The wisdom of crowds

http://compcogscisydney.org/psyc3211/

A/Prof Danielle Navarro
d.navarro@unsw.edu.au
compcogscisydney.org

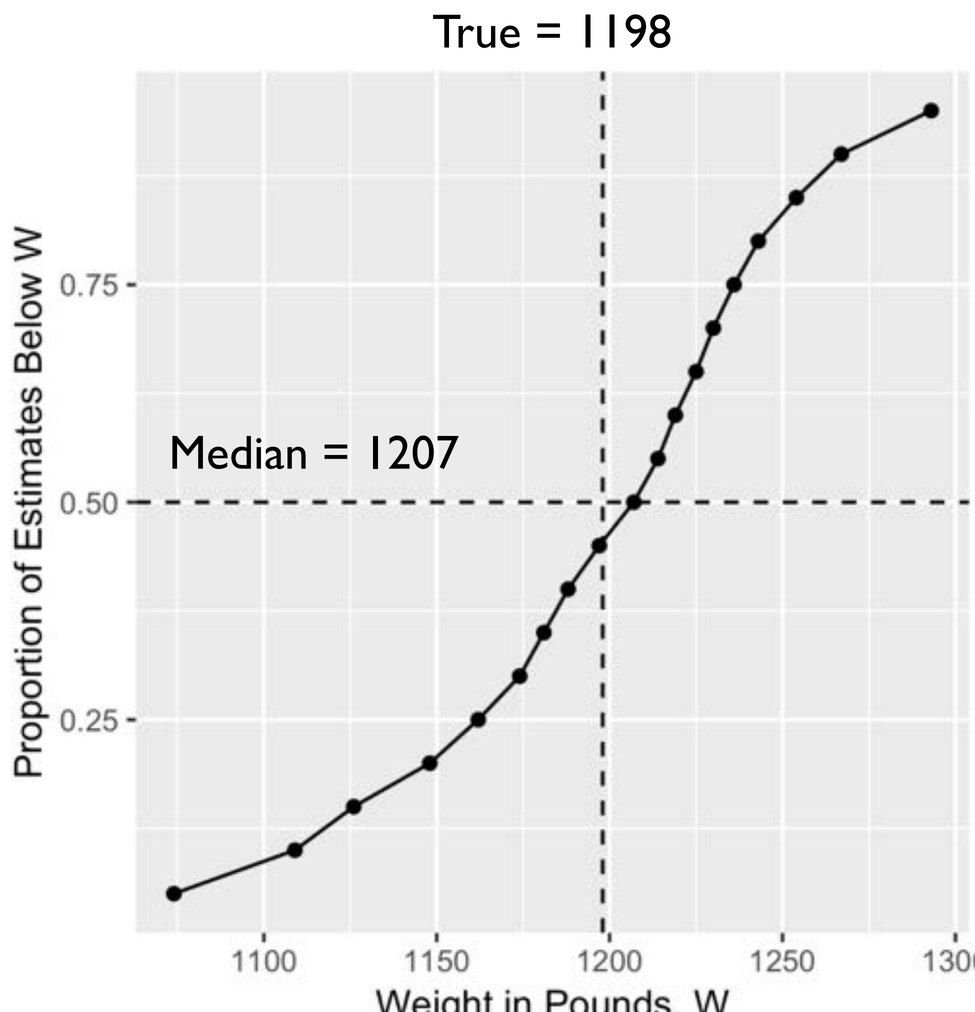
Where are we?

- L1: Connectionism
- L2: Statistical learning
- L3: Semantic networks
- L4: Wisdom of crowds
- L5: Cultural transmission
- L6: Summary

Structure of the lecture

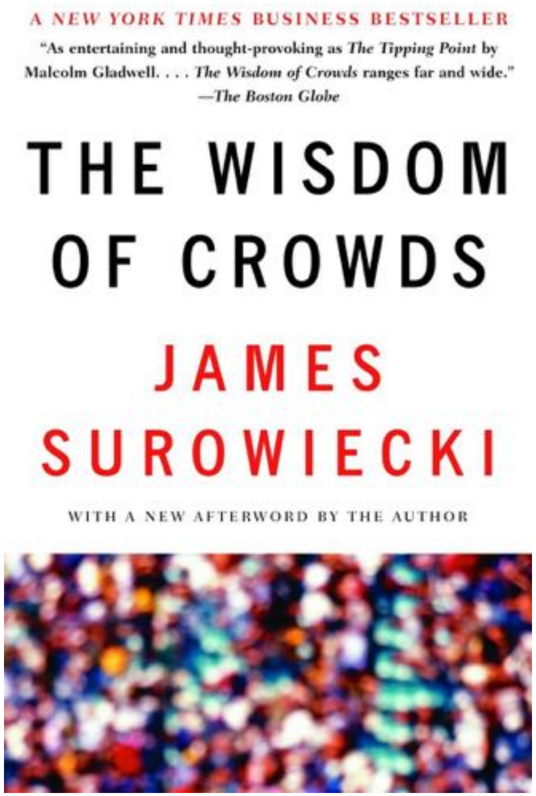
- The core idea (Galton, Surowiecki)
- Ranking tasks
- Categorisation tasks
- Combinatorial optimization tasks
- Application to forensic science

Vox populi
(Galton 1907)



The wisdom of crowds

(Surowiecki 2004)



Criterion	Description
Diversity	Each person should have their own personal knowledge to rely on
Independence	Each person should form this opinion without any information about the opinions of others
Decentralization	Each person should be able to draw on different sources to form their opinion
Aggregation	There should be a sensible mechanism for combining the different opinions

Ranking tasks

https://www.educationaltoysplanet.com/presidents-write-on-learning-placemat.html

Ranking tasks

(Steyvers et al 2009; Lee et al 2012)

Use a drag-and-drop interface to sort US presidents into chronological order

George Washington
John Adams
Thomas Jefferson
James Monroe
Andrew Jackson
Theodore Roosevelt
Harry Truman
Dwight Eisenhower

Variety of problems: books, city population, country landmass, country population, hardness, holidays, movies, US presidents, rivers, US states locations, superbowl, US constitution ten amendments, Bible ten commandments

Ranking tasks

(Steyvers et al 2009; Lee et al 2012)

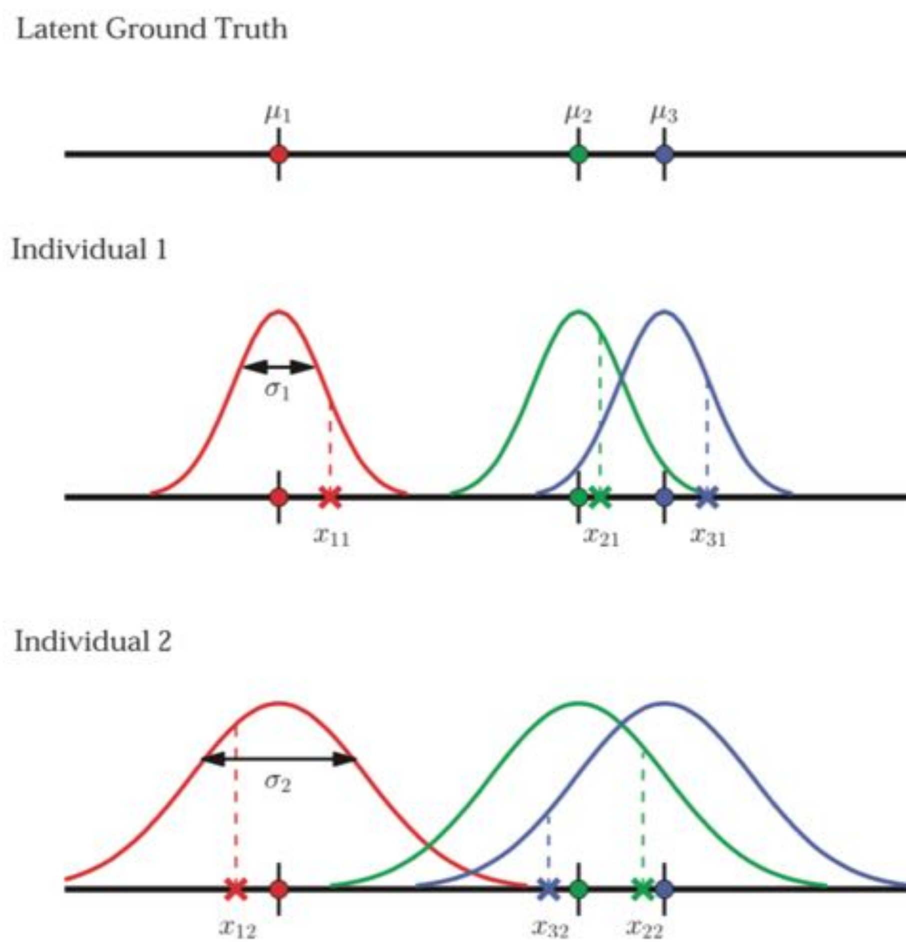
We assume the existence of a "*__latent__* ground truth" … needs to be estimated statistically from responses

These items are close together and are easy to mix up (e.g., Monroe & Jackson)

This item is very distant from the others, so it's easy to get correct (e.g., George Washington as 1st US president)

Ranking tasks

(Steyvers et al 2009; Lee et al 2012)



Some people have good knowledge of this (low noise)

Other people have poor knowledge of this (high noise)

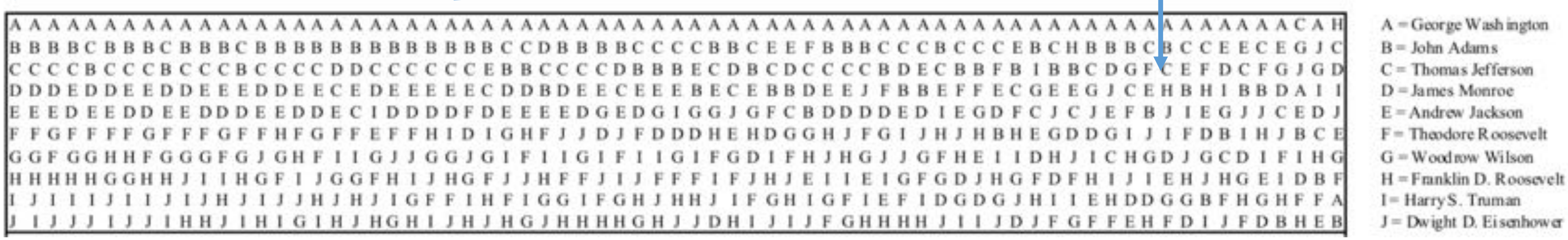
Ranking tasks

(Steyvers et al 2009; Lee et al 2012)

One person gets the exact ordering!

Almost everyone gets Washington correct

Few people know where to place Monroe



- Data from all 78 subjects on the US presidents question
- Variation in expertise of individuals
- Variation in difficulty of items

Ranking tasks

(Steyvers et al 2009; Lee et al 2012)

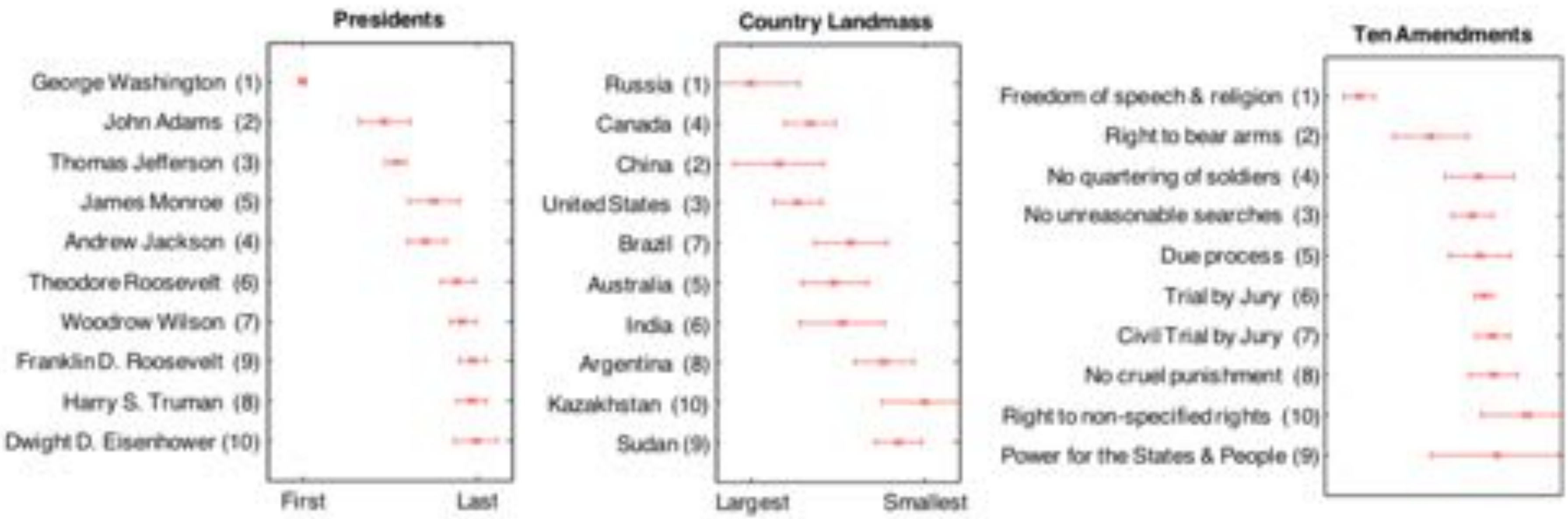
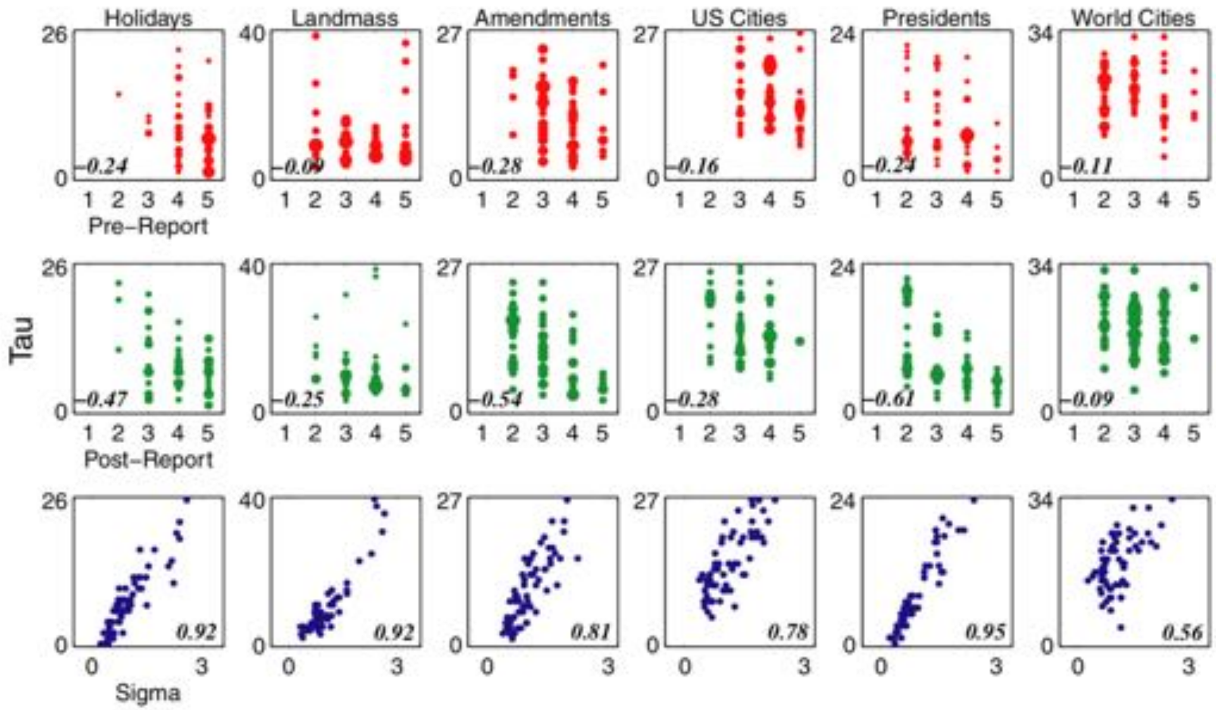


Figure 3. Sample Thurstonian inferred distributions. The vertical order is the ground truth ordering, while the numbers in parentheses show the inferred group ordering

Ranking tasks

(Steyvers et al 2009; Lee et al 2012)

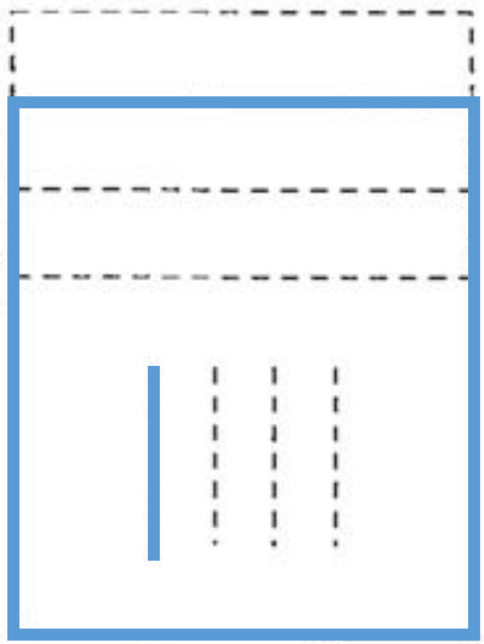


- Tau: Agreement with the true ordering
- Sigma: Expertise (noise) estimated by the model
- Report: Judgment of own knowledge before (pre) and after (post) doing the task

A category learning example

Category learning

(Kruschke 1993)



Stimuli are rectangles with different ***heights***

With an internal line with different horizontal ***positions***

Figure 5. Stimuli. Solid lines show one combination of rectangle height and lateral position of interior segment. Dotted lines show alternative heights and positions.

Category learning
(Kruschke 1993)

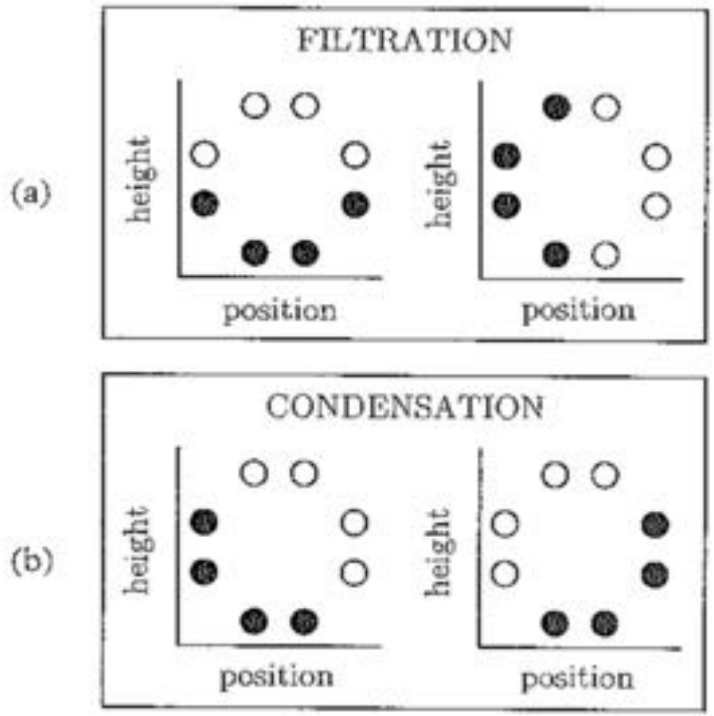


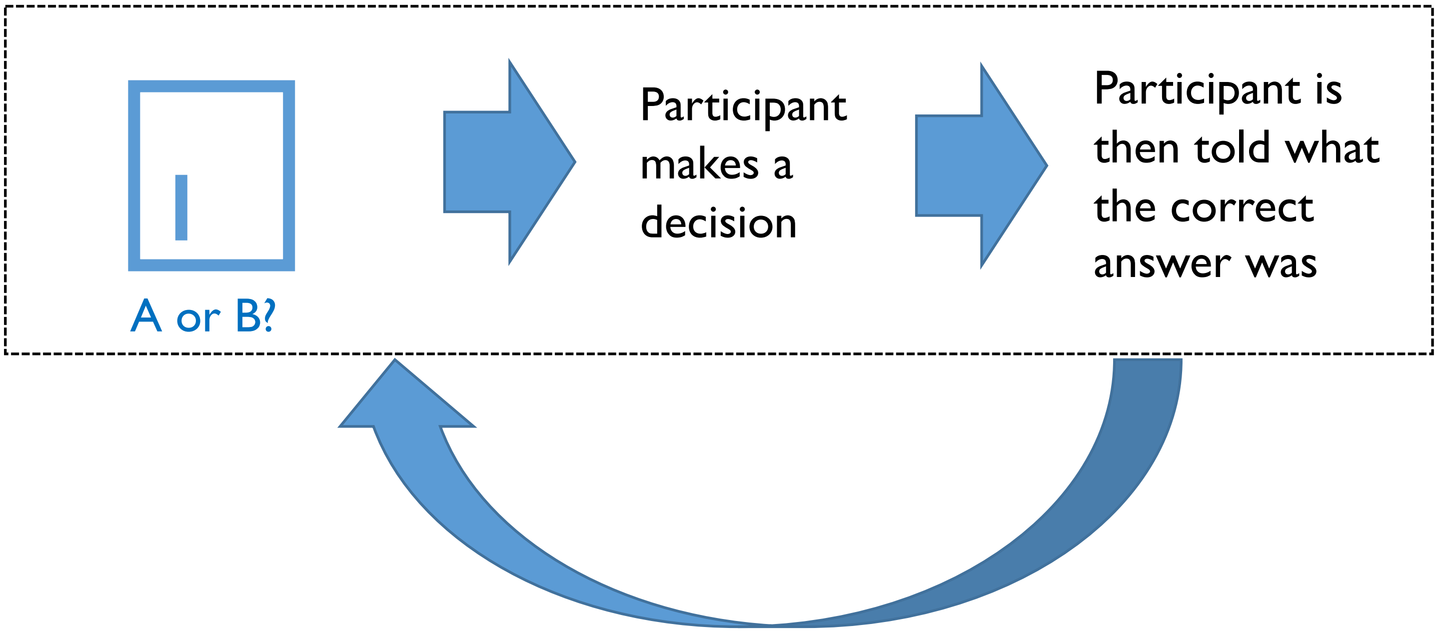
Figure 6. Structure of the filtration and condensation categories. Open circles denote one category, filled circles the other.

Black circles are "category A"
White circles are "category B"

Category learning

(Kruschke 1993)

Structure of a trial…



- Repeat for 8 trial blocks/epochs
- Each block/epoch presents each of the 8 items once

Category learning

(Kruschke 1993)

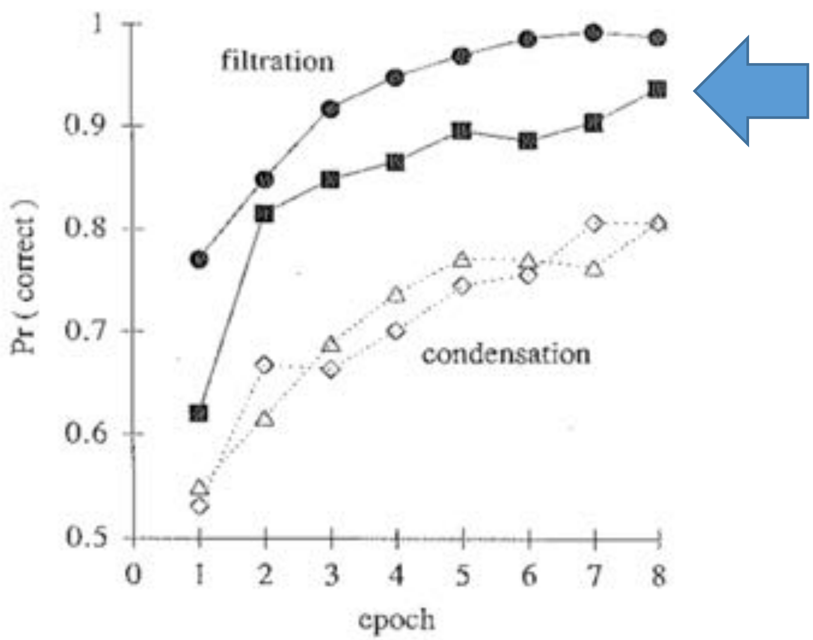


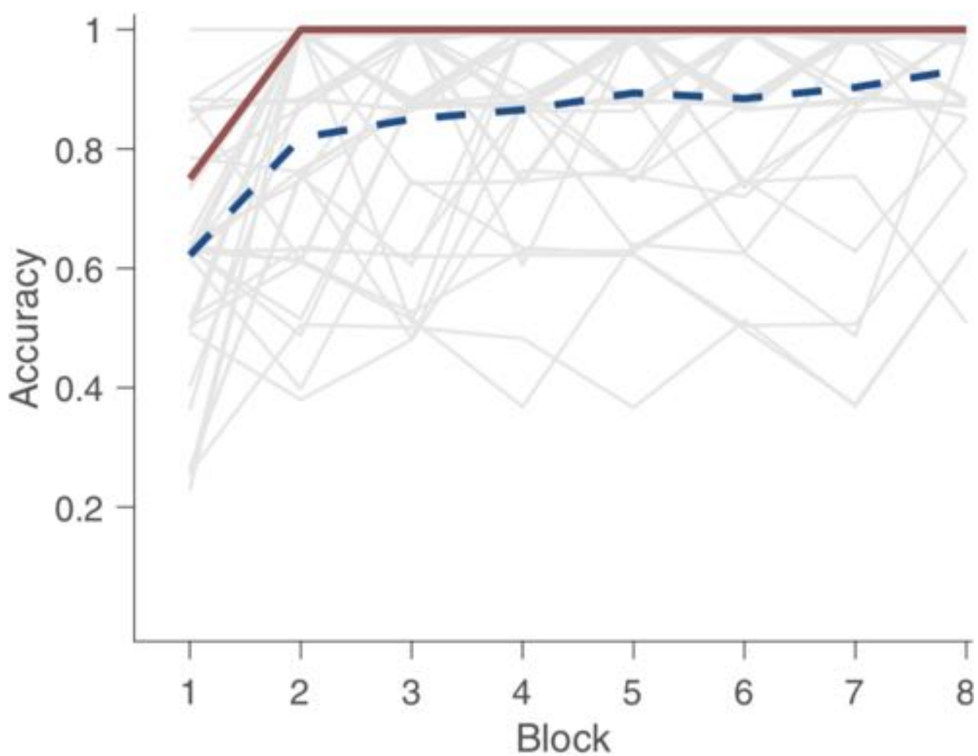
Figure 7. Human learning data for the four category structures shown in Figure 6. One 'epoch' is one sweep through the eight different stimuli. Filled circles show the position-relevant filtration; filled squares show the height-relevant filtration. Open markers show results from the two condensation conditions.

There's a separate literature focusing on why these conditions are different from each other but let's pick one condition and look at the individual differences…

Category learning

(Danileiko & Lee 2017)

The red line is what would happen if you always chose with the majority on every trial

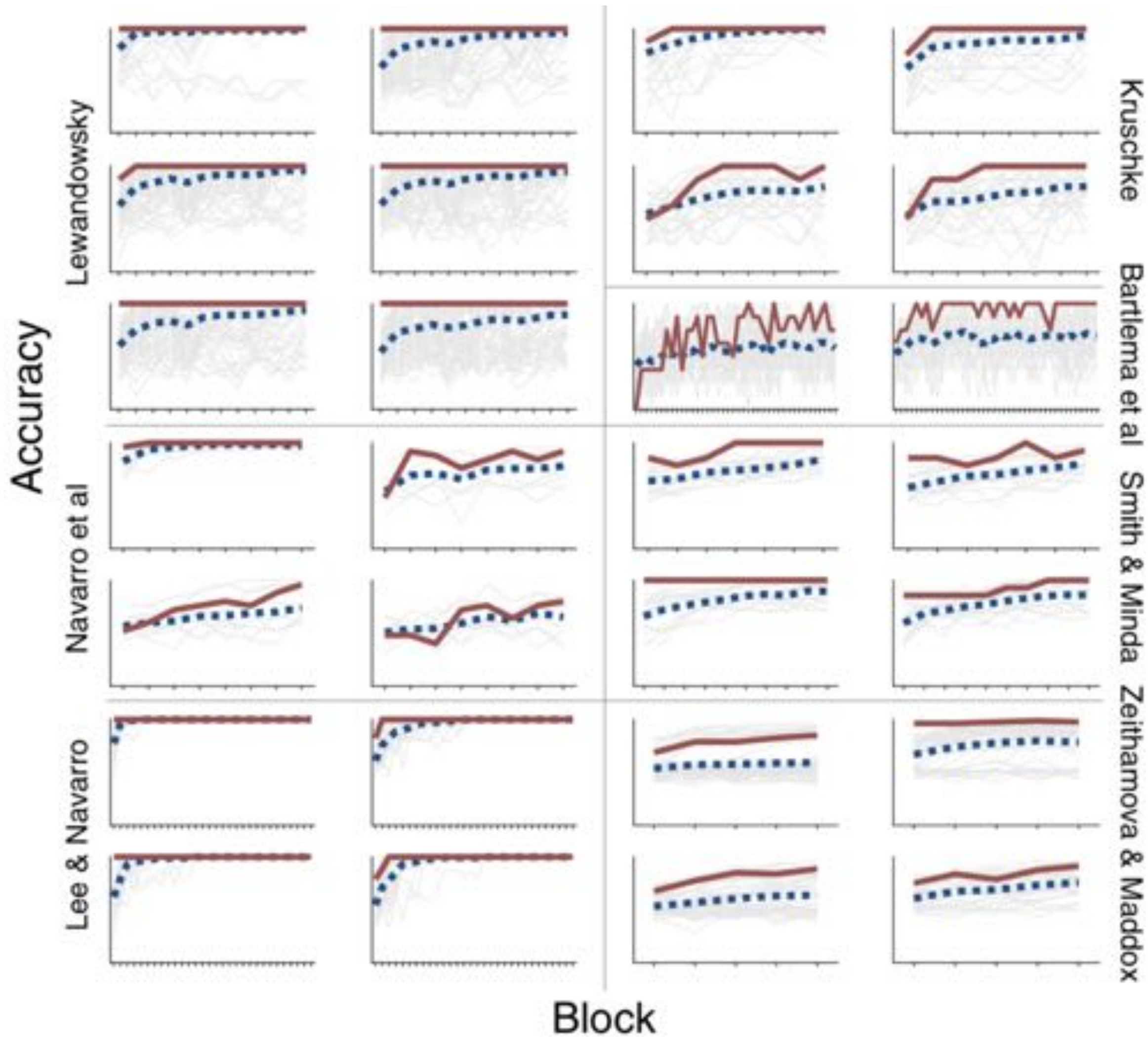


The blue line is the average accuracy of across people

Each grey line is the classification accuracy of one person over time

Category learning

(Danileiko & Lee 2017)



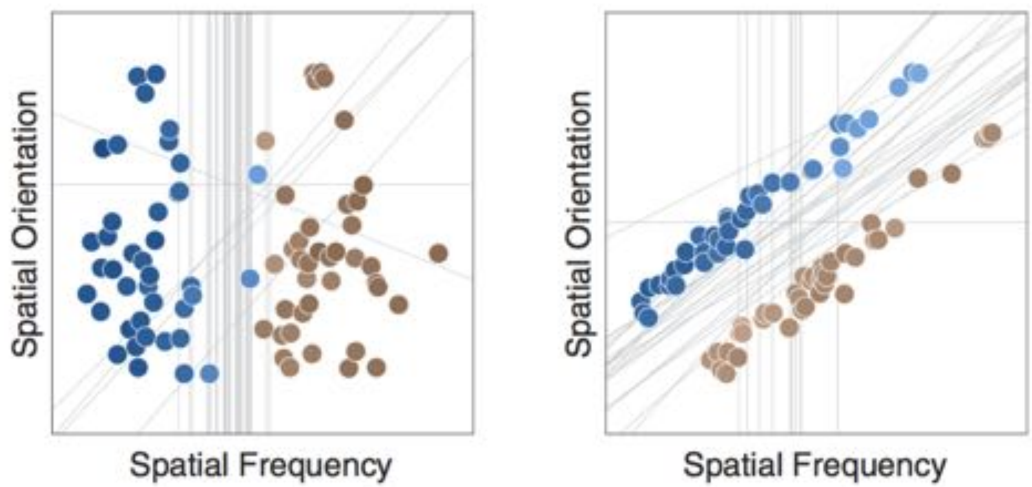
Here's the same thing for 28 data sets

Mostly good, but there are some cases where it fails

Category learning

(Danileiko & Lee 2017)

- Problem? How do we generalize from crowd knowledge????
- Solution: Instead of aggregating at the level of each *response*, estimate the categorization *rule* each person was applying, and average* those



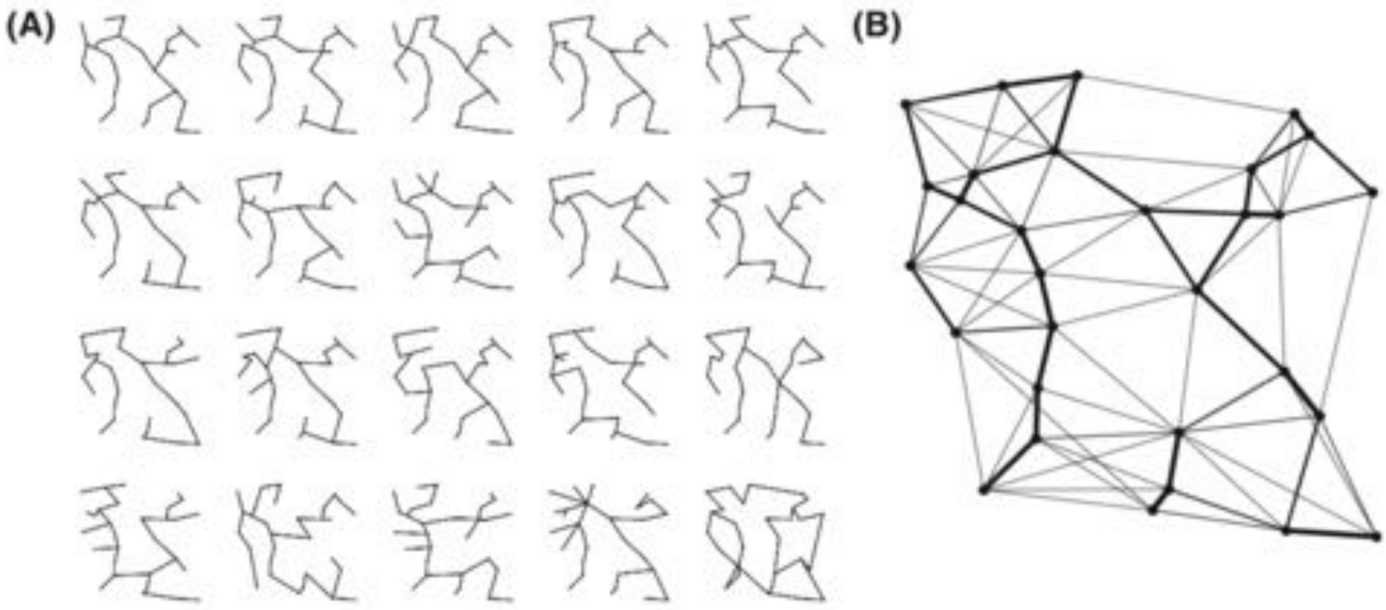
*sort of

Complications on the wisdom of crowds phenomenon?

Minimum spanning trees
(Yi et al 2012)

Individual solutions to the minimum spanning tree problem

The solution that maximises overall agreement* with individual choices is closer to optimal than any person's solution



*details omitted

Travelling salesperson problem

(Yi et al 2012)

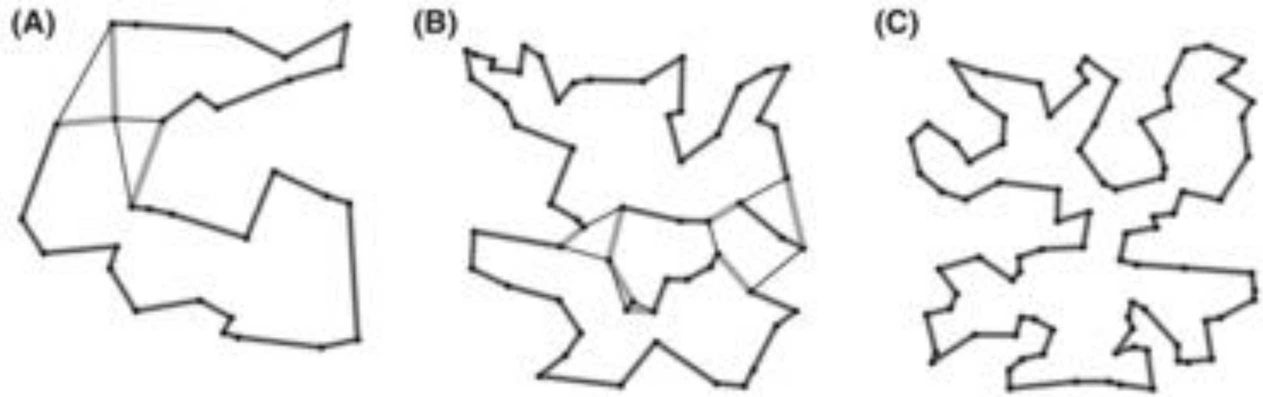


Fig. 6. Solution paths for the best-performing aggregate method parameters (thin black) and the optimal TSP (thick gray) for the (A) 30-node, (B) 60-node, and (C) 90-node problems.

Same thing for the TSP!

"The Price is Right"

(Lee et al 2010)

Fig. 1 A Price Is Right competition, with four players placing bids to win a stereo. Player 2 is the winner, because their bid of $675 is the closest to the true price of $960, without having exceeding the true price

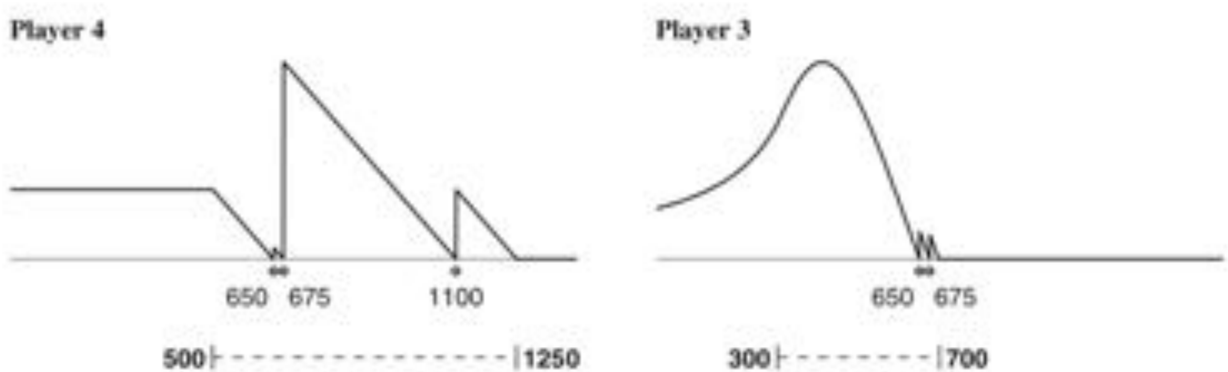
“The Price is Right”

(Lee et al 2010)

This is tricky because participants have a motivation not to give their best guess

A strategic bid from Player 3 depends on what Players 1 an 2 did, AND what they think Player 4 will do… so this is messy, socially-rich competitive environment!

Fig. 2 Rational decision-making for the fourth (left panel) and third (right panel) players, given the existing bids, shown by circles, and bounded price knowledge, shown by the dashed-line interval



“The Price is Right”

(Lee et al 2010)

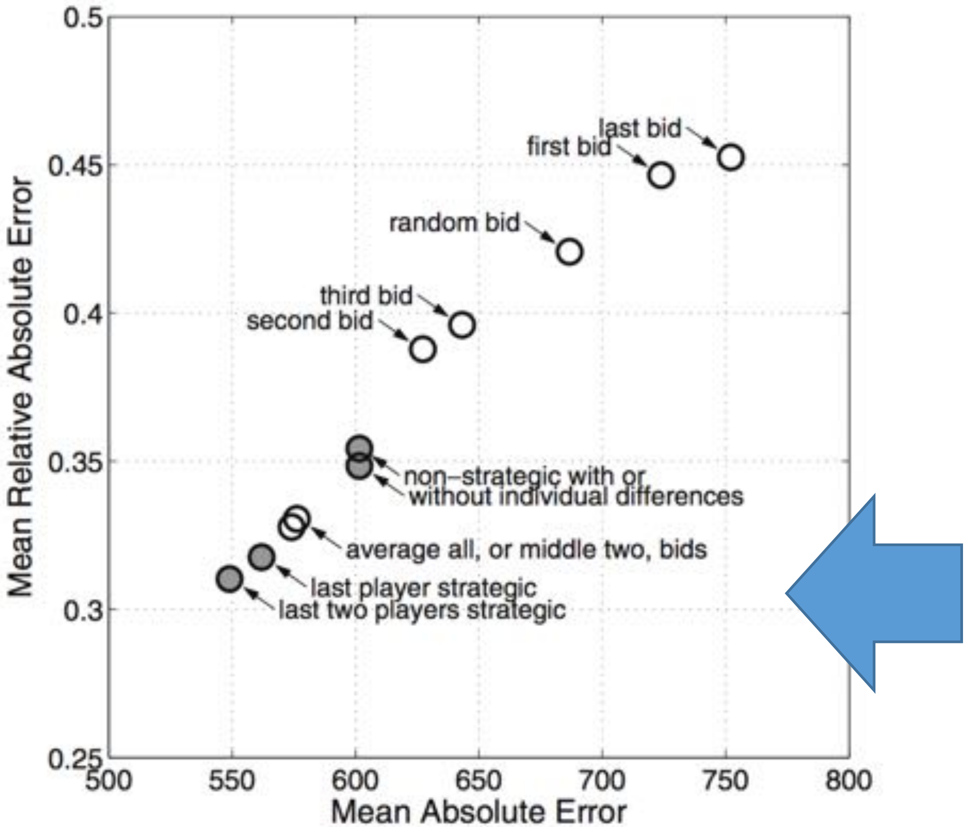


Fig. 4 Pverformance of the simple (unshaded circles) and model-based (shaded circles) methods for aggregate estimation, in terms of both mean absolute error and mean relative absolute error

Again we see wisdom of crowds, but the effect is strongest when aggregation is done using a cognitive model that assumes the last two players are betting strategically!

A forensic science application

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

A variety of questions

- The *process* problem: were these written in the same *way*? (e.g., disguising one's handwriting)
- The *authorship* problem: were these written by same *person*?
- The *feature match* problem: what are the relevant features, and do the samples match?

Specific case: how good are people at evaluating whether a feature match is informative? Do we know which features in handwriting are common and which are rare?

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.

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

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

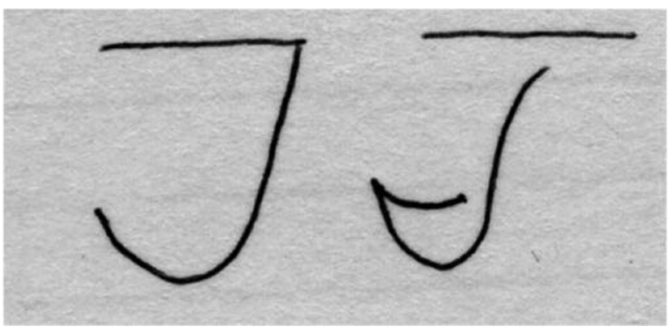
Funding information

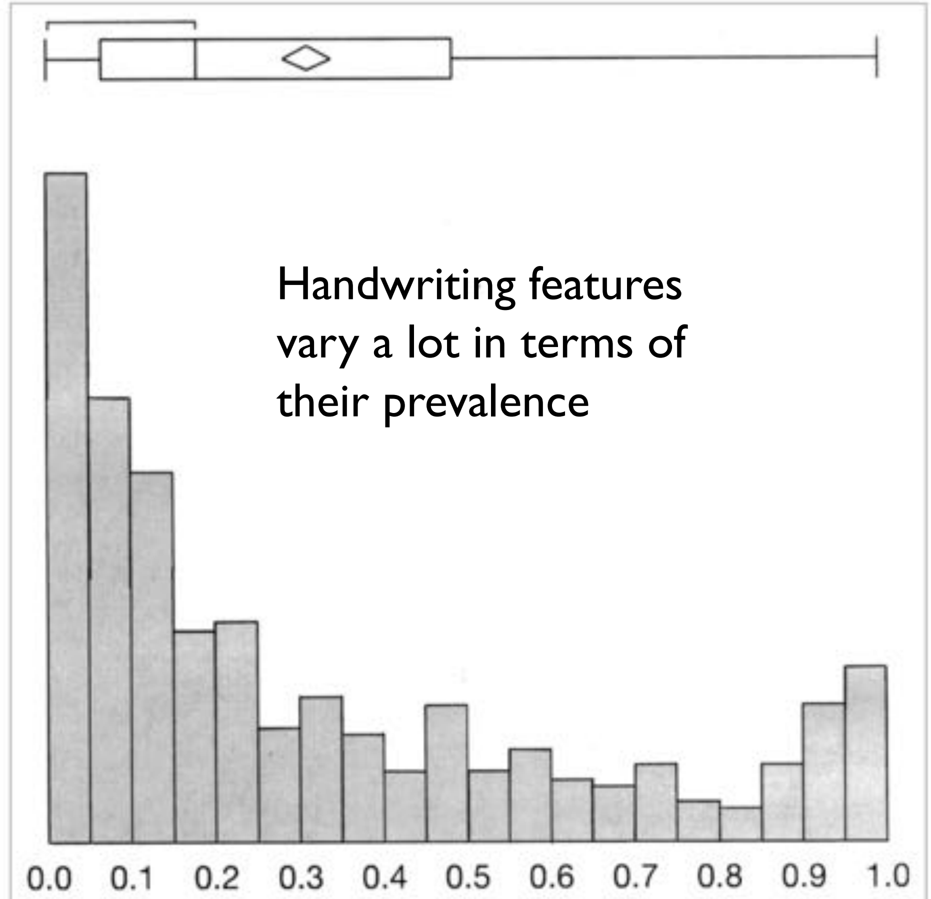
[†]Presented at the 67th Annual Meeting of the American Academy of Forensic Sciences, February 16-21, Orlando, FL.

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

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.





Handwriting features vary a lot in terms of their prevalence

Figure 1. Open in figure viewer | Download Powerpoint slide

Histogram of features present in the cursive project sample.

Johnson et al (2016)

What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts

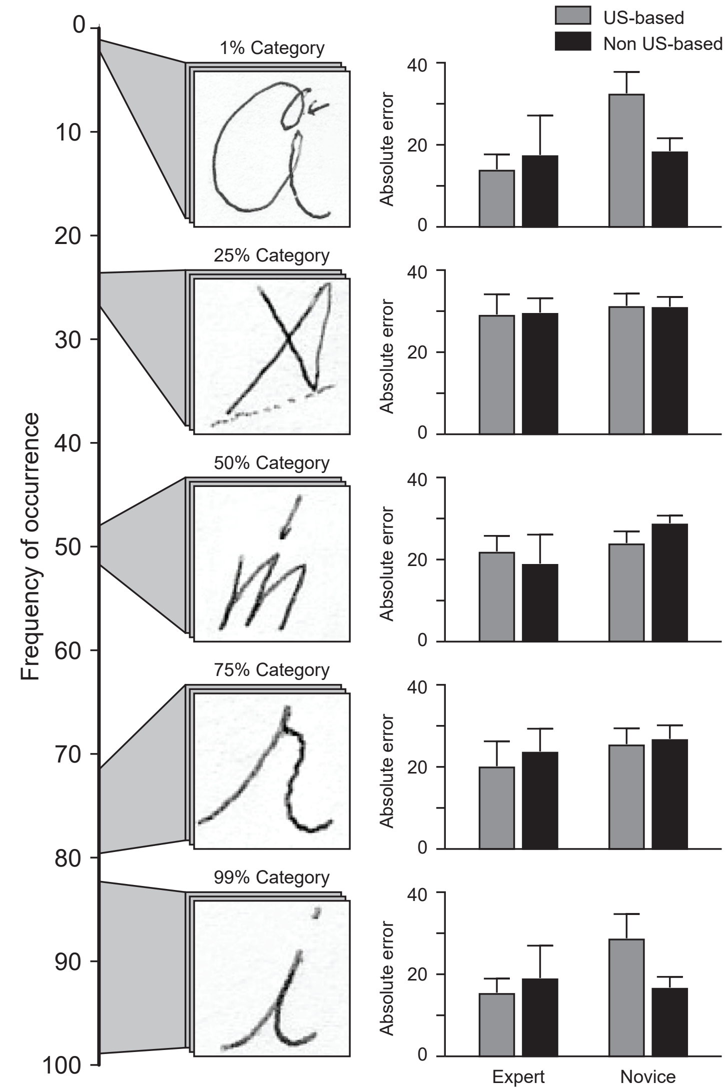
Kristy A. Martire
School of Psychology
University of New South Wales

Bethany Growns
School of Psychology
University of New South Wales

Danielle J. Navarro
School of Psychology
University of New South Wales

Abstract

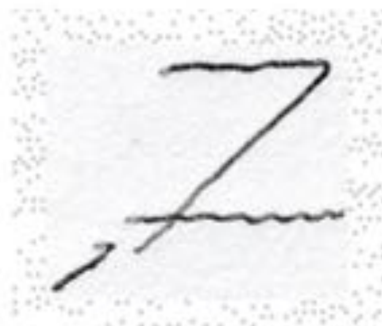
Forensic handwriting examiners currently testify to the origin of questioned handwriting for legal purposes. However, forensic scientists are increasingly being encouraged to assign probabilities to their observations in the form of a likelihood ratio. This study is the first to examine whether handwriting experts are able to estimate the frequency of US handwriting features more accurately than novices. The results indicate that the absolute error for experts was lower than novices, but the size of the effect is modest, and the overall error rate even for experts is large enough as to raise questions about whether their estimates can be sufficiently trustworthy for presentation in courts. When errors are separated into effects caused by miscalibration and those caused by imprecision, we find systematic differences between individuals. Finally, we consider several ways of aggregating predictions from multiple experts, suggesting that quite substantial improvements in expert predictions are possible when a suitable aggregation method is used.



Looked at people's accuracy for a variety of handwriting features

Forensic document examination

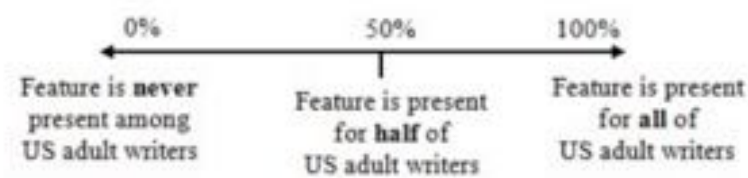
(Martire et al, in press)



Printed lower case ‘z’ is two strokes

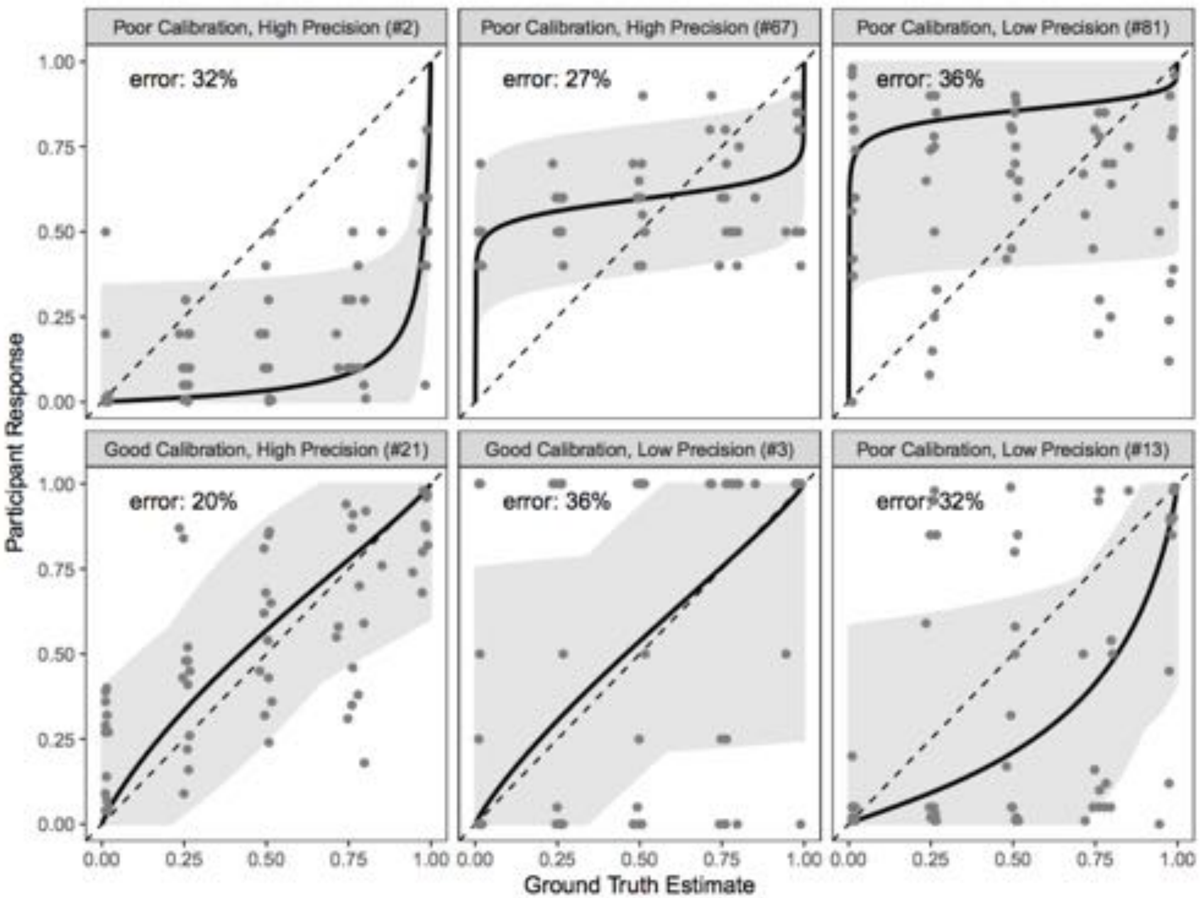
What percentage of the US adult population of adult writers have this feature in their handwriting?

Please type a number between 0 and 100 in the box below.



Forensic document examination

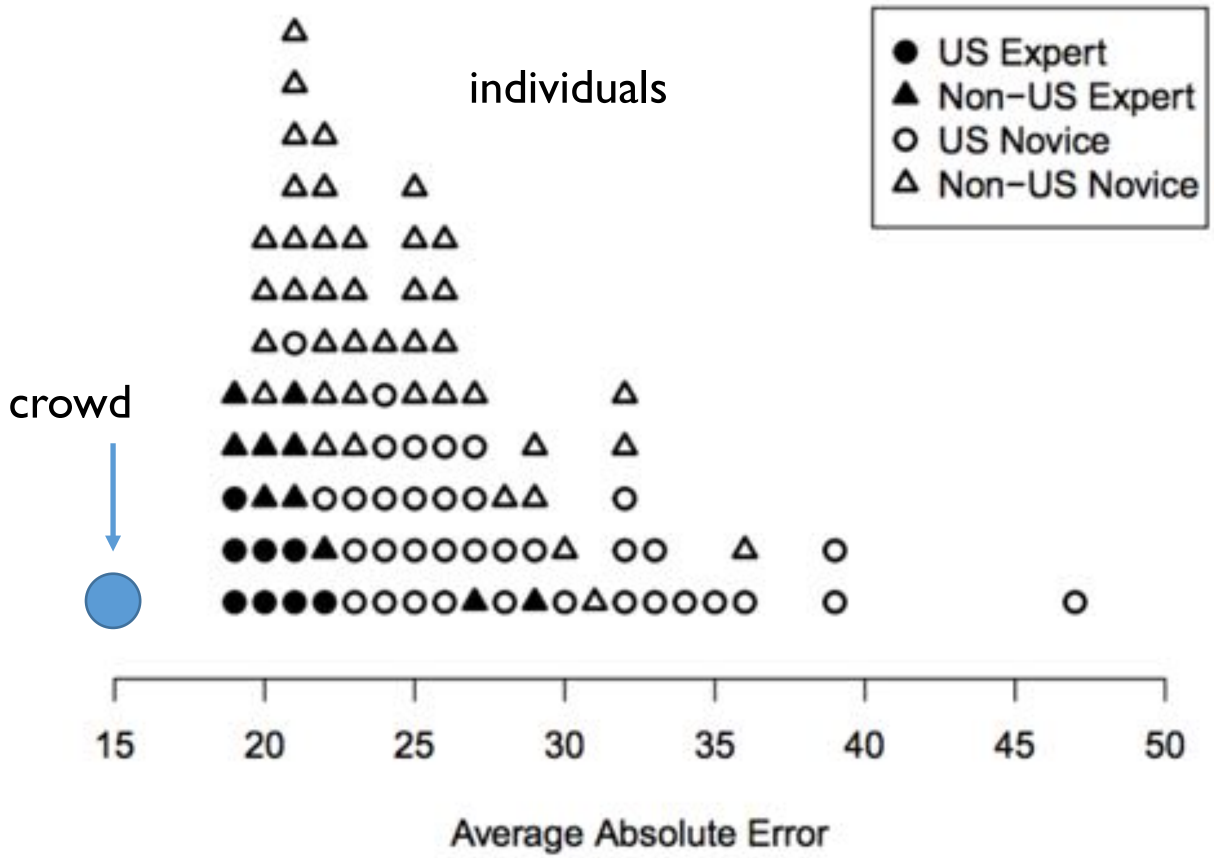
(Martire et al, in press)



Individual (both novices and experts) do know something about this, but judgments are noisy and there's a lot of variability in how much people know

Forensic document examination

(Martire et al, in press)



Forensic document examination

(Martire et al, in press)

Some aggregation methods work better than others…

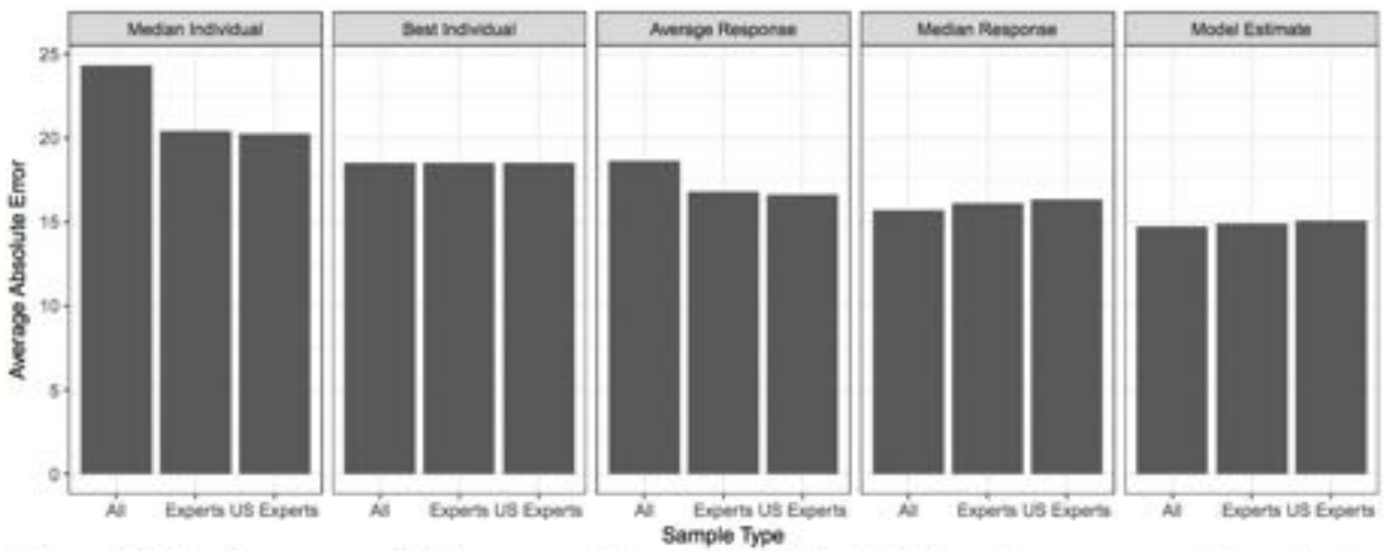


Figure 6. Performance of the aggregation model (far right), when compared to simpler aggregation approaches that averages responses (middle) or takes the median response (near right). For comparison purposes, both are plotted against the performance of the best individual participant (near left), and the median performance of all respondents (far left). Within each panel, three versions are plotted: one where we included all 94 participants, one where we used responses from the 17 experts, and a third where we used only the 8 US-based experts.

Thanks