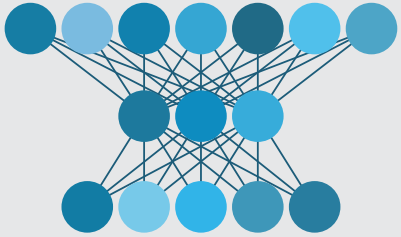
Consensus on validation of forensic-comparison systems in the context of casework

Geoffrey Stewart Morrison

Forensic Data Science Laboratory, Aston University

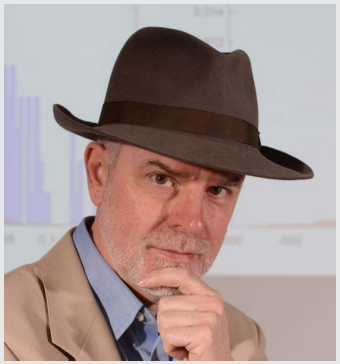
Forensic Evaluation Ltd

Aston University
BIRMINGHAM UK



$$\frac{p(E|H_p)}{p(E|H_d)}$$

Dr Geoffrey Stewart Morrison BSc MTS MA PhD FCSFS



Present:

- Director, Forensic Data Science Laboratory, Aston University
  - – forensic voice comparison
  - – forensic anthropology
  - – cell-site analysis
  - – fired cartridge cases
  - – communication of likelihood ratios
  - – fingerprints
- Director & Forensic Consultant, Forensic Evaluation Ltd
- Chair, Forensic Science Committee, British Standards Institution

Past:

- Scientific Counsel, Office of Legal Affairs, INTERPOL

Acknowledgments

- This research was supported by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2022.

- Thanks to the Netherlands Forensic Institute for hosting the initial meeting.

- Thanks to all the authors and supporters of the *Statement of consensus*.

- Thanks to my collaborators who helped develop the E3FS3α system, results from which are presented in *Calibration* and *Tippett plot* sections below.

- Thanks to the organizers of the AFORE webinar.

Disclaimer

- All opinions expressed are those of the presenter and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the presenter is associated.

Publication

- Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021). **Consensus on validation of forensic voice comparison**. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

- Supporters: Carlström Plaza F., González-Rodríguez J., Ramos D., Roberts P., Rose P., Solewicz Y., Vergeer P.

Contents

- Scope
- Methodology
- Document
- Key points
- Calibration
- log-likelihood-ratio cost (C_{llr})
- Tippett plots
- Reflections on broader issues related to validation and standard/guidelines

Scope

Scope

- Validation for the purpose of **demonstrating whether**, in the context of specific cases, forensic-comparison systems* are (or are not) **good enough for their output to be used in court**.

* The original states “forensic-voice-comparison systems”, but by making a few minor changes in wording (as I have done in this presentation) the *Statement of consensus* is **applicable across multiple branches of forensic science**.

Scope

- Addresses **scientific** matters that could have a bearing on legal decisions, but does not address legal matters directly.

- Validation of forensic-comparison **systems that are based on relevant data, quantitative measurements, and statistical models**,* and that **output numeric likelihood ratios**.

*** Also applicable to systems based on human perception and subjective judgement** if the system is calibrated.

explanation to follow

Methodology

Methodology

- **Invited participants** were individuals who when brought together could be considered **representative of the relevant scientific community**.

- Included individuals who had knowledge and experience of validating forensic-voice-comparison systems in **research** and/or **casework** contexts.

- Included individuals who **had actually presented validation results to courts**.

- Also included individuals who could **bring a legal perspective** on these matters, and individuals with **knowledge and experience of validation in forensic science more broadly**.

Methodology

- **Two-day meeting** in September 2019
  - **discuss concepts and attempt to reach consensus**
  - all participants given opportunities to speak, notes made on consensus reached
- Lead author drafts document and circulates to participants
- **Three online meetings** January, February, March 2020
  - participants provide feedback on draft and **discuss concepts** during meetings
  - lead author revises draft based on discussion and circulates revised draft
- **Five online meetings** April through August 2020 (**formal ISO-type process**)
  - participants submit comments and **concrete proposals** for changes of wording
  - proposals discussed and **decision to adopt or reject** made during meeting
  - earlier rounds focussed on specific sections, later rounds on whole document

Document

Document

- Introduction

- **Statement of consensus**
  - **recommendations** “should”

- **Informational appendices**
  - Appendix A: The likelihood-ratio framework
  - Appendix B: Recording conditions
  - Appendix C: Tippett plots and C_{llr}
  - Appendix D: Methodology

- References

Document

- Statement of consensus

    2.1. Scope

    2.2. Calculating a likelihood ratio: Propositions, relevant population, and conditions

    2.3. Calculating a likelihood ratio: Calibration

    2.4. Validation procedures

    2.5. Validation data

    2.6. Decision as to whether calibration and validation data are sufficient

Document

- Statement of consensus

  2.7. Anticipatory and case-by-case validation

  2.8. Presenting validation results

  2.9. Relationship between conditions and performance

  2.10. Validation threshold for C_{llr}

  2.11. Decision as to whether the likelihood-ratio value for the comparison of the questioned-source and known-source items is supported by the validation results

  2.12. Summary of key points

Key points

Key points

- Statement of consensus

2.12.1. The forensic practitioner **should communicate** to the court what **propositions** the forensic practitioner has adopted for the case, including what they have adopted as the **relevant population**.

2.12.2. The forensic practitioner **should communicate** to the court what the forensic practitioner understands the **conditions of the questioned-source and known-source items** to be.

Key points

- Statement of consensus

  2.12.3. The forensic-comparison system **should be well calibrated**.

explanation to follow

Key points

- Statement of consensus

2.12.4. **Validation data should be representative of the relevant population** for the case, and **reflective of the conditions** of the questioned-source and known-source items in the case.

2.12.5. The forensic practitioner's **decision** as to whether the validation data are sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-source and known-source items in the case, will be a **subjective judgement**.

Key points

- Statement of consensus

2.12.6. **Validation results should be presented as a Tippett plot and a C_{llr} value**. These **should be examined for signs of miscalibration**.

2.12.7. **The validation threshold (acceptance criterion) for C_{llr} should be 1**. As long as C_{llr} is less than 1, the system is providing useful information.

explanation to follow

Key points

- Statement of consensus

2.12.8. To decide whether the **likelihood-ratio value** calculated for the comparison of the questioned-source and known-source items is **supported by the validation results**, it **should be compared with the values shown in the Tippett plot**.

explanation to follow

Progress

- Scope ✓
- Methodology ✓
- Document ✓
- Key points ✓
- Calibration
- log-likelihood-ratio cost (C_{llr})
- Tippett plots
- Reflections on broader issues related to validation and standard/guidelines

Calibration

Calibration

- What is a well-calibrated likelihood-ratio system?

- A system for which **the likelihood ratio of the likelihood ratio is the likelihood ratio**.

$$LR = \frac{f(LR \mid H_s)}{f(LR \mid H_d)}$$

Calibration

- Perfectly calibrated ln(LR) distributions

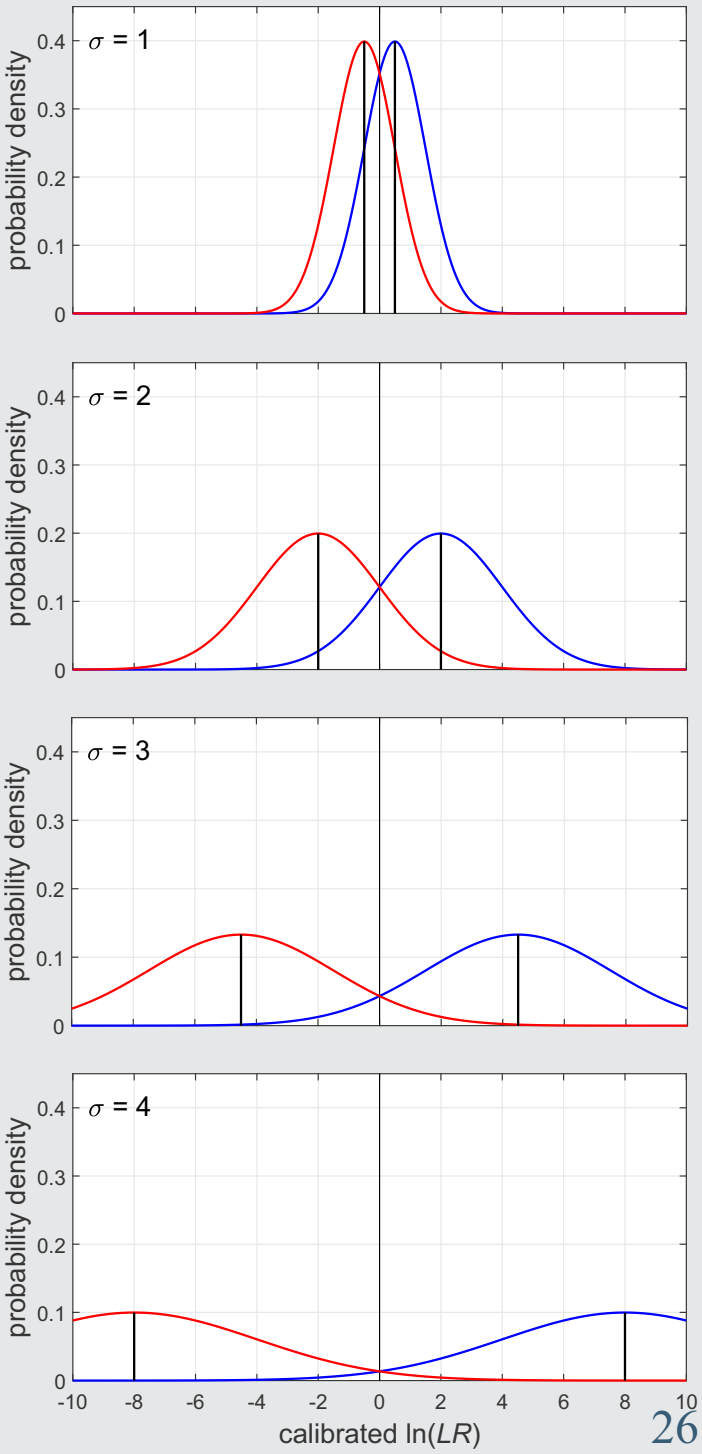
- different-source distribution and

  same-source distribution:

  – Gaussian

  – same variance

$$\mu_d = -\frac{\sigma^2}{2} \qquad \mu_s = +\frac{\sigma^2}{2}$$



Calibration

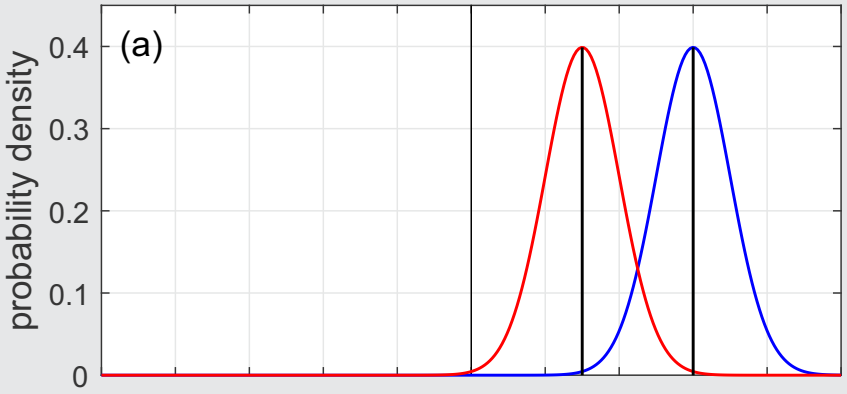
(a)

Uncalibrated scores

$\mu_d = 3$

$\mu_s = 6$

$\sigma = 1$



Vocabulary:

"score" = "uncalibrated log likelihood ratio"

"score" ≠ "similarity score"

Calibration

(a)

Uncalibrated scores

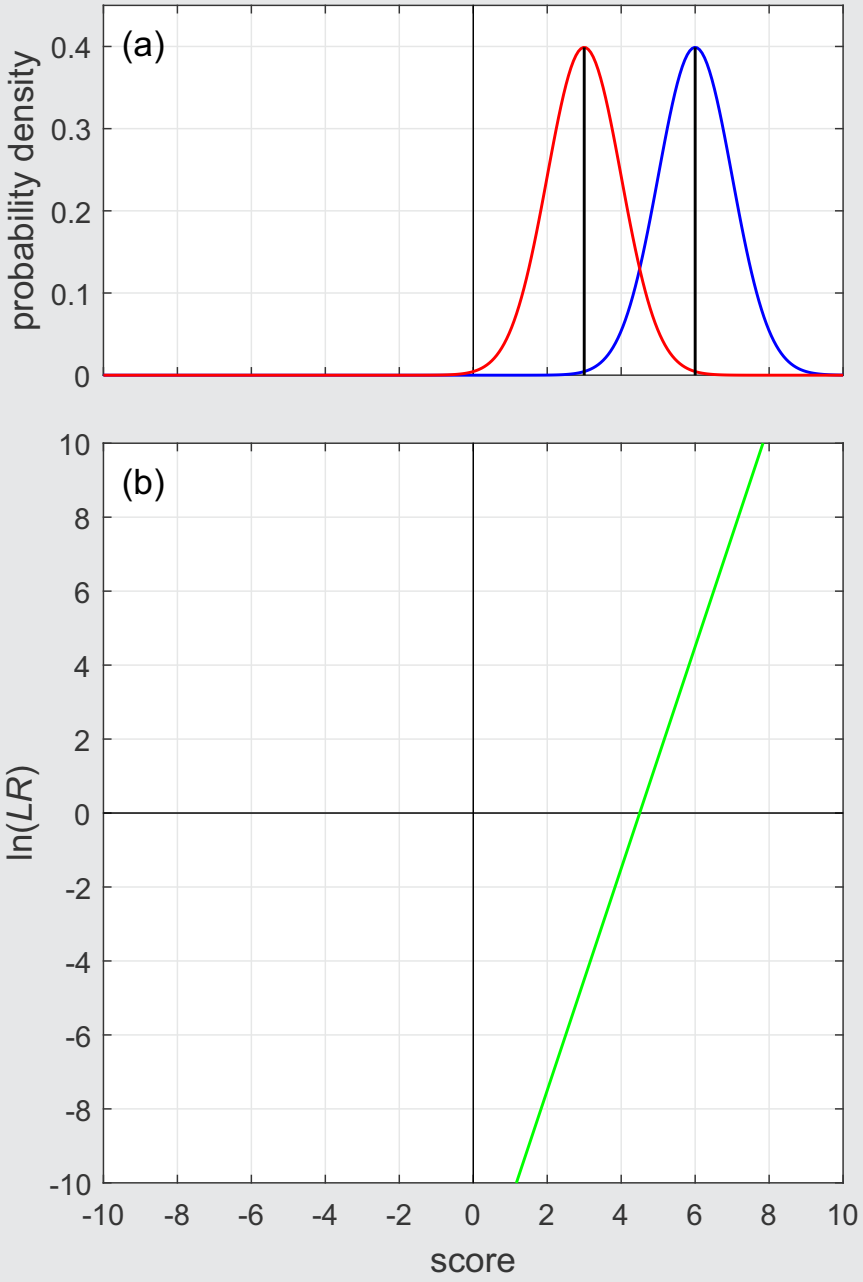
$\mu_d = 3$

$\mu_s = 6$

$\sigma = 1$

(b)

Score to ln(*LR*) mapping function



Calibration

(c)

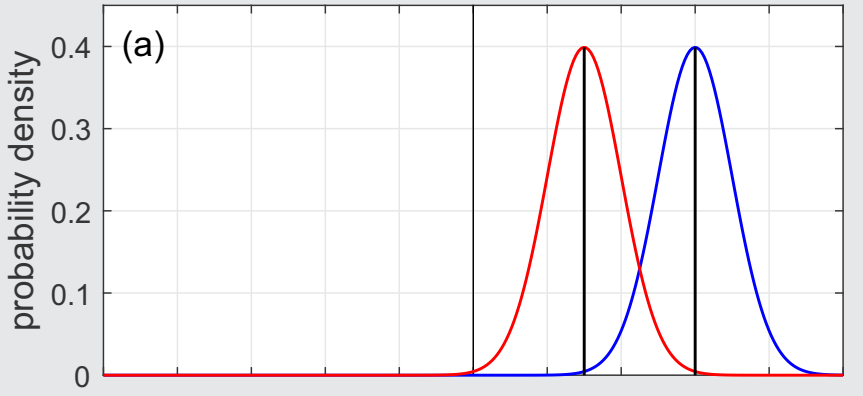
Calibrated

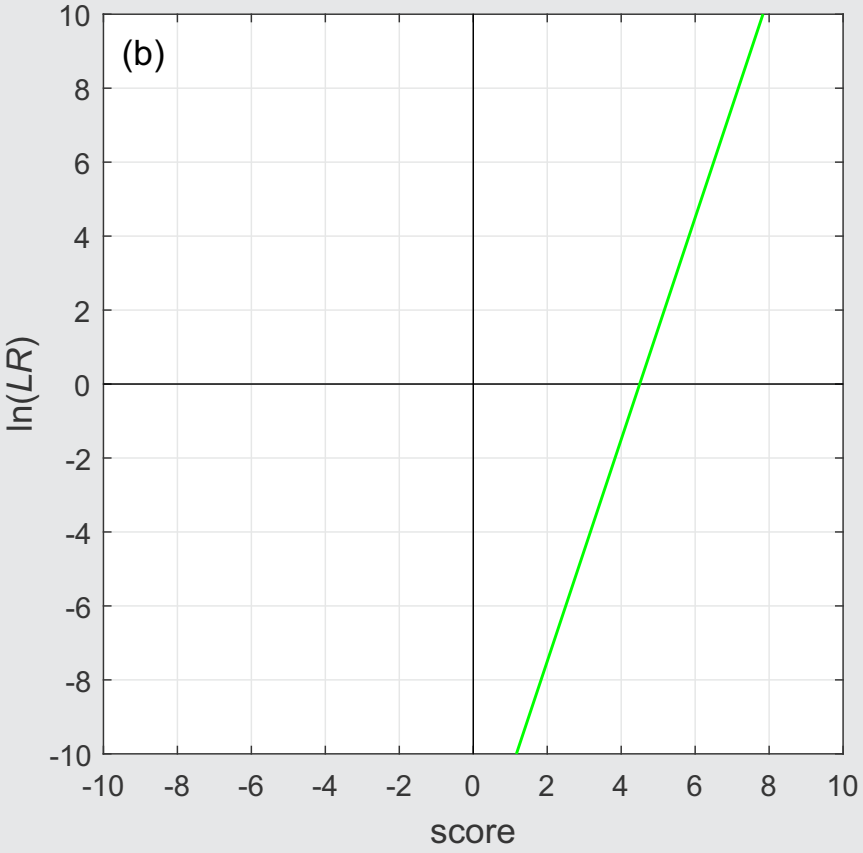
$\ln(LR)$

$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$





Calibration

(c)

Calibrated ln(*LR*)

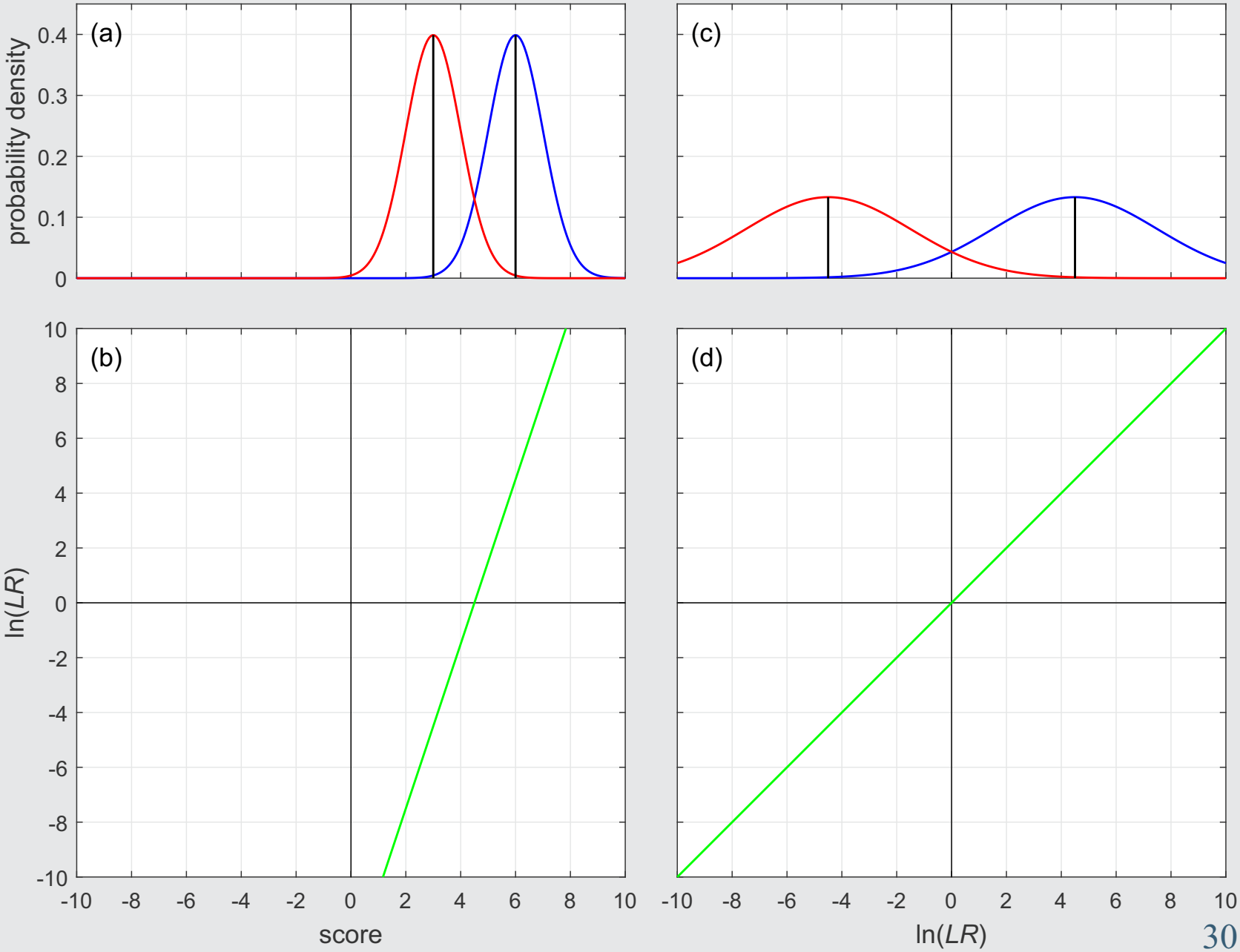
$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$

(d)

ln(*LR*) to ln(*LR*) mapping function



Calibration

- Score [x] to ln(LR) [y] mapping function:

$$y = a + bx$$

$$a = -b\frac{\mu_s + \mu_d}{2} \qquad b = \frac{\mu_s - \mu_d}{\sigma^2}$$

where μ_s, μ_d, σ are the statistics for the scores

Calibration

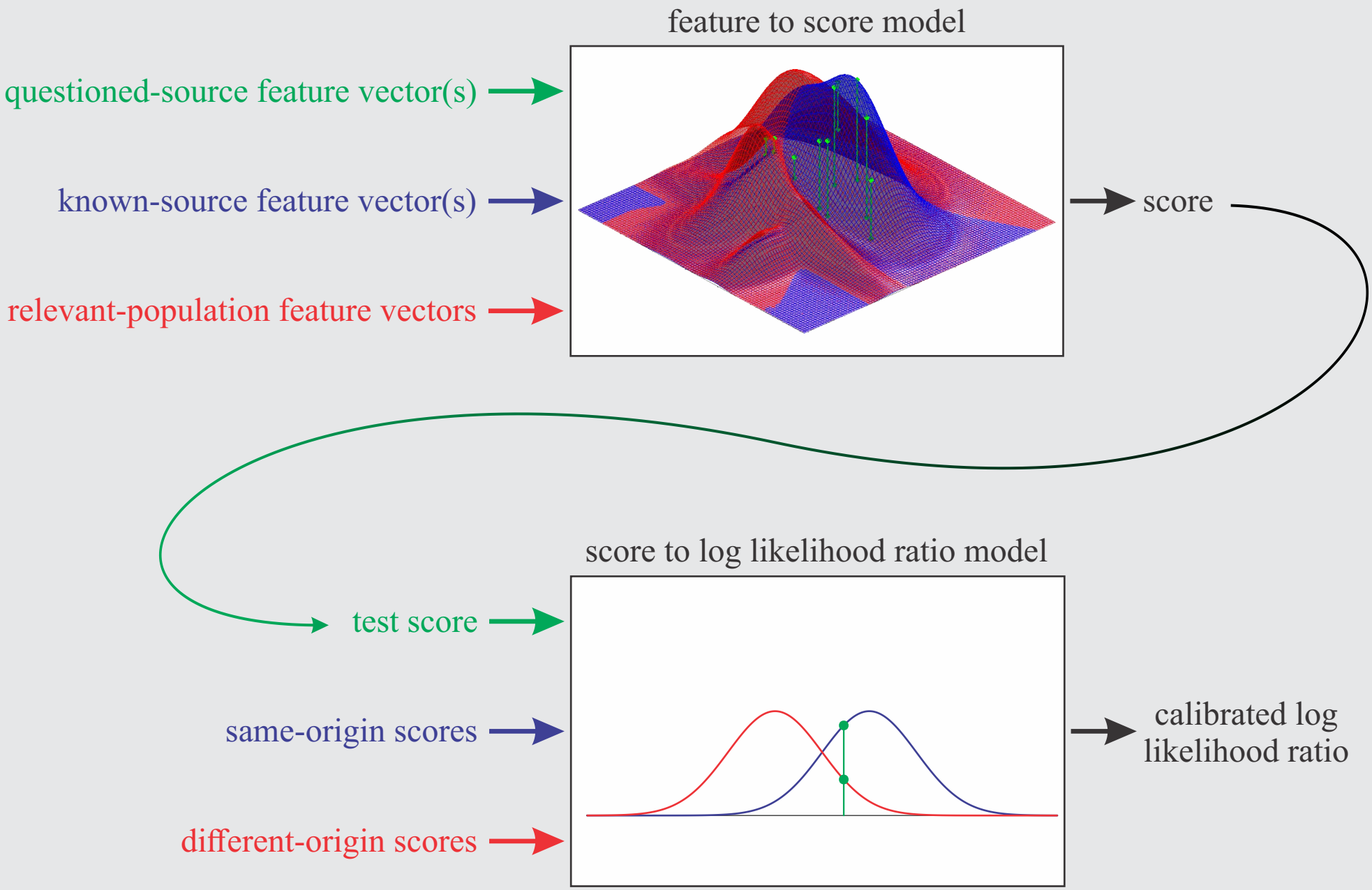
- Score [x] to ln(LR) [y] mapping function:

$$y = a + bx$$

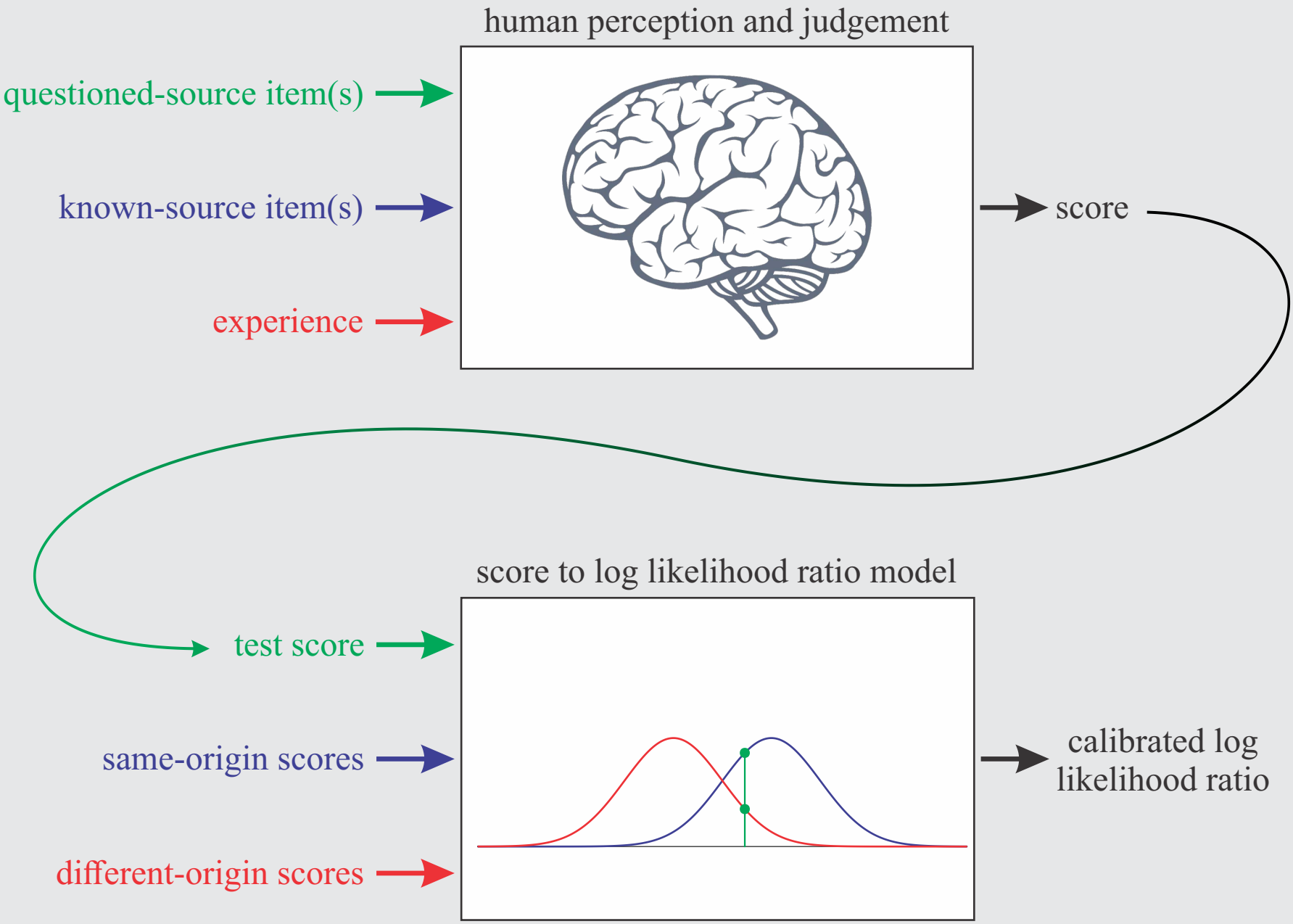
- In practice, **logistic regression** is used to calculate a and b.

- It is more robust to violations of the assumptions of Gaussian distributions with the same variance.

Calibration



Calibration



Calibration

- Calibration data are needed to train the calibration model, but **calibration data are free**.

- Use **validation data** with leave-one-item-out / leave-two-items-out **cross-validation**.

Calibration

- More information:

Morrison G.S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45, 173–197. http://dx.doi.org/10.1080/00450618.2012.733025 [Preprint: https://arxiv.org/abs/2104.08846]

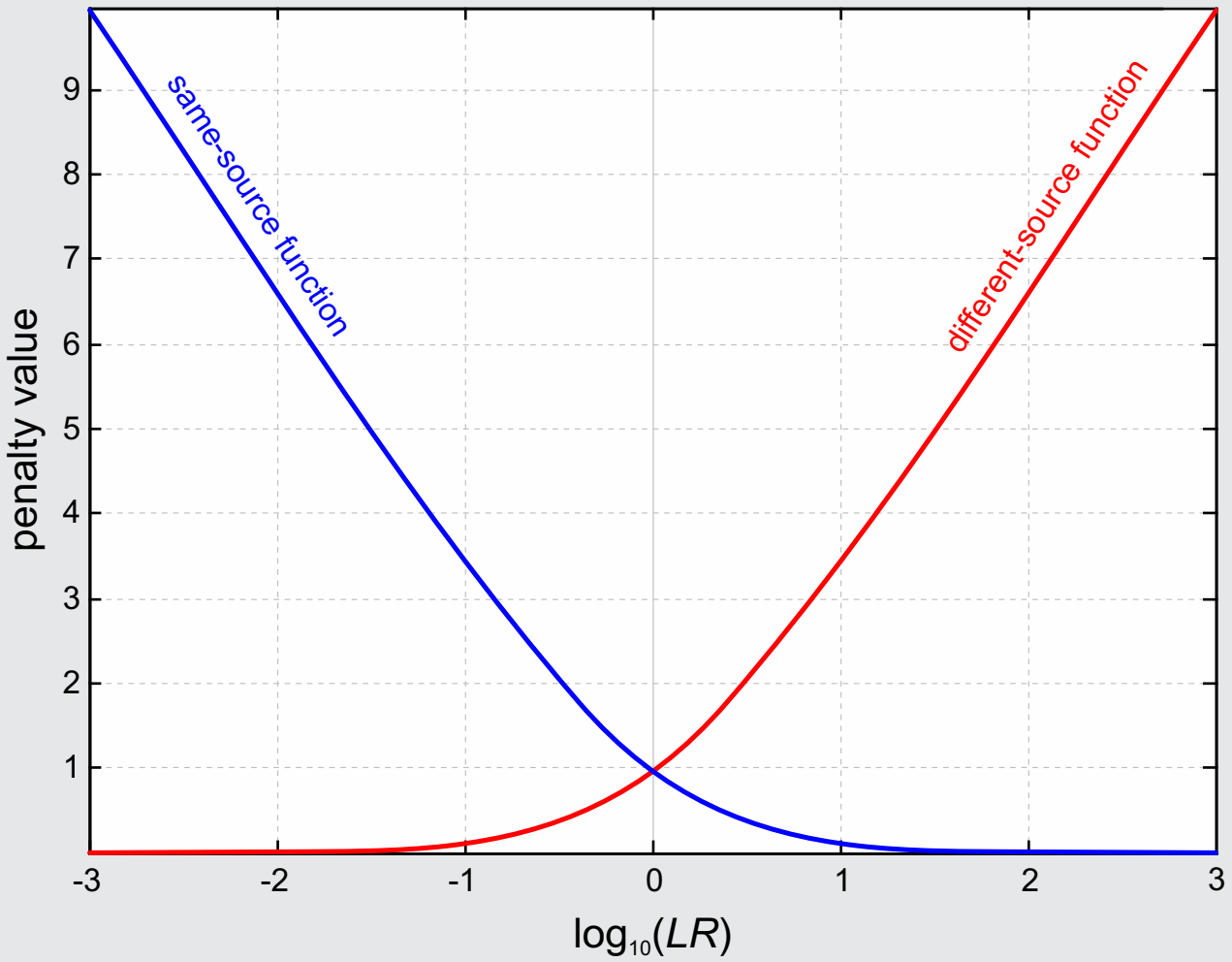
Morrison G.S. (2021). In the context of forensic casework, are there meaningful metrics of the degree of calibration? *Forensic Science International: Synergy*, 3, article 100157. https://doi.org/10.1016/j.fsisyn.2021.100157

Morrison G.S., Ferrer L., Ramos D., Vergeer P., Puch-Solis R., Ypma R.J.F. (2021) Symposium on calibration in forensic science. http://forensic-evaluation.net/symposium_on_calibration/

log-likelihood-ratio cost (C_{llr})

log-likelihood-ratio cost

- Penalty functions for calculating C_{llr}



log-likelihood-ratio cost

- Formula for calculating C_{llr}

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{s}}} \sum_{i=1}^{N_{\text{s}}} \log_2\left(1 + \frac{1}{LR_{\text{s}_i}}\right) + \frac{1}{N_{\text{d}}} \sum_{j=1}^{N_{\text{d}}} \log_2\left(1 + LR_{\text{d}_j}\right) \right)$$

log-likelihood-ratio cost

- **The better the performance of the system, the smaller the C_{llr} value**

- A system that **always responds with a likelihood-ratio value of 1** irrespective of the input provides **no useful information**
  – **the posterior odds will alway equal the prior odds**

- This system will have $C_{llr} = 1$

log-likelihood-ratio cost

- **The better the performance of the system, the smaller the C_{llr} value**

- A **well-calibrated system** will have $\boldsymbol{C_{llr} \leq 1}$

- If $C_{llr} < 1$, the system is providing useful information

- $C_{llr} > 1$ can occur for uncalibrated or miscalibrated systems
  – this can be fixed by appropriately calibrating the system

log-likelihood-ratio cost

- Examples of different systems validated on the same validation data

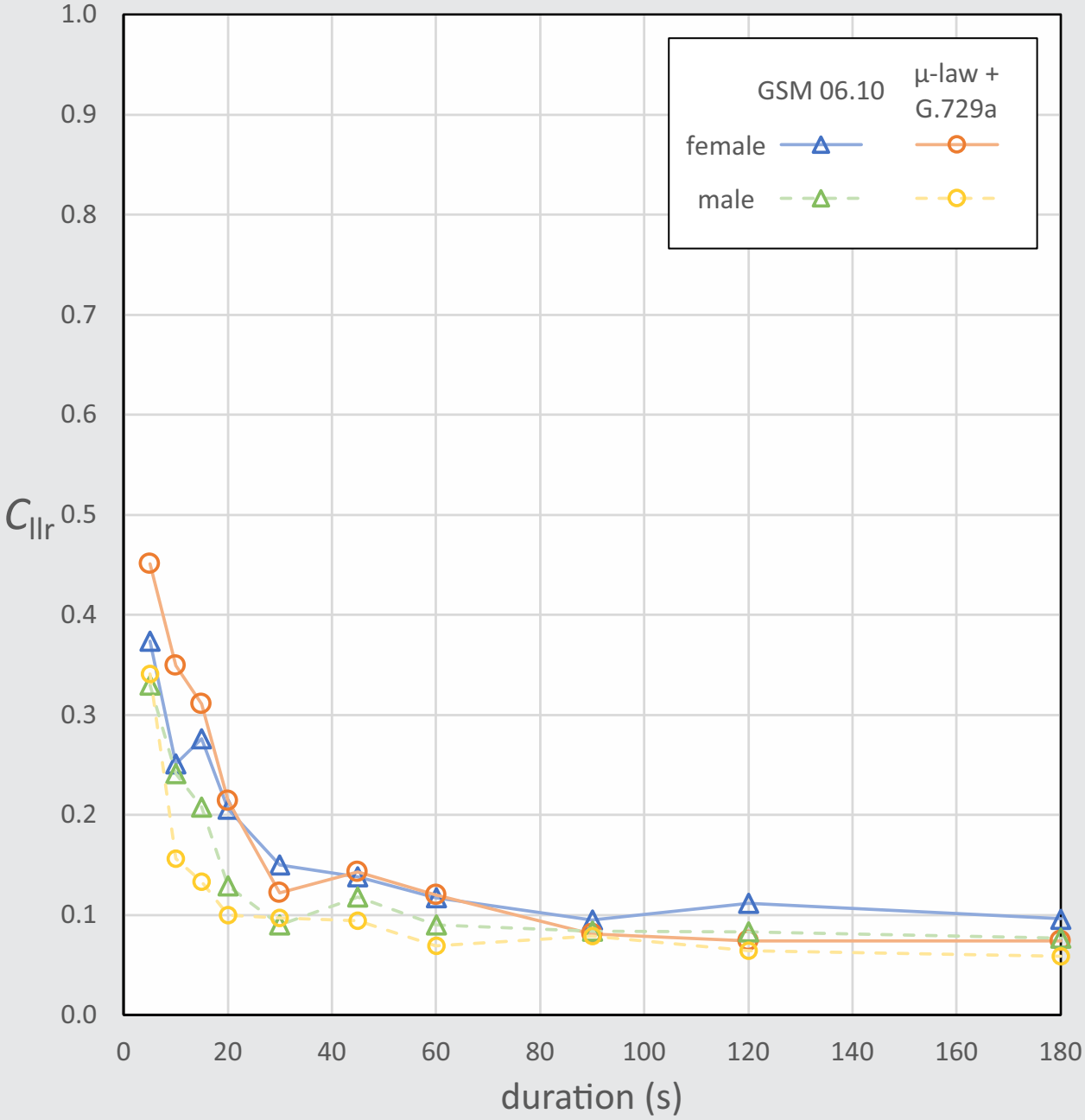
- **Baseline** validation

System name	System type	C_{llr}
Batvox 3.1	GMM-UBM	0.59
MSR GMM-UBM	GMM-UBM	0.58
MSR GMM i-vector	GMM i-vector	0.45
Batvox 4.1	GMM i-vector	0.37
Nuance 9.2	GMM i-vector	0.29
VOCALISE 2017B	GMM i-vector	0.27
VOCALISE 2019A	x-vector	0.25
E3FS3α	x-vector	0.21
Phonexia BETA4	x-vector	0.21

log-likelihood-ratio cost

- Examples of the same base system (E3FS3α) validated on data with different **case-specific** conditions

- Each system variant included cross-validated calibration



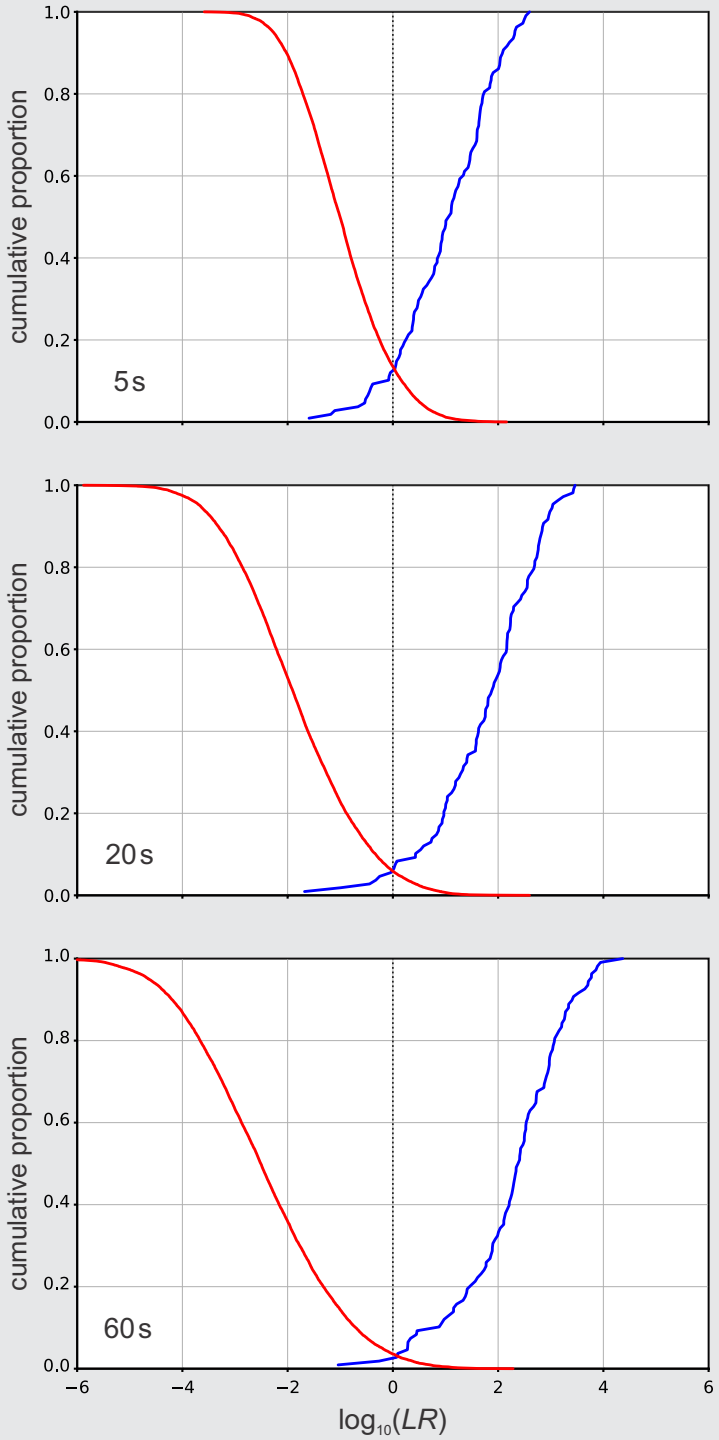
Tippett plots

Tippett plots

- Tippet plots for same base system (E3FS3α) validated on different conditions

- Each system variant included cross-validated calibration

- Corresponding C_{llr} values:
  - 0.45
  - 0.21
  - 0.12



Reflections

Reflections

- Validation is black-box testing of the whole system

- **Validation requirements should be exactly the same irrespective of the internal architecture of the forensic-comparison system**
  – relevant data, quantitative measurements, and statistical models
  – human perception and subjective judgement

- If it is practically difficult to validate a system because of its internal architecture, this should not excuse the system from being validated according to the same requirements imposed on other systems.

Reflections

- **Calibration is an essential thing to do as part of evaluation**

  – either using an initial model that is naturally well-calibrated, or using an overt calibration model as the final stage of the system

- Once a system has been calibrated, there is no meaningful metric of its degree of calibration

  Morrison G.S. (2021). In the context of forensic casework, are there meaningful metrics of the degree of calibration? *Forensic Science International: Synergy*, 3, article 100157. https://doi.org/10.1016/j.fsisyn.2021.100157

Reflections

- **There is no one-size-fits-all validation**

- Anticipatory validation:
  – Is the relevant population and are the conditions for this case sufficiently similar to those of an existing validation report?

- Case-by-case validation:
  – Can sufficient calibration and validation data be obtained that are sufficiently representative of the relevant population and sufficiently reflective of the conditions for this case?

Reflections

- **Validation should be fit for purpose**

- At least as traditionally interpreted, it could be that validation based on ISO 17025 is not fit for the purpose of validating forensic-comparison systems in the context of casework

- Potential solutions:
  - different interpretation
  - develop a new ISO standard on validation of forensic-comparison systems

Reflections

- **Consensus development**

☺ researchers and practitioners figure out what is **fit-for-purpose practice** → consensus → guidelines and standards

☹ existing practice → consensus → guidelines and standards → **fit-for-purpose practice**

Morrison G.S., Neumann C., Geoghegan P.H. (2020). Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 2, 206–209. https://doi.org/10.1016/j.fsisyn.2020.06.005

Morrison G.S., Neumann C., Geoghegan P.H., Edmond G., Grant T., Ostrum R.B., Roberts P., Saks M., Syndercombe Court D., Thompson W.C., Zabell S. (2021). Reply to Response to Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 3, article 100149. https://doi.org/10.1016/j.fsisyn.2021.100149

Reflections

- **Courts should require meaningful validation**

- Courts should look beyond whether validation was performed
  - the court should consider as to **whether the validation data were representative of the relevant population and reflective of the conditions for the case**
  - the court should consider **whether the validation results are good enough for the court's purpose**
  - the court should consider **whether the likelihood-ratio value for the comparison of the questioned-source item and known-source item is supported by the validation results**

Thank You

http://geoff-morrison.net/