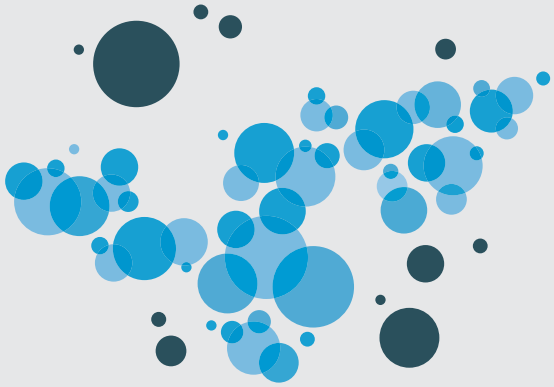
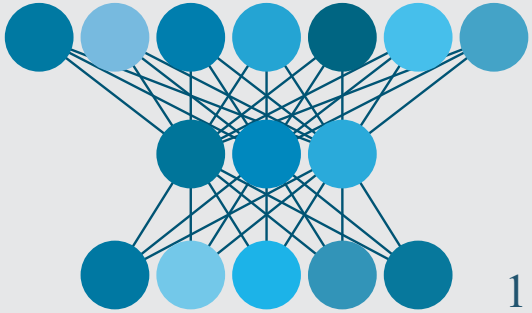
Symposium on calibration in forensic science

Forensic Data Science Laboratory
Aston Institute for Forensic Linguistics



Aston University
BIRMINGHAM UK



Calibration in forensic science

Geoffrey Stewart Morrison

Forensic Data Science Laboratory
Aston Institute for Forensic Linguistics

Calibration in weather forecasting

- Weather forecaster predicts:

  Probability of precipitation for tomorrow is 40%.

- The next day it either rains or it doesn't rain.

- Looking at lots of days for which the weather forecaster's PoP was 40%, on what percentage of those days did it actually rain?

Calibration in weather forecasting

Well calibrated:

- Prediction: 40%
- Actual: 40%

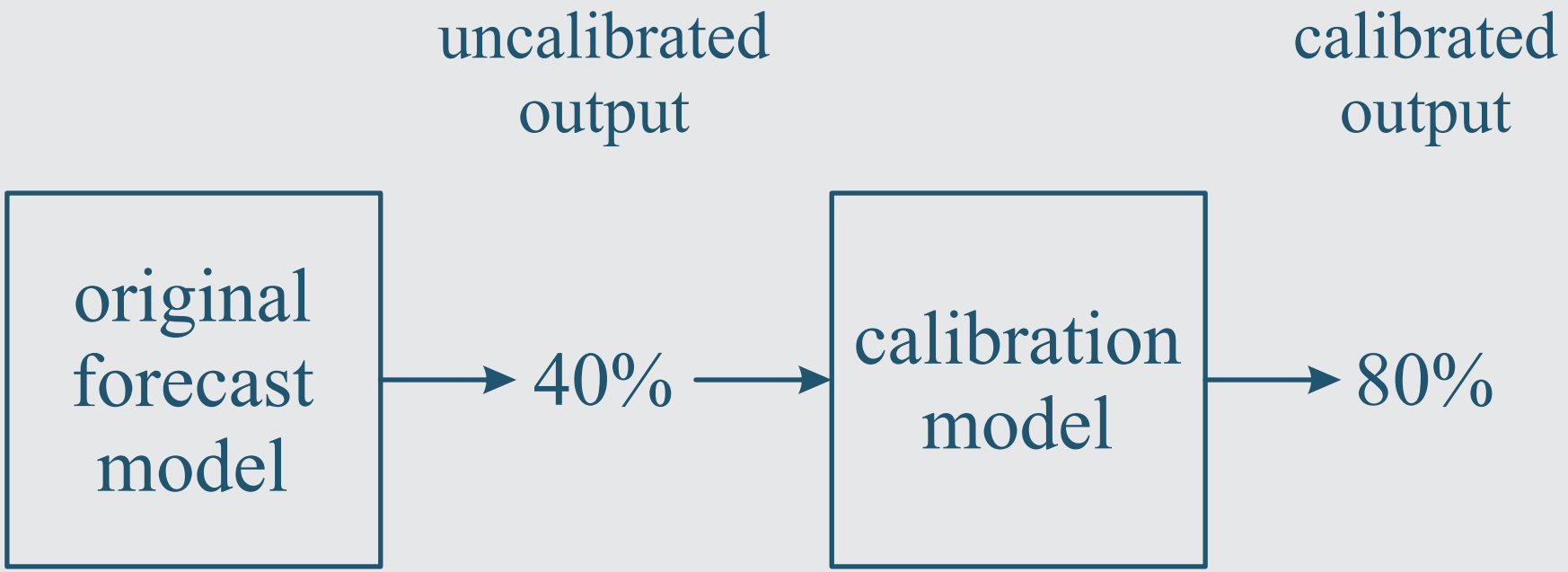
Not well calibrated:

- Prediction: 40%
- Actual: 80%

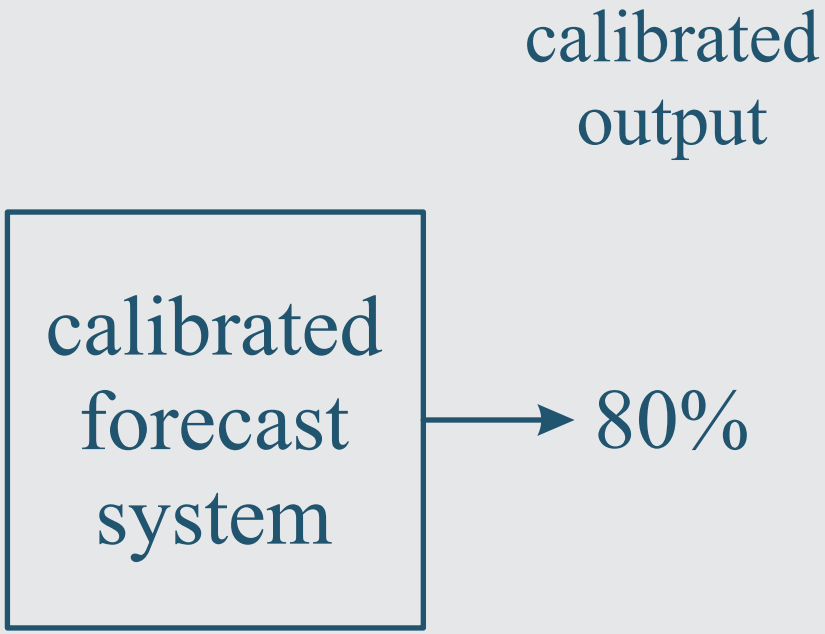
Calibration in weather forecasting

- Solution:
- Collect data from a large number of past days.
- For each day collect: prediction actual weather
- Train a calibration model.
- Use model to calibrate future predictions.

Calibration in weather forecasting



Calibration in weather forecasting



Calibration in general

- Peterson W.W., Birdsall T.G., Fox W.C. (1954). **The theory of signal detectability**. *Transactions of the IRE Professional Group on Information Theory*, 4, 171–211. https://doi.org/10.1109/TIT.1954.1057460

- Birdsall T.G. (1973). **The theory of signal detectability: ROC curves and their character**. Technical Report No. 177. Cooley Electronics Laboratory, Department of Electrical and Computer Engineering, The University of Michigan. Ann Arbor, Michigan.

- Dawid A.P. (1982). **The well-calibrated Bayesian**. *Journal of the American Statistical Association*, 77, 605–610. http://doi.org/10.1080/01621459.1982.10477856

- Good I.J. (1985). **Weight of evidence: A brief survey**. In Bernardo J.M., DeGroot M.H., Lindley D.V., Smith A.F.M. (Eds.), *Bayesian Statistics 2* (pp. 249–270). Elsevier.

Calibration in general

- If a model is a parsimonious parametric model.
- If there is a large amount of training data relative to the number of parameter values to be estimated.
- If the assumptions of the model are not violated by the population distributions.
- Then the output of the model will be well calibrated.

Calibration in forensic science

- Models often fit complex distributions to high-dimensional data.
- The amount of case-relevant training data is often small relative to the number of parameter values to be estimated.
- The assumptions of the models may be violated.
- The output of the models are often not well calibrated.

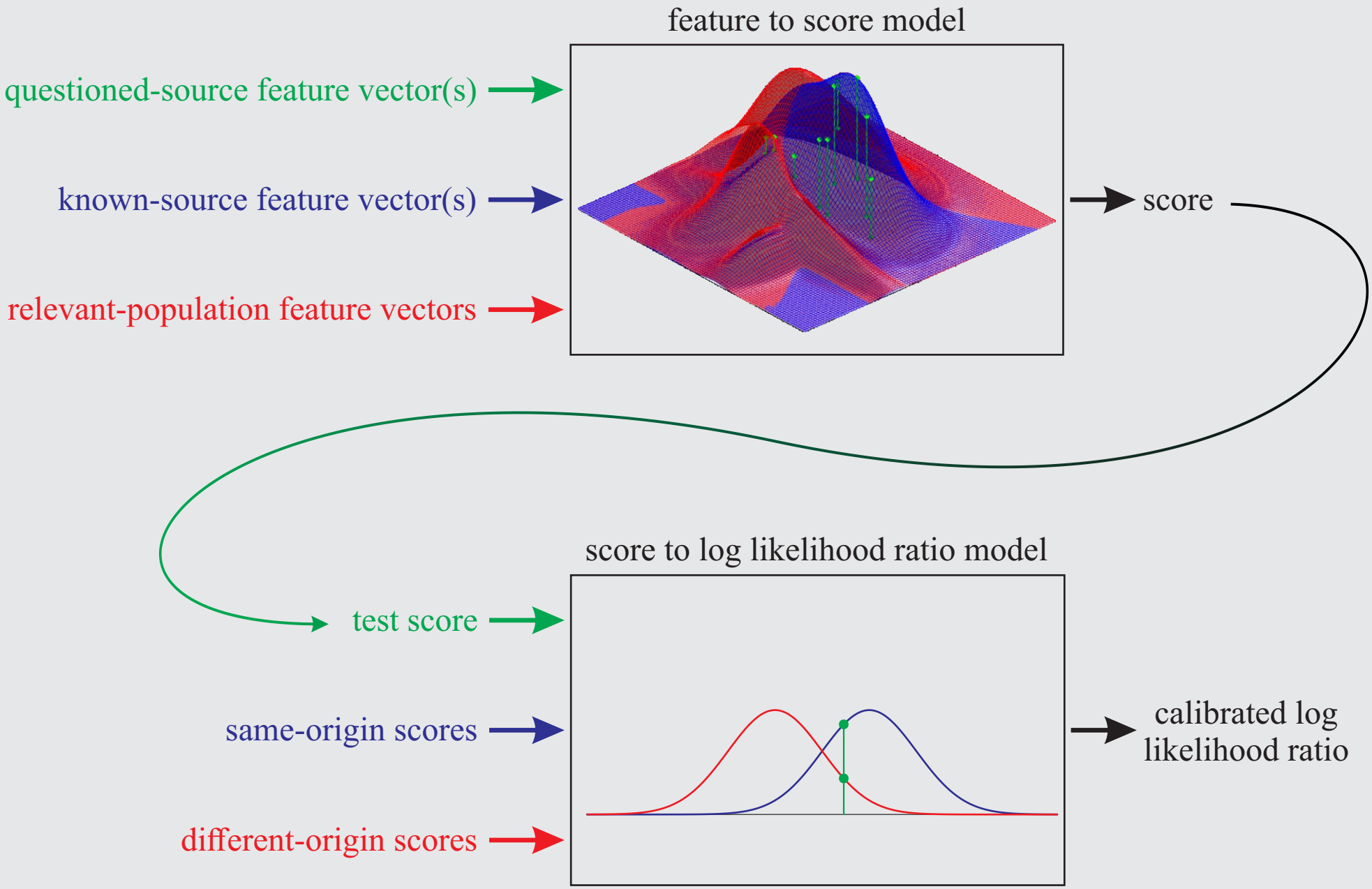
Calibration in forensic science

- Solution:

- Treat the output of the first (complex) model as uncalibrated log likelihood ratios (scores).

- Use a parsimonious model to convert the scores to calibrated log likelihood ratios.

Calibration in forensic science



Calibration in forensic science

- Take data that reflect the relevant population and conditions of the questioned-source specimen and known-source sample in the case.
- Construct same-source pairs and different-source pairs.
- Use the first model to calculate a score for each pair.
- Use the resulting same-source scores and different-source scores to train the calibration model.

Calibration in forensic science

- The scores are unidimensional.
- The calibration model is parsimonious.
- There is a large amount of data relative to the number of parameter values to be estimated.
- The output of the calibration model is well calibrated.

Calibration in forensic science

- Important condition:

- The data used for training the calibration model must reflect the relevant population and the conditions of the questioned-source specimen and known-source sample in the case.

- If not, the system will be miscalibrated.

Calibration in forensic science

- Important condition:

- The first model must output scores which are uncalibrated log likelihood ratios. They must take account of both:
  - the **similarity** between the questioned-source specimen and the known-source sample
  - their **typicality** with respect to the relevant population

- Similarity-only scores cannot be used.

Calibration in forensic science

- Similarity-only scores cannot be used.

- Morrison G.S., Enzinger E. (2018). **Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality**. *Science & Justice*, 58, 47–58. http://dx.doi.org/10.1016/j.scijus.2017.06.005

- Neumann C., Ausdemore M. (2020). **Defence against the modern arts: The curse of statistics –Part II: 'Score-based likelihood ratios'**. *Law, Probability and Risk*, 19, 21–42. http://dx.doi.org/10.1093/lpr/mgaa006

- Neumann C., Hendricks J., Ausdemore M. (2020). **Statistical support for conclusions in fingerprint examinations**. In Banks D., Kafadar K., Kaye D.H., Tackett M. (Eds.) *Handbook of Forensic Statistics* (Ch. 14, pp. 277–324). Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

Calibration in forensic science

- forensic voice comparison
- fingerprints
- DNA
- mRNA
- glass fragments
- mobile telephone colocation
- human perception and judgement

Calibration in forensic science

- González-Rodríguez J., Rose P., Ramos D., Toledano D.T., Ortega-García J. (2007). **Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition**. *IEEE Transactions on Audio, Speech, and Language Processing*, 15, 2104–2115. https://doi.org/10.1109/TASL.2007.902747

- 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 https://arxiv.org/abs/2104.08846

- Ypma R.J.F., Maaskant - van Wijk P.A., Gill R., Sjerps M., van den Berge M. (2021). **Calculating LRs for presence of body fluids from mRNA assay data in mixtures**. *Forensic Science International: Genetics*, 52, article 102455. https://doi.org/10.1016/j.fsigen.2020.102455

Well-calibrated likelihood-ratio systems

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

- The likelihood ratio of the likelihood ratio is the likelihood ratio.

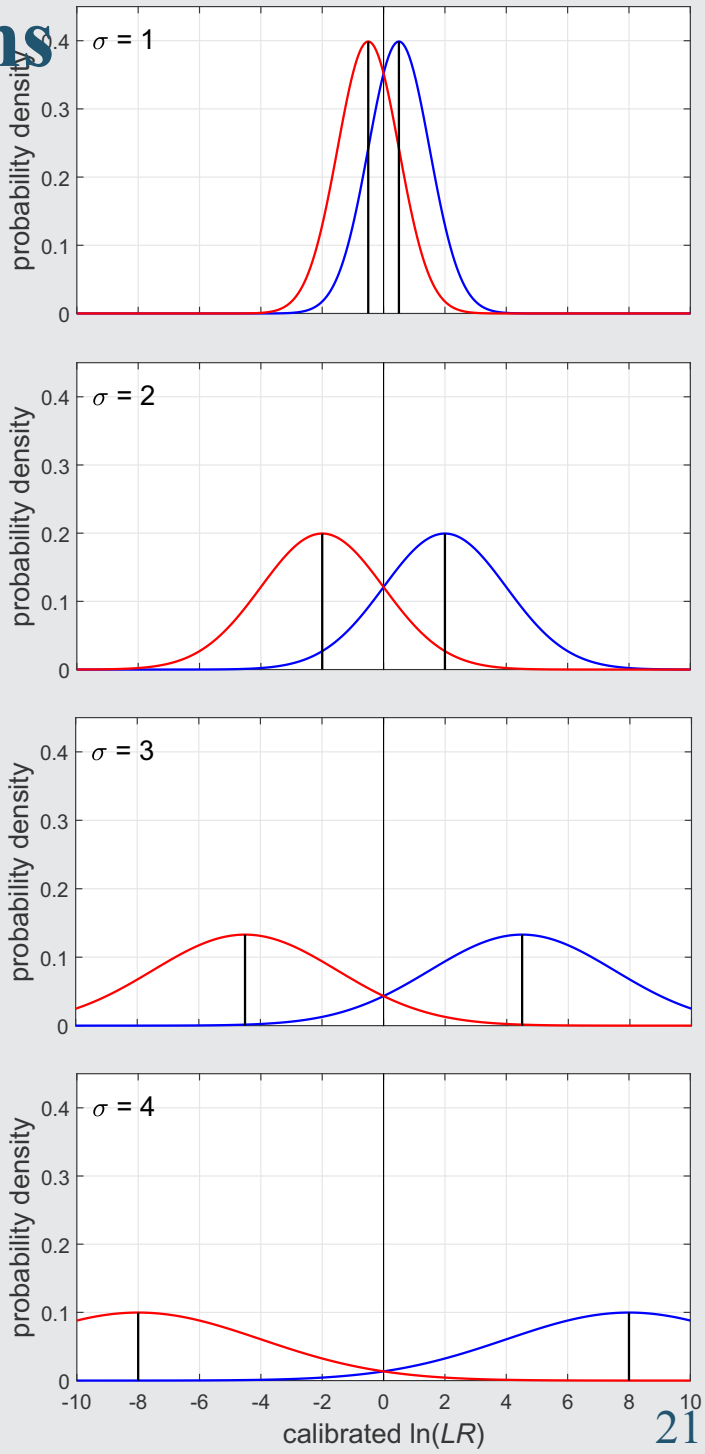
$$LR = \frac{f(\,LR \mid H_{\mathrm{s}}\,)}{f(\,LR \mid H_{\mathrm{d}}\,)}$$

Well-calibrated likelihood-ratio systems

- Perfectly calibrated ln(LR) distributions.

- Both same-source and different-source distributions are Gaussian, and they have the same variance.

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



Calibrating likelihood-ratio systems

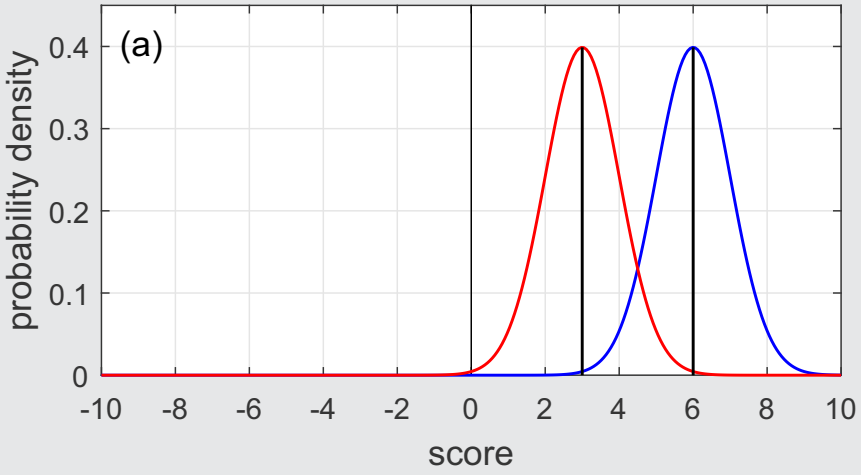
(a)

Uncalibrated scores.

$\mu_d = 3$

$\mu_s = 6$

$\sigma = 1$



Calibrating likelihood-ratio systems

(a)

Uncalibrated scores.

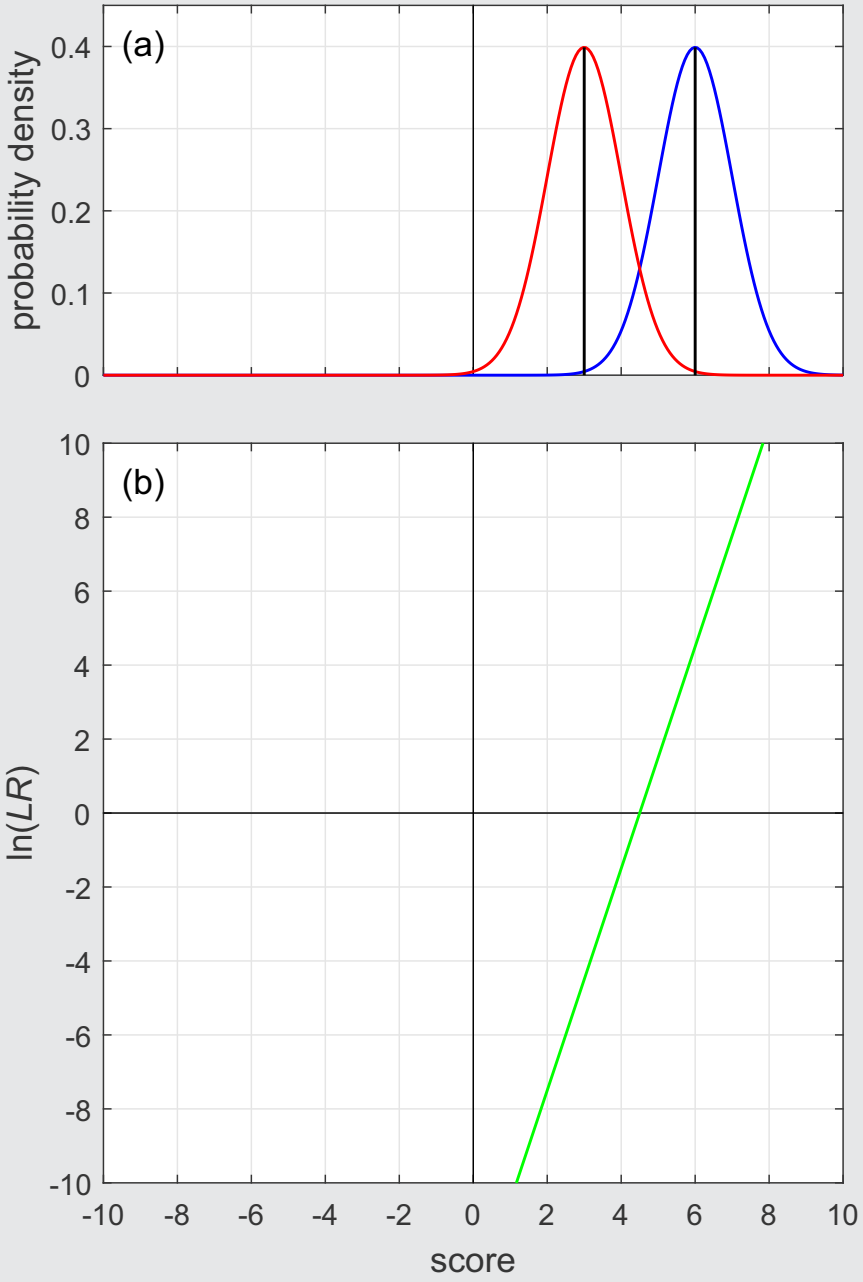
$\mu_d = 3$

$\mu_s = 6$

$\sigma = 1$

(b)

Score to ln(*LR*) mapping function.



Calibrating likelihood-ratio systems

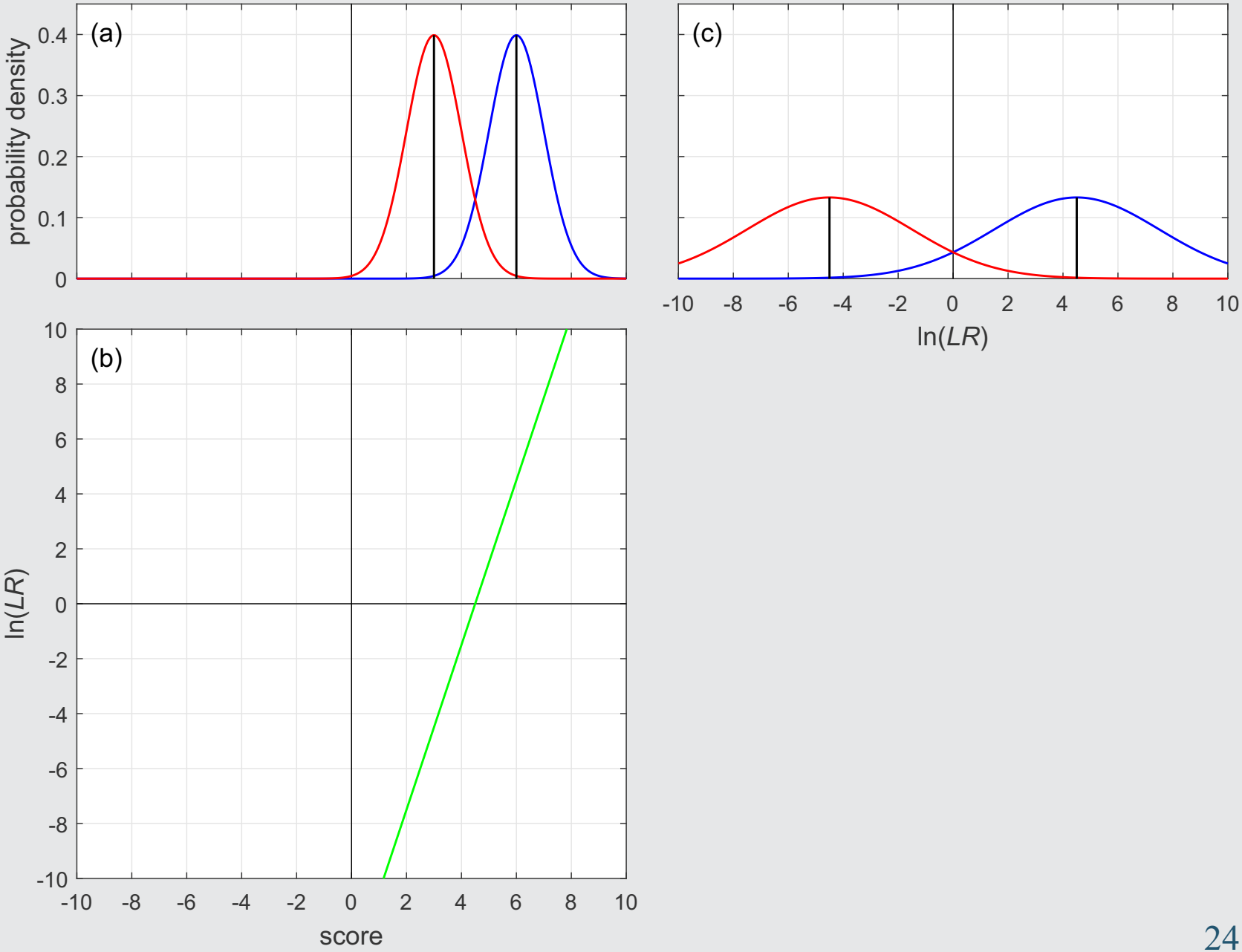
(c)

Calibrated ln(*LR*).

$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$



Calibrating likelihood-ratio systems

(c)

Calibrated ln(*LR*).

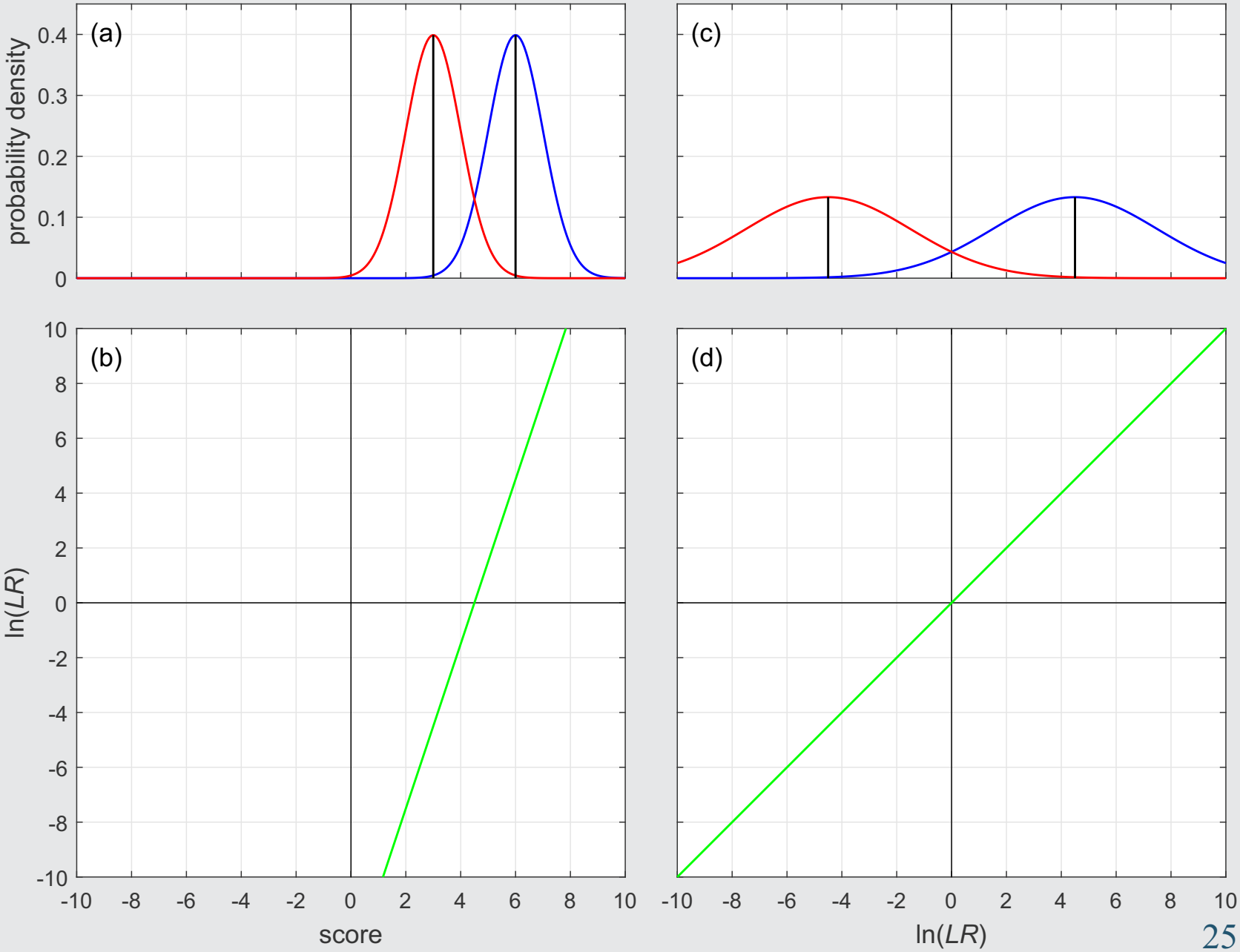
$\mu_d = -4.5$

$\mu_s = +4.5$

$\sigma = 3$

(d)

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



Calibrating likelihood-ratio systems

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

$$y = a + bx$$

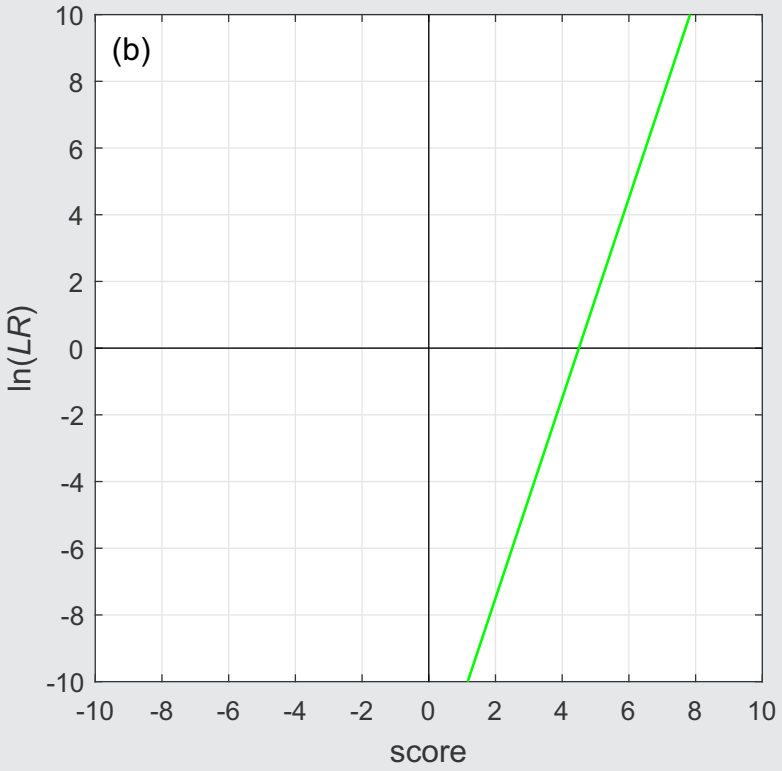
$$a = -b\frac{\mu_s + \mu_d}{2} \qquad b = \frac{\mu_s - \mu_d}{\sigma^2}$$

- Where μ_s, μ_d, σ are the statistics for the scores.

Calibrating likelihood-ratio systems

- Score [x] to ln(LR) [y]

  mapping function:



$$y = a + bx$$

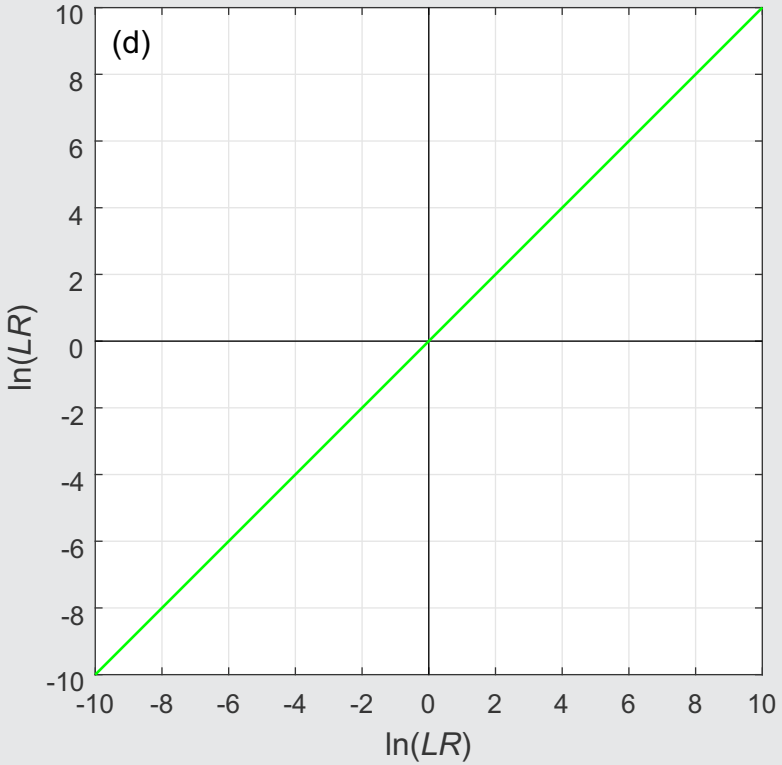
$$a = -b\frac{6+3}{2} \qquad b = \frac{6-3}{1^2}$$

$$a = -3 \times 4.5 \qquad b = 3$$

Calibrating likelihood-ratio systems

- $\ln(LR)$ [x] to $\ln(LR)$ [y]

  mapping function:



$$y = a + bx$$

$$a = -\,b\,\frac{4.5 + (-4.5)}{2} \qquad b = \frac{4.5 - (-4.5)}{3^2}$$

$$a = 0 \qquad b = 1$$

Calibrating likelihood-ratio systems

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

$$y = a + bx$$

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

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

Consensus on validation

- 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. [Carlström Plaza F., González-Rodríguez J., Ramos D., Roberts P., Rose P., Solewicz Y., Vergeer P.] (2021). **Consensus on validation of forensic voice comparison**. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Consensus on validation

- “In order for the forensic-voice-comparison system to answer the specific question formed by the propositions in the case, the output of the system should be well calibrated.”

- “A forensic-voice-comparison system should be calibrated using a statistical model that forms the final stage of the system”

Consensus on validation

- "Data used for training the calibration model ... should be sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, that, when the system is used to compare the questioned- and known-speaker recordings, the resulting likelihood ratio will be a reasonable answer to the question posed by the propositions."

Development of evaluative opinions

- Forensic Science Regulator (2021). **Codes of practice and conduct: Development of evaluative opinions** (FSR-C-118 Issue 1). Birmingham, UK: Forensic Science Regulator. https://www.gov.uk/government/publications/development-of-evaluative-opinions

Development of evaluative opinions

"probabilities have been assigned on the basis of a data set of sufficient relevance, quality and size"

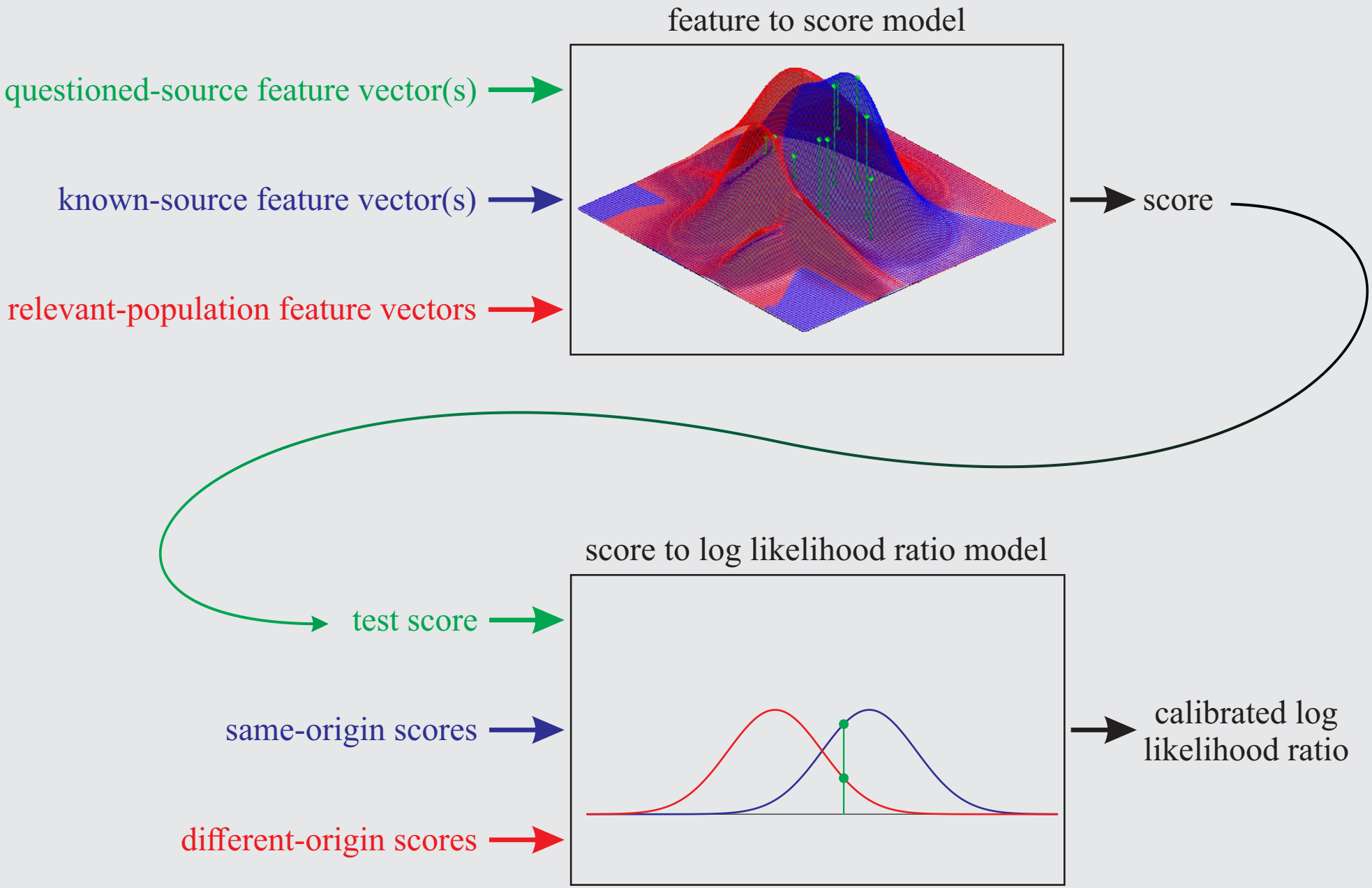
"probabilities have been assigned on the basis of structured data set(s) which are limited in their relevance, quality and/or size but are available for inspection by another expert"

"probabilities have been assigned on the basis of unstructured observations from experience, which are not available for inspection by another expert"

Development of evaluative opinions

"probabilities have been assigned on the basis of a data set of sufficient relevance, quality and size"

Development of evaluative opinions



Development of evaluative opinions

"probabilities have been assigned on the basis of structured data set(s) which are limited in their relevance, quality and/or size but are available for inspection by another expert"

Development of evaluative opinions

- "The validity of a structured data set (including any local data set) from previous casework, a 'knowledge base' ..., shall be calibrated regularly by conducting studies using ground truth data as described by Evett [22]."

- "Knowledge Base

  A structured database of information and assigned probabilities, ordered according to casework conditions. The knowledge base is calibrated through regular review of its content through experimentation under controlled conditions [22]."

Development of evaluative opinions

- "Calibration involves regular review of sections of the content by conducting experimentation using ground truth data under controlled conditions and comparing to relevant sections of the knowledge base.

- Such ground truth experimentation enables the knowledge base to be updated and expert opinions to be checked against a snapshot of known-source data."

Development of evaluative opinions

- [22] Evett I.W. (2015). **The logical foundations of forensic science: Towards reliable knowledge**. *Philosophical Transactions of the Royal Society B*, 370, article 20140263. http://dx.doi.org/10.1098/rstb.2014.0263

- This is a high-level review paper.

- It does not provide detail about how to implement:

  – calibration of a knowledge base

  – use of a knowledge base to assign probabilities in the context of a case

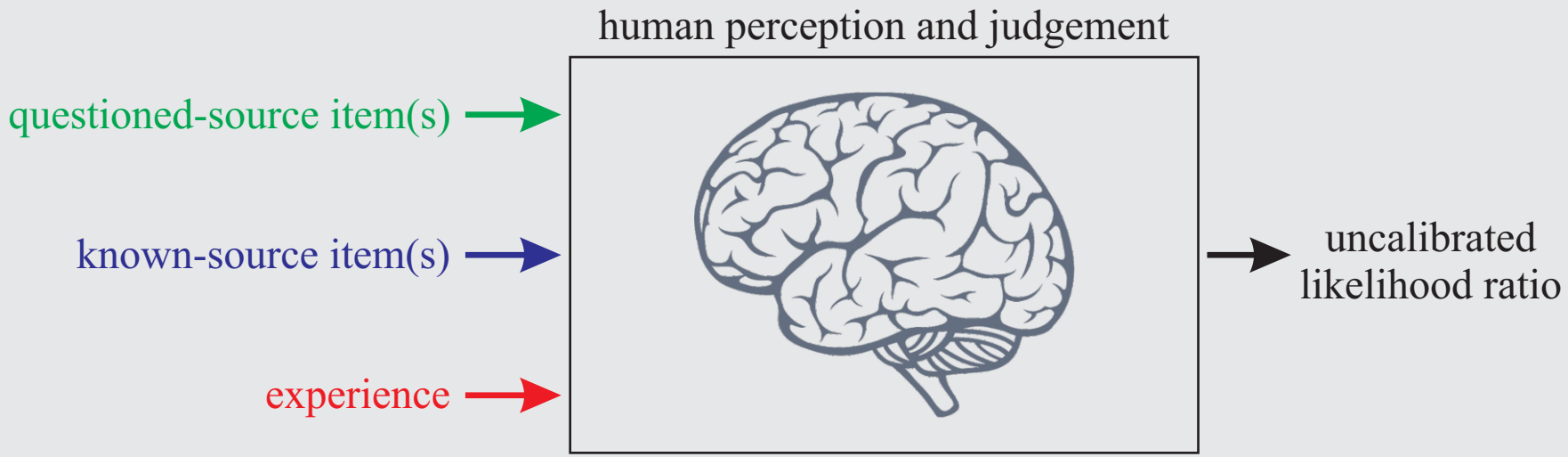
Development of evaluative opinions

"probabilities have been assigned on the basis of unstructured observations from experience, which are not available for inspection by another expert"

Development of evaluative opinions

- “In instances where an expert is unable to demonstrate any ... calibration of their expertise, the commissioning party and the court shall be made aware that their opinion is uncalibrated. ”

Development of evaluative opinions



ipse dixt

Development of evaluative opinions

- "Where the expert does not have relevant and robust experimental data to inform probabilities, they may have sufficient personal experience and knowledge to enable them to compare the relative frequencies of their observations given that each of the propositions were true. **The manner in which an expert should justify the use of observations from experience is by regular calibration of their expertise**"

Development of evaluative opinions

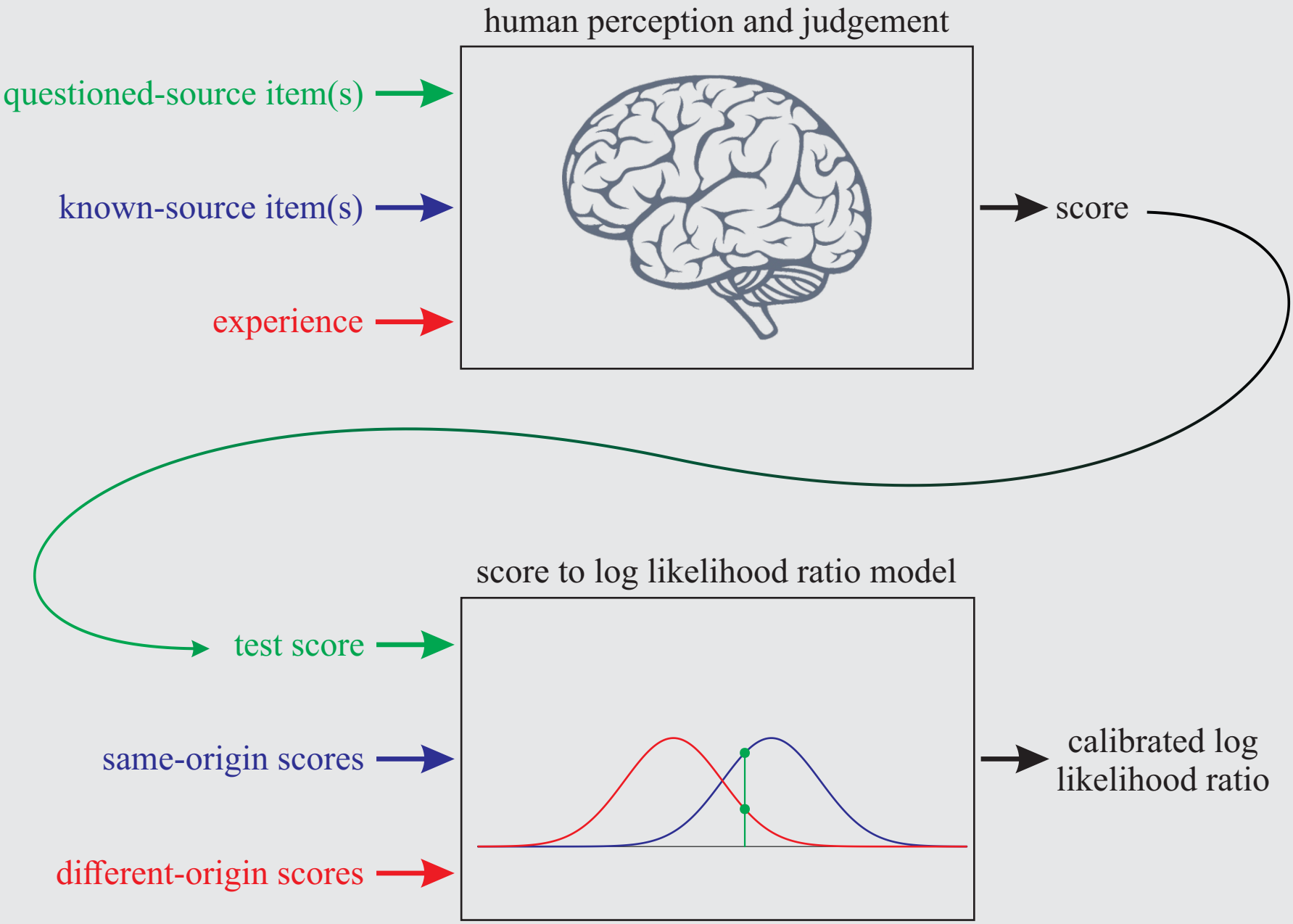
- "Experts should participate in regular calibration of their expertise [22] [23] through, for example, **blind proficiency tests that are representative of the complexity encountered in casework**."

Development of evaluative opinions

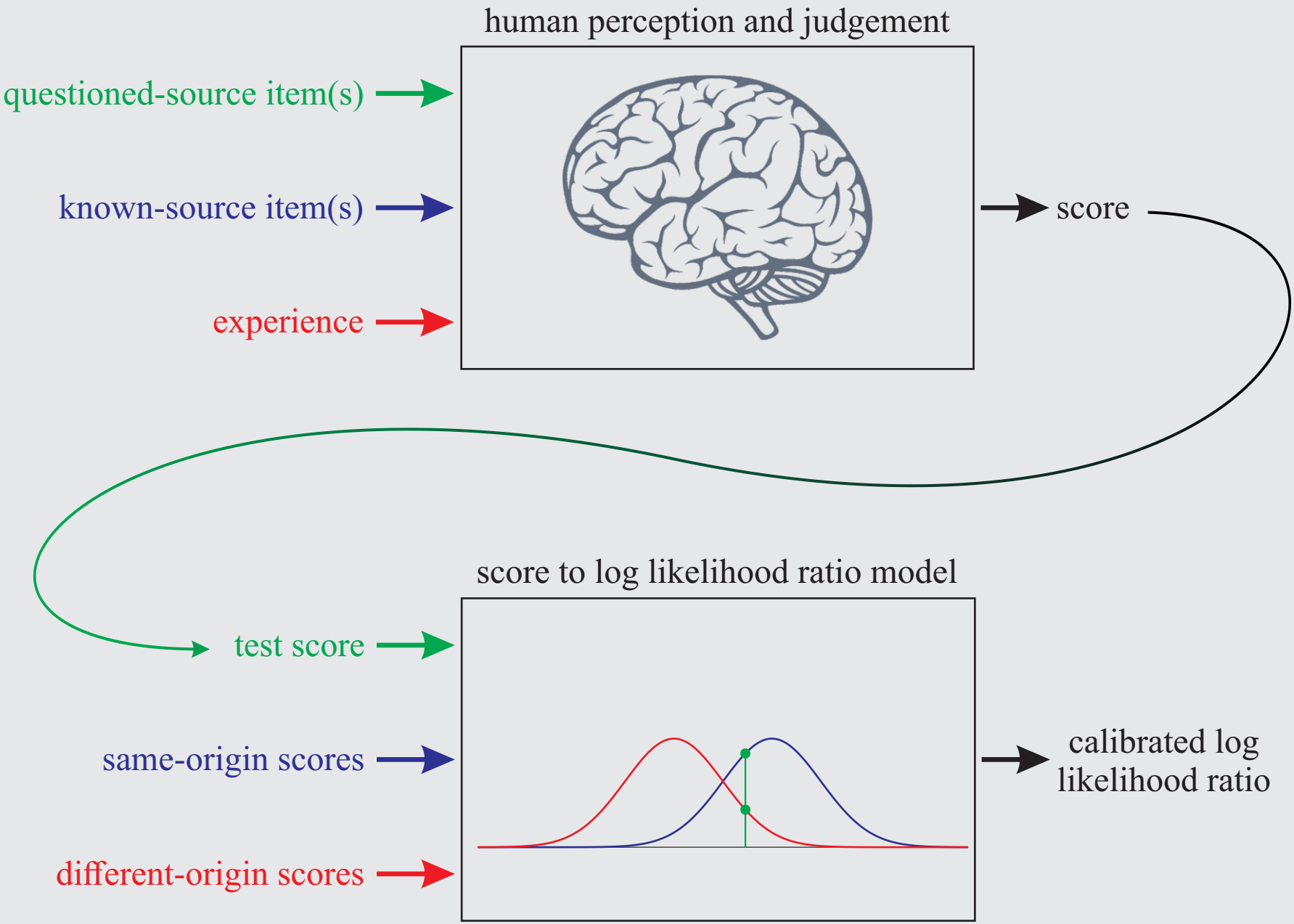
- [22] Evett I.W. (2015). **The logical foundations of forensic science: Towards reliable knowledge**. *Philosophical Transactions of the Royal Society B*, 370, article 20140263. http://dx.doi.org/10.1098/rstb.2014.0263

- "subjective assignments of probability are central to the forensic science paradigm but the driving principle for progress is that they should be conditioned not by casework experience, but by calibration under controlled conditions."

Development of evaluative opinions



Development of evaluative opinions



Thank You