Experiment 4

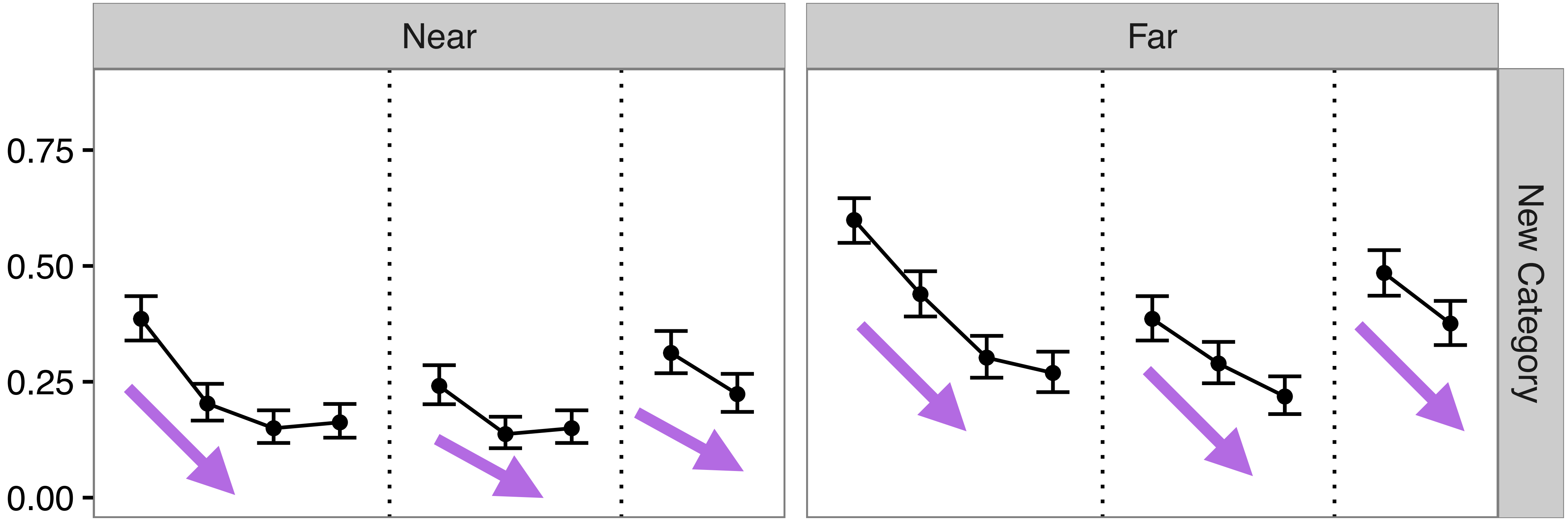


Which category does this belong to?

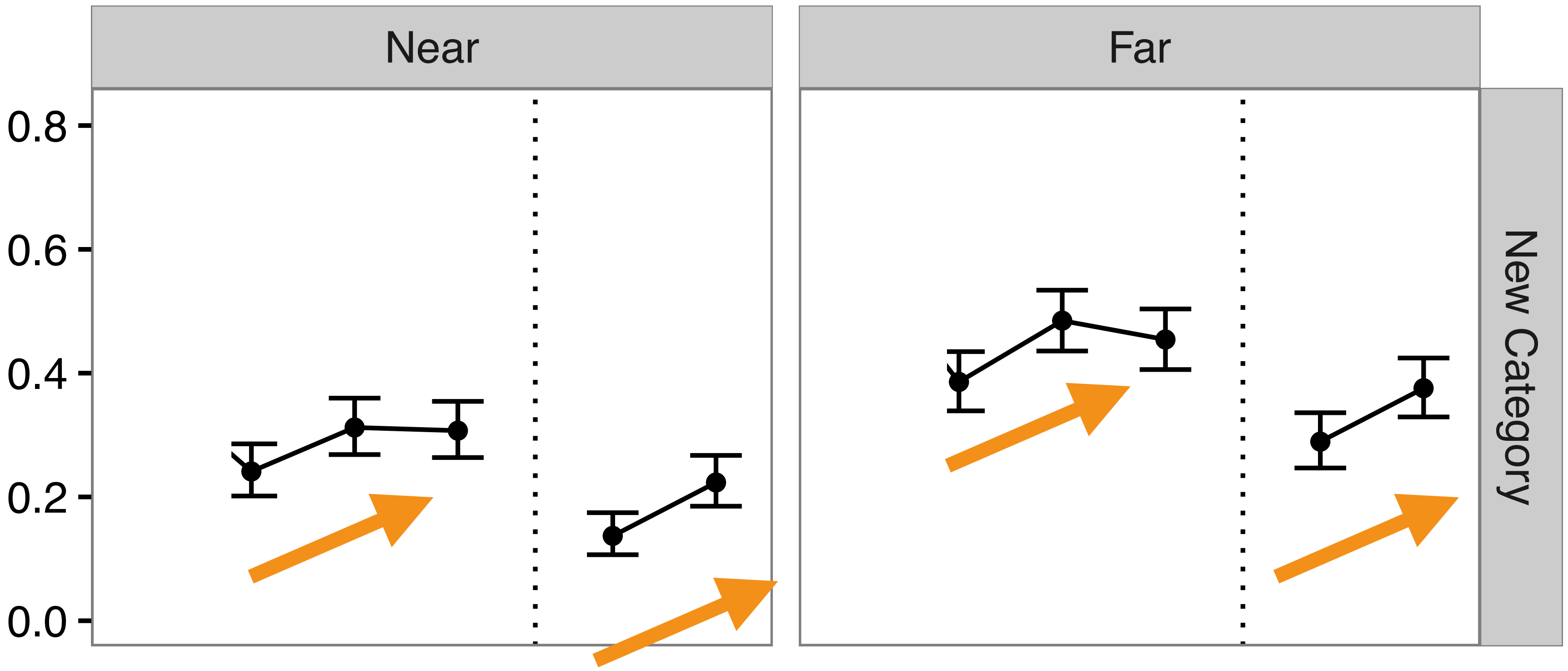
Similarity effect



Familiar addition



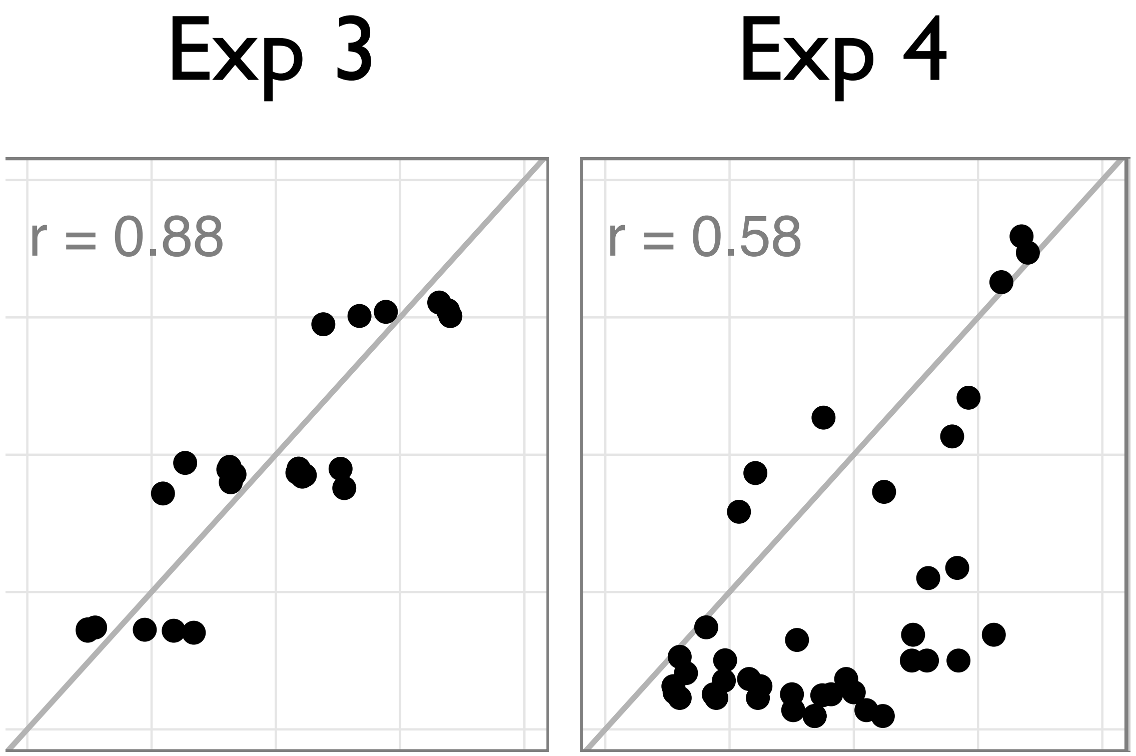
Novel addition*



* I'm hiding the [1] condition (ask me why!)

Parameters estimated from
Exp 3 and fixed for Exp 4

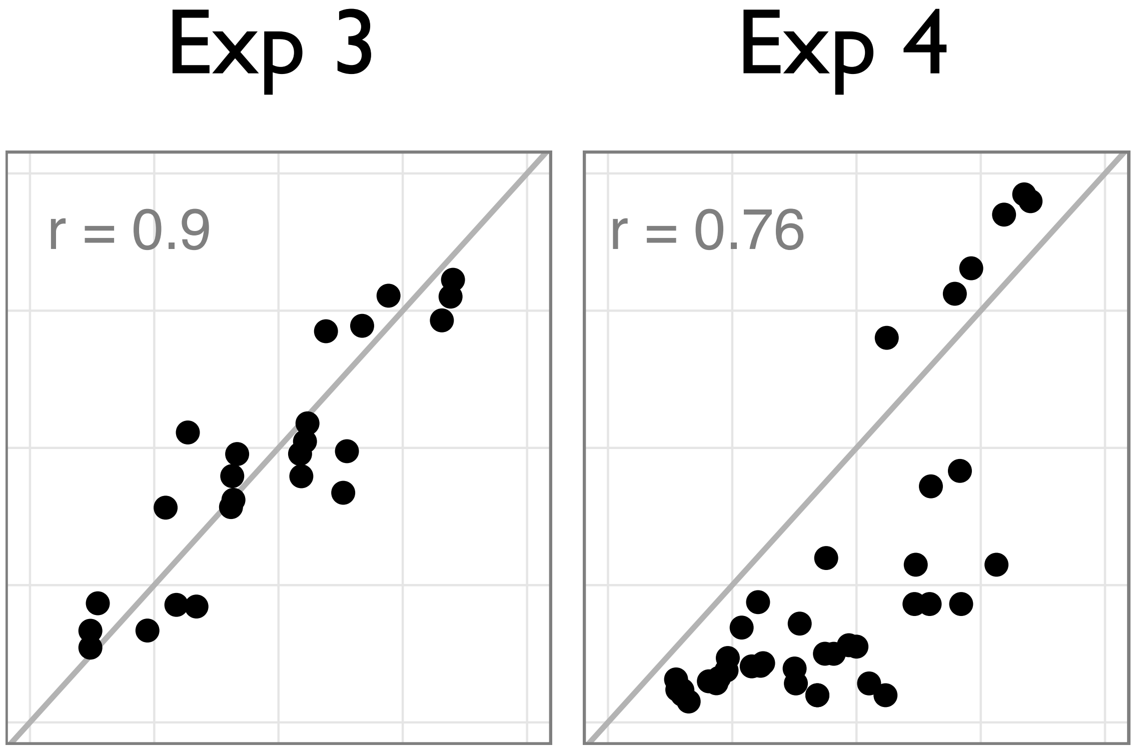
CRP performs poorly



x

Parameters estimated from
Exp 3 and fixed for Exp 4

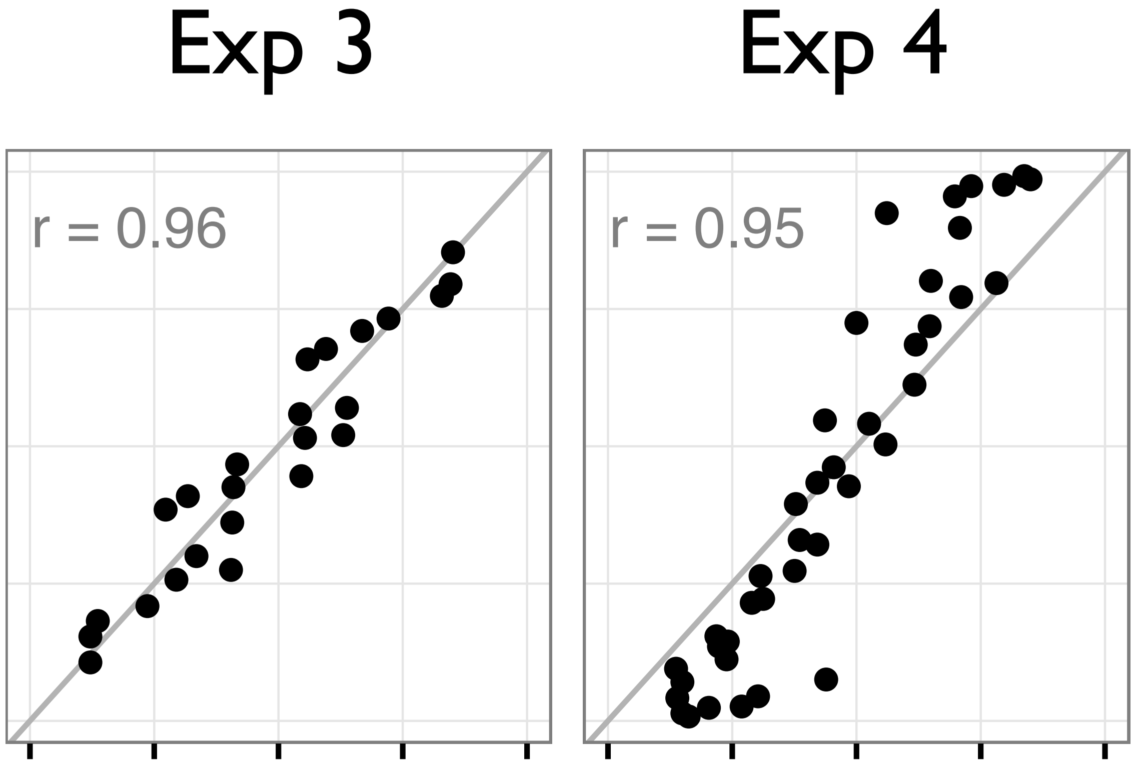
G-CRP model does slightly better



✗

Parameters estimated from
Exp 3 and fixed for Exp 4

HG-CRP model is easily the best

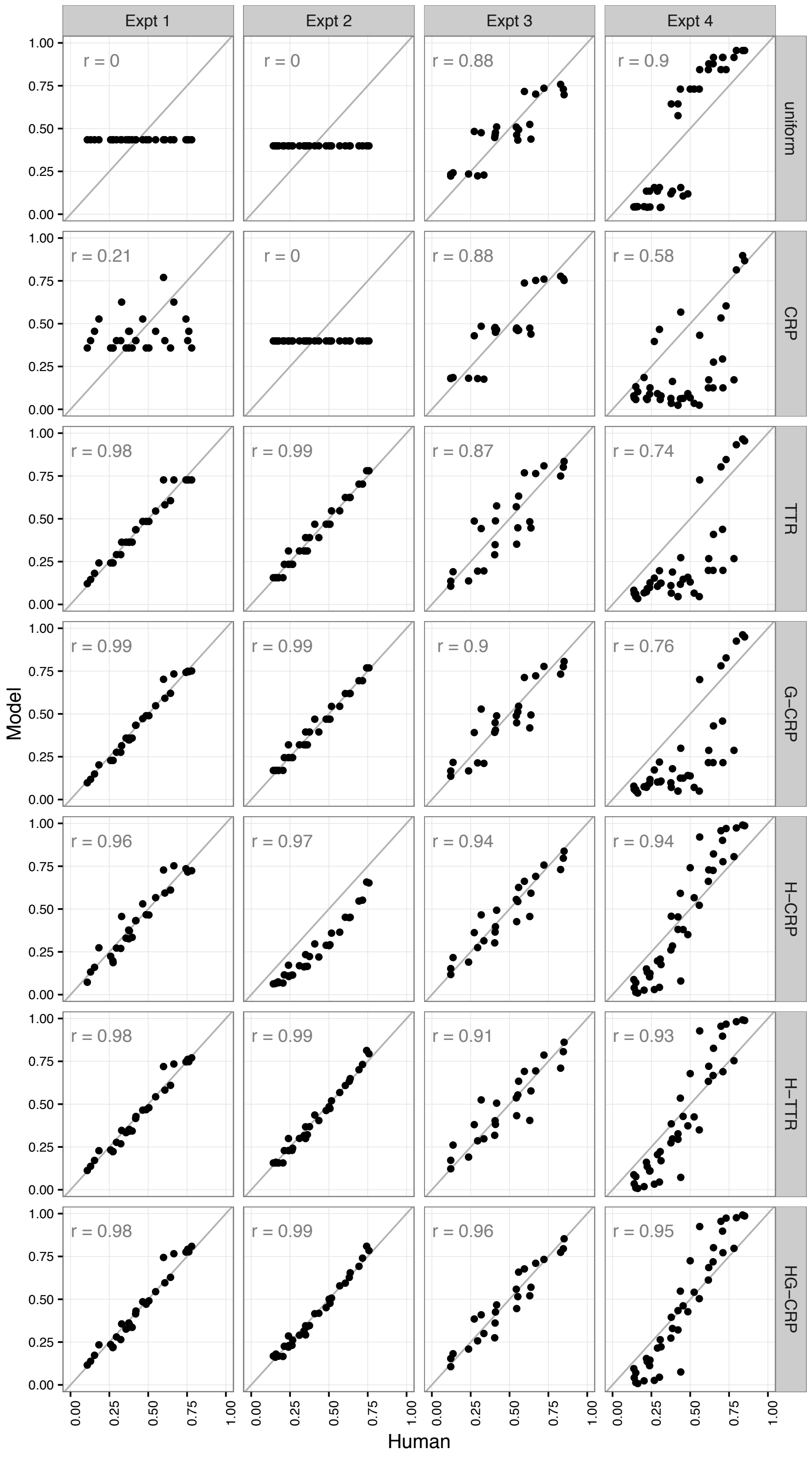


Exp 3

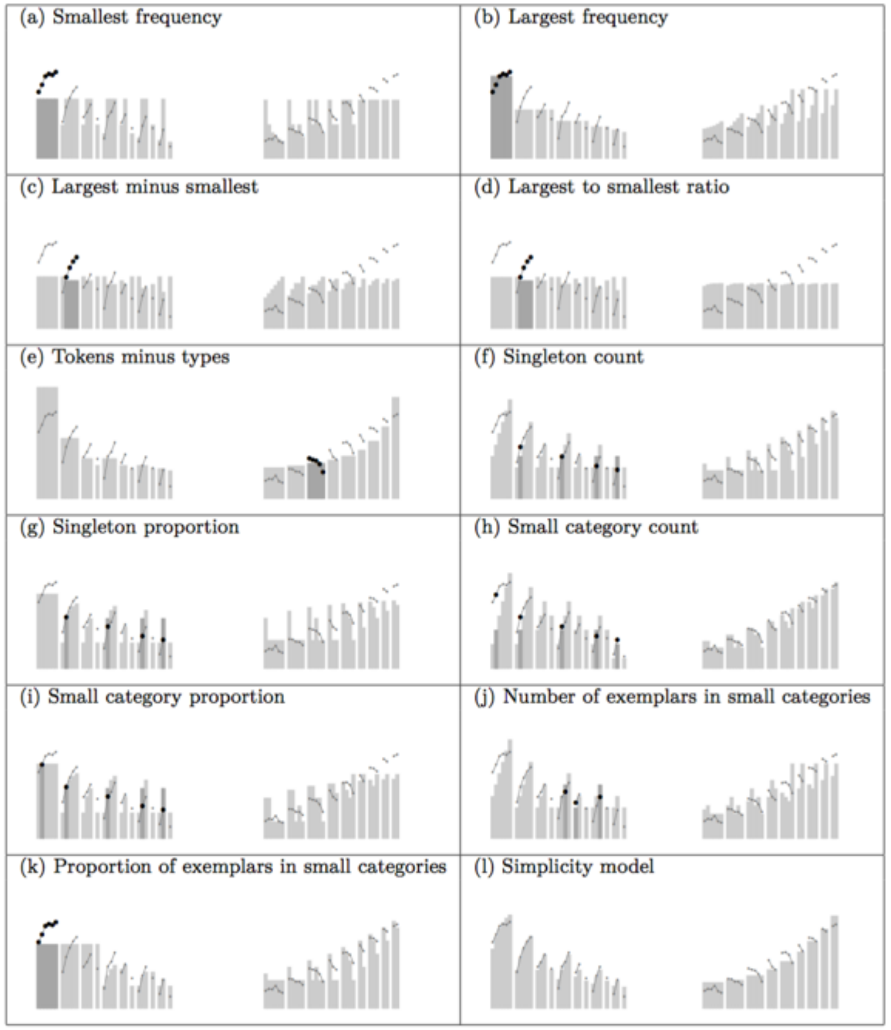
Exp 4

✓

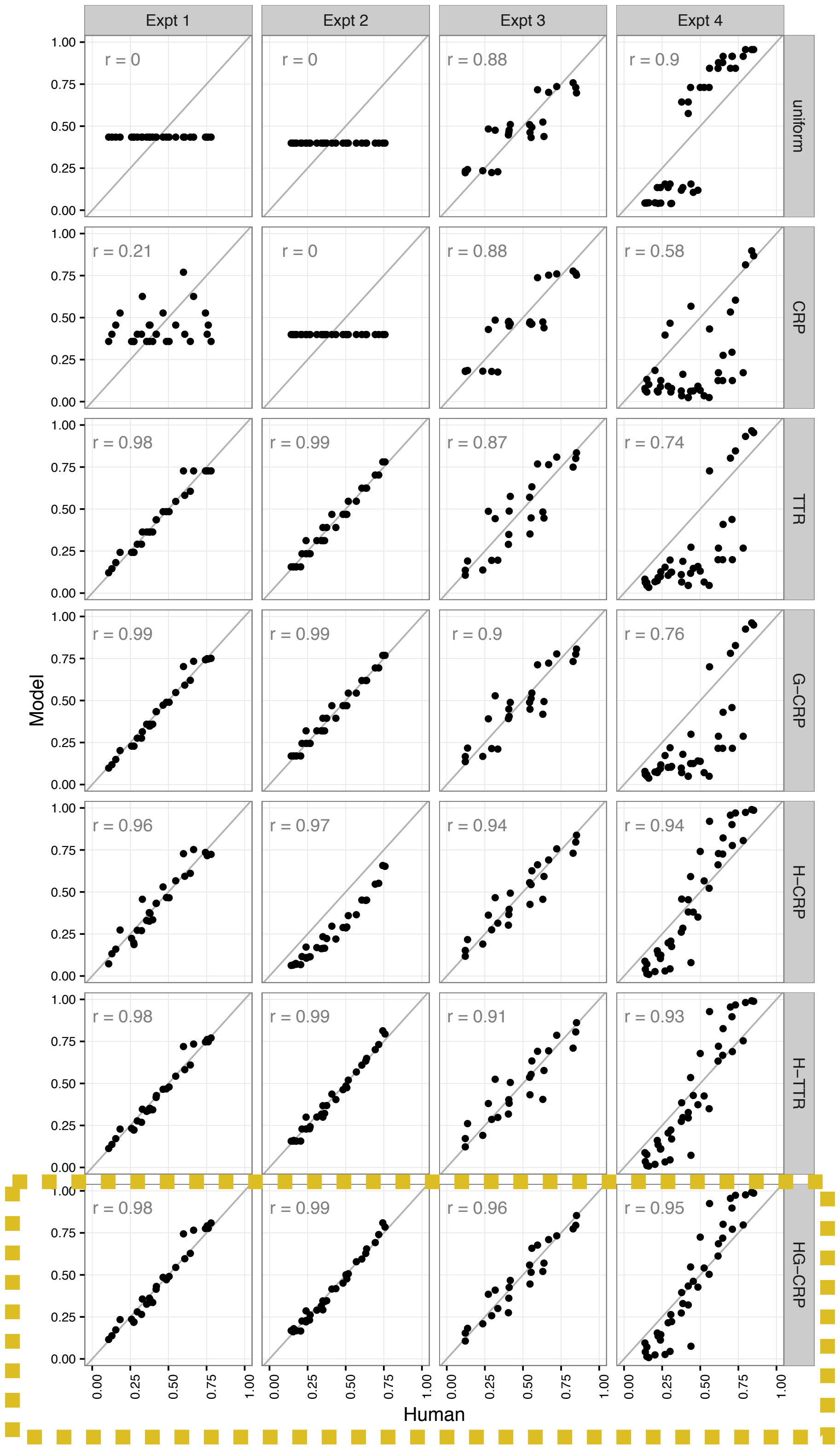
Summary

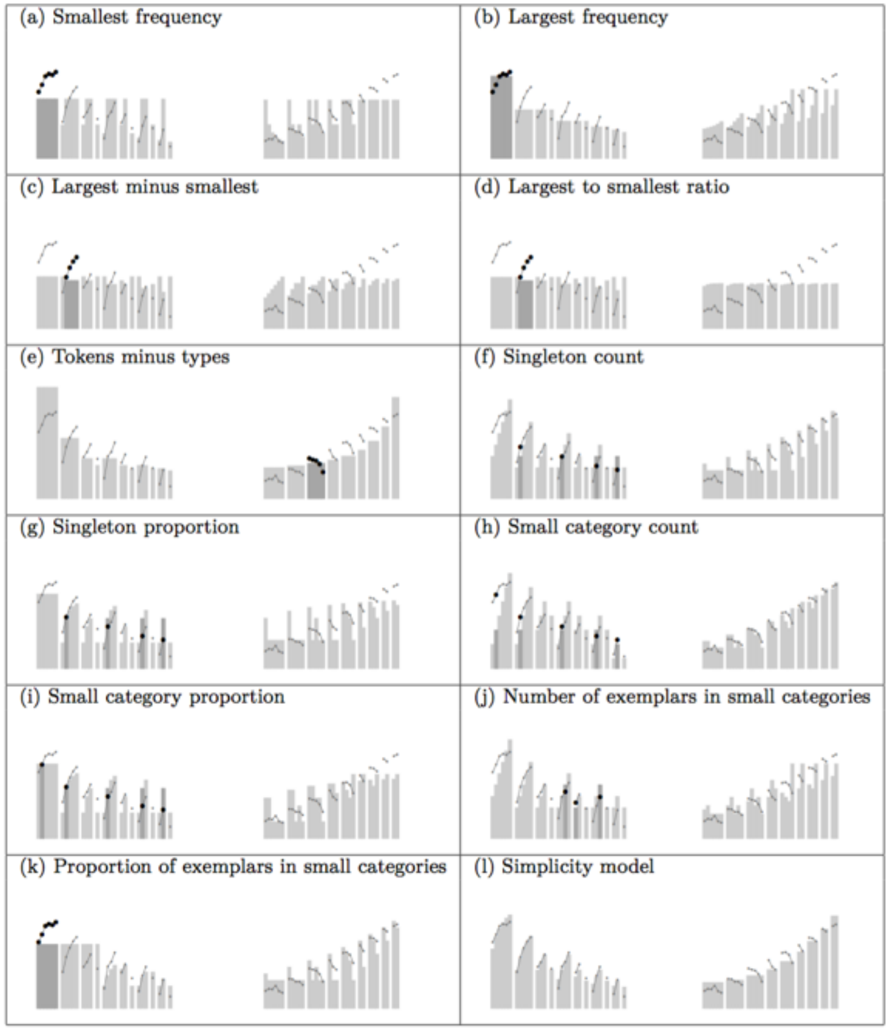


4 experiments, 20+ models, 100+ experimental conditions, 1000+ participants later…



… a model you've never heard of is the winner!




(a) Smallest frequency

(b) Largest frequency

(c) Largest minus smallest

(d) Largest to smallest ratio

(e) Tokens minus types

(f) Singleton count

(g) Singleton proportion

(h) Small category count

(i) Small category proportion

(j) Number of exemplars in small categories

(k) Proportion of exemplars in small categories

(l) Simplicity model

A *better* summary

How do I know this device needs a new label?

How similar is it to familiar things?

How often do I tend to run into novel categories?

How often do I encounter familiar categories?

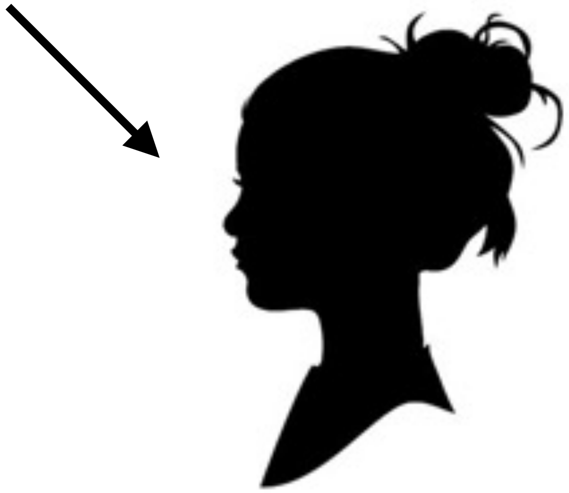
What does the distribution of objects across categories tell me?

Similarity effects

Novel addition effects

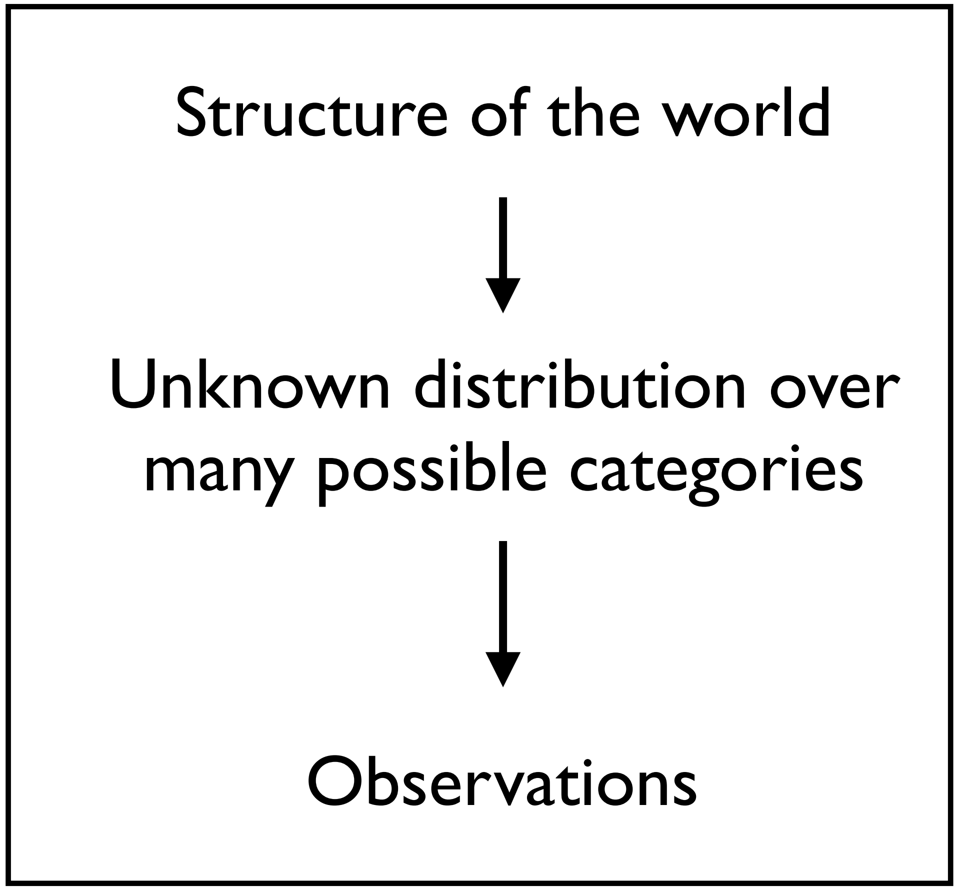
Familiar addition effects

Transfer effects



A theory of novelty detection needs to accommodate all these things

HG-CRP works because it also learns “*what kind of world is this?*” when asked “*is this novel?*”



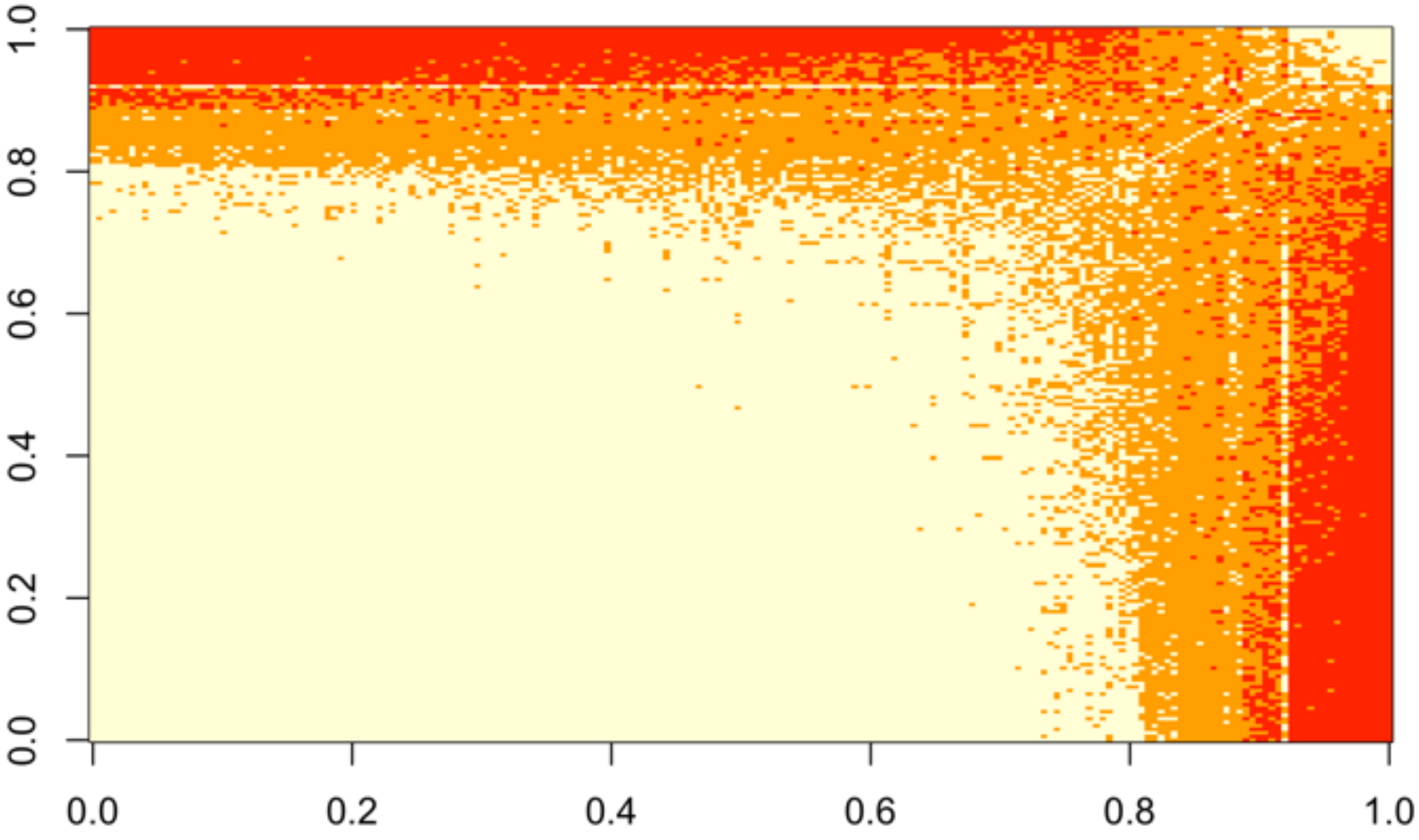
Structure of the world

Unknown distribution over many possible categories

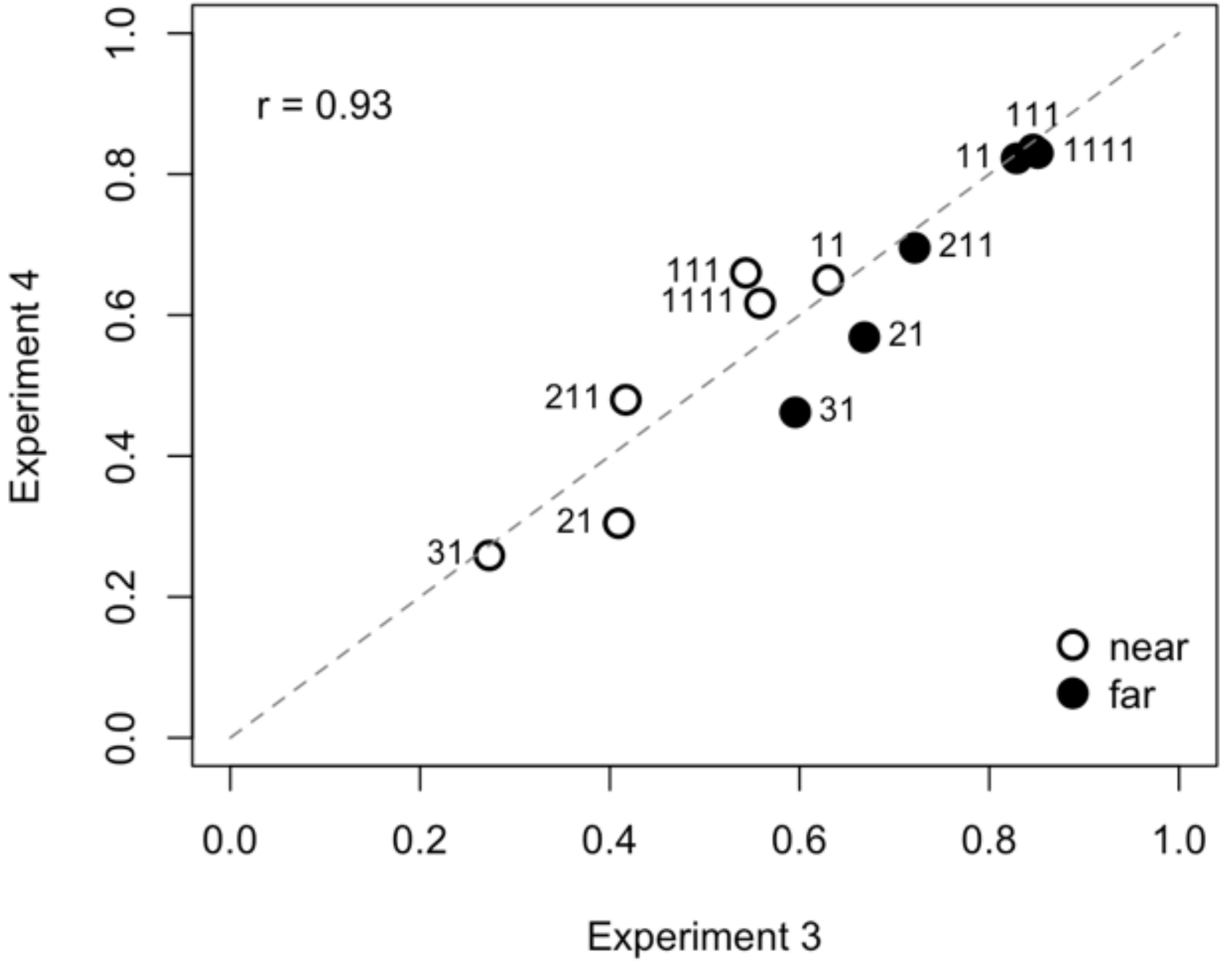
Observations

Thanks

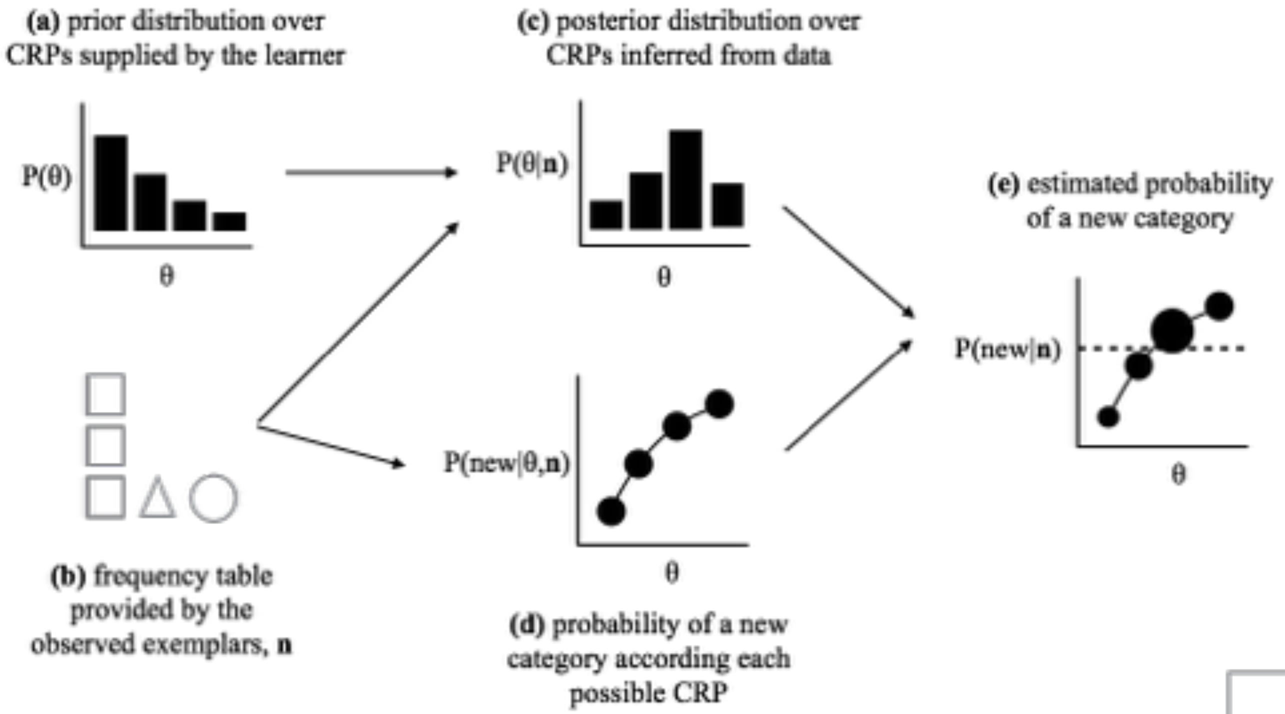
Individual differences (E2)

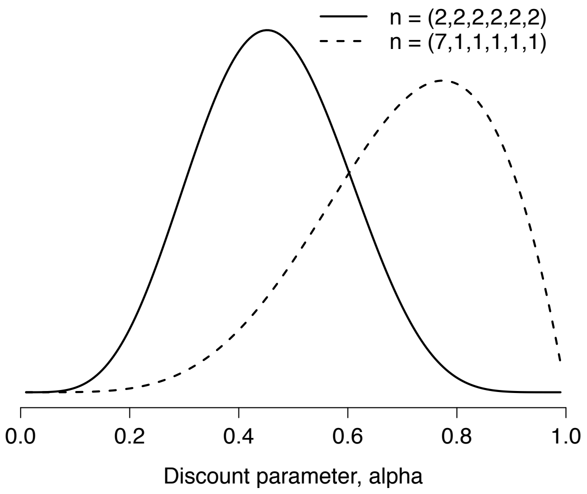


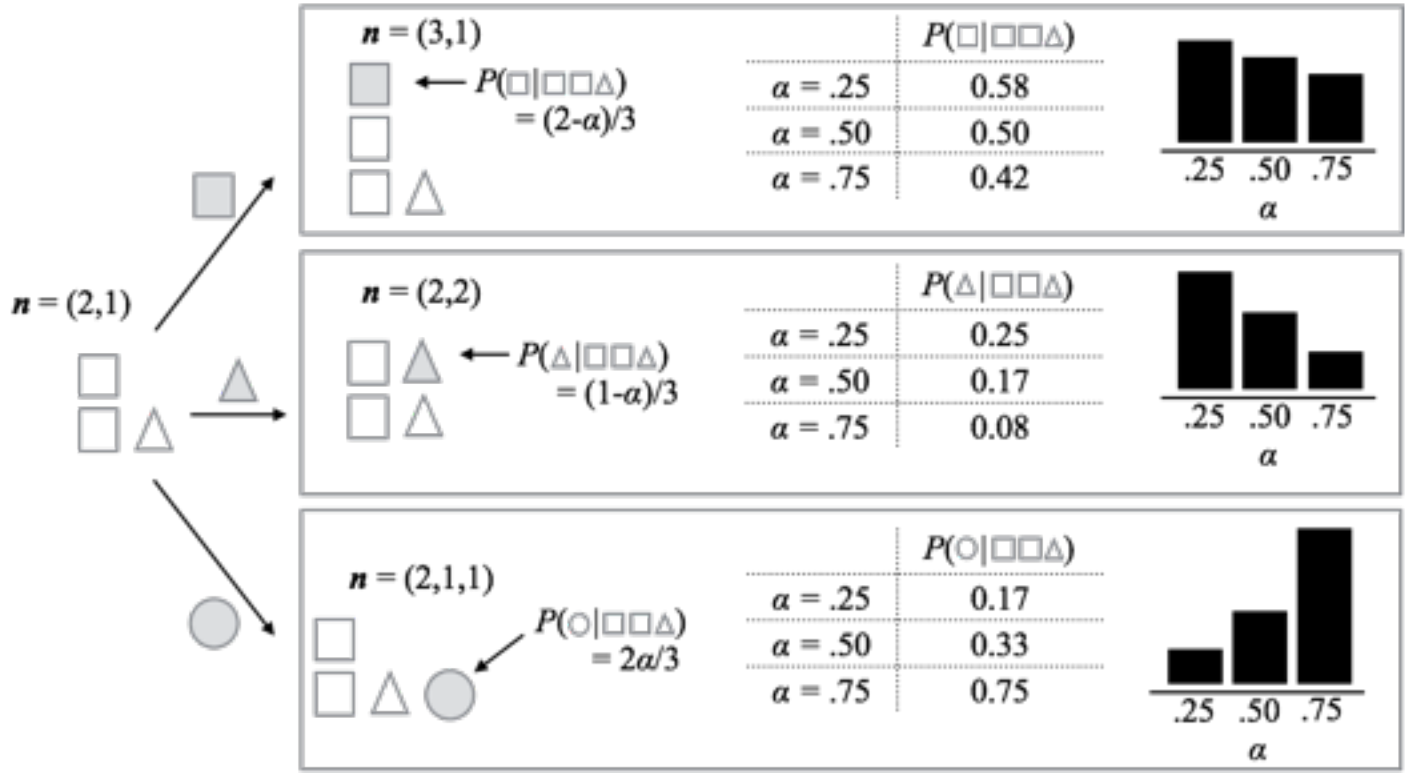
Replication check:



Why does the transfer effect exist? Learning distributional shape on the fly


(a) prior distribution over CRPs supplied by the learner
P(θ)
θ
(c) posterior distribution over CRPs inferred from data
P(θ|n)
θ
(e) estimated probability of a new category
P(new|n)
θ
(b) frequency table provided by the observed exemplars, n
P(new|θ,n)
θ
(d) probability of a new category according each possible CRP





	P(□|□□△)
α = .25	0.58
α = .50	0.50
α = .75	0.42

	P(△|□□△)
α = .25	0.25
α = .50	0.17
α = .75	0.08

	P(○|□□△)
α = .25	0.17
α = .50	0.33
α = .75	0.75