Bayes factor analyses with informed priors:

Examples from artificial orthography experiments and implications for literacy research

Anna Samara

School of Human Sciences, University of Greenwich

11/06/2020

$\mathcal{H}_0$ in psychological research

- Discussed in the context of false-positive psychology
- Selective preference for publishing studies that reject $\mathcal{H}_0$

- Today: Within NHST, accepting vs. rejecting $\mathcal{H}_0$ are not “on equal footing”

The problem

- **Theory predicts that X is impossible**
- **Evidence that X did happen**
- **Theory is incorrect**

“Nothing could force me to stay home for 3 months”

observation 1
observation 2
observation 3
observation 4
observation …
observation 8
observation 9
observation 10

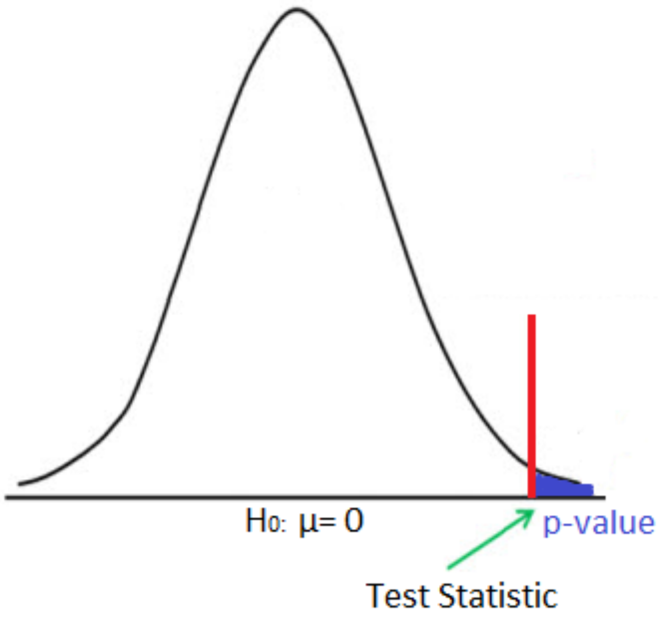
The problem

- **Theory predicts that X is impossible**
- **No evidence that X happened**
- **Theory is correct ?**

observation 1
observation 2
observation 3
observation 4
observation …
observation 8
observation 9
observation 10

“Nothing could force me to stay home for 3 months”

H_0
(no effect)



H_1
(non-zero effect)

Evidence for H_1

H_0

(no effect)

H_1

(non-zero effect)

p values cannot distinguish between these states of knowledge

Evidence for H_0

No evidence at all

Evidence for H_1

Children are powerful “statistical learners”

Children do not pick up on spelling patterns (until “late” stage)

Statistical learning skill is impaired dyslexia

No evidence of statistical learning impairment

Better ‘statistical learners’ are better readers/spellers

No link between lab-based learning skill and literacy

Today's talk

- Bayes Factors: **Bayesian** measure of evidence for something existing versus not existing (**hypothesis testing**)
  - Is there an effect of instructional method X on spelling; is there an interaction between learning ability and age?
- Different from Bayesian parameter estimation e.g. using brms (Buerkner, 2016)
  - How big is the effect of instructional method X (knowing that it exists)

Today’s talk

- BF approach advoked by Zoltan Dienes (2008, 2014, 2015, 2019)
- With thanks to Liz Wonnacott (UCL) and lab members (https://languagelearninglab-ucl.com/)
- “B for every p”
- You can only evidence that something does not exist given a claim of how big it could be **if it did exist:** I will illustrate how to do this using prior-informed distributions (cf. default distributions)

Why use BFs a measure of evidential strength?

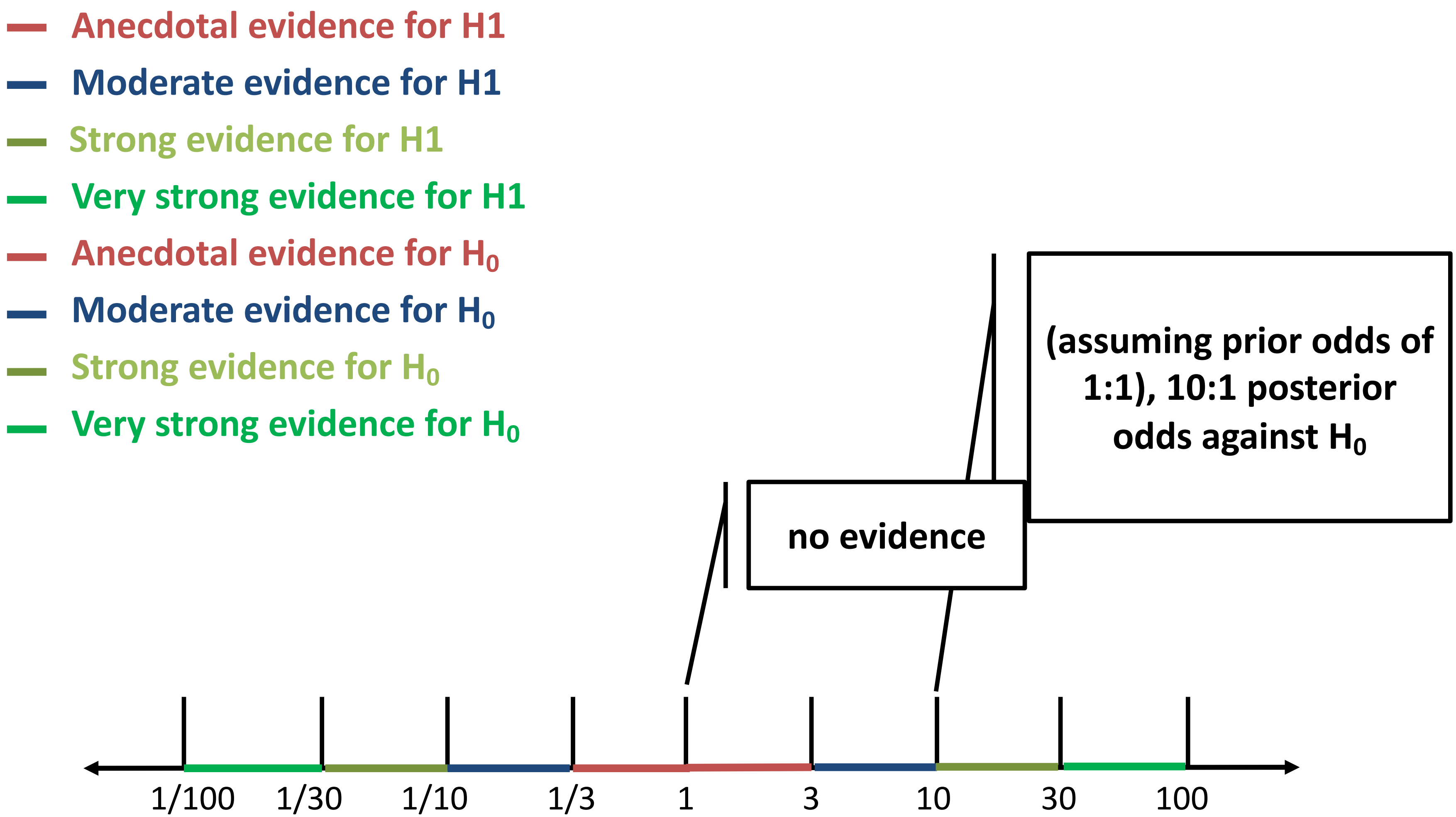
1. BFs can quantify evidence for $\mathcal{H}_1$ **<u>and $\mathcal{H}_0$</u>**
2. Prior informed BFs can be interpreted meaningfully when optional stopping is used (Dienes, 2016; Rouder, 2014)
3. Differences between BFs are meaningful
   - e.g., BF of 3 → posterior probability of .75 for $\mathcal{H}_1$
4. To specify your priors, you ***have to*** meaningfully think about your theory and its hypotheses

Bayes Factors: a brief tutorial

- Measure of strength of belief, based on the idea that the evidence supports the theory that most strongly predicts it
- BF = Expresses how much should a dataset sway our belief from one hypothesis (e.g., H_0) to another (e.g., H_1).

$$\frac{P(H1 \mid D)}{P(H0 \mid D)} = \boxed{\frac{P(D \mid H1)}{P(D \mid H0)}} \times \frac{P(H1)}{P(H0)}$$

Jeffreys (1939)



Computing Bayes Factors

- **R Code provided by …**
  - Baguley & Kaye (2010)
  - Stefan Wiens
  - Bence Palfi (2018)
- **Also check ZD's online calculator**
  http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm

Specifying your predictions

- **Three values!**

1. "If there were an effect how big would it be"

- Summary of your data

2. Effect size (e.g. mean difference between conditions)
3. Associated measure of standard error

- Based on a pilot
- Based on your own previous work
- Based on others’ previous work
- Based on ‘intuitive’ knowledge of max/min effects
- Based on the data itself

Dienes (2019). 10.31234/osf.io/yqaj4

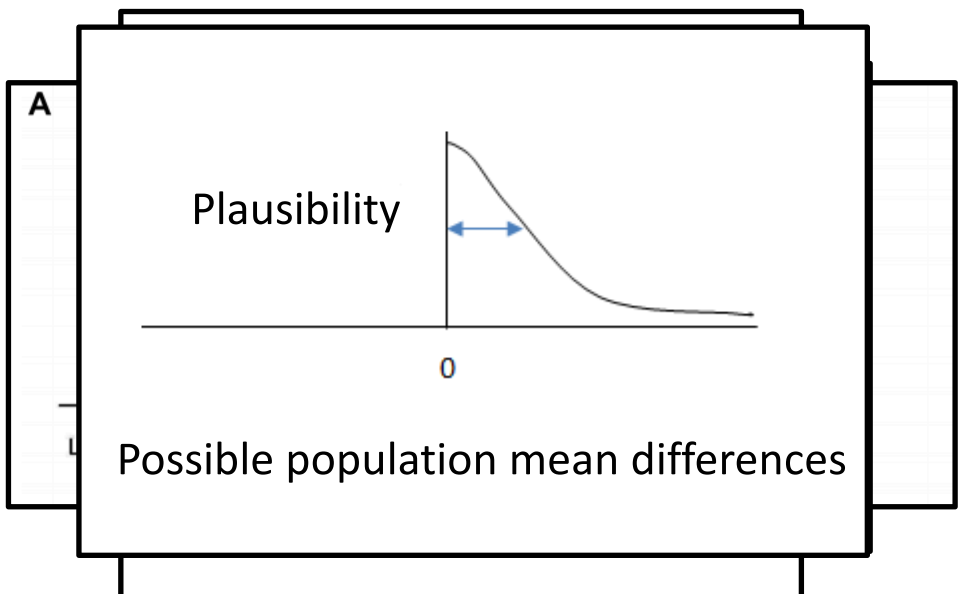
Robustness regions: RR[min, max]

- Often, there is several (equally good) ways of modelling H_1 to represent the same theory
  - Identify the range of ***scientifically plausible*** scale factors
    - e.g. better-than-chance accuracy: 50.1 - 100%
  - RR: range of ***scale factors that support the same qualitative conclusion*** (e.g. substantial evidence for H_1 over H_0)
- Infer **robustness** via comparing the former against the latter
  - If RR contains plausible range, conclusion is "safe"
  - Readers can also check whether conclusion holds against own scientific intuitions
- **BUT … it is still a heuristic**

Representing your predictions

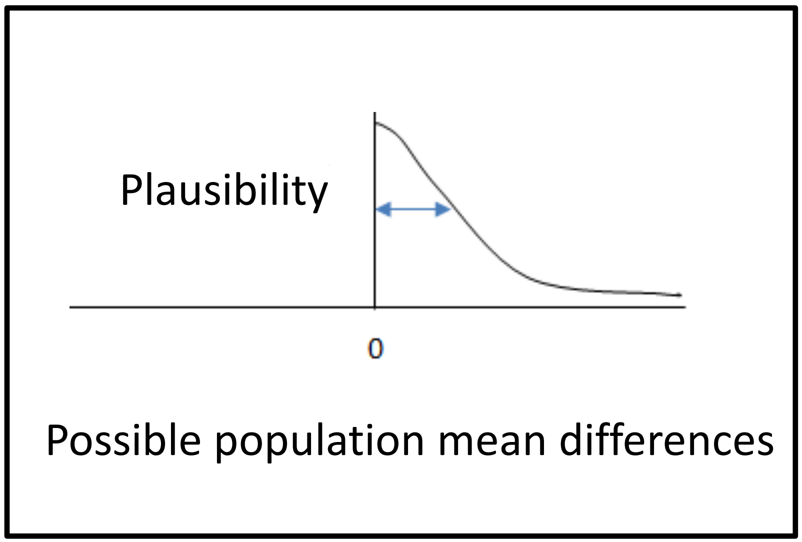
- Distribution of your prediction theory
  – Plausibility of population parameter values given the theory [p(population effect|theory)] (not distribution of data – assumed to be normally distributed)

- Uniform
- Normal
- Cauchy
- Half-normal



Often, precise shape of the distribution can make little difference

The half-normal distribution



$B_{HN}(0,SD)$

- Directional effect predicted
- Distribution centred on zero
  - Smaller effects are increasingly likely (good for developmental research)
- *SD* = likely predicted value
- Max = 2*SD*

Example from Samara et al. 2019

- Can 7-year-old children discriminate between items that differ in legality? (**main effect of legality**)

```
Bf(seBF, meanBF, uniform = 0, meanoftheory = 0, sdtheory = h1mean, tail = 1)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.289	0.154	1.883	0.060
legality.ct	0.419	0.139	3.013	0.003
condition.ct	0.157	0.311	0.504	0.614
legality.ct:condition.ct	-0.068	0.280	-0.241	0.809

```
## [1] 0.0307042
##
```

beta estimate from a previous study carried with 7-year-olds and comparable methods

For half-normal:
uniform = 0
meanoftheory = 0

- Hopefully, I have by now convinced you that using prior-informed BFs is….

- Useful
- Intuitive
- Straightforward to implement

Bayes factors in action!

Singh, Samara, & Wonnacott (under review). Statistical and explicit learning of graphotactic patterns with no phonological counterpart: Evidence from an artificial lexicon study with 7-8-year-olds and adults

Spelling development

1. Knowledge acquired **explicitly**
   - ✓ Patterns that are easy to verbalize (e.g., i before e except after c)

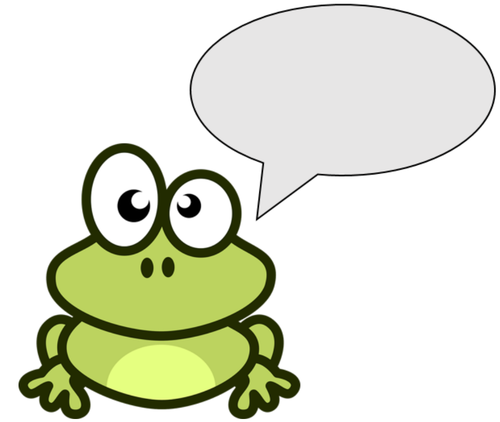
2. Knowledge acquired via **statistical learning processes**: basis of humans' ability to extract statistical patterns of varying complexity from the input

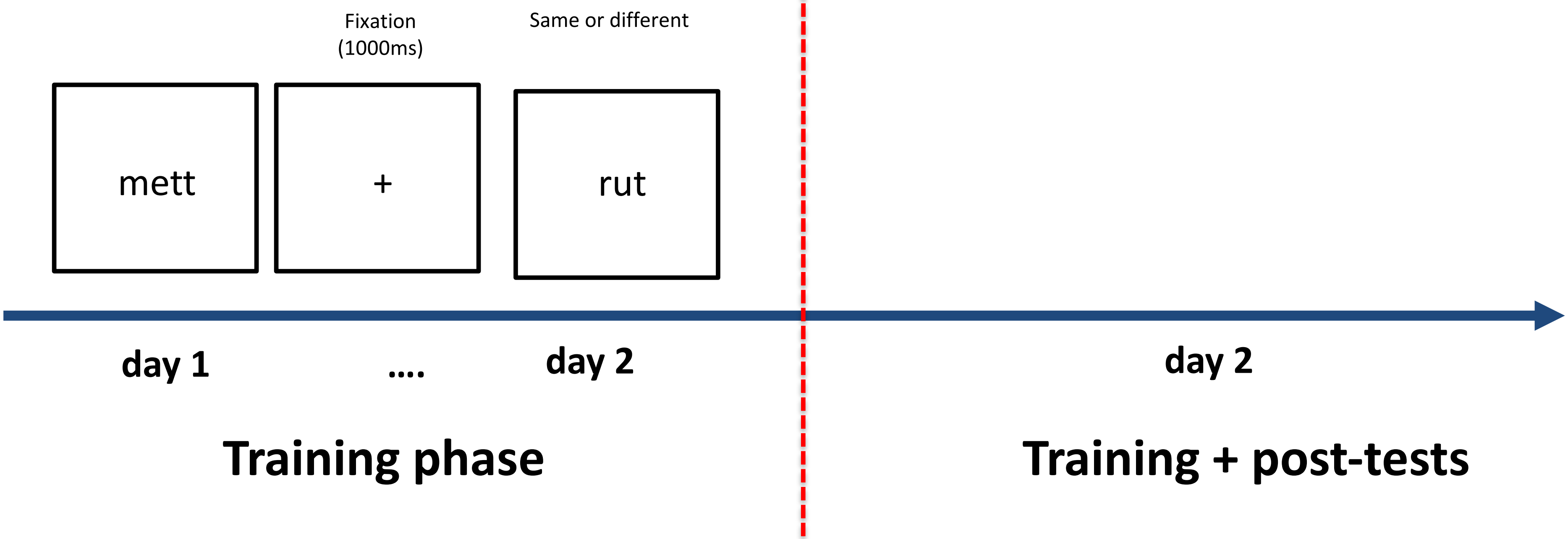
Spelling development

- **Written language is highly patterned (Kessler & Treiman, 2001)**
  - **Phoneme context:** e.g. medial /ɛ/ is 43% of time spelled with an ea before /d/ (e.g., h**ea**d)
  - **Visual constraints mirrored by spoken language constraints:** e.g., spoken/written English words never begin with *ng
  - **Purely visual constraints**: e.g., doubling occurs more often after 1-letter-vowel spellings than 2-letter-vowel spellings (e.g. bedding vs. heading)

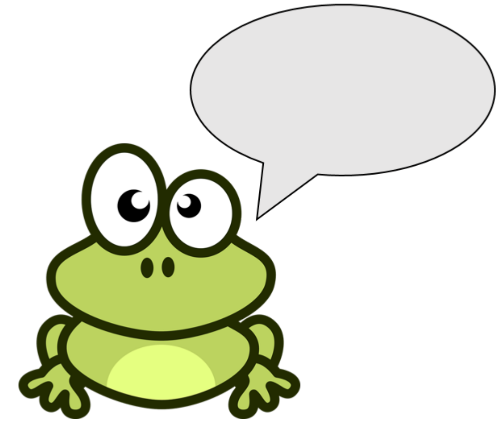
- Many ways to measure children and adults's sensitivity to such patterns

Samara & Caravolas (2014). JECP
Samara et al. (2019). Cognition





Samara & Caravolas (2014). JECP
Samara et al. (2019). Cognition



Is this consistent with previous words?

Fixation (1000ms)

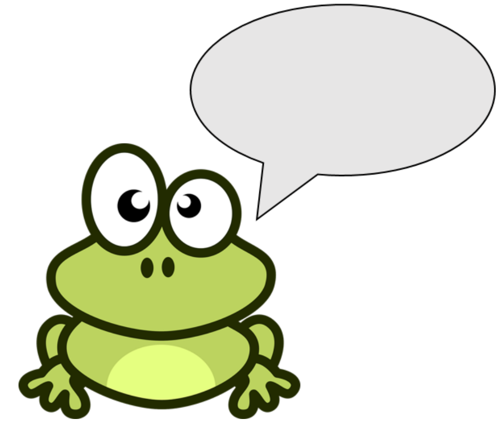
duff

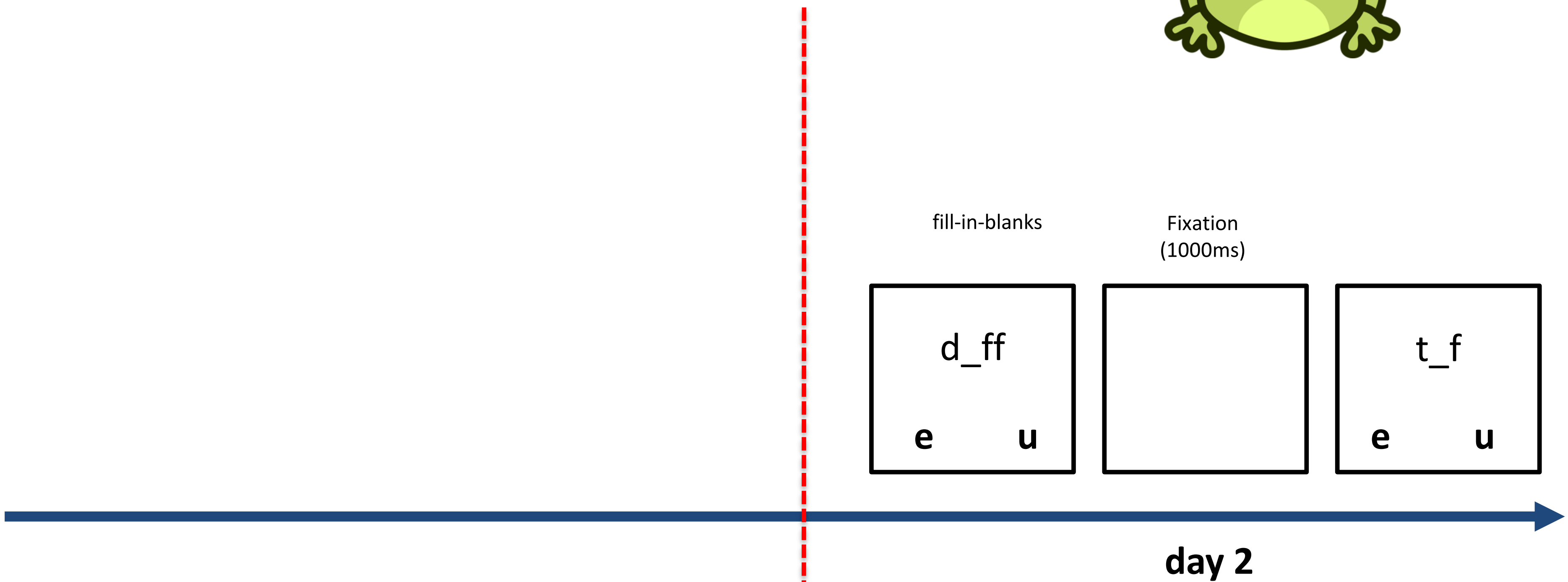
tef

day 2

Legality judgments

Samara & Caravolas (2014). JECP
Samara et al. (2019). Cognition



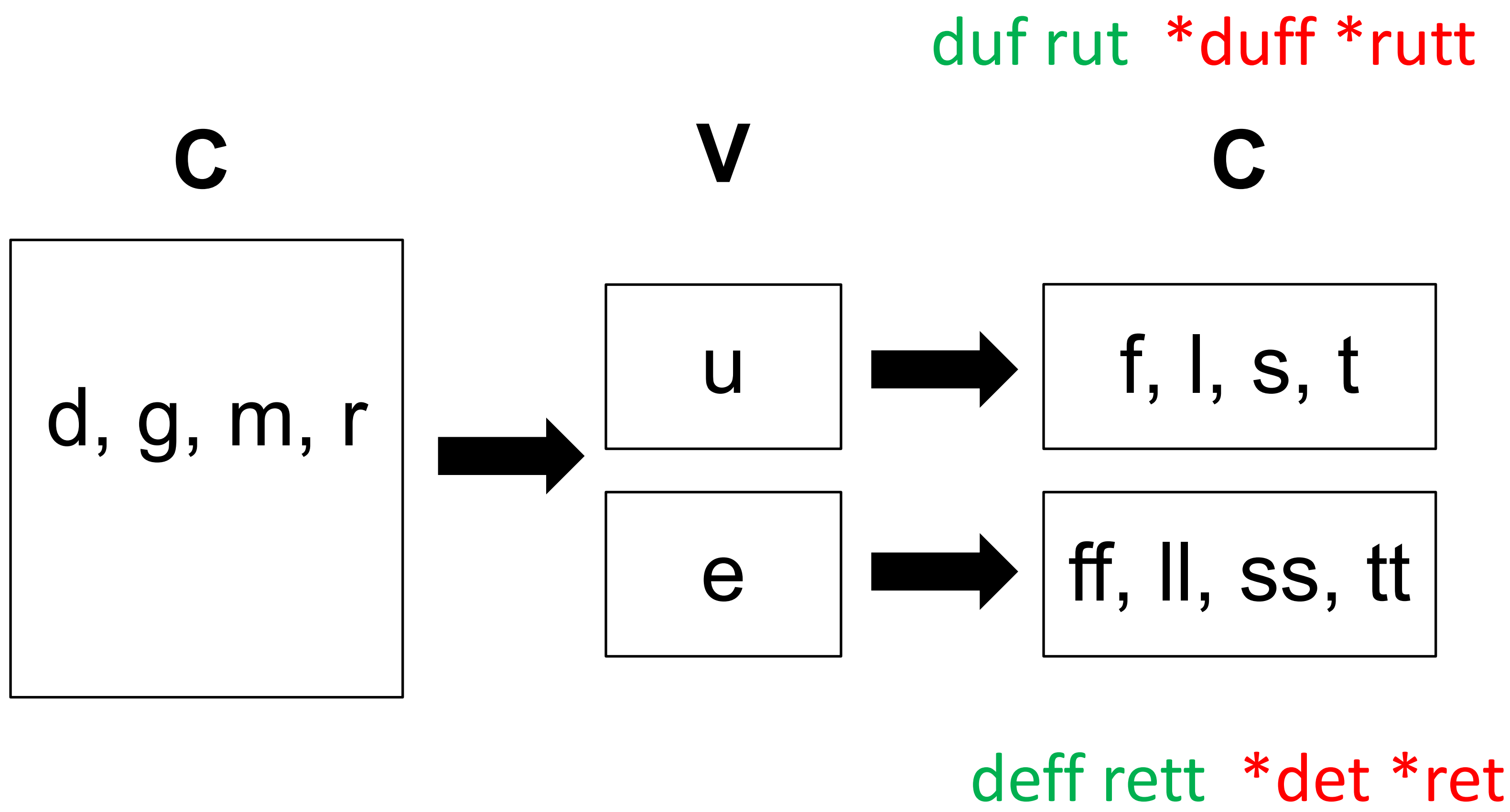


Fill-in-the-blanks

Research questions

1. Can children pick up on various frequency-based spelling patterns via statistical learning?
2. How does this ability compare to children's learning of patterns under explicit instructions?
3. Are better (statistical/explicit) learners better readers or spellers?

Training items



Test items

Generalization performance tested via production or judgments of unseen stimuli that either conform to pattern of violate patterns

LU	IL
rus	guff
mett	mel

frame
r_s
m_tt

Task order counterbalanced across participants and lists

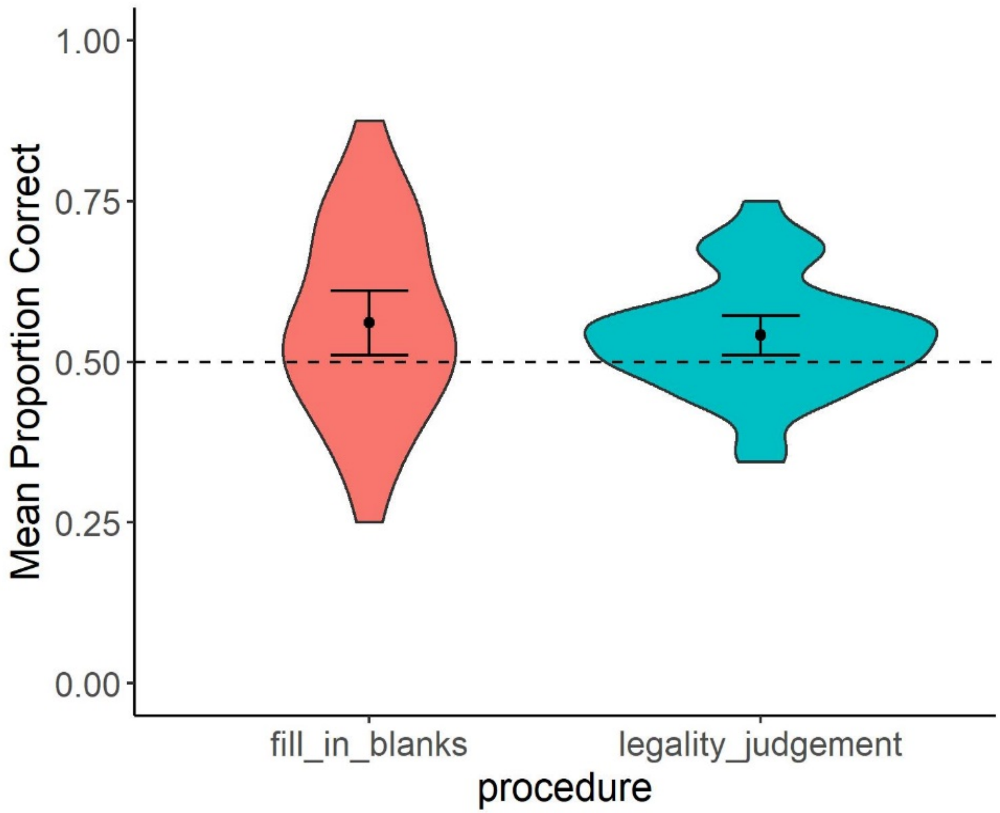
BF specification

predictions of H1 were modeled as a half-normal distribution with a SD of x

Question	**H_0**	**H1**
1. Above chance incidental learning?	• **Judgments:** chance (50%) accuracy in ability to discriminate legal from illegal items • **Production:** chance (50%) accuracy in producing correct vowel	• **Judgments:** 54% correct; **rough estimate** from Samara et al. (2019) used as *SD* • **Production:** 60% correct; rough maximum from pilot study. *SD* = max/2

Note we are working in log-odds space

Incidental learning performance



- **Fill in-blanks**
  - *p* = .01
  - BF = 8.72
  - RR: [0.05, 1.40]
- **Legality judgments**
  - *p* = .01
  - BF = 21.50
  - RR: [0.03, 2.21]

35 TD children
mean age = 6.6 years

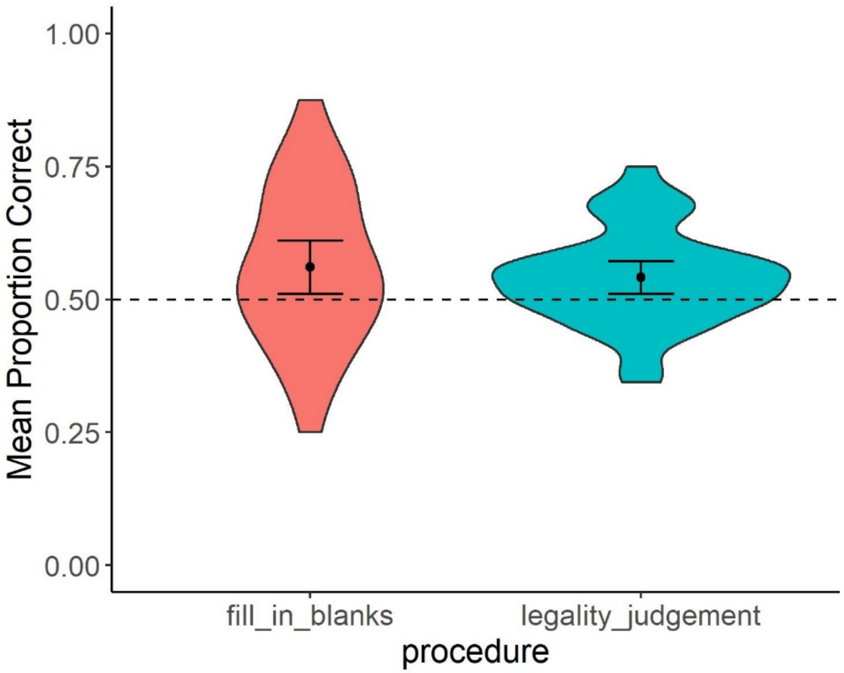
BF specification

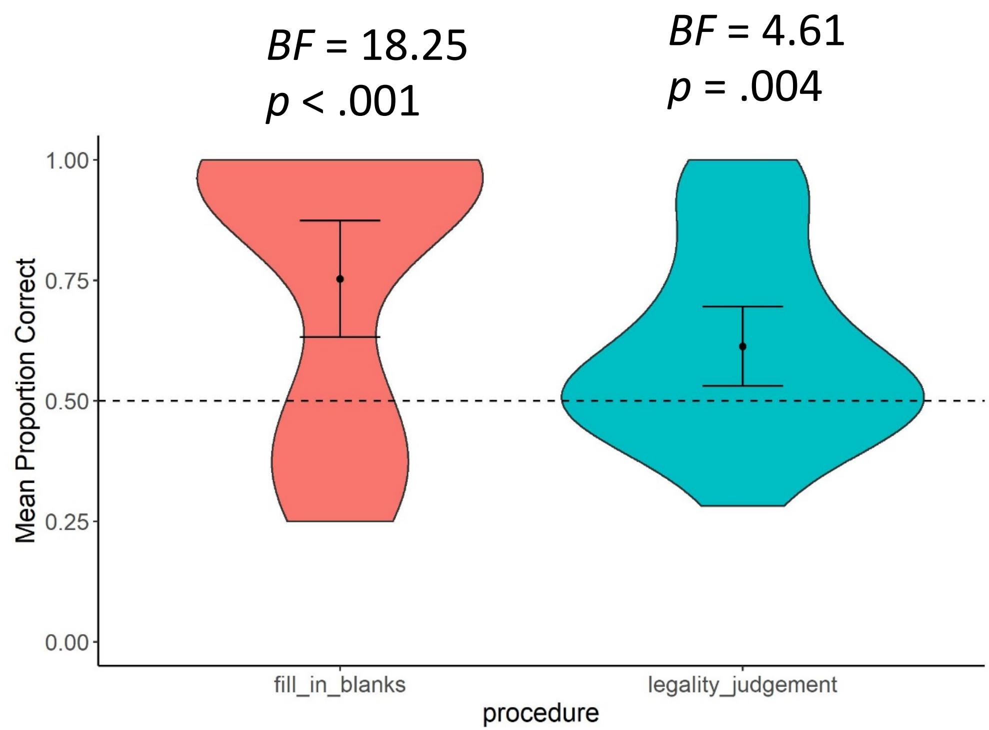
Question	**H_0**	**H1**
2. Explicit > incidental learning?	• **Production:** explicitly proportion correct = incidentally learnt proportion correct • **Judgments:** explicitly proportion correct = incidentally learnt proportion correct	**No previous data to infer a roughly predicted effect**

Specifying plausible maximum

- **Plausible maximum of Explicit minus Implicit Learning performance:**
  - Explicit performance carries the entire learning effect condition vs. incidental performance being at chance
  - Equivalent to*twice* the grand mean effect of the intercept
  - *SD* = HALF this value, i.e. grand mean effect of the intercept

Incidental vs. explicit learning





- **Fill in-blanks:** $p < .001$; BF = 389, RR: 0.10, > 4.59
- **Legality judgments:** $p = .03$, BF = 5.93, RR: 0.13, 1.32

BF specification

Question	**H_0**	**H1**
3. Positive association correlation?	• **Explicit condition**: no correlation between task performance and measure of reading/spelling • **Incidental condition:** no correlation between task performance and measure of reading/spelling	• **Explicit condition:** roughly estimated r = .40 (from literature) • **Incidental condition:** rough maximum from Explicit condition study = .52

Explicit learning

		reading	spelling
fill-in-the-blanks	*BF* [RR]	**3.46**[b] [0.20, 0.62]	1.40 [0, >4.59]
	P	.06	.22
	z_r(SE_{z_r})	0.40(0.21)	0.26(0.21)
legality judgment	*BF* [RR]	**9.42**[b] [0.12, 2.67]	**5.74**[b] [0.14, 1.38]
	P	.02	.03
	z_r(SE_{z_r})	0.52(0.21)	0.46(0.21)

Incidental learning

		reading	spelling
fill-in-the-blanks	*BF* [RR]	0.17 [0, >4.59]	**0.33**[a] [0, >4.59]
	P	.29	.95
	z_r(SE_{z_r})	-0.19(0.18)	0.01(0.18)
legality judgment	*BF* [RR]	0.76 [0, >4.59]	0.39 [0, >4.59]
	P	.35	.80
	z_r(SE_{z_r})	0.16(0.18)	0.05(0.18)

To sum up

- Children generalize knowledge of novel statistical patterns (akin to those seen in natural orthographies) when these are presented under brief incidental conditions
- Clear advantage of explicit instruction over incidental (statistical) learning (at least for verbalizable patterns tested here)
- Substantial evidence for theory linking explicitly induced learning (generalization) performance and reading
- Some support for H_0 regarding link between incidental learning performance and reading or data insensitivity

Bringing it all together

- Worth considering Bayes factors as a measure of evidential strength
- In many cases, same conclusion is supported by frequentist and BF analyses
- But only Bayes can provide “evidence of absence”
  - In many cases, proving the null is of theoretical interest
  - At minimum, important to tell apart null findings from instances of data insensitivity
- Thinking about your priors is good exercise!
  - Report openly and consider preregistering to further mitigate concern!

Thank you for your attention!

a.samara@greenwich.ac.uk