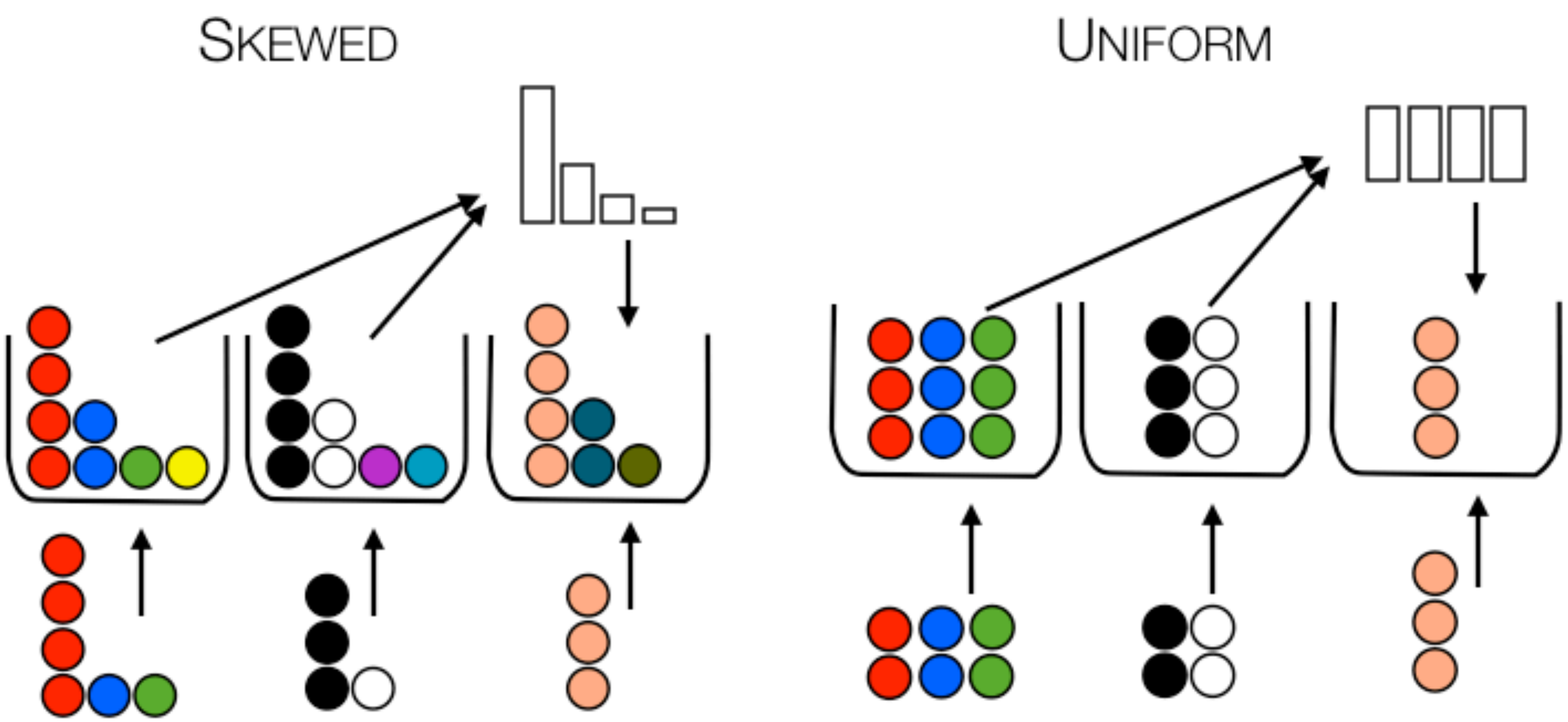
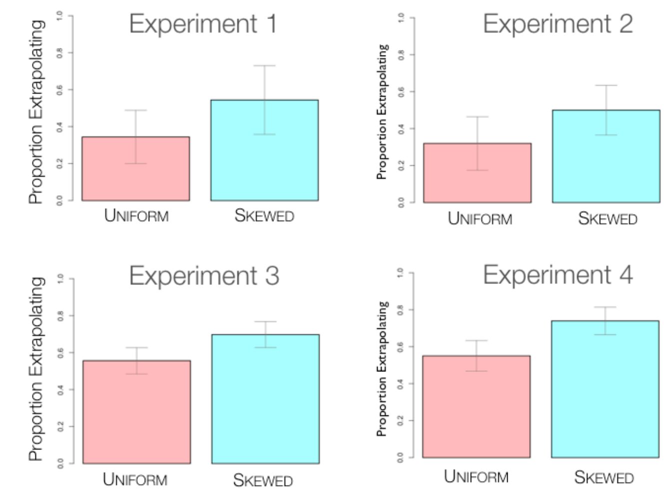
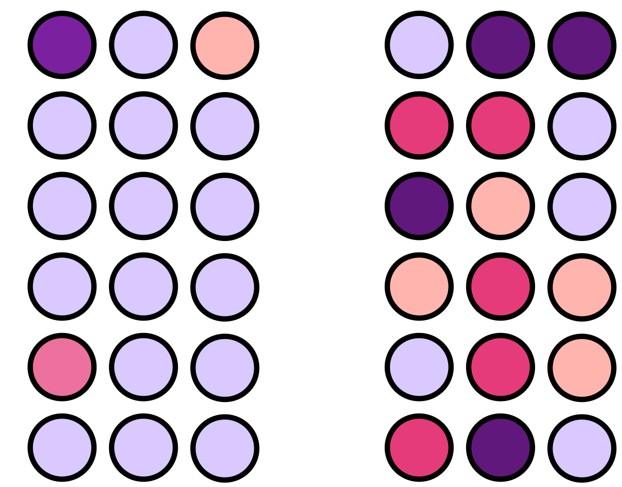
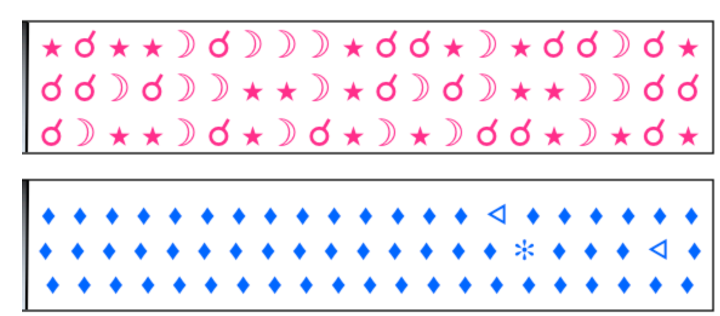
# Computational Cognitive Science



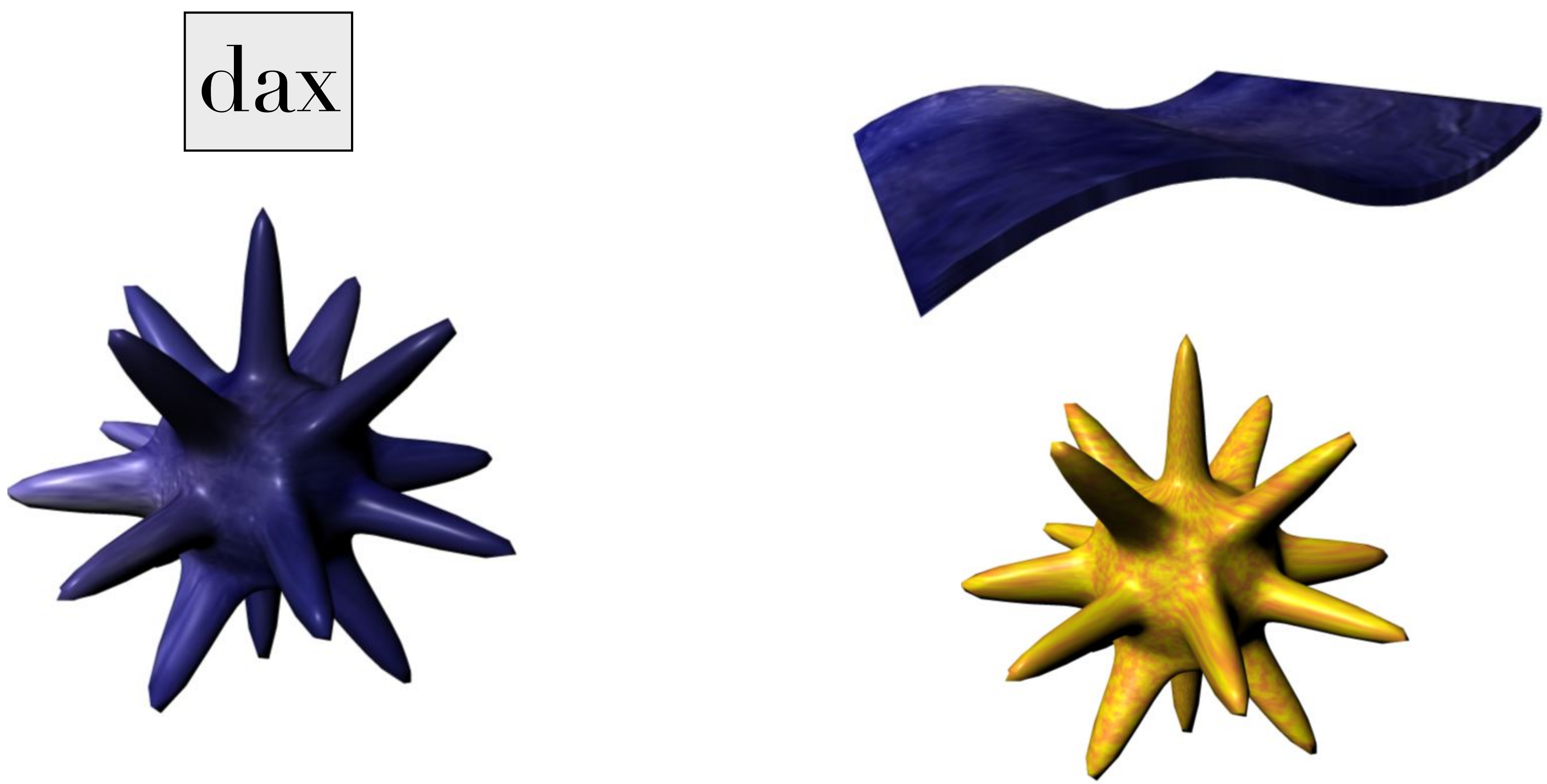





Lecture 12: Higher order knowledge 2

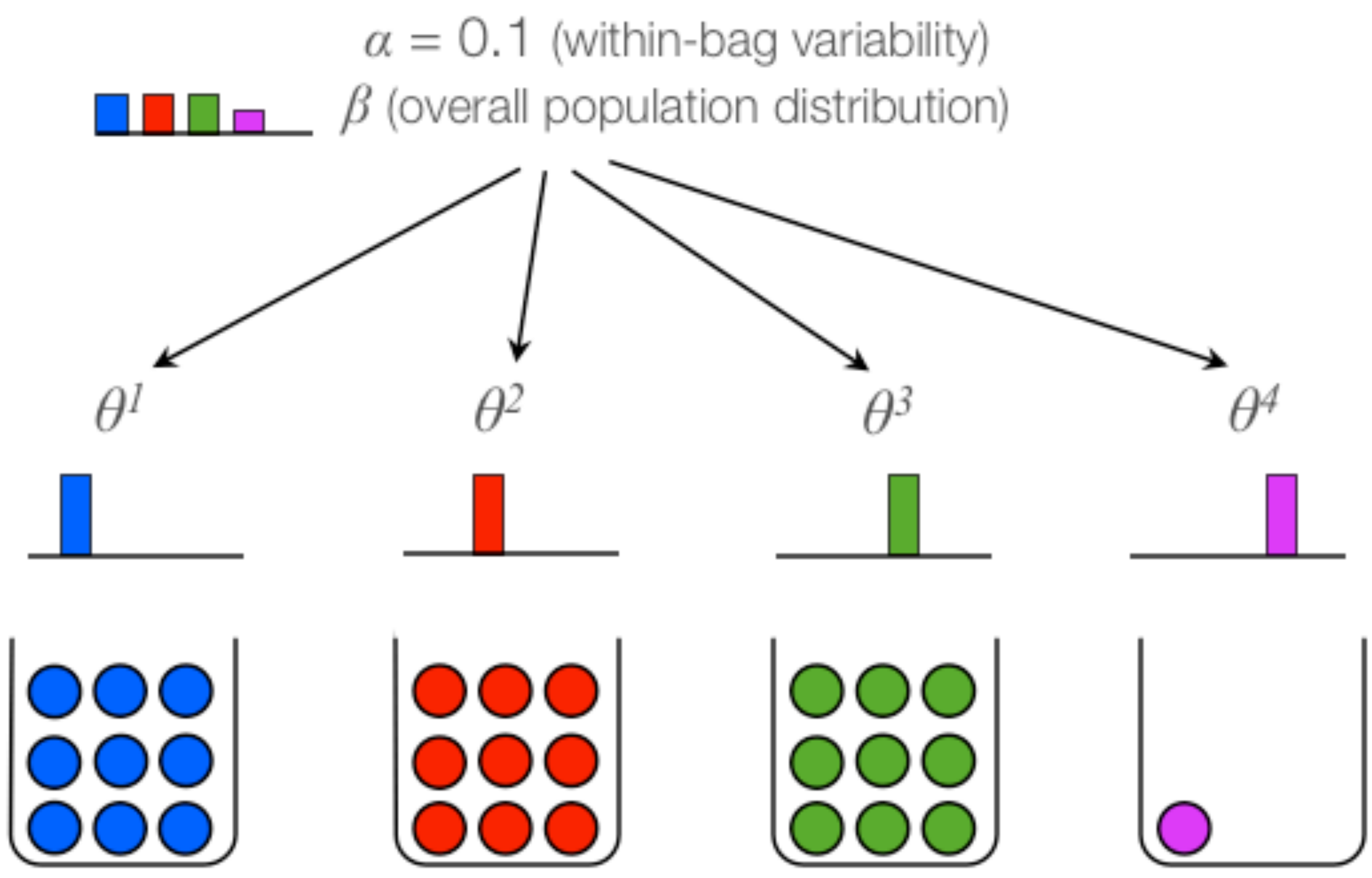
# Higher order knowledge so far

- Last lecture we saw how people can learn higher-order knowledge about hypotheses (called overhypotheses), which licenses inferences based on just one datapoint



# Higher order knowledge so far

- We also saw that **hierarchical Bayesian models** can capture this sort of learning (although they tend not to capture human limitations, at least not without additions)



# Higher order knowledge so far

- We also saw that hierarchical Bayesian models can capture this sort of learning (although they tend not to capture human limitations, at least not without additions)
- That model (and learning) had to do with the *variability* of different features within categories

- There are lots of other kinds of overhypotheses
- Today we consider another -- about the *distribution* of types of things within a category or domain

# Lecture outline (next three lectures)

- Last time: Learning about category variability
  - This kind of learning in children and adults
  - A model for this kind of learning
  - Limitations of this model
- Today: Learning about distributions of categories
  - This kind of learning in adults
  - Failure of current models
  - A model for this kind of learning
- Lecture 13: Learning about category structure
  - A model for this kind of learning
  - This kind of learning in people

# Lecture outline (next three lectures)

- Last time: Learning about category variability
  - This kind of learning in children and adults
  - A model for this kind of learning
  - Limitations of this model
- ➡ Today: Learning about distributions in categories
  - ➡ This kind of learning in adults
  - Failure of current models
  - A model for this kind of learning
- Lecture 13: Learning about category structure
  - A model for this kind of learning
  - This kind of learning in people

# Learning about distributions in categories

There are several different ways our ability to learn about distributions is evident

- You see an 18-year old man. How old do you think he'll be when he dies?

  70–80 probability > 120 = 0

- You hear that a movie has earned $10M so far, but you don't know how long it's been running. How much do you think it will make in total?

  $40M? $100M? could be $500M, $800M

These answers are based on knowing the distribution of ages and movie grosses.

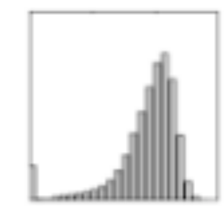
# Learning about distributions in categories

Distributions are useful for making predictions about expected values of common events

Life spans: You see an 18-year-old man. How old do you think he'll be when he dies?

Movie run times: You have been watching a movie for 30 minutes. How long do you think the movie is?

Gaussian



Movie grosses: You hear that a movie has earned $10M so far, but you don't know how long it's been running. How much do you think it will make in total?

Poem lengths: Your friend reads line 5 from a poem. How long do you think the poem is?

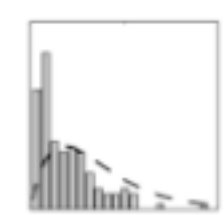
Reigns of pharoahs: You read that a particular pharoah had been reigning for 11 years in 4000BC. How long did he reign in total?

Power law

Cake baking times: A cake has been in the oven for 35 minutes. How long do you think its total baking time is?

Terms of US Representatives: A particular rep. has served for 15 years. How long do you think heis total term will be?

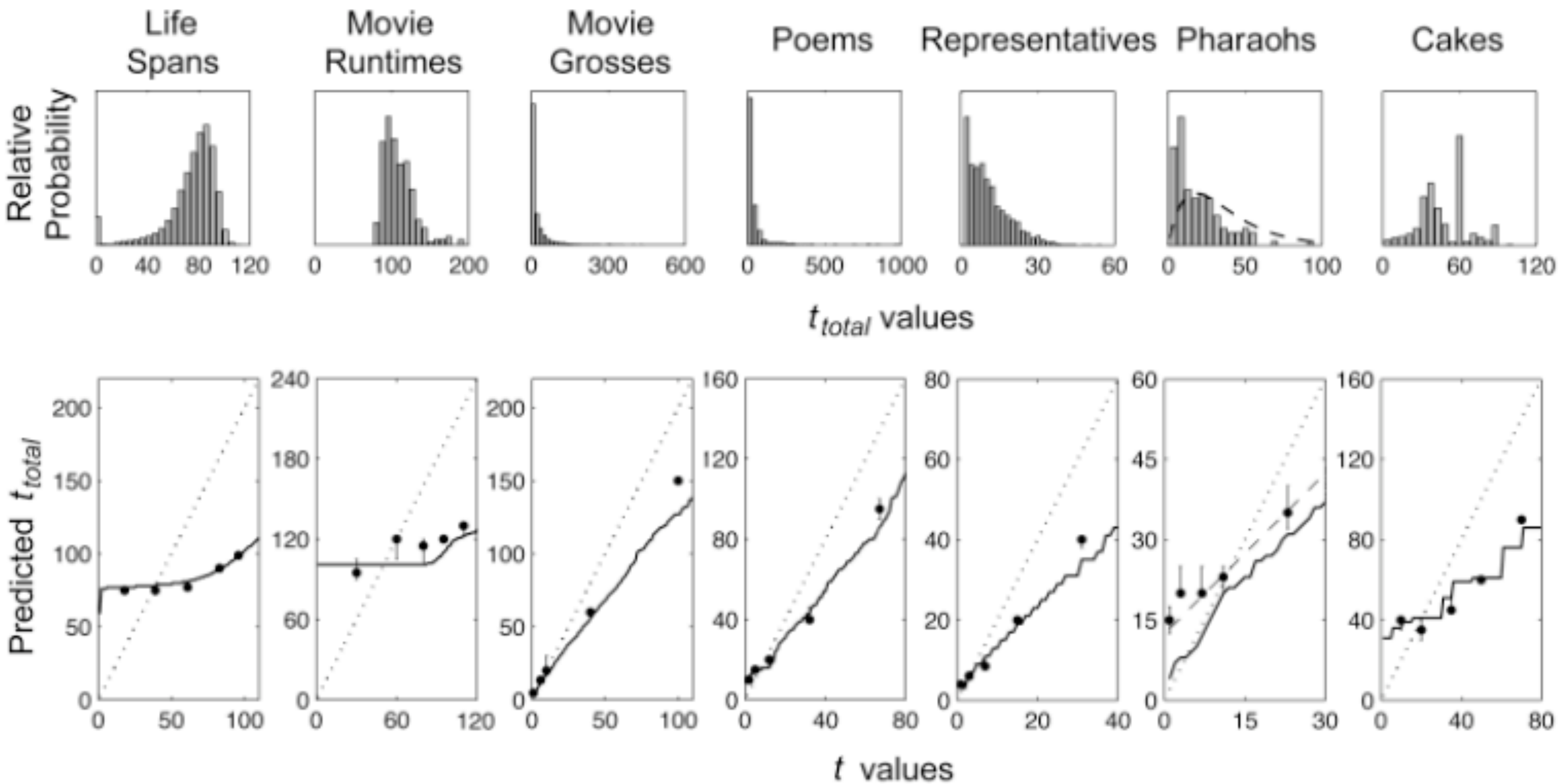
Erlang



Griffiths & Tenenbaum, 2006

# Learning about distributions in categories

People are very good at predicting the length of time for common events, based on abstract knowledge about the nature of the distribution




line = model that uses that distribution; points are people's guesses

Griffiths & Tenenbaum, 2006

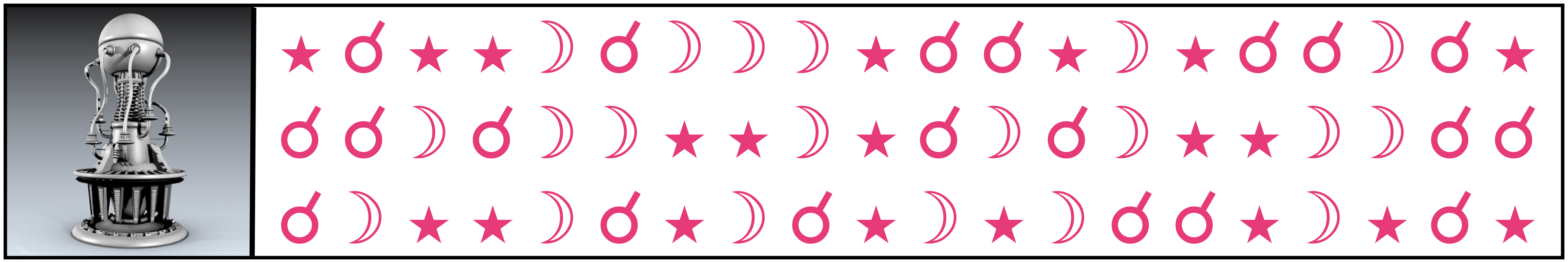
# Learning about distributions in categories

Another thing that knowing about the distribution of things in categories is useful for is predicting how many things you haven’t yet seen

An alien machine

# Learning about distributions in categories

This machine outputs symbols in an alien alphabet...

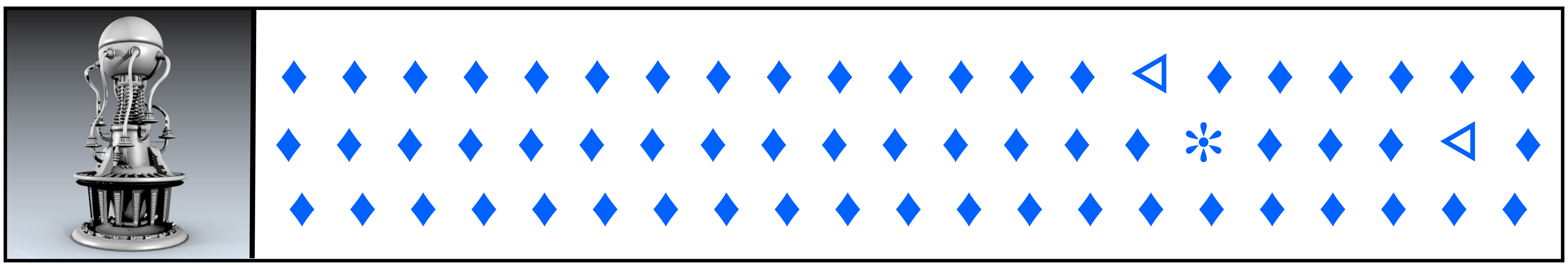


How surprised would you be if
the next symbol was + ?

# Learning about distributions in categories

This machine outputs symbols in an alien alphabet...





What about now?

Would a + be a big surprise now?

# This is obviously a frequency effect

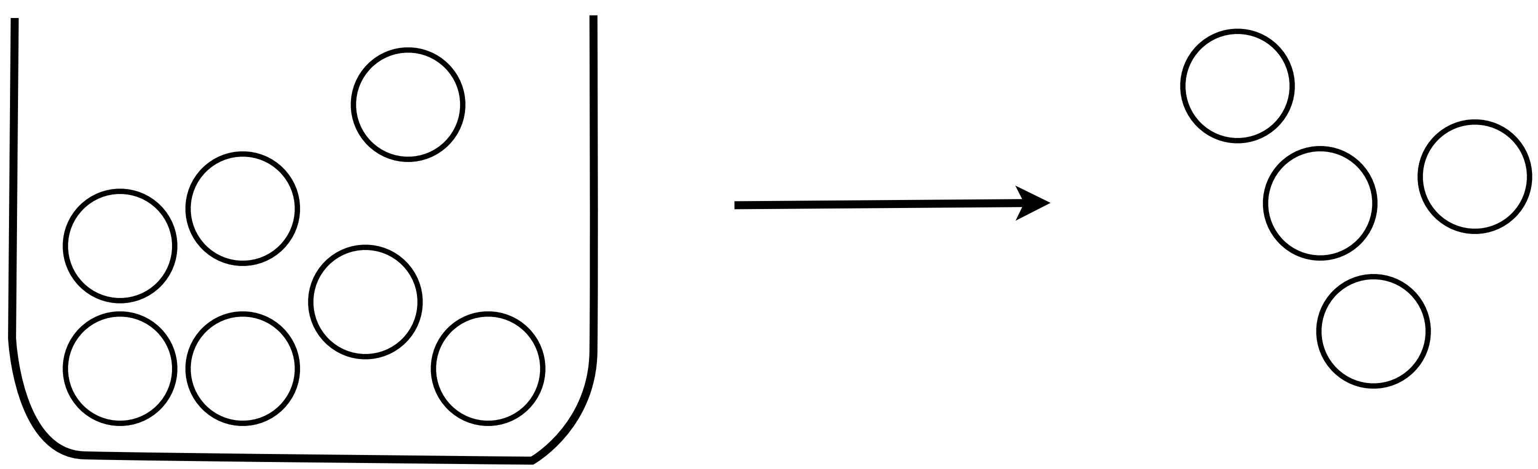
| | |
|---|---|
| ★ 20 instances | ♦ 57 instances |
| ☌ 20 instances | ☍ 2 instances |
| ☽ 20 instances | ☼ 1 instance |

Same number of instances (60),
same number of exemplars (3),
but different *distributions*

But is this a real phenomenon, and not just a thought experiment in class?

# A simple experiment: The basic idea

Simplified task: how many types (colours) of marble are there in a bag?

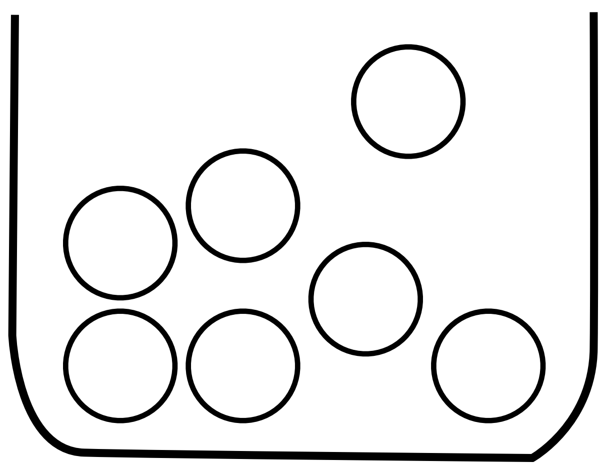


Bag full of marbles (of many types?)

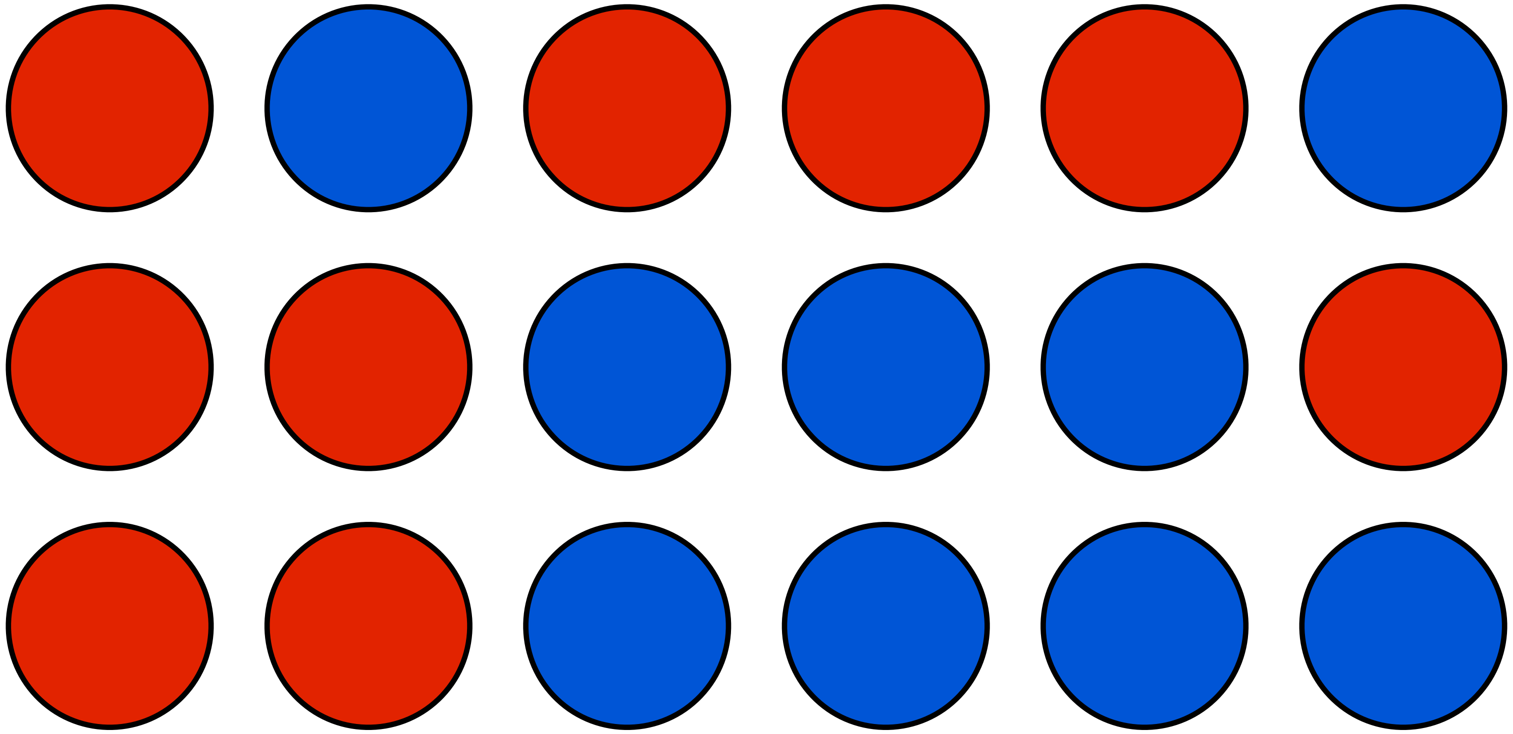
Draw some from the bag and make a guess

# A simple experiment: The basic idea

Bag A contains 100 marbles...



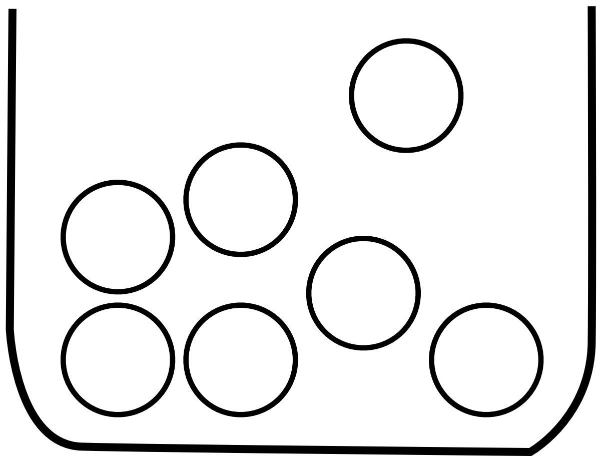
# A simple experiment: The basic idea



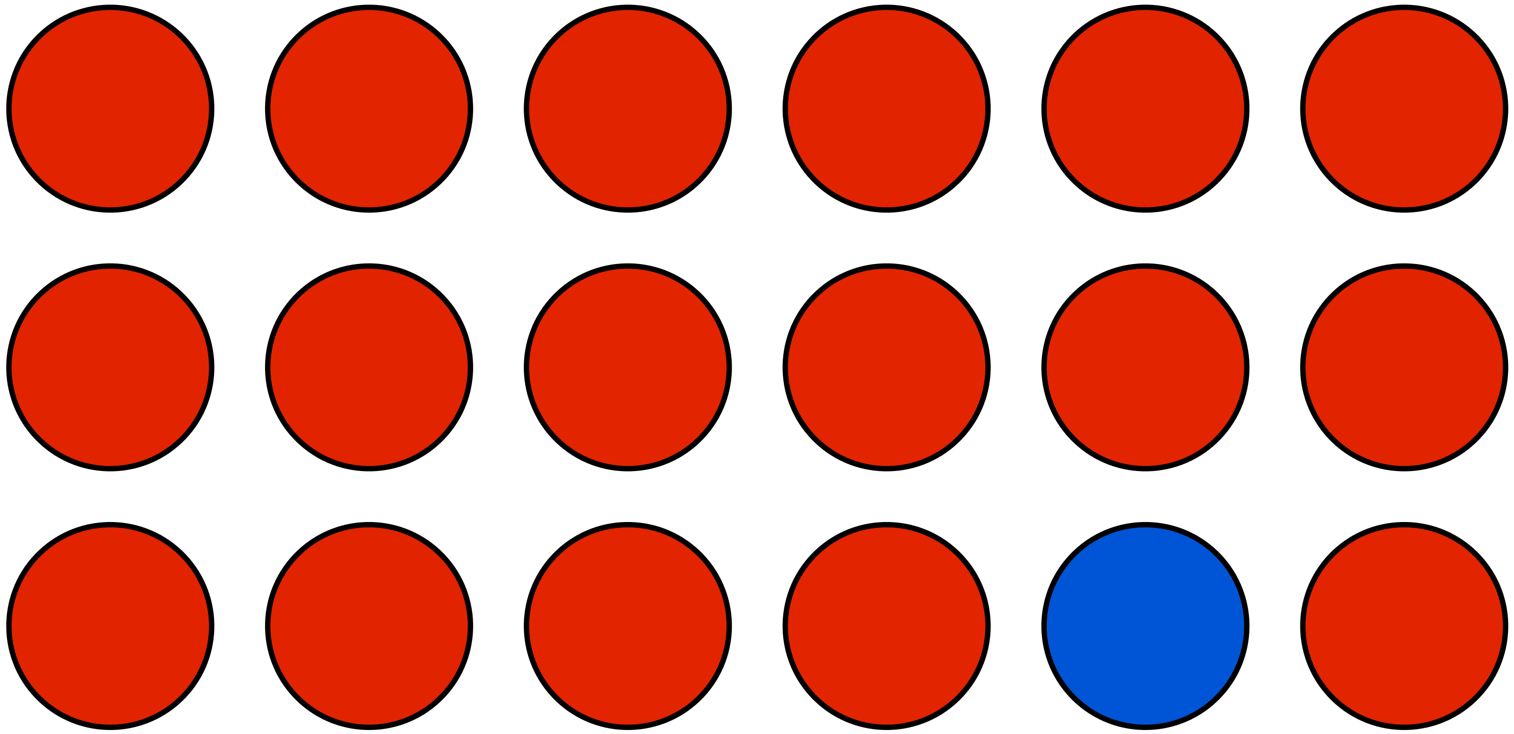
Probably only two types of marble in bag

# A simple experiment: The basic idea

Bag B also contains 100 marbles...



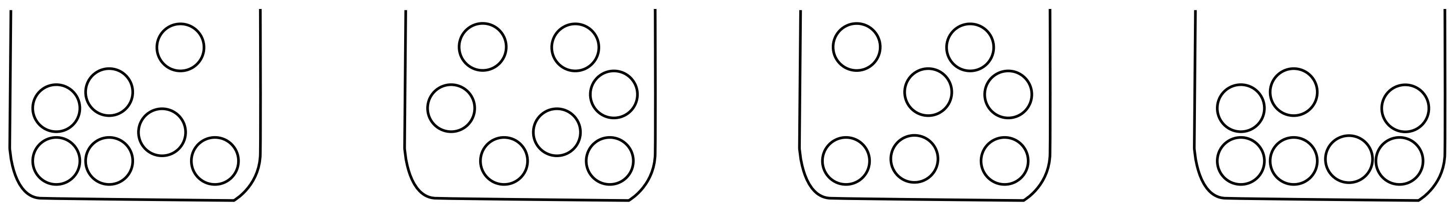
# A simple experiment: The basic idea



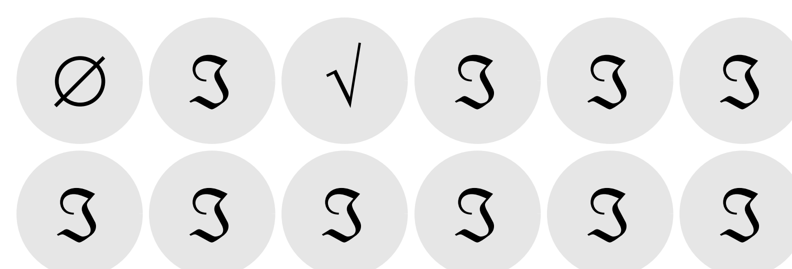
Maybe two types? Maybe more?

# Experiment structure

- Participants see a series of bags (each with 100 marbles)



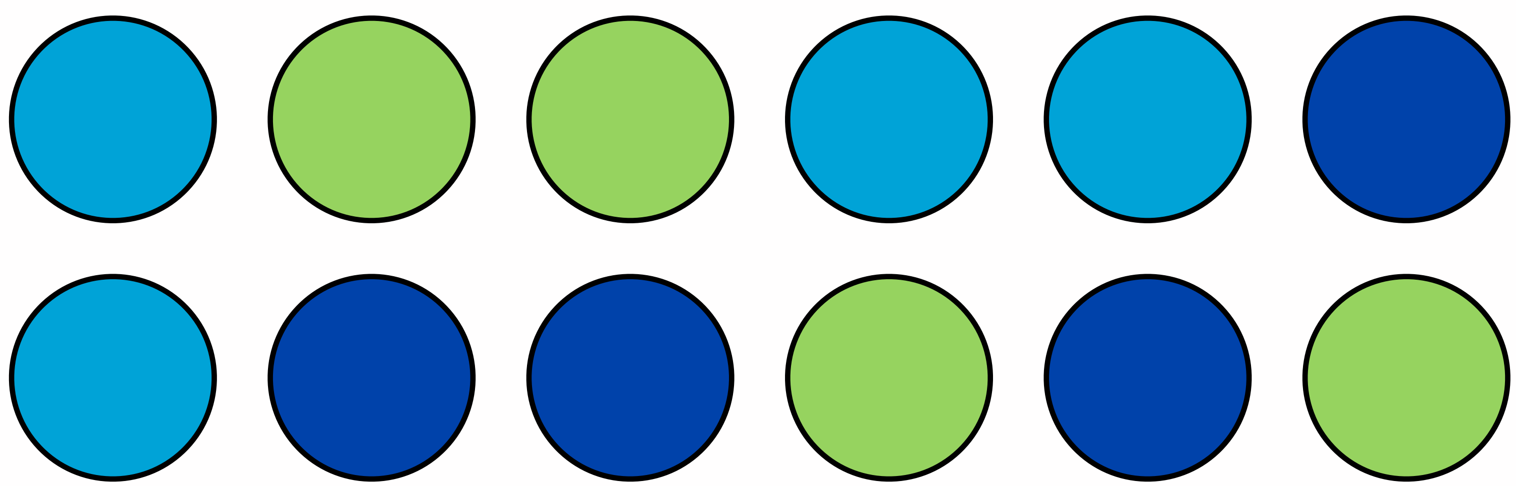
- For each bag, participants are shown a sample and asked to guess how many types were in the full bag

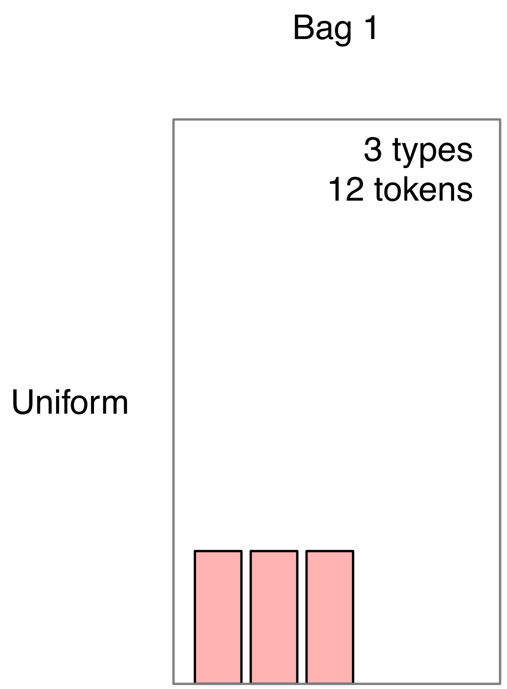


- Two conditions:
  - **UNIFORM**: in each bag, there are approximately the same number of items of each colour
  - **SKEWED**: in each bag, the vast majority are one colour, and the others occur at very low frequency

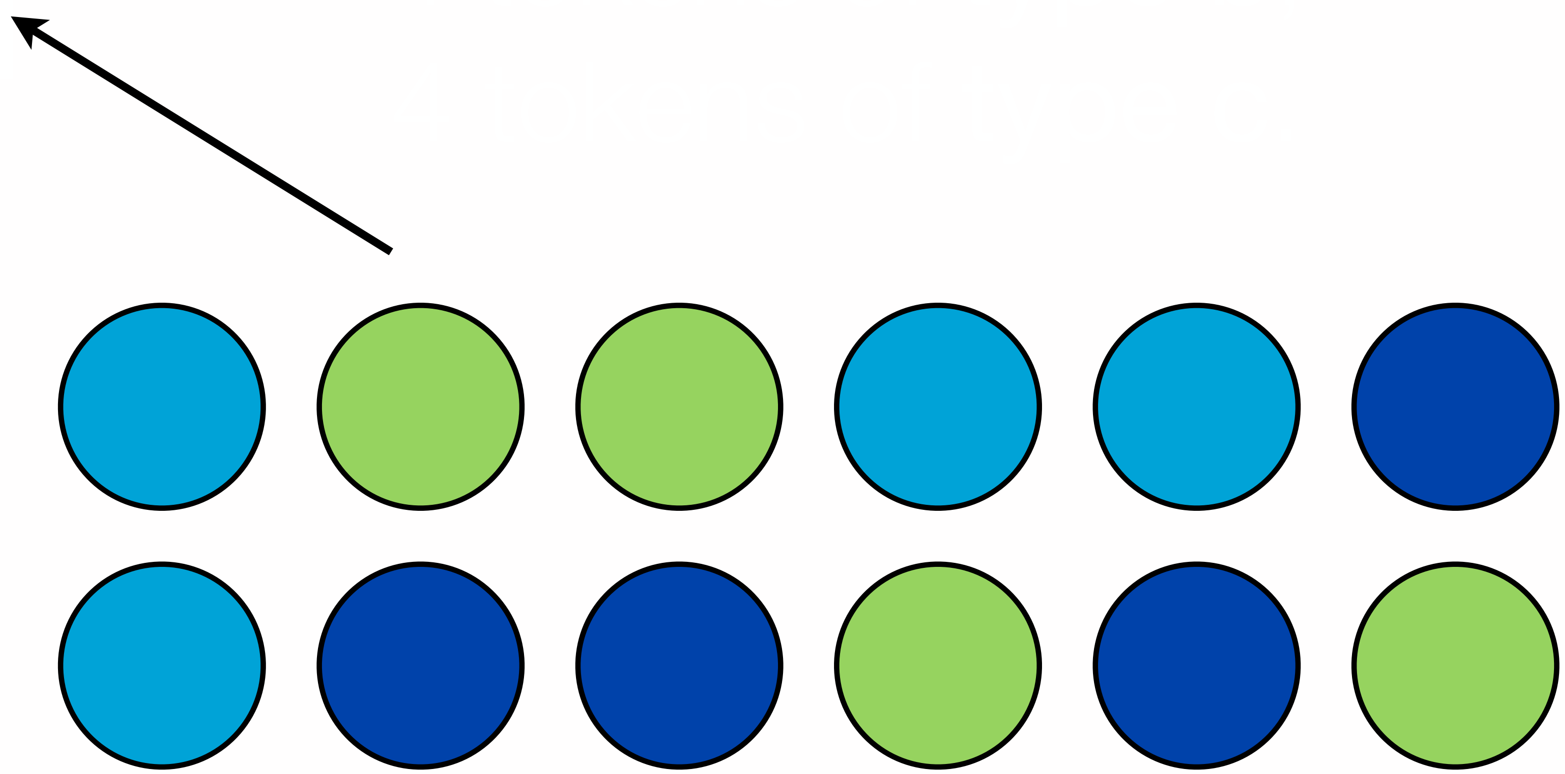
UNIFORM condition, bag #1...

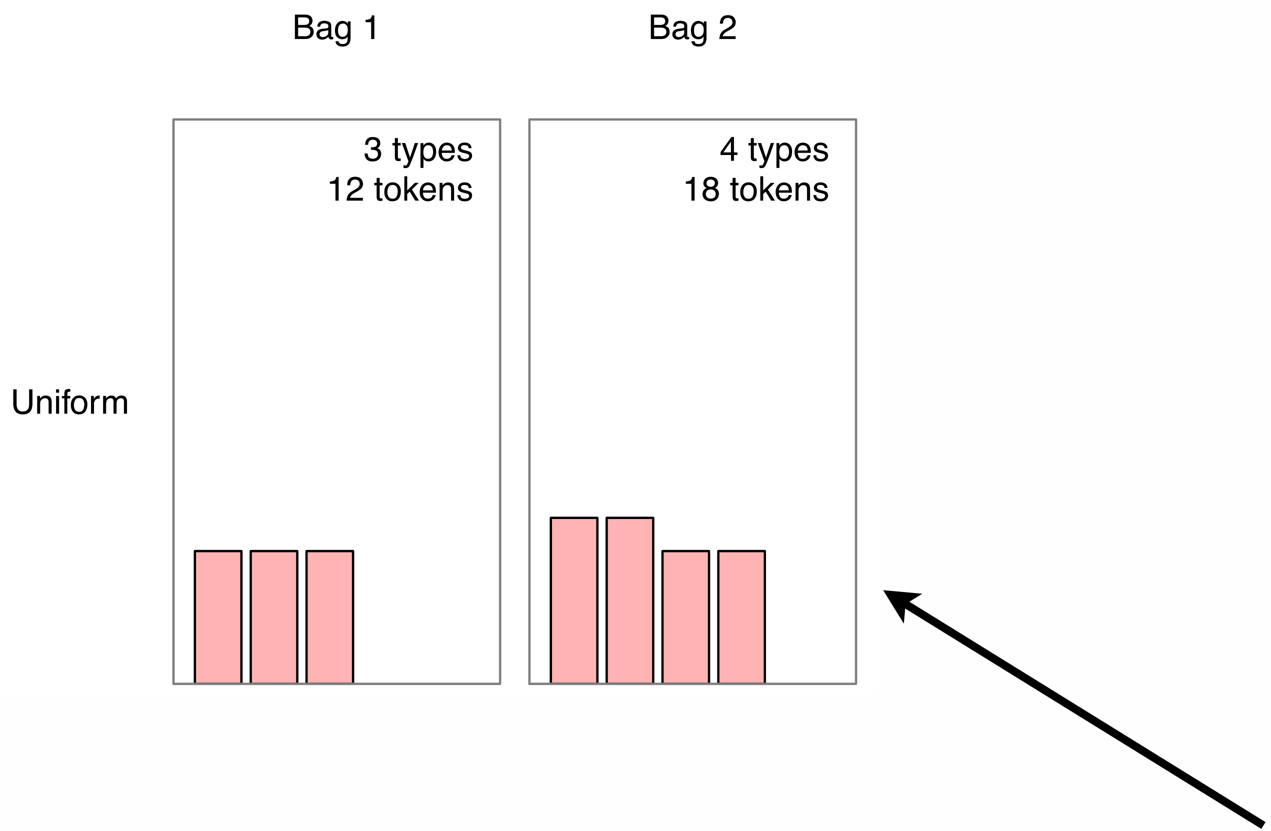
4 tokens of type a,
4 tokens of type b,
4 tokens of type c.



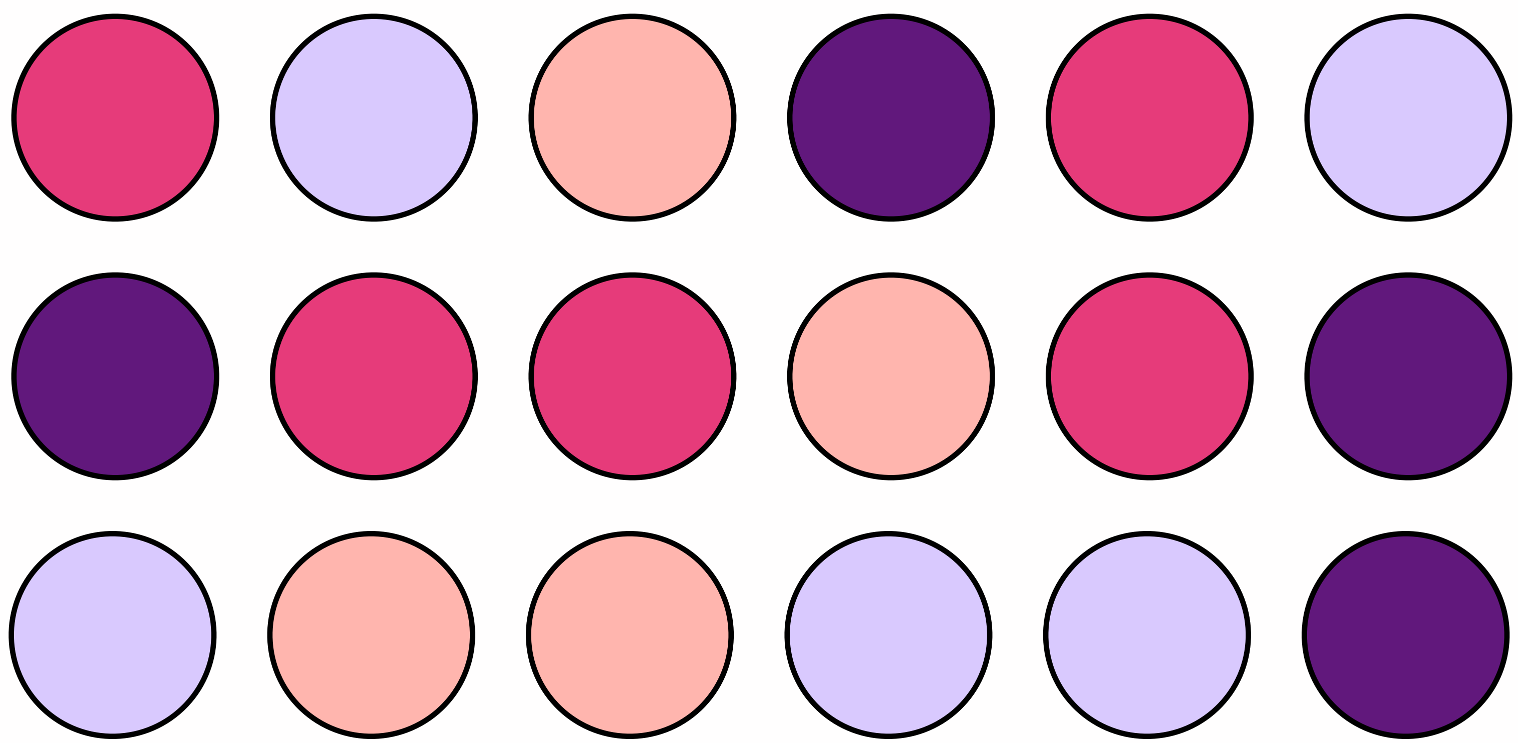


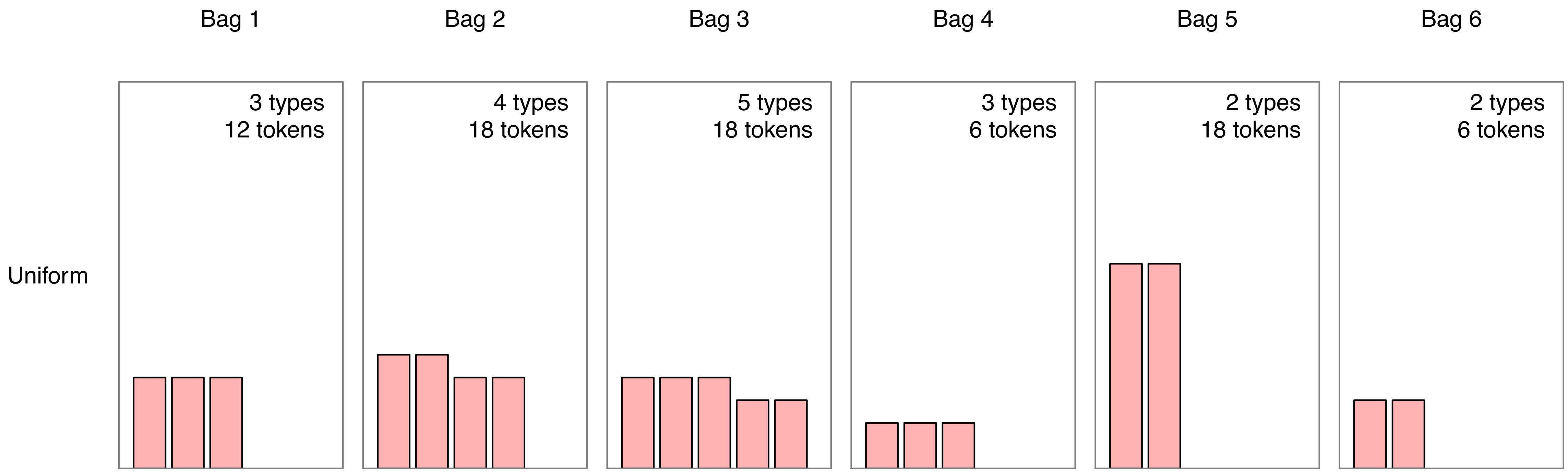

# UNIFORM condition, bag #1...

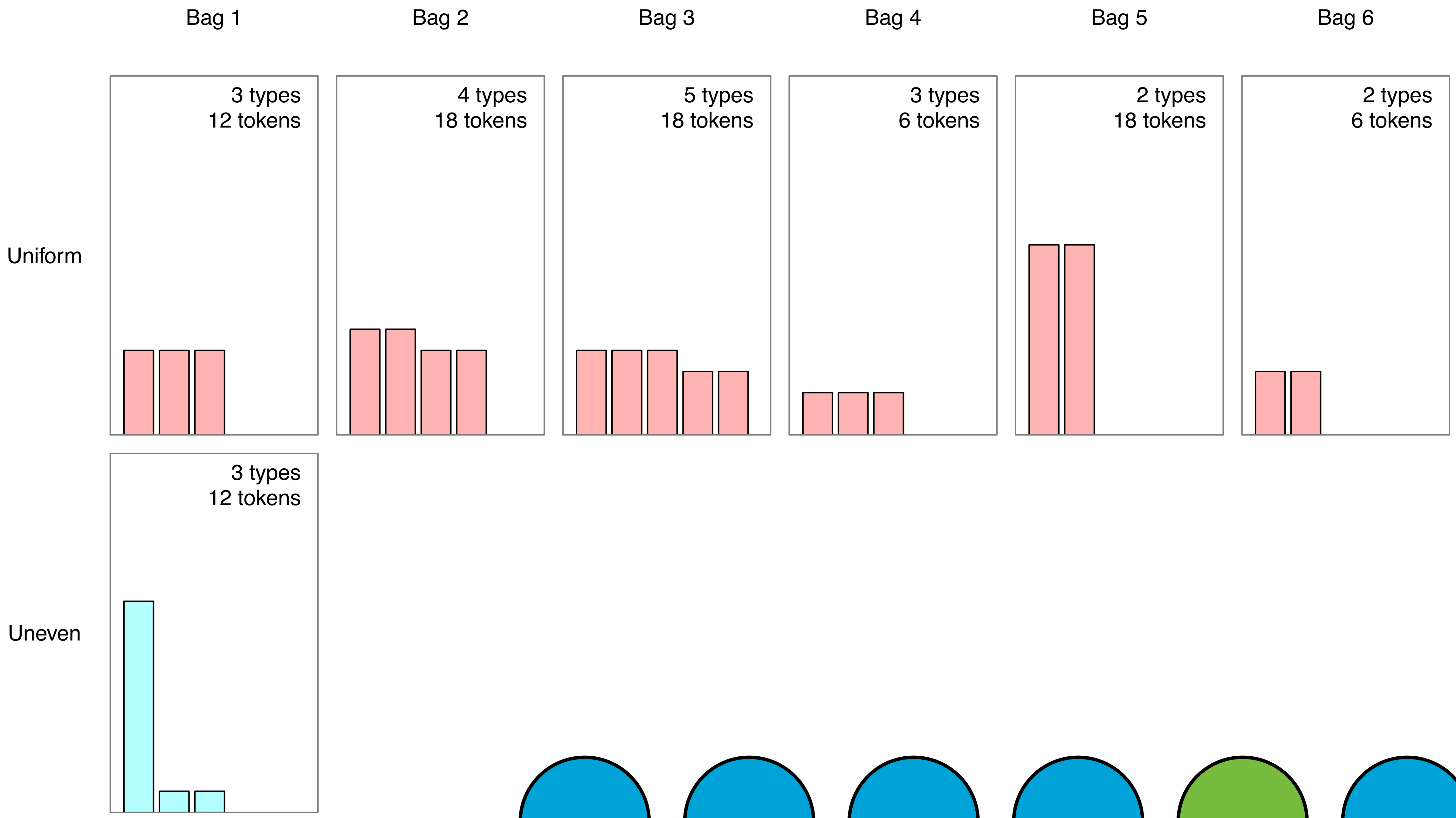


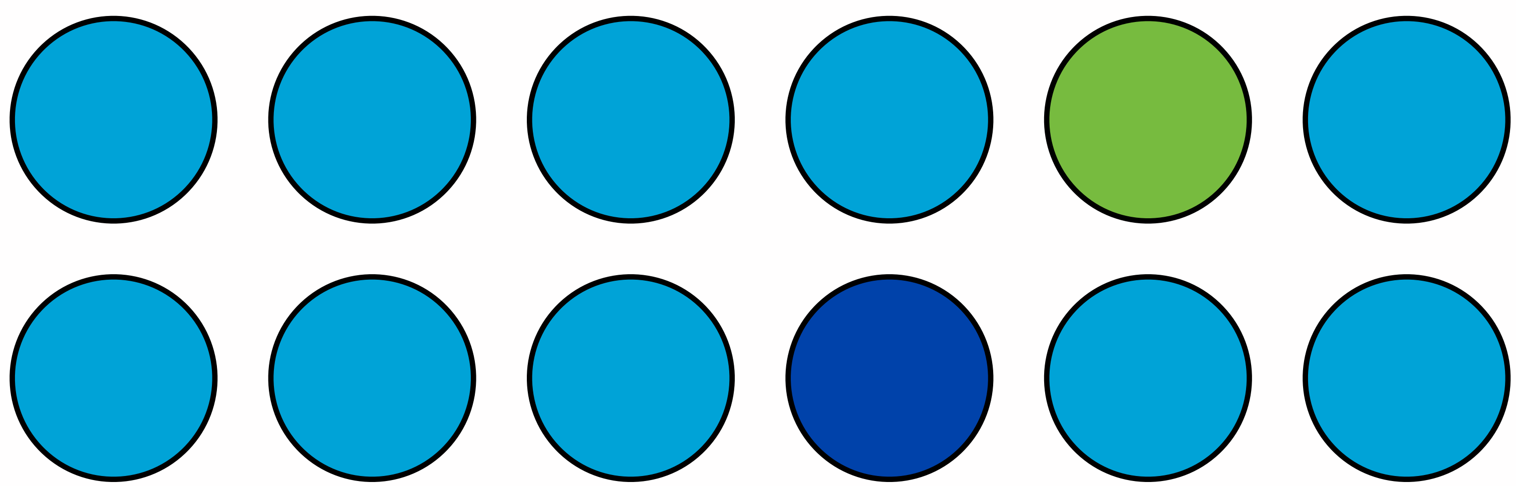



UNIFORM condition, bag #2...

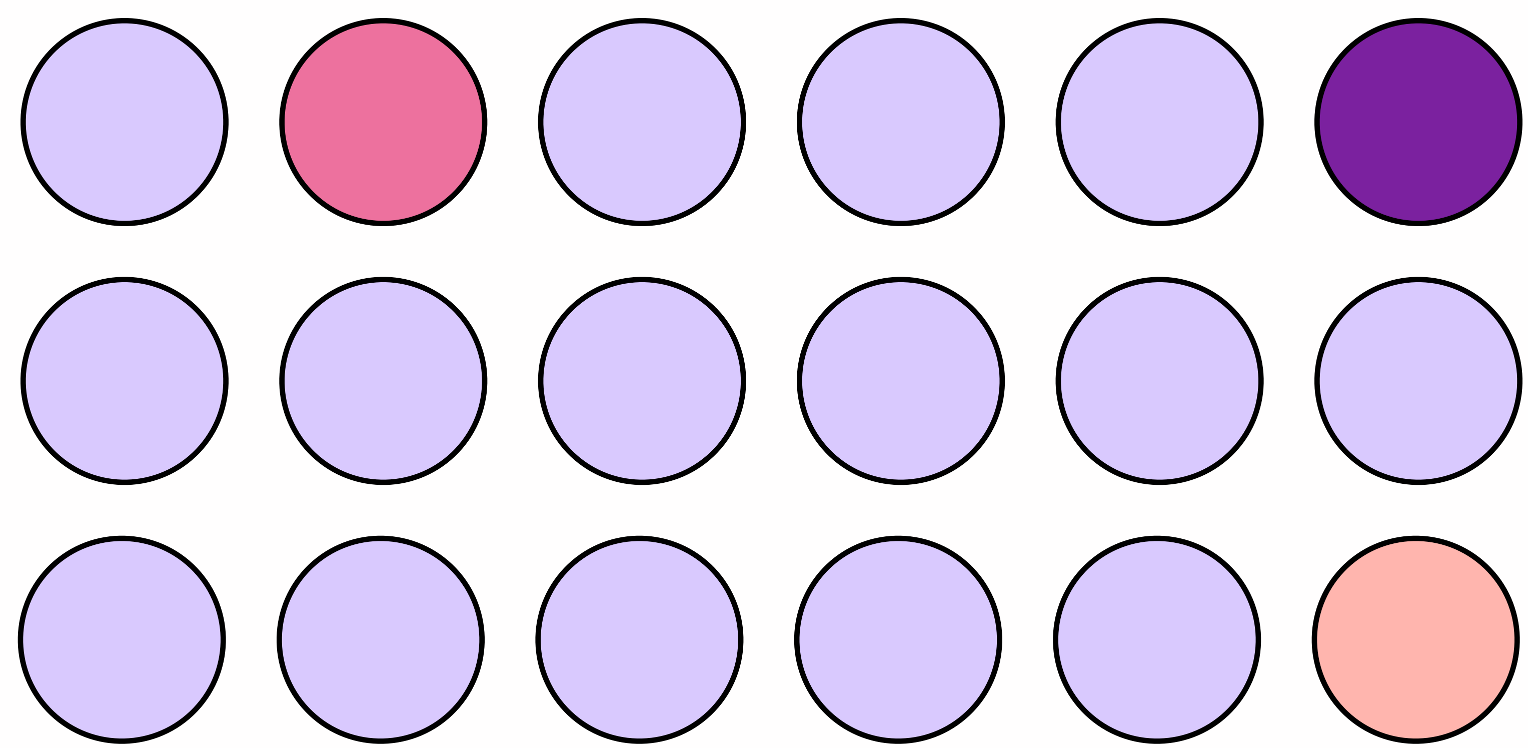



Bag 1
Bag 2
Bag 3
Bag 4
Bag 5
Bag 6
Uniform
3 types
12 tokens
4 types
18 tokens
5 types
18 tokens
3 types
6 tokens
2 types
18 tokens
2 types
6 tokens

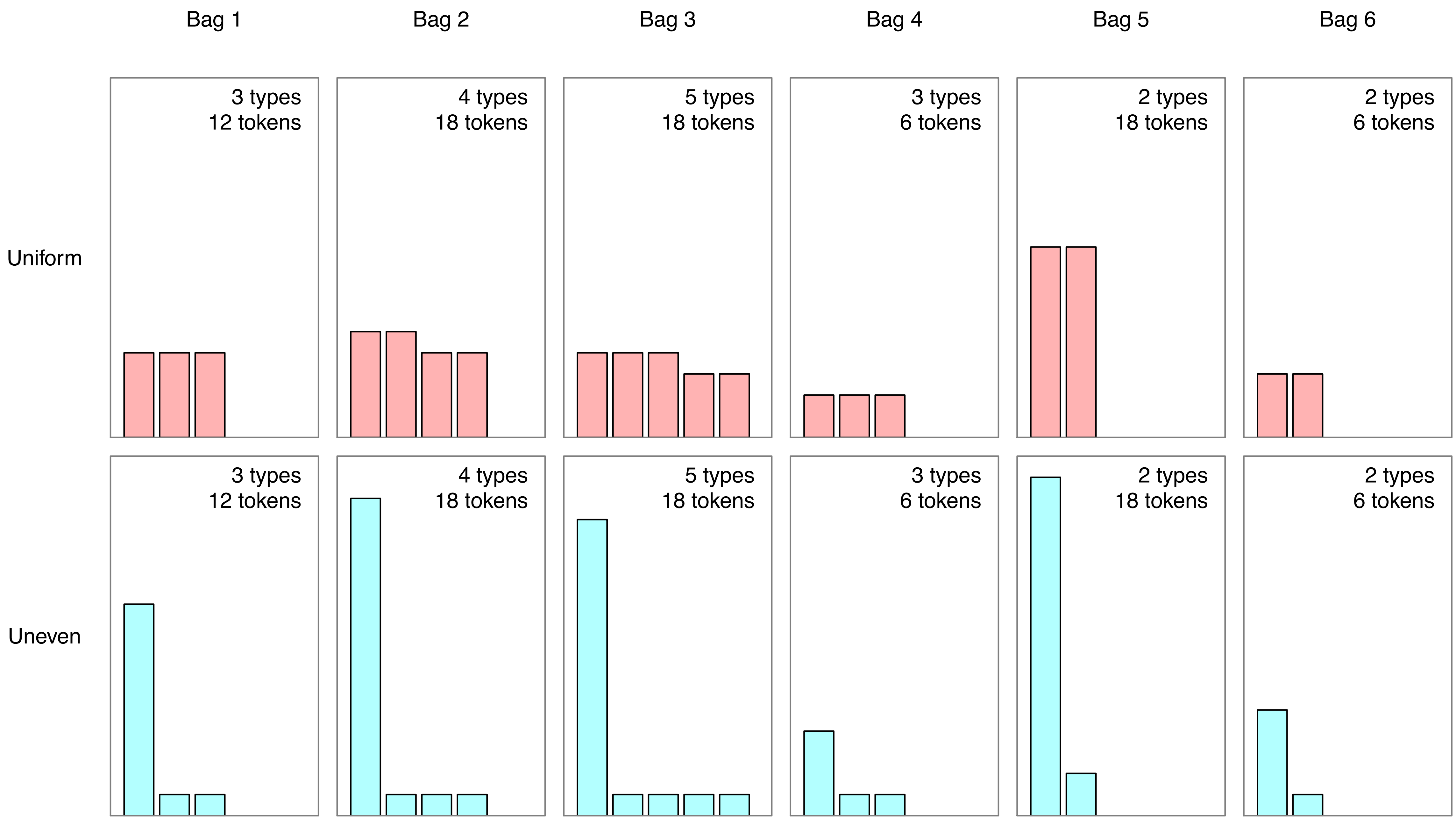


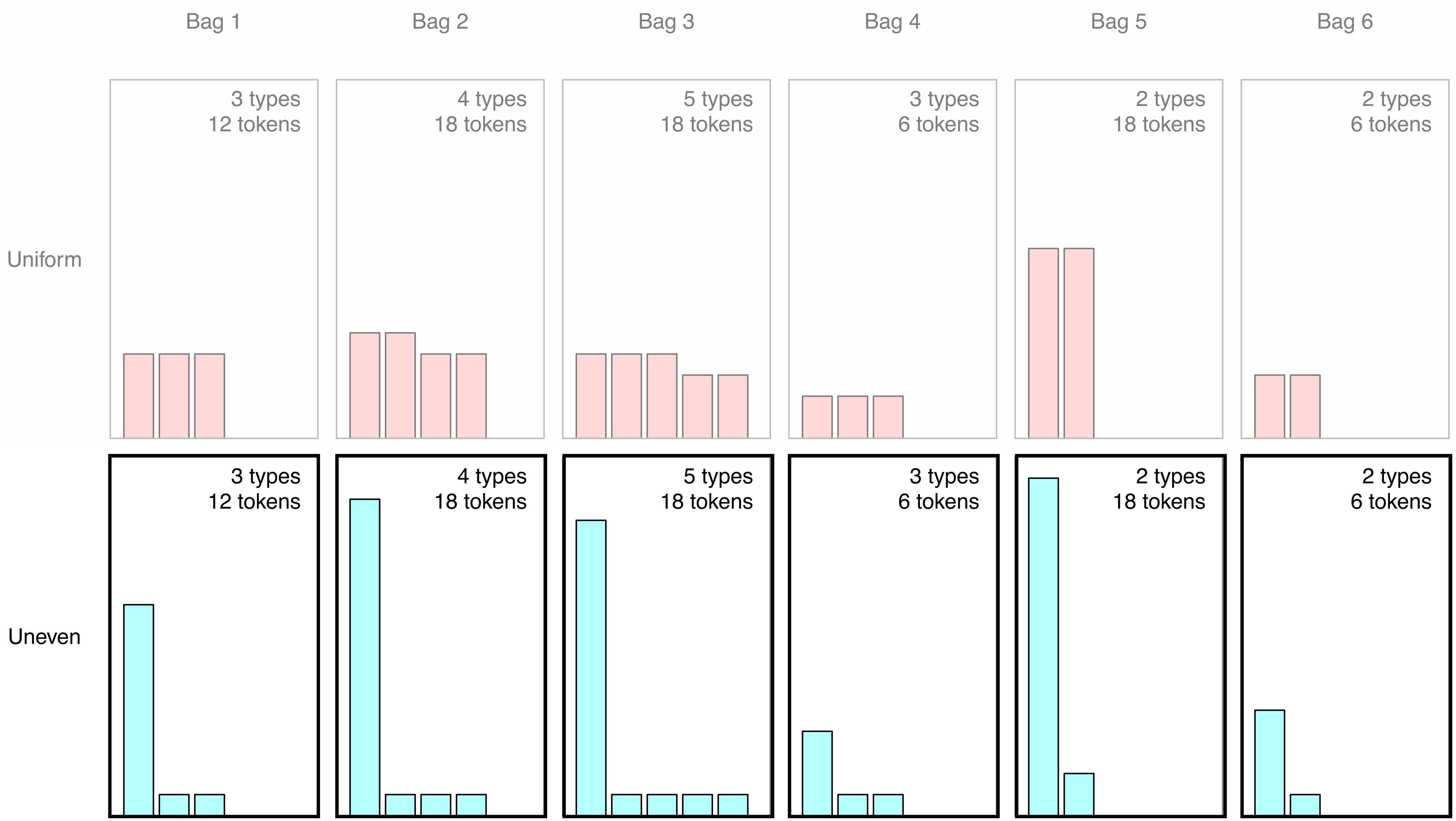
SKEWED condition, bag #1...

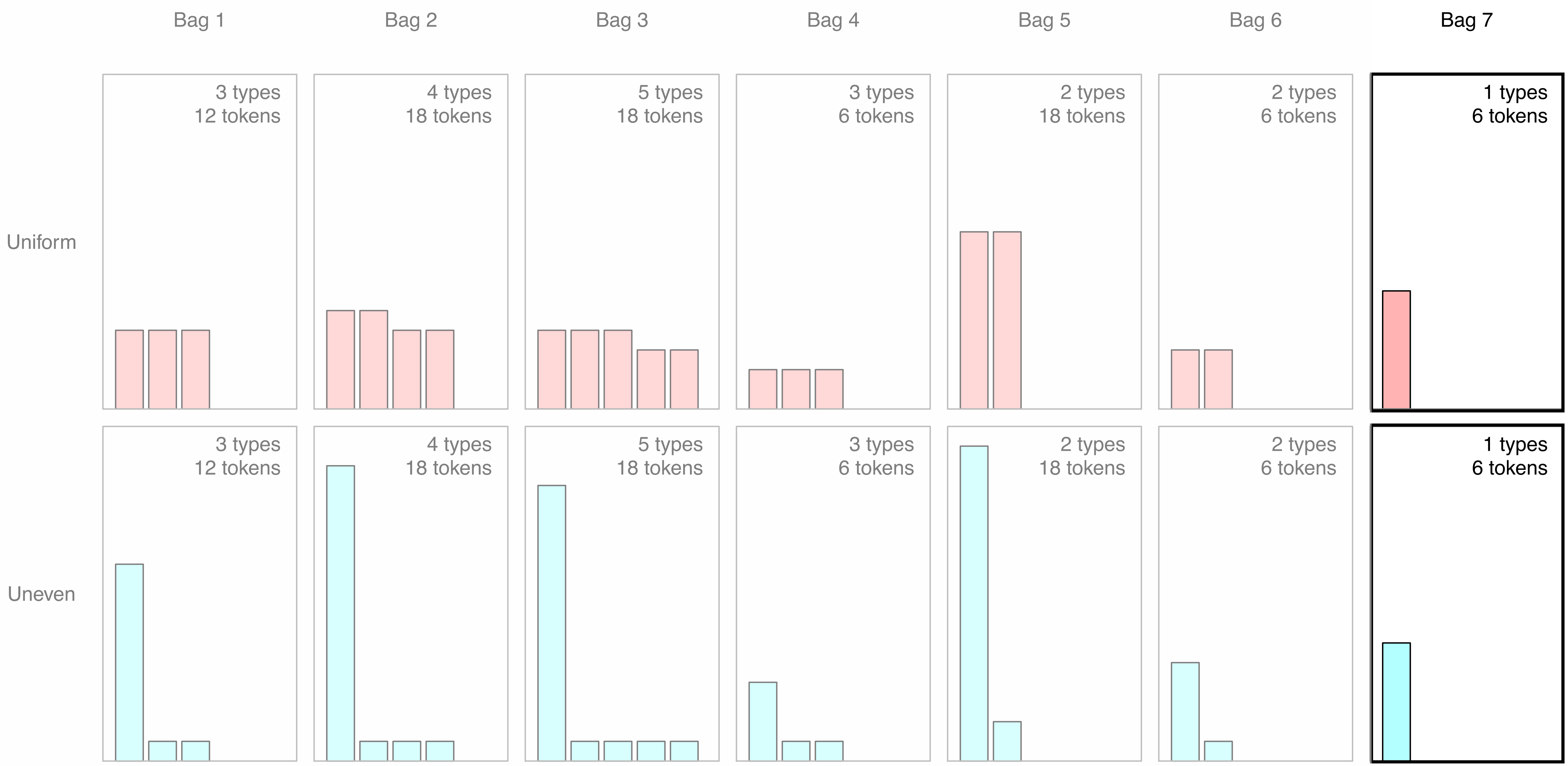
# SKEWED condition, bag #2...



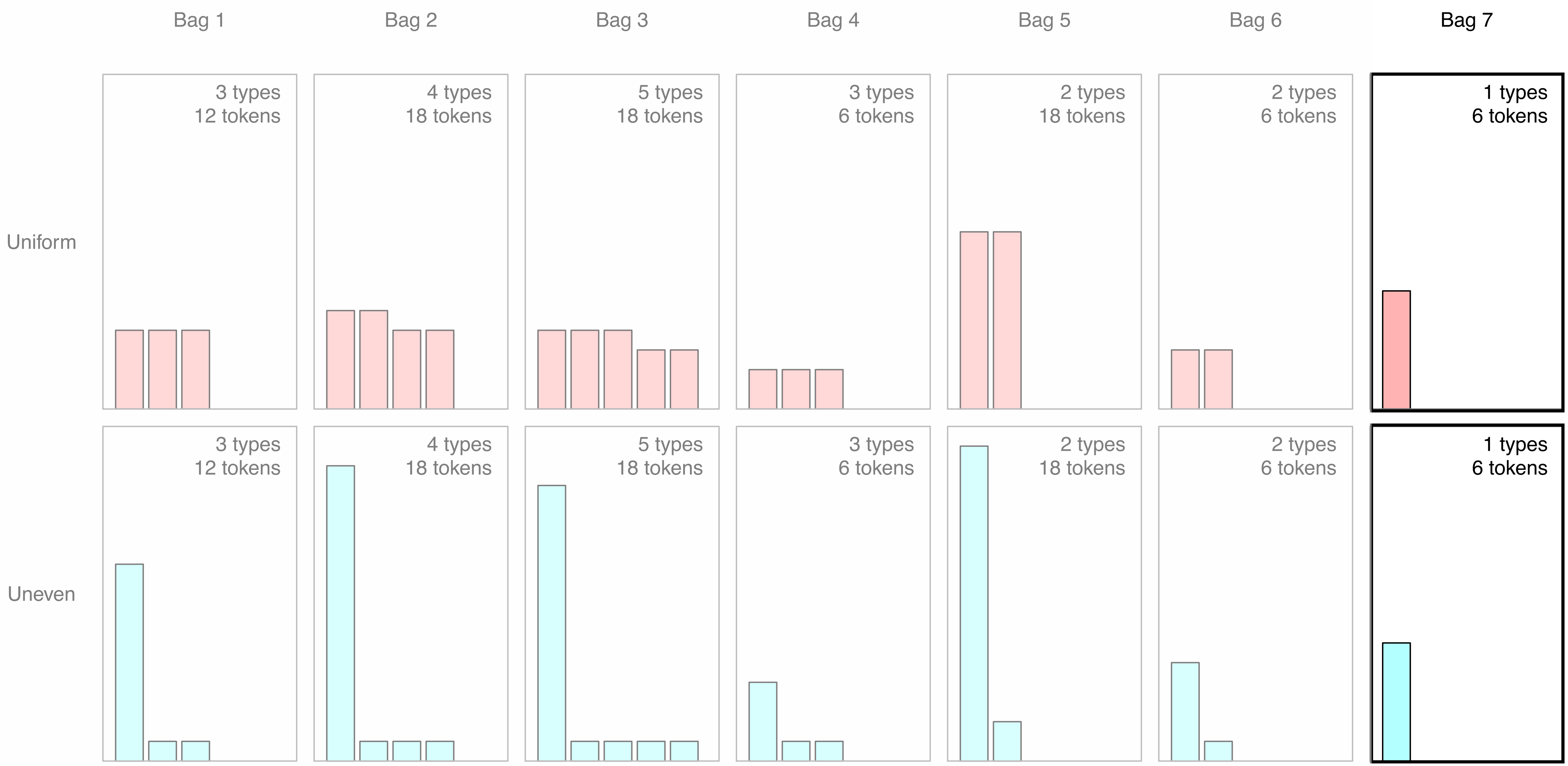

Conditions always matched on number of types and number of tokens




*Prediction*: people are more likely to think the category contains unobserved types in the uneven condition




Include a test trial at the end, identical for both conditions




**Exploratory question:**
Do people learn across bags?
Do people make different responses on bag 7?

# Experiment 1

Task:

- Paper and pencil questionnaire

Participants:

- 44 University of Adelaide students
- Participation as a class exercise
- Included undergraduates and postgraduates

# Experiment 2

Task:

- Task presented on computer
- Same stimuli as Experiment 1
- More detailed instruction set

Participants:

- 57 paid participants (mostly ex-undergrads)
- Paid $10 across multiple bundled experiments

# Experiment 3

Task:

- Run online via Amazon Mechanical Turk
- Intention was to use the same stimuli. Order of bags 1 and 2 was reversed due to “coding” error

Participants:

- 163 US-based Turkers
- Paid $0.50 for 10 min task

# Experiment 4

Task:

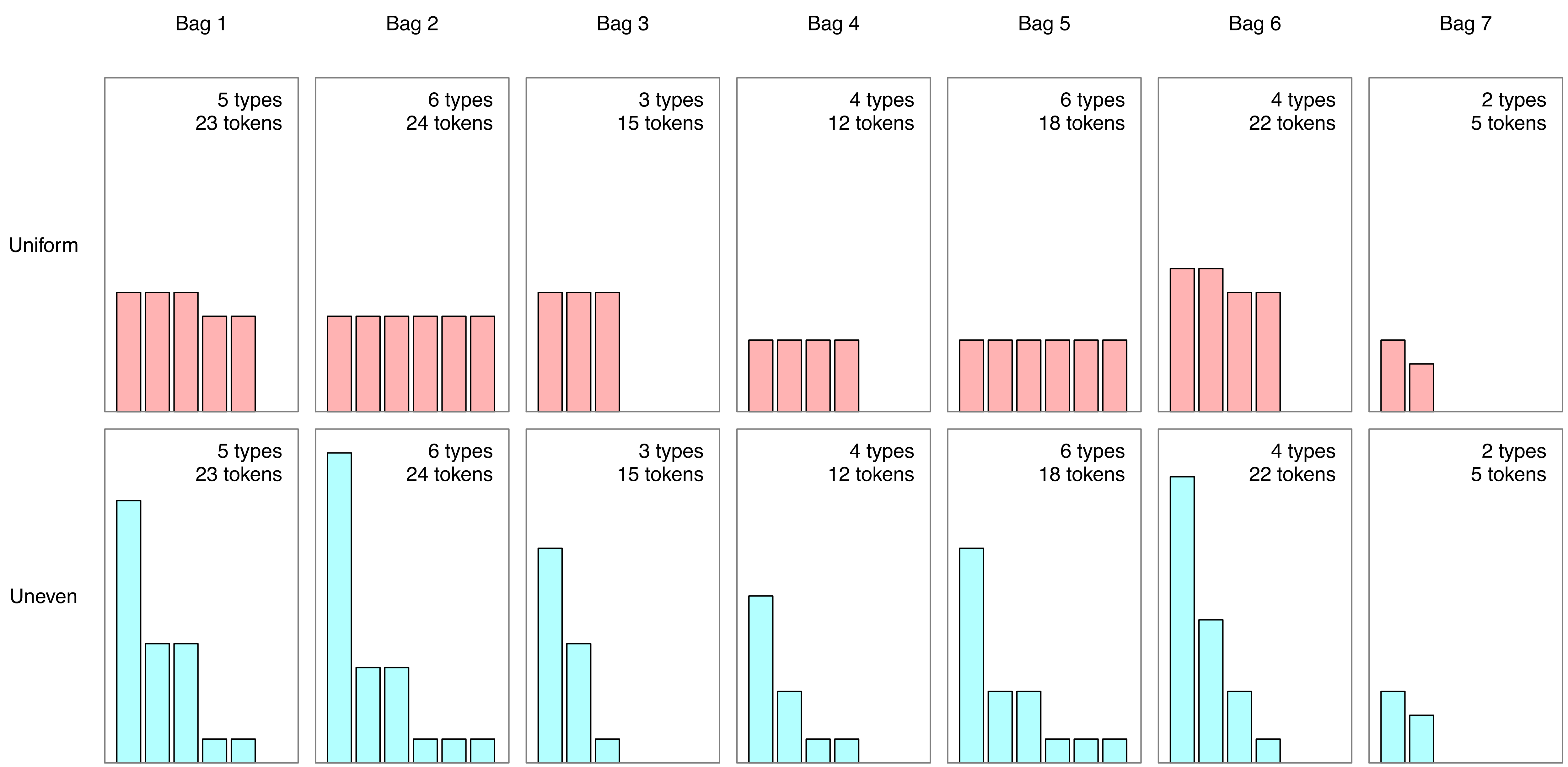
- Run online via Amazon Mechanical Turk
- New stimulus design with more types and more tokens. Check that the results generalise

Participants:

- 142 US-based Turkers
- Paid $0.50 for 10 min task

# Experiment 4

More types, more tokens, less extreme unevenness

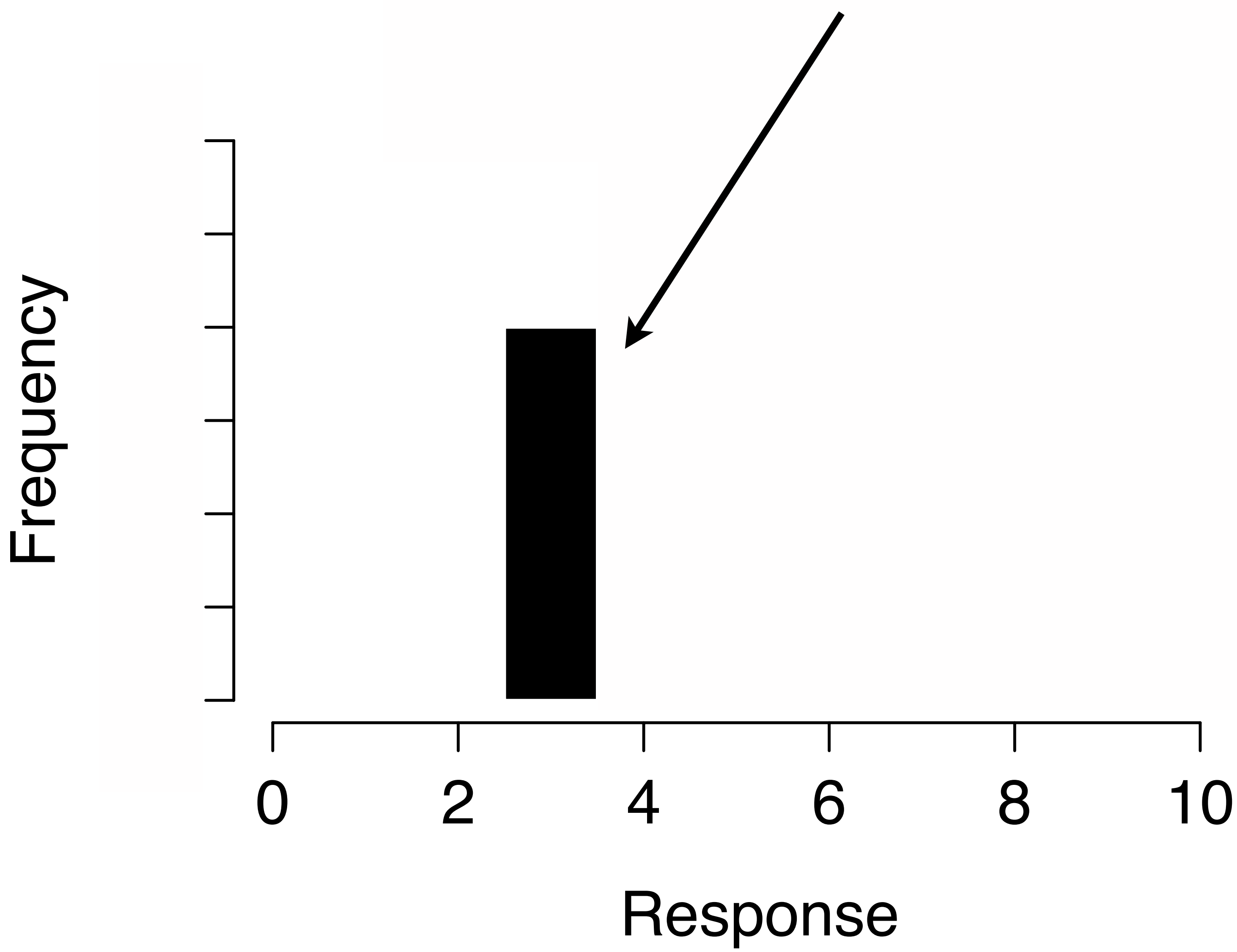


# Experiment 4

How should we measure people's beliefs about what the true number of marble types is?

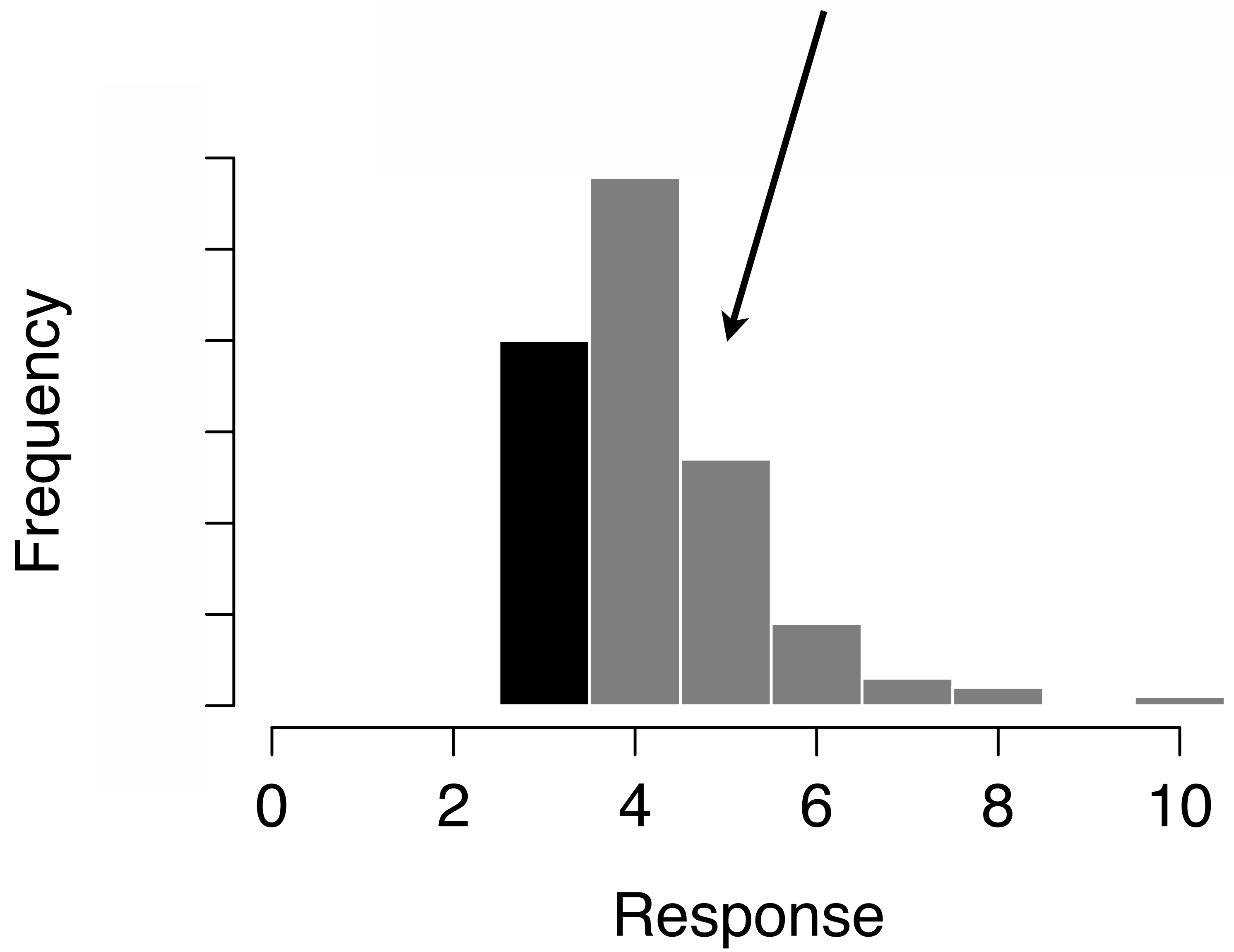
# Experiment 4

A response of “3” implies “no unobserved types”



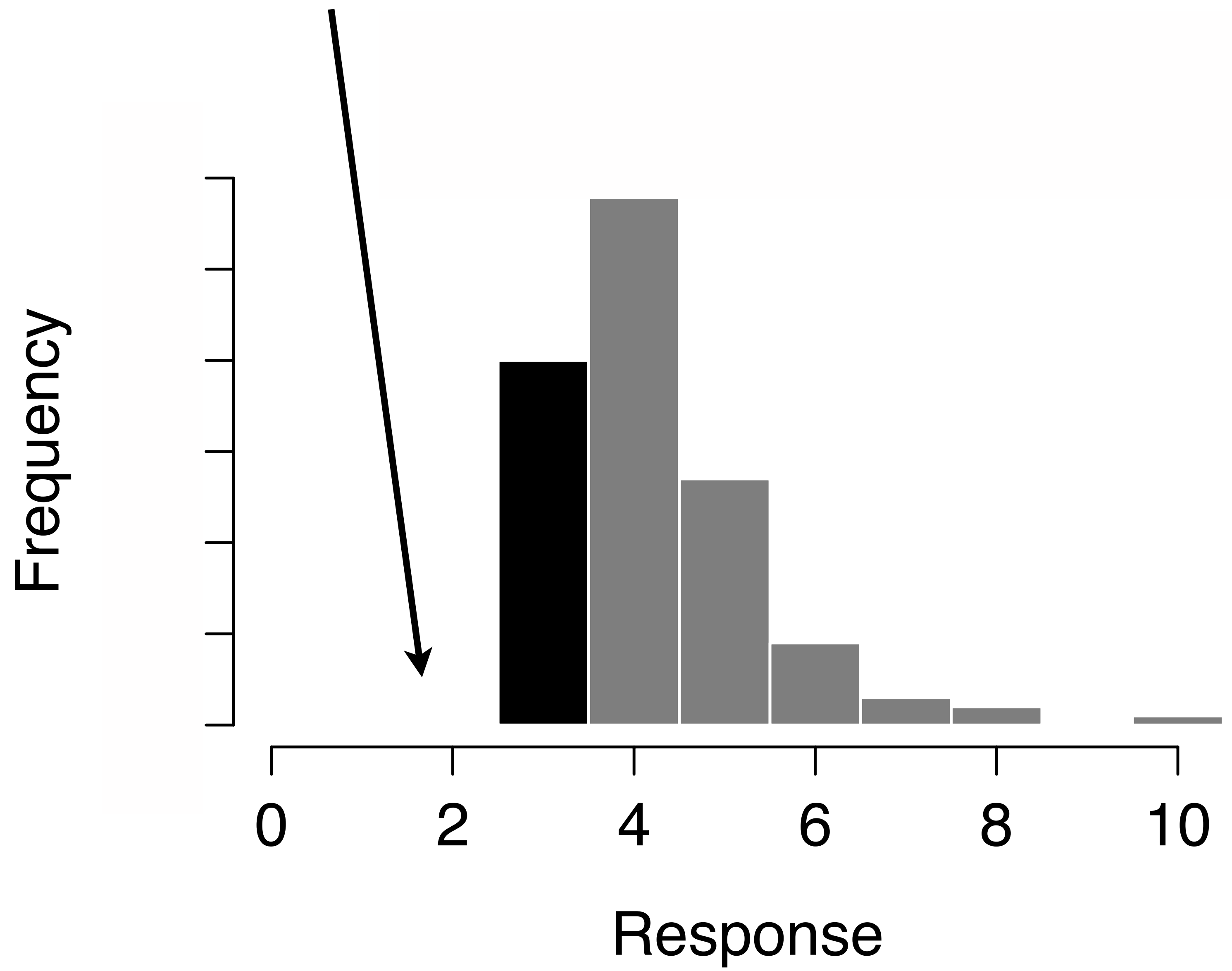
# Experiment 4

A response greater than 3 implies some unobserved types



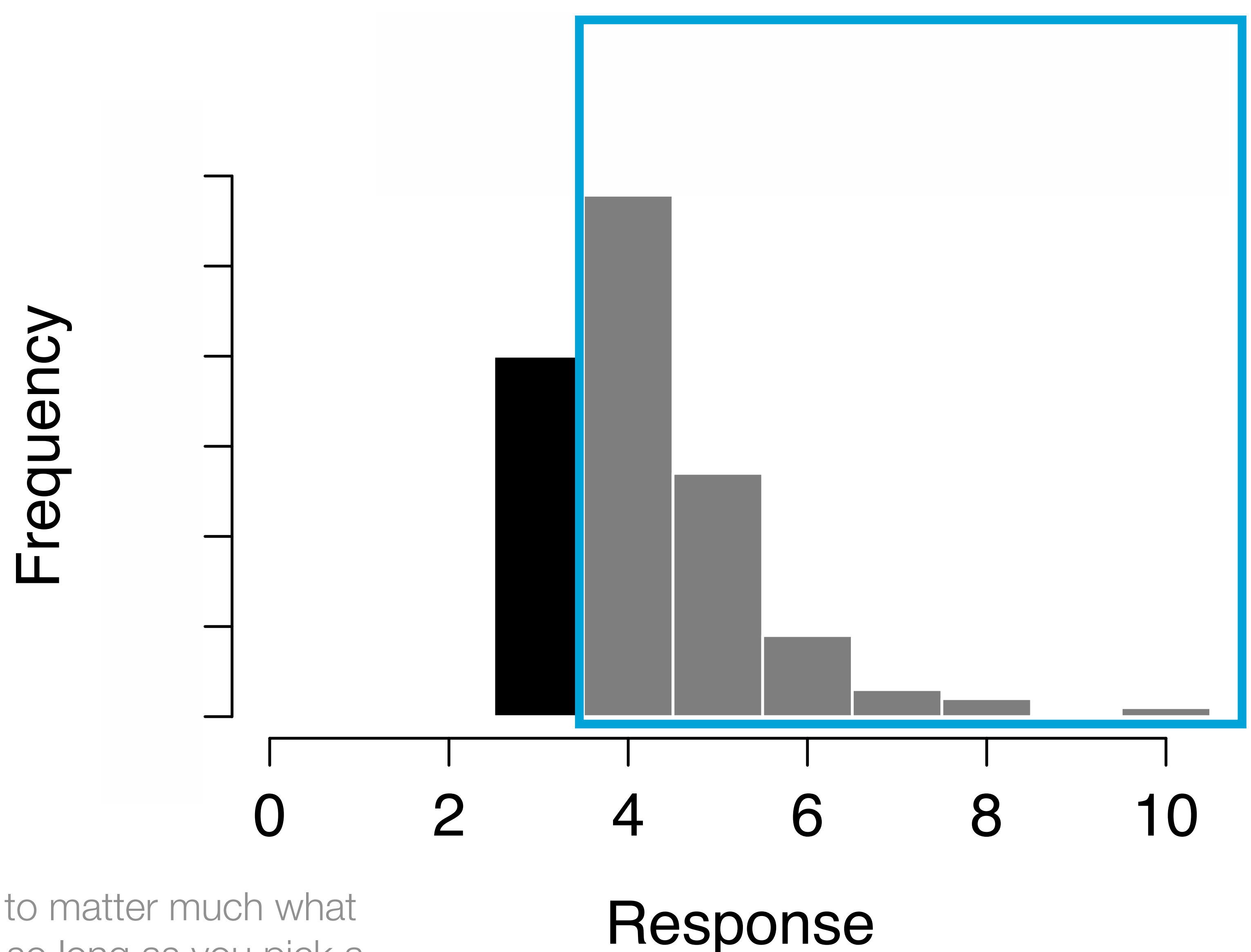
# Experiment 4

Less than three implies that a
mistake was made (rare)



# Experiment 4

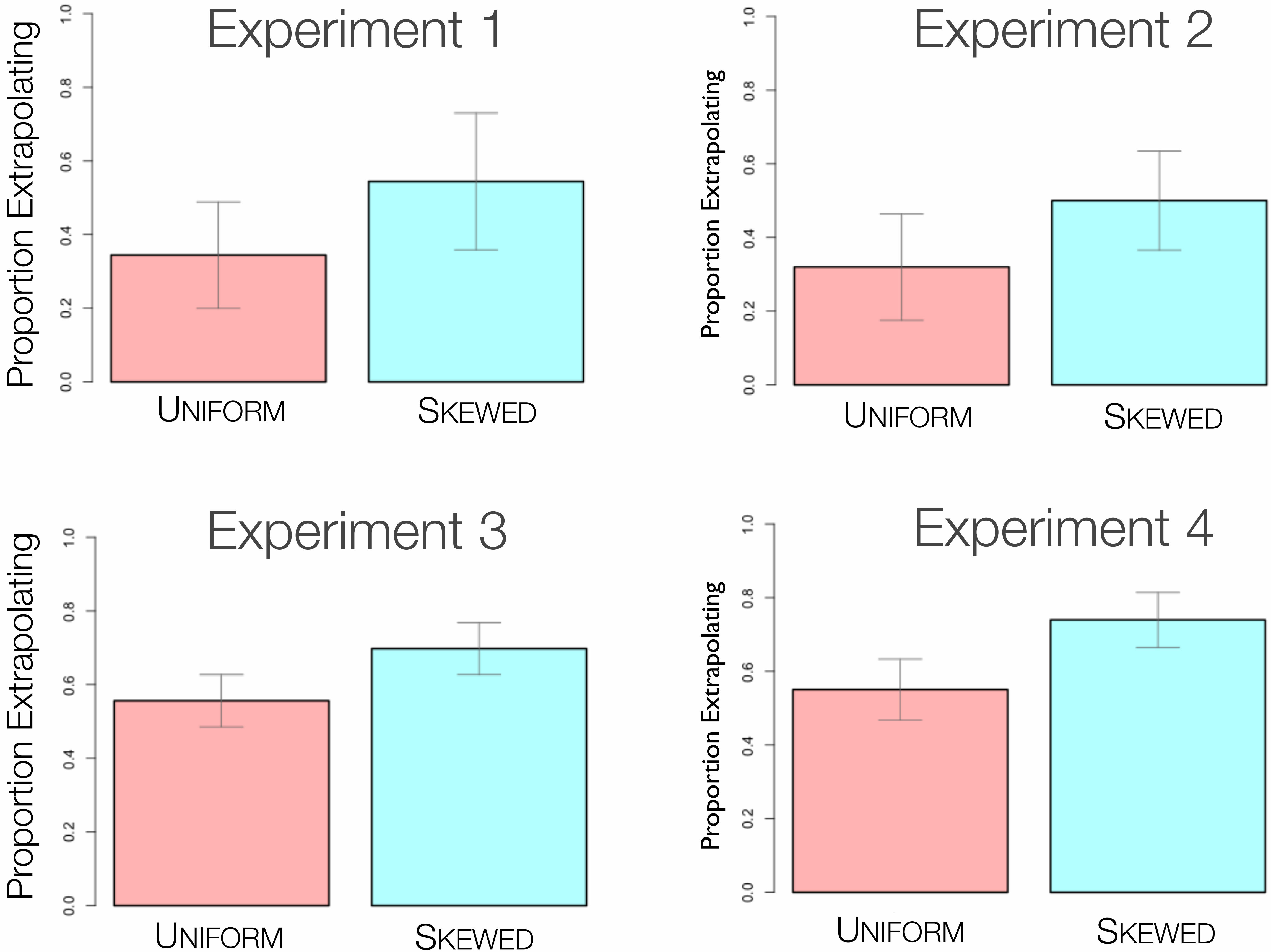
We care about what proportion of people are extrapolating




* It turns out not to matter much what statistic you use, so long as you pick a measure that is robust under skewness

# Analysis of all experiments

Are people more likely to believe that unobserved exemplar types exist when the sample has an uneven frequency distribution?

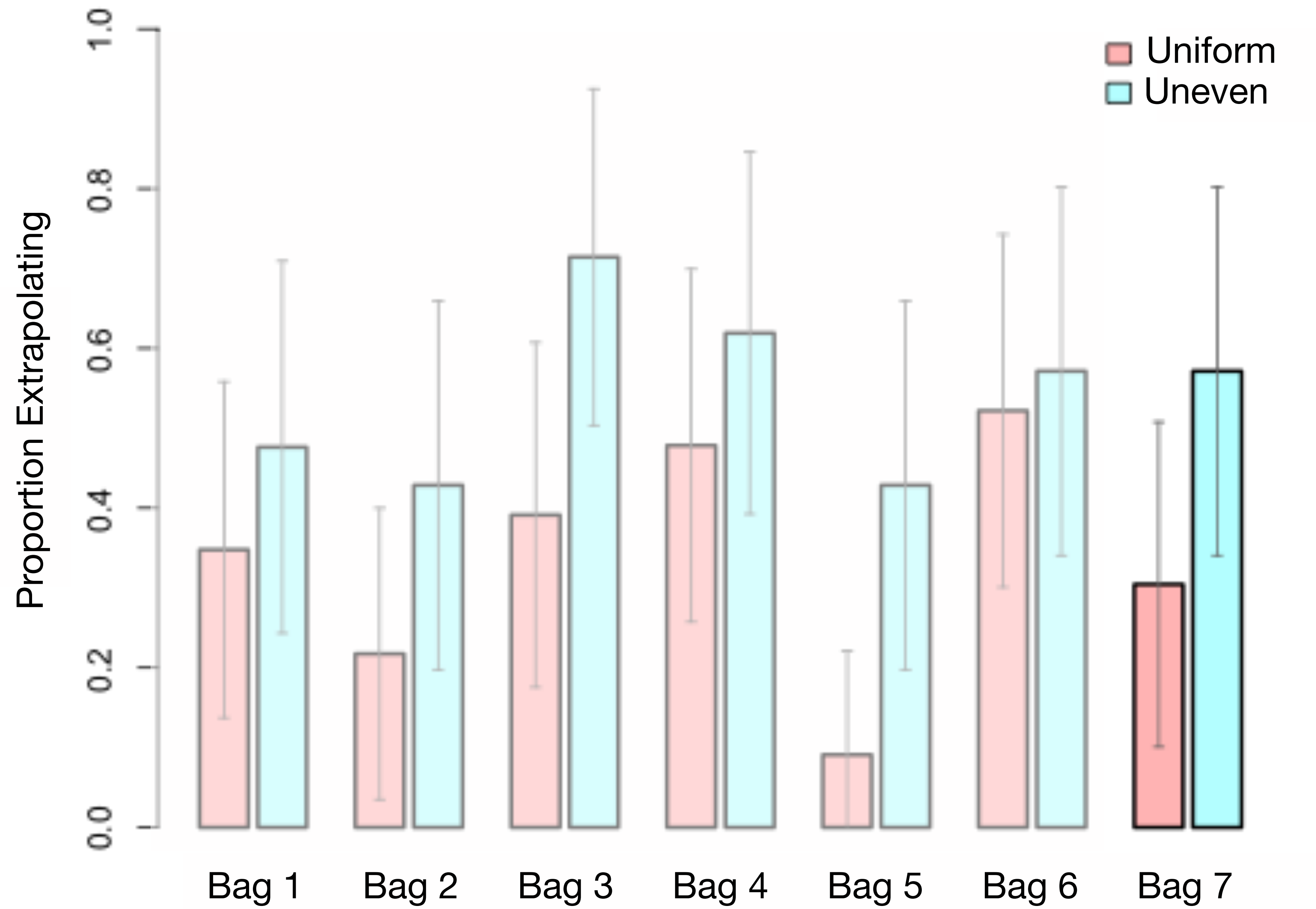


Experiment 1
Proportion Extrapolating
0.0
0.2
0.4
0.6
0.8
1.0
UNIFORM
SKEWED
Experiment 2
Proportion Extrapolating
0.0
0.2
0.4
0.6
0.8
1.0
UNIFORM
SKEWED
Experiment 3
Proportion Extrapolating
0.0
0.2
0.4
0.6
0.8
1.0
UNIFORM
SKEWED
Experiment 4
Proportion Extrapolating
0.0
0.2
0.4
0.6
0.8
1.0
UNIFORM
SKEWED

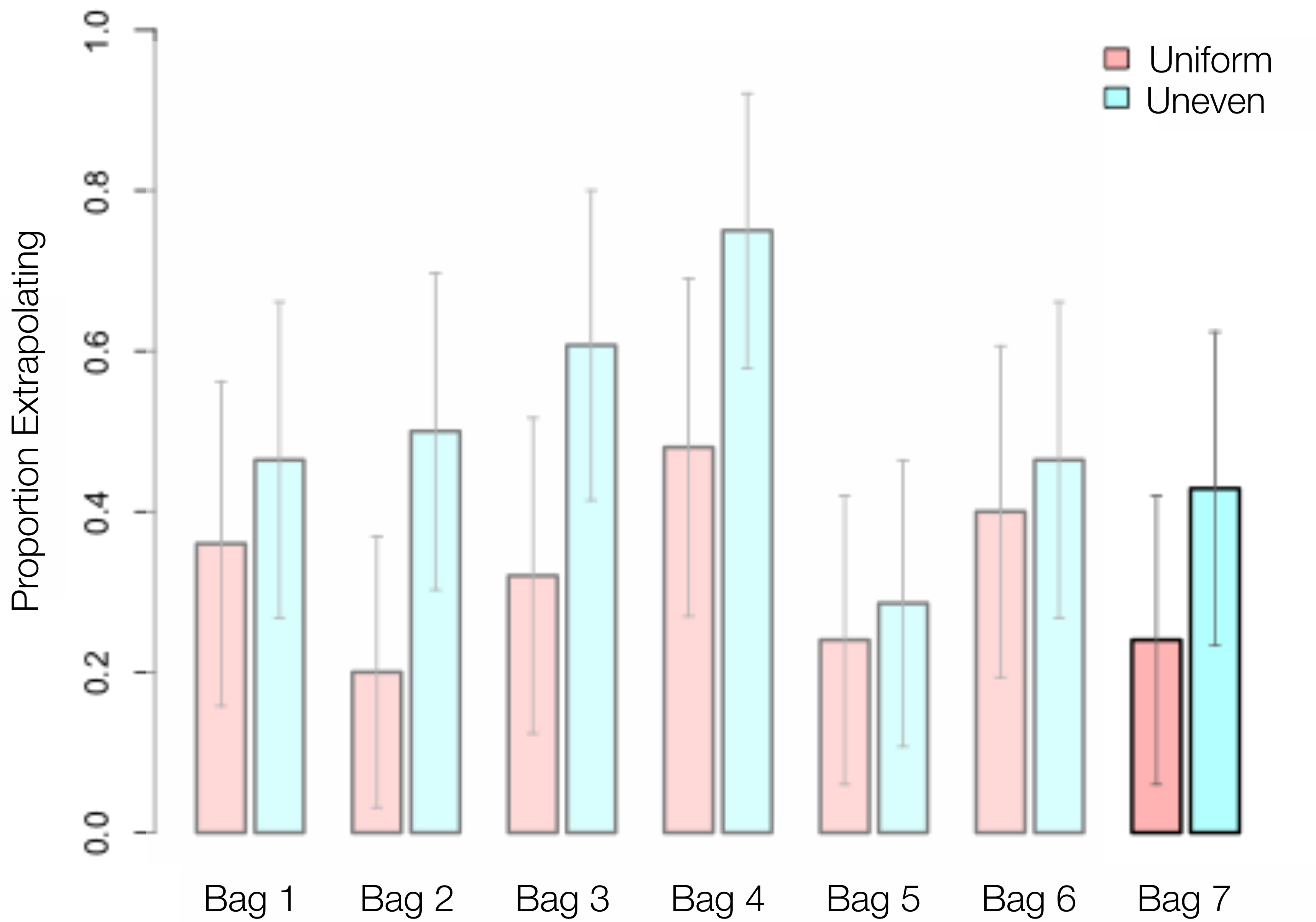
# Analysis of all experiments

Is the effect specific to any particular "bag" or is it robust across all trials in the experiment?

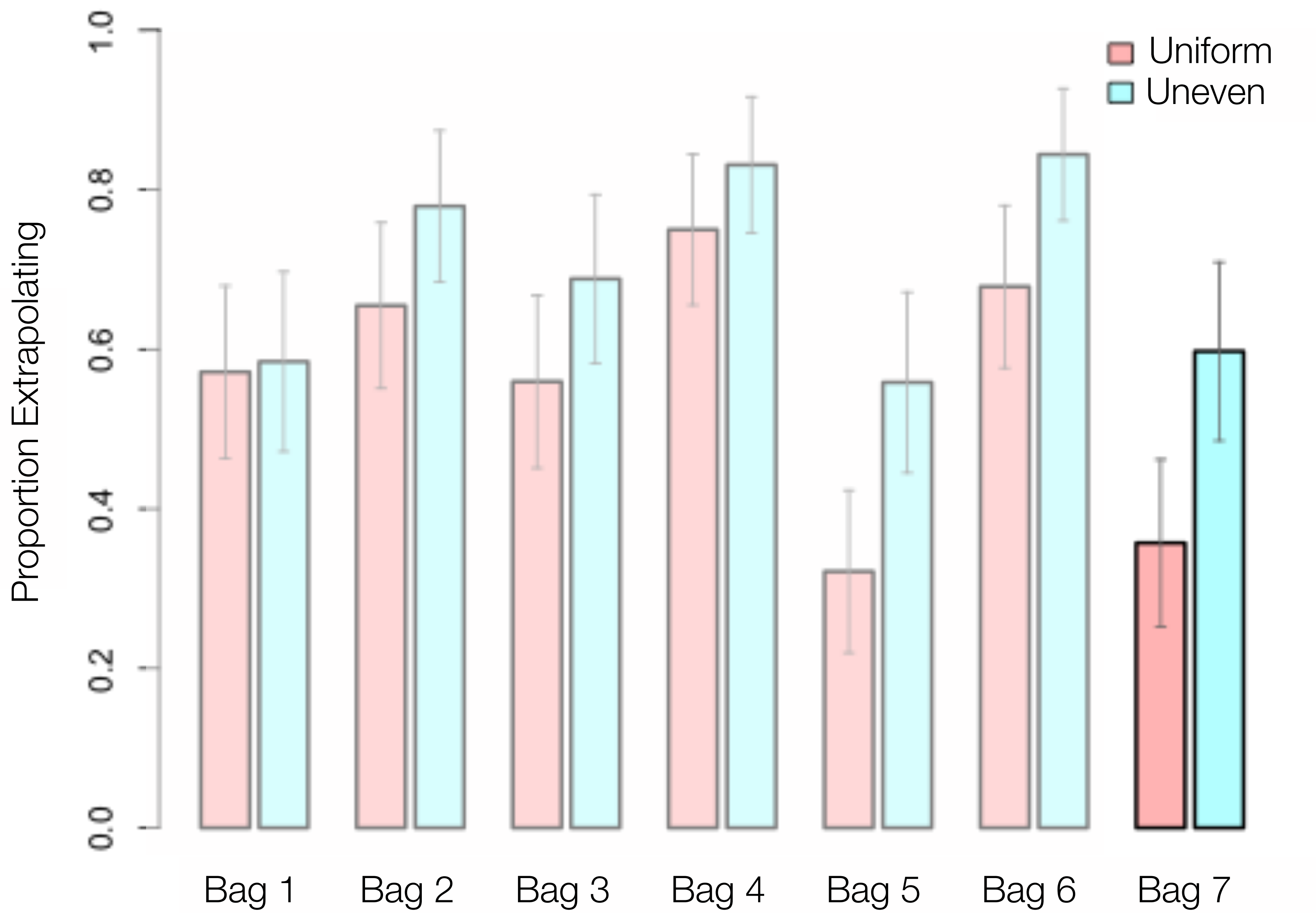
# Experiment 1



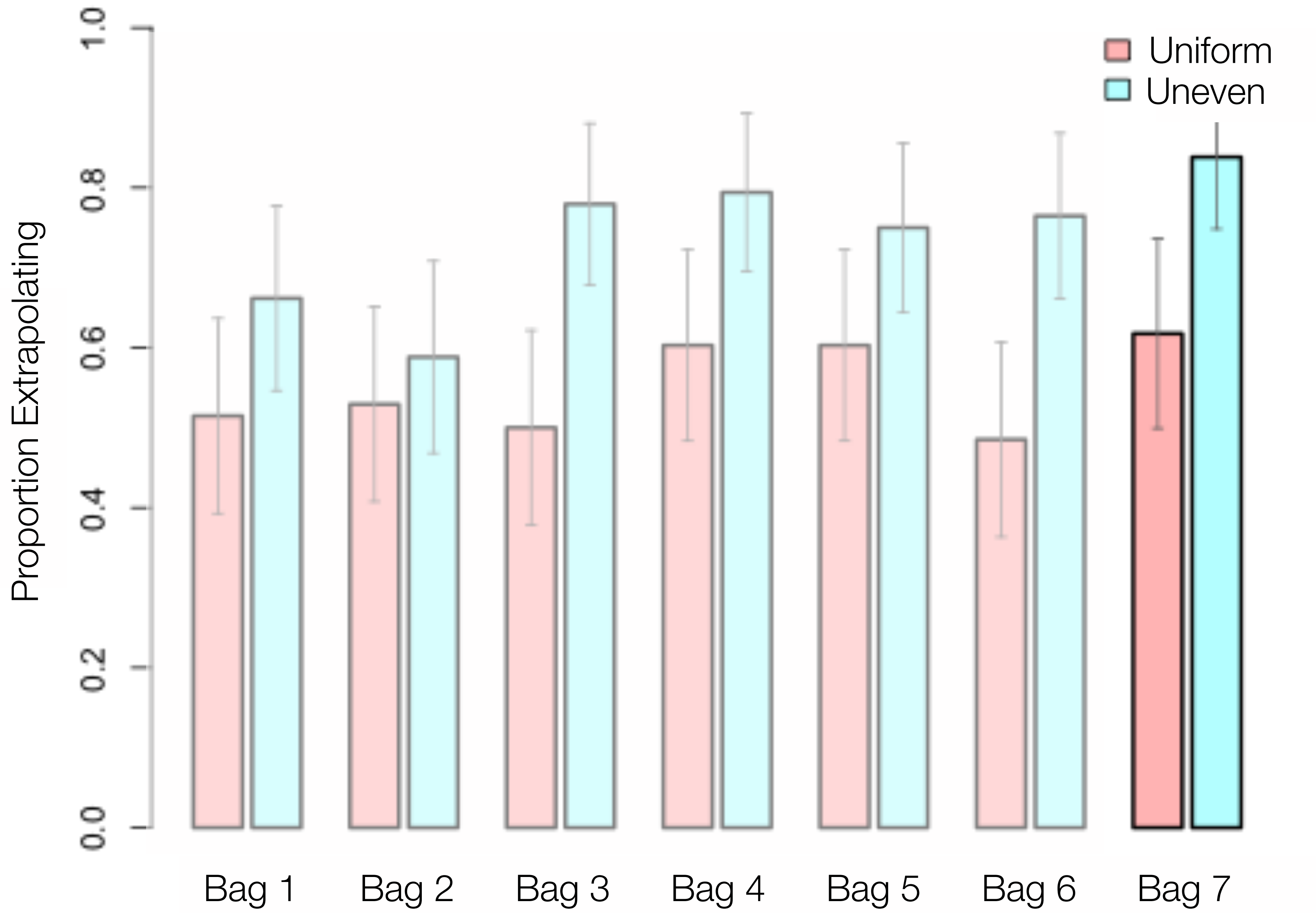
# Experiment 2



# Experiment 3



# Experiment 4



# Experiment 4

This seems to be a real effect.

Can we account for it with the category learning models we have seen so far?

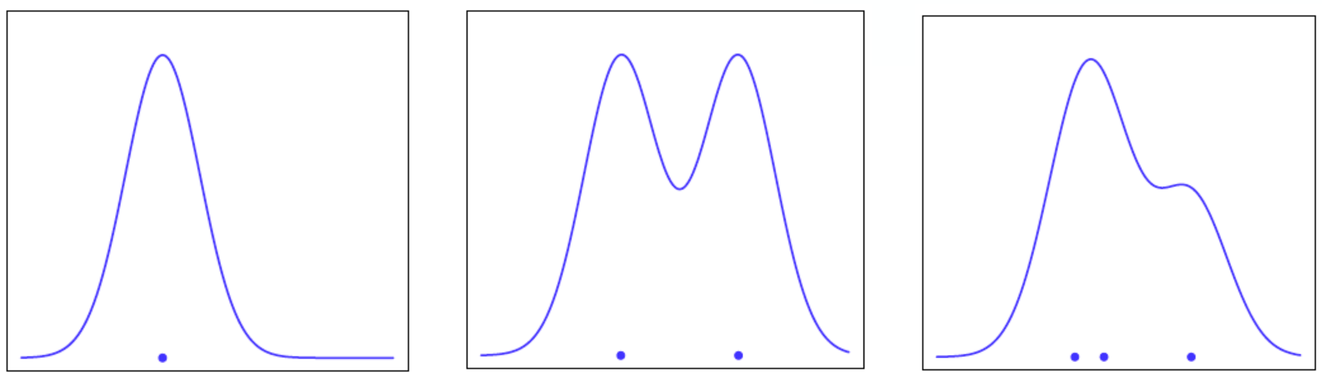
# Lecture outline (next three lectures)

- Last time: Learning about category variability
  - This kind of learning in children and adults
  - A model for this kind of learning
  - Limitations of this model
- ➡ Today: Learning about distributions of categories
  - This kind of learning in adults
  - ➡ Failure of current models
  - A model for this kind of learning
- Lecture 13: Learning about category structure
  - A model for this kind of learning
  - This kind of learning in people

# Standard exemplar model fails

This is a kind of kernel density estimator

kernel density estimator: $P(y|x_1, \ldots, x_n) \propto \sum_{i=1}^{n} K(y - x_i)$



standard exemplar model: $T(y) = \sum_{k=1}^{K} n_k \, S(x_k, y)$

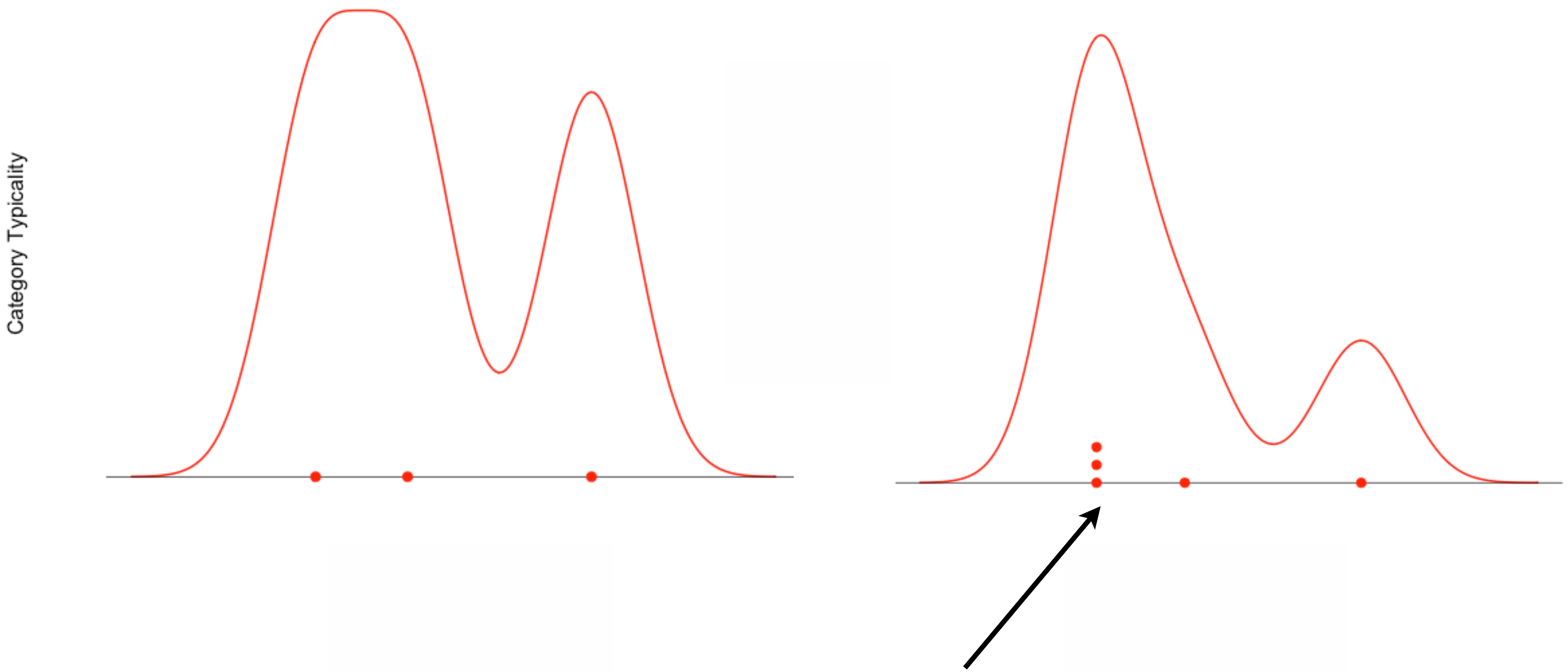
typicality

frequency

similarity

# Standard exemplar model fails

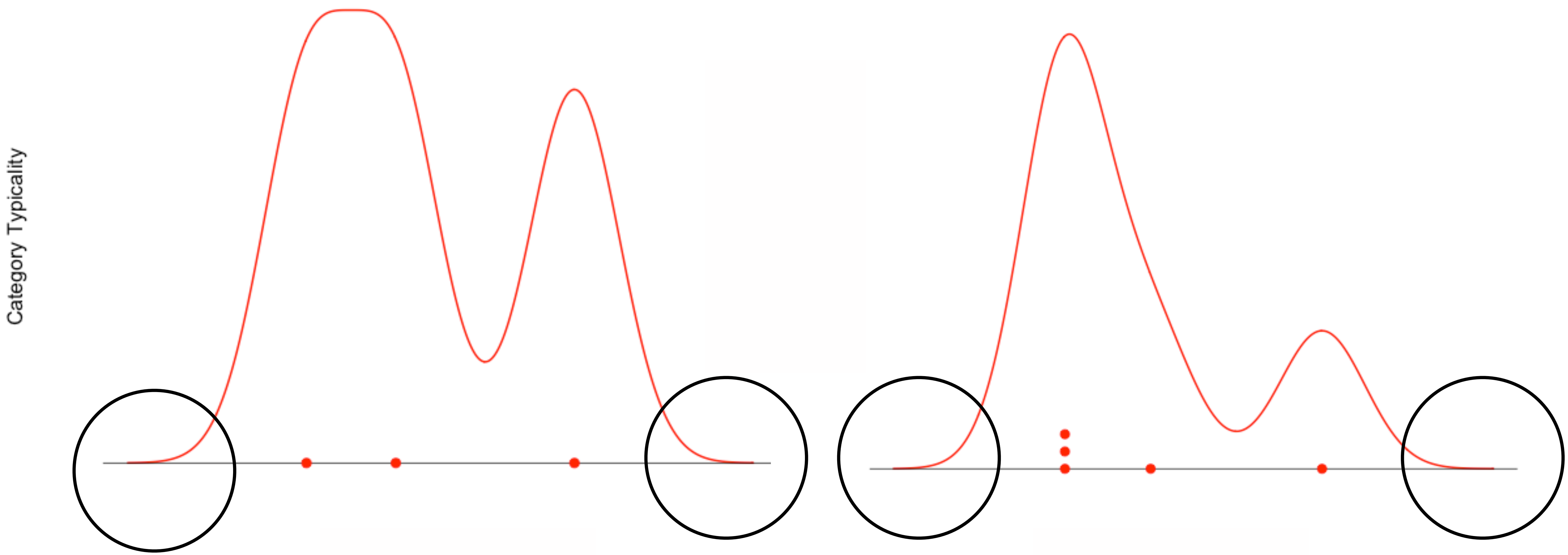
There *is* an effect of frequency, but it's carried by similarity




Increasing exemplar frequency makes
that exemplar and those similar to it more typical

# Standard exemplar model fails

There *is* an effect of frequency, but it's carried by similarity



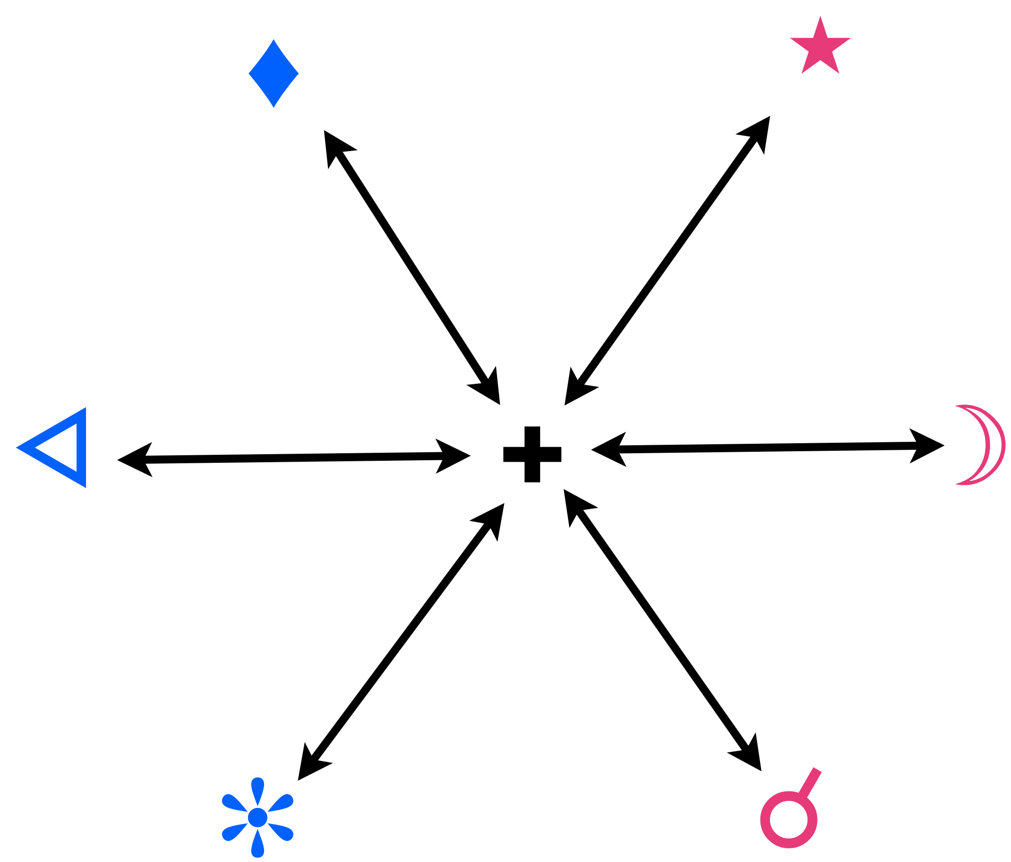

No similarity effects?
Then no frequency effects

# Standard exemplar model fails

We can illustrate what happens in the alien alphabet situation

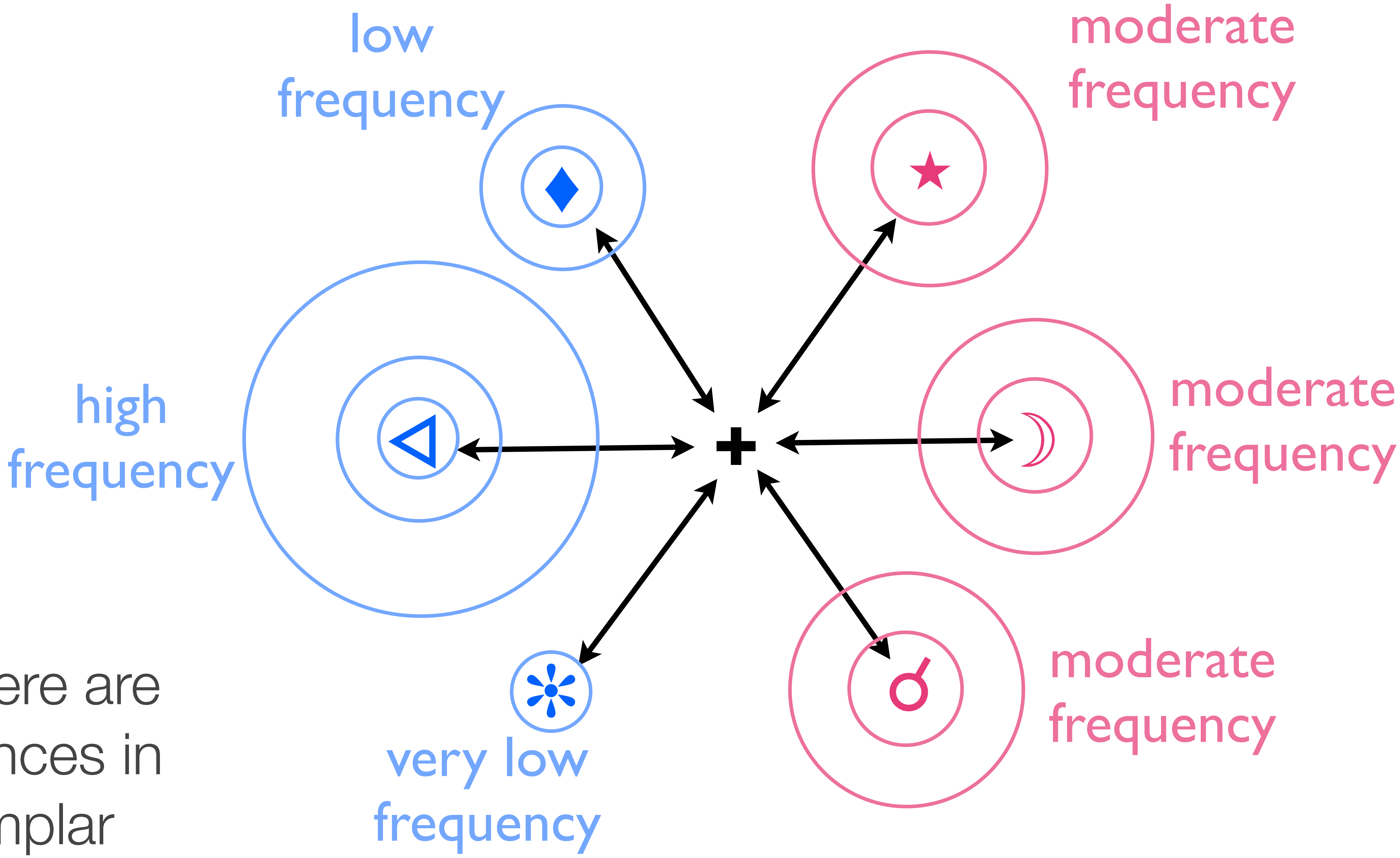
No similarity differences:

Training exemplars are roughly equally distant from the target item



# Standard exemplar model fails

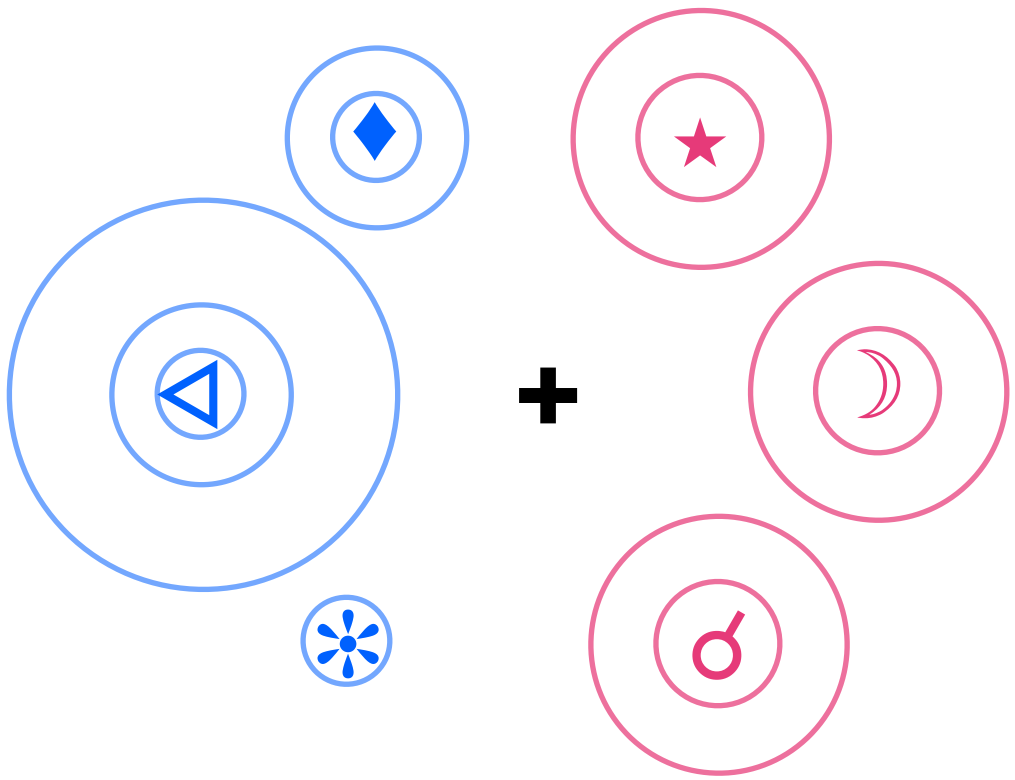
We can illustrate what happens in the alien alphabet situation




But there are differences in exemplar frequency

# Standard exemplar model fails

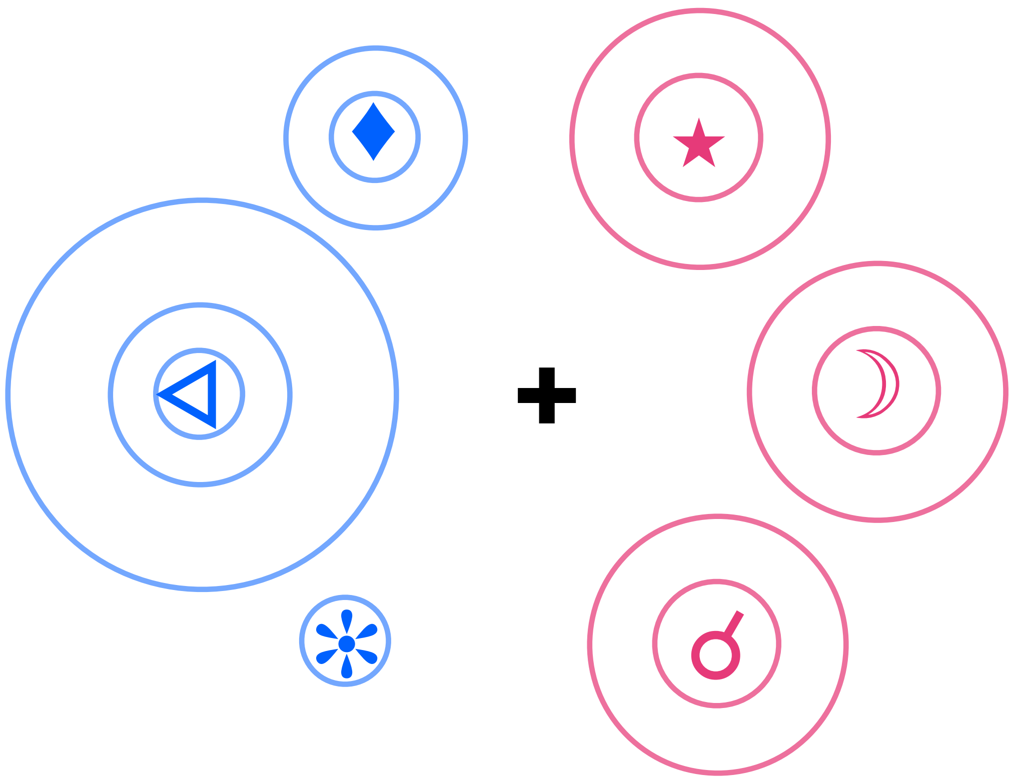
How does an exemplar model explain the alien alphabet effect?



$$T(y) = \sum_{k=1}^{K} n_k \, S(x_k, y)$$

# Standard exemplar model fails

How does an exemplar model explain the alien alphabet effect?

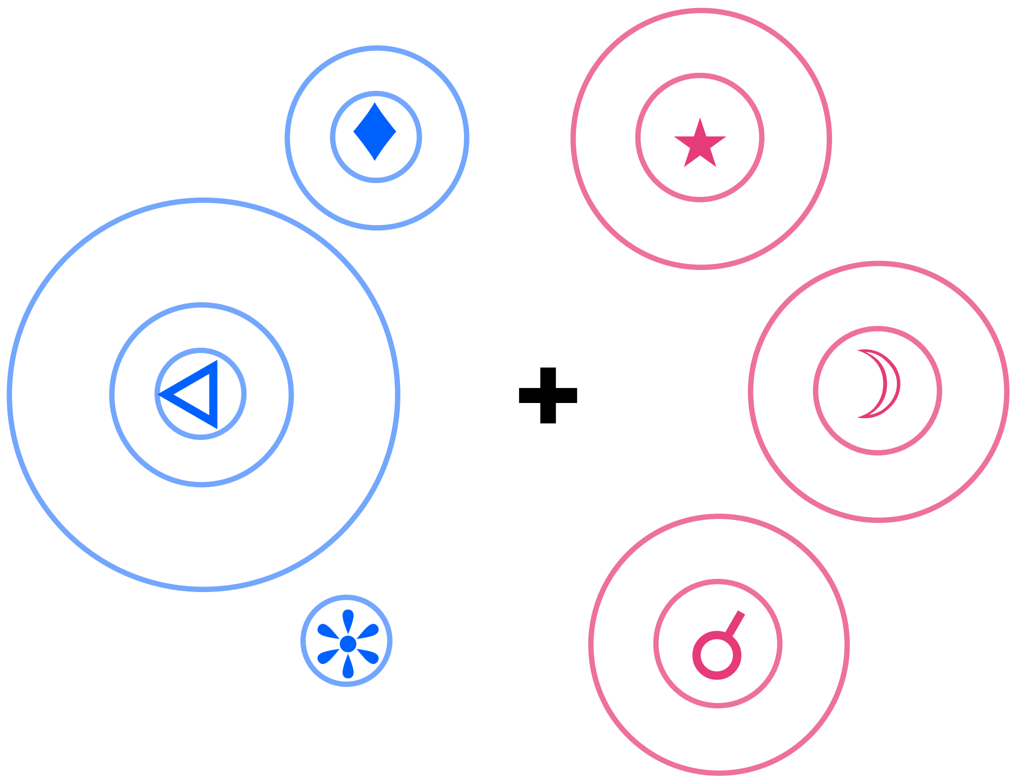


$$T(y) = \sum_{k=1}^{K} n_k \, S(x_k, y)$$

The exemplar frequency term is the one that needs to have an influence

# Standard exemplar model fails

It can't explain the effect.



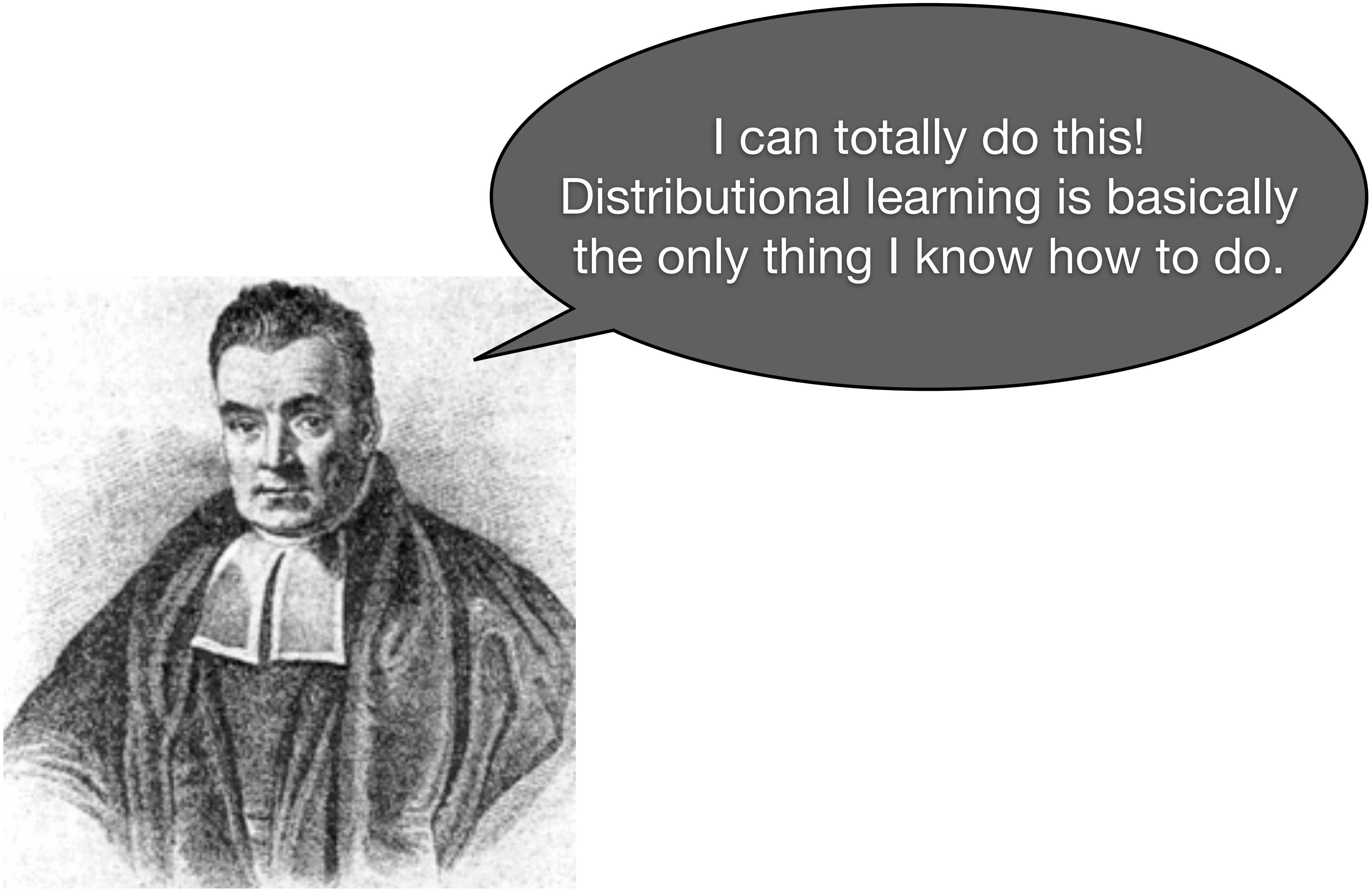
$$T(y) = \sum_{k=1}^{K} n_k \, S(x_k, y)$$

$$= \sum_{k=1}^{K} n_k \, s$$

$$= \; Ns$$

Unless there is a similarity effect in play, category typicality depends on the total number of instances N, but does *not* depend on the frequencies of specific exemplars, nk

# Maybe the RMC (rational model)?



# Maybe the RMC (rational model)?

Nope. Just look at the equations!

$$P(\text{new cluster}) = \frac{\alpha}{\alpha + N}$$

The probability of a novel type depends on the *total number* of instances $N$, and a free parameter $\alpha$. Again, there's *no effect of individual exemplar frequency,* $n_k$

As before, I'm assuming there's no similarity effect going on.

# Maybe the RMC (rational model)?

Nope. Just look at the equations!

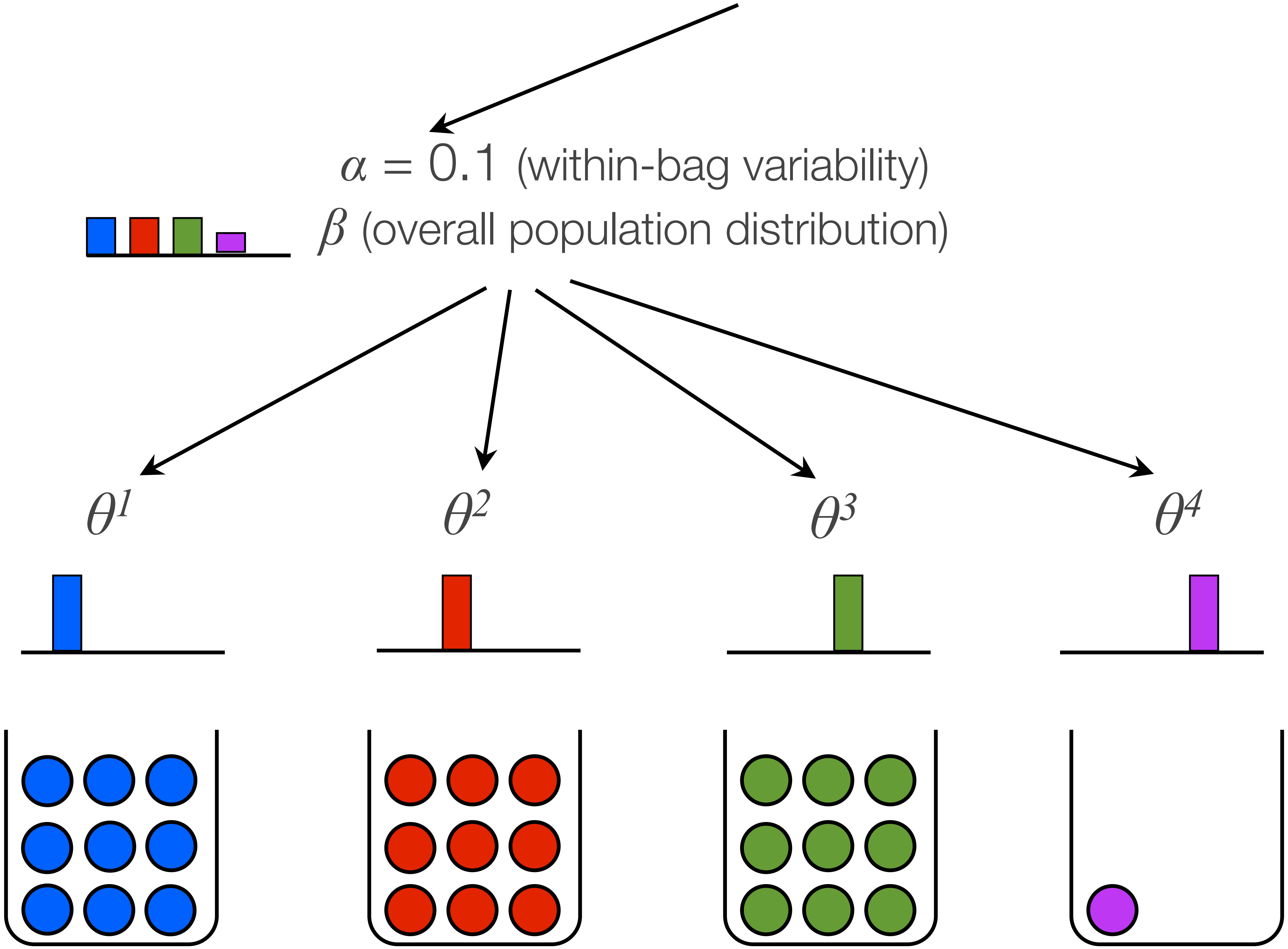
$$P(\text{new cluster}) = \frac{\alpha}{\alpha + N}$$

The probability of a novel type depends on the *total number* of instances $N$, and a free parameter $\alpha$. Again, there's *no effect of individual exemplar frequency,* $n_k$

As before, I'm assuming there's no similarity effect going on.

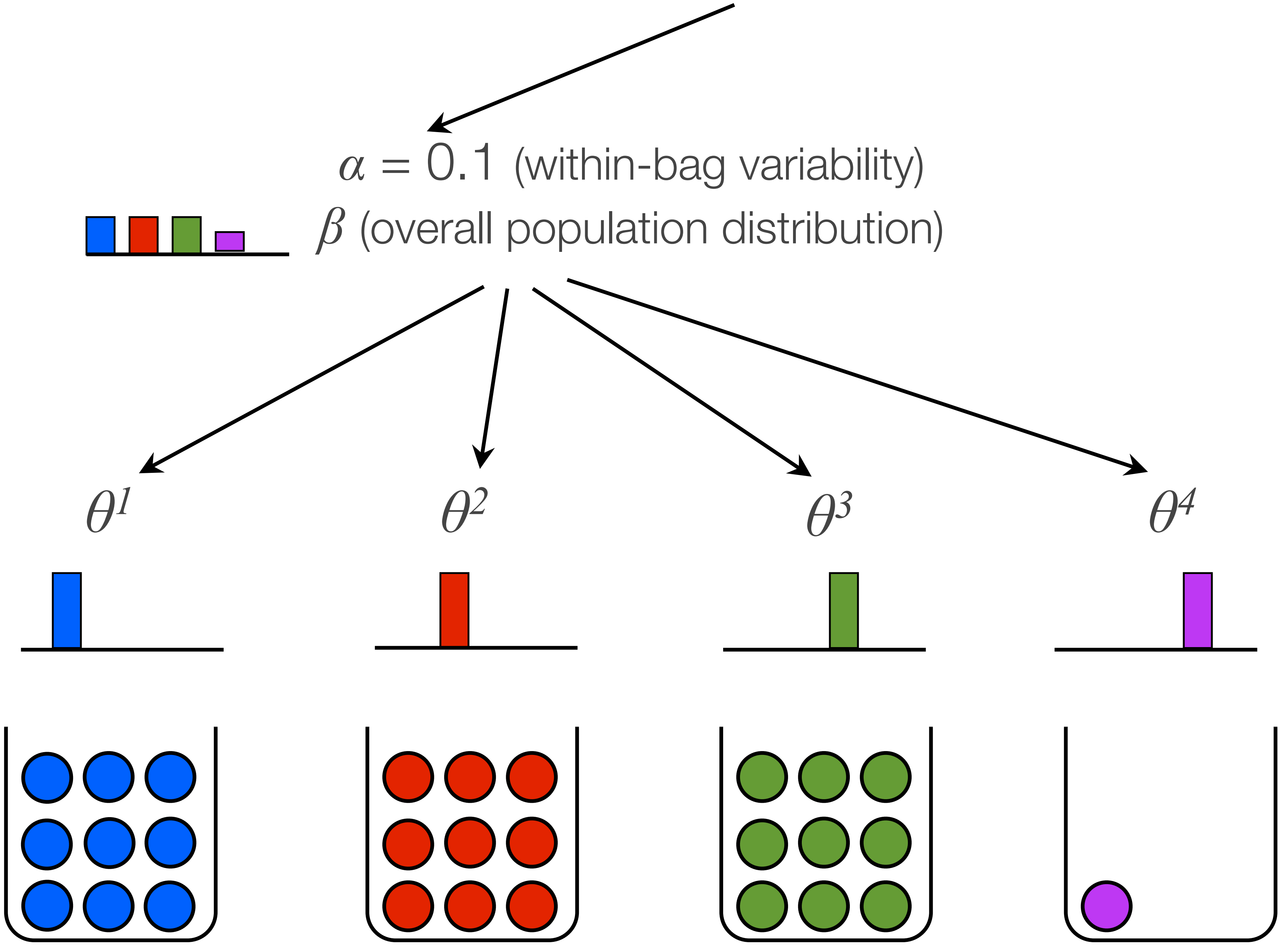
# How about the overhypothesis model?

It seems like learning about within-bag variability would certainly help...



# How about the overhypothesis model?

It seems like learning about within-bag variability would certainly help...

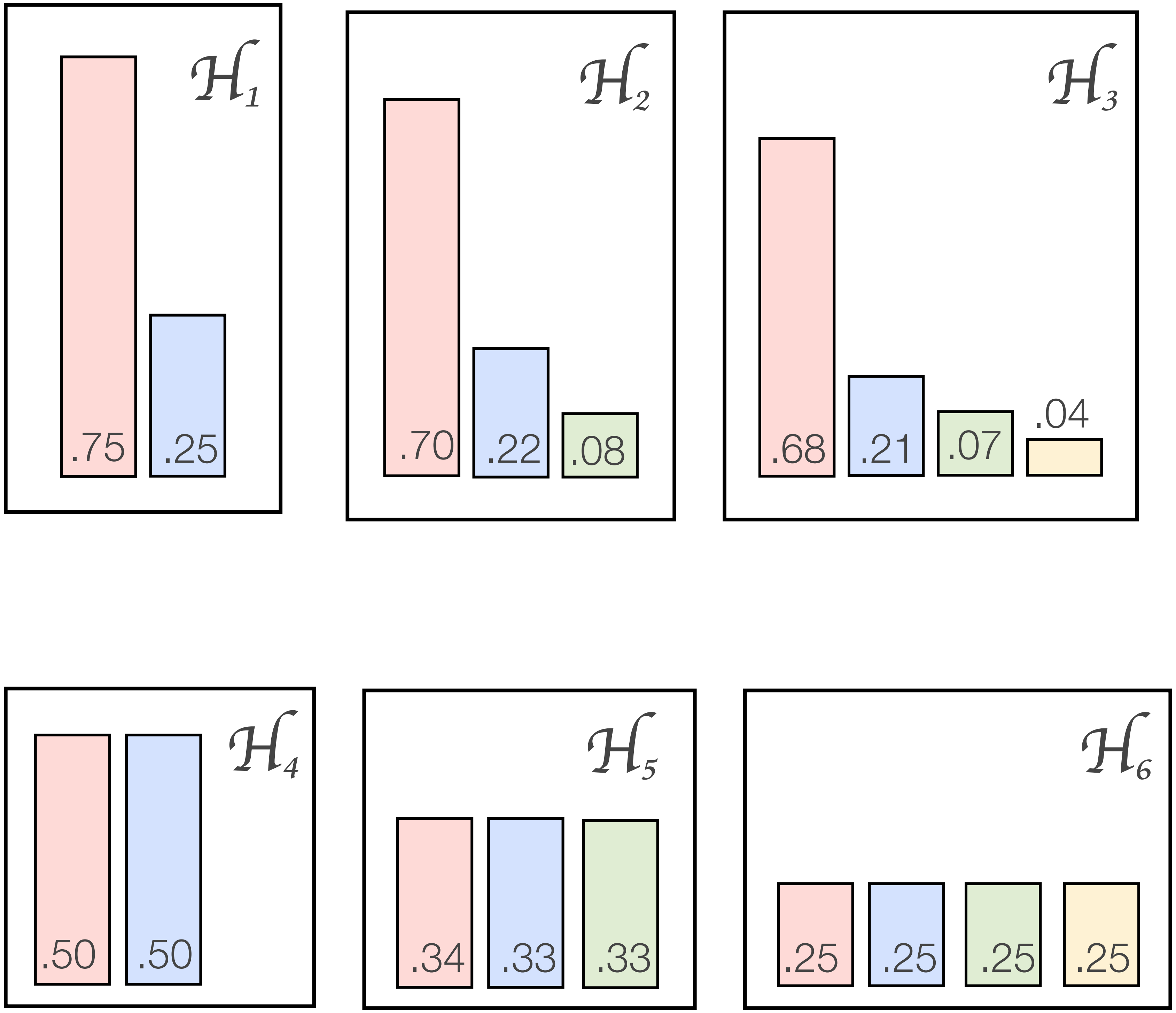



But there is nothing in this model about the number of types in each bag!

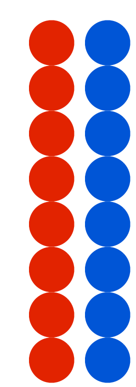
# Lecture outline (next three lectures)

- Last time: Learning about category variability
  - This kind of learning in children and adults
  - A model for this kind of learning
  - Limitations of this model
- ➡ Today: Learning about distributions of categories
  - This kind of learning in adults
  - Failure of current models
  - ➡ A model for this kind of learning
- Lecture 13: Learning about category structure
  - A model for this kind of learning
  - This kind of learning in people

# What do we *want* our model to do?




draw some data from a bag...

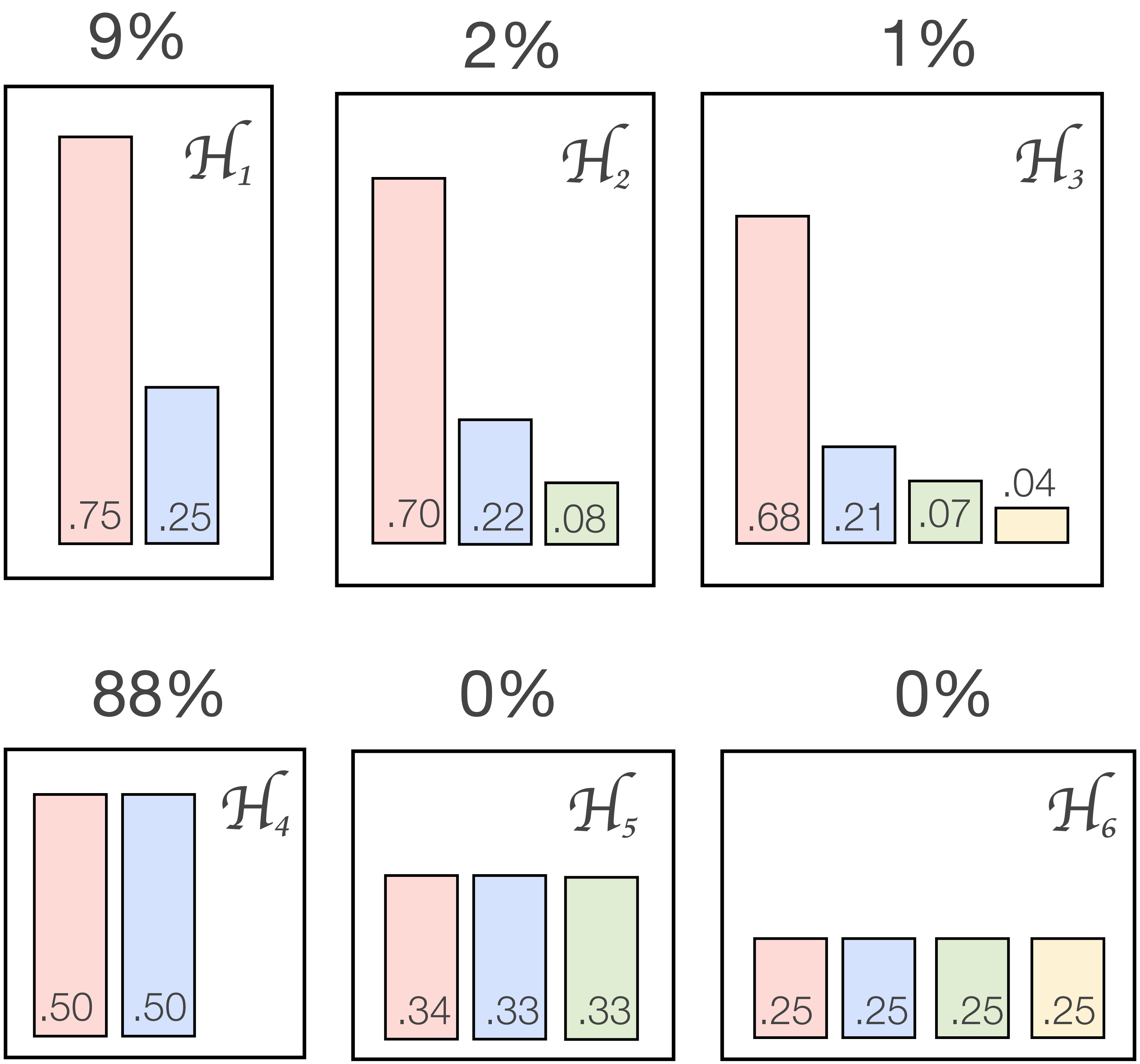


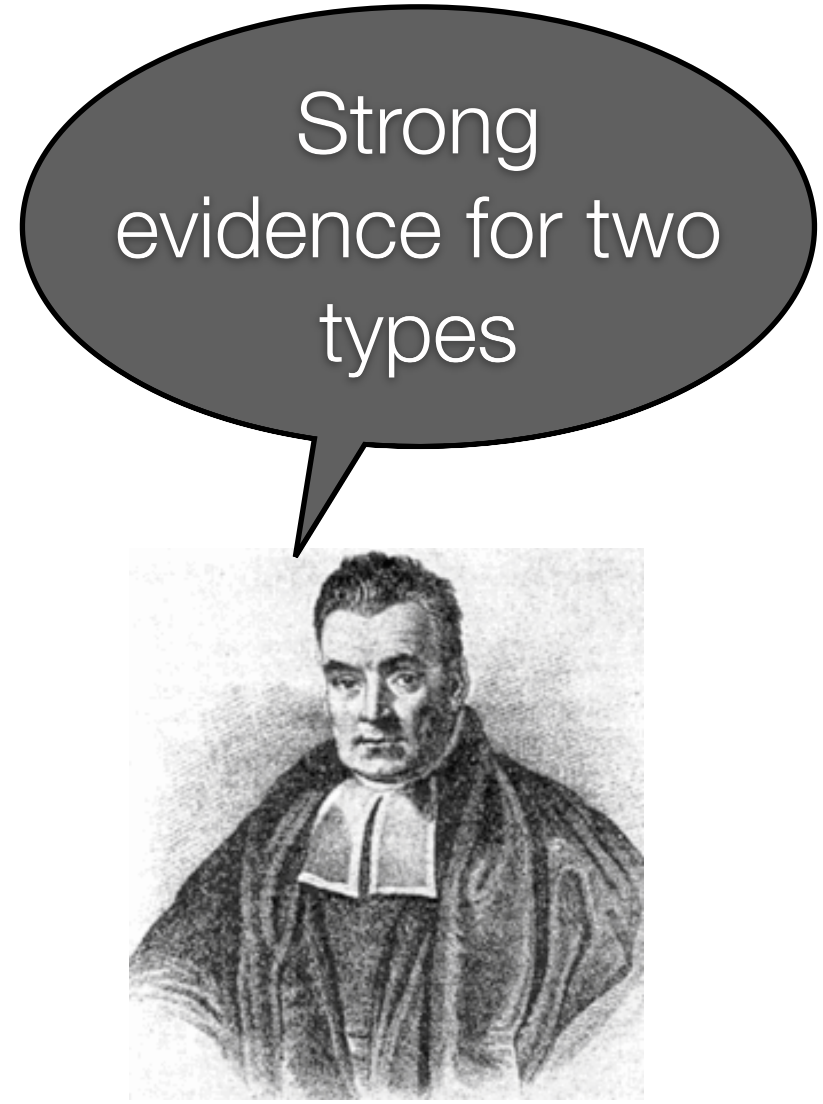
# What do we *want* our model to do?

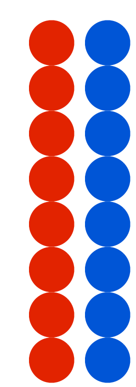
Use Bayes' rule to evaluate the relative plausibility of all hypotheses

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

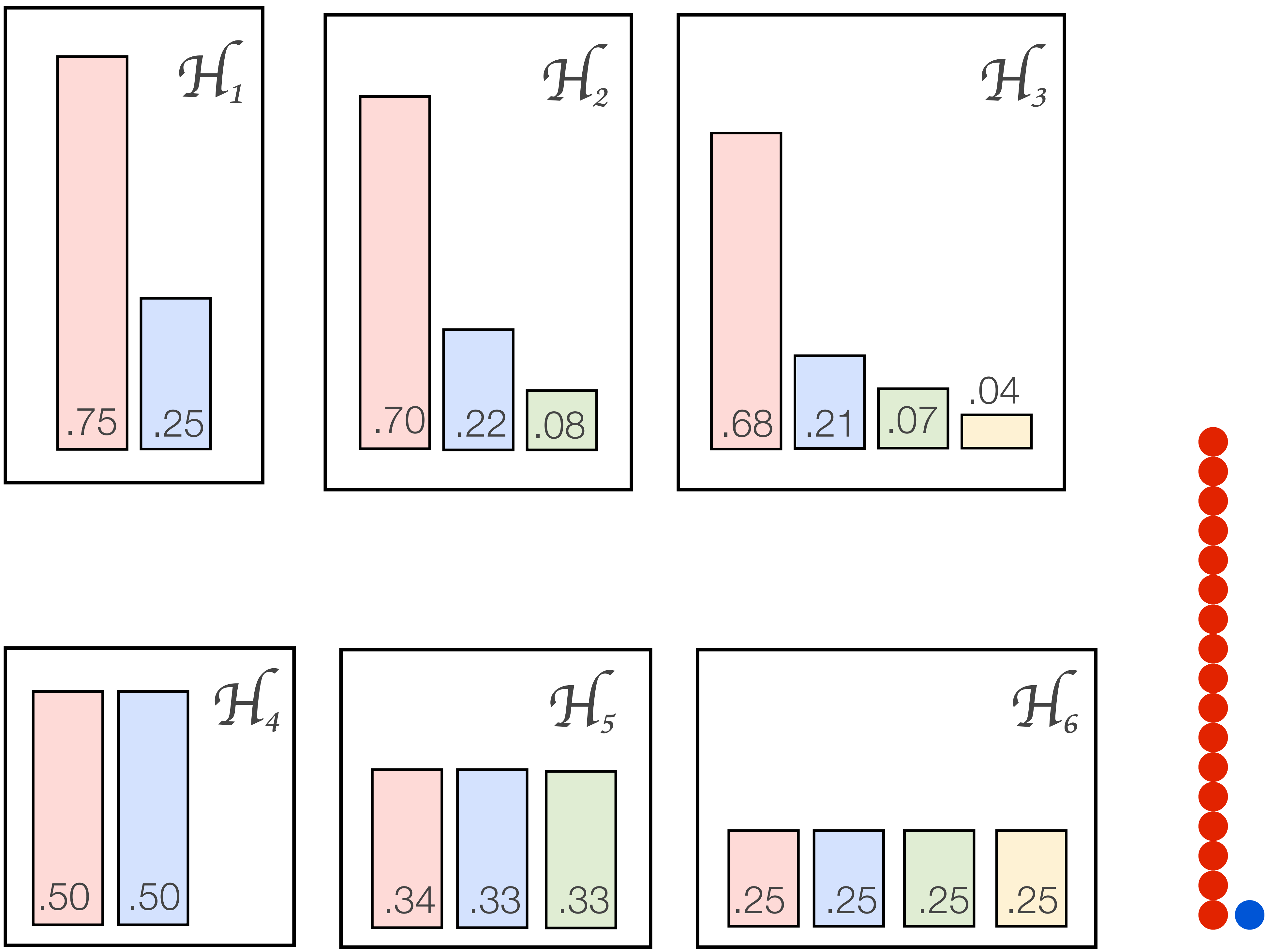
# What do we *want* our model to do?



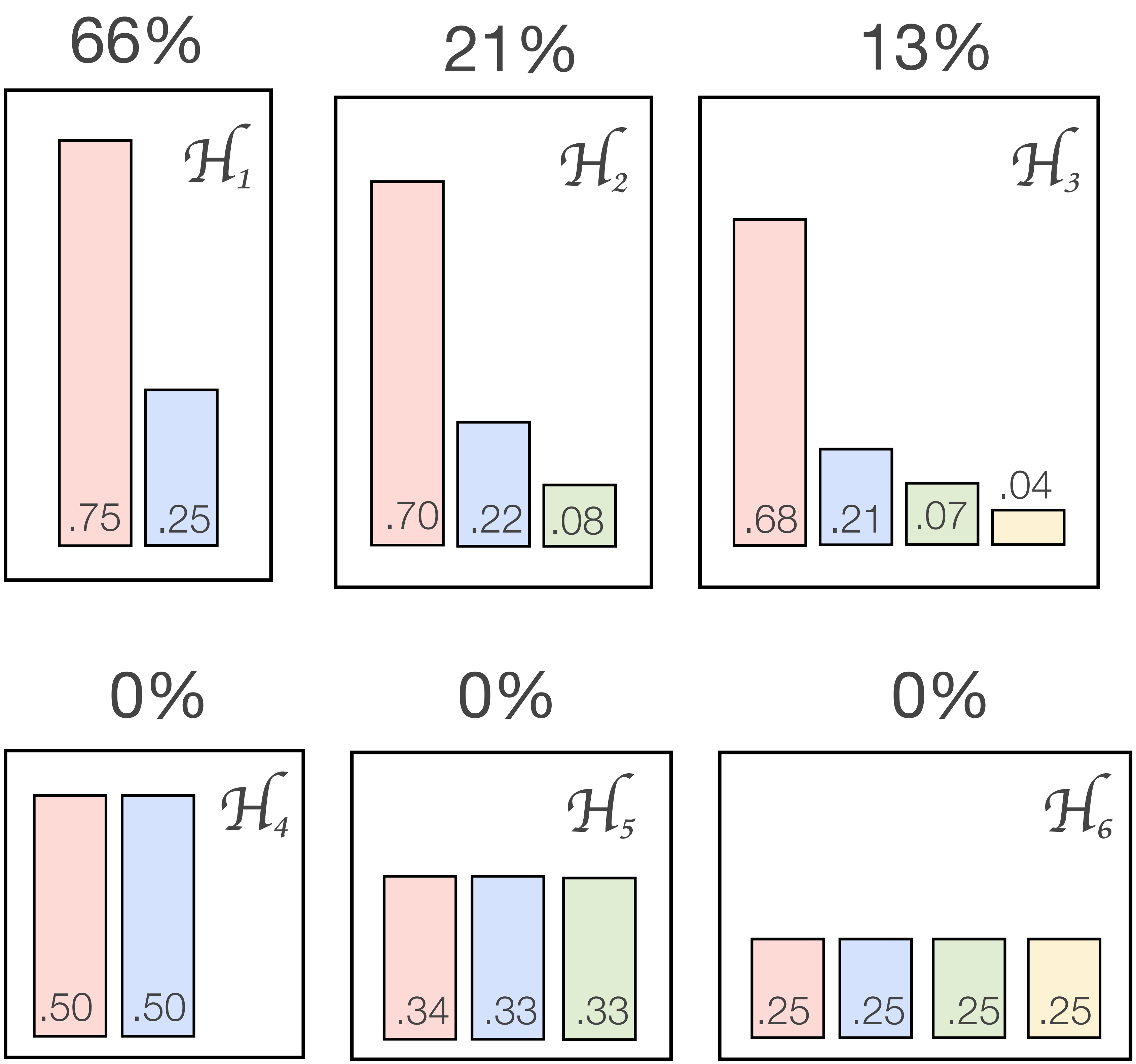



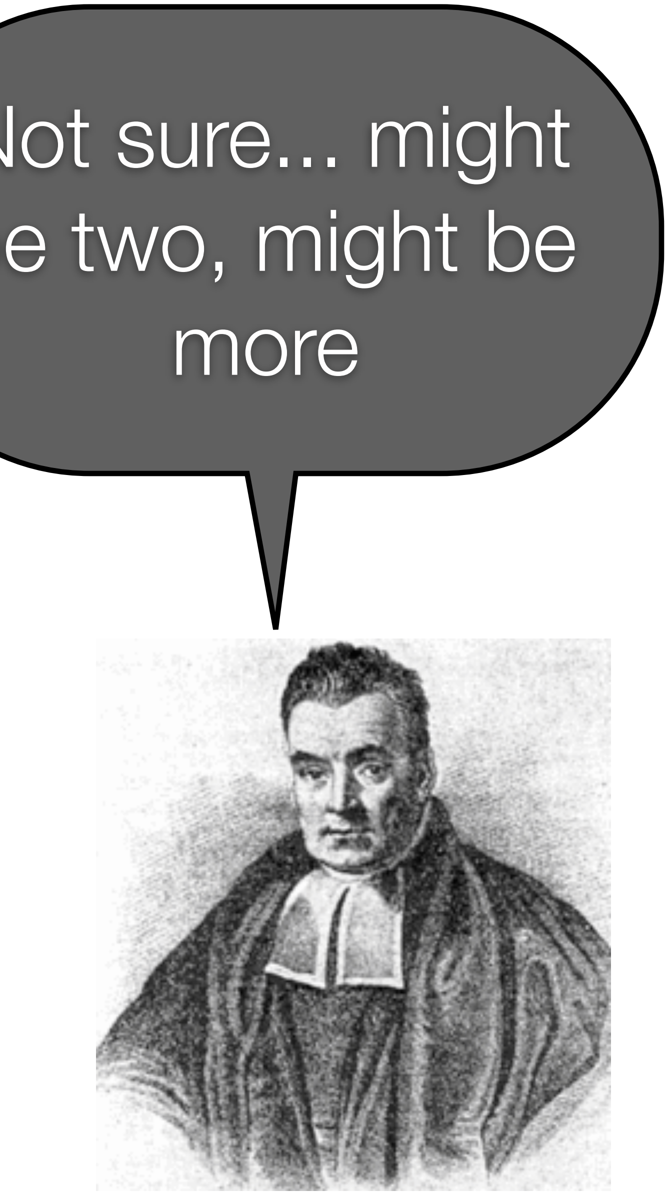



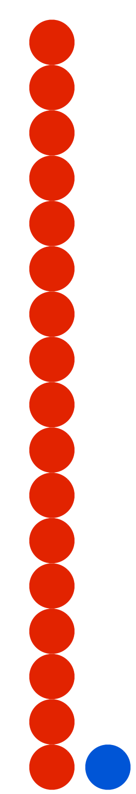
# What if the data are uneven?



# What if the data are uneven?

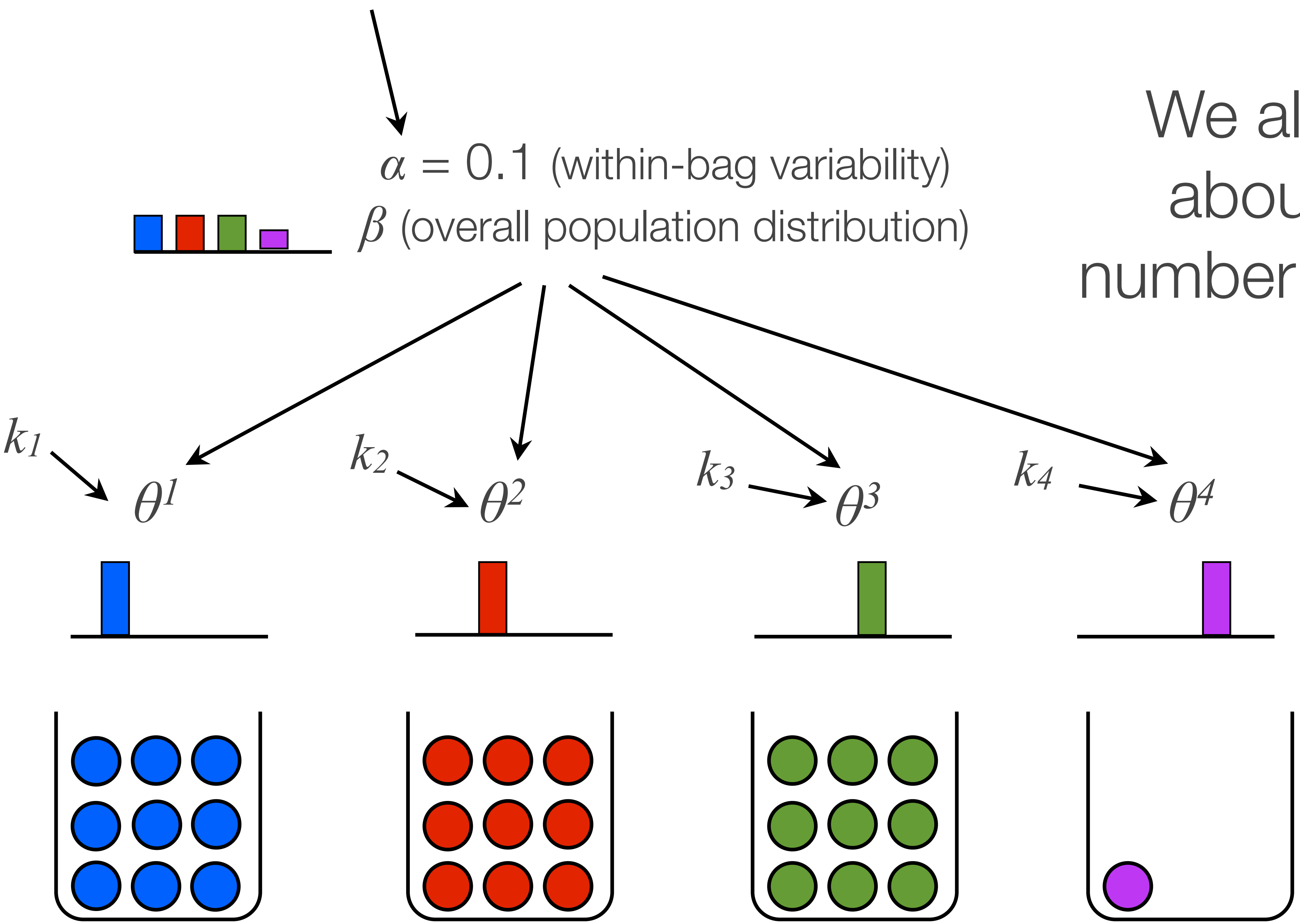


# What we want our model to do

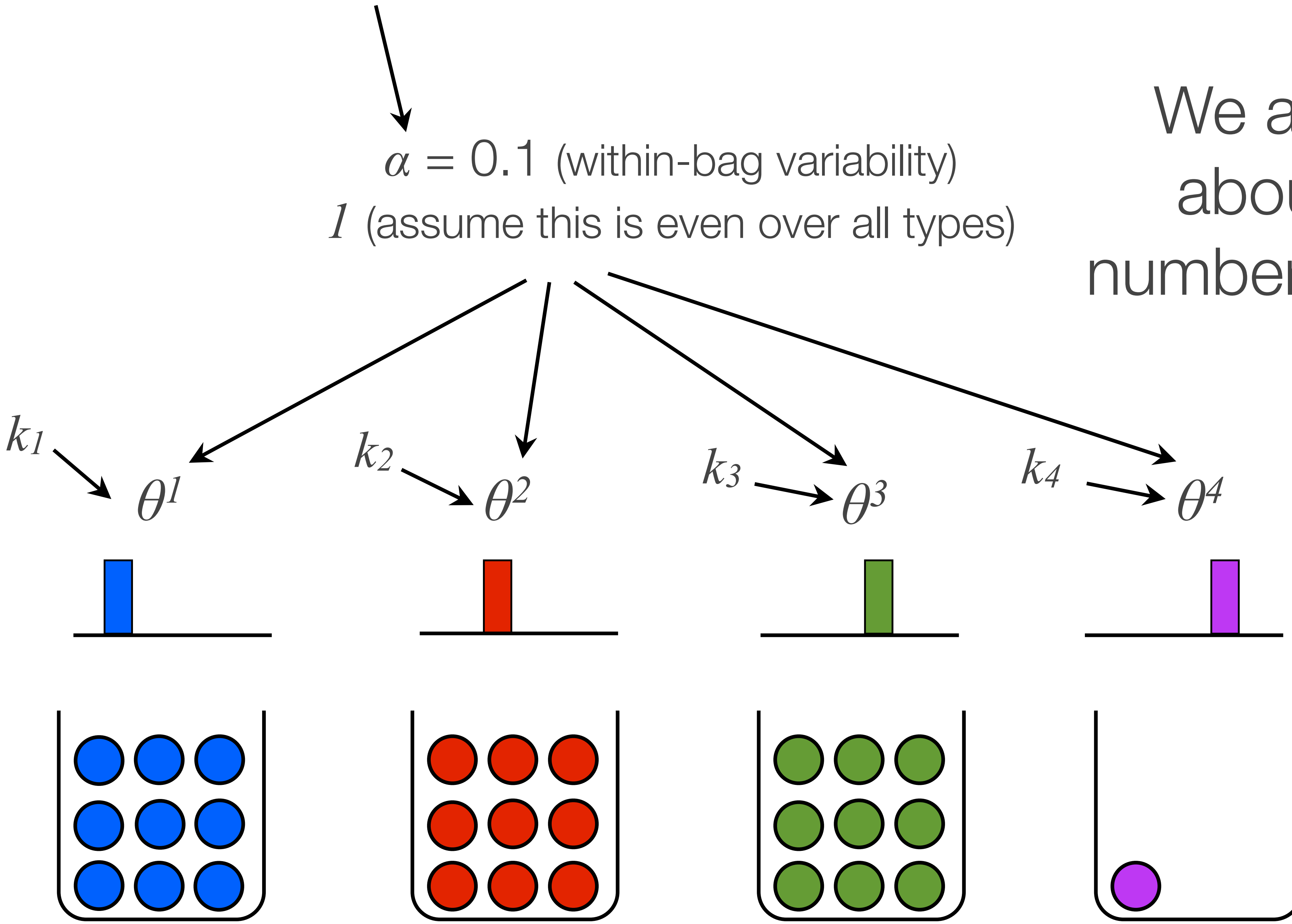
In addition to learning about category variability...




We also want to learn about the estimated number of types $k$ per bag

# What we want our model to do

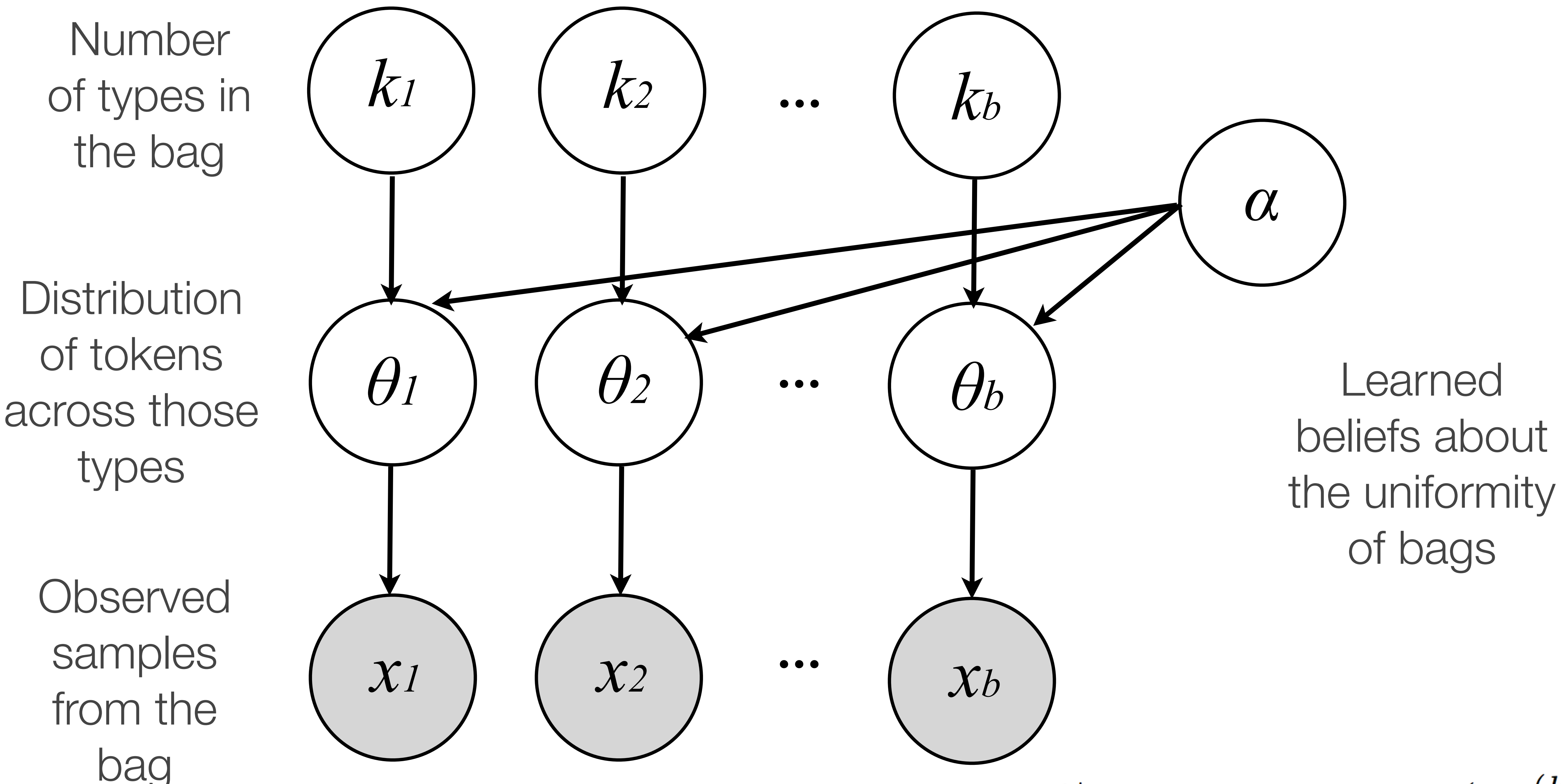
In addition to learning about category variability...




We also want to learn about the estimated number of types $k$ per bag

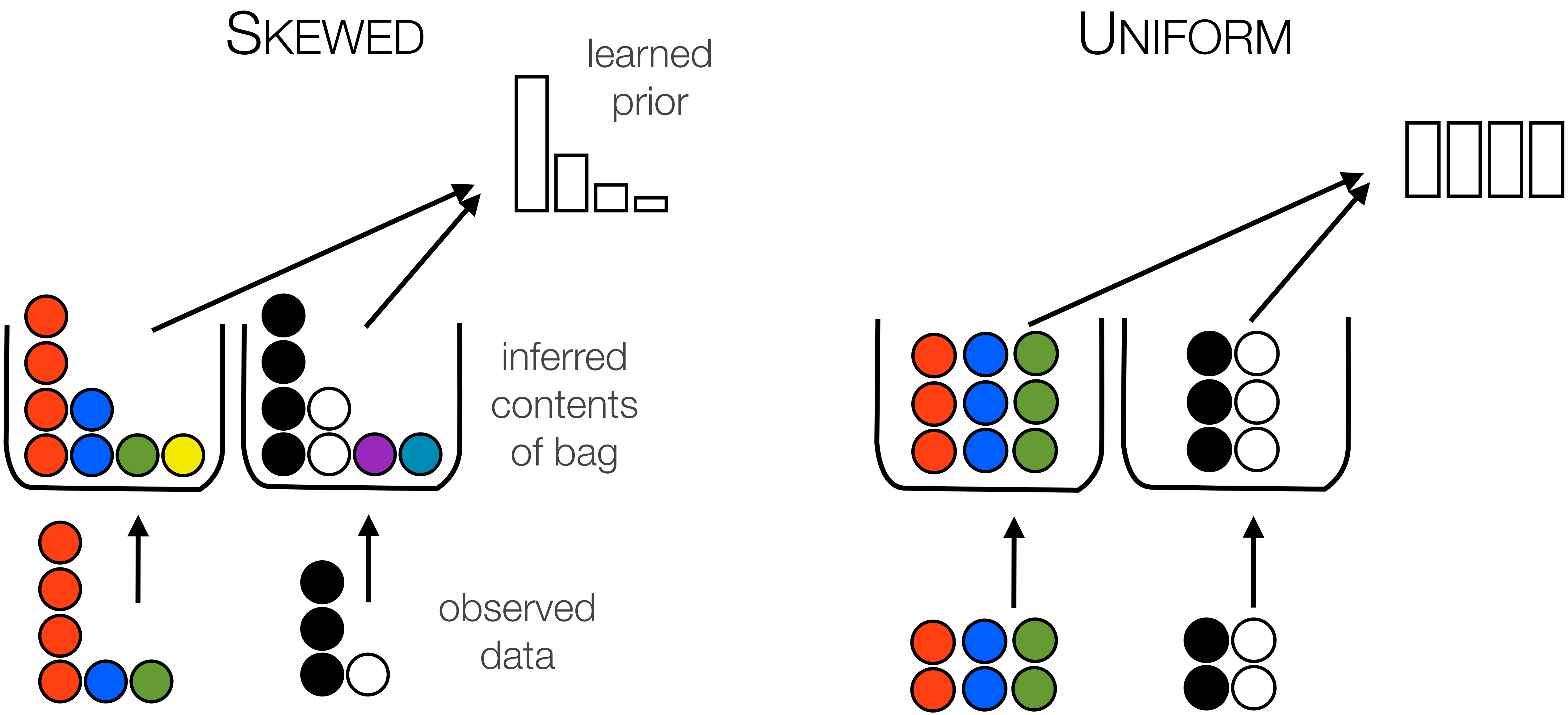
# What we want our model to do

## Another way of viewing the same model (plate diagram)



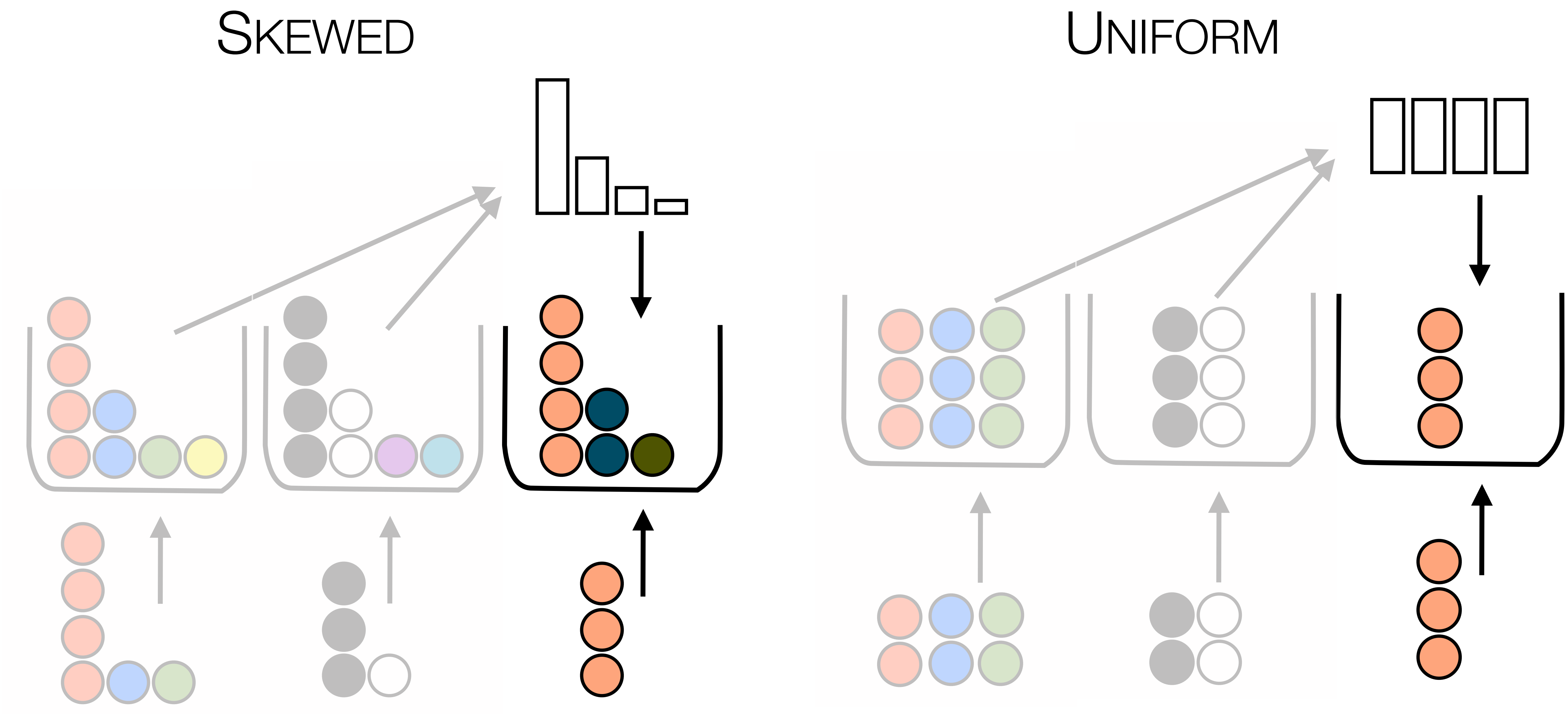

$$
\begin{aligned}
\boldsymbol{\theta} | k, \alpha &\sim \text{Dirichlet}(\alpha \mathbf{1}^{(k)}) \\
\boldsymbol{x} | \boldsymbol{\theta} &\sim \text{Multinomial}(\boldsymbol{\theta}, n)
\end{aligned}
$$

# What is this model doing?




Previous experience with the “marble world” shapes expectations (learned biases)

# What is this model doing?

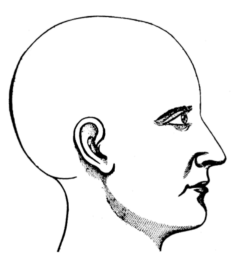


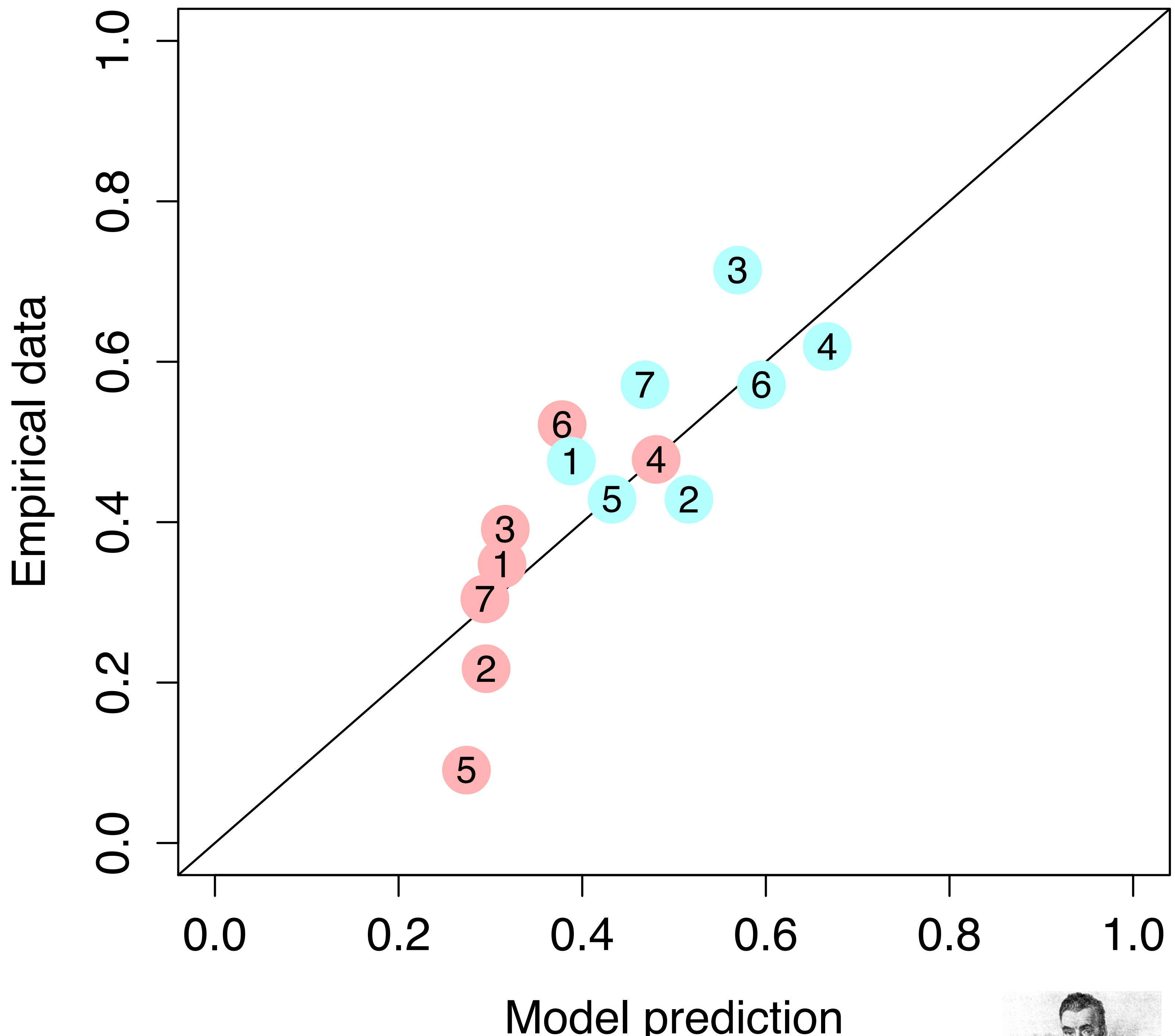

Different expectations license different inferences about the same stimulus

# What is this model doing?

Does the model make the same judgments as human participants?

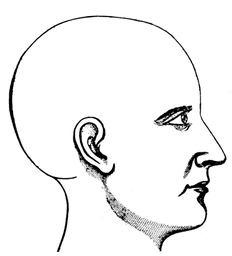
# Experiment 1

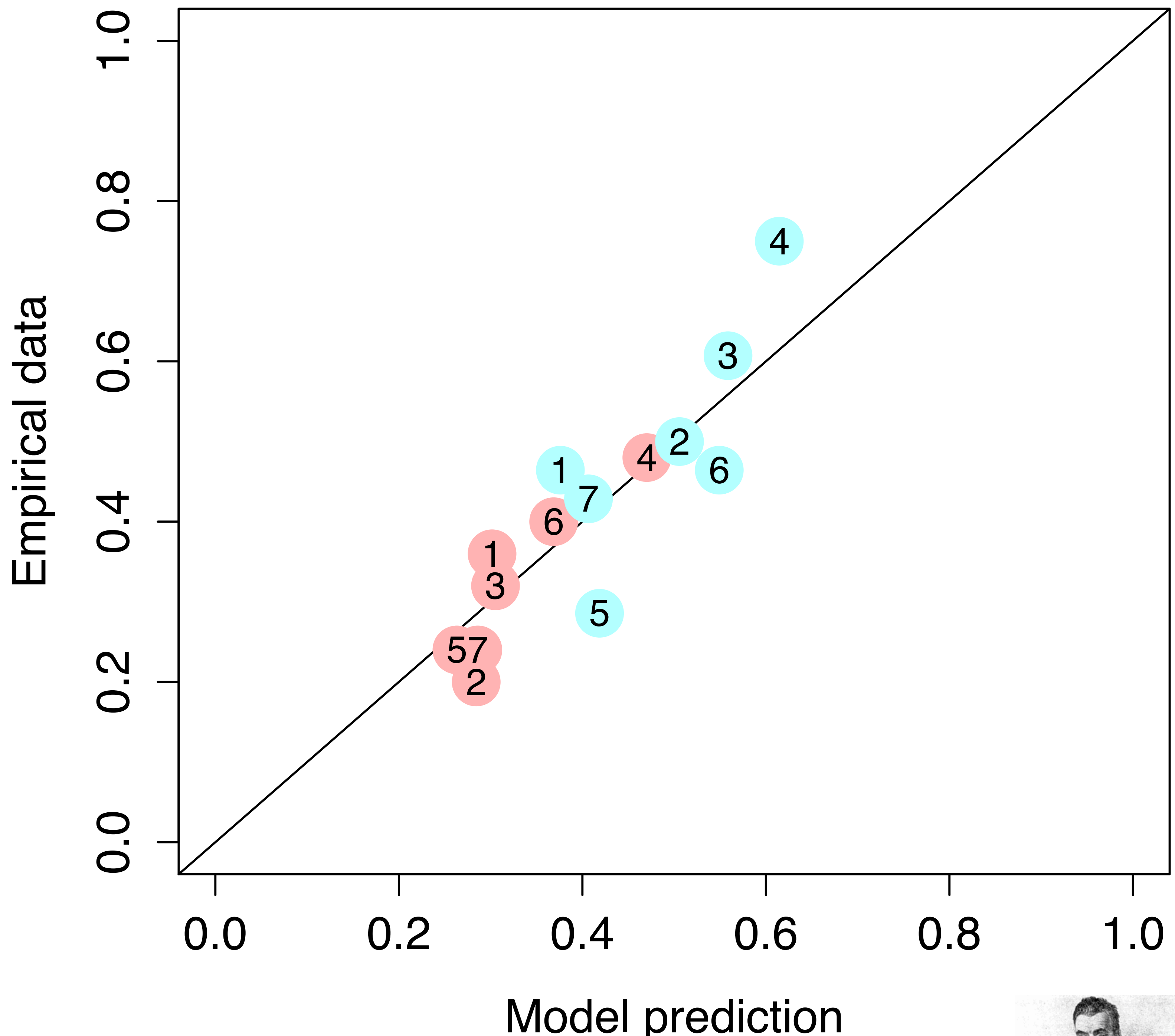





r = .82, $p < .001$

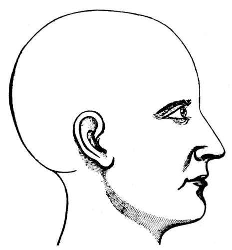
# Experiment 2

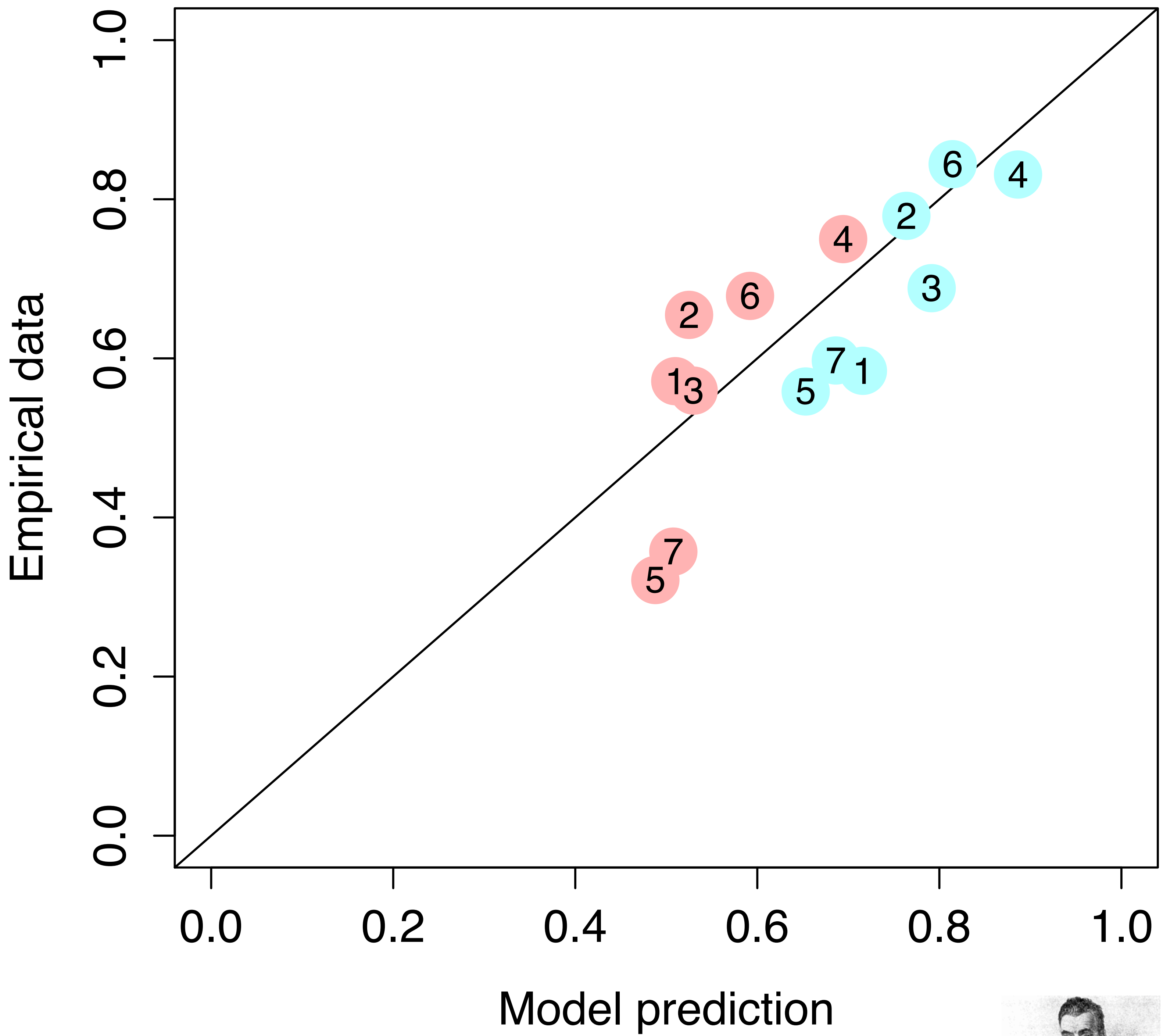





$r = .82$, $p < .001$

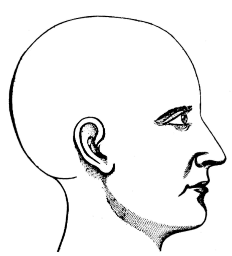
# Experiment 3

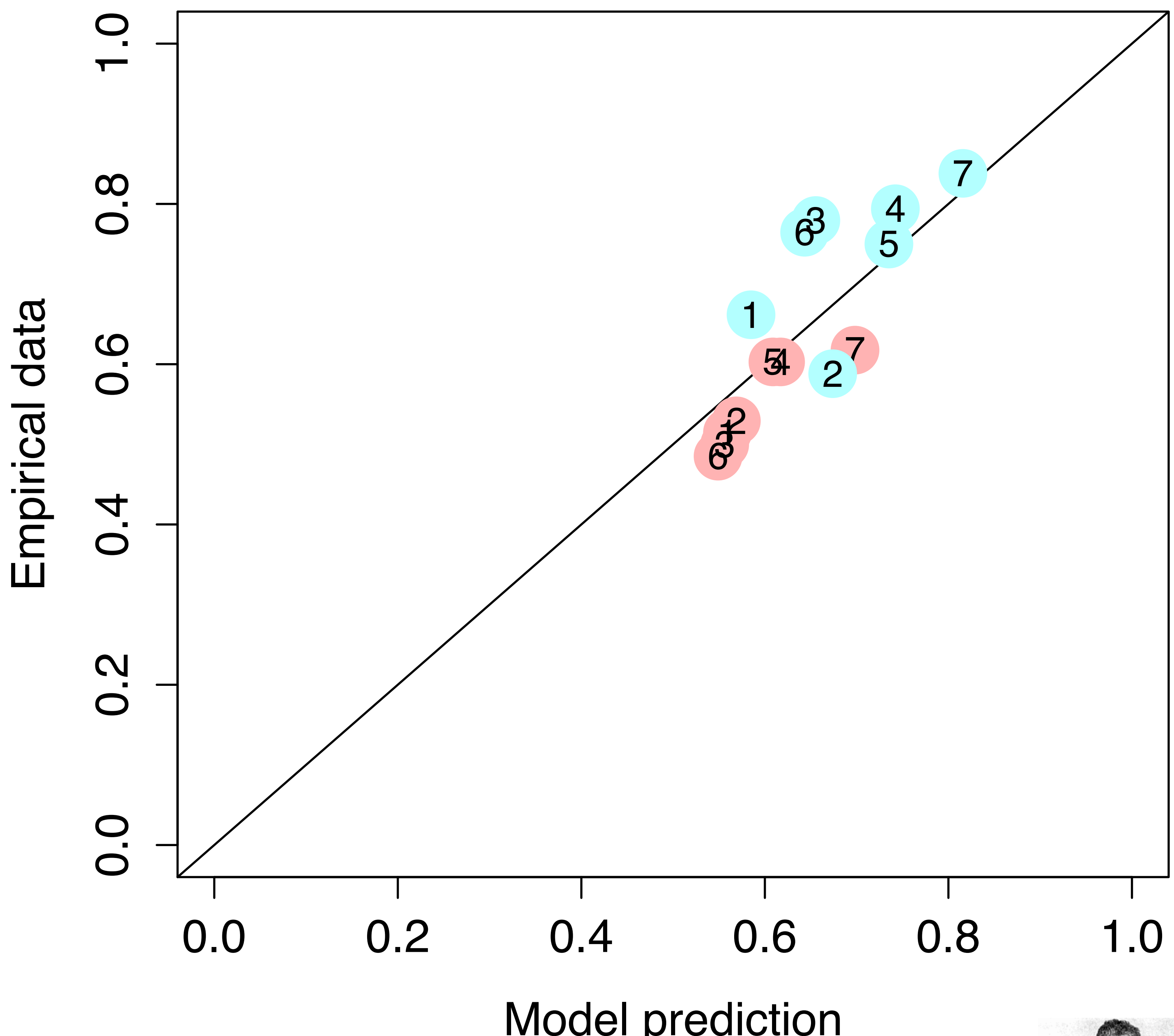





r = .82, p < .001

# Experiment 4

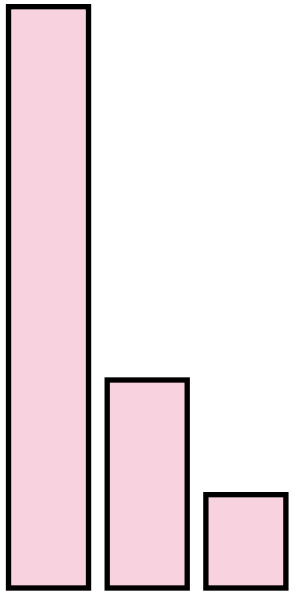




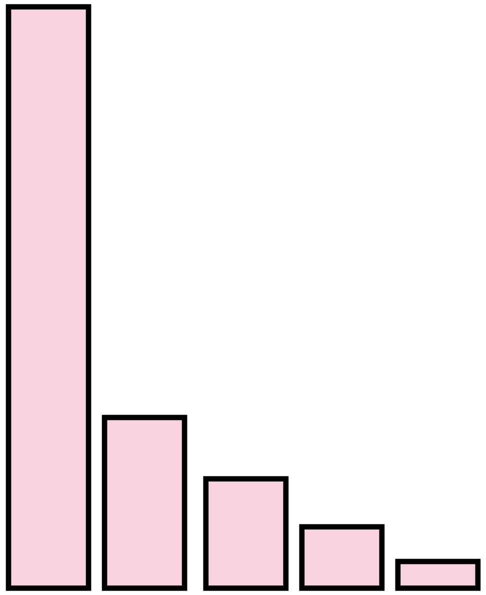

r = .82, $p < .001$

# Intuitively, what's going on?

Suppose the learner starts out being very close-minded about the structure of the distribution



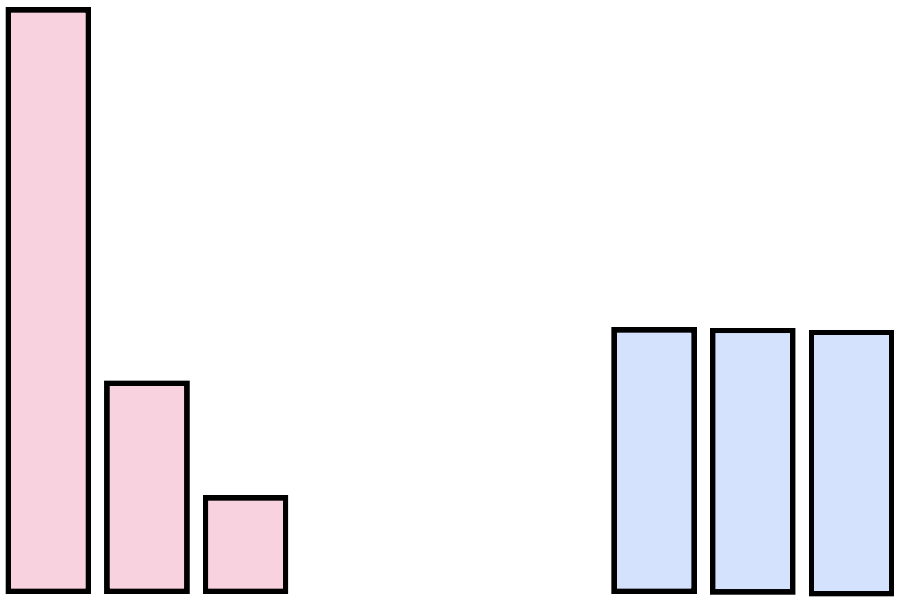
Observed shape is uninformative because the learner already "knows"



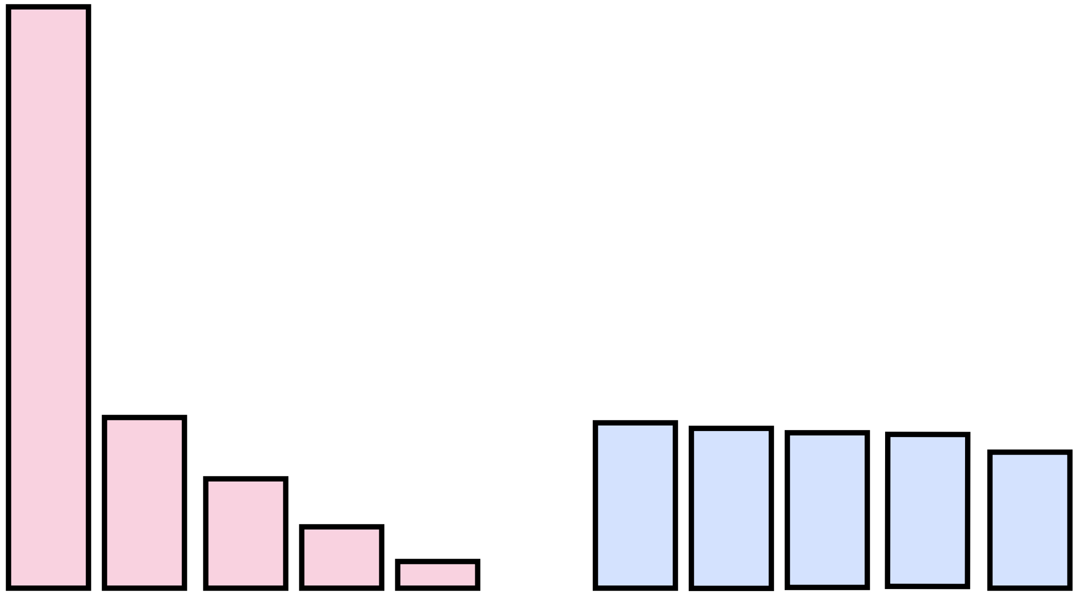
This is why the RMC and overhypothesis model fails

# Intuitively, what's going on?

If the learner is a bit more open-minded...



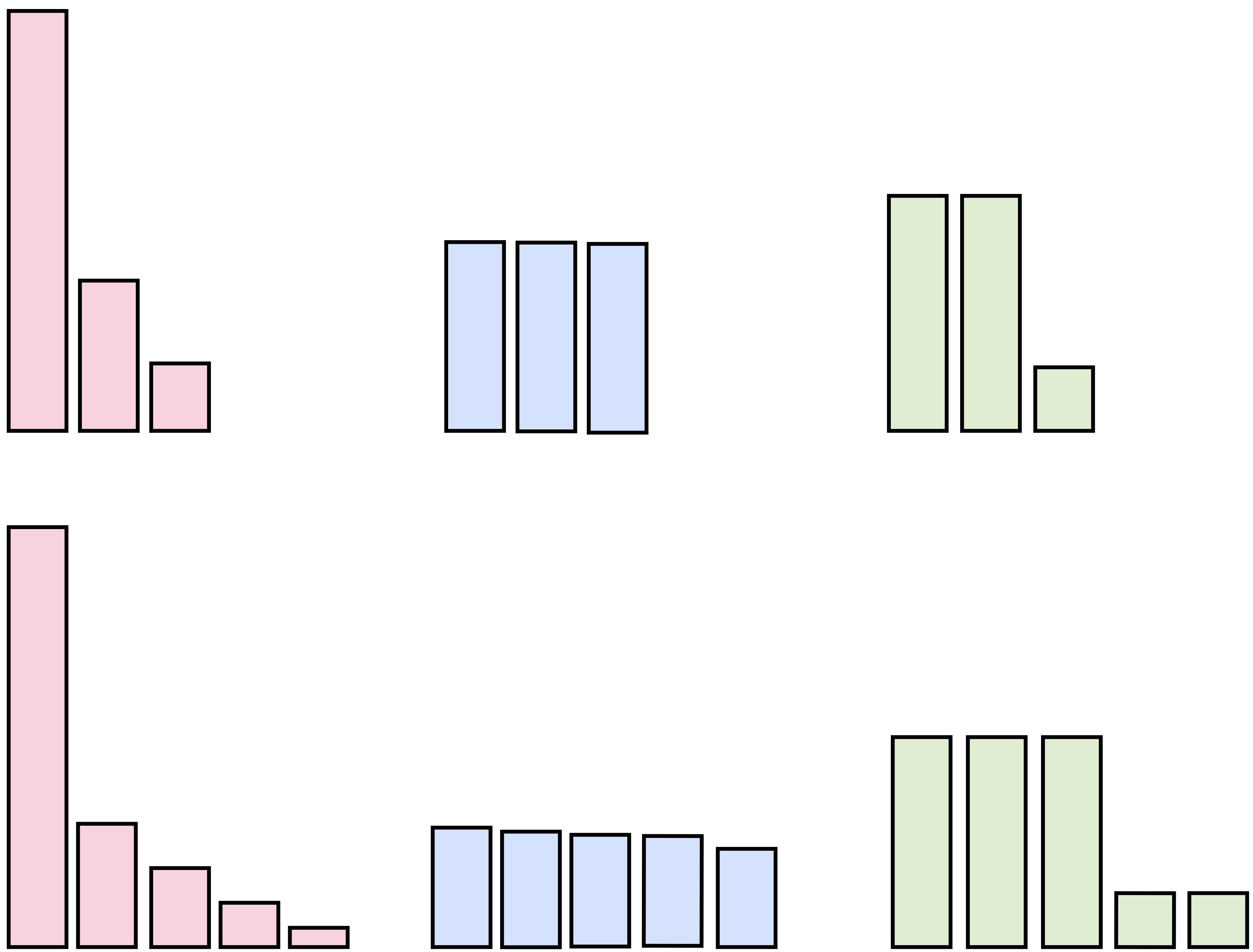
Observed shape is informative



This is why the model we presented works!

# Intuitively, what’s going on?

But if the learner is too open-minded...

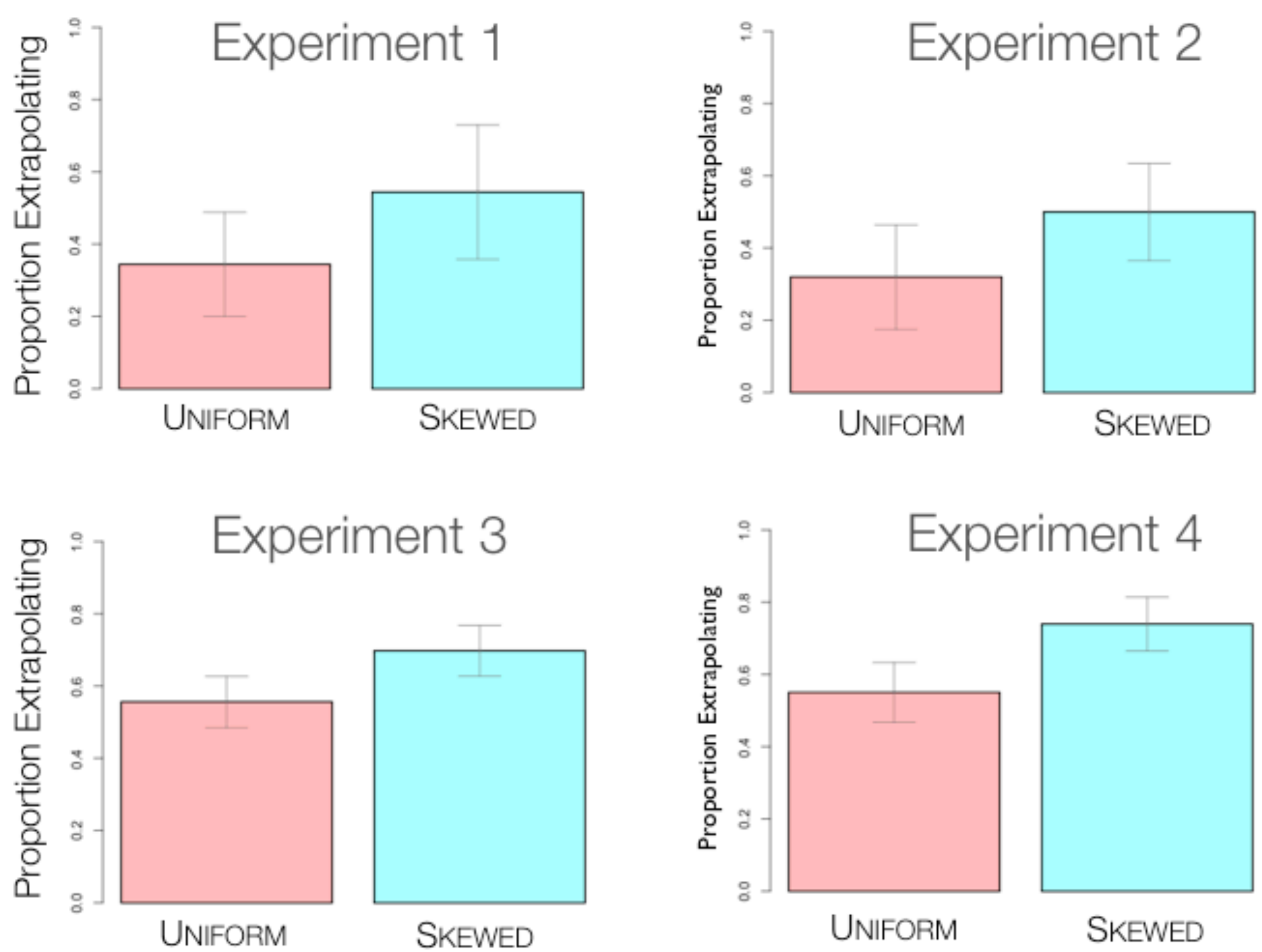


The learner can’t use the shape to guide their judgments

This is why the exemplar model fails!

# Summary

- People can learn a lot about the *distribution* of items within categories, and use that to make inferences about how many types they haven't seen



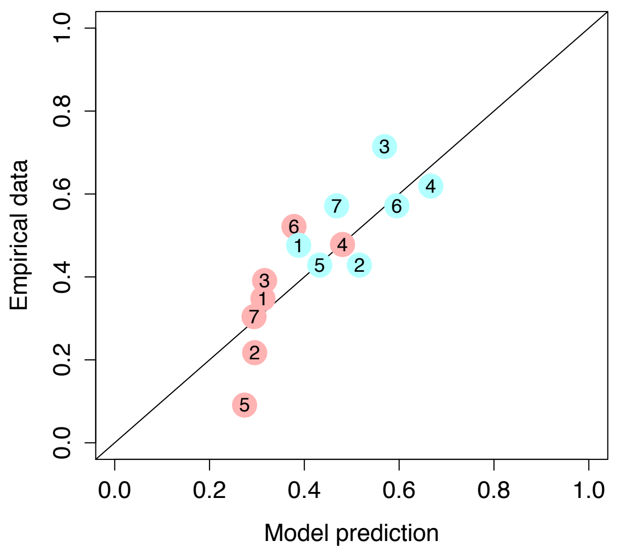
# Summary

- People can learn a lot about the *distribution* of items within categories, and use that to make inferences about how many types they haven’t seen
- The models we’ve seen so far can’t capture it, because they don’t make use of bag-specific frequency info

$$
\begin{aligned}
T(y) &= \sum_{k=1}^{K} n_k \, S(x_k, y) \\
&= \sum_{k=1}^{K} n_k \, s \\
&= Ns
\end{aligned}
$$

# Summary

- People can learn a lot about the *distribution* of items within categories, and use that to make inferences about how many types they haven't seen
- The models we've seen so far can't capture it, because they don't make use of bag-specific frequency info
- Adding to the overhypothesis model, by also learning overhypotheses about the number of types per bag, captures human performance



# Summary

- People can learn a lot about the *distribution* of items within categories, and use that to make inferences about how many types they haven't seen
- The models we've seen so far can't capture it, because they don't make use of bag-specific frequency info
- Adding to the overhypothesis model, by also learning overhypotheses about the number of types per bag, captures human performance
- Next time: one more kind of higher-order knowledge: learning about structure

# Additional references (not required)


- Navarro, D. (2013). Finding hidden types: Inductive inference in long-tailed environments. In M. Knauff, M. Pauen, N. Sebanz, and I Wachsmuth (eds). *Proceedings of the Annual Conference of the Cognitive Science Society*: 1061-1066