What do the experts know?

A hierarchical Bayesian approach for assessing and aggregating the predictions of forensic handwriting experts

Dani Navarro, Kristy Martire, Kaye Ballantyne,
Bethany Growns

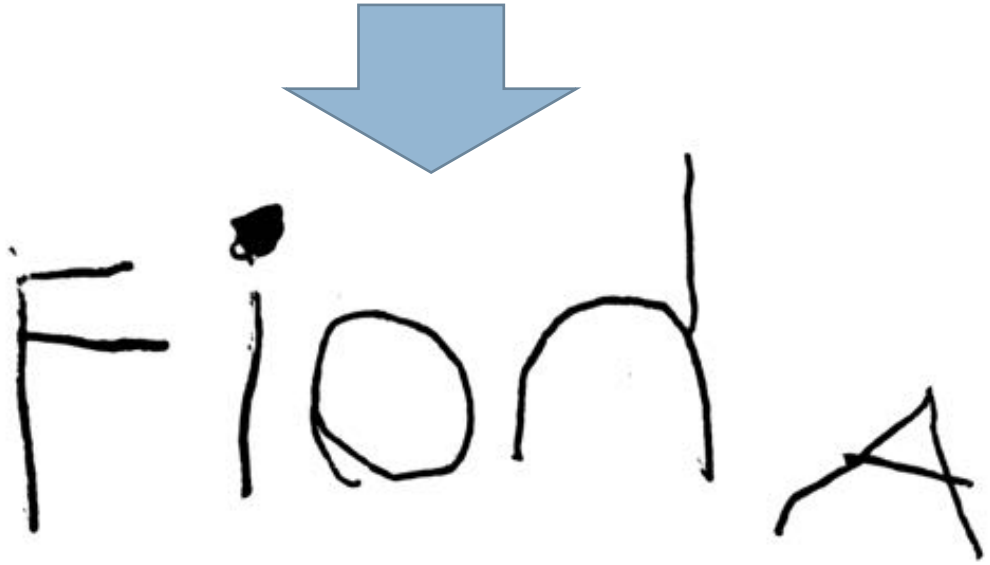
A crime has been committed

A crime has been committed

We have suspects

A note is found near
the crime scene

The police have a sample of handwriting from one of our suspects



The *process* problem: were these written in the same *way*?

Is the author trying to "disguise" their handwriting or mimic someone else's… or is it a "natural" process?

The *authorship* problem: were these written by same *person*?

Can you tell the difference between a forgery, a genuine sample, and someone trying to disguise their identity?

The *feature match* problem: what are the relevant features, and do the samples match?

The *feature probability* problem: how likely is it that a random sample of handwriting has this feature?

On the *feature probability* question

What have the document examiners (implicitly) learned about the statistics of the environment?

How common is it to see a backwards "n"?

How common is it to see backwards sloping letters?

Probability judgment is linked* to authorship judgment: the evidentiary value of a matching feature depends on how commonplace it is…

* Importantly though, they're not the same thing

An opportunistic data collection exercise

Kristy Martire

Bethany Growns

JOURNAL OF FORENSIC SCIENCES

Explore this journal ›

Paper

Measuring the Frequency Occurrence of Handwriting and Handprinting Characteristics[†,‡]

Mark E. Johnson Ph.D., Thomas W. Vastrick B.S., Michèle Boulanger Ph.D., Ellen Schuetzner B.A.

First published: 16 November 2016 Full publication history

DOI: 10.1111/1556-4029.13248 View/save citation

Cited by (CrossRef): 0 articles Check for updates Citation tools ▾

2

Funding Information

[†]Presented at the 67th Annual Meeting of the American Academy of Forensic Sciences, February 16-21, Orlando, FL.

[‡]Funded by a grant from National Institute of Justice, Award Number 2010-DN-BX-K273.

View issue TOC
Volume 62, Issue 1
January 2017
Pages 142–163

Abstract

The premise of this study was to take a valid population sampling of handwriting and handprinting and assess how many times each of the predetermined characteristic is found in the samples. Approximately 1500 handwriting specimens were collected from across the United States and pared to obtain a representative sample of the U.S. adult population according to selected demographics based on age, sex, ethnicity, handedness, education level, and location of lower-grade school education. This study has been able to support a quantitative assessment of extrinsic and intrinsic effects in handwriting and handprinting for the six subgroups. Additional results include analyses of the interdependence of characteristics. This study found that 98.55% of handprinted characteristics and 97.39% of cursive characteristics had an independence correlation of under 0.2. The conclusions support use of the product rule in general, but with noted caveats. Finally, this study provides frequency occurrence proportions for 776 handwriting and handprinting characteristics.

Johnson et al (2016)

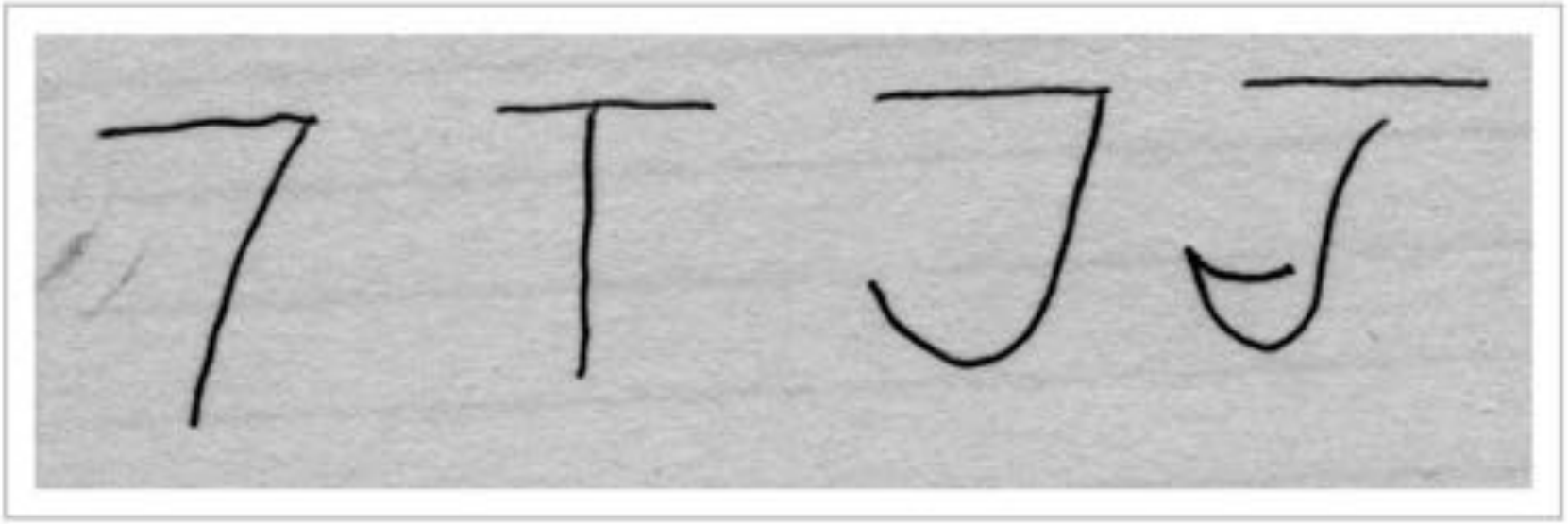


Figure 5.

Open in figure viewer | Download Powerpoint slide

Four different letter designs, all of which fulfill requirement for presence of feature CUCT 15, "disconnected cap is approximately straight." Reasons for noting presence or absence of feature are not necessarily homogenous.

Johnson et al (2016)

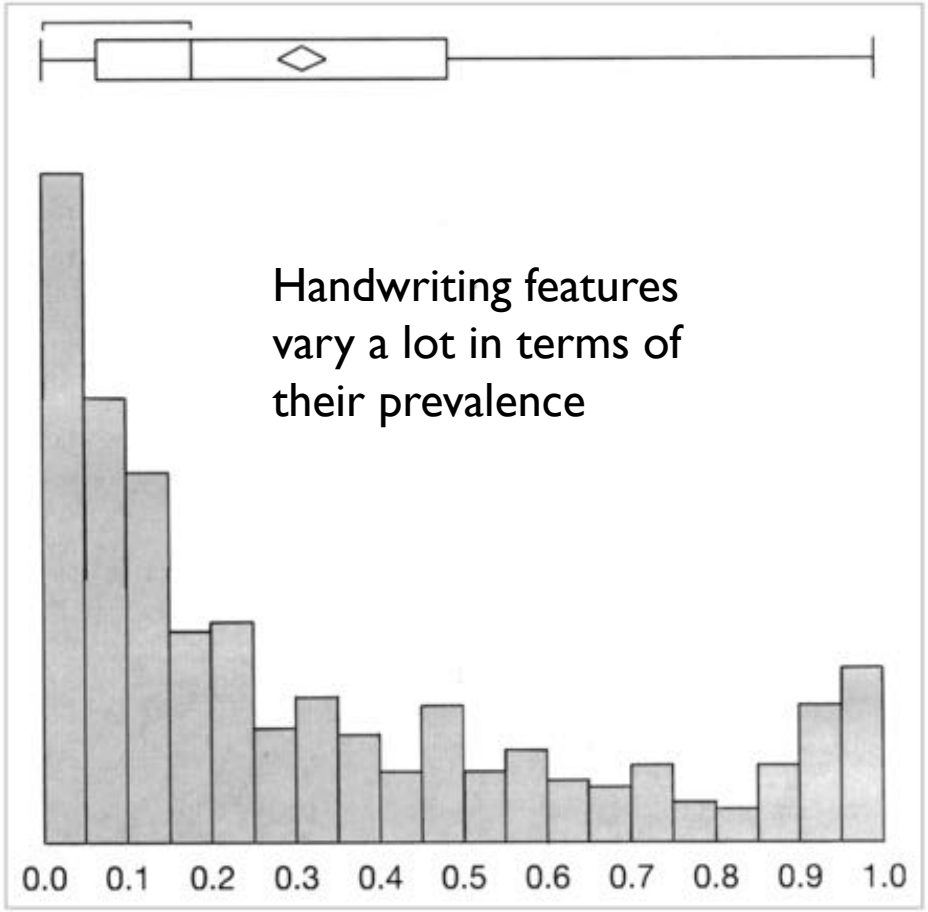


Figure 1. Open in figure viewer | Download Powerpoint slide

Histogram of features present in the cursive project sample.

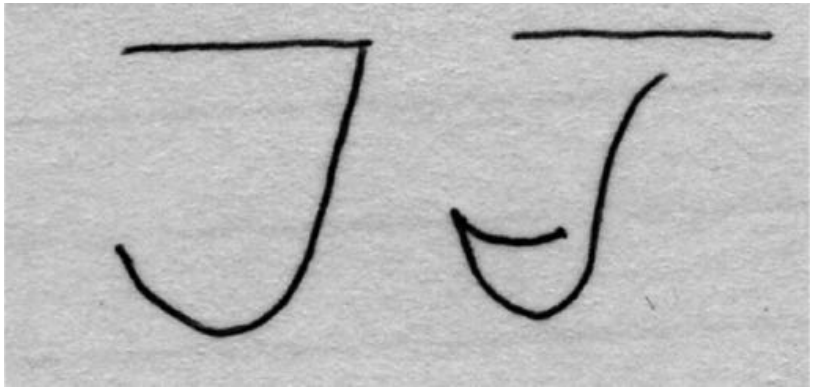
So… how much intuitive knowledge ***do*** document examiners have about feature frequency?

Johnson et al (2016)

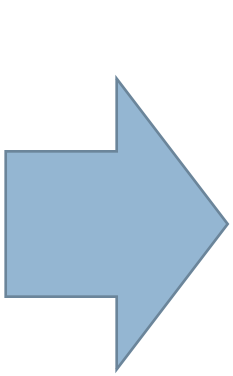
We've seen this data before the public release, so we can use it to design a study… if we're quick about it

???

What is the frequency of this feature in US handwriting?

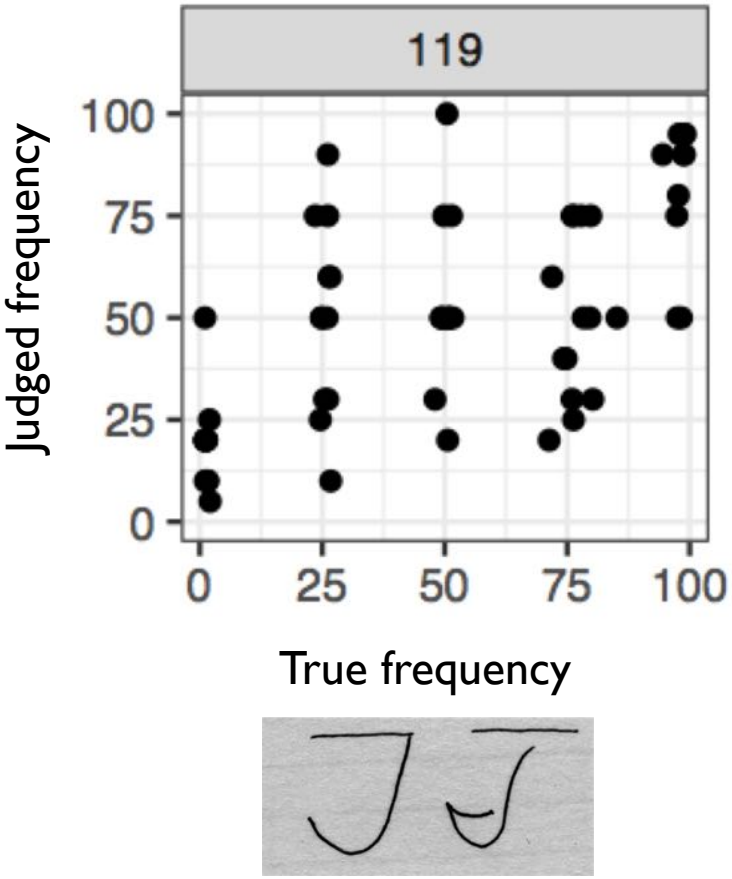


Blah blah description blah



- 60 judgments per person
- Items varied in true frequency

For example, one person might give responses like this…



US

Have some real world experience of the relevant environmental statistics?

Non US

Experience of the "handwriting world" is likely to reflect different statistics?

Experts

Have professional experience testifying about handwriting

$n = 8$	$n = 10$

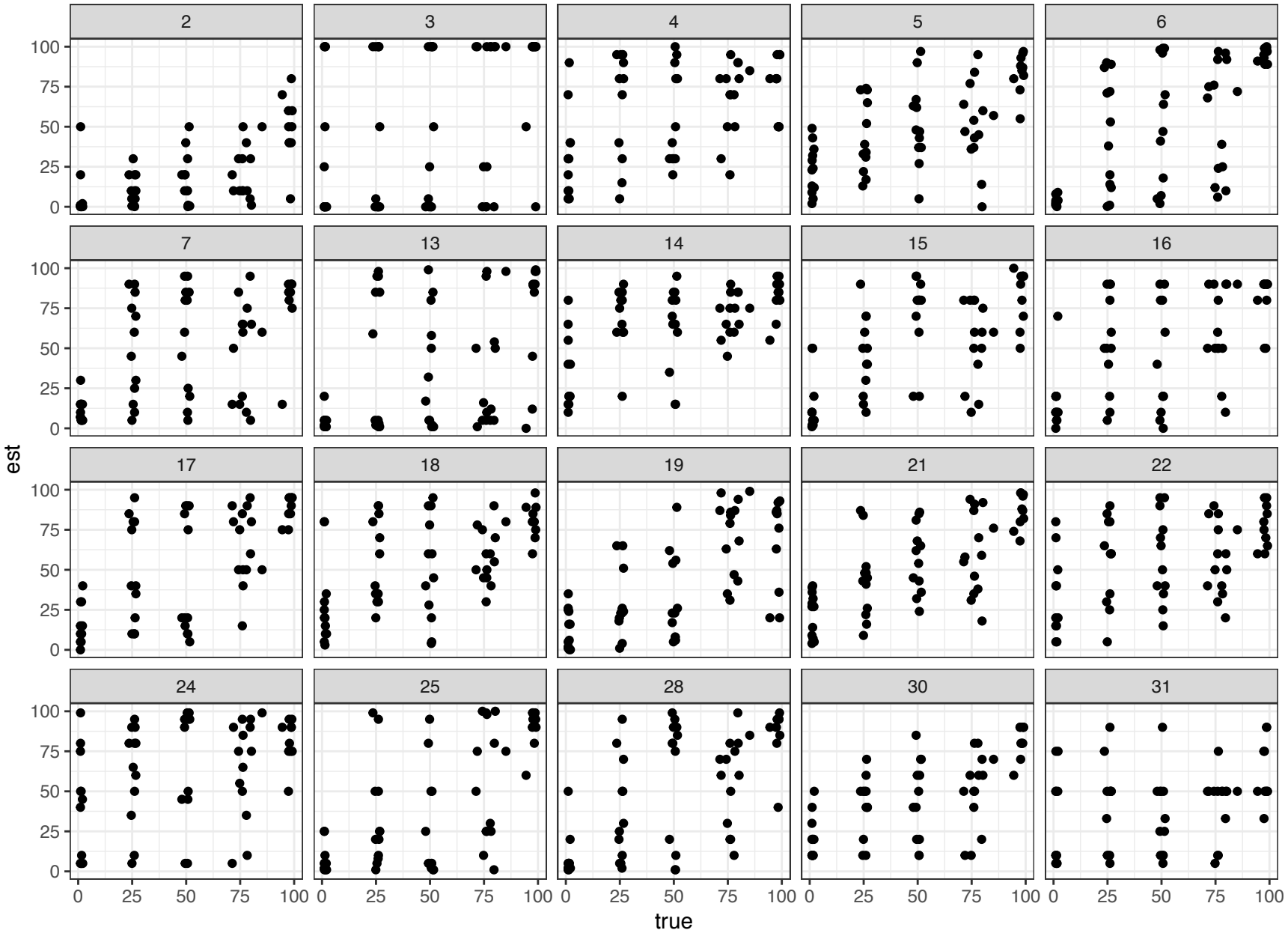
Novices

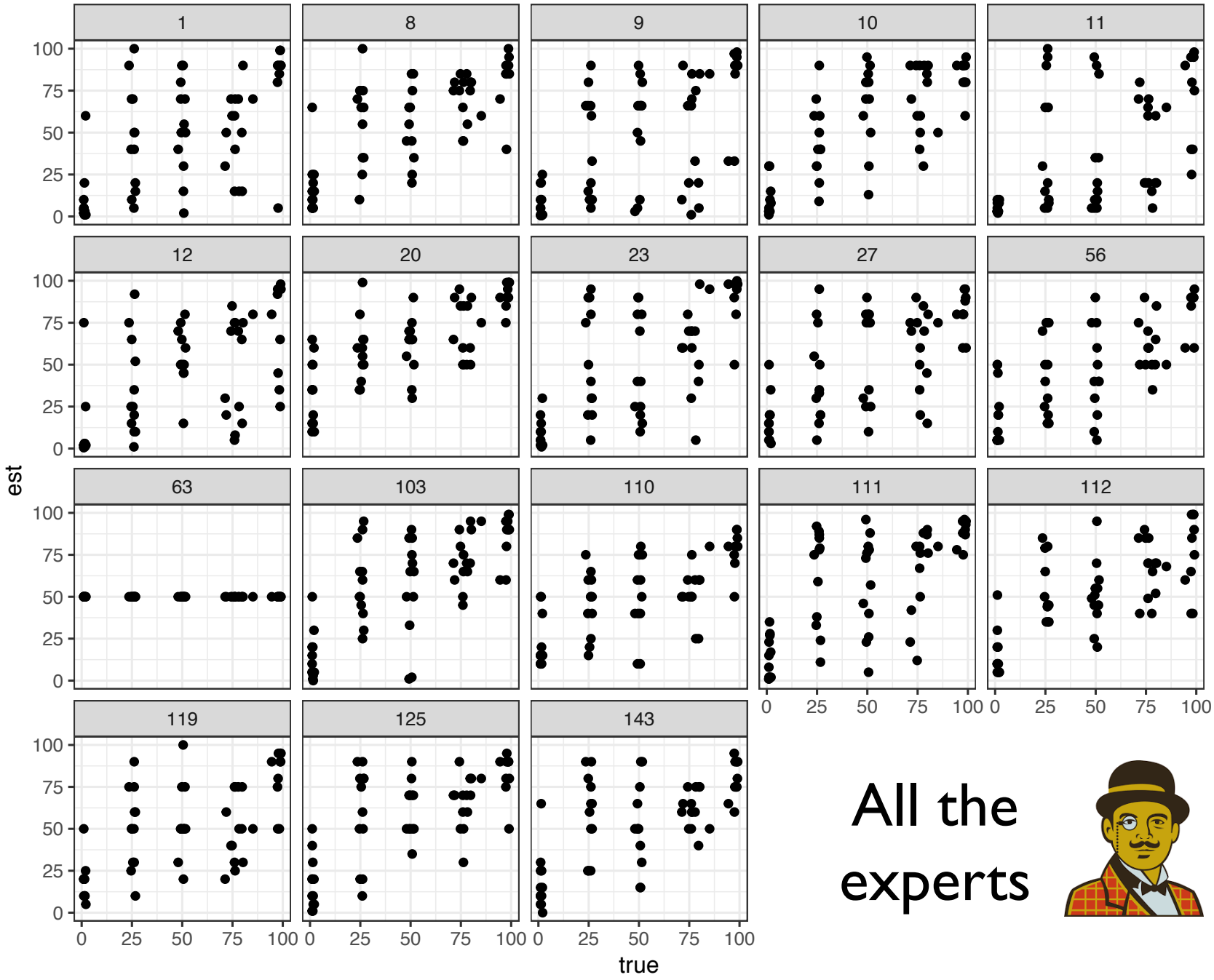
No such experience

$n = 36$	$n = 41$

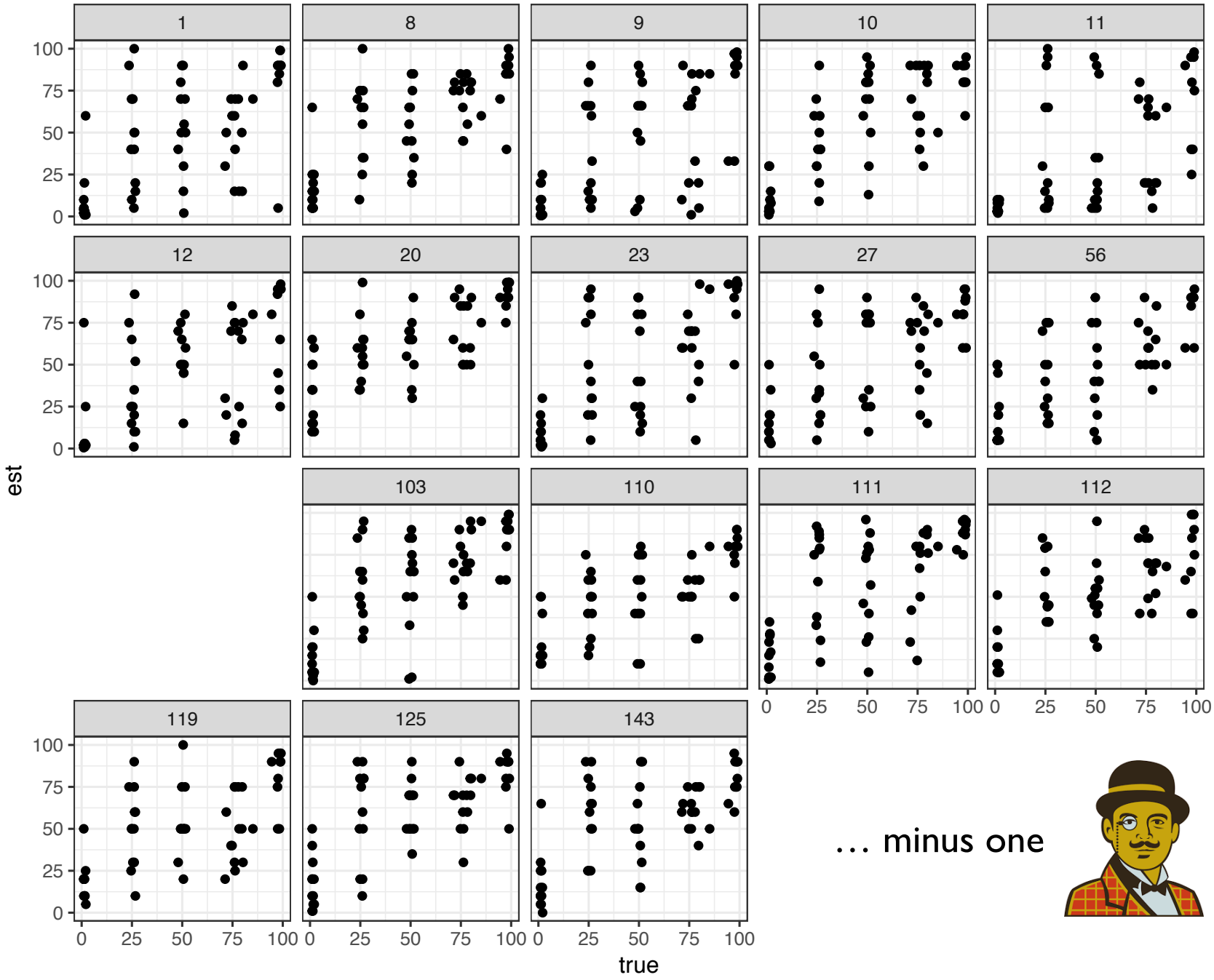
The data & some confirmatory analyses

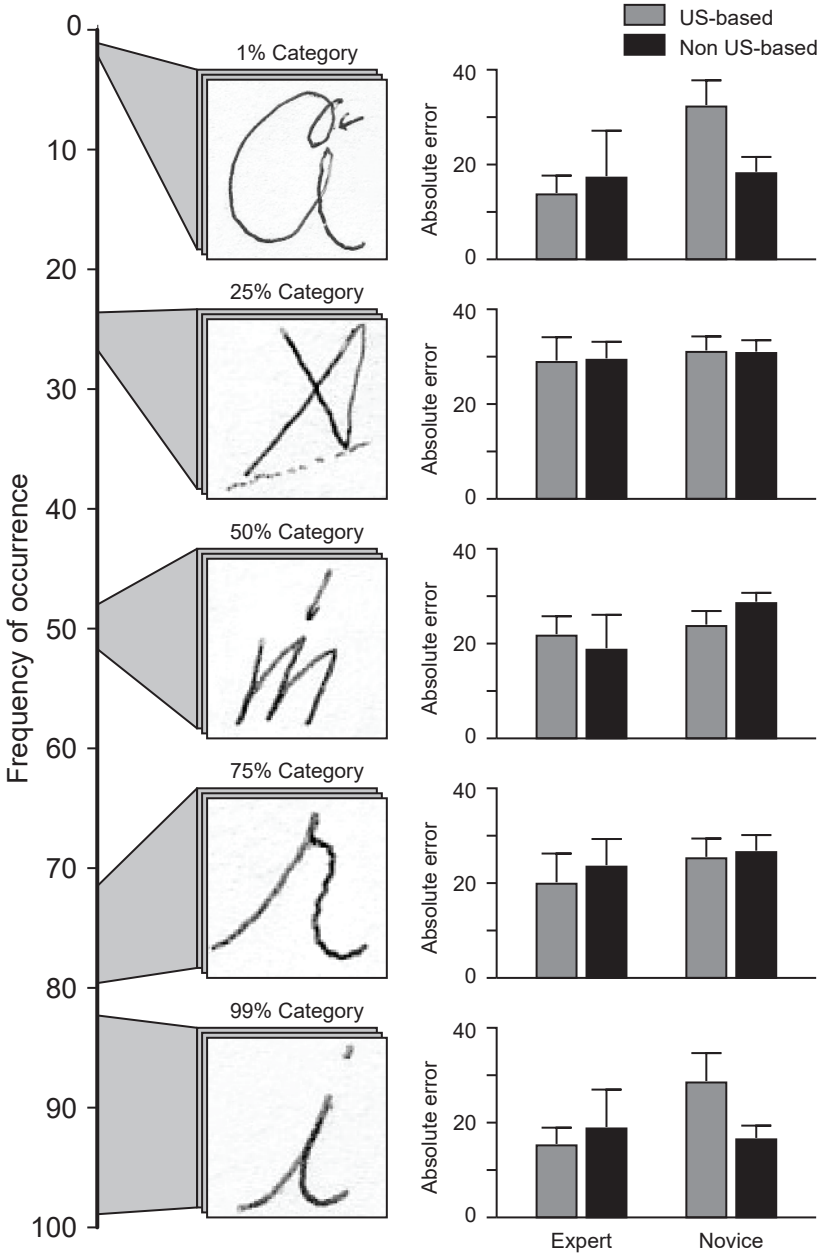
Some novices…

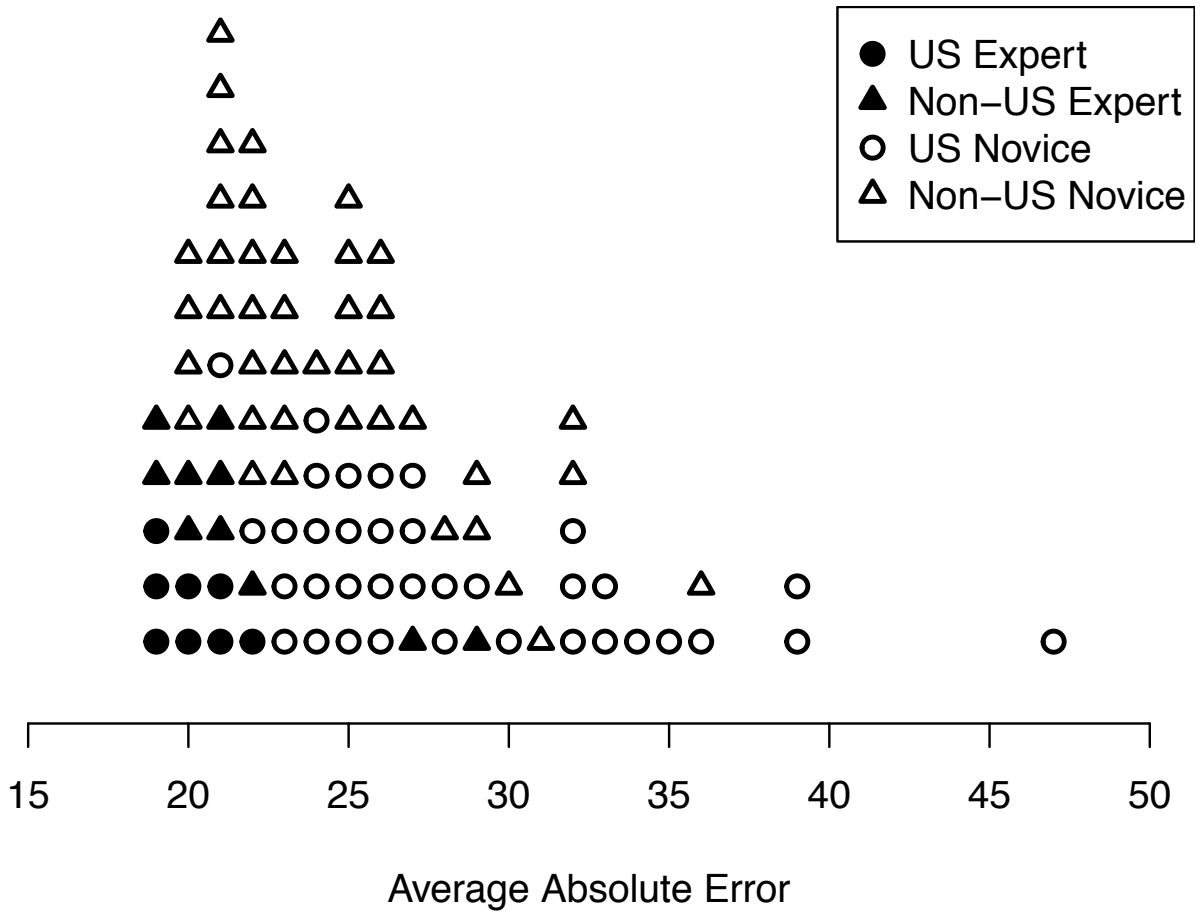




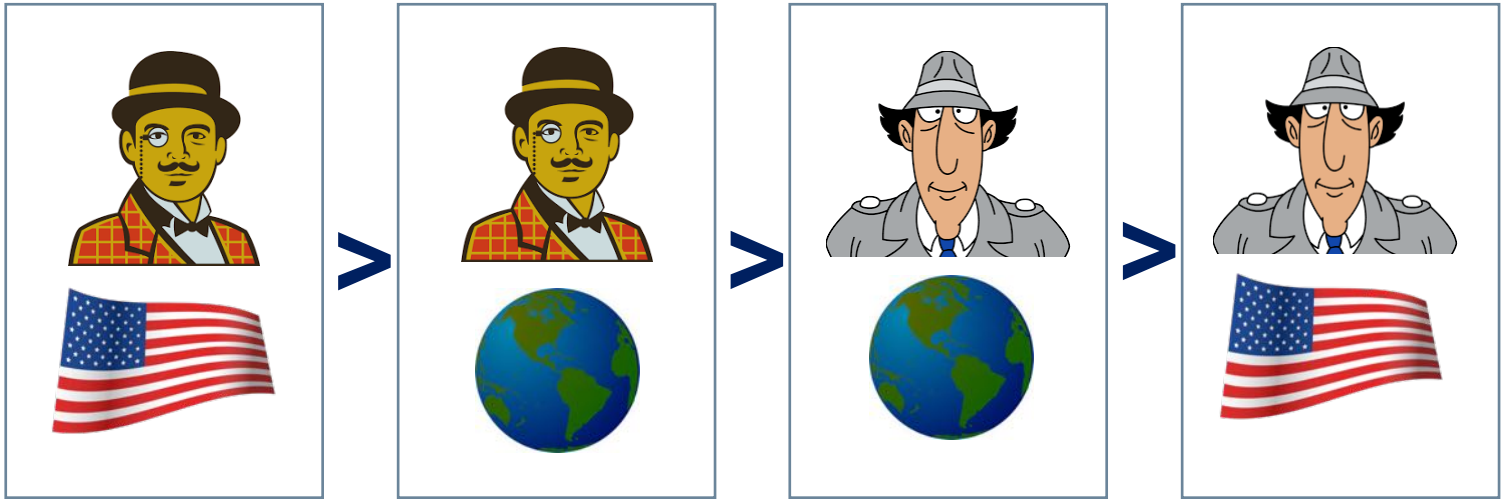
All the experts


1
8
9
10
11
12
20
23
27
56
103
110
111
112
119
125
143
est
true
0
25
50
75
100
… minus one

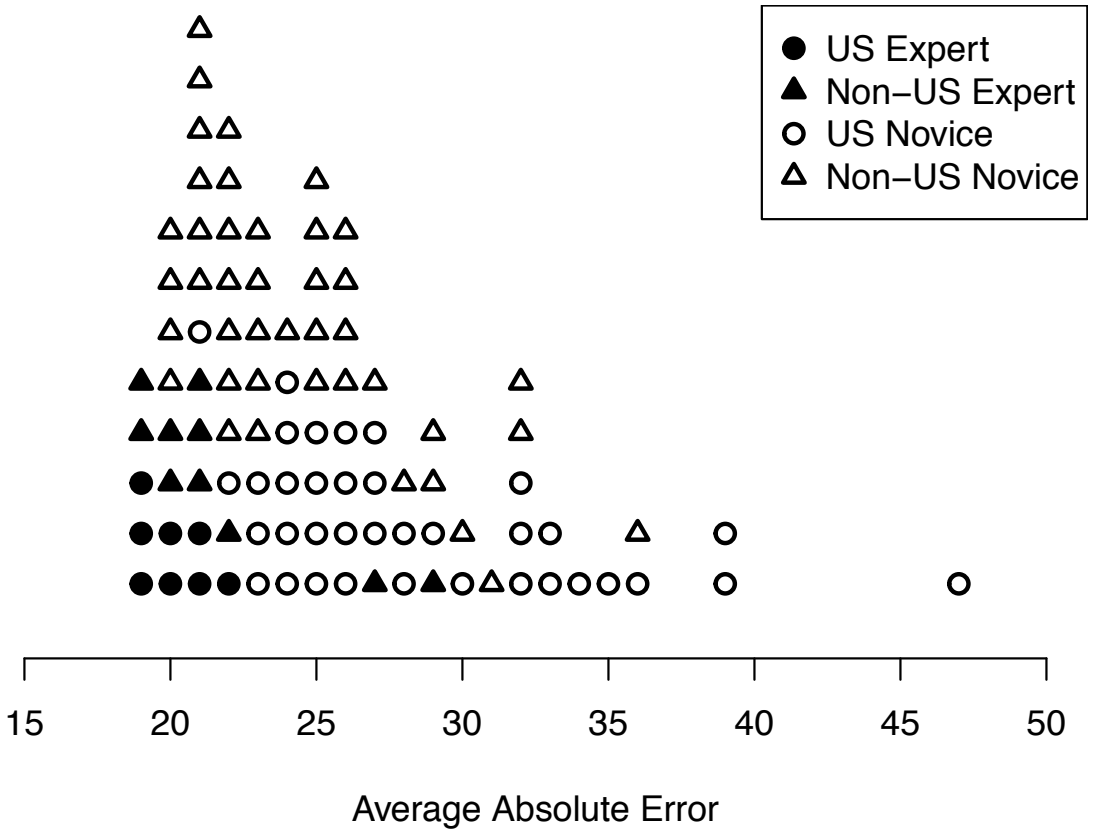

Frequency of occurrence
0
10
20
30
40
50
60
70
80
90
100
1% Category
25% Category
50% Category
75% Category
99% Category
US-based
Non US-based
Absolute error
40
20
0
Expert
Novice


US Expert
Non−US Expert
US Novice
Non−US Novice
15
20
25
30
35
40
45
50
Average Absolute Error

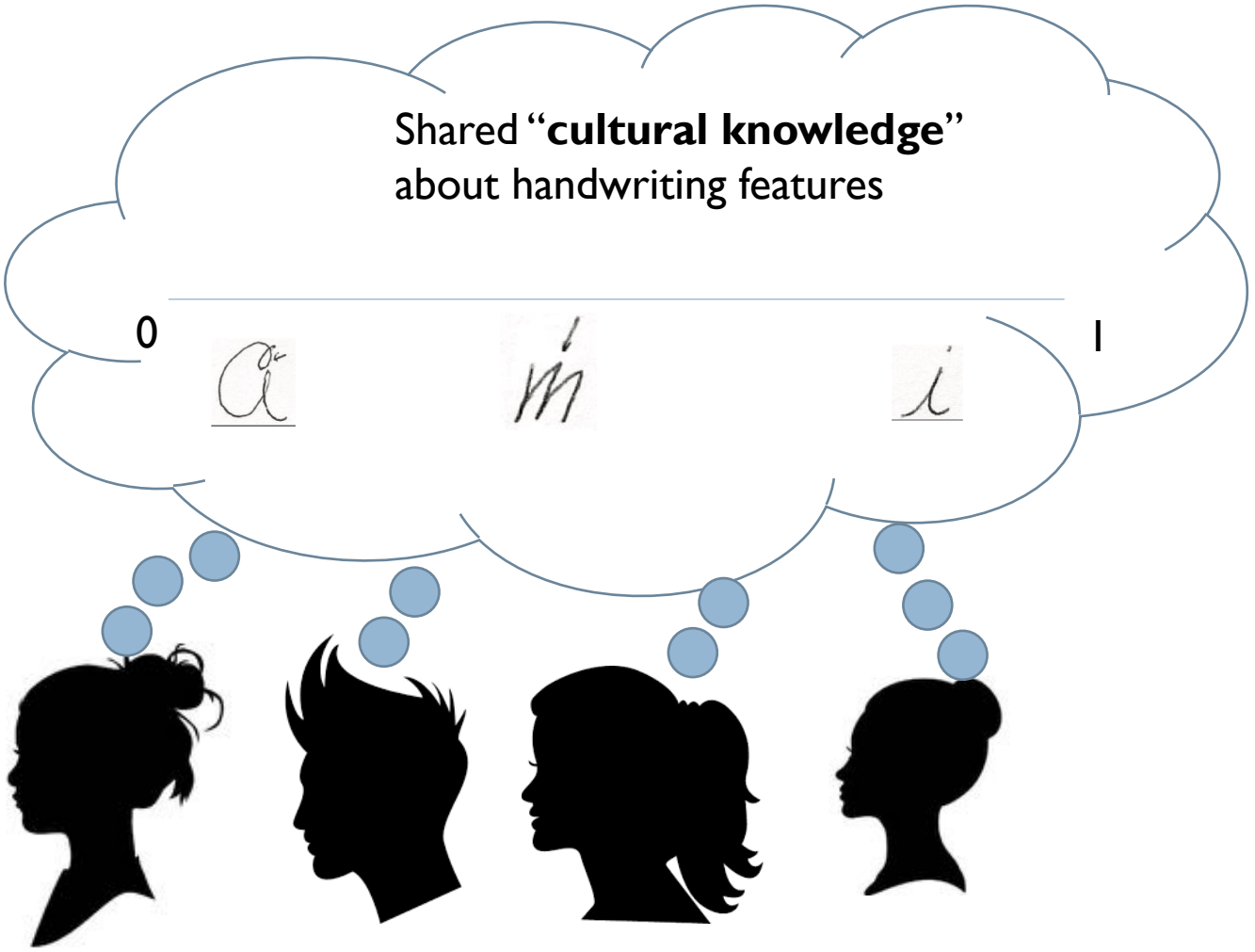
When analyzed in this fashion there is strong evidence (Bayes factor 39:1 against the baseline model including only the random effects) that the expert judges were more accurate – average error 21% on any given trial – than the novices, who produced errors of 26% on average. However, the best performing model was the 'full' model that considered all four groups (US experts, US novices, non-US experts, non-US novices) separately, with a Bayes factor of 300:1 against the baseline and 3.7:1 against a model that includes both main effects and no interaction. Consistent with this, the data show a clear ordering: the most accurate group were the US experts (20% error), followed by the non-US experts (22% error). The novices were both somewhat worse, but curiously the non-US novices performed better (24% error) than the US novices (28% error).



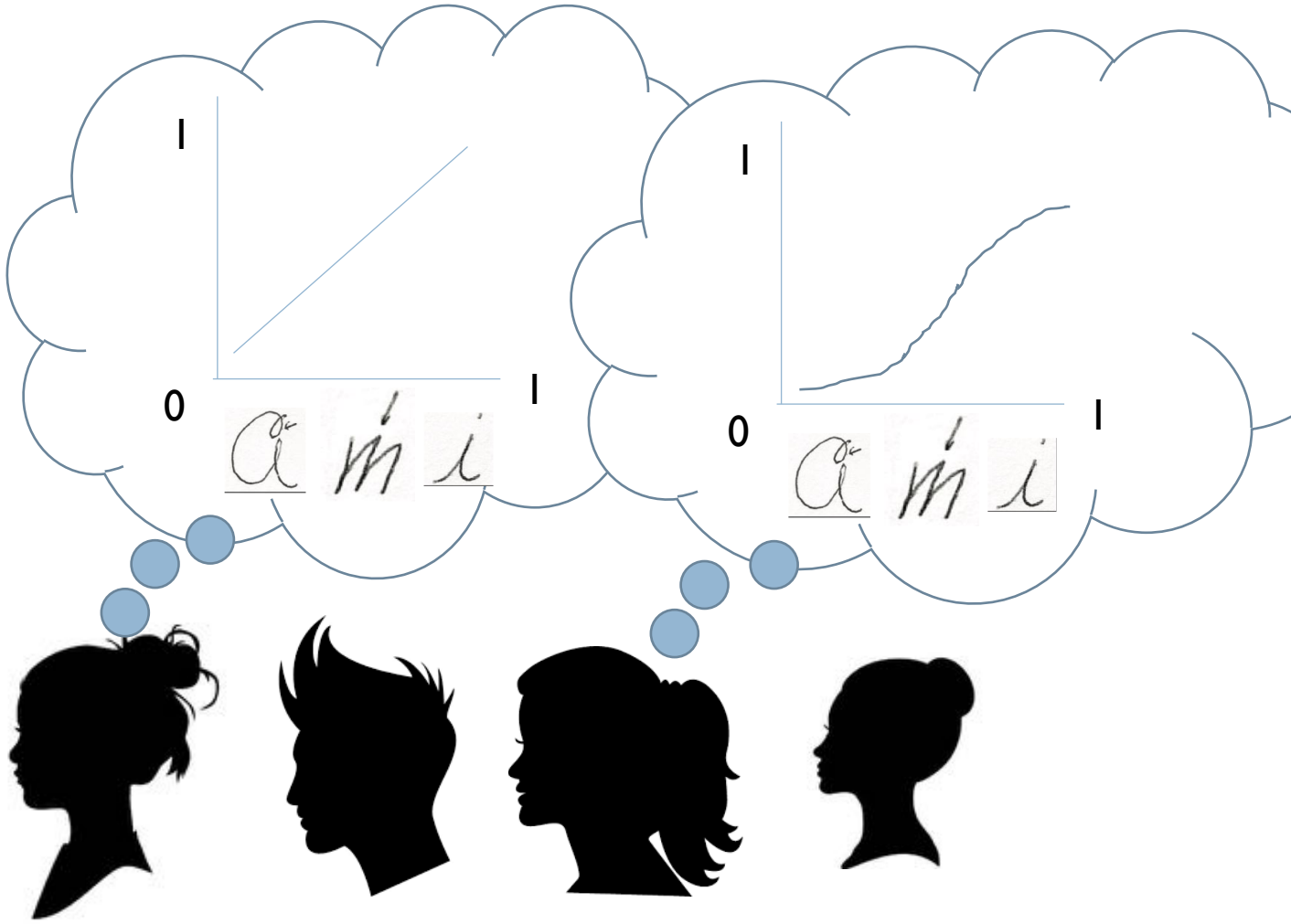
Okay, so expertise buys you something in this task… but what????



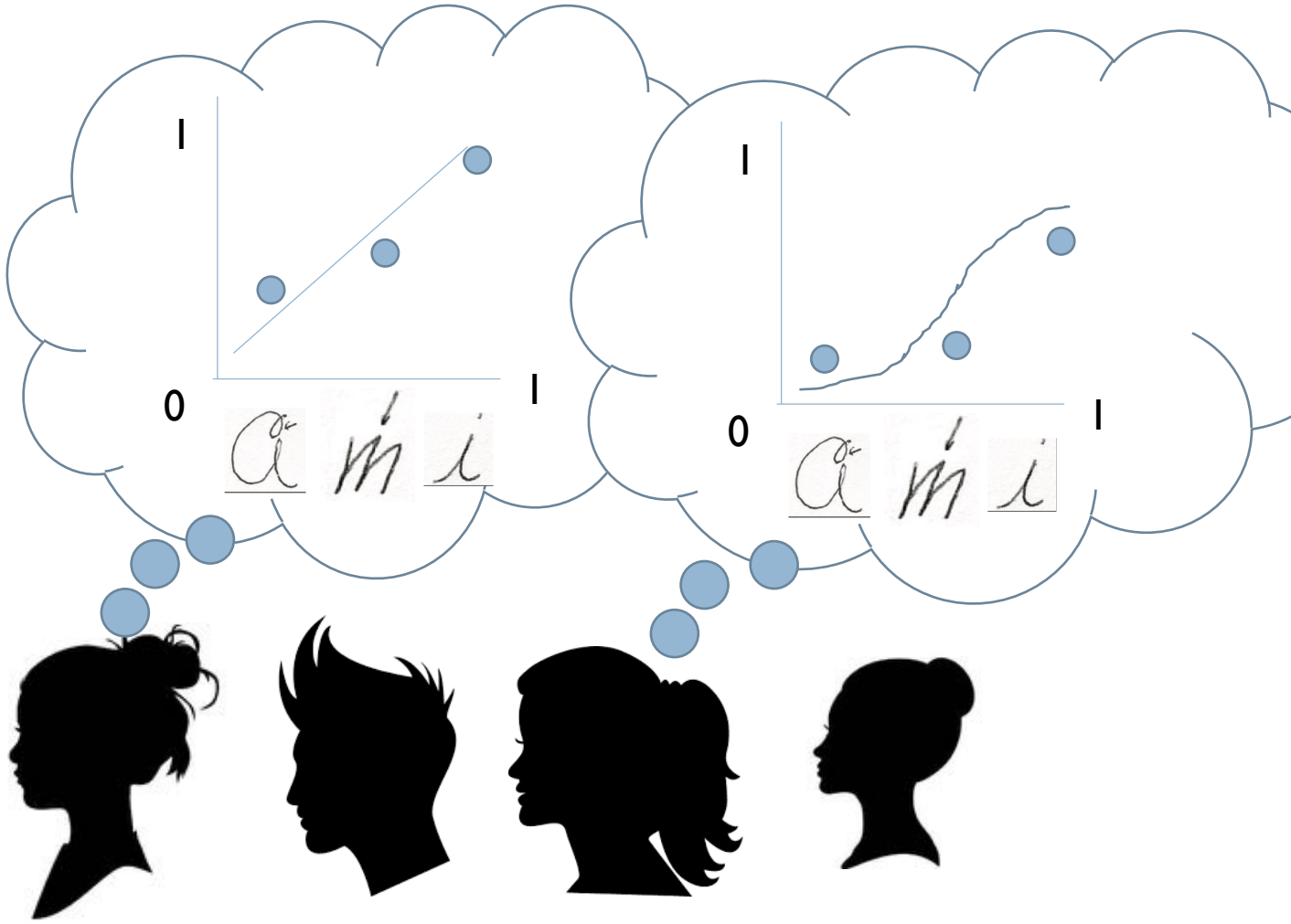
Exploratory data analysis using a hierarchical Bayesian model of probability judgment


Shared "**cultural knowledge**" about handwriting features
0
1

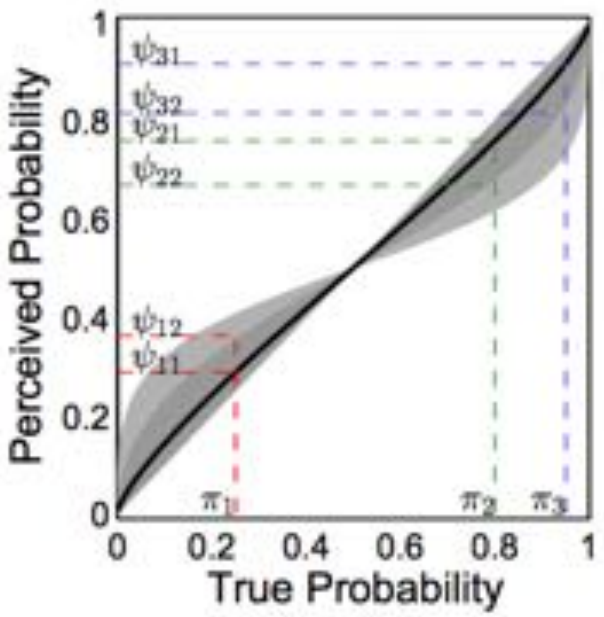
Idiosyncratic "**calibration**" function mapping beliefs to stated probabilities



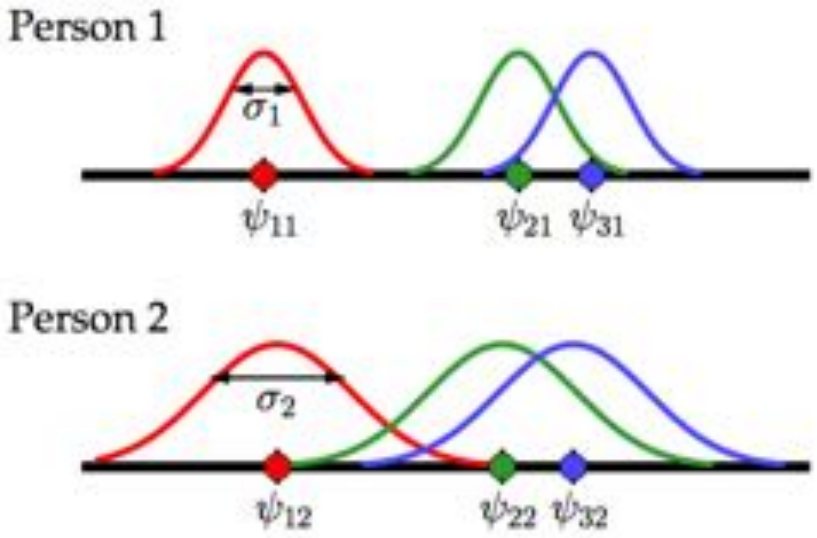
Idiosyncratic distortion (or noise) in the stated beliefs reflecting the level of precision with which each person can access the cultural knowledge



Calibration



Precision

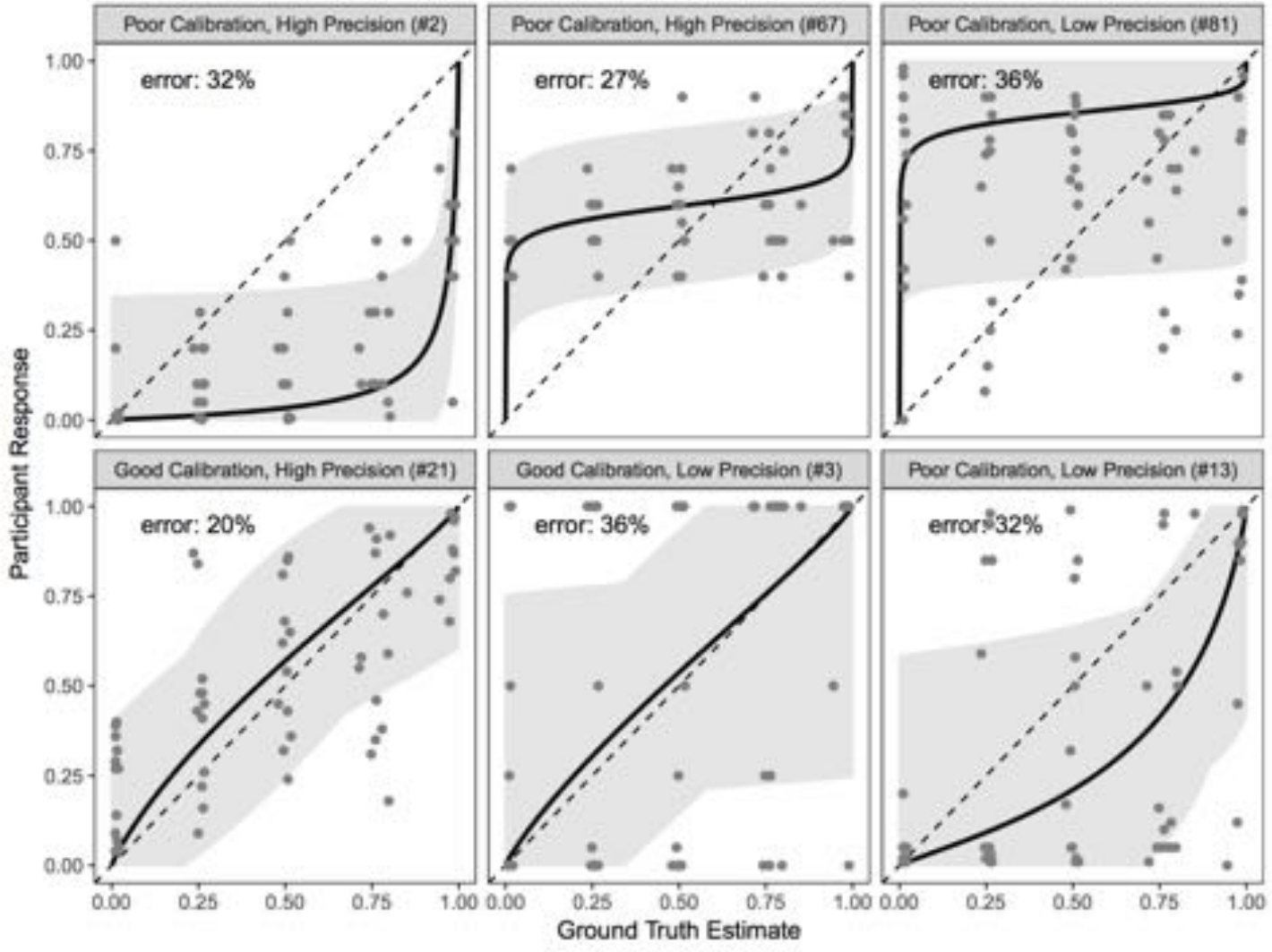


… it's essentially a version of the Bayesian Thurstonian model with a more flexible class of calibration models

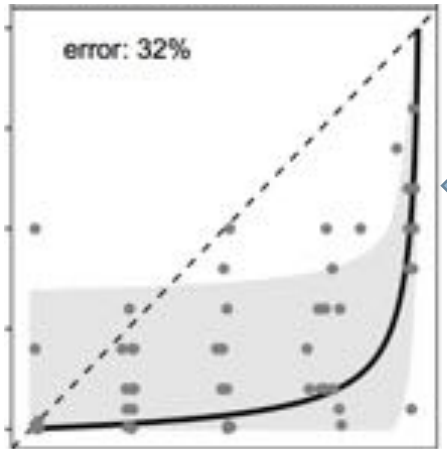
Lee & Danileiko (2014)

So what do we observe?

Analysis at the level of error rate masks a lot of individual variability



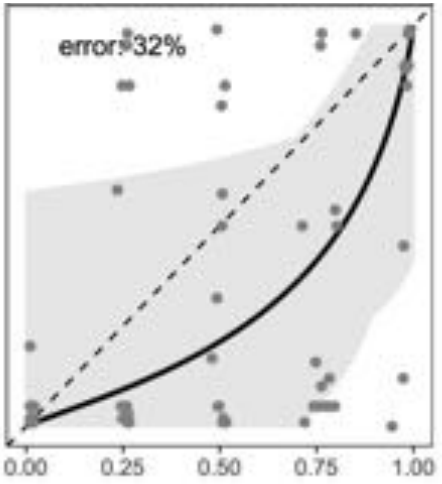
Sometimes there are tradeoffs



These people have calibration curves with the same curvilinear shape, and same overall error rate

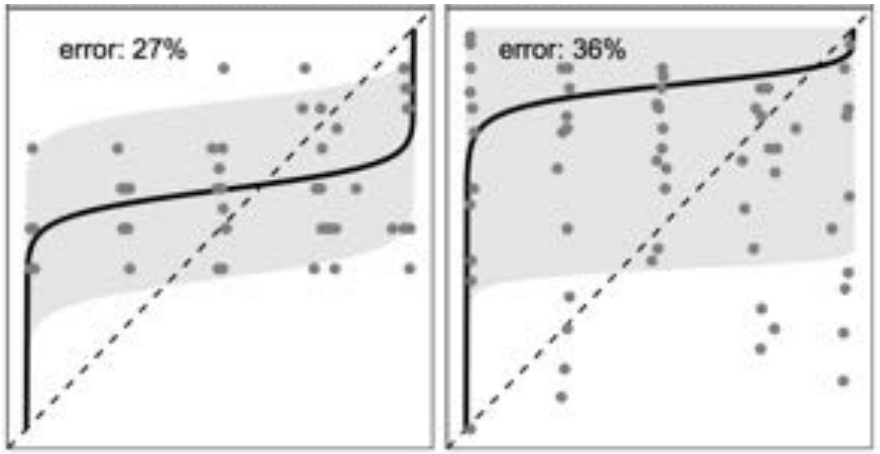
more bias,
less noise

less bias,
more noise



Sometimes one person is just better

These two have the same *shape* calibration function, but…



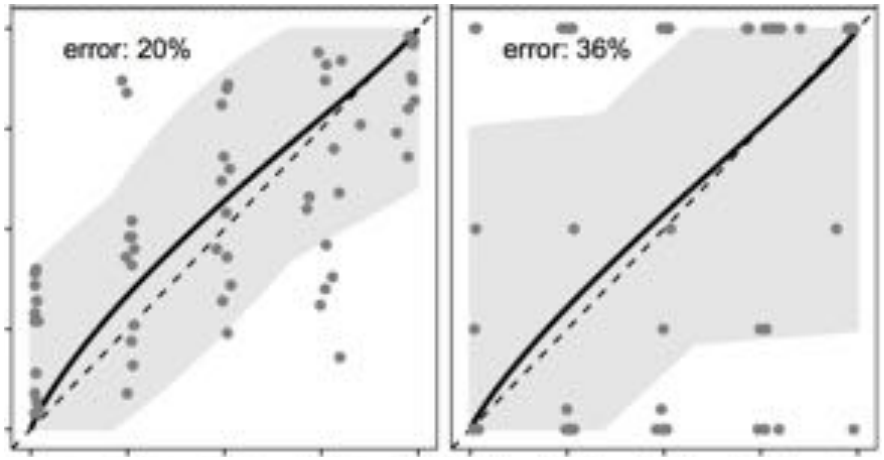
less bias,
less noise

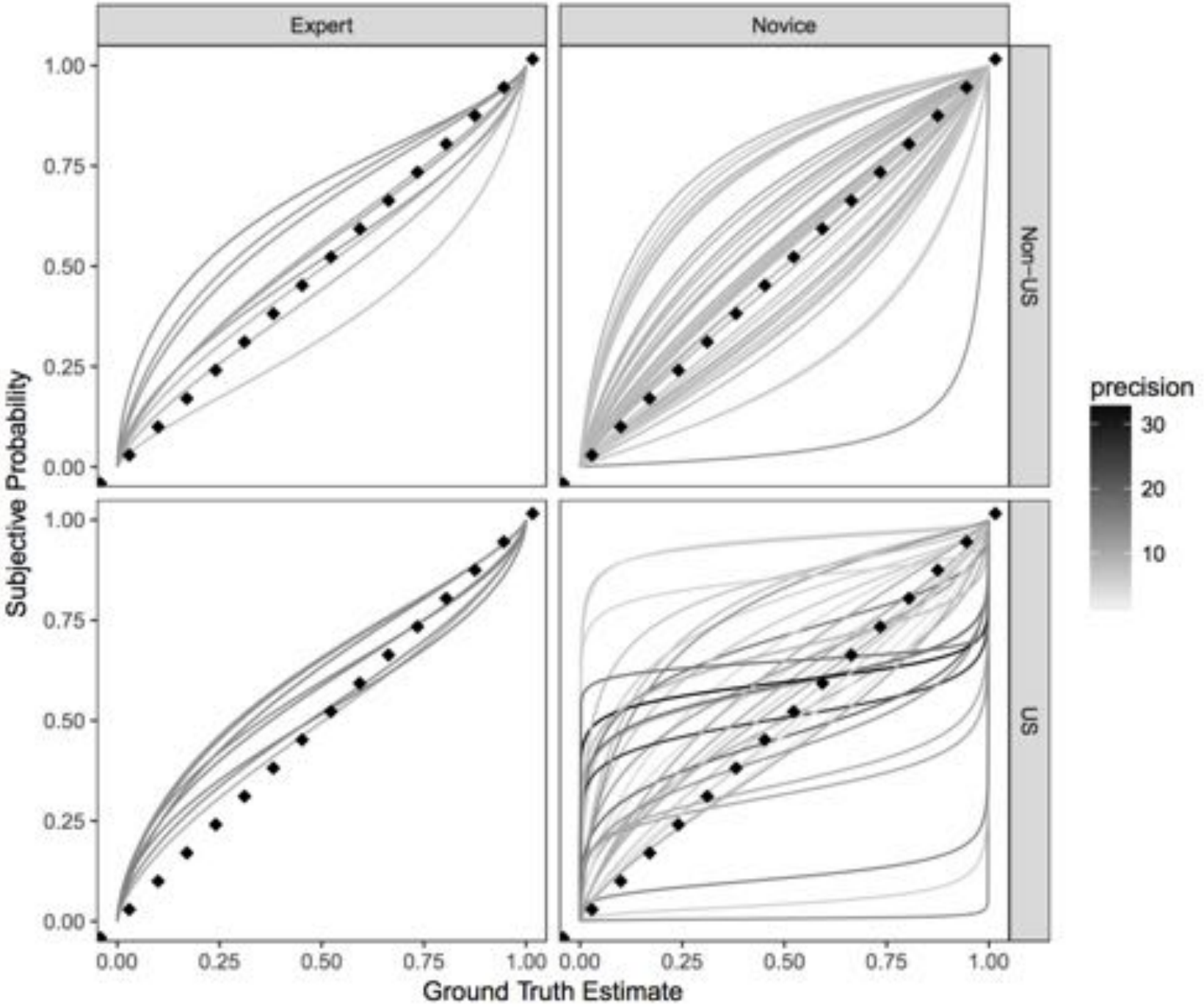
more bias,
more noise

Two people with very low bias, but very different levels of imprecision

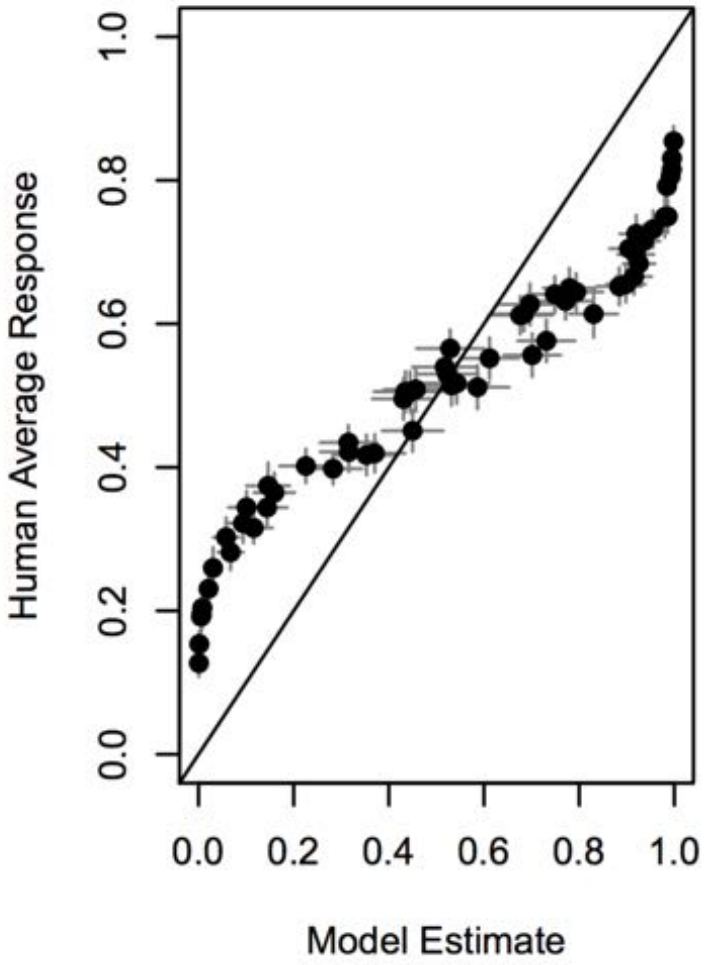
no bias, low(ish) noise

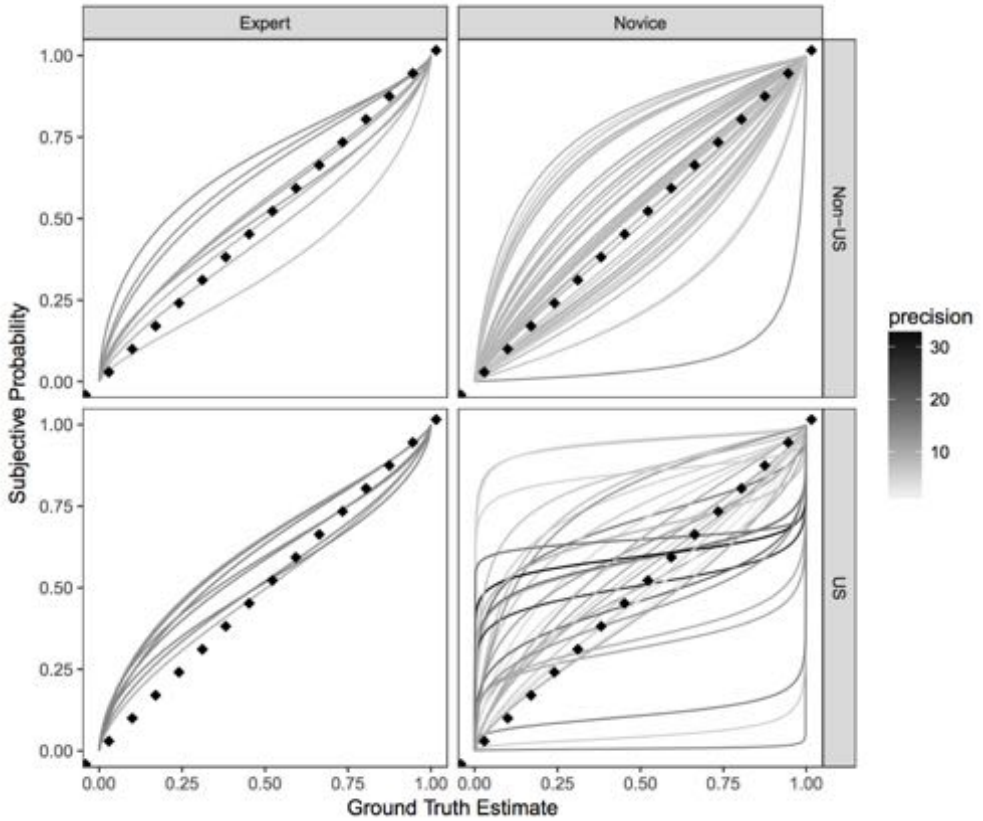
no bias, very large noise



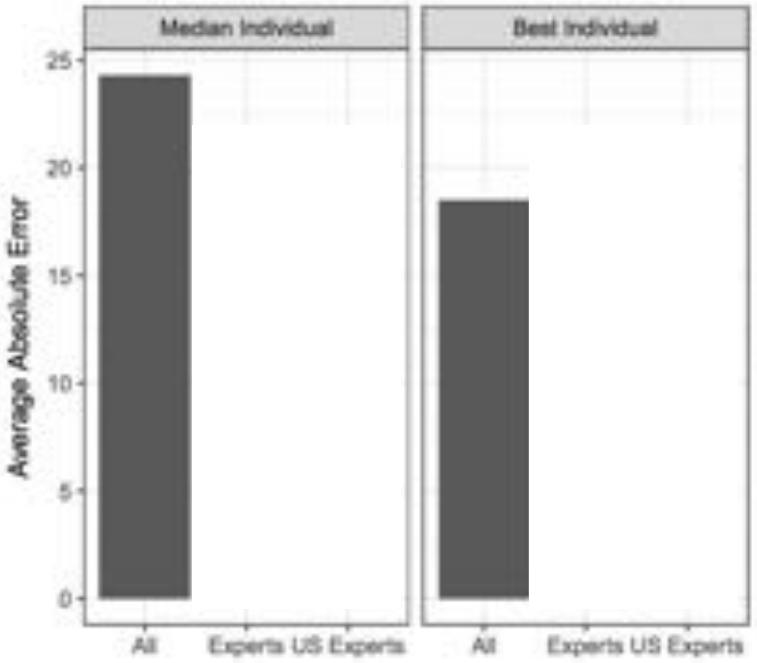

Expert
Novice
Non−US
US
Subjective Probability
Ground Truth Estimate
0.00
0.25
0.50
0.75
1.00
precision
30
20
10

Averaging responses masks the individual differences in calibration functions!

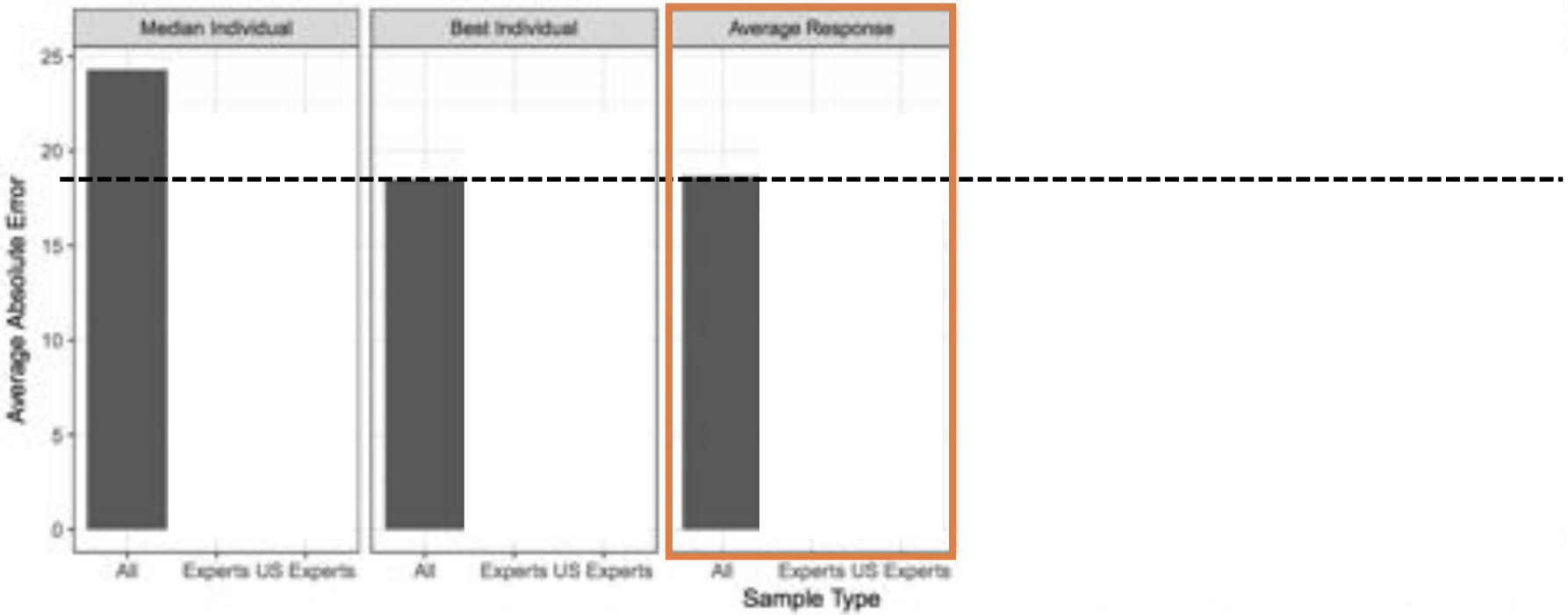




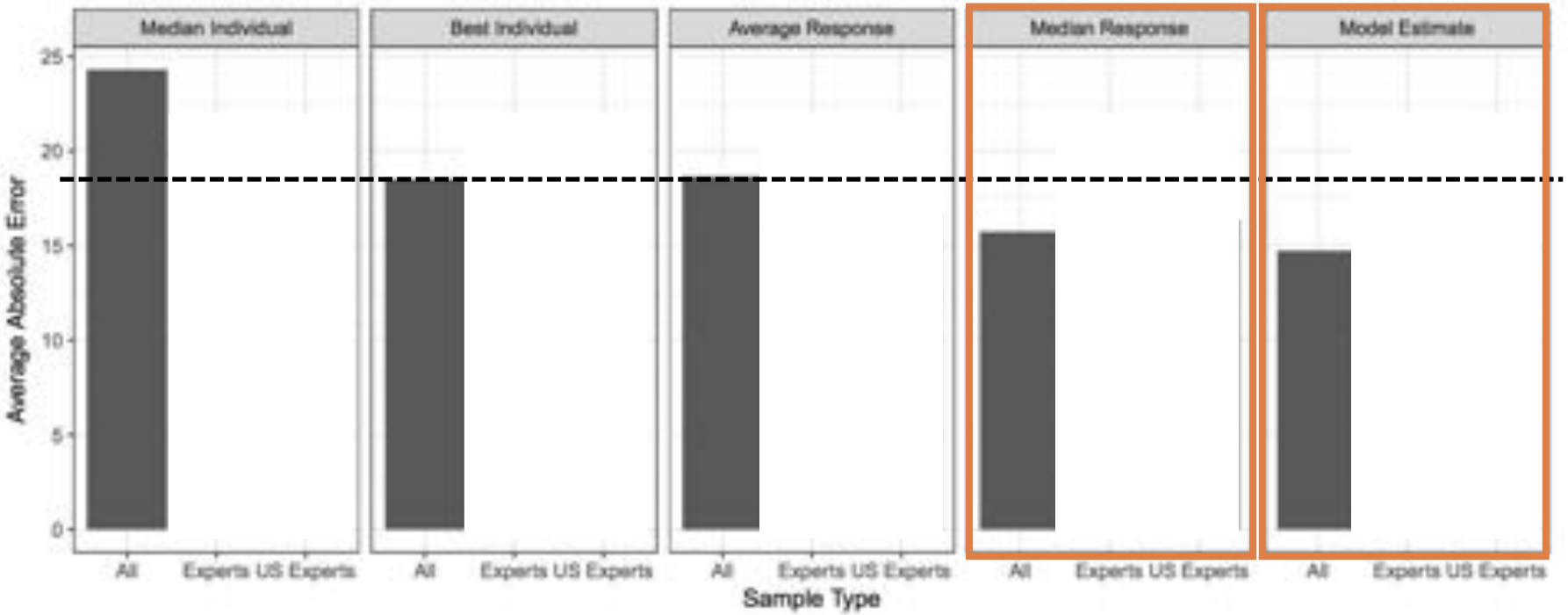
Is there a wisdom of crowds effect?



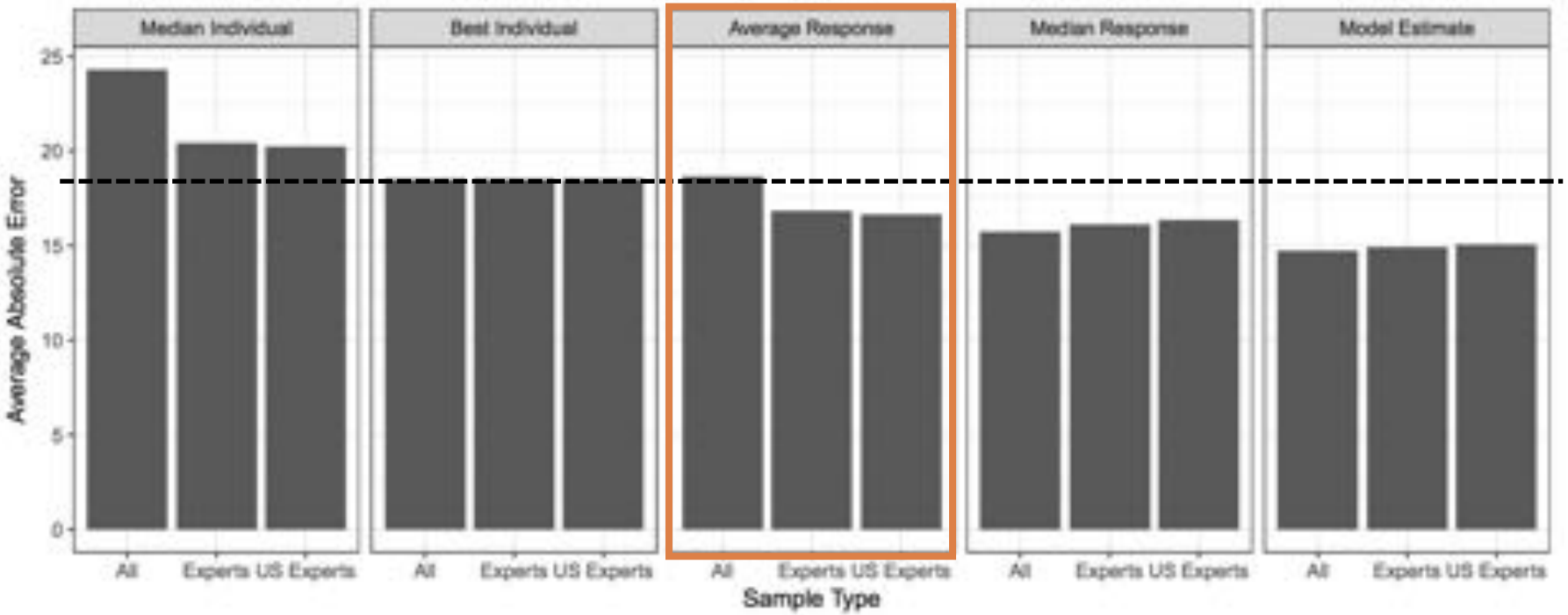
Not if you use the average response



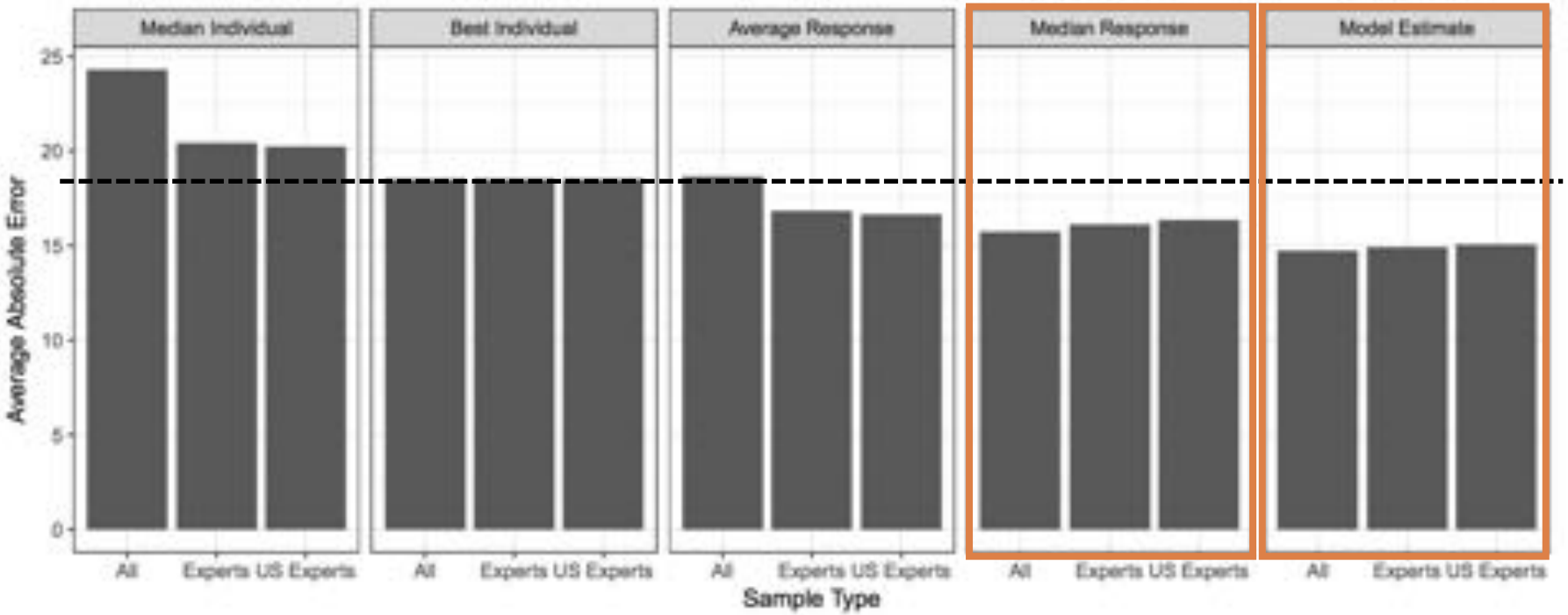
There is if you use the median or the Thurstonian model



Average recovers a bit if you know who the experts are and only include them



Median & Thurstonian are both helped slightly by the inclusion of novices?



Extending the approach when ground truth is harder to establish

Kristy Martire Kaye Ballantyne

*okay fine I couldn't find pictures, but they both own cute puppies, so…

The *feature probability* problem allows straightforward benchmarking… get some handwriting, count the features!

JOURNAL OF FORENSIC SCIENCES

Explore this journal >

Paper

Measuring the Frequency Occurrence of Handwriting and Handprinting Characteristics[†,‡]

Mark E. Johnson Ph.D., Thomas W. Vastrick B.S., Michèle Boulanger Ph.D., Ellen Schuetzner B.A.

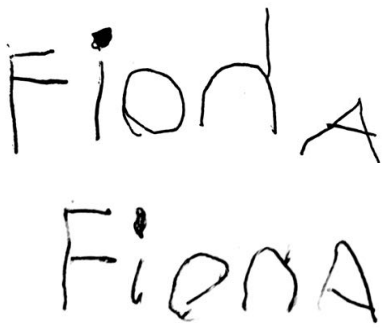
The *authorship* problem is trickier…

… are these equally "difficult" decisions? How would you know?

… and of course the difficulty generalises to the *process* problem too

Were these written by the same person?

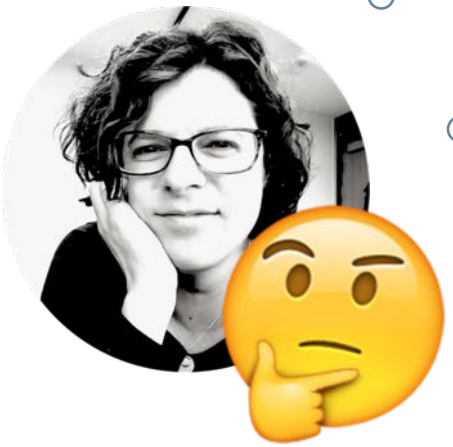
(*I'm soooo oversimplifying the data collection)



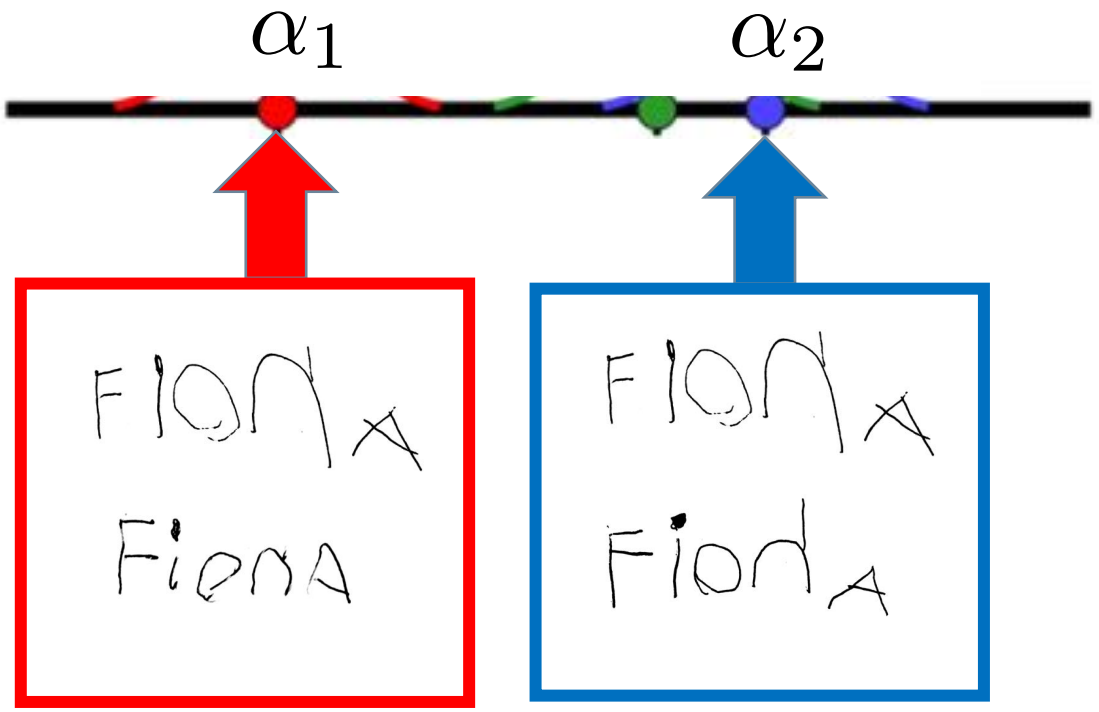
1. Very strong support for "yes"
2. Qualified support for "yes"
3. Evidence is inconclusive
4. Qualified support for "no"
5. Very strong support for "no"

Psst… Likert scales are tricky, so don't screw this up?

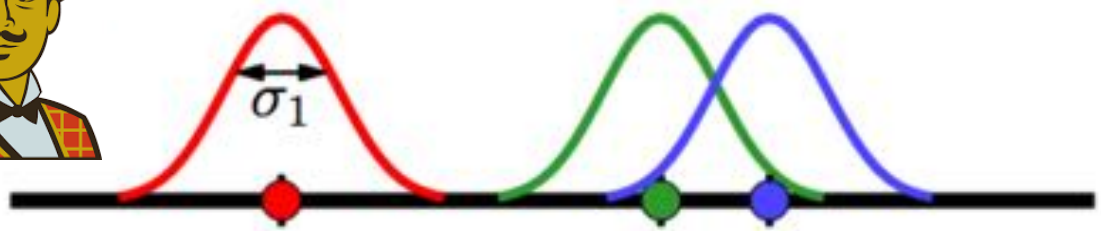
Wisdom of crowds models are tricky, so don't screw this up?

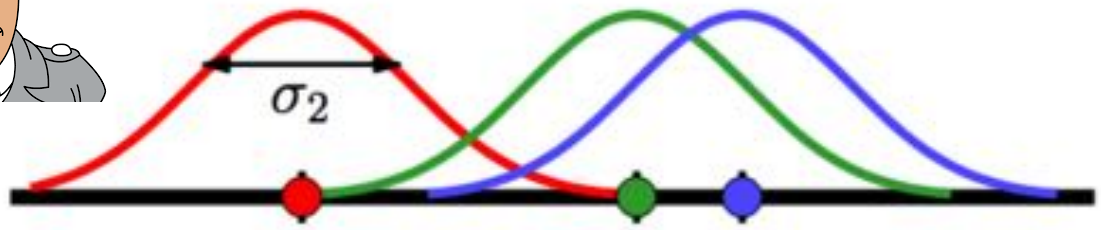


Experts don't grow on trees so don't screw this up?

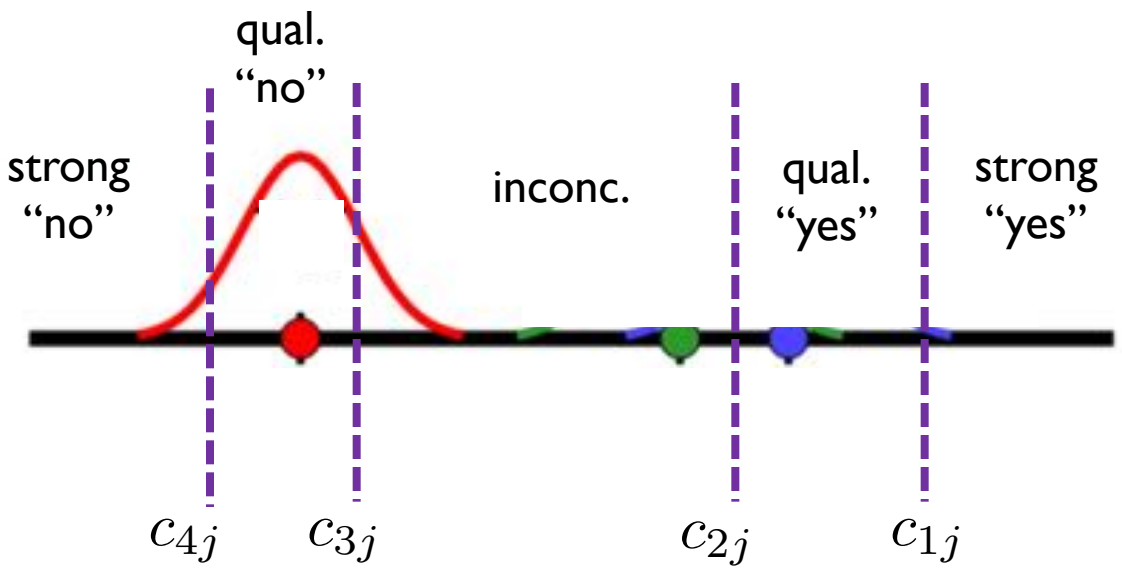


Okay, let's assume a latent "*authorness*" scale for each item

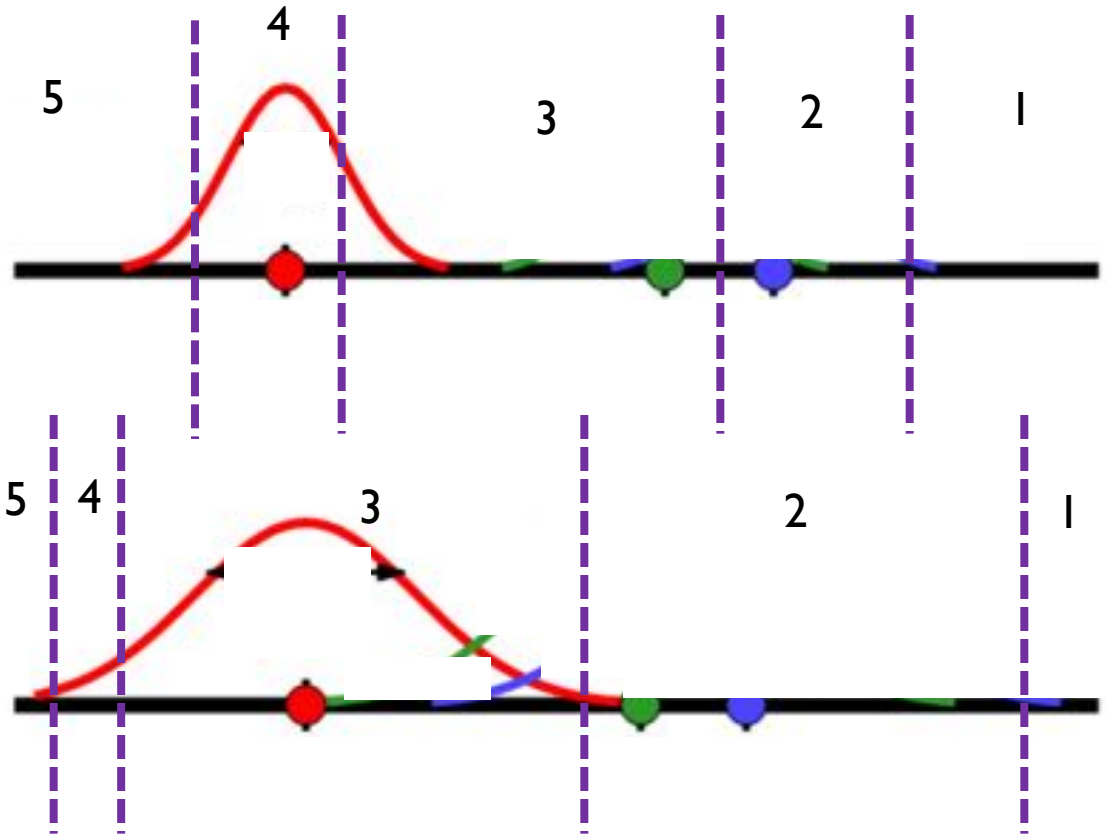




And a latent expertise for each person

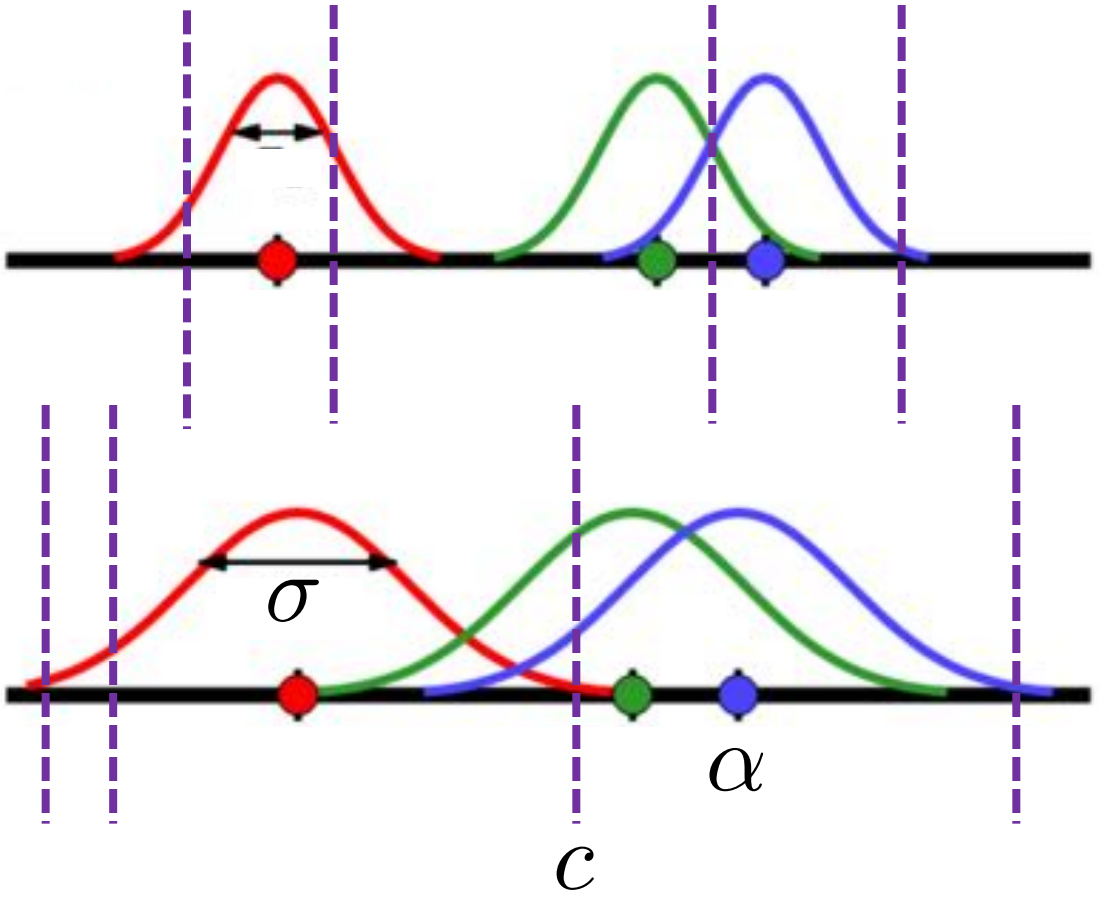


To model Likert responding, we assume each person sets decision thresholds



Each person can set
their own thresholds

The model as a whole



α

characteristics of item (unknown)

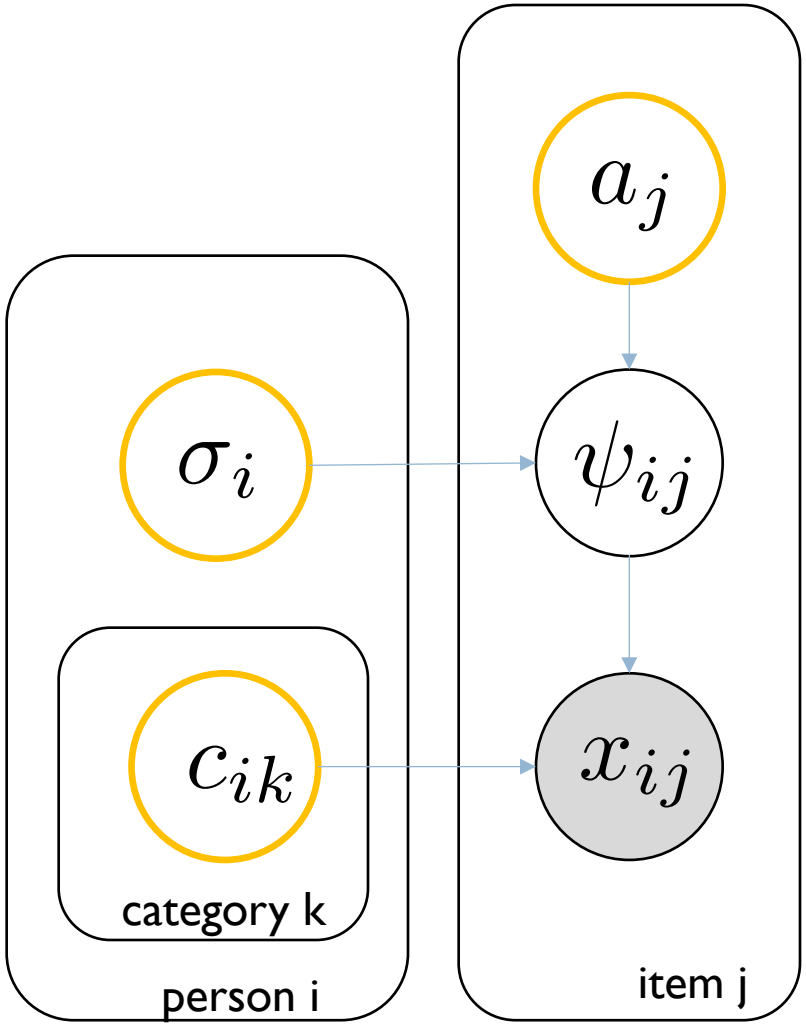
σ

expertise of decision maker (unknown)

c

response strategy adopted by the decision maker (unknown)

The model as a whole



true "authorness" of item j

expertise level of person i

subjective impression of authorness for item j felt by person i

decision threshold set by person i for response category k

response category selected by person i when shown item j

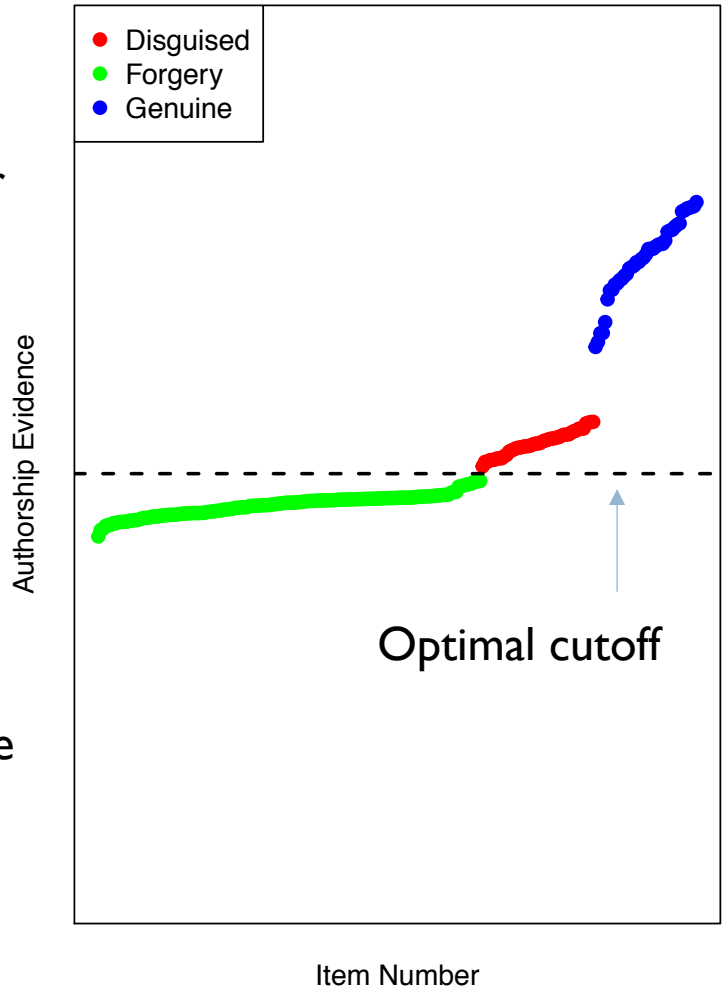
Are you going somewhere useful with this?

Oh, right…

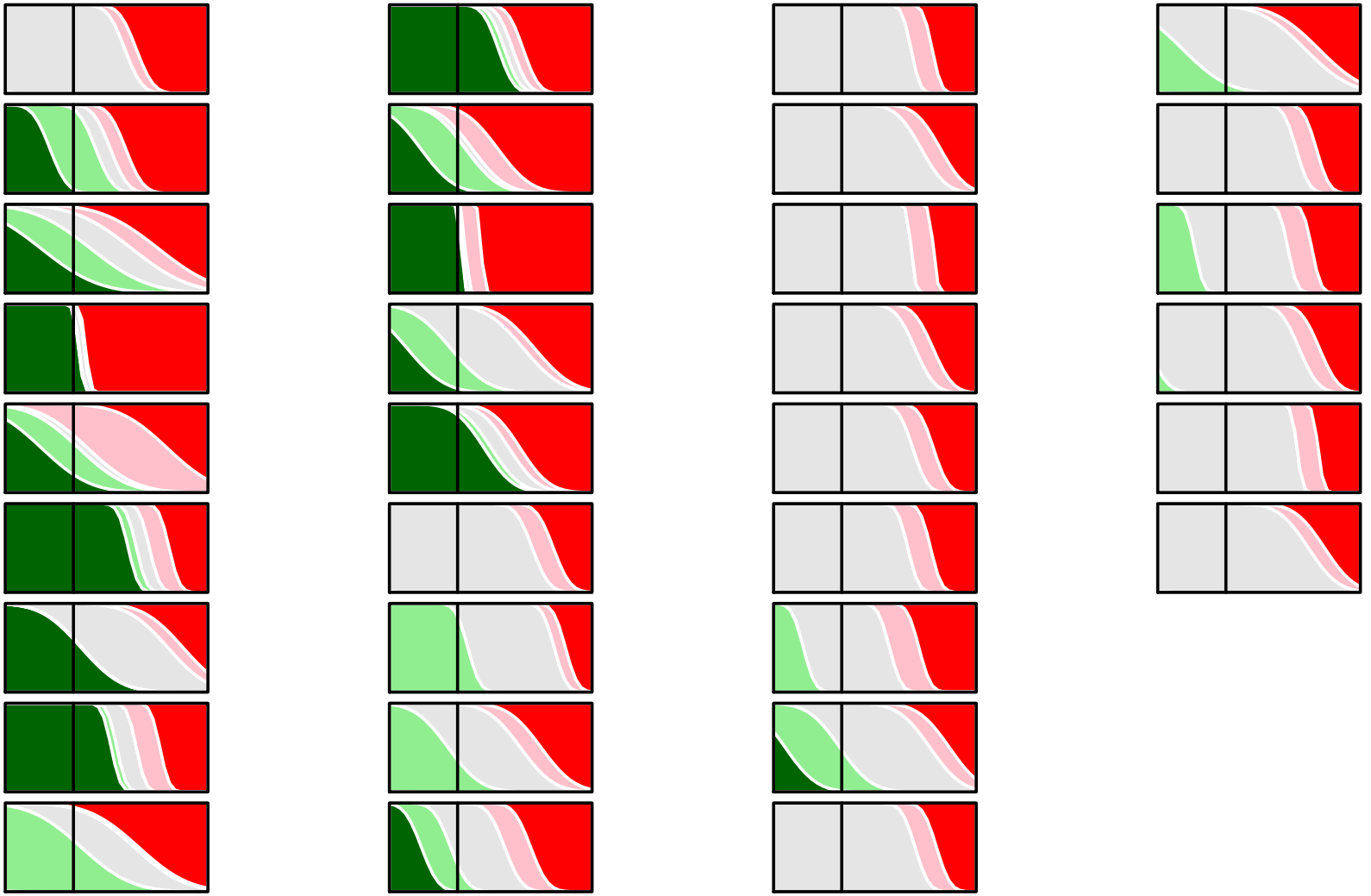
Estimates of "latent strength of evidence"

More likely to be same author

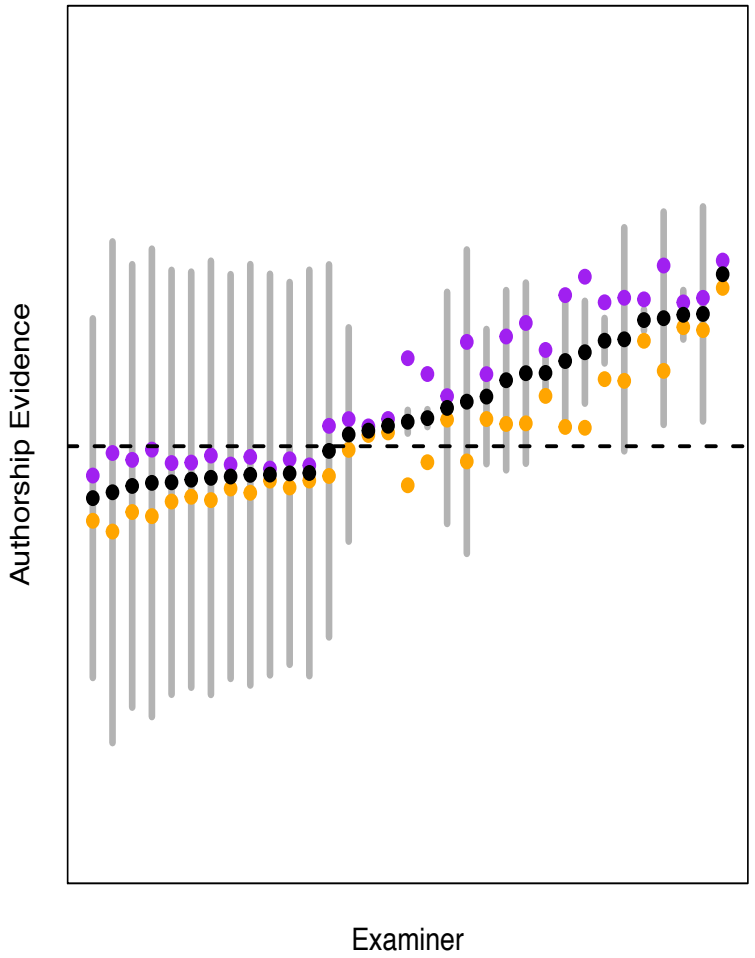
More likely to be different author



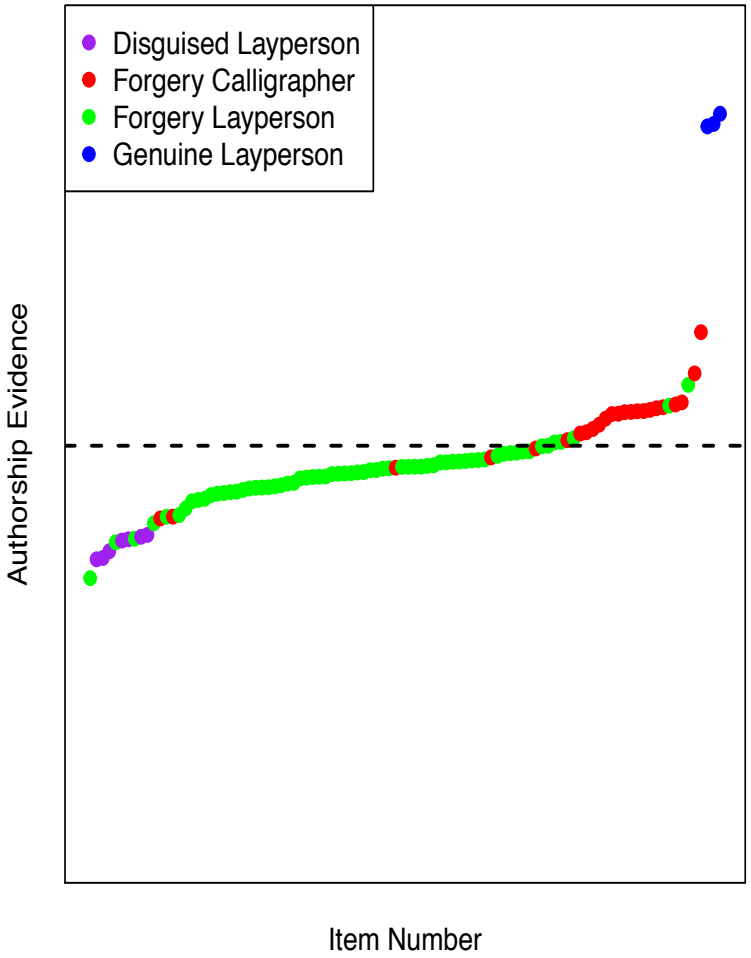
It estimates the decision strategy and expertise level for each person (training tool?)

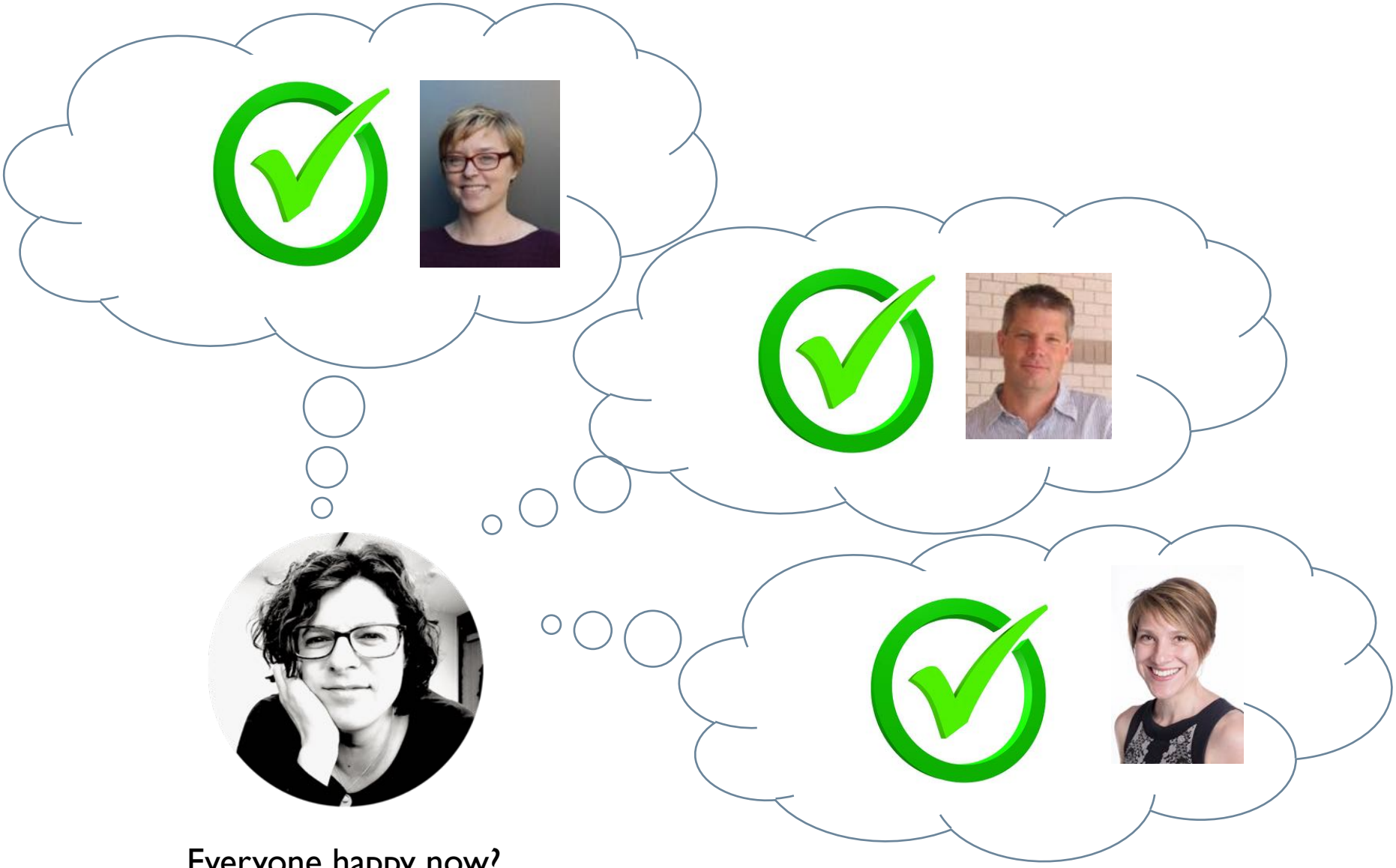


It looks at how experts adapt the response strategy when the data are "malicious" (lot of "inconclusives")



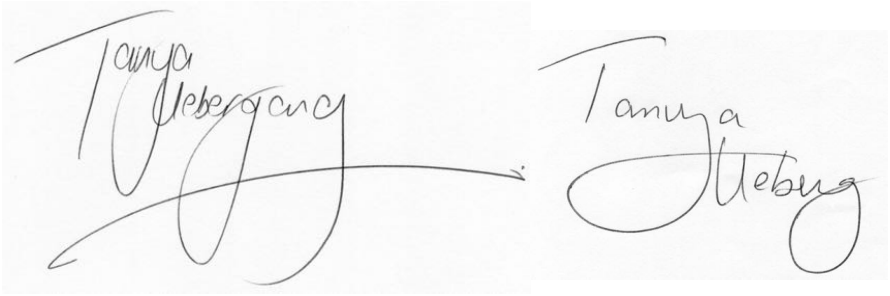
But it also reveals how "malicious" data still manages to mess with people





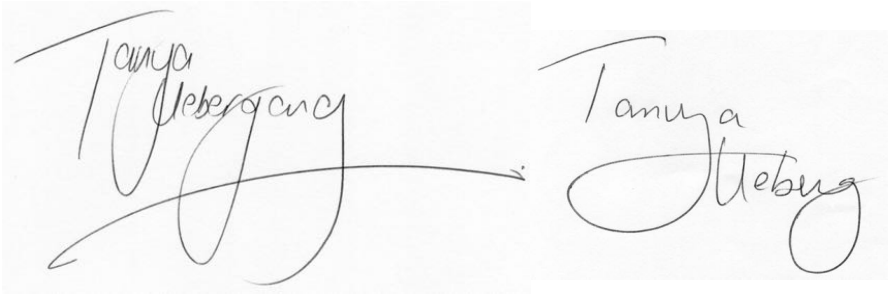
Everyone happy now?

Current directions?



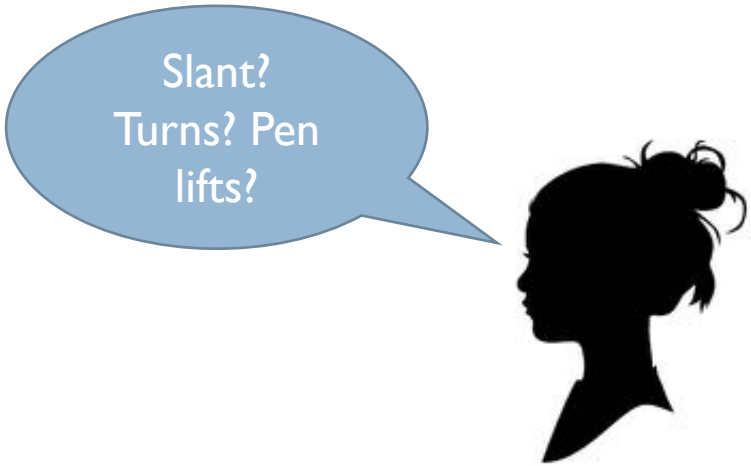
Extend the analysis to cover
wider range of data sets

Current directions?

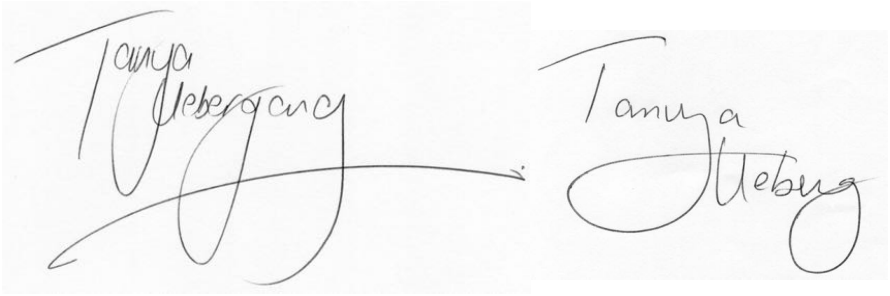


Extend the analysis to cover wider range of data sets

Add covariates: how does performance relate to features that experts verbally report relying on?

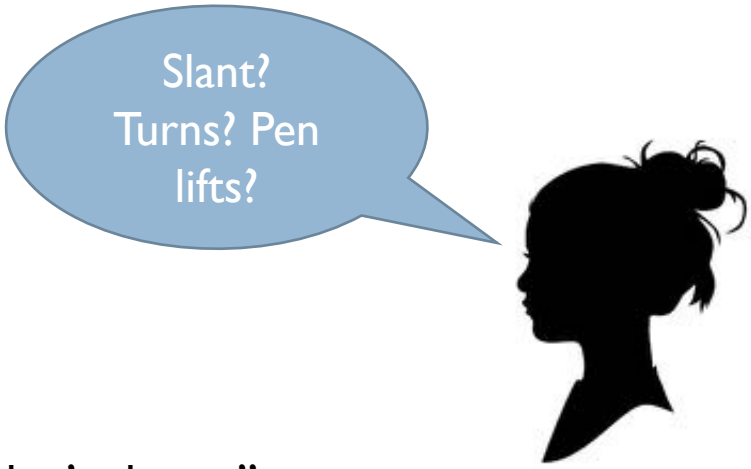


Current directions?

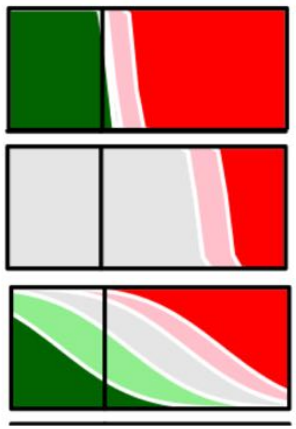


Extend the analysis to cover wider range of data sets

Add covariates: how does performance relate to features that experts verbally report relying on?



How effective are the visual "pepsi plot" representations as training tools (e.g., inducing criterion shift?)

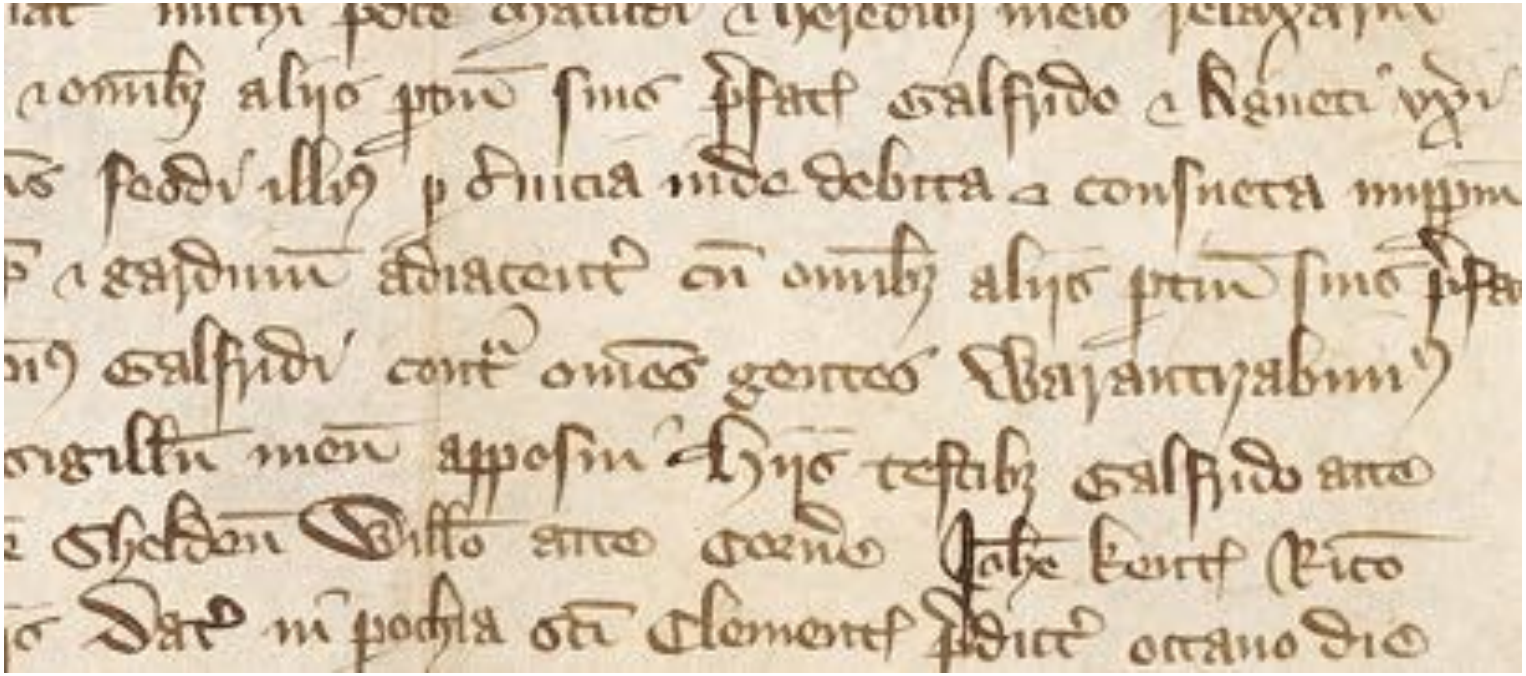


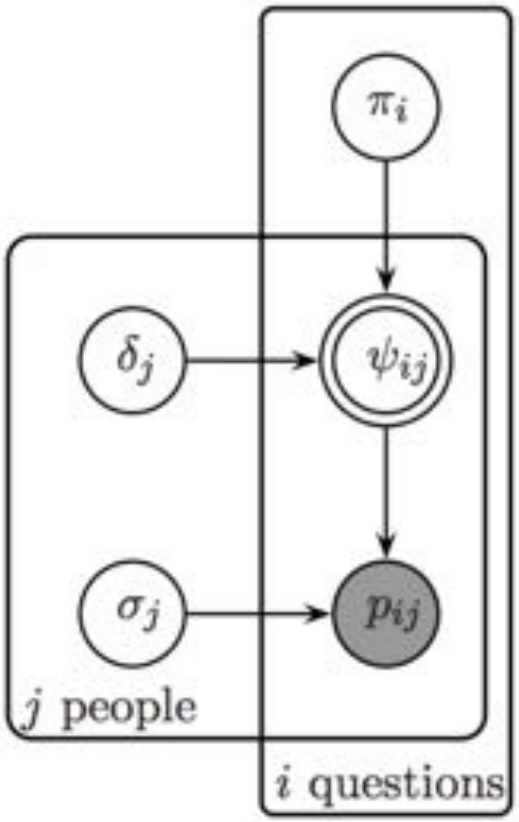
"don't change"

"call things forgeries more"

"revisit basic training?"

Thanks!



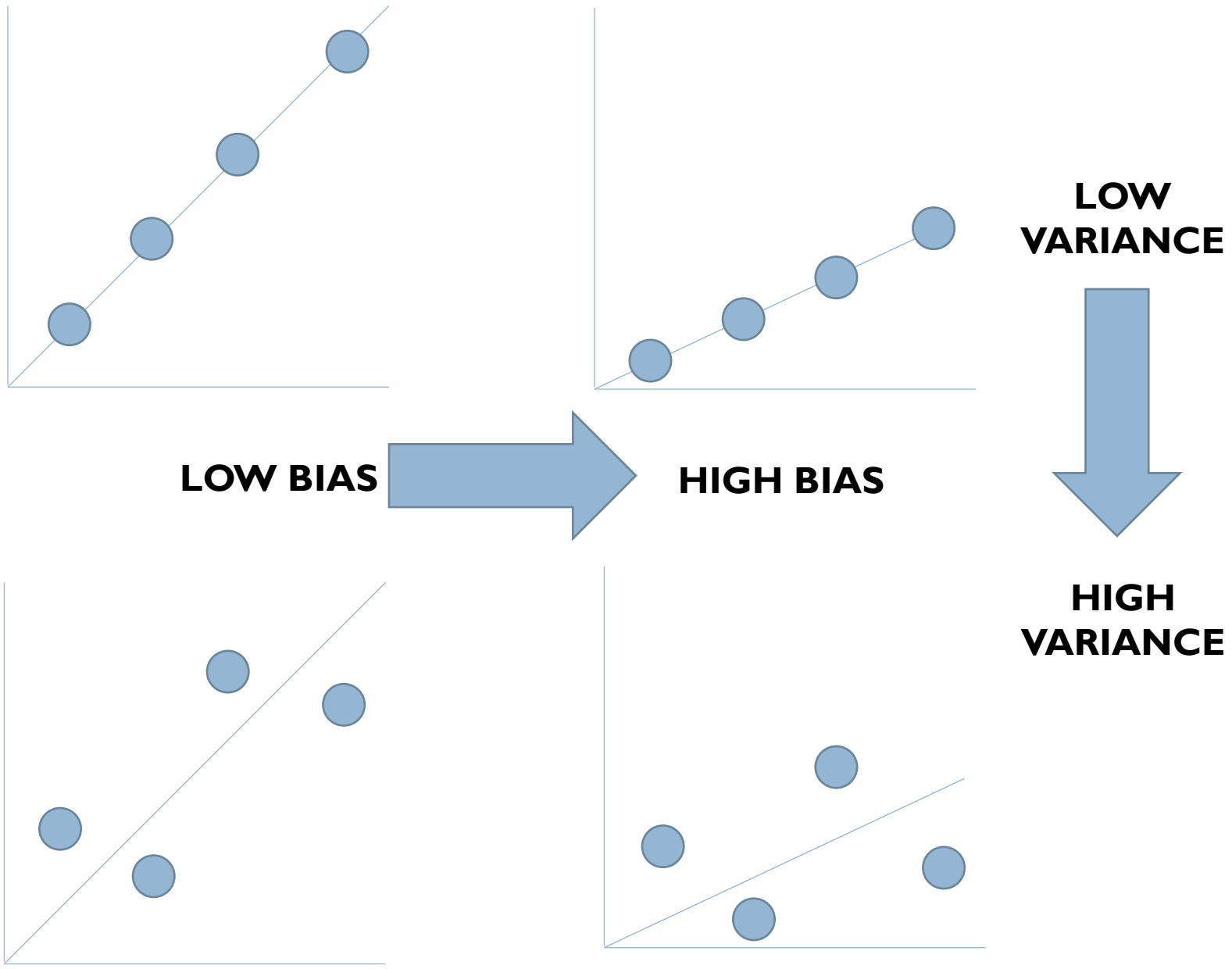


$$\delta_j \sim \text{Beta}(5, 1)$$

$$\psi_{ij} \leftarrow \delta_j \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\sigma_j \sim \text{Uniform}(0, 1)$$

$$p_{ij} \sim \text{Gaussian}\left(\frac{\exp(\psi_{ij})}{1+\exp(\psi_{ij})}, \frac{1}{\sigma_j^2}\right)$$


LOW BIAS
HIGH BIAS
LOW VARIANCE
HIGH VARIANCE