Calibrated Strength of Evidence Statements For Casework

Thomas Busey, PhD
Professor, Indiana University Bloomington
Meredith Coon, MS, CLPE

Strength of Evidence

When a friction ridge examiners says "Identification", how should fact finders update their beliefs about the facts of the case?

Quantitative measurements of fingerprint evidence by Neuman et al (ref) gave estimates of the strength at 10^5 to 10^{10} (100,000 to 10 billion).

These numbers are at odds with error rate studies, which put the strength of the evidence in the 100s or 1000s.

Strength of Evidence

What does “Strength of Evidence” mean?

The ability to distinguish between two alternative propositions or states of the world.

One proposition: the latent print and the exemplar print from AFIS came from the same finger.

Second proposition: the latent print and the exemplar print come from different fingers.

Support for Propositions

The fact finder wants to know which of these two propositions is correct.

Forensic scientists can only provide the *relative support* for each of the two propositions.

“The fingerprint evidence provides 1000 times more support for the prosecution proposition than the defense proposition.”

The fact finder will use this information along with other facts in the case to update their beliefs about the defendant.

Measuring Relative Support

When examiners say “Identification” how do we translate that word into a belief change?

Various ways researchers have approached this:

- 71% of laypeople think it means “to the exclusion of all others” (Swofford & Cino, 2017)
- PCAST- Measure error rates.
- FRStat- use assumptions about minutia constellation rarity to compute statistics. (Swofford et al, 2018)
- ENFSI- express a subjective likelihood ratio

  (Champod et al, 2016)

Challenges with “Identification”

The scientific consensus suggests moving away from definitive conclusions.

- Require knowing the prior probability of a mated pair.
- Decisions depend on an internal threshold that can be affected by things like expertise and personality.
- The cost of the two kinds of errors can affect the decisions.
- Subsumes the role of fact finder.
- Evaluative reporting has challenges as well.

Challenges with Error Rates

A system-wide error rate is a good starting point to know whether a discipline falls into “junk science” status.

However, most errors are the result of a few types of evidence (black powder on galvanized metal) or a few examiners (See FBI/Noblis Black Box 2).

Error rate studies are difficult to apply to individual examiners.

“But I’ve never made an error.” How would they know without knowing ground truth in casework?

Challenges/Opportunities with Quantitative Measures

Quantitative measures such as FRStat make important contributions.

Depend on the assumptions of the models, especially the creation of close non-mates.

Only reflect a subset of the available information (typically minutiae).

Values establish a lower bound that if high enough provide a good statistical foundation.

No reason not to use if you have access to validated tools.

Our approach: BASE-LR

Rely on human examiners to compute the strength of the evidence on a sample of images with known ground truth.

Provide the strength of the evidence for these images.

Develop and validate an interface that extends this strength of evidence to operational casework.

Free and open source. Self-validating and does not require exposure of casework materials outside the lab.

Our approach: BASE-LR

Stands for Benchmark Anchored Sorting Estimation Likelihood Ratios.

https://buseylab.org/BASE-LR/

Quick demo for context, then describe the validation of the tool.

Step 1: Measuring Strength of Evidence

With any comparison there are two ground truth states: mated and nonmated.

In casework we typically don't know ground truth, but we do in black box (error rate) studies.

Step 1: Measuring Strength of Evidence

You might say “Identification” to a clear latent with 17 minutia, a clear starting target group, and level three detail.

We would say that the support for the same-source (mated) proposition is **much greater** than the support for the different-source proposition.

Step 1: Measuring Strength of Evidence

You might say “Identification” to a really sketchy borderline latent with tonal inversion, distortion and no clear core.

We would say that the support for the same-source (mated) proposition is **somewhat greater** than the support for the different-source proposition.

Step 1: Measuring Strength of Evidence

In both cases the same word is used, but the strength of the evidence is likely higher in the first case.

Examiners likely set their decision threshold so high that the strength of the evidence is so high that the differences are negligible.

This is an empirical question that we will dispute.

Converting Words to Numbers

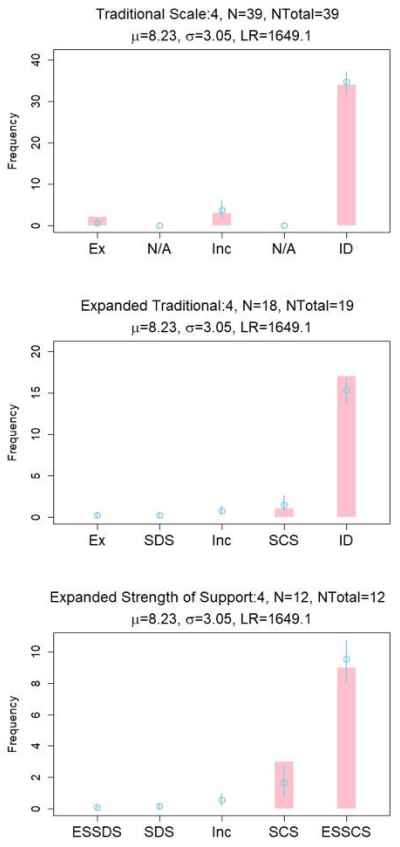
In the second case, not all examiners might reach “Identification”.

Some might say “Inconclusive” or even “Exclusion”.

In black box studies we can collect the distribution of the number of examiners who reach each conclusion.

That becomes the estimate of the strength of that comparison.

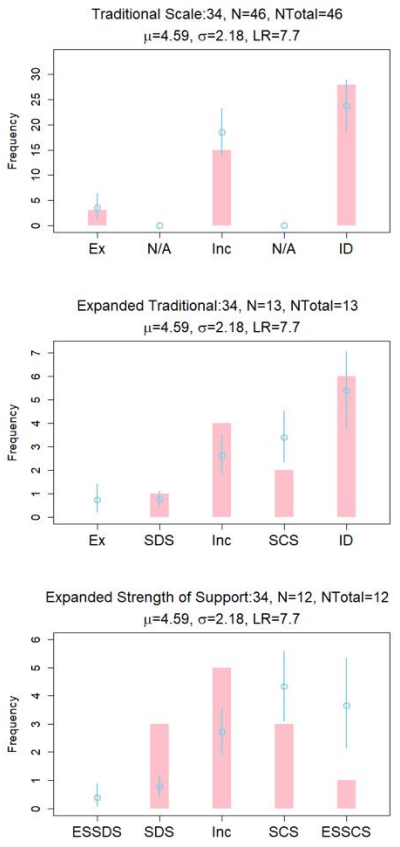
Examiner response distribution: Mated pair



34 Identification decisions
3 Inconclusive decisions
2 Exclusion decisions

Examiner response distribution: Mated pair

Pair 34 Rank 47 Mu = 4.59 [3.98 to 5.27] Sigma = 2.18 LR=7.7 Mated



28 Identification decisions
15 Inconclusive decisions
3 Exclusion decisions

How do we construct likelihood ratios?

First, make observations (e.g. the amount of perceived detail in agreement).

Next, consider the relative likelihood of the observations under two *propositions* (or alternative universes)*:*

1) The person of interest made the impression at the crime scene (i.e. the two images are **mated**).
2) Some unknown individual made the impression at the crime scene (i.e. the two images **nonmated**).

How do we construct likelihood ratios?

The likelihood ratio is the ratio of two conditional probabilities:

$$LR = \frac{p(observations|the\ two\ prints\ are\ mated)}{p(observations|the\ two\ prints\ are\ nonmated)}$$

Where:

$p(observations|the\ two\ prints\ are\ mated)$ represents how likely you would observe this amount of perceived detail in agreement given the two prints came from the same finger.

$p(observations|the\ two\ prints\ are\ nonmated)$ represents how likely you would observe this amount of perceived detail in agreement given the two prints came from different fingers.

If you don’t like math

The posterior odds represent *how much more guilty than innocent the person appears after you learn the likelihood ratio, as updated from how guilty the person was originally.*

Example:

Imagine you are on a jury. Based on an eyewitness, you feel that the defendant is twice as likely to be the criminal than some other person.

A likelihood ratio from forensics is presented, with a value of 1000.

You should interpret this as meaning that the defendant is 2 * 1000 = 2000 times more likely to be the criminal than some other person.

Where do these likelihood ratios come from?

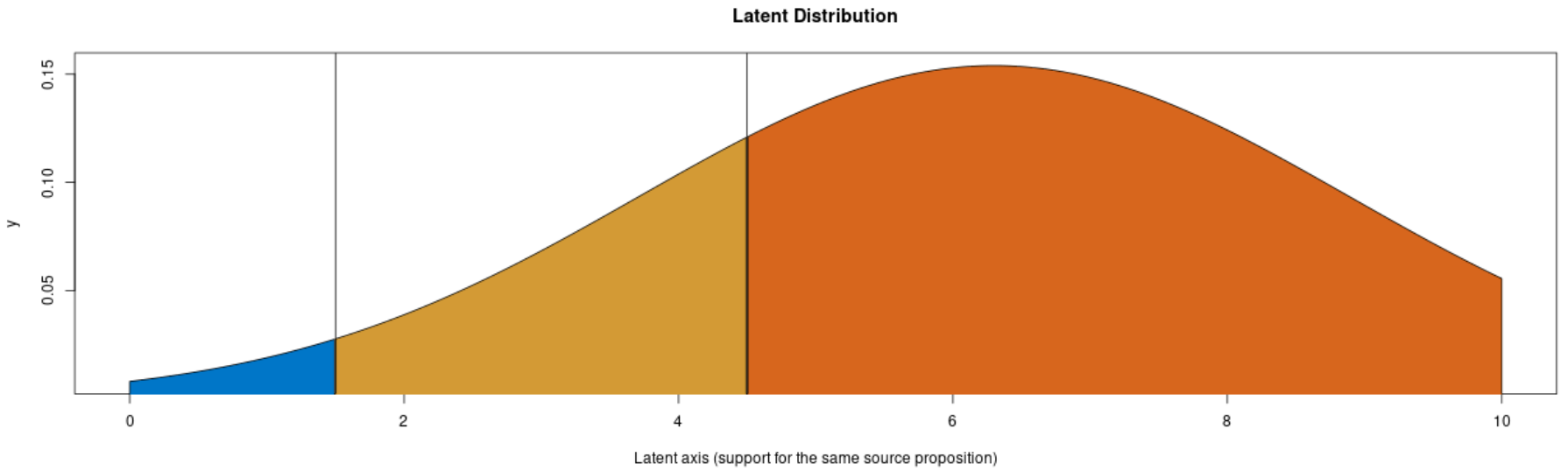
Converting examiner response distributions

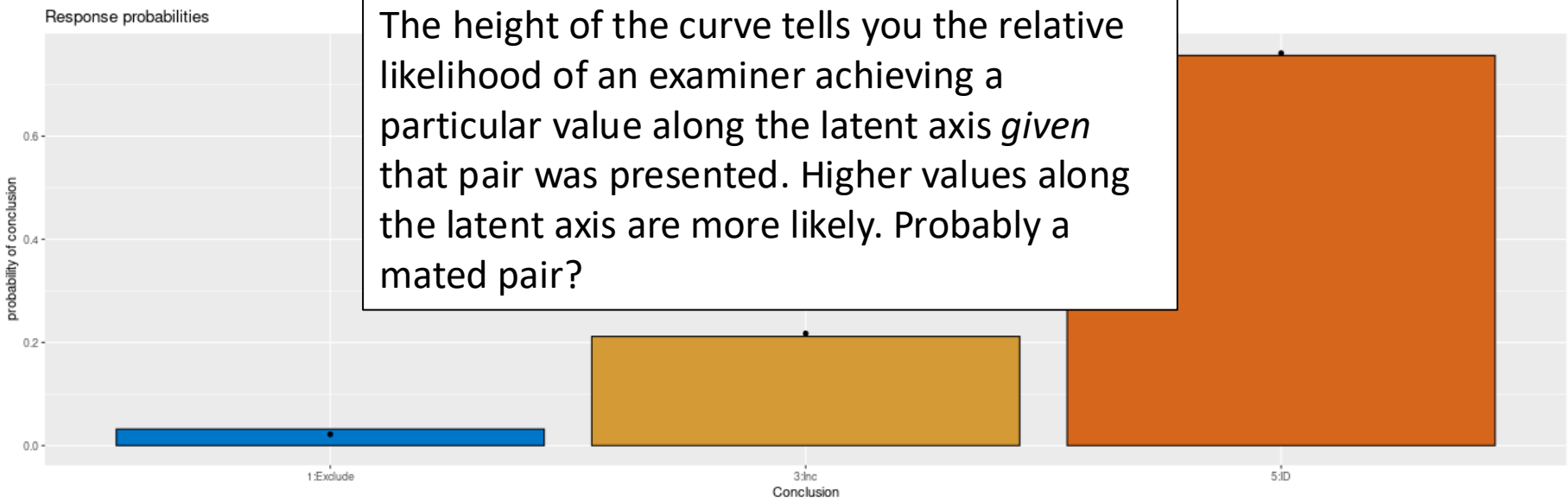
- While most examiners made an Identification for both of the image pairs, the *number* of examiners making Identification conclusions differed between comparisons.
- The support for the common source proposition **differs** across image pairs.
- Not all Identifications are equal.
- We can summarize the collections of responses using something called the Ordered Probit Model.

The Ordered Probit Model

- There is something not directly visible from the data in this study: how much support each examiner believes a particular comparison provides for the same source proposition. We call this the latent axis (not the same as a latent print).
- Imagine a magical electrode that determined how much support the examiner mentally determined for the same source proposition. Each examiner's brain would generate a number, a value along the latent axis.
- If we repeat this process with 20-30 examiners, we end up with a *distribution* of values along this latent axis.
- We can approximate this latent distribution with a normal (Gaussian) distribution.
- Two thresholds are applied to this latent distribution to produce the three conclusions (exclusion, inconclusive, or identification).
- That will help us capture the overall strength of evidence provided by an image pair.

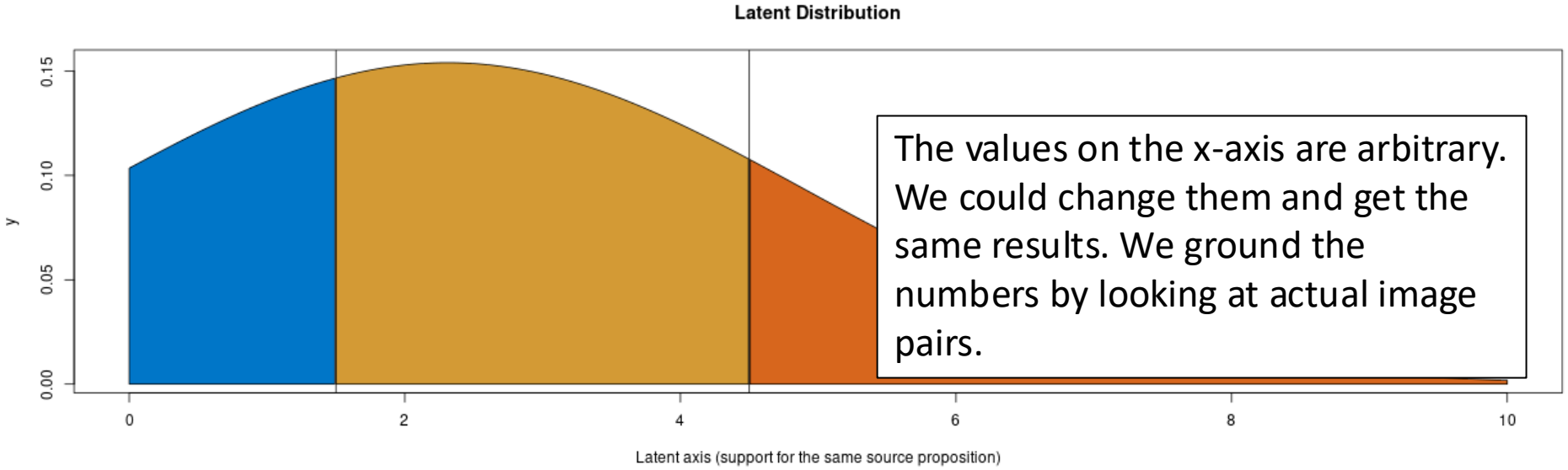
What does the normal curve represent?



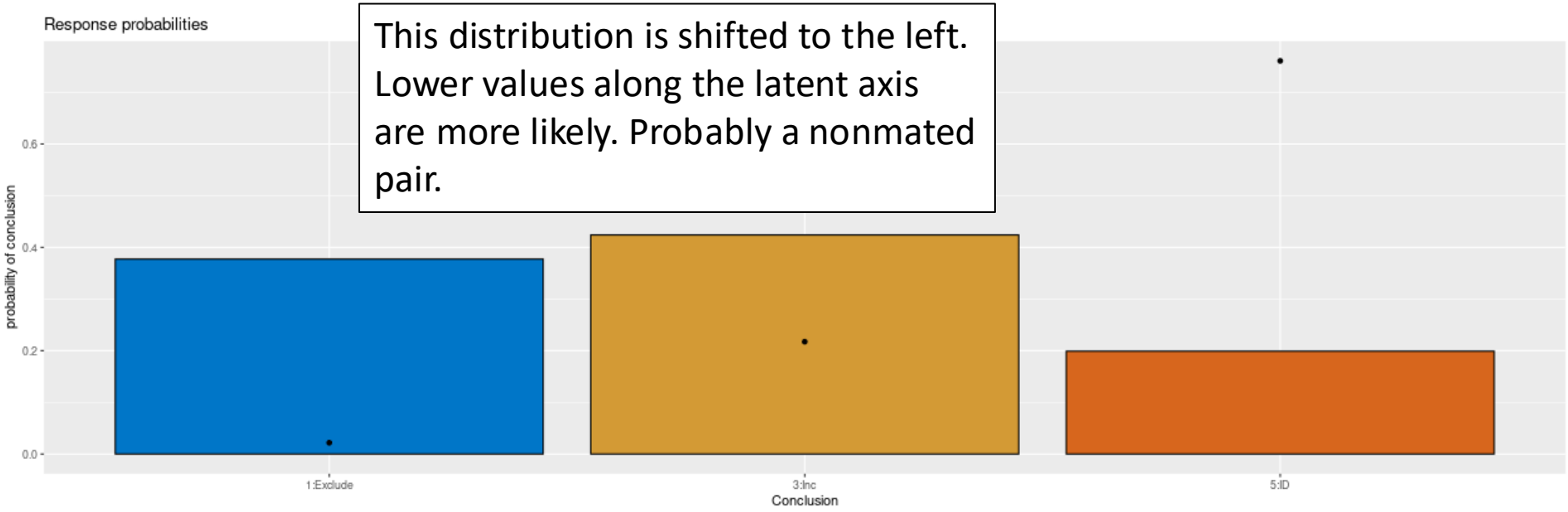


The height of the curve tells you the relative likelihood of an examiner achieving a particular value along the latent axis *given* that pair was presented. Higher values along the latent axis are more likely. Probably a mated pair?

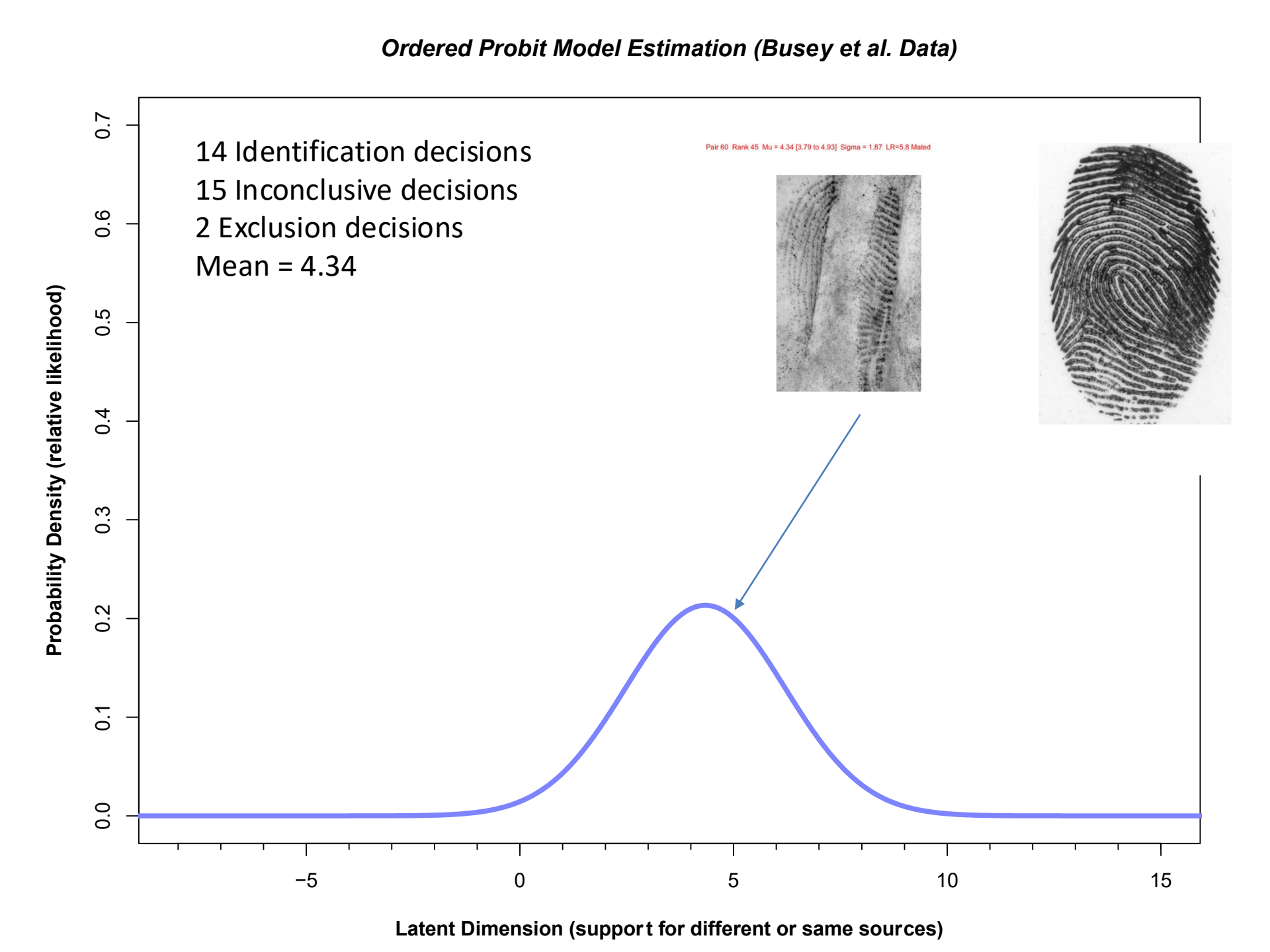
What does the normal curve mean?

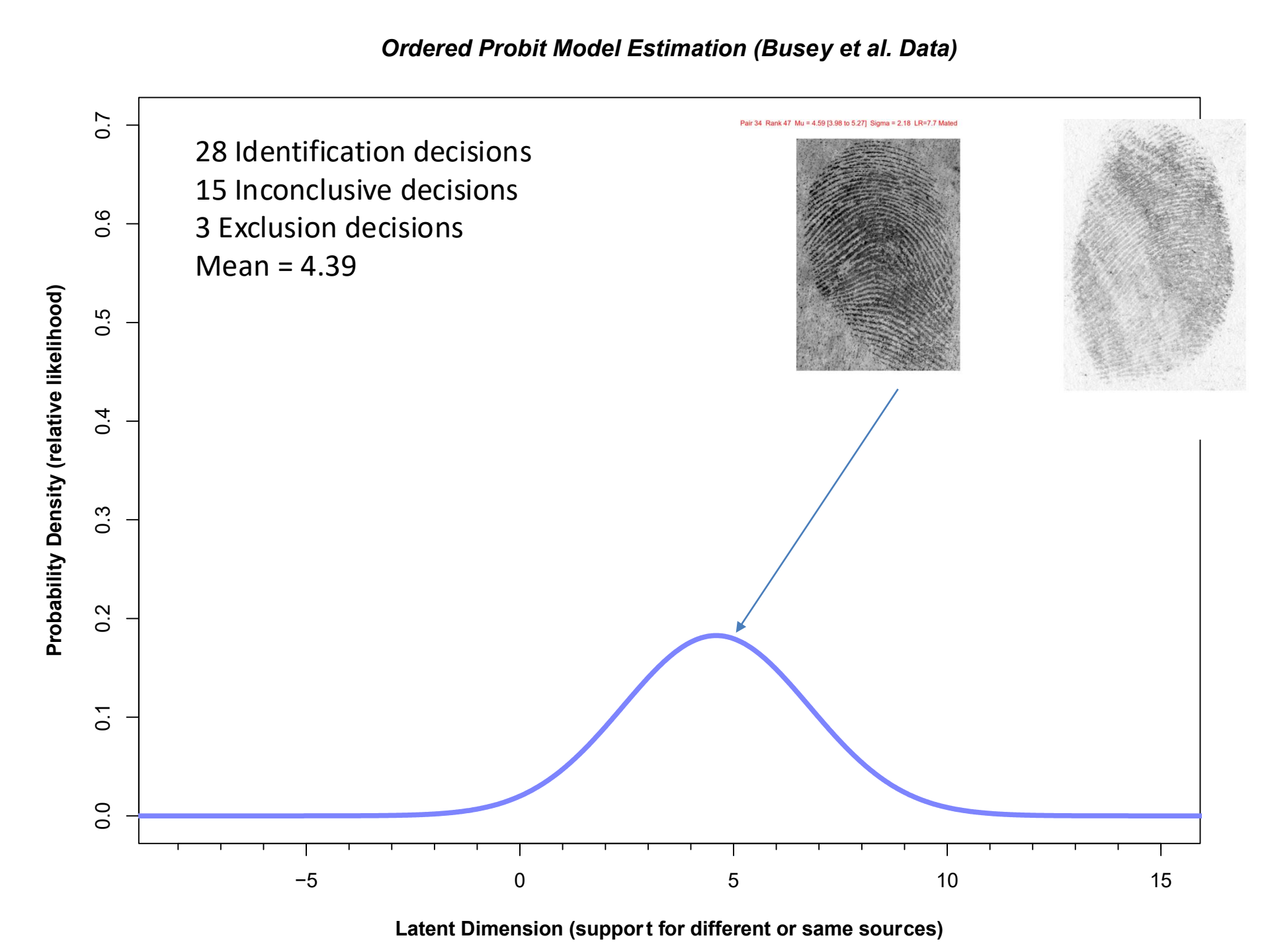


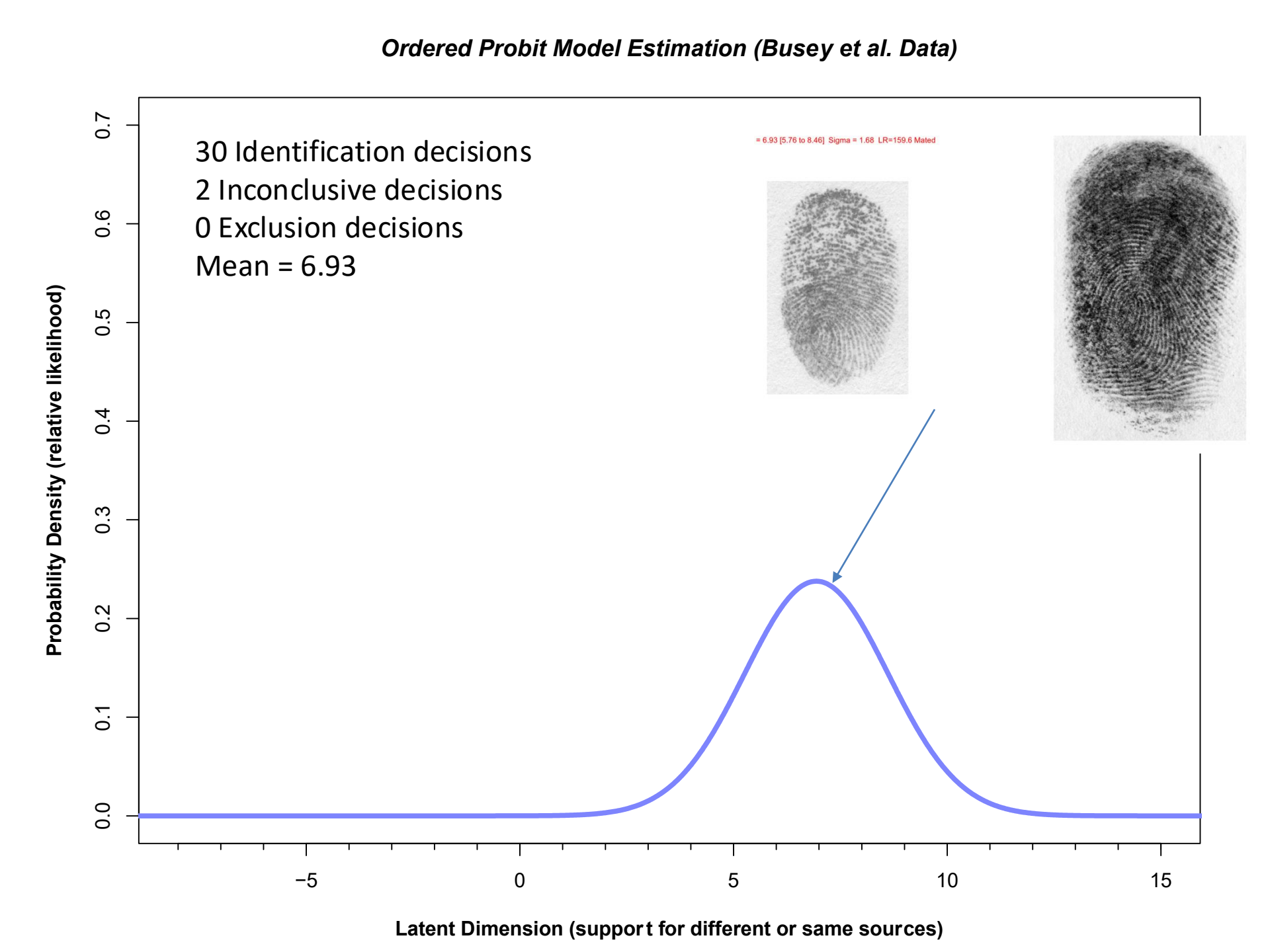
The values on the x-axis are arbitrary. We could change them and get the same results. We ground the numbers by looking at actual image pairs.



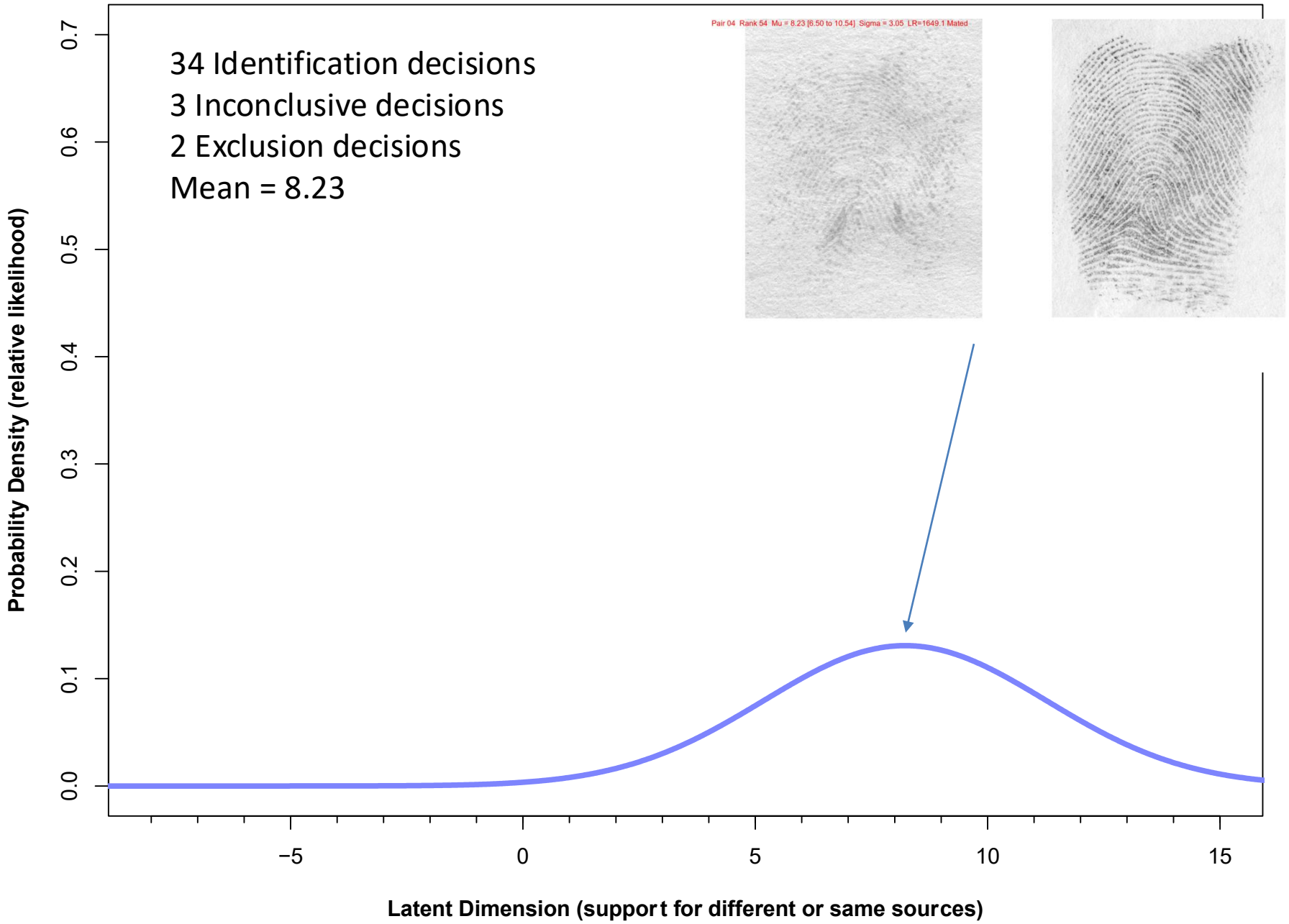
This distribution is shifted to the left. Lower values along the latent axis are more likely. Probably a nonmated pair.


Ordered Probit Model Estimation (Busey et al. Data)
14 Identification decisions
15 Inconclusive decisions
2 Exclusion decisions
Mean = 4.34
Pair 60 Rank 45 Mu = 4.34 [3.79 to 4.93] Sigma = 1.87 LR=5.8 Mated
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)


Ordered Probit Model Estimation (Busey et al. Data)
28 Identification decisions
15 Inconclusive decisions
3 Exclusion decisions
Mean = 4.39
Pair 34 Rank 47 Mu = 4.59 [3.98 to 5.27] Sigma = 2.18 LR=7.7 Mated
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)


Ordered Probit Model Estimation (Busey et al. Data)
30 Identification decisions
2 Inconclusive decisions
0 Exclusion decisions
Mean = 6.93
= 6.93 [5.76 to 8.46] Sigma = 1.68 LR=159.6 Mated
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)

Ordered Probit Model Estimation (Busey et al. Data)



Ordered Probit Model Estimation (Busey et al. Data)

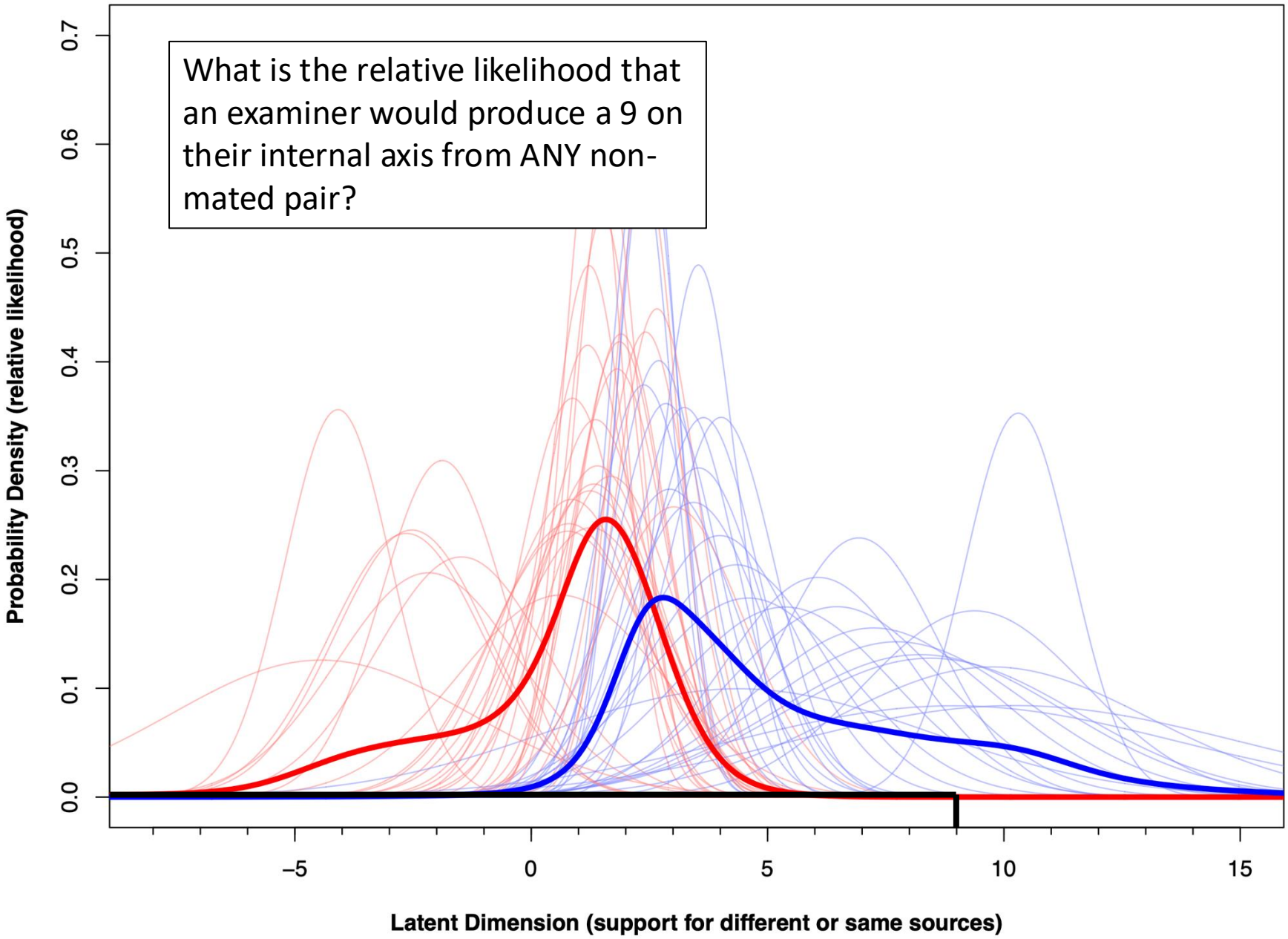
37 Identification decisions
0 Inconclusive decisions
0 Exclusion decisions
Mean = 10.3

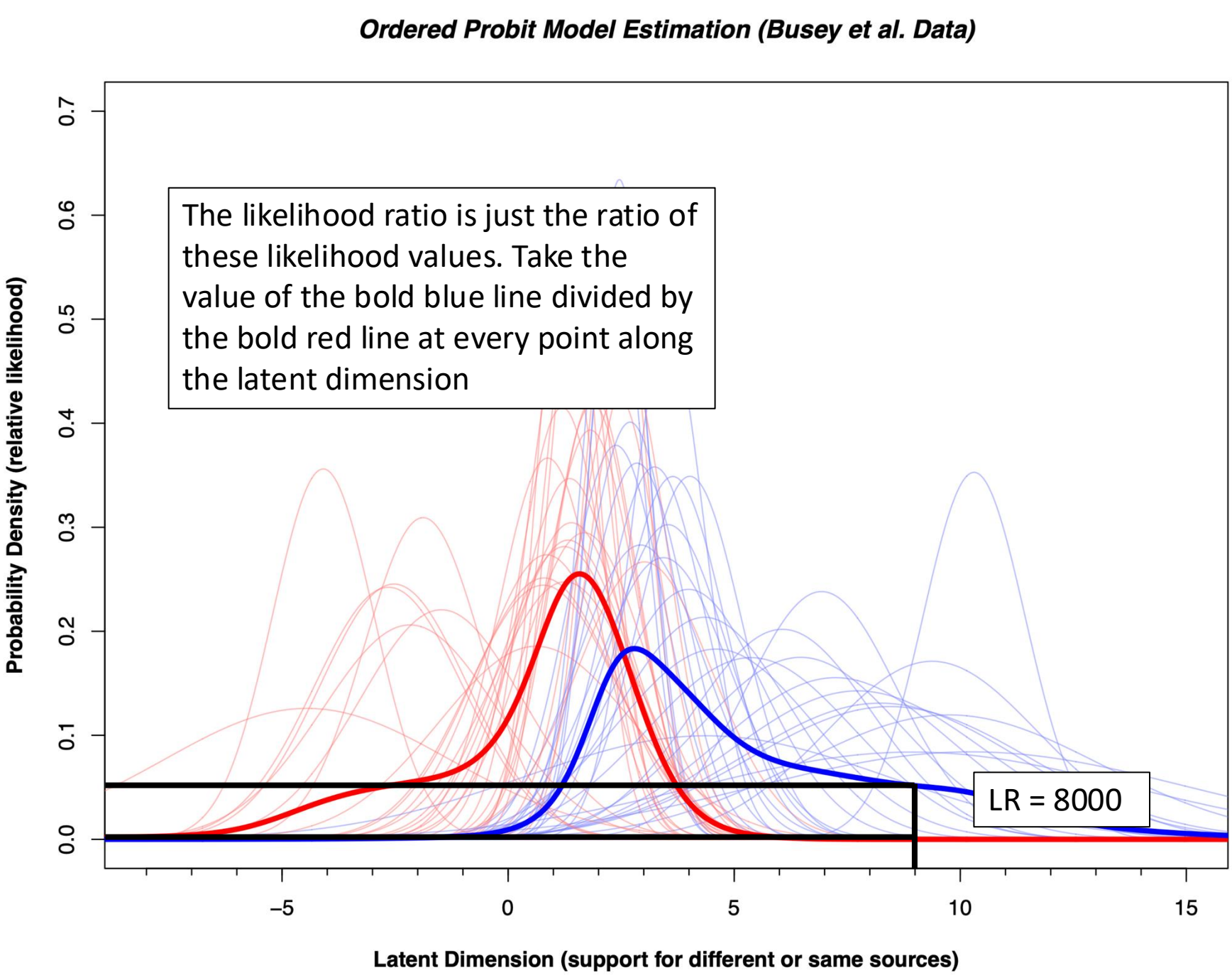
Probability Density (relative likelihood)

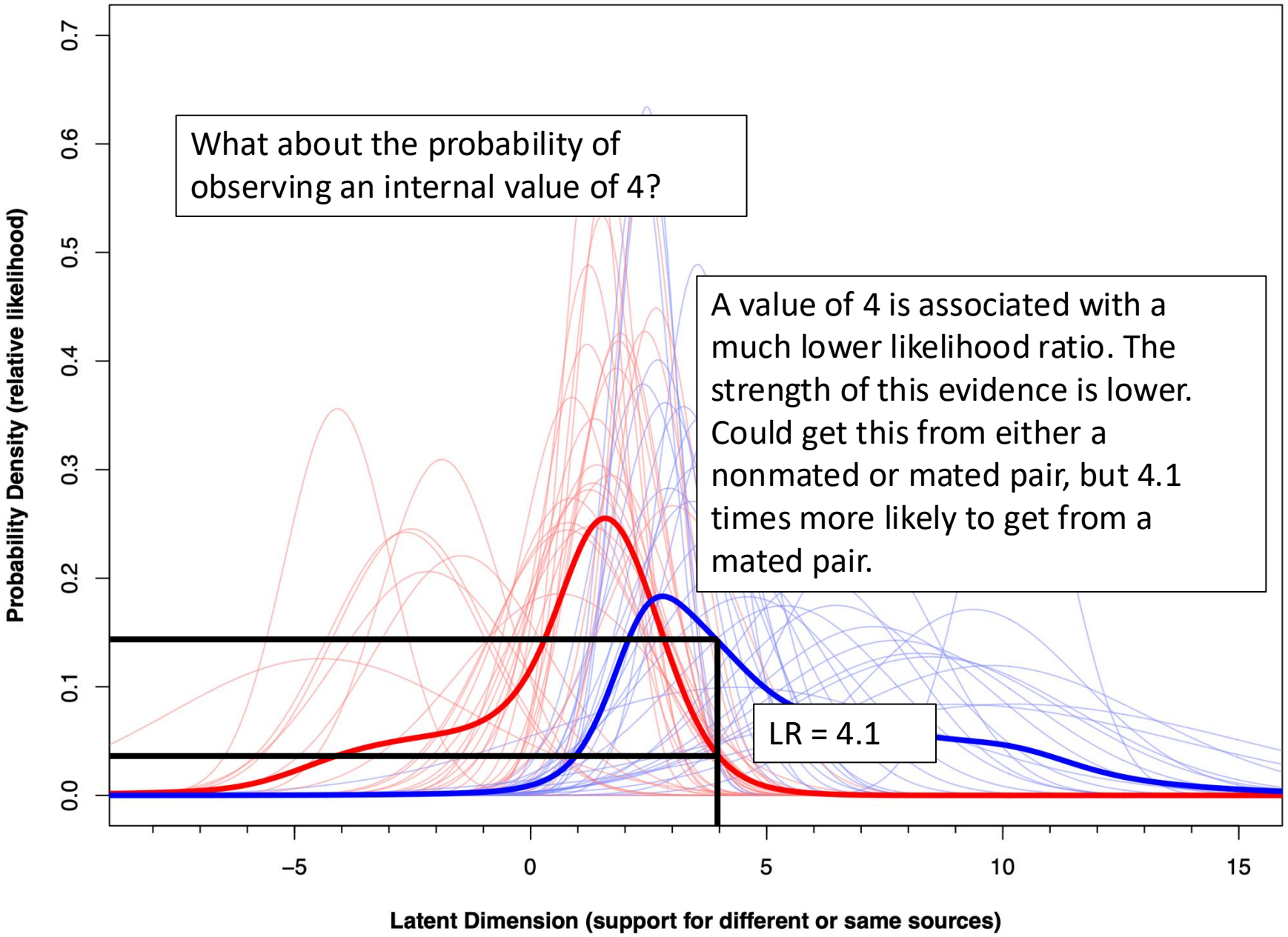
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7

−5 0 5 10 15

Latent Dimension (support for different or same sources)


Ordered Probit Model Estimation (Busey et al. Data)
What is the relative likelihood that an examiner would produce a 9 on their internal axis from ANY non-mated pair?
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)


Ordered Probit Model Estimation (Busey et al. Data)
The likelihood ratio is just the ratio of these likelihood values. Take the value of the bold blue line divided by the bold red line at every point along the latent dimension
LR = 8000
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)


Ordered Probit Model Estimation (Busey et al. Data)
What about the probability of observing an internal value of 4?
A value of 4 is associated with a much lower likelihood ratio. The strength of this evidence is lower. Could get this from either a nonmated or mated pair, but 4.1 times more likely to get from a mated pair.
LR = 4.1
Probability Density (relative likelihood)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−5
0
5
10
15
Latent Dimension (support for different or same sources)

Step 1: Measuring Strength of Evidence

That was a lot.

TL;DR- For each image pair in the database we know the likelihood ratio.

This represents the strength of support for the same-source proposition relative to the different-sources proposition.

But- some majority-ID image pairs have likelihood ratios less than 100.

Maybe Identification doesn't mean 10^5 or 10^{10}. Maybe it means 100.

Step 2: Generalize this to Casework

- We collected response data on 180 image pairs from 104 examiners.
- We know the likelihood ratio for each image pair.
- But- casework impressions will not be part of this database. How do we compute likelihood ratios for image pairs not in the database?

Step 2: Generalize this to Casework

- Key step- use the same tools that created the initial decision- the eyes and brain of the examiner.
- Currently the best pattern recognition system (never count out AI…).
- Previously we asked examiners to compare two images.
- Now we will ask them to compare *comparisons.*

Step 2: Generalize this to Casework

- We first ask examiners to conduct 6 comparisons and remember how much support each image pair provided for the same source proposition (i.e. is mated).
- They then use a graphical interface to place each image pair such that the *horizontal location corresponds to the relative strength of the evidence. (Further to the right = more support for same source)*

Sorting Image Pairs by Strength of Support

Done with sorting

29:16 Pause

Validation

- To validate this process, we ask examiners to conduct an additional “casework” analysis and comparison.
- They placed this new image pair relative to the original six image pairs.

Sorting Image Pairs by Strength of Support

Done with sorting

29:16 Pause

Validation Study- Ask Examiners to Place a “Casework” Image Pair

Validation- Key Insight

- We know the strength of evidence for the “Casework” image pair because it was part of the black box study.
- We can see how consistent examiners are with the group when placing the “casework” image pair by comparing its horizontal location to the computed strength of support using linear regression.

Experimental Design

- 64 latent print examiners eligible to testify in the US.
- Had previously completed our part 1 black box study.
- Two conditions for image pair placement:

1. **Free sort-** Subjects could place the six benchmark image pairs themselves.
2. **Fixed sort-** The six benchmark image pairs were placed relative to their strength of evidence and could not be moved.

Free vs. Fixed Sort

- All examiners completed comparisons of all six benchmark image pairs.
- Free sort subjects did their own sort of the image pairs.
- Fixed sort subjects had the images locked in place.

Free vs. Fixed Sort

- Both groups compared and sorted five “casework” image pairs amongst the six benchmark images.
- In both cases the horizontal locations of the benchmark images was frozen for “casework” image pairs. (The Free-Sort group was stuck with their previous decisions of placement).
- This procedure was repeated with 6 new benchmark image pairs and 5 new “casework” image pairs to increase experimental power.

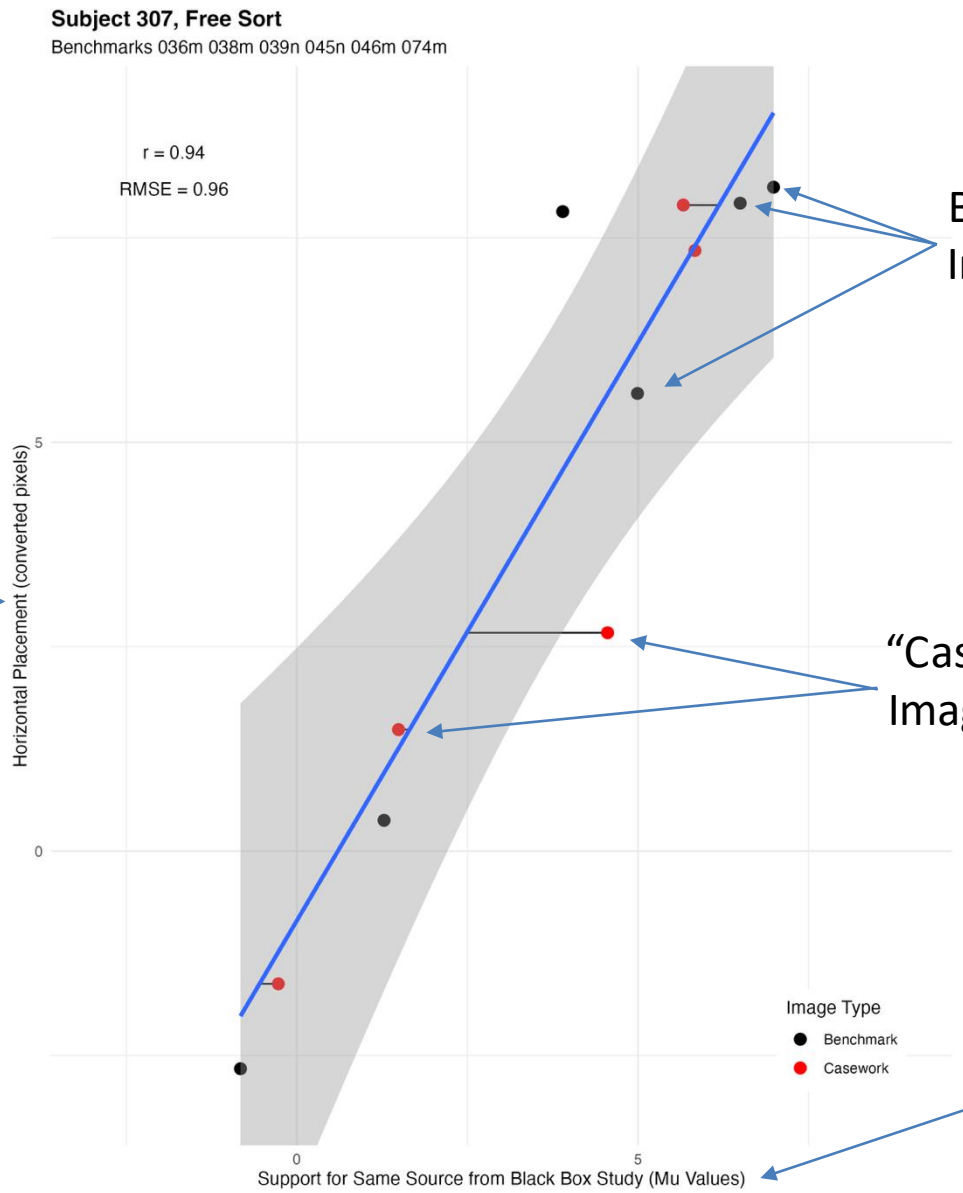
Free vs. Fixed Sort

- Which sort might be better?
- Fixed sort- there is no variation in placement of the benchmark image pairs. But- subjects must internalize the scale that the group consensus determined in the black box study as summarized with the ordered probit model.

Free vs. Fixed Sort

- Which sort might be better?
- Free sort- there is more variation in placement of the benchmark image pairs. But- subjects may spend more time internalizing the underlying scale as represented by the benchmark image pairs.
- We can compare the estimated strength of evidence to the actual strength of evidence for casework image pairs to see which sort is better.

How Good Are Subjects?



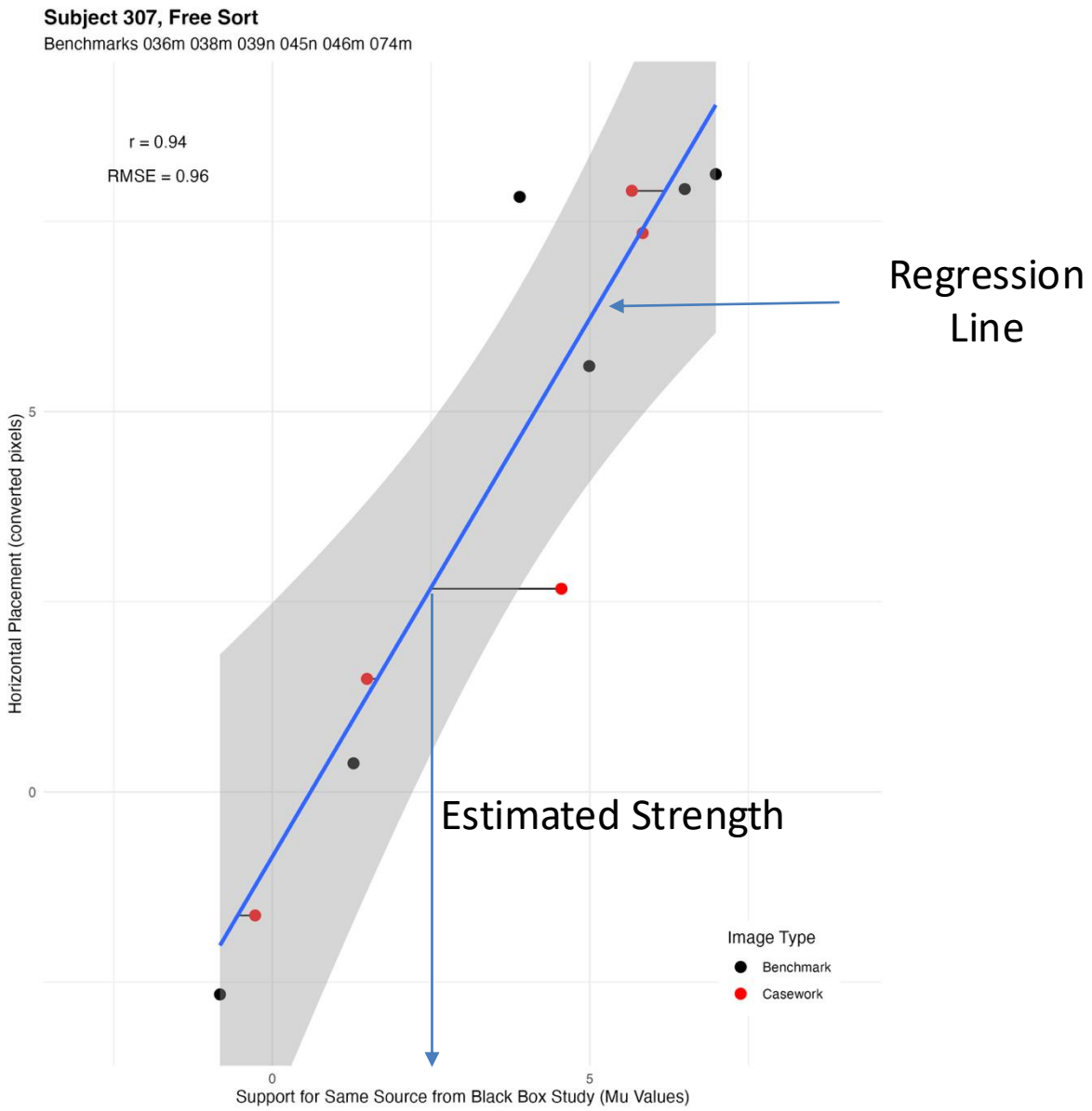
Horizontal Placement

Benchmark Image Pairs

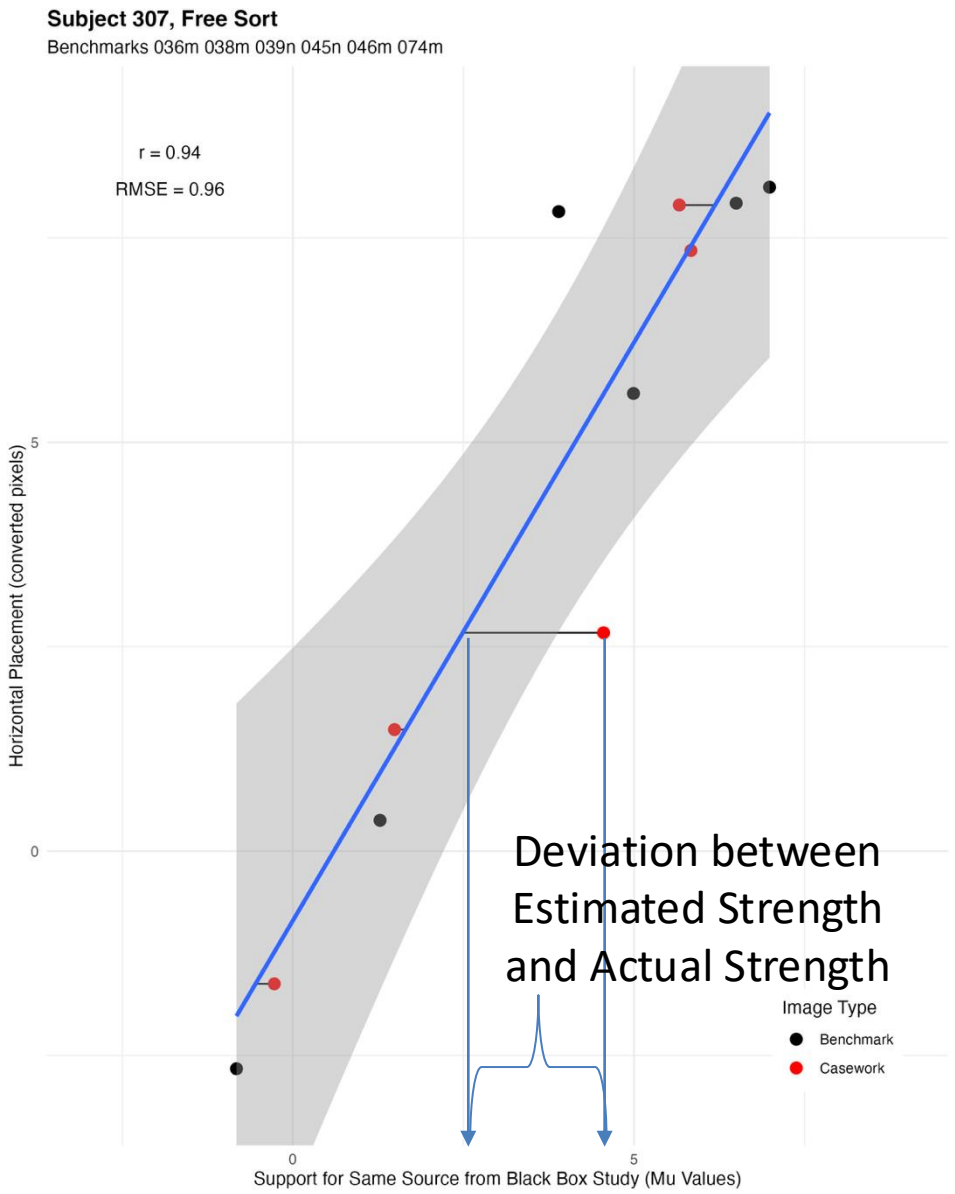
"Casework" Image Pairs

Strength of Evidence from Black Box Study

How Good Are Subjects?



How Good Are Subjects?



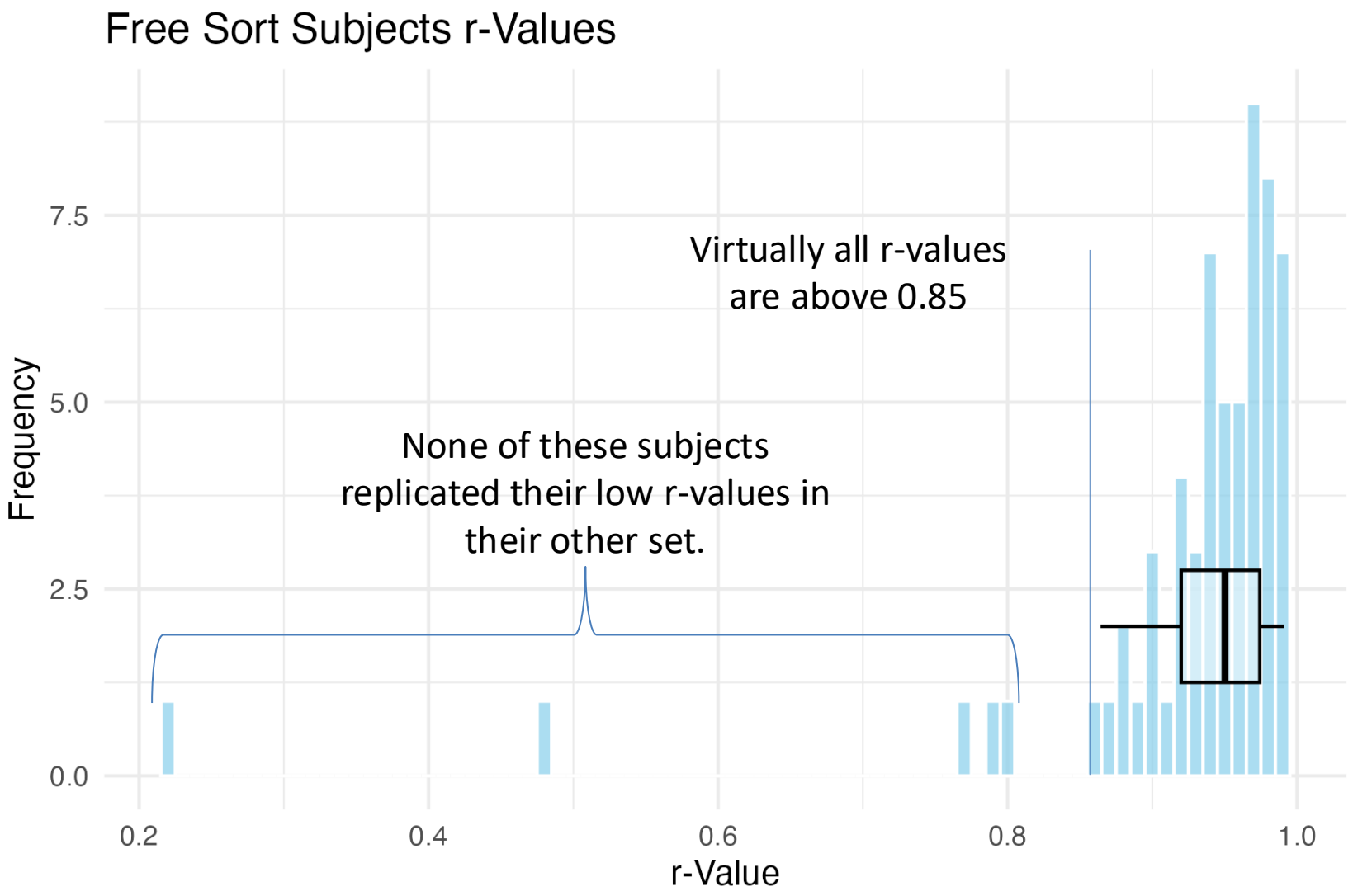
Free vs. Fixed Sort

- There is an additional advantage to the free sort condition:
- Because a subject's placement of the six benchmark image pairs will never exactly correspond with the values calculated from the ordered probit model, we can use this to prune bad subjects.

Validation

- By comparing the estimated strength of the casework image pair against the actual strength from the error rate study we can see how accurate our sorting method is.
- How accurate are subjects on average?
- Which sort might be more accurate?
- How strong are the regression coefficients for the benchmark image pairs in the free sort condition?

Free Sort Regression Values



Self-Validation

- The free sort condition is self-validating.
- If a user doesn’t agree with the group consensus for strength of evidence of the benchmarks, we simply don’t proceed.

Validation

We compute the difference between the estimated and actual strength of evidence (mu value) as a root mean squared error (RMSE).

Free sort has lower error values than the fixed sort

Sort Type	Image Set	RMSE
Free Sort	002m 003n 006m 008n 017m 058m	1.11
Free Sort	036m 038m 039n 045n 046m 074m	1.07
Fixed Sort	002m 003n 006m 008n 017m 058m	1.86
Fixed Sort	036m 038m 039n 045n 046m 074m	1.31

Free Sort Is Best

- The free sort condition is self-validating.
- The free sort condition produces lower error, possibly because users better internalize the underlying scale represented by the six benchmark image pairs.

What About Accuracy?

Give the RMSE values we see, how accurate are the computed strength of evidence values?

Lower mu Value	Higher mu Value	Difference of Log10(LR)
1	2	0.47
2		0.79
3		1.11
4	5	1.05
5	6	0.83

We are about 1 order of magnitude within the true value.

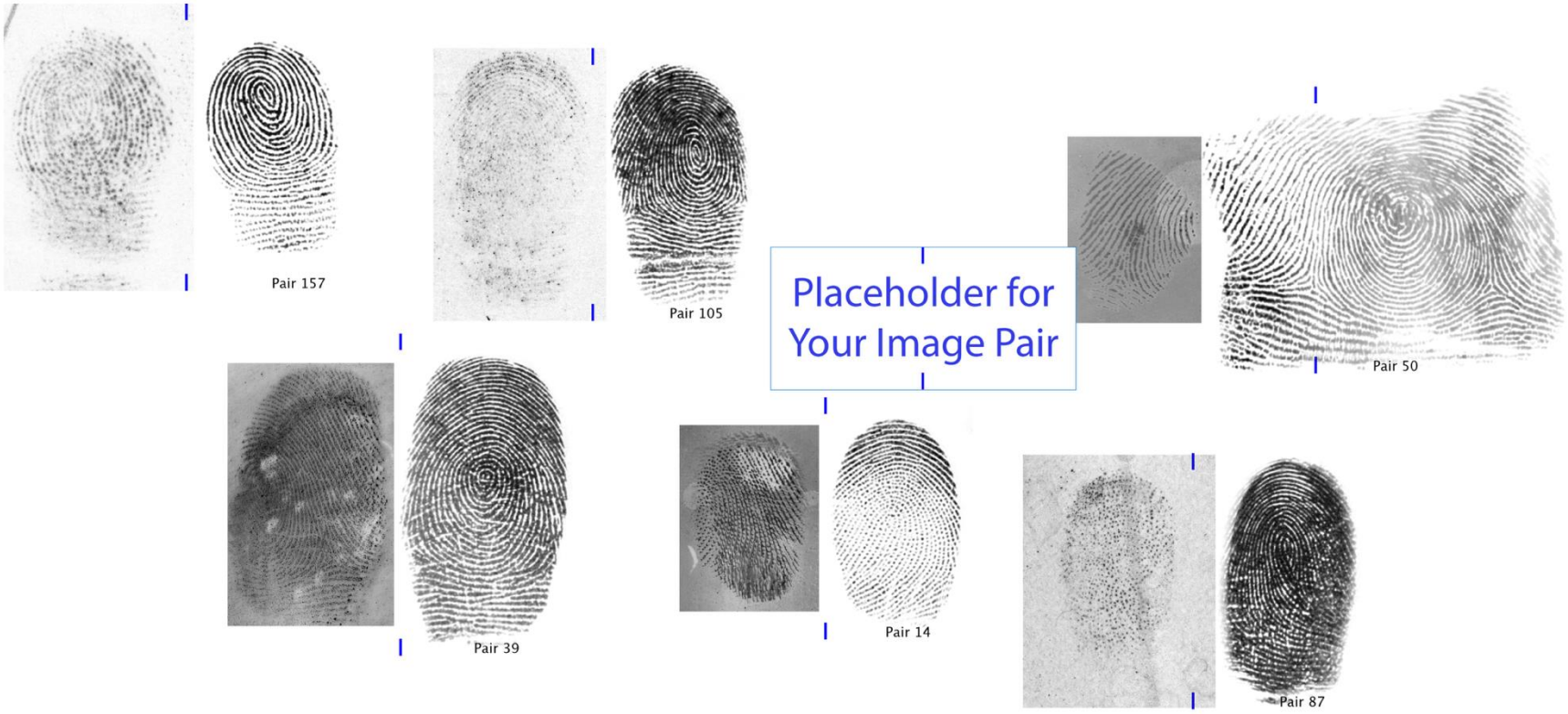
BASE-LR

- We have converted this to an online tool that you can use today!
- https://buseylab.org/BASE-LR/
- Benchmark Anchored Sorting Estimation Likelihood Ratios

BASE-LR

- https://buseylab.org/BASE-LR/
- It will take you through the benchmark comparisons and let you sort them.
- You then use a placeholder that represents where your current case falls relative to the other benchmark image pairs.

BASE-LR: Placeholder for your case



BASE-LR

- If your benchmark sort meets the criteria, we then report a likelihood ratio certificate.
- This can be annotated and downloaded.

BASE-LR

BASE-LR LIKELIHOOD RATIO CERTIFICATE

Date: May 19, 2026 at 11:39 AM
User ID: knowngood

Correlation (r-value): **0.98**

Interpolated mu-value: **5.19**

Likelihood Ratio: **207.2**

> The support for the same source proposition is **207.2** times more than the support for the different source proposition.

Note: example LR

LIKELIHOOD RATIOS GENERATED BY THE BENCHMARK ANCHORED SORTING ESTIMATION LIKELIHOOD RATIO (BASE-LR) METHOD

BASE-LR: When to Use

- This tool is best suited for borderline same-source cases.
- It can be used for exclusions, but the likelihood ratio tends to confuse people.
- The ordered probit model breaks down with really clear impressions. These are not necessary anyway in most cases.
- Most of the likelihood ratios will be in the 100’s to 1000’s.

Advantages of BASE-LR

- BASE-LR builds on the community-based database of responses in the black box study.
- It quantifies the strength of support for the same source proposition relative to the different source proposition.
- BASE-LR values can be reported directly, or used to augment traditional reporting scales.
- If someone wants to know the strength of your evidence, you can now estimate it.

Advantages of BASE-LR

- It smoothly handles borderline cases.
- Instead of worrying about “is it enough for an ID” you simply place your casework relative to the other benchmark image pairs.
- You can’t make an error, because no decision is made. You are just reporting a number.
- It gets around the debate about “Inconclusive”. You just report a weak LR.

Challenges of BASE-LR

- Likelihood ratios tend to be confusing for consumers.
- If we replace categorical conclusion reporting, we will have to rethink validation, proficiency, verification.
- These are fixable, known engineering challenges, whereas scientists view the alternative—categorical opinions—as having a fatal, unscientific flaw.

Questions and a plea for future research

https://buseylab.org/BASE-LR/

This research is only made possible if we have participants in studies. To be added to an email list for announcements about future research opportunities, visit this link: