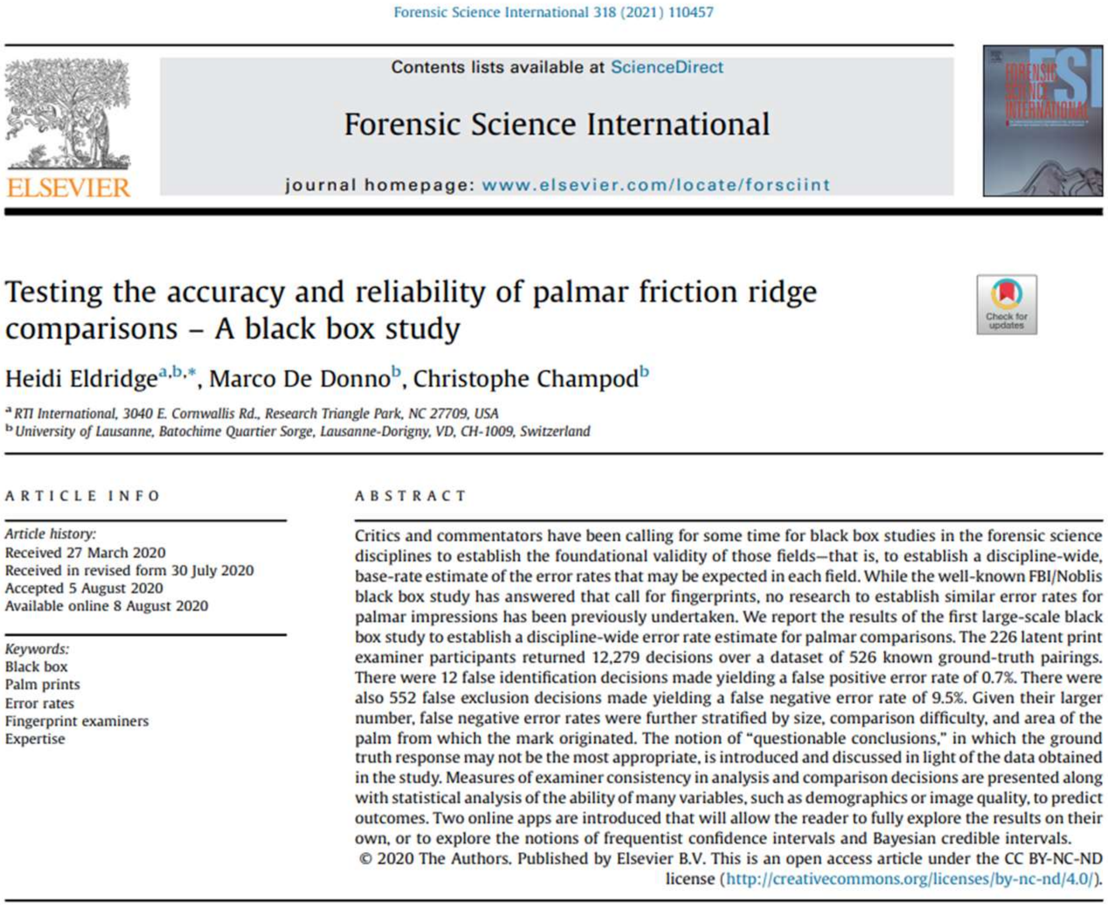
# Palmprint Ordered Probit Model

Meredith Coon, MFS, CLPE
Thomas A Busey, PhD, Indiana University, Bloomington

The views expressed in this presentation are my own and based on my training, analyses, and experiences. Any opinions expressed are not reflective of any past or current employer

# Data Set



Forensic Science International 318 (2021) 110457

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

## Testing the accuracy and reliability of palmar friction ridge comparisons – A black box study

Heidi Eldridge[a,b,*], Marco De Donno[b], Christophe Champod[b]

[a] RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC 27709, USA
[b] University of Lausanne, Batochime Quartier Sorge, Lausanne-Dorigny, VD, CH-1009, Switzerland

ARTICLE INFO

*Article history:*
Received 27 March 2020
Received in revised form 30 July 2020
Accepted 5 August 2020
Available online 8 August 2020

*Keywords:*
Black box
Palm prints
Error rates
Fingerprint examiners
Expertise

ABSTRACT

Critics and commentators have been calling for some time for black box studies in the forensic science disciplines to establish the foundational validity of those fields—that is, to establish a discipline-wide, base-rate estimate of the error rates that may be expected in each field. While the well-known FBI/Noblis black box study has answered that call for fingerprints, no research to establish similar error rates for palmar impressions has been previously undertaken. We report the results of the first large-scale black box study to establish a discipline-wide error rate estimate for palmar comparisons. The 226 latent print examiner participants returned 12,279 decisions over a dataset of 526 known ground-truth pairings. There were 12 false identification decisions made yielding a false positive error rate of 0.7%. There were also 552 false exclusion decisions made yielding a false negative error rate of 9.5%. Given their larger number, false negative error rates were further stratified by size, comparison difficulty, and area of the palm from which the mark originated. The notion of “questionable conclusions,” in which the ground truth response may not be the most appropriate, is introduced and discussed in light of the data obtained in the study. Measures of examiner consistency in analysis and comparison decisions are presented along with statistical analysis of the ability of many variables, such as demographics or image quality, to predict outcomes. Two online apps are introduced that will allow the reader to fully explore the results on their own, or to explore the notions of frequentist confidence intervals and Bayesian credible intervals.

# Data Set

- 226 Participants compared Palm impressions
- 12,279 decisions on 526 ground-truthed pairings
- Examiners provided with a single latent impression and provided with a single hand of an individual (right or left hand)
- Examiners asked to decide:
  - Value/Suitability for Comparison
  - Exclusion, Inconclusive or Identification

# Sample Characteristics

27.1% of the samples were ground truth nonmates

72.9% of the samples were ground truth mates

On average, 23 examiners viewed each sample

Only samples with 16+ conclusions were modeled in this analysis

# Conclusion Characteristics

| | Nonmated Pairs | Mated Pairs |
|---|---|---|
| Identification | 10 (0.4%) | 5244 (78.5%) |
| Inconclusive | 733 (29.6%) | 924 (13.8%) |
| Exclusion | 1727 (70.0%) | 515 (7.7%) |

Decisions by participants
"Value for Exclusion only" decisions excluded

# Ordered Probit Model Likelihood Ratios

The goal of our approach is to calculate the relative strength of support for the same and different sources propositions for each sample based upon the distribution of responses provided by the group of practitioners to that sample.

Most black box studies aggregate all samples and practitioners to allow for each type of error to be calculated using a percentage.

This approach calculates the strength of evidence for **each individual sample.**

# Fingers vs Palms

What makes palm comparisons different from finger comparisons?

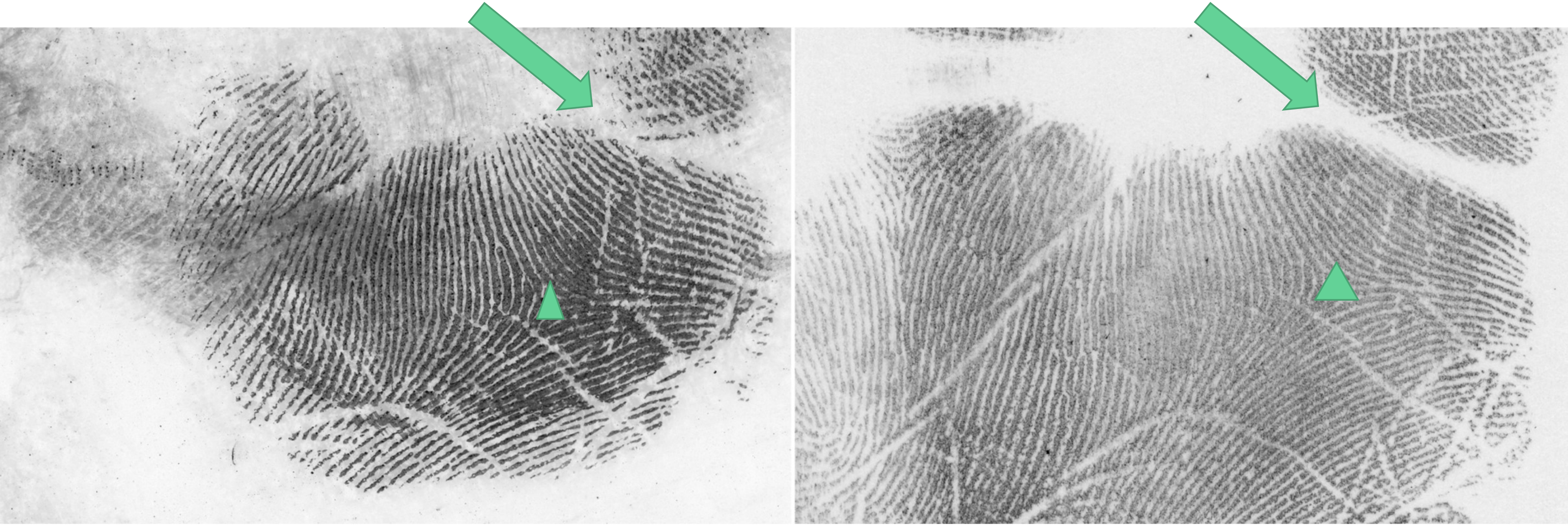
Palm latent impressions can be:

- Ambiguous in orientation
- Ambiguous in handedness
- A larger known region to search
    - Up to 10x larger
    - 8x more minutiae features when fully recorded
- More subtle or infrequent anchor points (cores or delta formations, primary creases)

# Fingers vs Palms

- Participants were 2x as likely to report **Inconclusive** in Palm latents as compared to finger latents when the samples were **non-mates**.
- Mated fingers were **10%** unanimous ID (Ulery 2011), Mated palms were **25%** unanimous ID.
- **36** ground truth mated pairs were majority **Exclusions**
    - Only 1 sample like this in the black box finger data

# More Information = Easier Comparison



This impression contains a delta formation, a primary crease and is of high contrast

# Less Information = Harder Comparison

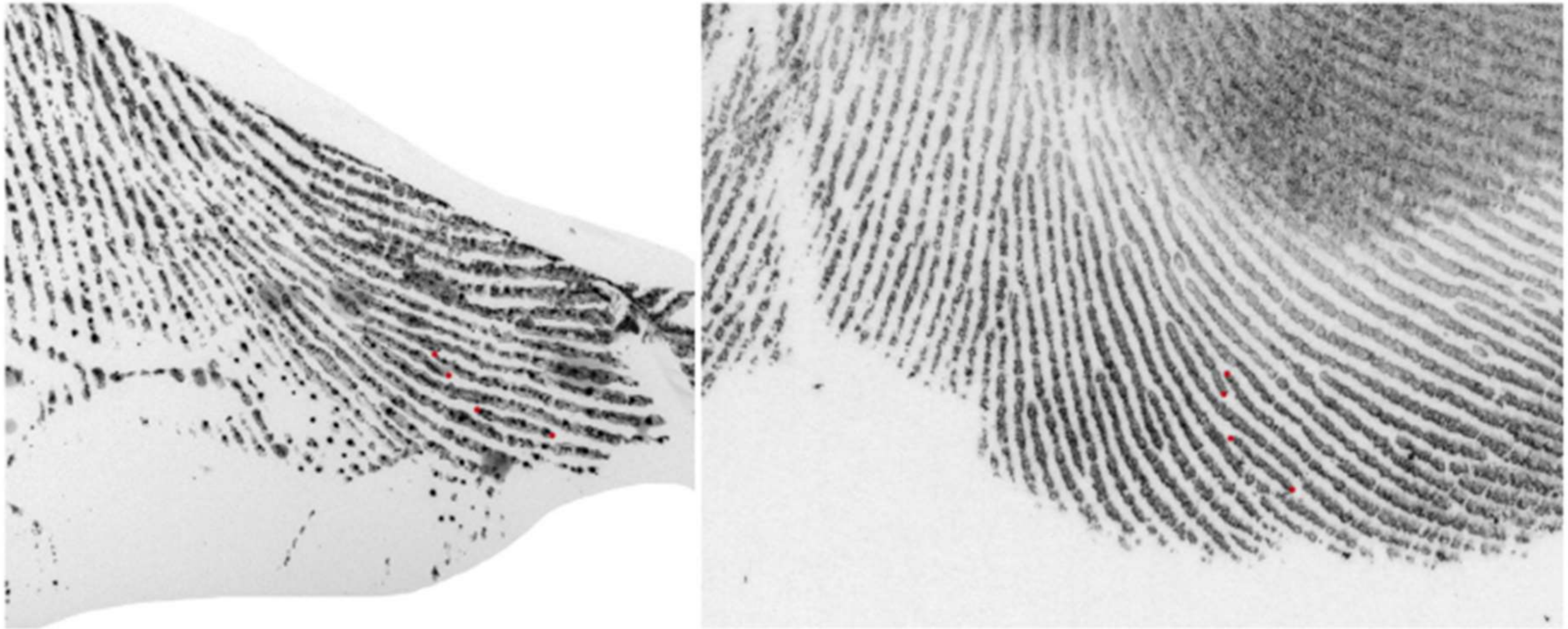


**Fig. 15.** Comparison starting point for case 0458. Four minutiae in common have been marked in red to get the reader started (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

This impression is ambiguous in its orientation and contains no primary features to anchor.

# Strength of Support

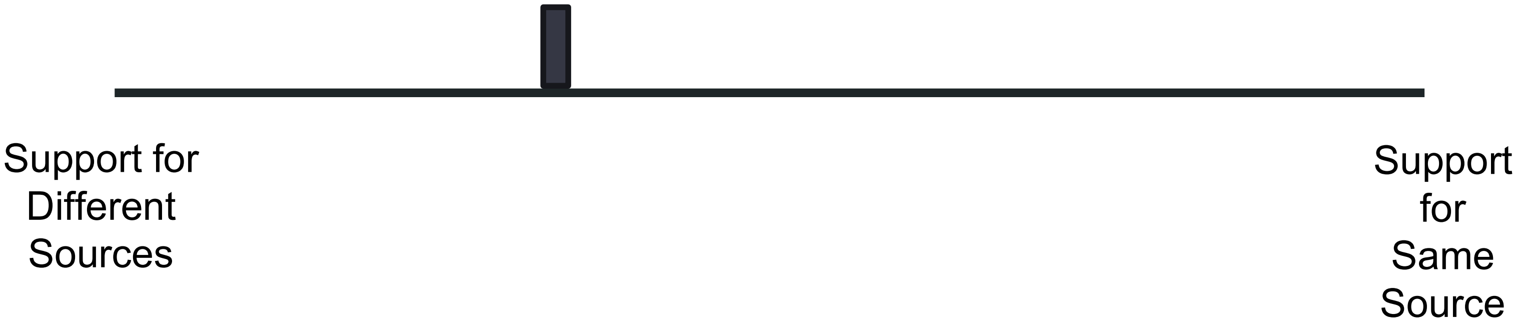
What we're most interested in is the support for the same source **proposition**. A **proposition** is a state of the world, either the impression came from the same source, or it didn’t. In casework we can never *know* the answer, but in black box studies we can.

The number of examiners reaching a particular conclusion can provide a **sense** of the support for the same source proposition.

All examiners reporting Identification probably indicates more support for the same source proposition than if half of the examiners said Identification and half said Inconclusive.

How do we group decisions like Identification, Inconclusive or Exclusion though? They’re not numerical…

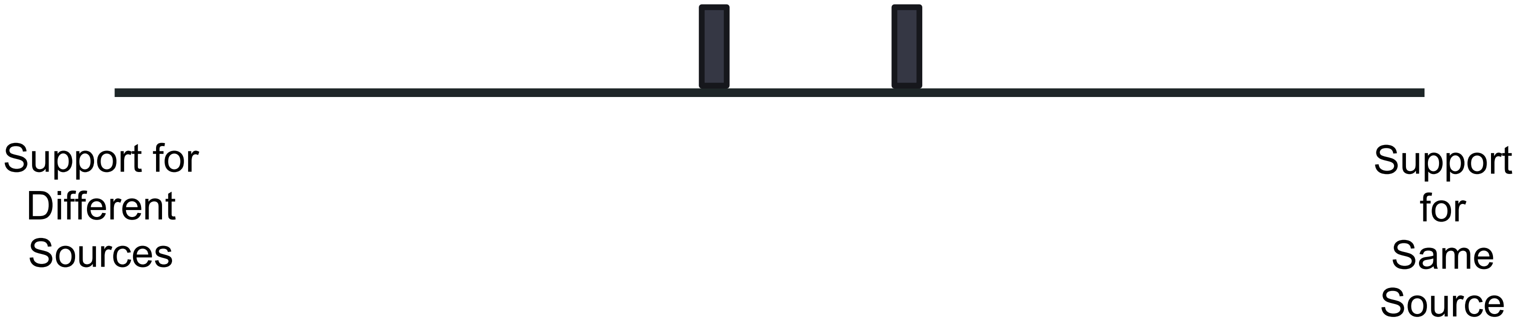
# In the Minds of Examiners



Support for Different Sources

Support for Same Source

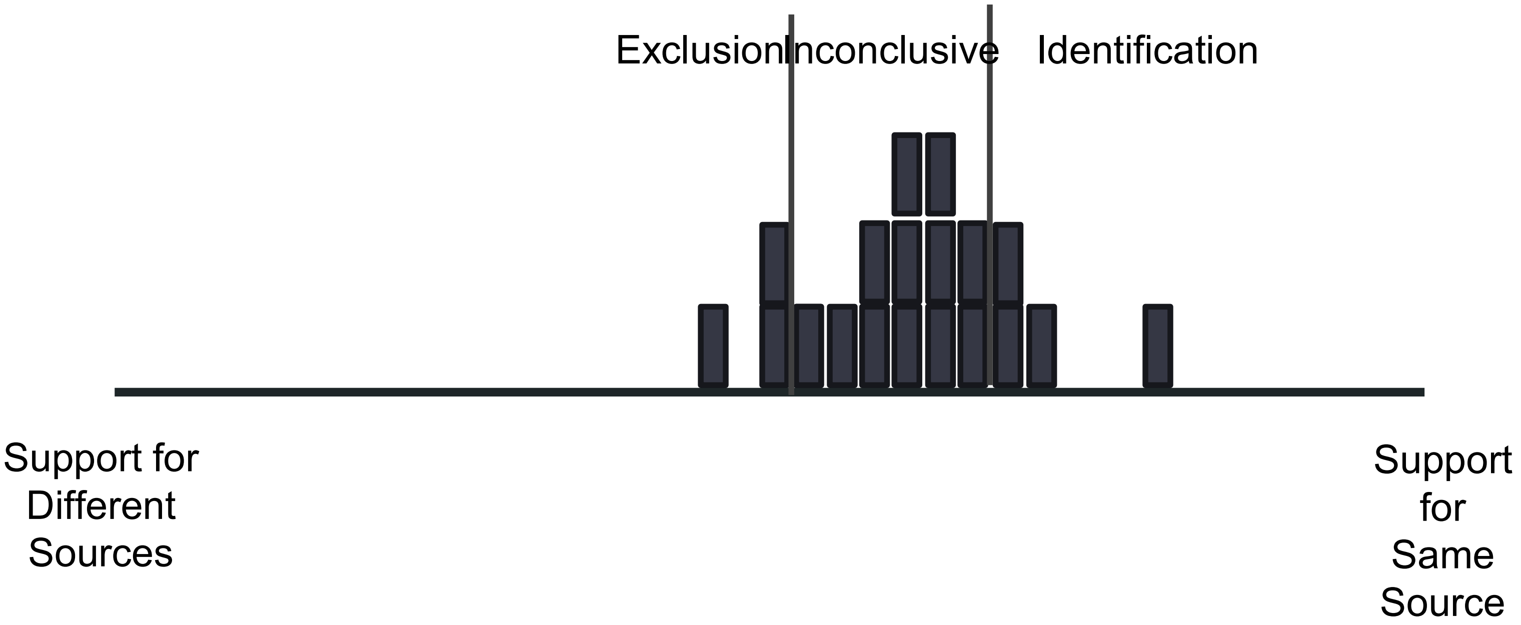
# A second examiner



Support for Different Sources

Support for Same Source

# 20 Examiners

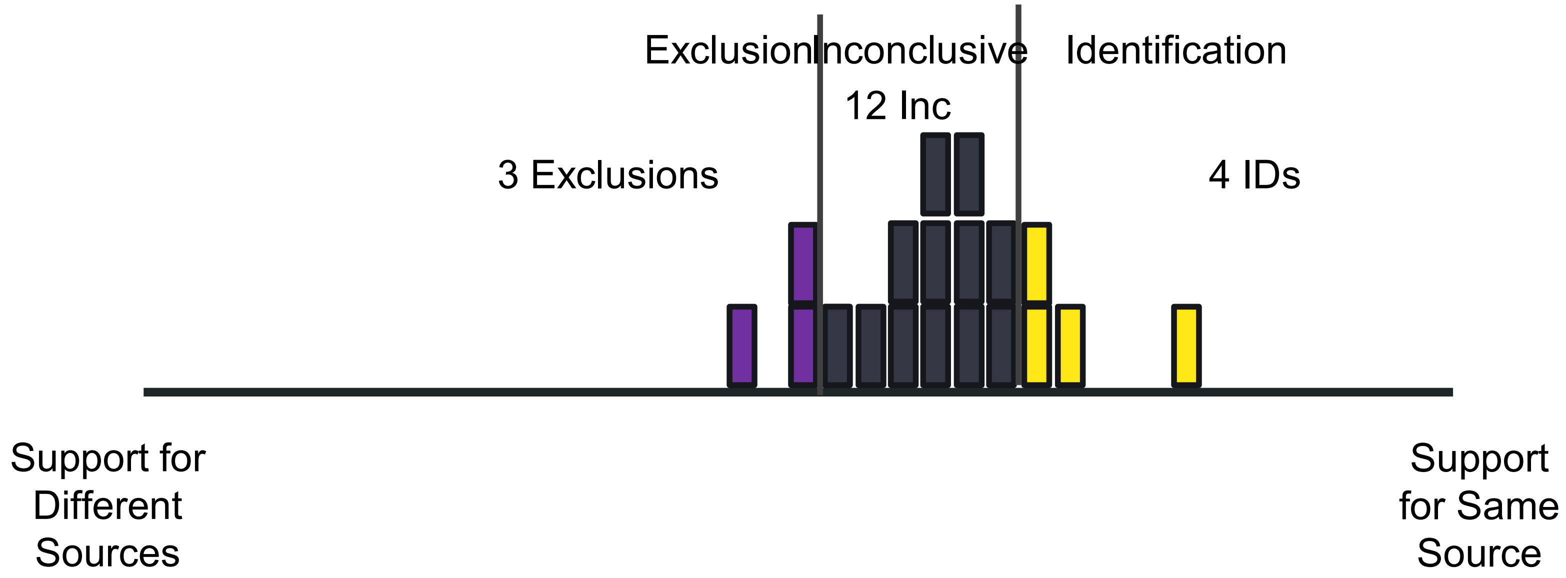


Exclusion Inconclusive Identification

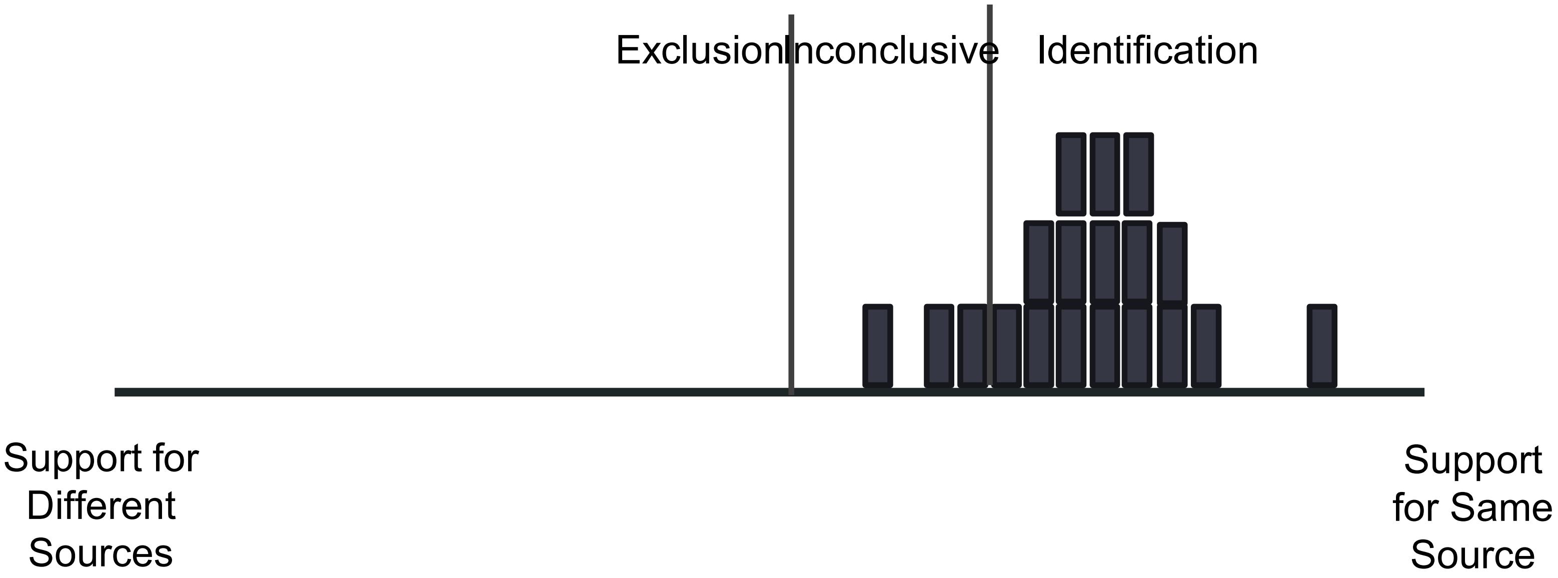
Support for
Different
Sources

Support
for
Same
Source

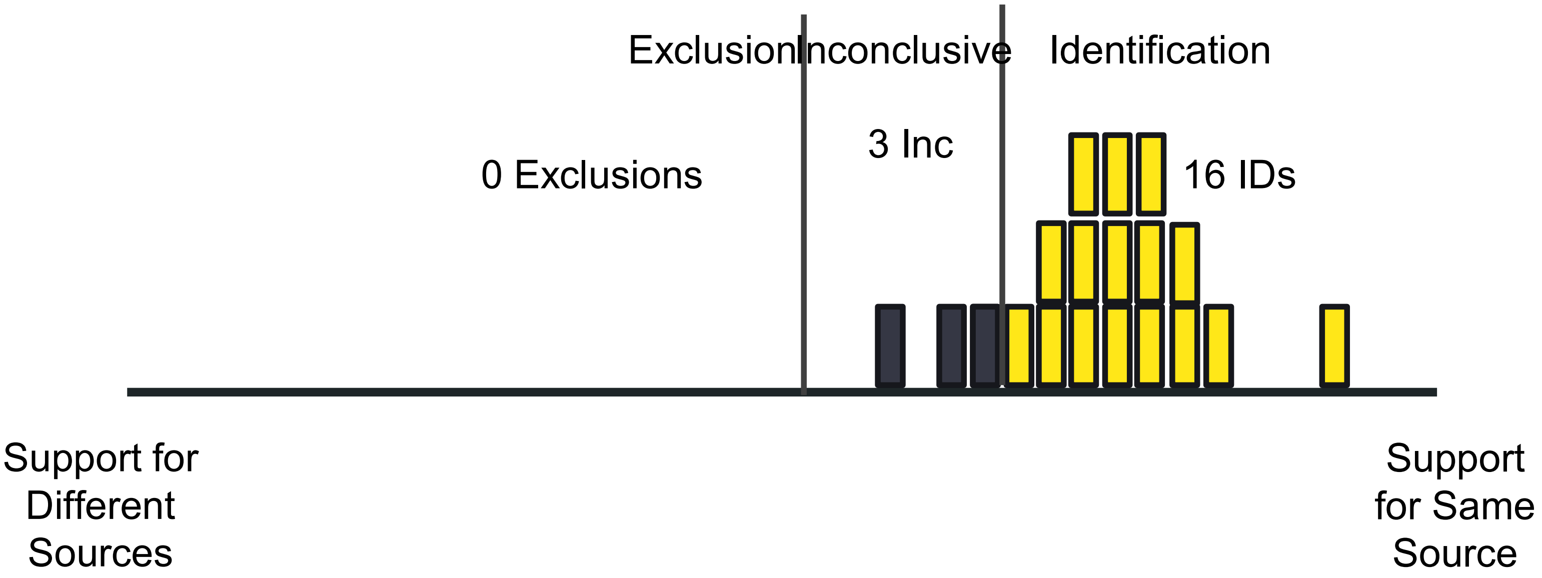
# Apply decision thresholds



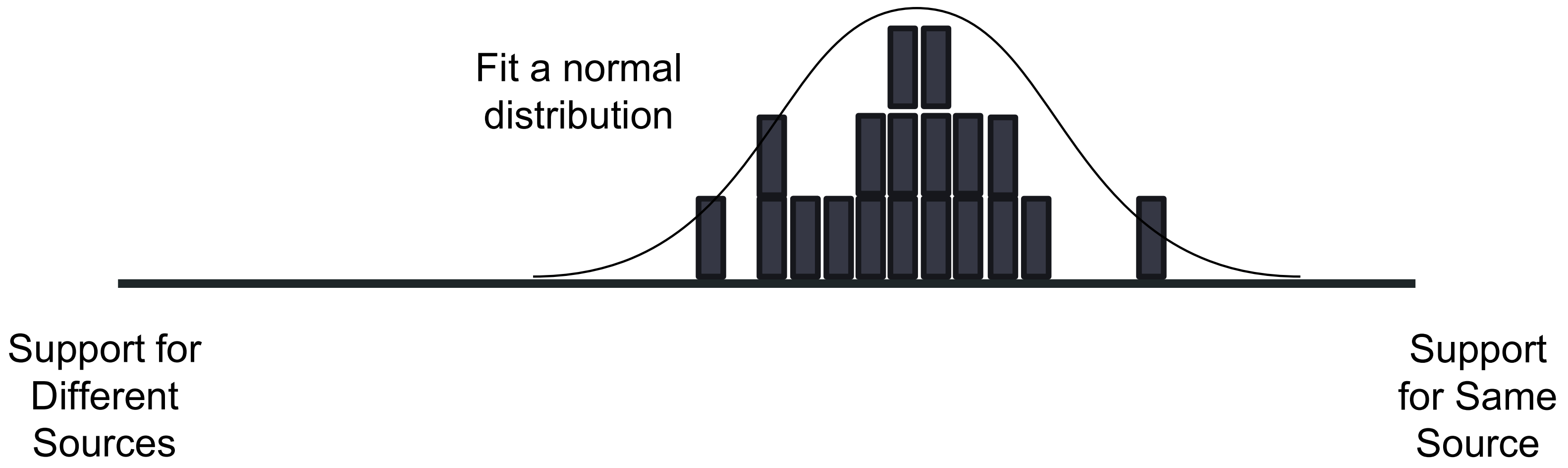
# A print pair with more specificity/ “uniqueness”



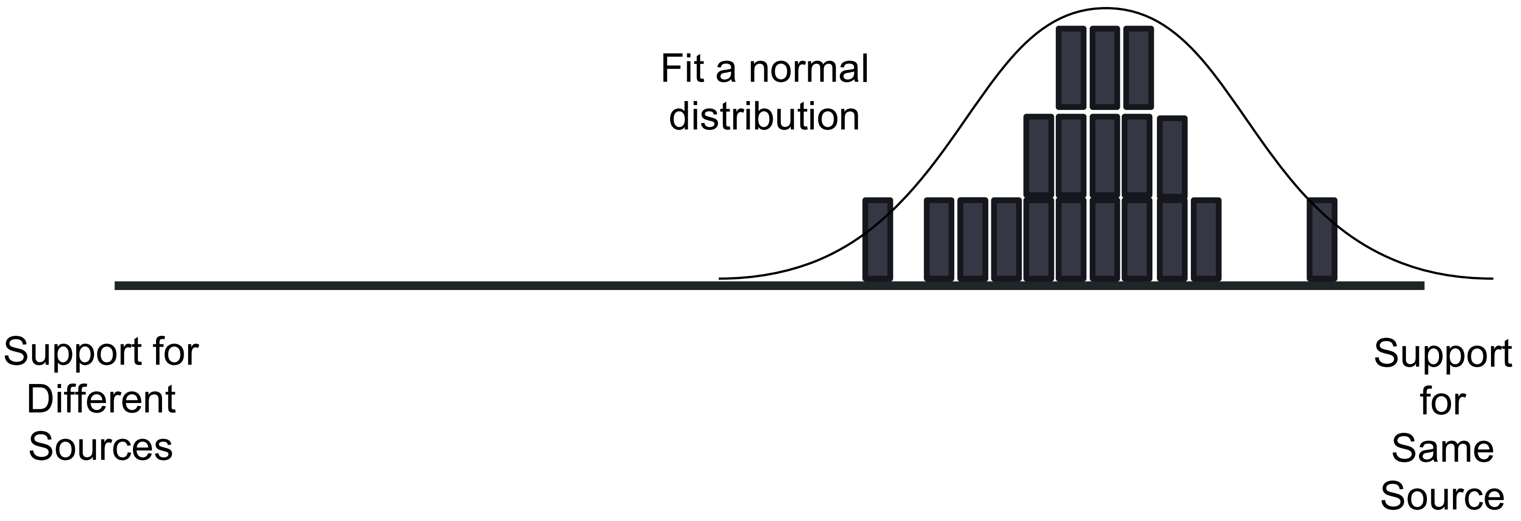
# Apply decision thresholds

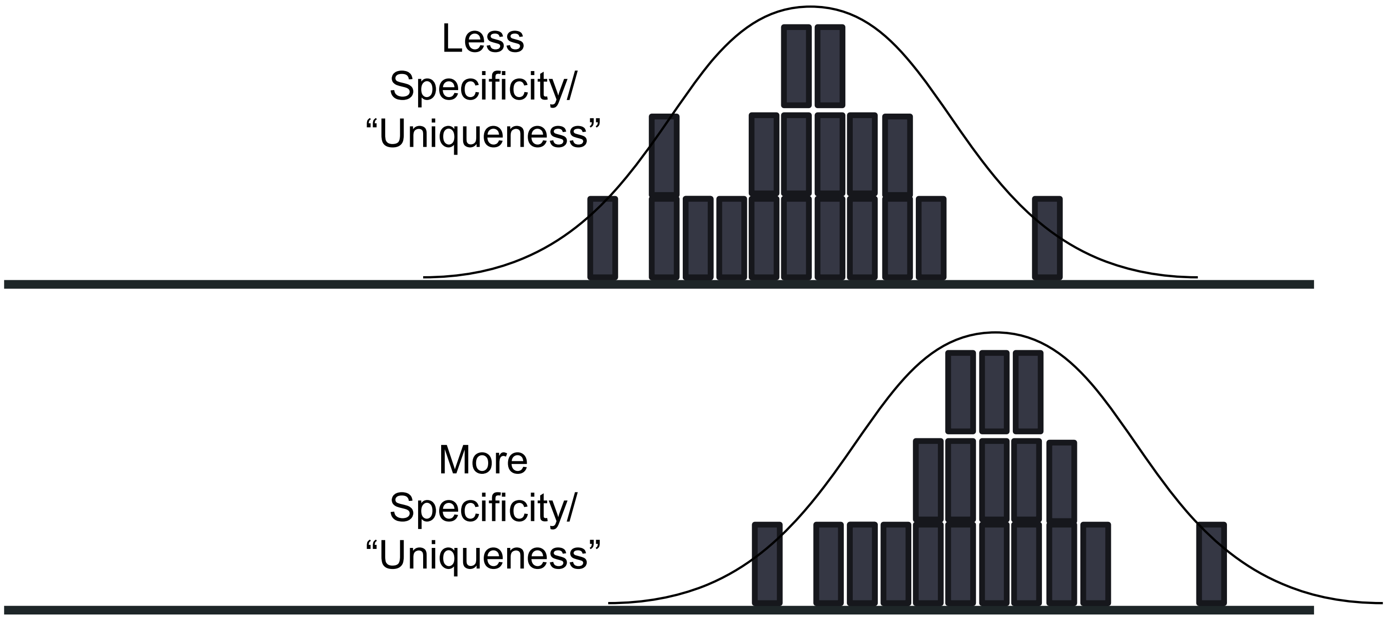


# How can we summarize the distribution?



# A print pair with more specificity/ “uniqueness”





Less
Specificity/
“Uniqueness”

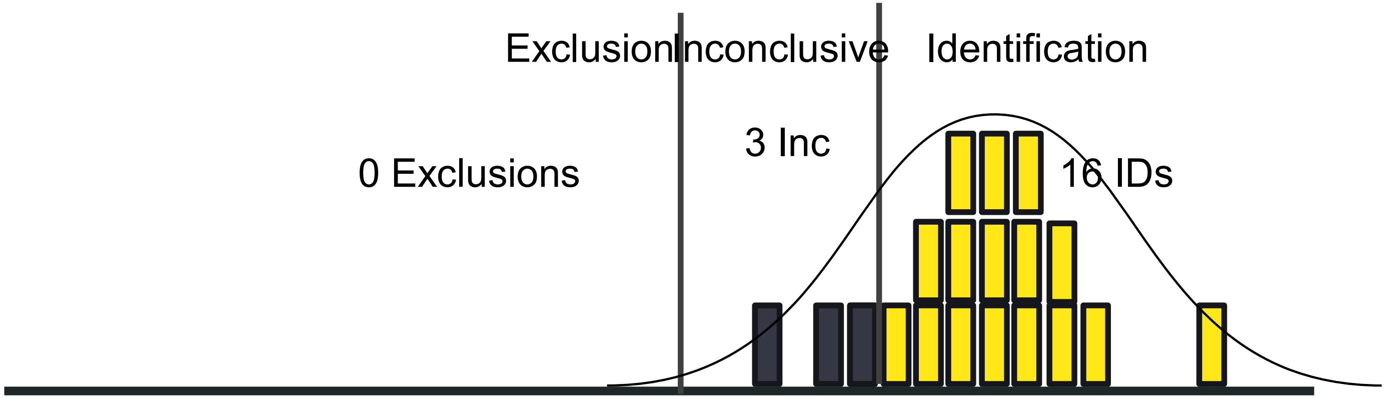
More
Specificity/
“Uniqueness”

Support for
Different
Sources

Support
for
Same
Source

# But we don’t have a “magic electrode”...

We can work backward from the distribution of responses to infer what the underlying distribution might have been.



Exclusion

Inconclusive

Identification

0 Exclusions

3 Inc

16 IDs

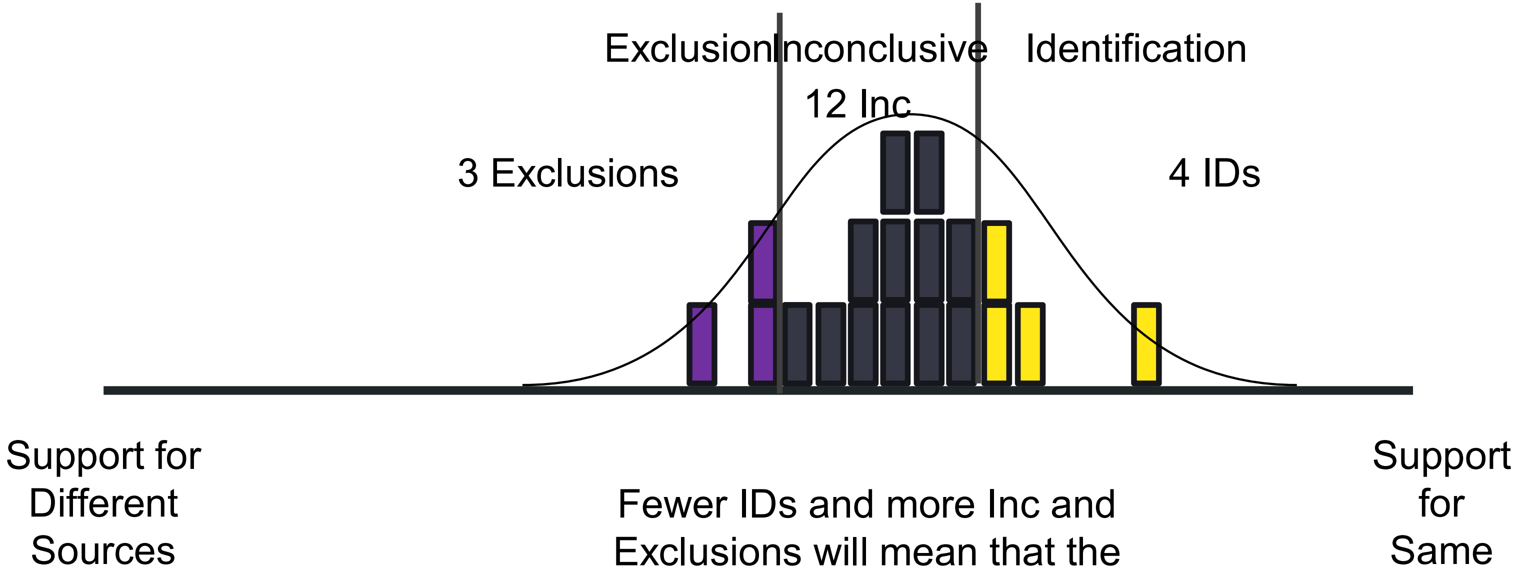
Support for Different Sources

The normal distribution summarizes the typical support for the same source proposition

Support for Same Source

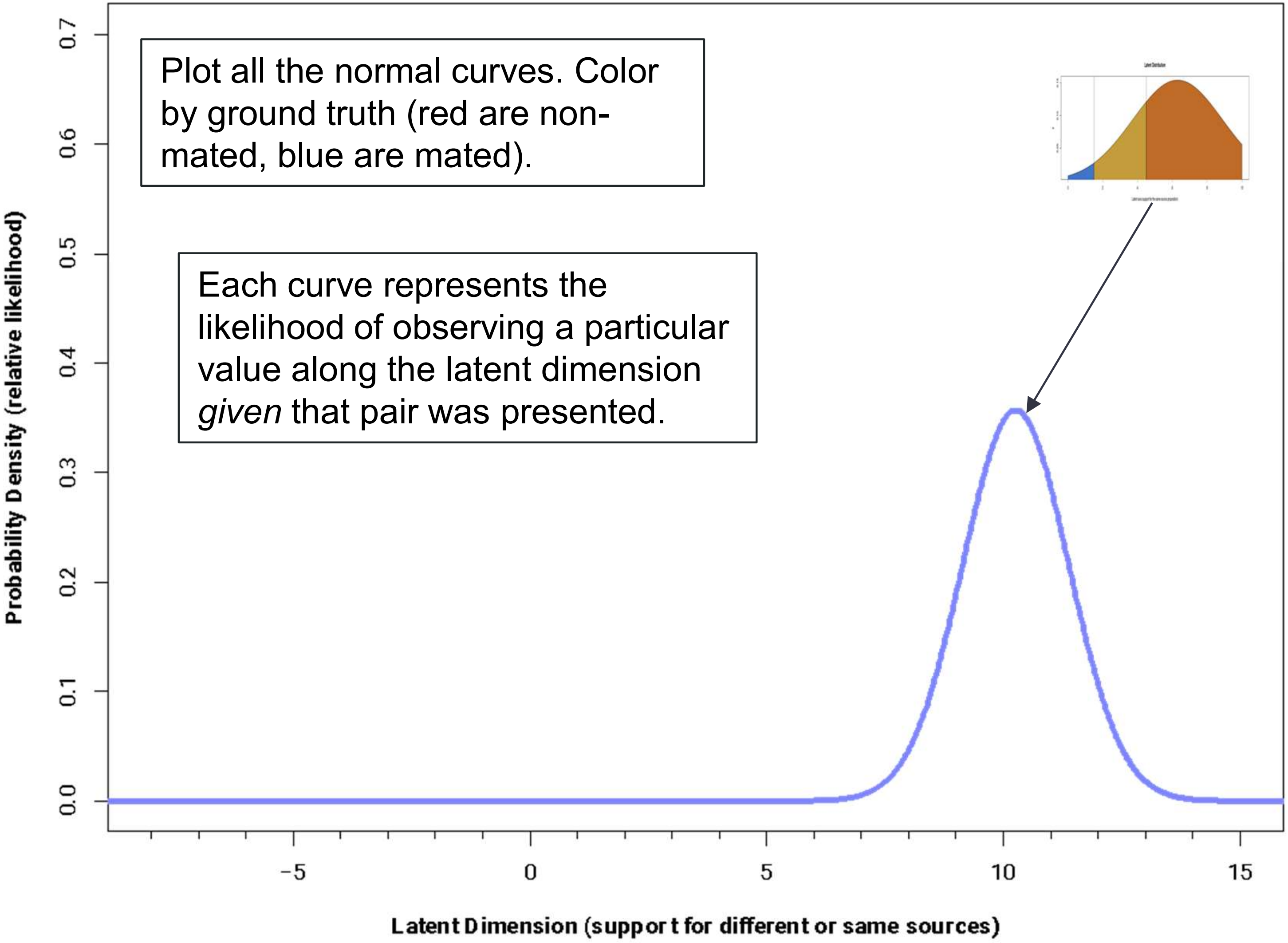
# But we don’t have a “magic electrode”...

We can work backward from the distribution of responses to infer what the underlying distribution must have been.



Fewer IDs and more Inc and Exclusions will mean that the normal distribution is shifted to the left.

# Ordered Probit Model Estimation (Busey et al. Data)



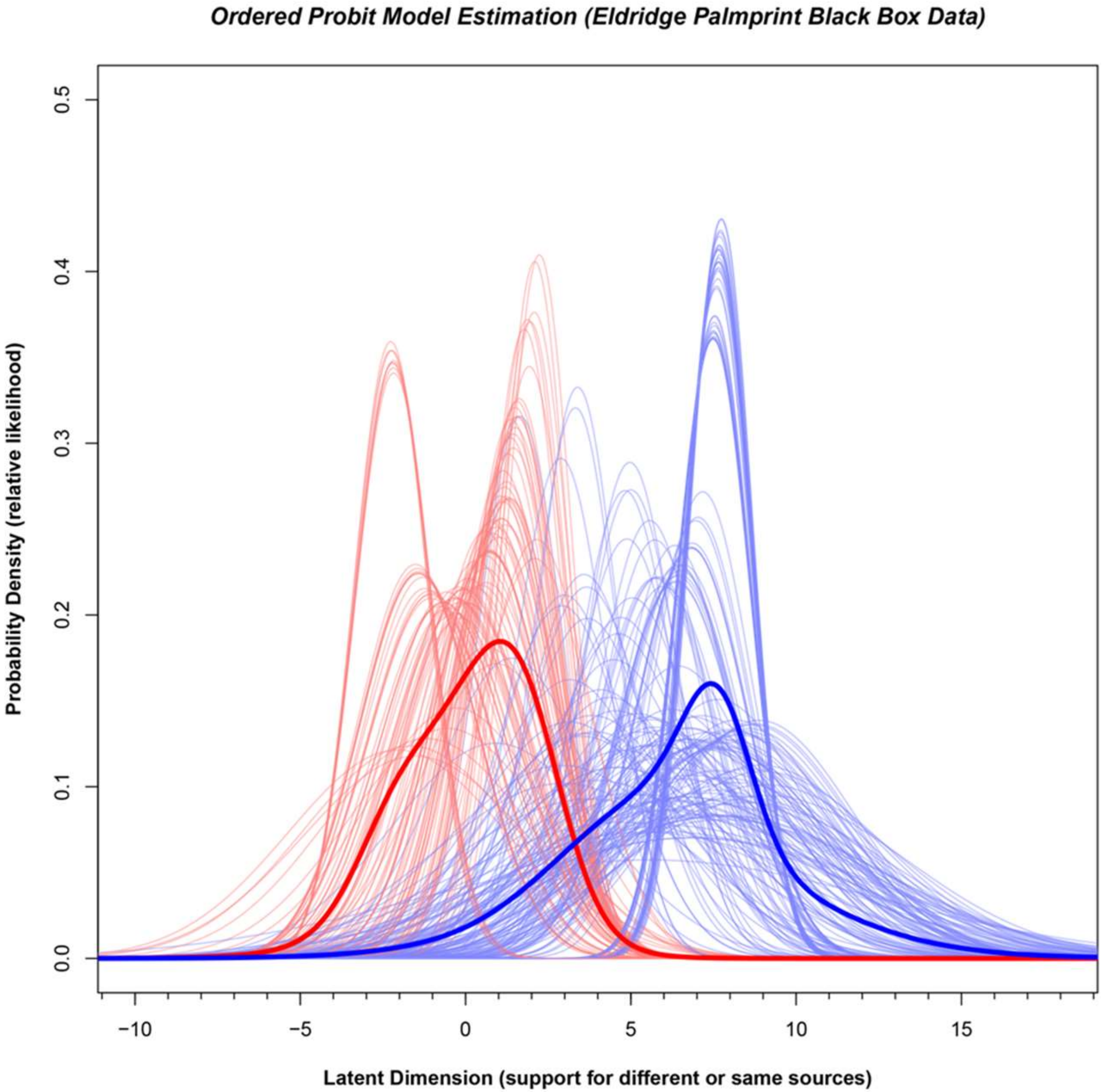
Plot all the normal curves. Color by ground truth (red are non-mated, blue are mated).

Each curve represents the likelihood of observing a particular value along the latent dimension *given* that pair was presented.

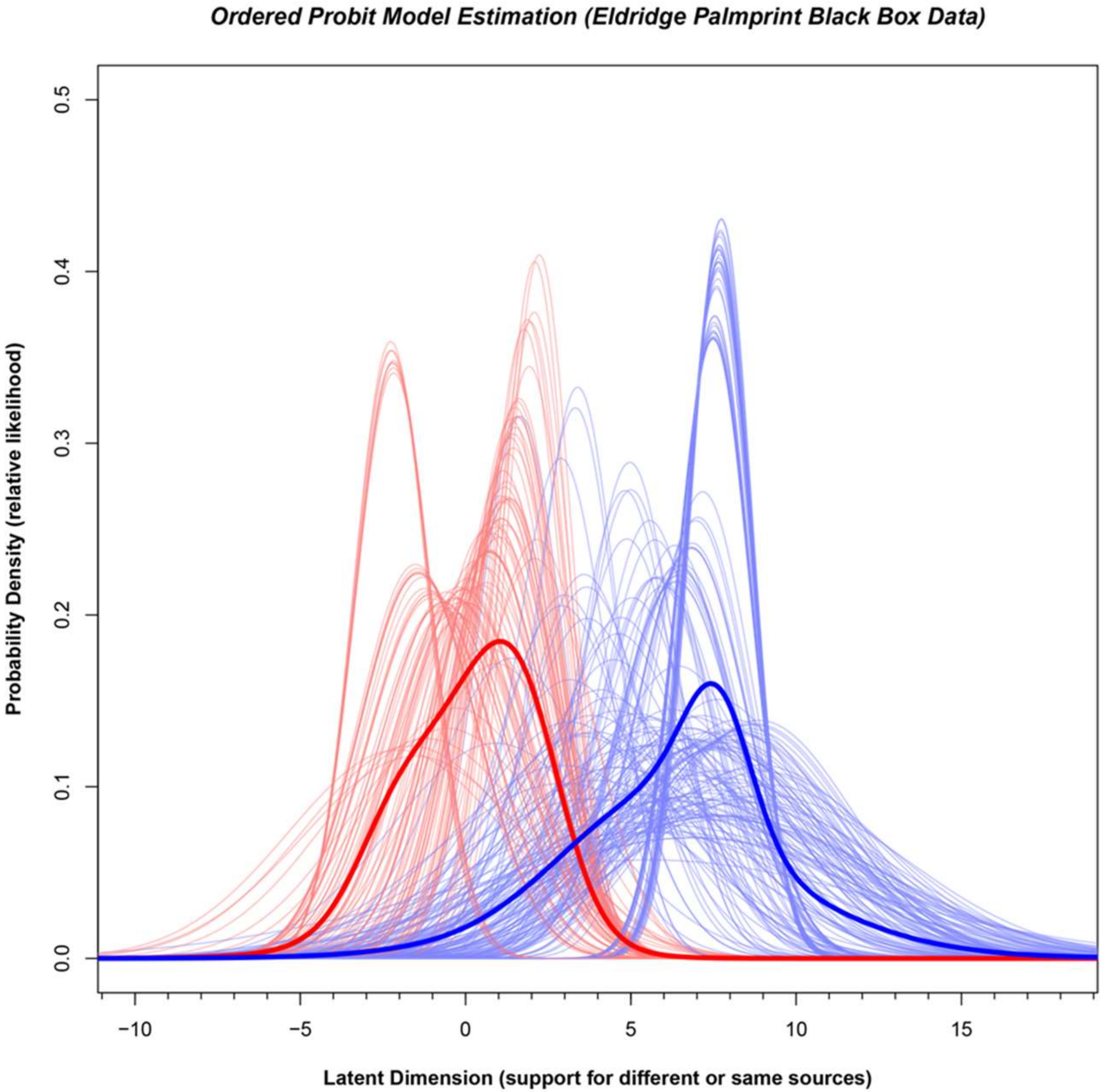
Red curves indicate samples which were non-mated.

Blue curves indicate samples which were mated.

Bold curves indicate the summation of the colors into a combined distribution for non-mated and mated. This assumes all trials are independent and we use the "or" rule to justify summing them.

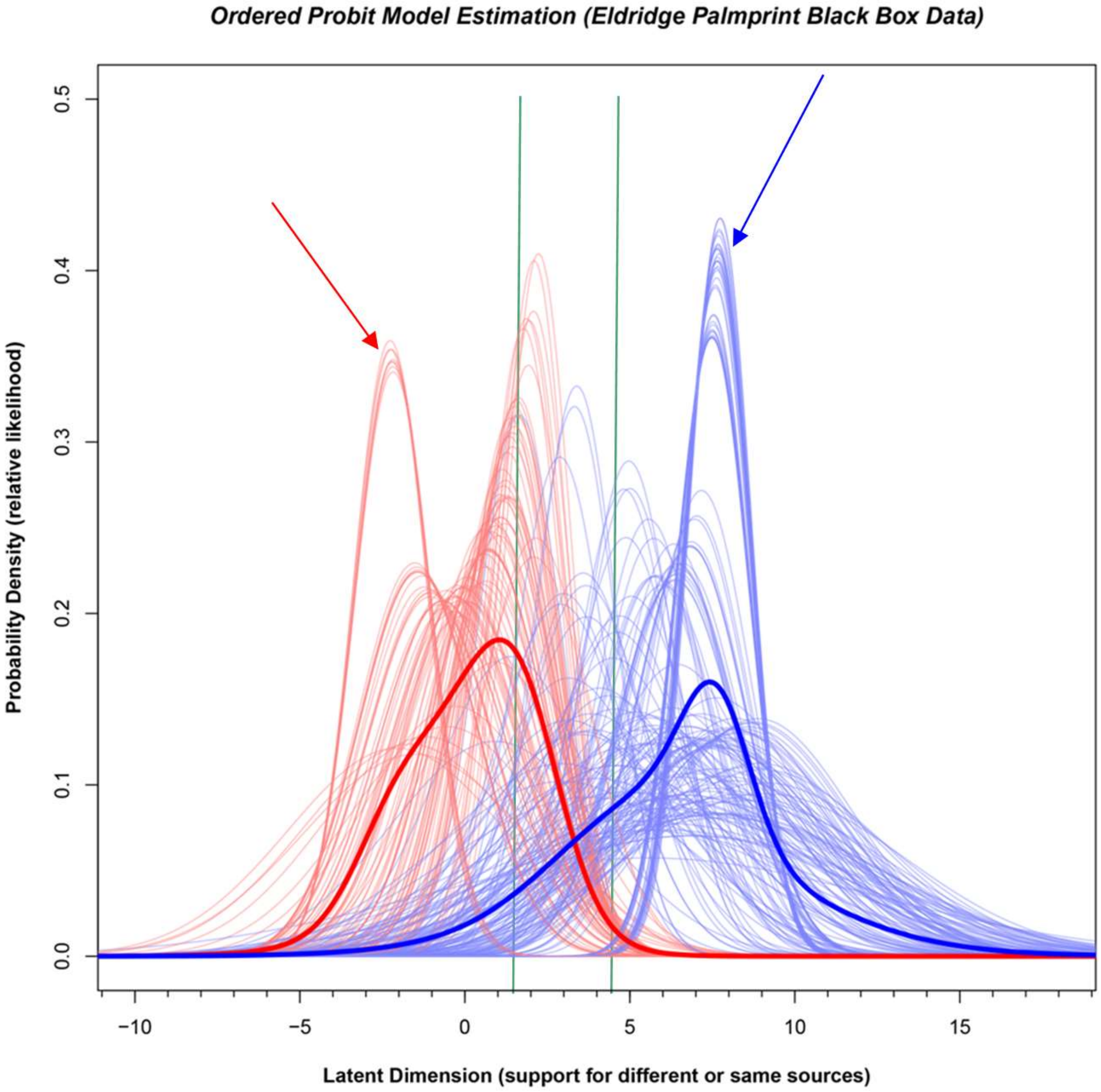


Curves which are further to the left indicate those which have less support for the same source proposition, those further to the right have more support for the same source proposition.



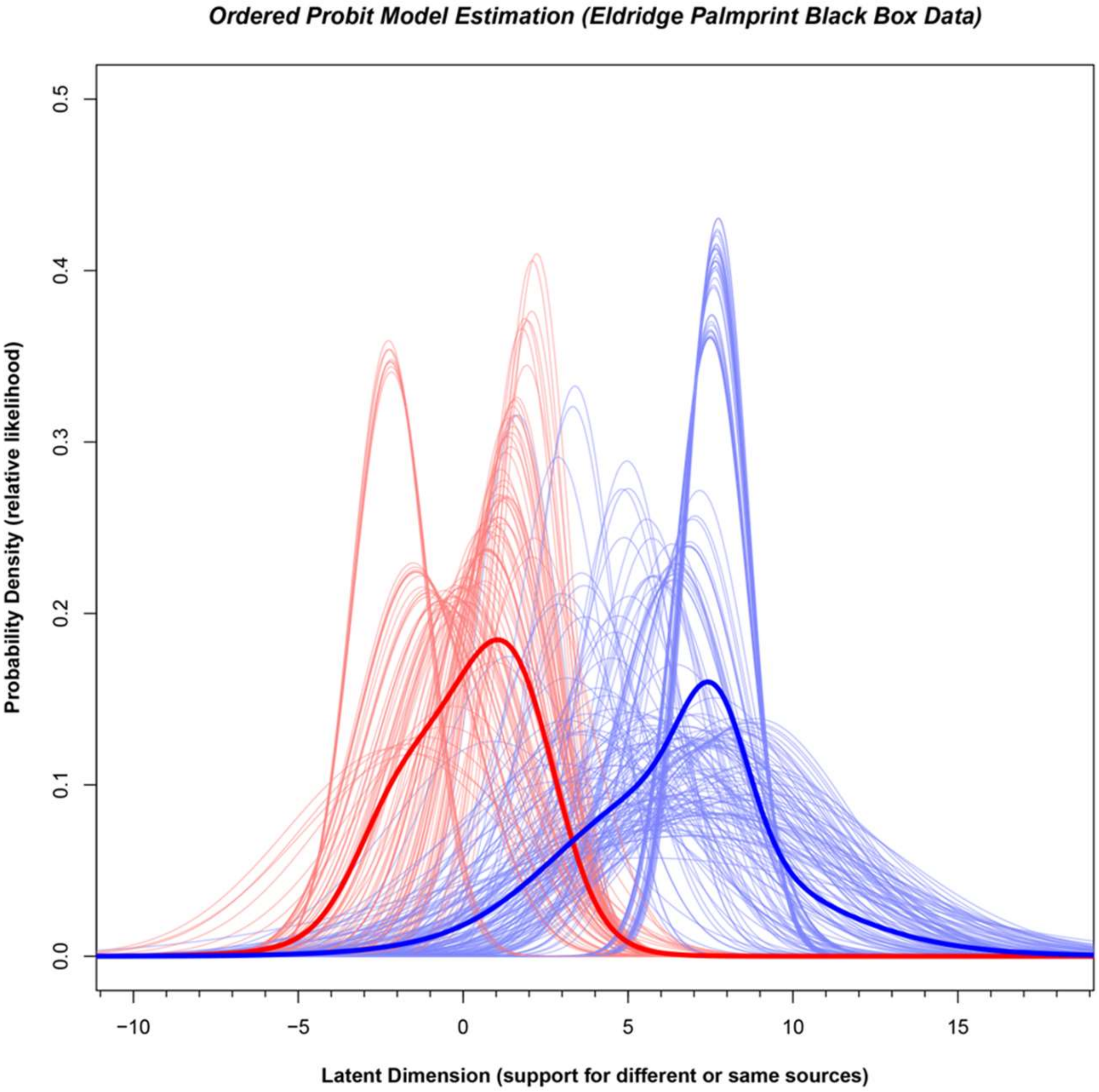
Green lines indicate the thresholds surrounding the inconclusive decision.

Arrows indicate samples where a unanimous agreement was reached amongst the participants for Exclusion or Identification decisions.



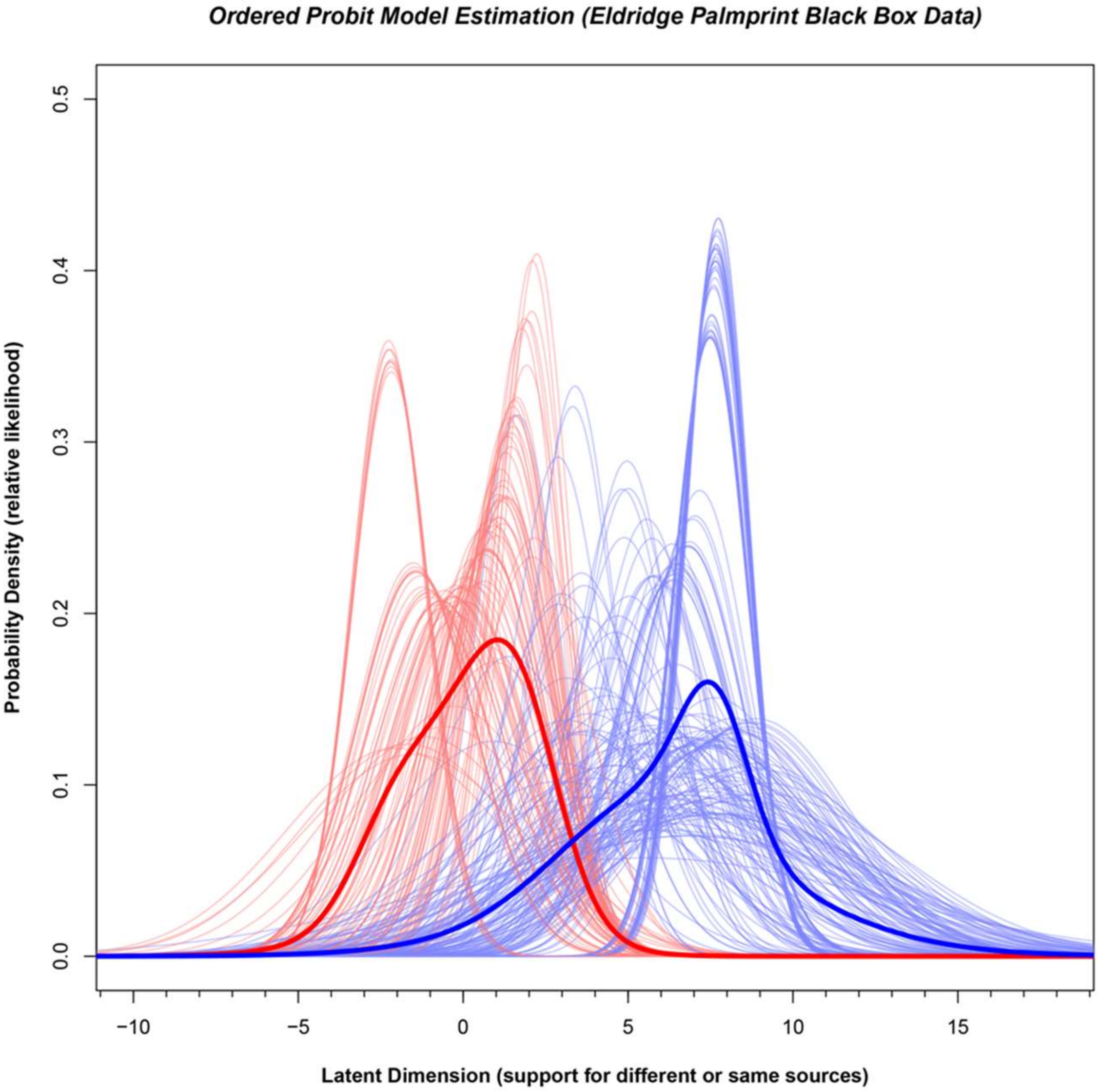
The variation between the samples is demonstrated by each curve's position along the x-axis.

Some samples created a great deal of disagreement between examiners, and this is expressed with wider standard deviations.



On mated samples, the Identification decision was unanimous **25%** of the time (peak at 8)

On nonmated samples, the Exclusion decision was unanimous **7.1%** of the time. (Peak at –2.5)



# Likelihood Ratios

In general, likelihood ratios are defined as the probability of the data observed ***given*** the truth of the world. In any given comparison, the prints are either mated, or nonmated. Prints are either from the same source of skin or they’re not.

We are calculating the likelihood ratio or odds of a **mated pair**

$$\frac{\text{p(Data observed given Mated)}}{\text{p(Data observed given Nonmated)}}$$

What's the probability we would observe this amount of correspondence in the features (for example 12 corresponding minutiae) between the latent and the known ***given*** these two possible universes?

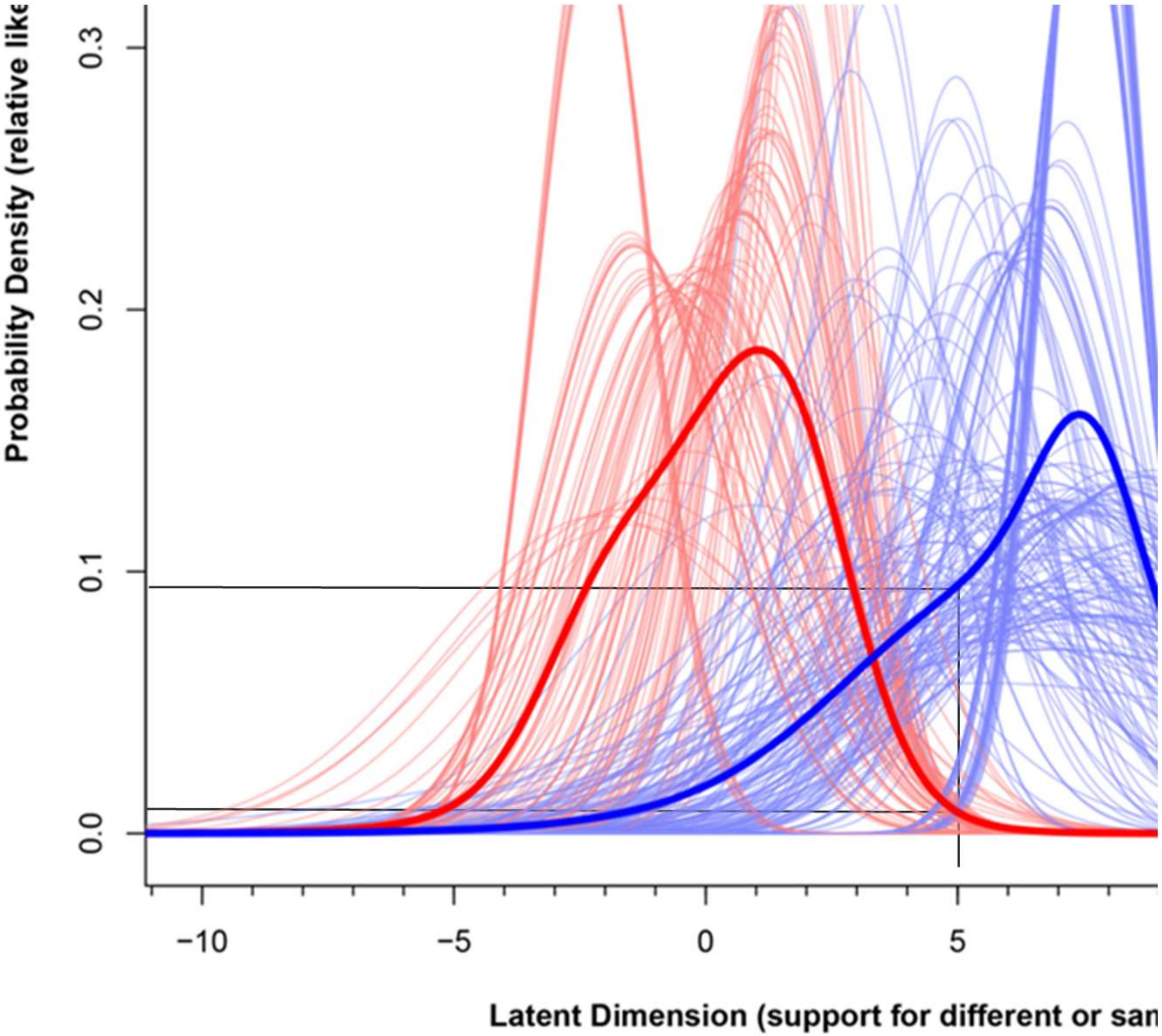
Since we use more than just minutiae, and rely on examiners to interpret information, the location of a distribution along the x-axis is the **data observed** in our approach.

The bold blue line is the numerator in our likelihood ratio.

The bold red line is the denominator.

At each point along the x-axis, the values of the blue and red lines can be compared.

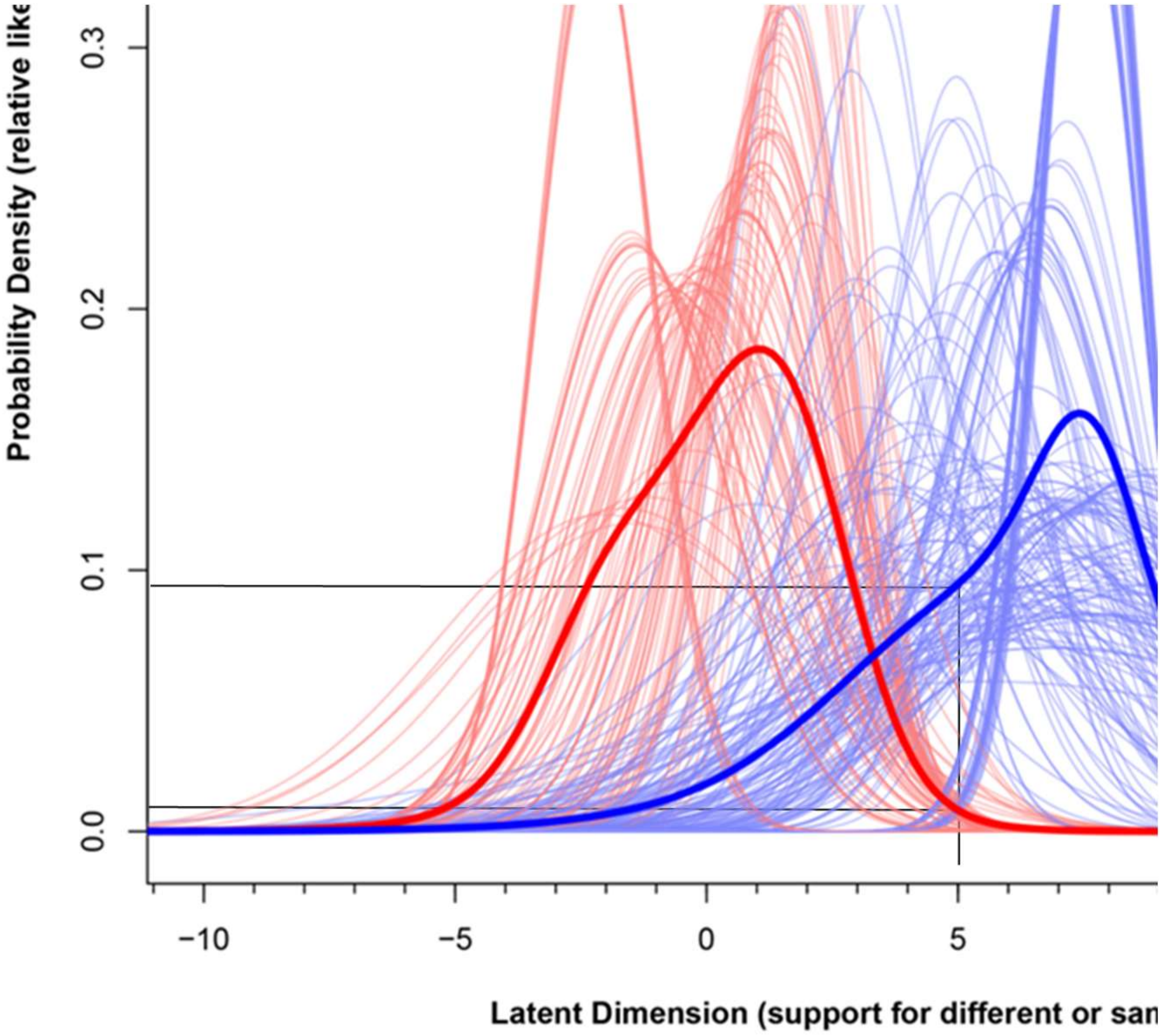
At a value of 5 on the Latent Dimension, we observe a value for the blue bold line, and a red bold line on the y-axis.



The height of the blue curve at a latent value of 5 indicates the probability of observing a 5 *given* a mated pair.

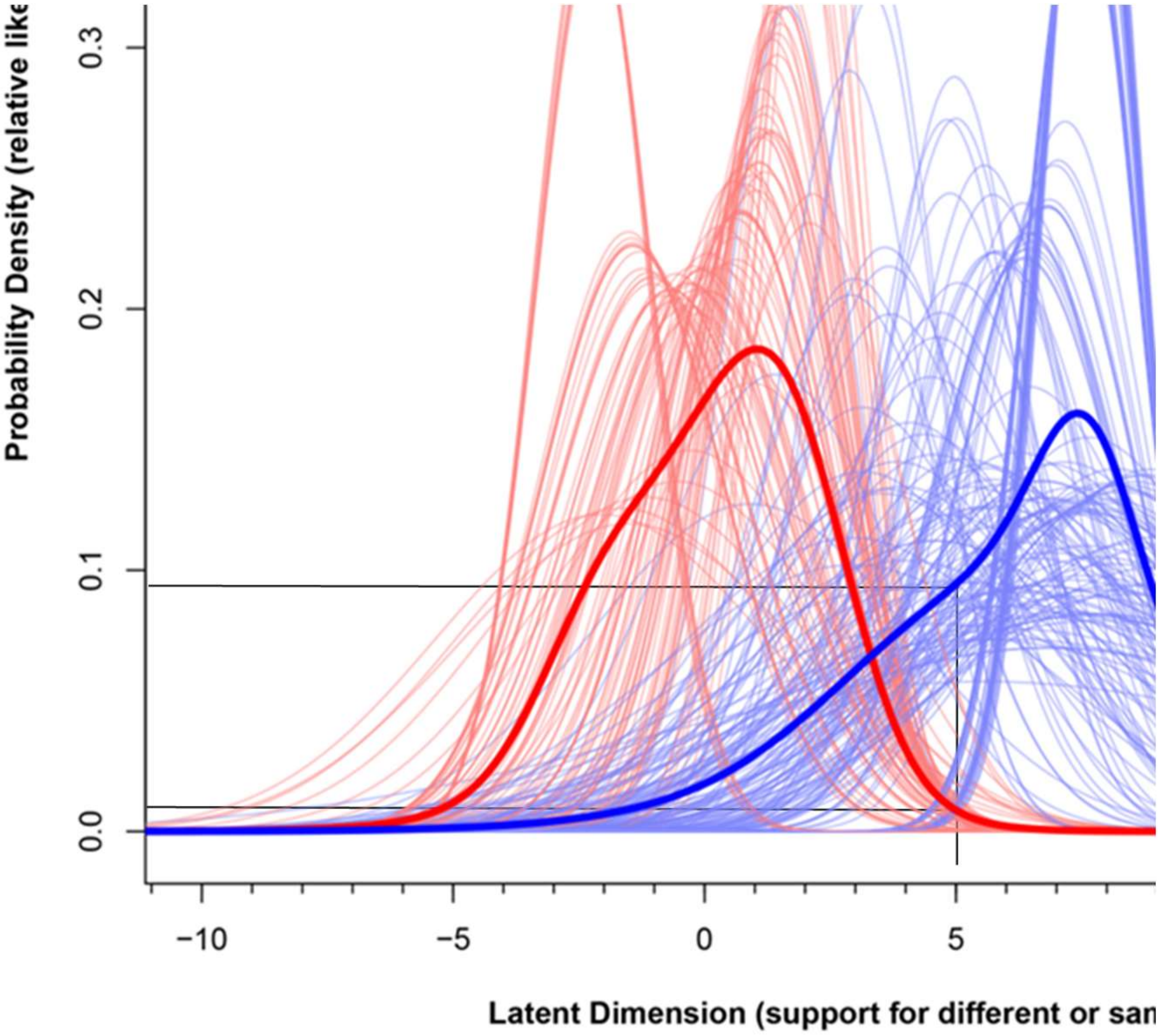
The heigh of the red curve at a latent value of 5 indicates the probability of observing a 5 given a nonmated pair.

How likely is it we would see 5 minutiae in common given we are viewing a mated pair? What about a nonmated pair?



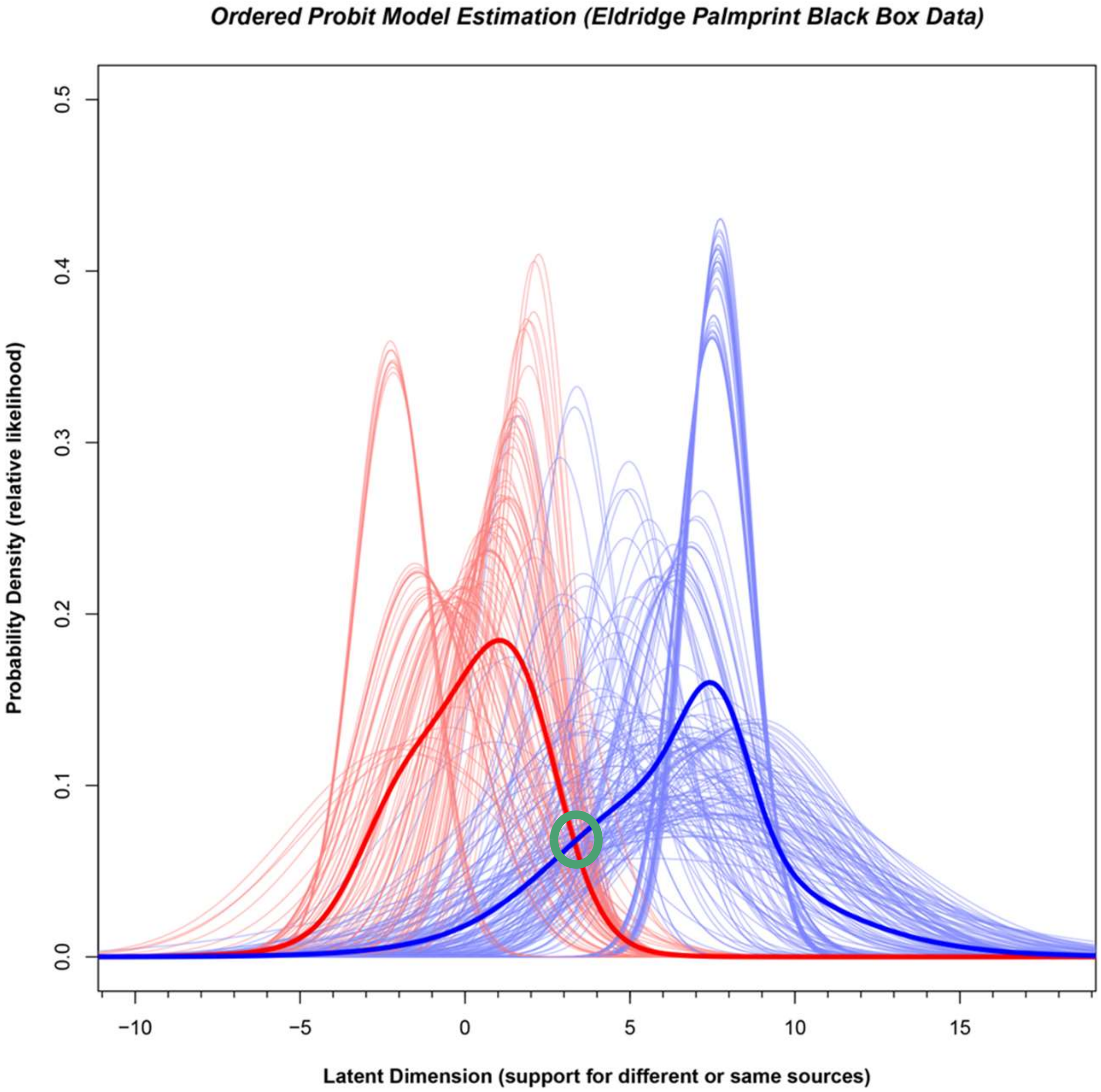
The Likelihood Ratio is weighing the probability of the observations given two possible states of the world.

We divide the value of the blue line over the red line and we calculate a unitless number.



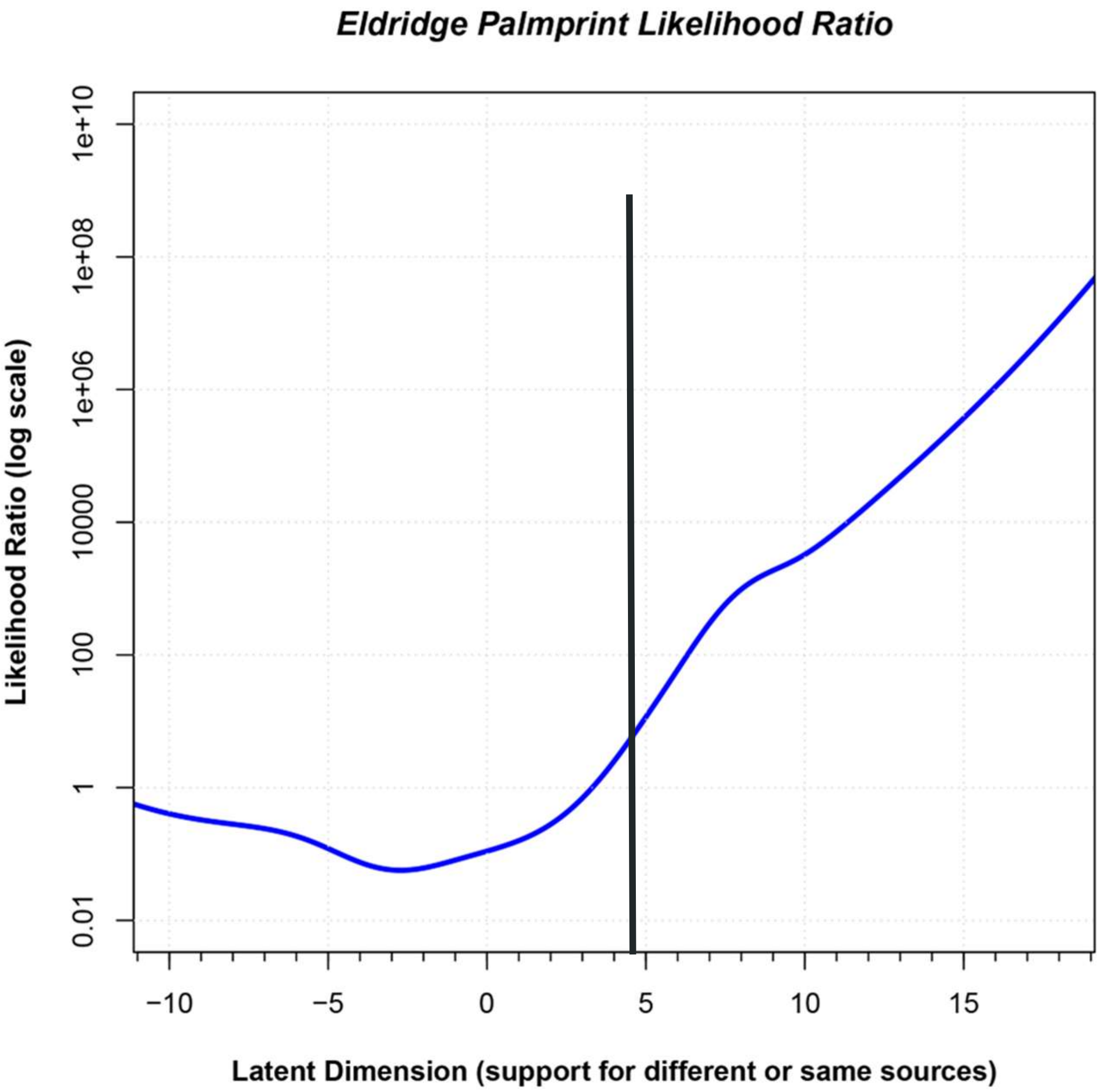
All values to the right of the green circle will create likelihood ratios greater than 1, and therefore indicate more support for the same source proposition.

All values to the left of the green circle will be between 0 and 1, and indicate more support for the different sources proposition.

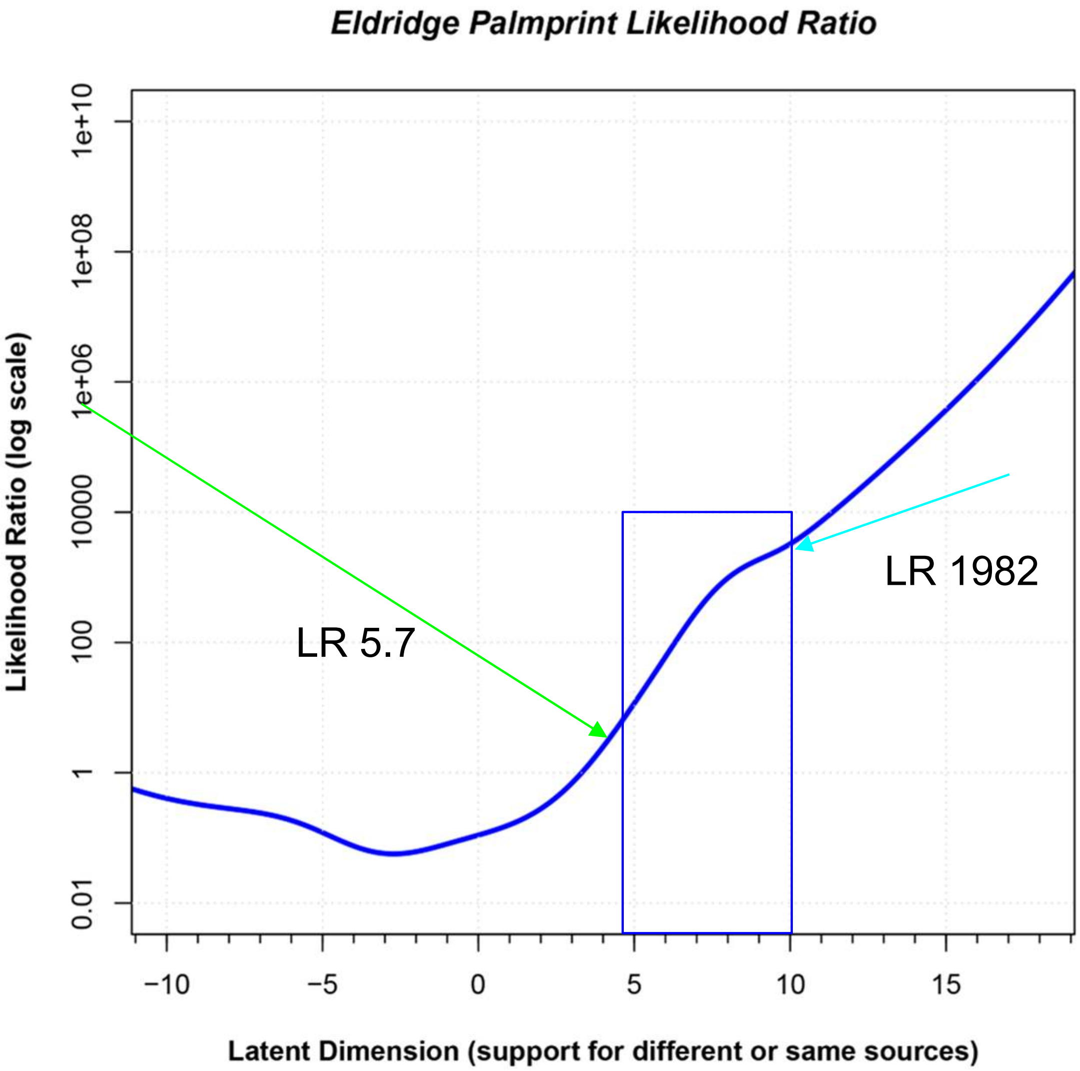


Calculating this likelihood ratio at **every point** along the two curves gives us this graph.

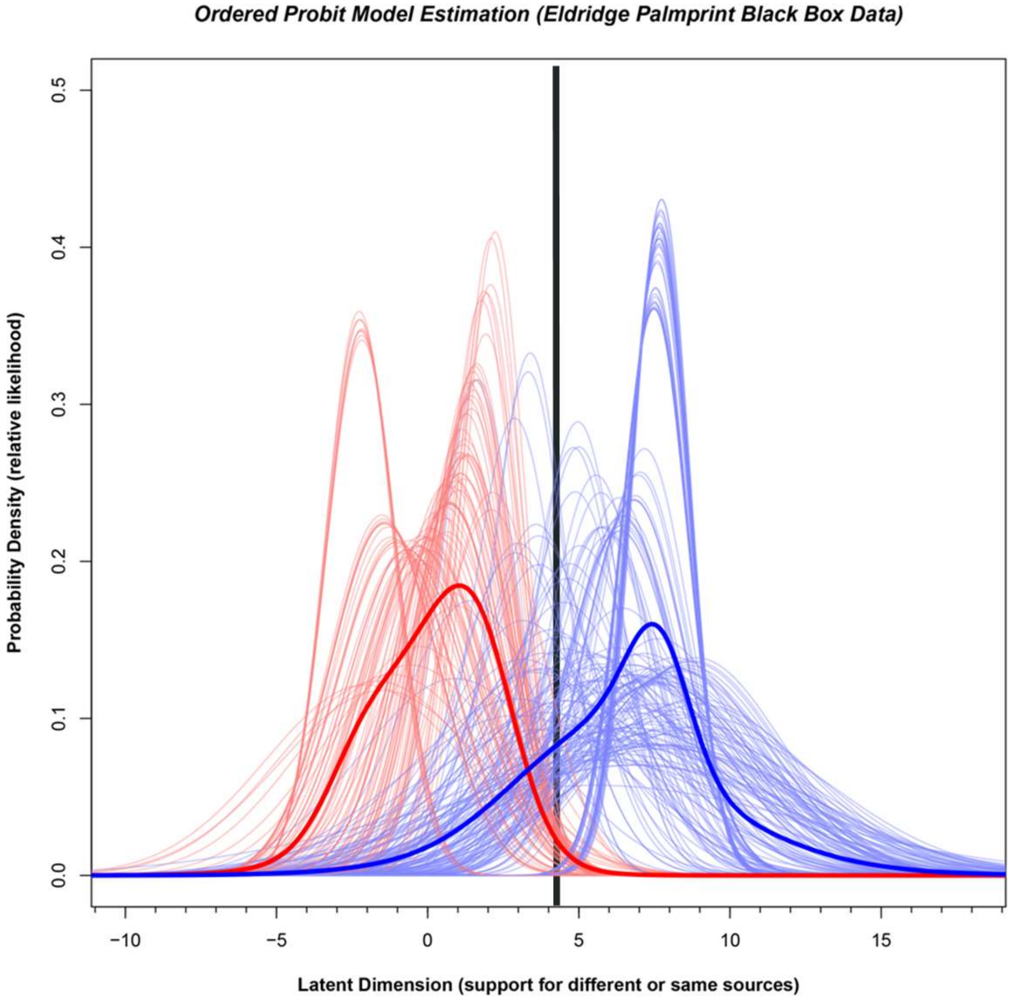
When the majority (50%) of participants use the conclusion "Identification", this falls around 4.5 for our model, indicated by the black line.

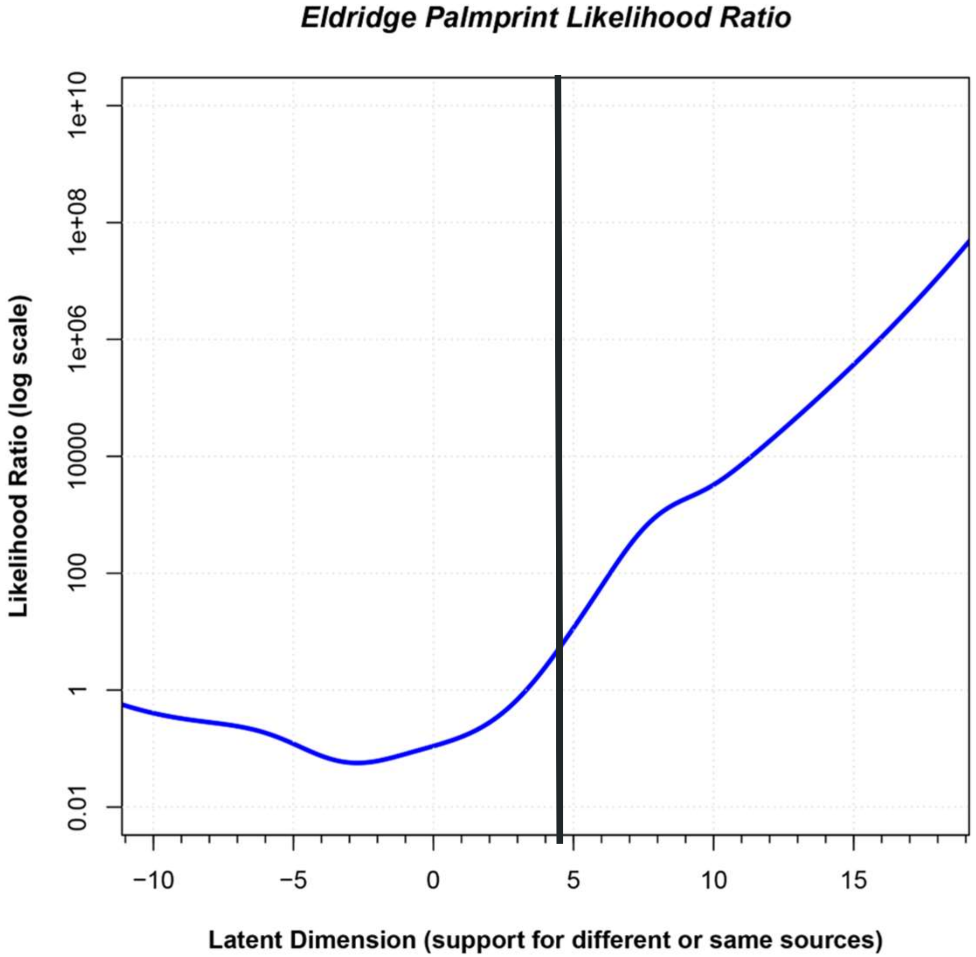


The left side of the blue box indicates the LR values for image pairs where more than 50% of examiners said “ID”

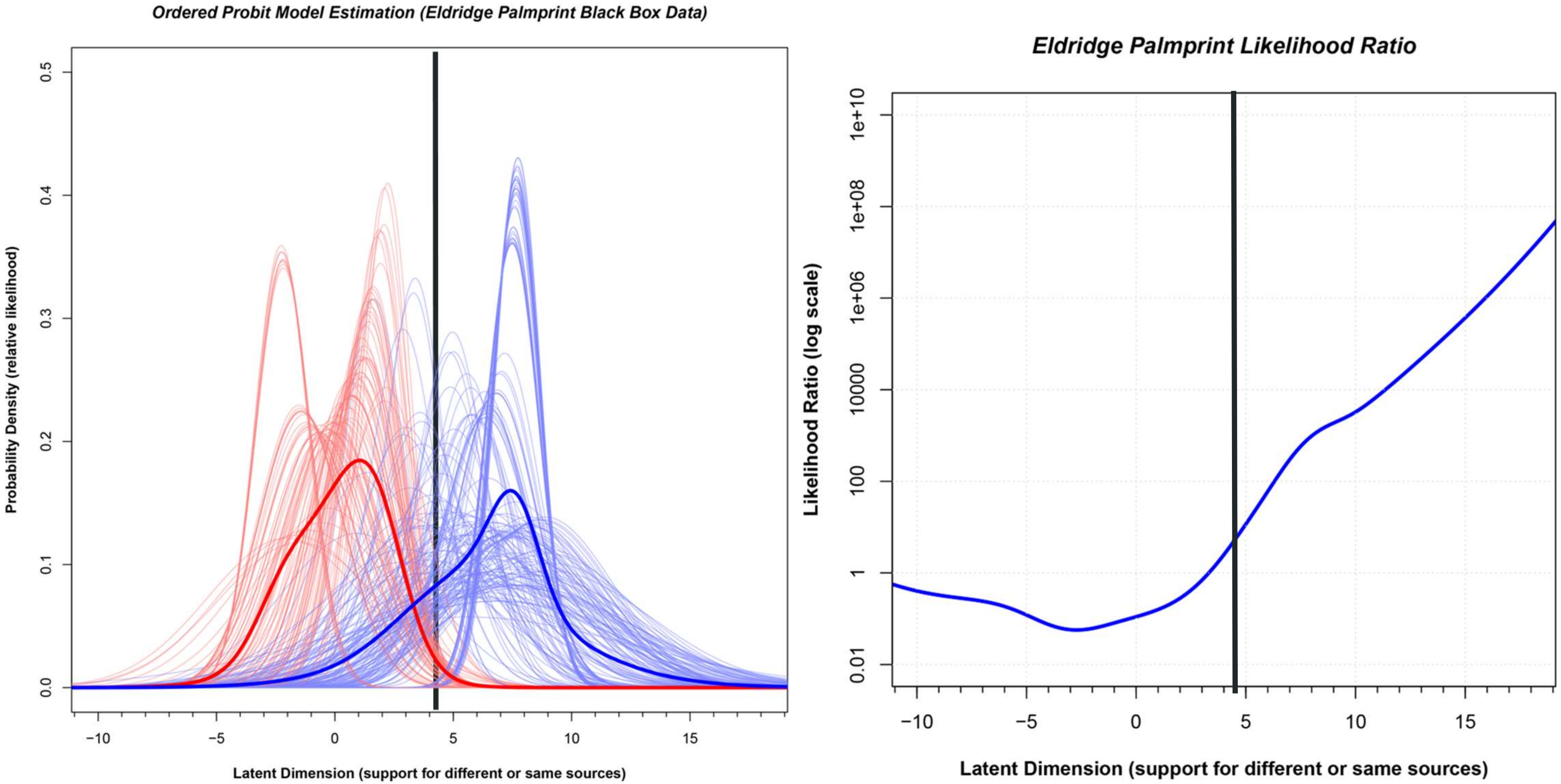


The right side of the blue box indicates the upper limit of unanimous decisions (prints where everyone said ID. Couldn’t go higher than 10.)

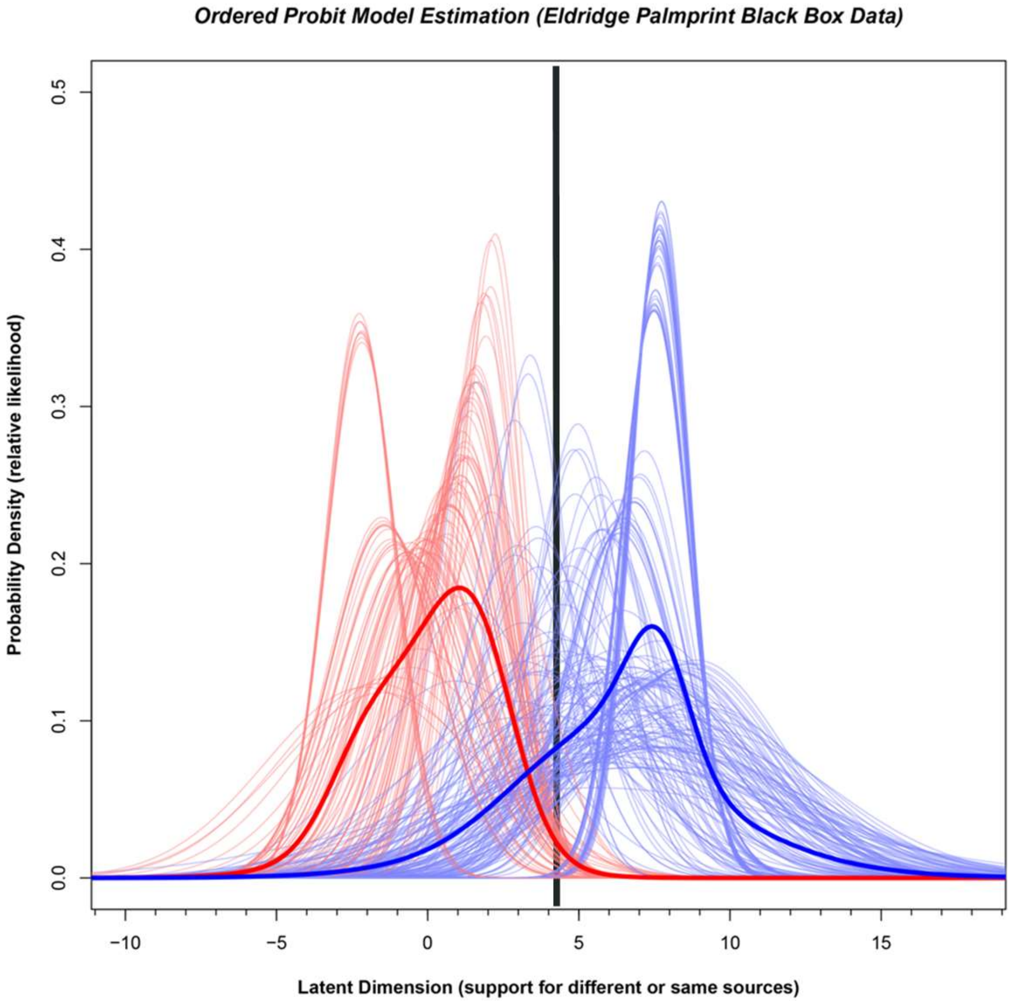

Ordered Probit Model Estimation (Eldridge Palmprint Black Box Data)


Eldridge Palmprint Likelihood Ratio

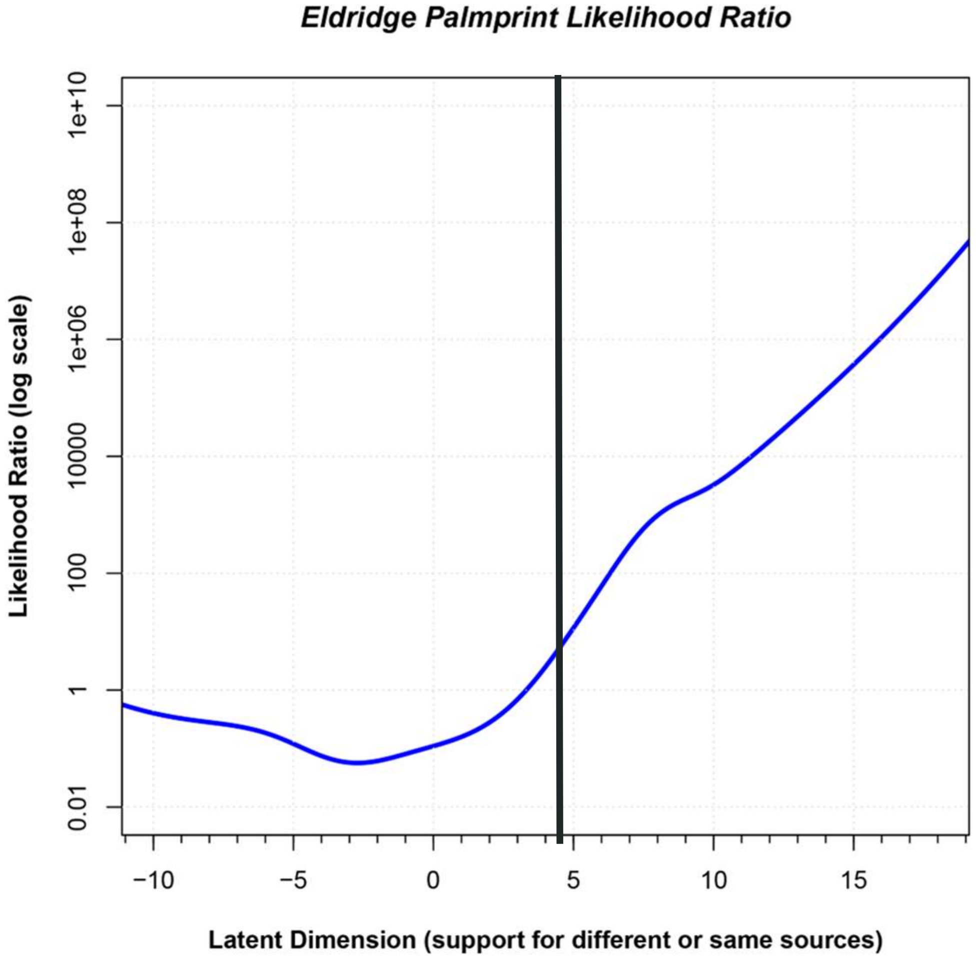
Even for samples which received a majority ID, the amount of support for the same source proposition can vary dramatically (5 – 1982).



The samples which contain a majority of Identification decisions could be reported by a laboratory as "Identification" but vary in the strength of support for the same source proposition.



Ordered Probit Model Estimation (Eldridge Palmprint Black Box Data)



Eldridge Palmprint Likelihood Ratio

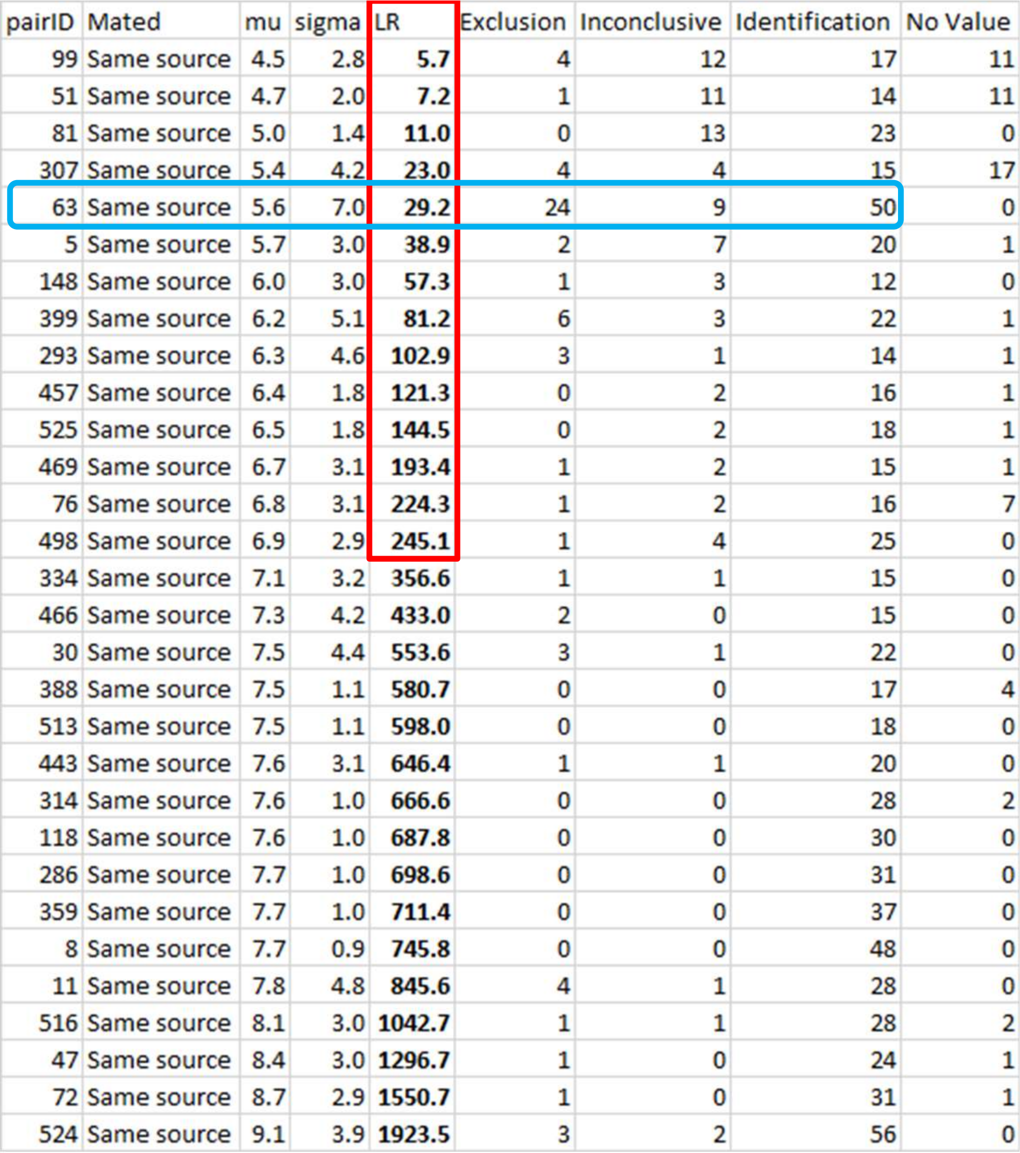
The Likelihood Ratio value at the majority ID boundary is **5.7**. Which means that reported Identifications might have a likelihood ratio as low as 5.7.

**5.7** times more support for the same sources than for different sources.

# Likelihood Ratios

The values in this table which have a majority of Identification decisions have been bolded.

The unanimous samples seem irrelevant to calculate a LR since they have so much support for the same sources proposition. The samples we are most interested in are those with the most potential to mislead the jury.



| pairID | Mated | mu | sigma | LR | Exclusion | Inconclusive | Identification | No Value |
|---|---|---|---|---|---|---|---|---|
| 99 | Same source | 4.5 | 2.8 | **5.7** | 4 | 12 | 17 | 11 |
| 51 | Same source | 4.7 | 2.0 | **7.2** | 1 | 11 | 14 | 11 |
| 81 | Same source | 5.0 | 1.4 | **11.0** | 0 | 13 | 23 | 0 |
| 307 | Same source | 5.4 | 4.2 | **23.0** | 4 | 4 | 15 | 17 |
| 63 | Same source | 5.6 | 7.0 | **29.2** | 24 | 9 | 50 | 0 |
| 5 | Same source | 5.7 | 3.0 | **38.9** | 2 | 7 | 20 | 1 |
| 148 | Same source | 6.0 | 3.0 | **57.3** | 1 | 3 | 12 | 0 |
| 399 | Same source | 6.2 | 5.1 | **81.2** | 6 | 3 | 22 | 1 |
| 293 | Same source | 6.3 | 4.6 | **102.9** | 3 | 1 | 14 | 1 |
| 457 | Same source | 6.4 | 1.8 | **121.3** | 0 | 2 | 16 | 1 |
| 525 | Same source | 6.5 | 1.8 | **144.5** | 0 | 2 | 18 | 1 |
| 469 | Same source | 6.7 | 3.1 | **193.4** | 1 | 2 | 15 | 1 |
| 76 | Same source | 6.8 | 3.1 | **224.3** | 1 | 2 | 16 | 7 |
| 498 | Same source | 6.9 | 2.9 | **245.1** | 1 | 4 | 25 | 0 |
| 334 | Same source | 7.1 | 3.2 | **356.6** | 1 | 1 | 15 | 0 |
| 466 | Same source | 7.3 | 4.2 | **433.0** | 2 | 0 | 15 | 0 |
| 30 | Same source | 7.5 | 4.4 | **553.6** | 3 | 1 | 22 | 0 |
| 388 | Same source | 7.5 | 1.1 | **580.7** | 0 | 0 | 17 | 4 |
| 513 | Same source | 7.5 | 1.1 | **598.0** | 0 | 0 | 18 | 0 |
| 443 | Same source | 7.6 | 3.1 | **646.4** | 1 | 1 | 20 | 0 |
| 314 | Same source | 7.6 | 1.0 | **666.6** | 0 | 0 | 28 | 2 |
| 118 | Same source | 7.6 | 1.0 | **687.8** | 0 | 0 | 30 | 0 |
| 286 | Same source | 7.7 | 1.0 | **698.6** | 0 | 0 | 31 | 0 |
| 359 | Same source | 7.7 | 1.0 | **711.4** | 0 | 0 | 37 | 0 |
| 8 | Same source | 7.7 | 0.9 | **745.8** | 0 | 0 | 48 | 0 |
| 11 | Same source | 7.8 | 4.8 | **845.6** | 4 | 1 | 28 | 0 |
| 516 | Same source | 8.1 | 3.0 | **1042.7** | 1 | 1 | 28 | 2 |
| 47 | Same source | 8.4 | 3.0 | **1296.7** | 1 | 0 | 24 | 1 |
| 72 | Same source | 8.7 | 2.9 | **1550.7** | 1 | 0 | 31 | 1 |
| 524 | Same source | 9.1 | 3.9 | **1923.5** | 3 | 2 | 56 | 0 |

# Sample #53

Mated Pair:

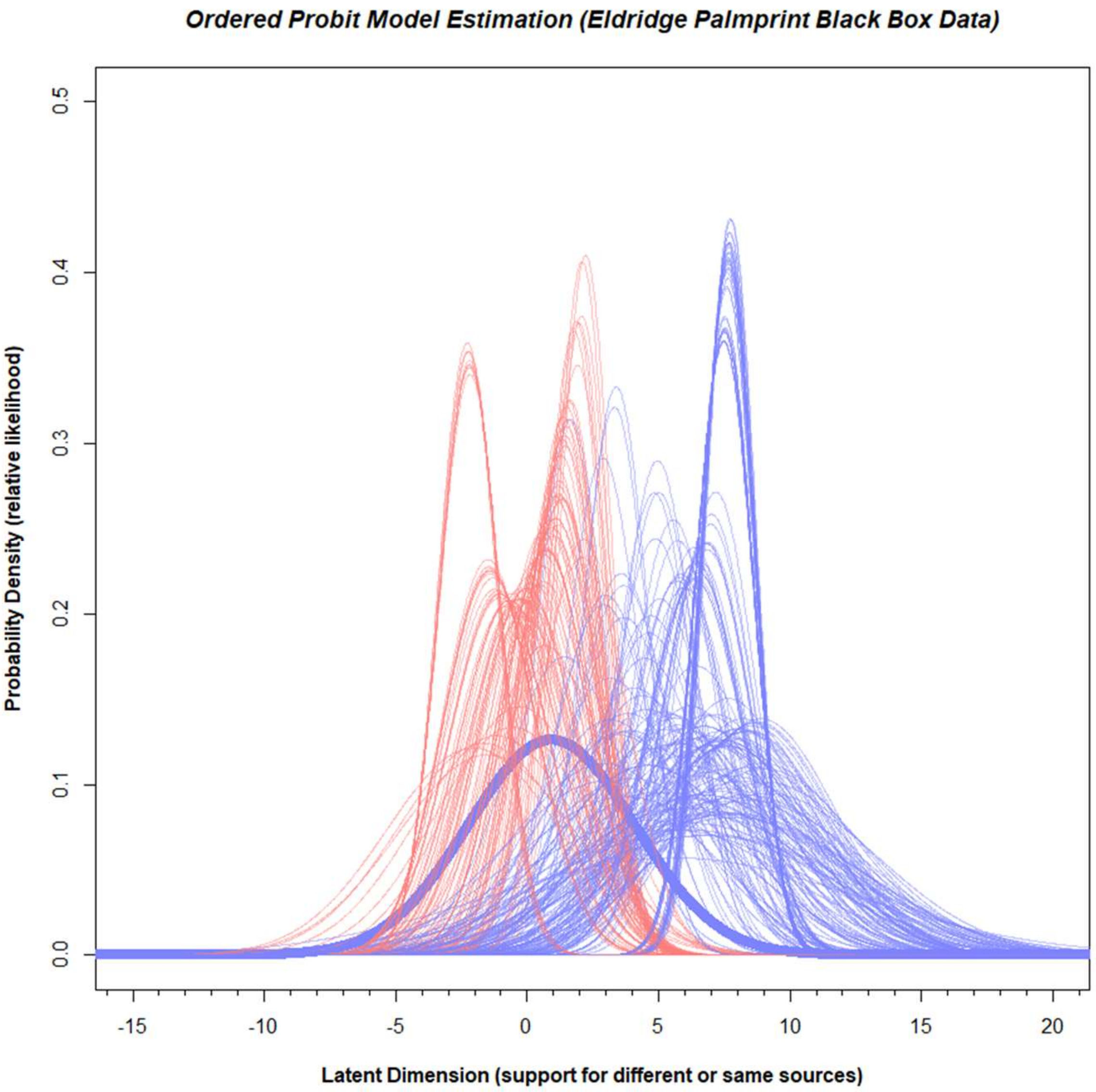
10 Exclusions

5 Inconclusives

2 Identifications

**LR of 0.15**

(has more support for different sources)



# Considerations

- More **difficult/less informative** comparisons can cause more inconclusive decisions, shifting the blue and red curves closer together.
- Examiners who erroneously exclude shift that sample's curve to the left. The more erroneous exclusions there are, the more the bold blue curve shifts to the left.
- The closer together the bold curves are (vertically), the lower the LR values.
- **Unanimous** decisions are hard to model. The model wants to push these data points to infinity (so we have to limit them somehow), so a prior on the means is applied to the model.
- LRs on samples with less support for the same sources proposition can actually go **down** if we allow for the unanimous decisions to move far to the right due to the normalization of all curves (bold lines).

# Considerations

- We assume that the internal values reached by examiners are **normally distributed** (most things in the world do) as opposed to other models like a t-distribution.
- Were the images used in these studies **casework quality**? The participants in this study indicated the prints they received were casework quality.
- The model is still **subjective**, not automated and not based entirely on minutiae. Ultimately it would be up to the examiner to calculate a LR for a casework print.

# Likelihood Ratio Ranges

For image pairs where the **majority** decision was ID (at least half of examiners said ID):

Fingerprint Samples:

Busey Study (2022) -  Likelihood ratios were from 20 to 100,000.

Black Box (Ulery 2011) - Likelihood ratios were from 50 to 20,000

Palmprint Samples

Eldridge Study - Likelihood ratios were from 5 to 2,000

# Likelihood Ratio Ranges

**A word of caution**: is the print you're "ID"ing a LR of **10**?

If there is only 10 times more support for the same source proposition than the difference sources proposition, is this consistent with the term Identification?

A LR of **20** is equivalent to a person in the population having green eyes and brown hair.

If you were a juror: Would you convict someone if the **only** information you had was "green eyes and brown hair"?

Latent examination **shouldn't be the only factor** in a trial. We can't control how much weight a jury gives our evidence when we say "ID". But we are doing them a disservice by not appropriately weighting and explaining to the jury how much weight they **should** be giving our evidence.

# Limitations

**This is a theoretical model!**

How do we apply it to casework? We would need:

- More research/Validation studies
- Conversations about usage (LR alone? With categorical decisions? Verifiers using it too? Range of LRs? What about conflicts?...)
- Training for examiners
- Jury interpretation studies (in progress)
- Investigator/Attorney interpretations
- Proficiency Testing
- Methods to apply the model every day

# Benefits of LR

- No erroneous IDs (but misleading LRs are still possible)
- Increases ability to communicate nuance to peers and jury
- A LR could be generated to any number of subjects. (Both Mayfield and Daoud could have been assigned a LR, and both would likely have been very low because the latent is terrible, low strength of support)
- Becomes a “multiplying” factor for the jury. Testimony could assist with information they’ve already heard but not stand on its own. (Other evidence pointed to defendant, latents are just one more factor. Or person we associated has been dead for 20 years and therefore the prior odds are 0 to begin with. Even 100,000 times 0 is 0!)

# Future Directions

I’m working with Dr Busey on an NIJ grant studying how examiners feel about ranking comparisons!

- Which ones have more support and which ones have less?
- Should we use numbers instead of words? More confusing to the consumer?

If you want to assist in this development, you can be one of our guinea pigs!

# Questions?
# Comments?
# Tomatoes?

If you would like to participate in future research!