



**ISHI 2020 Validation Workshop**  
**Friday September 18th, 2020 // 9:00 am - 12:30 pm**

# **Validation**

## **Principles, Practices, Parameters, Performance Evaluations, and Protocols**

**John M. Butler and Hari K. Iyer**

National Institute of Standards and Technology



# Disclaimers

**Points of view are those of the presenter** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

## **Identification does not imply endorsement**

Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

# Acknowledgments

I acknowledge many meaningful discussions on validation over the years with Dave Duewer, Margaret Kline, Robin Cotton, Catherine Grgicak, and Charlotte Word

I appreciate input from four practitioners (**Teresa Cheromcha, Kristy Kadash, Kate Philpott, and Janel Smith**) on ideas of what would be helpful in this workshop

I appreciate input on our DNA mixture interpretation scientific foundation review from **Rich Cavanagh, Mike Coble, Katherine Gettings, Hari Iyer, John Paul Jones, Willie May, Niki Osborne, Rich Press, Robert Ramotowski, Sarah Riman, Shyam Sunder, Melissa Taylor, Pete Vallone, and Sheila Willis – and a 13-member DNA Mixture Resource Group (met with the NIST team 12 times from Dec 2017 to June 2019)**

**This work is Congressionally-funded  
through the NIST Special Programs Office**

# Workshop Outline

Topics	Presenter
<b>Module 1:</b> Introduction, Guidance Documents, and Terminology	John
<b>Module 2:</b> Reliability Assessment of LR Systems: General Concepts	Hari
<b>BREAK</b>	---
<b>Module 3:</b> Validation Plans and Experimental Design	John
<b>Module 4:</b> Reliability Assessment of LR Systems: Data Examples	Hari
<b>Module 5:</b> Summarizing, Using, and Communicating Validation Data	John

**A Live  
Q&A is  
Planned  
After  
Each  
Module**



**ISHI 2020 Validation Workshop**  
**Friday September 18th, 2020 // 9:00 am - 12:30 pm**

Validation Principles, Practices, Parameters,  
Performance Evaluations, and Protocols  
**Introduction, Guidance Documents,  
& Terminology**

**Module 1**

**John M. Butler**

National Institute of Standards and Technology



## Module 1 (John)




- Introduction
  - Why this workshop? Why now?
  - Our previous experience
  - Input received for this workshop
- Available Guidance Documents on Validation
  - FBI QAS & SWGDAM
  - Other groups: OSAC/ASB, ANAB, ISO, ILAC-G19, ISFG, UKFSR, NIFS, ENFSI
- Terminology
  - Validation & internal validation
  - Issues with “validated” (when used in a binary sense)
  - Reliability (to be covered by Hari in Module 2)

# Introduction

# My Interest in Validation Grew at ISHI 16 Years Ago

I gave a talk at ISHI in October 2004 ([https://strbase.nist.gov/pub\\_pres/PromegaTalkOct2004.pdf](https://strbase.nist.gov/pub_pres/PromegaTalkOct2004.pdf))




**Can the Validation Process  
in Forensic DNA Typing  
Be Standardized?**

John M. Butler<sup>1</sup>, Christine S. Tomsey<sup>2</sup>, Margaret C. Kline<sup>1</sup>

*Proceedings of the 15<sup>th</sup> International Symposium on Human Identification. Available at [https://strbase.nist.gov/pub\\_pres/PromegaPaperOct2004.pdf](https://strbase.nist.gov/pub_pres/PromegaPaperOct2004.pdf) and <https://promega.media/-/media/files/resources/conference-proceedings/ishi-15/oral-presentations/butler.pdf?la=en>*

**STRBase Site:** <https://strbase.nist.gov/validation.htm>



15<sup>th</sup> Internat

PROFILES IN DNA

<https://www.promega.com/resources/profiles-in-dna/2006/debunking-some-urban-legends-surrounding-validation-within-the-forensic-dna-community/>

**Debunking Some Urban Legends Surrounding  
Validation Within the Forensic DNA Community**

By John Butler  
National Institute of Standards and Technology, Gaithersburg, Maryland, USA

September 2006  
*Profiles in DNA* **9(2)**, 3-6



## D.N.A. BOX 7.2

## REVIEW OF URBAN LEGENDS

In September 2006 I published an article reviewing eight 'urban legends' surrounding validation (Butler 2006). The urban legends discussed in this article included the following:

1. Hundreds or thousands of samples are required to fully validate an instrument or method.
2. Validation is uniformly performed throughout the community.
3. Each component of a DNA test or process must be validated separately.
4. Validation should seek to understand everything that could potentially go wrong with an instrument or technique.
5. Learning the technique and training other analysts are part of validation.
6. Validation is boring and should be performed by summer interns since it is beneath the dignity of a qualified analyst.
7. Documenting validation is difficult and should be extensive.
8. Once a validation study is completed you never have to revisit it.

As technology advances and new methods are developed, there will always be something to validate in a laboratory. A primary purpose is writing the *Urban Legends* article was to help analysts appreciate that validation requires common sense and is best performed (where possible) with

well-characterized samples through concordance to results produced from previous methods. Some aspects of validation can be achieved with a minimal amount of DNA samples while other aspects will require more extensive studies. In November 2010, the European Network of Forensic Science Institutes (ENFSI) DNA Working Group QA/QC subcommittee released a document building on the *Urban Legends* article and provided more detail on various aspects of the DNA typing process (ENFSI 2010).

Treating validation as a one-time event that is performed by a single individual (perhaps a summer intern who leaves the lab after performing the measurements) can lead to problems. Every analyst that is interpreting DNA typing data should be familiar with and understand the validation studies that hopefully underpin the laboratory's standard operating procedures. Validation defines the scope of a technique and thus its limitations. Making measurements around the edges of what works well will help better define the reliable boundaries of the technique. While developmental validation may be broadly applicable, internal validation is not transferrable in the same way. The performance characteristics and limitations of an instrument, a software program, and a DNA typing assay are important to understand in order to effectively interpret forensic DNA data.

ENFSI DNA Working Group (2010). Recommended minimum criteria for the validation of various aspects of the DNA profiling process. Available at <http://www.enfsi.eu>.

**Sources:**

Butler, J.M. (2006). *Debunking some urban legends surrounding validation within the forensic DNA community*. Profiles in DNA, 9(2), 3-5. Available at <http://www.promega.com/profiles/>.

# My Comments on My Urban Legends

“Treating validation as a one-time event that is performed by a single individual (perhaps a summer intern who leaves the lab after performing the measurements) can lead to problems. **Every analyst that is interpreting DNA typing data should be familiar with and understand the validation studies that hopefully underpin the laboratory's standard operating procedures.** Validation defines the scope of a technique and thus its limitations. Making measurements around the edges of what works well will help better define the reliable boundaries of the technique. While developmental validation may be broadly applicable, internal validation is not transferrable in the same way.”

“**The performance characteristics and limitations of an instrument, a software program, and a DNA typing assay are important to understand in order to effectively interpret forensic DNA data.**”

# STRBase Validation Site:

<https://strbase.nist.gov/validation.htm>

**NIST** National Institute of  
Standards and Technology  
U.S. Department of Commerce



## Validation Information to Aid Forensic DNA Laboratories

[[Presentation at Promega 2004 meeting](#)] [[Promega 2004 meeting publication](#)]  
[[Questionnaire used](#)]

[President's DNA Initiative Validation Workshop Materials](#) for workshop held at  
NFSTC August 24-26, 2005

[Validation Workshop \(208 slides\)](#) presented for Applied Biosystems' HID  
University/Future Trends in Forensic DNA Technology in Albany, NY, May 10, 2006

*To provide information or suggest improvements to this section of STRBase, please  
contact John Butler <[john.butler@nist.gov](mailto:john.butler@nist.gov)>.*

---

[[Explanation of Validation](#)] [[Standards/Guidelines](#)] [[Validation Summary Sheets](#)]  
[[Internal Validation Studies](#)] [[Helpful Information](#)] [[Literature Summary](#)]

**This website was  
initially created to  
support the 2004  
SWGDM Revised  
Validation Guidelines  
(the website is out of date  
and needs updating)**

# My Motivation for Doing This Validation Workshop

1. Growth and changes in the field in the past 15 years
  - My Urban Legends article needs revamping (I have seen it misused to oversimplify the purpose and process of validation)
  - Study of terminology as part of OSAC and NIST scientific foundation reviews
  - NIST is planning a workshop **Validation in Forensic Science** for June 2021
2. Review of literature on DNA mixture interpretation and PGS
  - Need for more information to help forensic DNA analysts and TLs strengthen their work
  - Desire to have information available for review to assess the degree of reliability of PGS systems – defense challenges and admissibility hearings have increased in recent years
3. Chapters in new DNA books sparked interest in revisiting validation
  - Bright & Coble (2020) Chapter 8 “Considerations on Validation of Probabilistic Genotyping Software”
  - Gill et al. (2020) Chapter 9 “Validation”

# Previous Workshops/Webinars on Validation (1)

ISHI 2004 Talk: [https://strbase.nist.gov/pub\\_pres/PromegaTalkOct2004.pdf](https://strbase.nist.gov/pub_pres/PromegaTalkOct2004.pdf)

Created STRBase Validation Page: <https://strbase.nist.gov/validation.htm>

1. Workshop filmed at NFSTC (Aug 24-26, 2005) *with Robyn Ragsdale*
  - <https://strbase.nist.gov/validation/validationworkshop.htm>
2. AAFS 2006 (Feb 20, 2006) and Massachusetts State Police Crime Laboratory (Apr 27-28, 2006) *with Bruce McCord*
  - [https://strbase.nist.gov/pub\\_pres/AAFS2006\\_validation.pdf](https://strbase.nist.gov/pub_pres/AAFS2006_validation.pdf)
3. HID University (May 10, 2006)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWorkshop\\_May2006.pdf](https://strbase.nist.gov/pub_pres/ValidationWorkshop_May2006.pdf)
4. New Jersey State Police (Dec 5-6, 2006)
  - [https://strbase.nist.gov/pub\\_pres/NJSP2006\\_ValidationEssentials.pdf](https://strbase.nist.gov/pub_pres/NJSP2006_ValidationEssentials.pdf)
5. Pennsylvania State Police (June 5, 2007)
  - [https://strbase.nist.gov/pub\\_pres/PSP\\_Validation\\_June2007.pdf](https://strbase.nist.gov/pub_pres/PSP_Validation_June2007.pdf)
6. International Society for Forensic Genetics (Aug 21, 2007)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWorkshopISFG2007.pdf](https://strbase.nist.gov/pub_pres/ValidationWorkshopISFG2007.pdf)
7. Webinar for Legal Medical Service in Santiago, Chile (Aug 26, 2008)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWebinar\\_Aug2008.pdf](https://strbase.nist.gov/pub_pres/ValidationWebinar_Aug2008.pdf)

<https://strbase.nist.gov/validation/Introductions.pdf>

**NIJ** National Forensic Science Technology Center  
President's DNA Initiative - Workshops



## Validation Workshop

Robyn Ragsdale, PhD  
Florida Department of Law Enforcement (FDLE)

John M. Butler, PhD  
National Institute of Standards and Technology (NIST)

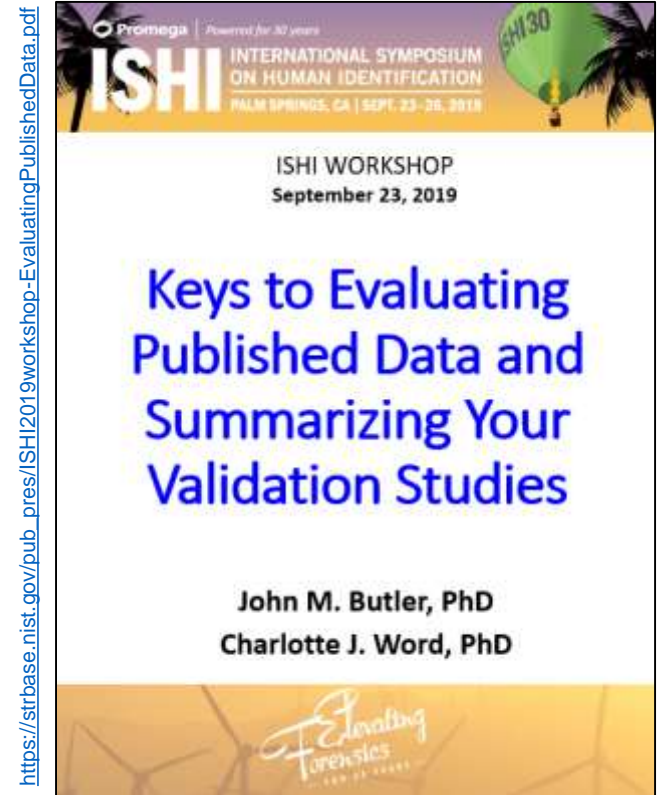


# Previous Workshops/Webinars on Validation (2)

8. ISHI 2007 Workshop (Validation: What Is It, Why Does It Matter, and How Should It Be Done?)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWorkshop\\_Promega2007.pdf](https://strbase.nist.gov/pub_pres/ValidationWorkshop_Promega2007.pdf)
9. ISHI 2009 Workshop
  - [https://strbase.nist.gov/pub\\_pres/ValidationWorkshop-Promega2009.pdf](https://strbase.nist.gov/pub_pres/ValidationWorkshop-Promega2009.pdf)
10. Florida International University *with Bruce McCord* (July 20-24, 2009)
11. ISFG Workshop *with Pete Vallone* (Sept 15, 2009)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWorkshopISFG2009.pdf](https://strbase.nist.gov/pub_pres/ValidationWorkshopISFG2009.pdf)

## CODIS Core Loci Working Group & FBI Consortium Validation Project (2010-2012)

12. NIST Mixture Webinar (April 12, 2013)
  - [https://strbase.nist.gov/training/MixtureWebcast/9\\_LowTemplateValidation-Butler.pdf](https://strbase.nist.gov/training/MixtureWebcast/9_LowTemplateValidation-Butler.pdf)
13. NIST DNA Analyst Webinar (Aug 6, 2014)
  - [https://strbase.nist.gov/pub\\_pres/ValidationWebinar-Butler-Aug2014.pdf](https://strbase.nist.gov/pub_pres/ValidationWebinar-Butler-Aug2014.pdf)
14. ISHI 2014 Workshop (Oct 2, 2014)
  - [https://strbase.nist.gov/training/ISHI2014\\_New-Loci-Kits-Workshop.htm](https://strbase.nist.gov/training/ISHI2014_New-Loci-Kits-Workshop.htm)
15. ISHI 2019 Workshop *with Charlotte Word* (Sept 23, 2019)
  - [https://strbase.nist.gov/pub\\_pres/ISHI2019workshop-EvaluatingPublishedData.pdf](https://strbase.nist.gov/pub_pres/ISHI2019workshop-EvaluatingPublishedData.pdf)



Cited in *Gissantaner Amicus* 2020 brief

# Previous Workshops/Webinars on Validation (3)

## 16. Improving Biometric and Forensic Technology: The Future of Research Datasets (Jan 26-27, 2015)

- Hari Iyer presentation on “Rule of 3 and Rule of 30” regarding experimental design
- <https://www.nist.gov/system/files/documents/forensics/Iyer-Presentation.pdf>

## 17. ISHI 2019 Mixture Workshop (Sept 26, 2019)

- Hari Iyer presentation on reliability considerations and PGS LR validation
- [https://strbase.nist.gov/pub\\_pres/ISHI2019-MixtureWorkshop.pdf](https://strbase.nist.gov/pub_pres/ISHI2019-MixtureWorkshop.pdf) (slides 38-125)



**Hari K. Iyer**  
NIST Statistical  
Engineering  
Division

### ISO/IEC 19795-1:

*Sufficient samples shall be collected per test subject so that the total number of attempts exceeds that required by the **Rule of 3** or **Rule of 30** as appropriate*

- What is the **RULE OF 3** and how is it applied when determining sample sizes?
- What is the **RULE OF 30** and how is it applied when determining sample sizes?

January 2015

### Some Factors Affecting Reliability of an LR System

1. Sample
  - a) Sample amount (contributor template amounts)
  - b) Sample quality (degradation level)
2. Labs
  - a) Kits used
  - b) Equipment Used
  - c) Number of PCR cycles
  - d) Analyst
  - e) Choice of Analytical Threshold (AT)
3. Probabilistic Genotyping (PG) Model
  - a) Choice of model
  - b) Choice of laboratory specific parameters for use in the PG model
  - c) **Propositions Chosen ( $H_p$  and  $H_d$ )**
4. Software Implementing the PG Model
  - a) Choice of numerical methods for computing LR (MCMC, Numerical Integration)
  - b) Choice of number of iterations OR numerical integration parameters (such as grid size)

**FACTOR  
SPACE**

September 2019

# Some Specific Input Received for This Workshop

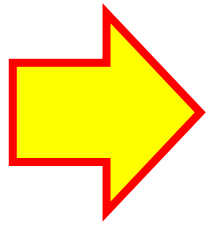
- **Teresa Cheromcha** (Colorado Bureau of Investigation-Grand Junction)
  - Assistant TL for CBI system with 5 laboratories
- **Kristy Kadash** (Jefferson County Regional Crime Laboratory, Colorado)
  - Member of SWGDAM and OSAC and former TL
- **Kate Philpott** (Adjunct Faculty/Research Analyst, VCU Forensic Science Program)
  - Legal and scientific consultant; recently co-authored the June 2020 *Gissantaner* amicus brief
- **Janel Smith** (Phoenix Police Department)
  - DNA Technical Leader for a large city laboratory; member of OSAC

**Their input  
is discussed  
in Module 3**

*I reached out to each of them and asked for **ideas of things we should cover** to best assist DNA analysts and TLs and specifically **what information on the topic of validation would be most helpful** to them in their work*

# Many Laboratory Activities Need Validation

- DNA Extraction Robotic Process
- Quantitation Kits or Assays
- New STR Kits
- CE Instruments
- Genotyping Software
- Rapid DNA Instrument
- NGS Instrument



- **Probabilistic Genotyping Software (PGS)**



**Hyperlinks to Each Document Are Included in  
the PDF Version of These Presentation Slides**

# **Guidance Documents**

# Documents that Govern and Influence DNA Operations in Accredited Forensic Laboratories

Document	Authority	Who Creates	Who Uses or Enforces
----------	-----------	-------------	----------------------

## Forensic Discipline-Specific Efforts (DNA)

<b>Quality Assurance Standards (QAS)</b> 1998/1999 updated in 2009, 2011, 2020	Law passed by Congress in 1994; issued by FBI Director	Originally DAB (1995-2000), now SWGDAM	FBI and ANAB auditors to assess U.S. forensic laboratories
<b>Guidelines &amp; Best Practices</b>	Forensic practitioner community	<b>SWGDAM</b> , ENFSI DNA WG, ISFG DNA Commission	Forensic laboratories and practitioners (not required)

## National and International Standards Groups

<b>ILAC G19 (2014) and ISO/IEC 17025 (2017)</b>	Standards community	ISO committee	Accrediting bodies (ANAB; formerly ASCLD/LAB)
<b>ANSI/ASB Standards (and OSAC Registry)</b>	SDOs with forensic practitioner community input	SDOs (ASB, ASTM) and OSAC	Accrediting bodies as they are adopted

## Country-Specific or Region-Specific Forensic Science Efforts

<b>UK Forensic Science Code of Practice</b>	UK Forensic Science Regulator	UK Forensic Science Regulator WGs	UK forensic laboratories and practitioners
<b>ENFSI</b>	European forensic laboratories	ENFSI WGs	European forensic laboratories

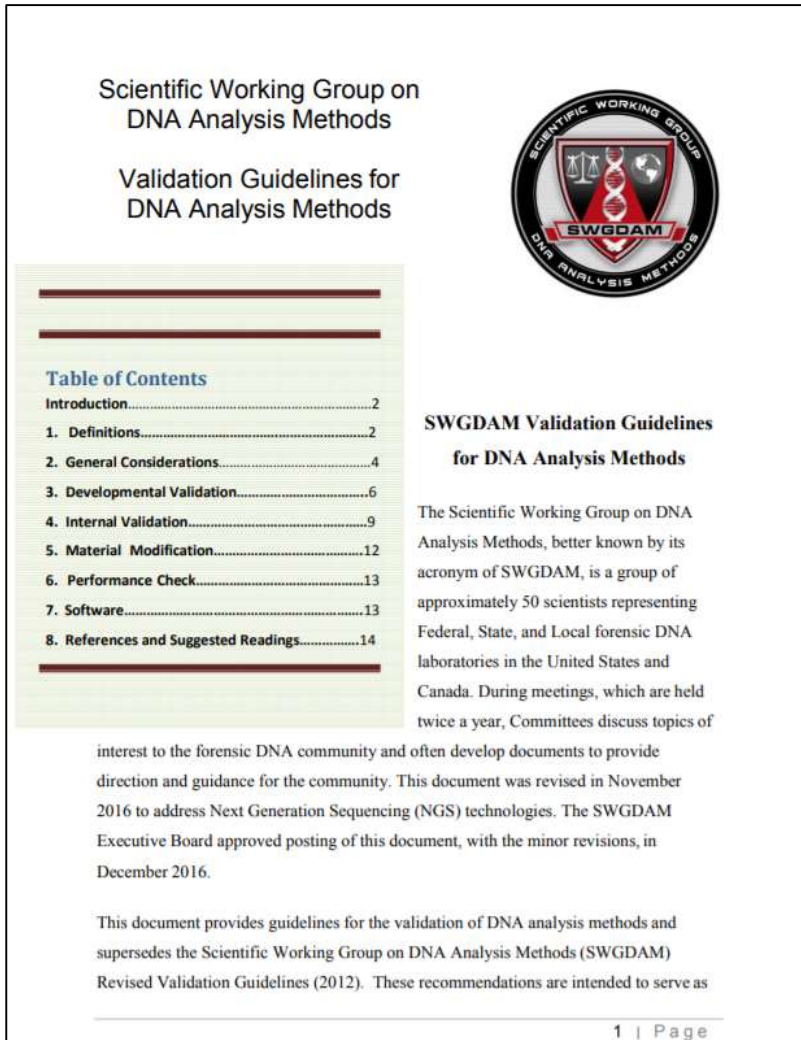
**Abbreviations Defined**

**ANAB** = ANSI National Accreditation Board  
**ANSI** = American National Standards Institute  
**ASB** = AAFS Standards Board  
**ASCLD/LAB** = American Society of Crime Laboratory Directors Laboratory Accreditation Board  
**ASTM** = American Society for Testing and Materials  
**DAB** = DNA Advisory Board  
**ENFSI** = European Network of Forensic Science Institutes  
**IEC** = International Electrotechnical Commission  
**ILAC** = International Laboratory Accreditation Cooperation  
**ISFG** = International Society for Forensic Genetics  
**ISO** = International Organization for Standardization  
**OSAC** = Organization of Scientific Area Committees for Forensic Science  
**SDO** = Standards Developing Organization  
**SWGDAM** = Scientific Working Group for DNA Analysis Methods  
**WG** = working group

# Validation Guidance Documents from Forensic Discipline-Specific Efforts (DNA)

- **FBI Quality Assurance Standards** (1998/1999, 2009, 2011, 2020)
  - [Quality Assurance Standards for Forensic DNA Testing Laboratories](#)
  - [Quality Assurance Standards for DNA Databasing Laboratories](#)
  - [Guidance Document for the FBI QAS \(effective 07/01/2020\)](#)
    - Standard 8 Validation
- **SWGDM Validation Guidelines** (2004, 2012, 2016)
  - [Validation Guidelines for DNA Analysis Methods](#)
  - Section 3: Developmental Validation
  - **Section 4: Internal Validation**
  - Section 6: Performance Check
  - Section 7: Software

# SWGDM DNA Analysis Validation Guidelines (2016)



December 2016

**Developmental Validation** shall include, where applicable:

(3.1) Characterization of genetic markers, (3.2) species specificity, (3.3) sensitivity studies, (3.4) stability studies, (3.5) precision and accuracy, (3.6) case-type samples, (3.7) population studies, (3.8) mixture studies, (3.9) PCR-based studies, (3.10) NGS-specific studies

**Internal Validation** shall include these studies:

- ✓ (4.1) Known or mock evidence samples
- ✓ (4.2) Sensitivity and stochastic studies
- ✓ (4.3.1) Precision and accuracy: repeatability
- ✓ (4.3.2) Precision and accuracy: reproducibility
- ✓ (4.4) Mixture studies
- ✓ (4.5) Contamination assessment

*(4.4) Mixed DNA samples that are **representative of those typically encountered** by the testing laboratory should be evaluated*

# Validation Guidance Documents from National and International Groups

## Standards Groups

- International Laboratory Accreditation Cooperation (2002, 2014)
  - [ILAC G19:08/2014 Modules in a Forensic Science Process](#)
- International Organization for Standardization (2005, 2017)
  - ISO/IEC 17025: 2017 General Requirements for the Competence of Testing and Calibration Laboratories (see Section 7.2.2 Validation of methods)
- ANSI/ASB/OSAC (see next slide)

## Accreditation Body

ANAB Accreditation Requirements (2017, 2019)

- [ANAB ISO/IEC 17025:2017-Forensic Science Testing and Calibration Laboratories Accreditation Requirements \(AR 3125\)](#)

## Advisory Groups

- PCAST (President's Council of Advisors on Science and Technology) (2016, 2017)
  - [Report to the President – Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods](#)
  - [An Addendum to the PCAST Report on Forensic Science in Criminal Courts](#)
- OSAC Human Factors Committee (2020)
  - [Human Factors in Validation and Performance Testing in Forensic Science](#)

# OSAC Human Factors in Validation and Performance Testing in Forensic Science

OSAC Technical Series 0004



## Human Factors in Validation and Performance Testing of Forensic Science

<https://doi.org/10.29325/OSAC.TS.0004>

OSAC Human Factors Committee

March 2020



- The research strategies discussed here are helpful for establishing the range of validity of new forensic science methods and for demonstrating the range of validity of older methods.
- Defines and explains key terms: accuracy, consistency, reliability, **sensitivity, specificity**, validity, validation, black-box and white-box studies
- Reviews some key issues in designing, conducting, and reporting validation research
  - (1) Institutional Review Board review
  - (2) Study administration general issues
  - (3) Source of test specimens: created versus casework
  - (4) Evaluating test specimens regarding suitability and level of difficulty
  - (5) Adequacy of sample size
  - ...
  - (9) How to report the results of validation studies on methods used to reach categorical results
  - (10) Special problems in assessing the accuracy of likelihood ratios
  - (11) Sharing research findings in an open, transparent manner

# DNA Validation Guidance Documents from OSAC and ANSI/ASB

## **Published by Standards Developing Organization (SDO)**



1. [ANSI/ASB Standard 020 \(2018\): Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol](#)
2. [ANSI/ASB Standard 040 \(2019\): Standard for Forensic DNA Interpretation and Comparison Protocols](#)
3. [ANSI/ASB Standard 018 \(2020\): Standard for Validation of Probabilistic Genotyping Systems](#)

## **OSAC Draft/Proposed Standards** (under development by ASB)

1. [Standard for Internal Validation of Forensic DNA Analysis Methods](#) [ASB 38]
2. [Standard for Internal Validation of Human Short Tandem Repeat Profiling on Capillary Electrophoresis Platforms](#) [ASB 39]
3. [Best Practice Recommendations for Internal Validation of Human Short Tandem Repeat Profiling on Capillary Electrophoresis Platforms](#) [ASB 129]
4. [Best Practice Recommendation for Validation of Forensic DNA Software](#) [ASB 114]

# Validation Guidance Documents from Country-Specific or Region-Specific Efforts

- Eurachem (1998, 2014)
  - [The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics](#)
- ENFSI DNA Working Group (2010)
  - [Recommended Minimum Criteria for the Validation of Various Aspects of the DNA Profiling Process](#)
- ENFSI (2006, [2014](#))
  - [Guidelines for the Single Laboratory Validation of Instrumental and Human Based Methods in Forensic Science](#)
- UK Forensic Science Regulator (2014, 2020)
  - [Codes of Practice and Conduct \(2020, FSR-C-100, Issue 5\)](#)
    - see section 21 on test methods and method validation
  - [Validation Guidance \(2014, FSR-G-201, Issue 1\)](#)
  - [Validation Protocol – Use of Casework Material \(2016, FSR-P-300, Issue 1\)](#)
- ANZPAA NIFS (Australia New Zealand Policing Advisory Agency National Institute of Forensic Science) (2019)
  - [Empirical Study Design in Forensic Science: A Guideline to Forensic Fundamentals](#)



# PGS Software Validation Guidance Documents

- **SWGDM PGS Validation (2015)**
  - [Guidelines for the Validation of Probabilistic Genotyping Systems](#)
- ISFG DNA Commission (2016)
  - [Recommendations on the validation of software programs performing biostatistical calculations for forensic genetic applications](#)
- ENFSI DNA Working Group (2017)
  - [Best Practice Manual for the Internal Validation of Probabilistic Software to Undertake DNA Mixture Interpretation](#)
- UK Forensic Science Regulator (2018)
  - [Software Validation for DNA Mixture Interpretation \(FSR-G-223\)](#)
- ANSI/ASB (2020)
  - [Standard 018: Standard for Validation of Probabilistic Genotyping Systems](#)
- **FBI Quality Assurance Standards ([2020](#))**
  - See Standard 8.8

# Some Published Articles in Peer-Reviewed Journals on PGS and Likelihood Ratio Validation

## PGS

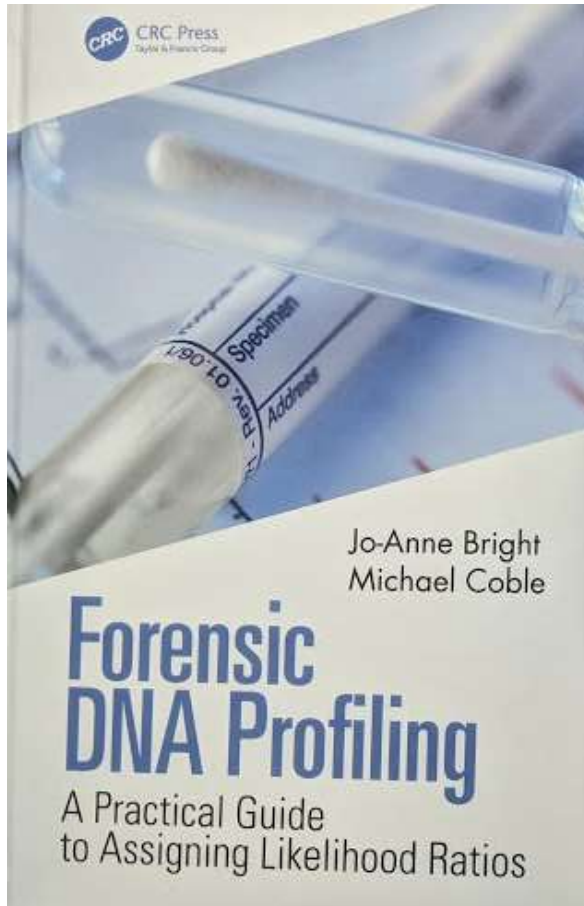
- Bright et al. 2015 (*Forensic Sci. Int. Genet.* 14:125-131)
  - [A series of recommended tests when validating probabilistic DNA profile interpretation software](#)
- Taylor et al. 2015 (*Forensic Sci. Int. Genet.* 16:165-171)
  - [Testing likelihood ratios produced from complex DNA profiles](#)
- Haned et al. 2016 (*Science & Justice* 56:104-108)
  - [Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations](#)
- Coble et al. 2016 (*Forensic Sci. Int. Genet.* 25:191-197)
  - [ISFG DNA Commission: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetic applications](#)

## LR

- Morrison 2011 (*Science & Justice* 51:91-98)
  - [Measuring the validity and reliability of forensic likelihood-ratio systems](#)
- Meuwly et al. 2017 (*Forensic Sci. Int.* 276:142-153)
  - [A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation](#)

# New Books to Assist with DNA Mixture Interpretation

CRC Press  
(January 2020)



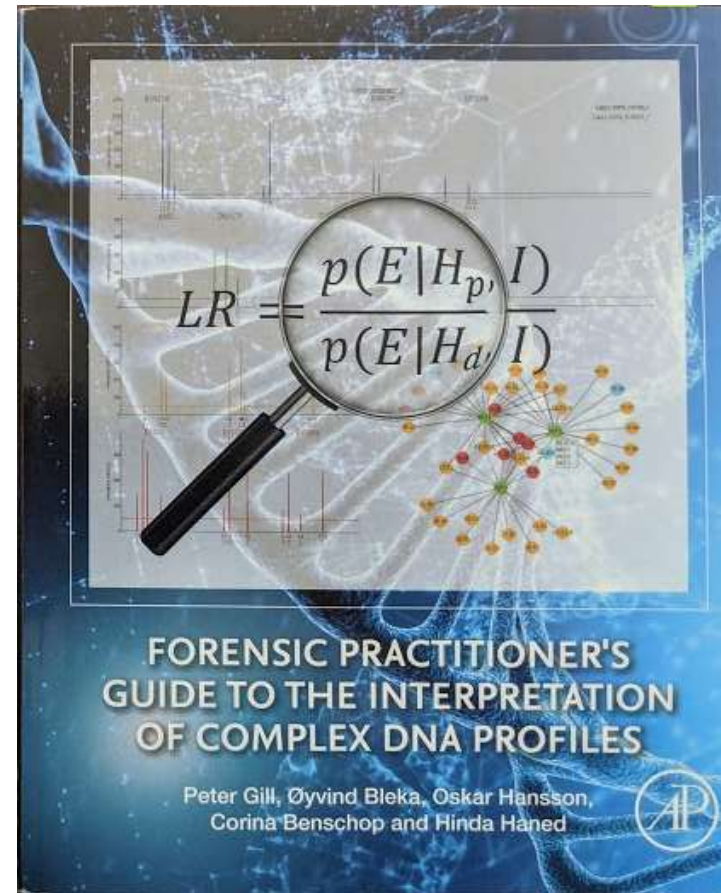
**Hardback:** 258 pages  
109 references cited

## Table of Contents

1. DNA Profiling Interpretation
2. Statistics and Proposition Setting
3. LR Single-Source Examples and Population Genetics
4. Binary LR for Mixtures
5. LRs Considering Relatives as Alternative Contributors
6. Probabilistic Genotyping: Semicontinuous Models
7. Probabilistic Genotyping: Continuous Models
8. **Considerations on Validation of PGS**

Appendix 1: Allele Frequencies  
Appendix 2: Model Answers

Elsevier Academic Press  
(June 2020)



**Paperback:** 530 pages  
362 references cited

## Table of Contents

1. Forensic Genetics Basics
2. DNA Profiles
3. Allele Drop-out
4. Low-template DNA
5. LRmix Model Theory
6. LRmix Studio
7. Continuous Model Theory
8. EuroForMix
9. **Validation**
10. DNAXs
11. *SmartRank* & *CaseSolver*
12. Interpretation & Reporting
13. Complex DNA Profiling by Massively Parallel Sequencing

Appendix A: Genotype Probabilities  
Appendix B: Probabilistic Models

# Terminology

# Some Definitions for Validation

**= Fit for Purpose**

Source	Definition of Validation
SWGDM 2016 Validation Guidelines	A process by which a procedure is evaluated to determine its <b>efficacy</b> and <b>reliability</b> for forensic casework and/or database analysis
FBI QAS 2020	A process by which a method is evaluated to determine its efficacy and reliability for forensic casework analysis and includes the following: (1) developmental validation... and (2) internal validation...
ISO/IEC 17025:2017	<b>Verification</b> , where the specified requirements are adequate for intended use [verification: provision of objective evidence that a given item fulfils specified requirements]
ILAC G19 (2014)	Validation is the <b>confirmation by examination</b> and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled
OSAC Lexicon ( <a href="http://lexicon.forensic-osac.org/">http://lexicon.forensic-osac.org/</a> ) one of the 14 definitions supplied	The process of performing and evaluating <b>a set of experiments that establish the efficacy, reliability, and limitations of a method</b> , procedure or modification thereof; establishing recorded documentation that provides a high degree of assurance that a specific process will consistently produce an outcome meeting its predetermined specifications and quality attributes. May include developmental and/or internal validation.

From Oxford Dictionary

<https://www.lexico.com/en/definition>

**Efficacy:** the ability to produce a desired or intended result

**Reliability:** the quality of being trustworthy or of performing consistently well; the degree to which the result of a measurement, calculation, or specification can be depended on to be accurate

# Some Definitions for Internal Validation

Source	Definition of Internal Validation
SWGDM 2016 DNA Validation Guidelines	An <b>accumulation of test data within the laboratory</b> to demonstrate that established methods and procedures <b>perform as expected in the laboratory</b>
FBI QAS 2020	An accumulation of test data within the laboratory to demonstrate that established methods and procedures <b>perform as expected in the laboratory</b>
SWGDM 2015 PGS Validation Guidelines	The accumulation of test data within the laboratory to demonstrate that the established <b>parameters, software settings, formulae, algorithms and functions perform as expected</b>
ASB018 Standard for Validation of PGS	The acquisition of test data within the laboratory <b>to verify the functionality of the system, the accuracy of statistical parameters, the appropriateness of analytical and statistical parameters, and the determination of limitations of the system</b>
ISFG DNA Commission (Coble et al. 2016)	Empirical studies performed either within a laboratory or outsourced to a third-party entity <b>to ensure that the software runs properly</b> within the relevant laboratory

What does it mean to “perform as expected”?

An expectation is set during developmental validation studies

## Users Decide When Sufficient Data Have Been Collected

- Validation studies/experiments performed in a laboratory provide information to make assessments regarding the degree of reliability for a specified method
- These studies are concluded and deemed sufficient when those performing them have *convinced themselves* that the results obtained are reliable for their application
  - In other words, when the intended users are happy with how things work compared with how they plan to use them
  - A determination of whether the amount and type of data available is satisfactory or sufficient to the user of the information is something that **must be decided by the user of the information not the provider.**

# Information Provider and User

## *Responsibilities and Examples*

**Provider**  **User**

Responsibilities	<b>Provides accessible data</b> to be used for assessment by the user and <b>explains relevance</b> and significance	<b>Determines validity</b> (whether method is fit-for-purpose) and <b>assesses degree of reliability and makes decision whether sufficient information exists</b> for the intended application
Example 1	<b>Product developer</b> of software or instrument	<b>Product user</b> of software or instrument ( <b>forensic scientist</b> )
Example 2	<b>Expert witness</b> providing testimony ( <b>forensic scientist</b> )	<b>Judge and lawyers</b> in a trial or admissibility hearing using provided testimony
Example 3	Documentary standard developer	Standard user, who makes it “regulatory” when adopting it



# Validation Studies Conducted vs. a “Validated” Method

- Guidance documents on validation in forensic science are typically **focused on types of tests to perform** in gathering the data rather than ways to assess the data.
- In our opinion, it is unwise to describe a method as “validated” in a generic fashion without some **context around the method’s use** and access to any underpinning data to support claims of validity and reliability for those who would like to independently review them



# Are We on the Right Side of the Equation?

*Systems Thinking is Looking at the Big Picture and How Inputs Impact Outputs...*



Component(s) + Process(es) = **Outcome**

*What?*

*How?*

*How well?*

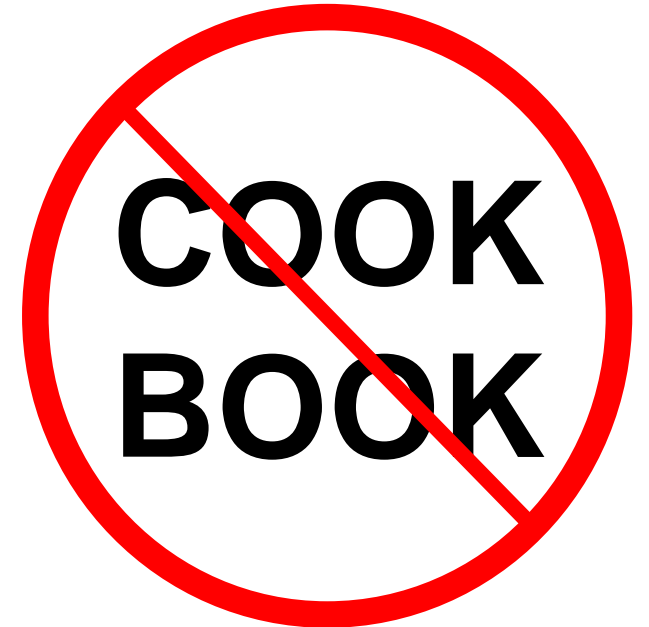
*So what?*

# Our Goal for This Workshop

To Review Important Principles to  
Aid Understanding of Validation...

## Key Aspects of Validation:

- How to **Design** Validation Studies
- How to **Perform** Validation Studies
- How to **Describe** Validation Studies
- How to **Utilize** Validation Data



In Module 2, Hari will discuss reliability and conceptual approaches to assessing the degree of reliability with LR results produced by PGS

# Thank you for your attention!

**Points of view are the presenters** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

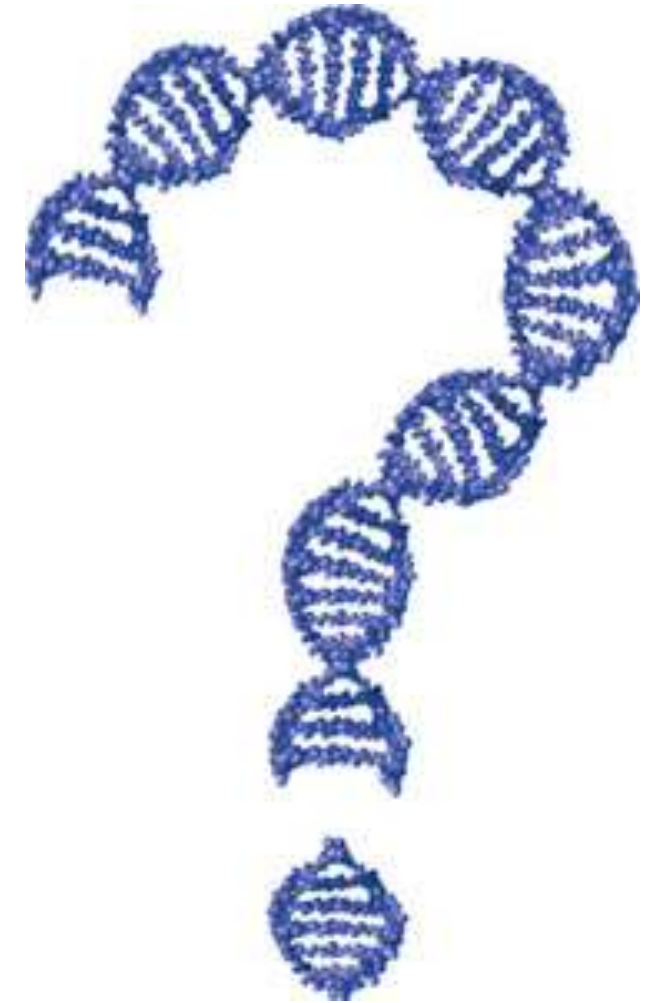
## Contact Information

**John M. Butler**

[john.butler@nist.gov](mailto:john.butler@nist.gov)

**Hari K. Iyer**

[hariharan.iyer@nist.gov](mailto:hariharan.iyer@nist.gov)



RESEARCH. STANDARDS. FOUNDATIONS.



**ISHI 2020 Validation Workshop**  
**Friday September 18th, 2020 // 9:00 am - 12:30 pm**

**Validation Principles, Practices, Parameters,  
Performance Evaluations, and Protocols**

**Reliability Assessment of  
LR Systems: General Concepts**

**Module 2**

**Hari K. Iyer**

National Institute of Standards and Technology



# Acknowledgments & Disclaimers

I would like to thank Steve Lund, William Guthrie, Antonio Possolo, Adam Pintar, Jan Hannig, Marty Herman, and other NIST colleagues, for many ongoing, meaningful discussions on foundational concepts in statistics.

I wish to also thank John Butler, Katherine Gettings, Niki Osborne, Rich Press, Sarah Riman, Melissa Taylor, Pete Vallone, and Sheila Willis – and the DNA Mixture Resource Group, for valuable discussions on Fundamentals of DNA mixture interpretation and related issues.

**Points of view are of the presenters** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

# Discussion Topics

- What is “Reliability”?
- How is reliability demonstrated/judged?
- Terms associated with Reliability:  
Accuracy, Precision, Repeatability, Reproducibility,  
Uncertainty, Error
- System Reliability vs Component Reliability
- Main criteria for Reliability:  
Discrimination power and Calibration Accuracy
- Introduction to Discrimination/Calibration concepts
- Summary

# Reliability

**Reliability**  
/re-ly-a-bi-li-ti/

1. To be able to produce good results time after time.
2. How much a person can be depended on.



# Reliability

The Cambridge Dictionary describes “Reliability” as “how **accurate** or able to be **trusted** someone or something is considered to be.”

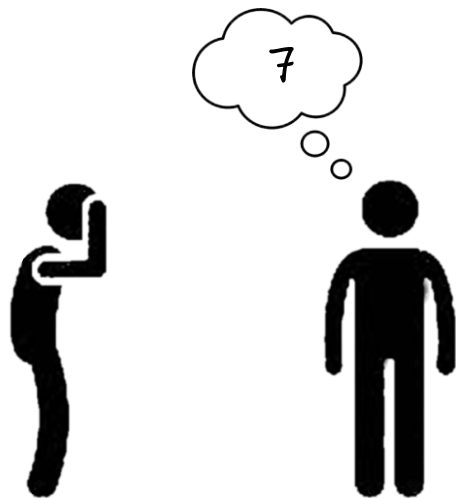
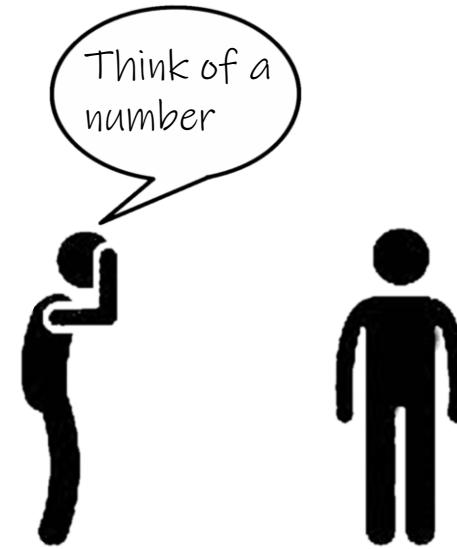
[<https://dictionary.cambridge.org/us/dictionary/english/reliability>]

# Trust

Trust can arise in several ways:

- Logic
- Empirical demonstrations of a claim in ground truth known situations; Making predictions and verifying if the predictions come true
- Belief (in another person's opinions, e.g. expert)

# Being Convinced is a Personal Matter



# Role of Science

- Absolute truth is difficult or impossible to establish but one can be “convinced” that something is true based on a combination of the above modes of forming trust.
- Each individual has his/her own thought processes involving combination of empirical knowledge with intuition and belief that lead him/her to form a degree of acceptance of a claim.
- To what extent one is convinced of the truth of a claim is a personal matter.

Science attempts to minimize the level of belief one needs to accept a claim by providing empirical demonstrations of the extent to which the claim is “correct”.

# Reliability vs Consistency

The plain English meaning of the word 'reliability' is 'trustworthiness'.  
This is the sense in which we use this term here.

In the fields of psychology and sociology the term **Reliability** is used to describe **Consistency**. This has led to much confusion.

Reliability implies consistency

**But consistency alone does not imply reliability**

**Reliability requires being consistently accurate**

# Judgements of Reliability

- **A Method is RELIABLE if it produces ‘good’ results time after time.**
- What is meant by ‘good’? Rather than give binary answers (reliable or not reliable) or personal assessments (method has a high degree of reliability) what we require are FACTS and DATA.
- **Personal Assessment:** “this surgical procedure has an excellent track record of being successful”.
- **Facts&Data:** “90 out of 100 patients who underwent this type of surgery survived and lived for at least 5 more years. The other 10 died on the operating table.”

Judgements of reliable/unreliable are personal.  
But facts and data are not personal.

# Terms Related to Reliability

- **Accuracy**
- Precision
- Repeatability
- Reproducibility
- Uncertainty
- Error

# Accuracy

- **Accuracy**
- Precision
- Repeatability
- Reproducibility
- Uncertainty
- Error

Accuracy: 'how close is the result to the true value?'  
or 'how often does this procedure lead to correct decisions (desired outcomes) or conclusions?'

Inaccuracy: 'how far is the result from the true value?'

True value can be an elusive quantity.

Usually substituted with 'highly trusted reference value'.  
[Standard Reference Materials (SRMs): values from NIST 😊 ]

Or a 'consensus value' based on various authoritative national metrology labs.



# Precision

- Accuracy
- **Precision**
- Repeatability
- Reproducibility
- Uncertainty
- Error

**Precision:** 'To what extent do repeated measurements of the 'same' quantity agree with one another?'

**Imprecision:** 'To what extent do repeated measurements of the 'same' quantity disagree with one another?'

When repeated measurements give different values (there is measurement variability) we can all see that the process does not produce perfectly accurate results. So the focus shifts to

- How variable are the different measurements of the same quantity?

# Repeatability/Reproducibility

**Repeatability** and **Reproducibility** explore the extent to which measurements of the 'same' quantity differ under varying conditions.

# Uncertainty

- Accuracy
- Precision
- Repeatability
- Reproducibility
- **Uncertainty**
- Error

# Measurement Uncertainty

NIST Technical Note 1900

## Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results

Antonio Possolo

- **Measurement uncertainty** is the doubt about the true value of the measurand that remains after making a measurement.
- Measurement uncertainty is described fully and quantitatively by a probability distribution on the set of values of the measurand.
- At a minimum, it may be described summarily and approximately by a quantitative indication of the dispersion (or scatter) of such distribution.

# Uncertainty

- Accuracy
- Precision
- Repeatability
- Reproducibility
- **Uncertainty**
- Error

Uncertainty is the doubt regarding the underlying truth that remains after considering all available relevant information.

# Error

- Accuracy
- Precision
- Repeatability
- Reproducibility
- Uncertainty
- **Error**

Conventional meaning: **Mistake**

Statistical usage: Difference between offered result and 'truth' or an authoritative 'reference value'

# Reliability: Models vs Empirical Data



If you toss this quarter twice, what is the probability that both tosses will give 'HEADS' ?

# Reliability: Models vs Empirical Data

- There are 4 possible outcomes:  
(Tail, Tail), (Tail, Head), (Head, Tail), (Head, Head).
- Only one of the 4 outcomes is what we want.
- Assuming, all 4 outcomes are equally likely,

The **probability of getting both heads** in two tosses of the coin must be  $\frac{1}{4}$ .

EXPERIMENT: A coin is tossed two times and the number of 'heads' is recorded (0 or 1 or 2). The experiment is repeated 1000 times. **Based on our "model" the expected frequencies are as follows:**

	Tail, Tail	Tail, Head	Head, Tail	Head, Head	TOTAL
EXPECTED	250	250	250	250	1000



# Reliability: Models vs Empirical Data

- There are 4 possible outcomes:  
(Tail, Tail), (Tail, Head), (Head, Tail), (Head, Head).
- Only one of the 4 outcomes is what we want.
- Assuming, all 4 outcomes are equally likely,

The probability of getting both heads in two tosses of the coin must be  $\frac{1}{4}$ .

EXPERIMENT: A coin is tossed two times and the number of 'heads' is recorded (0 or 1 or 2). The experiment is repeated 1000 times. But suppose the observed frequencies are very different !

	Tail, Tail	Tail, Head	Head, Tail	Head, Head	TOTAL
EXPECTED	250	250	250	250	1000
OBSERVED	360	237	243	160	1000



“It doesn’t matter how beautiful your theory is, it doesn’t matter how smart you are. If it doesn’t agree with experiment, it’s wrong.”

**Richard P. Feynman**

**Nobel Laureate, 1965**

*Quantum Electrodynamics & Physics of Elementary Particles*

# Federal Rules of Evidence 702 (FRE 702)

## Rule 702. Testimony by Expert Witnesses

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

- (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
- (b) the testimony is based on **sufficient facts** or **data**;
- (c) the testimony is the product of **reliable** principles and methods; and
- (d) the expert has **reliably** applied the principles and methods to the facts of the case.

[https://www.law.cornell.edu/rules/fre/rule\\_702](https://www.law.cornell.edu/rules/fre/rule_702)

# Daubert

Rule 702 has been amended in response to *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993

- In *Daubert* the Court charged trial judges with the responsibility of acting as gatekeepers to **exclude unreliable expert testimony**, and
- the Court in *Kumho* clarified that this gatekeeper function applies to all expert testimony, not just testimony based in science (*Kumho Tire Co. v. Carmichael*, 1999)

[https://www.law.cornell.edu/rules/fre/rule\\_702](https://www.law.cornell.edu/rules/fre/rule_702)

# Daubert “checklist”

*Daubert* set forth a non-exclusive [*non-exhaustive?*] checklist for trial courts to use in assessing the **reliability of scientific expert testimony**.

The specific factors explicated by the *Daubert* Court are

- 1) whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability;
- 2) whether the technique or theory has been subject to peer review and publication;
- 3) the **known or potential rate of error** of the technique or theory when applied;
- 4) the existence and maintenance of **standards and controls**; and
- 5) whether the technique or theory has been generally accepted in the scientific community.

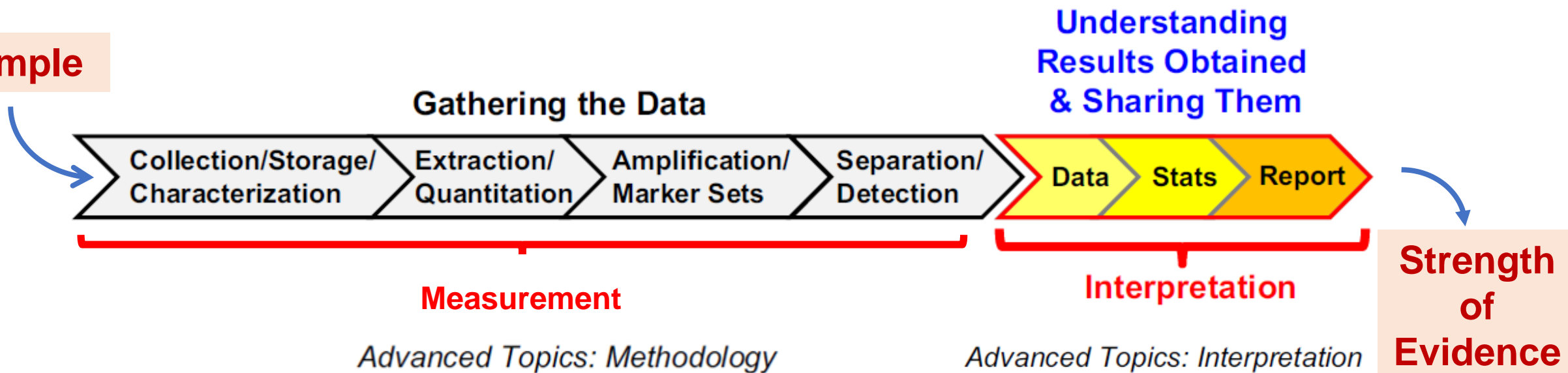
[https://www.law.cornell.edu/rules/fre/rule\\_702](https://www.law.cornell.edu/rules/fre/rule_702)

# **DNA Mixture Interpretation**

## **Reliability Considerations**

# DNA: Measurement & Interpretation

**Sample**

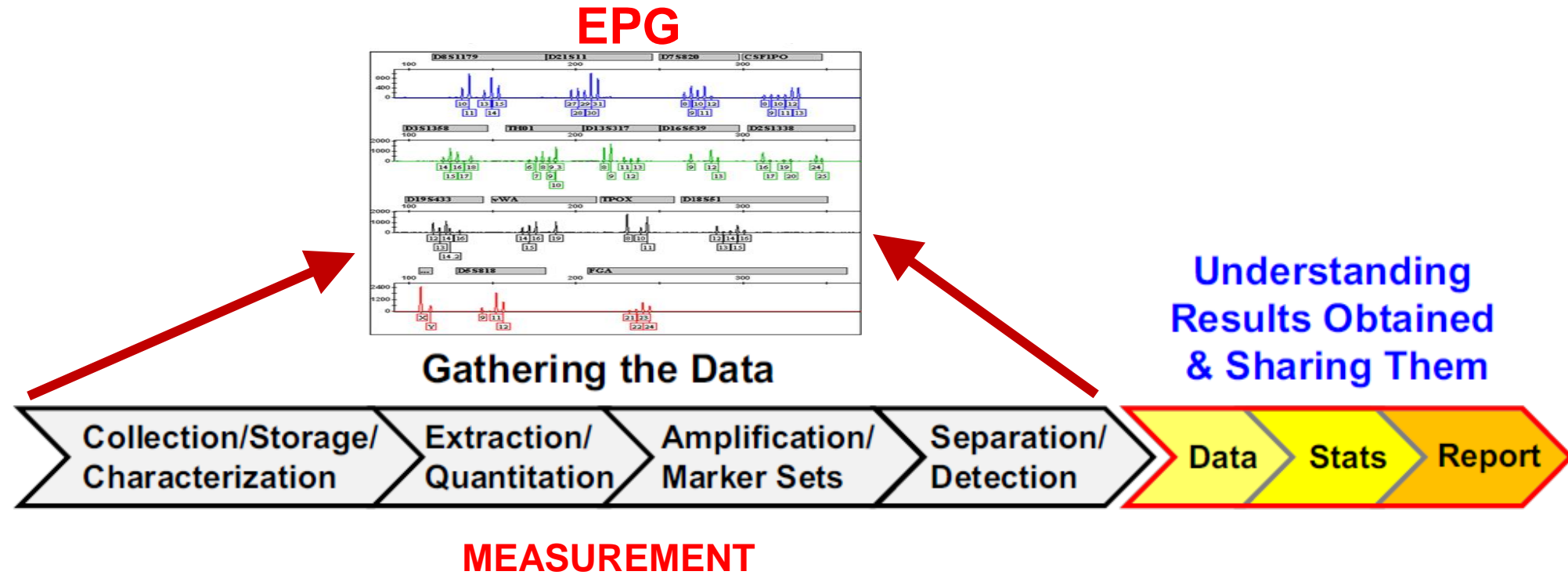


**FIGURE 1.1** Steps involved in the overall process of forensic DNA typing. This book focuses on understanding the data through data interpretation and statistical interpretation.

JOHN M. BUTLER  
National Institute of Standards and Technology  
Gaithersburg, Maryland, USA

*Advanced Topics in Forensic DNA Typing: Interpretation*  
<http://dx.doi.org/10.1016/B978-0-12-405213-0.00001-4>

# DNA: Measurement



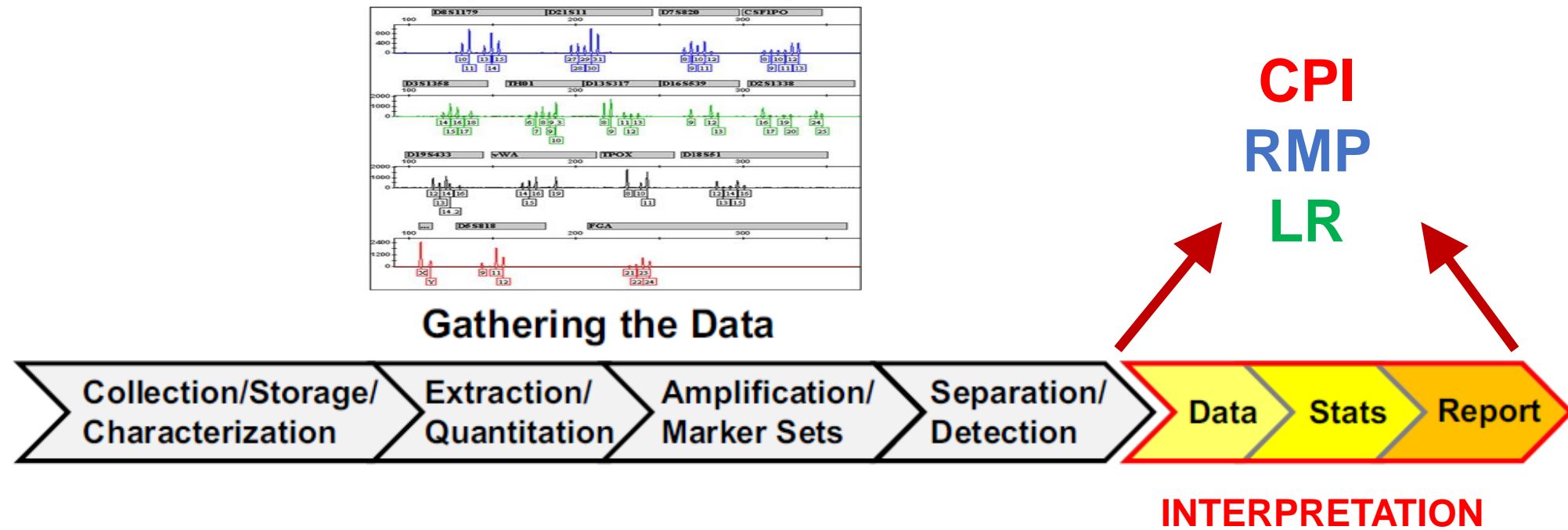
**FIGURE 1.1** Steps involved in the overall process of forensic DNA typing. This book focuses on understanding the data through data interpretation and statistical interpretation.

JOHN M. BUTLER  
National Institute of Standards and Technology  
Gaithersburg, Maryland, USA

*Advanced Topics in Forensic DNA Typing: Interpretation*  
<http://dx.doi.org/10.1016/B978-0-12-405213-0.00001-4>



# DNA: Interpretation

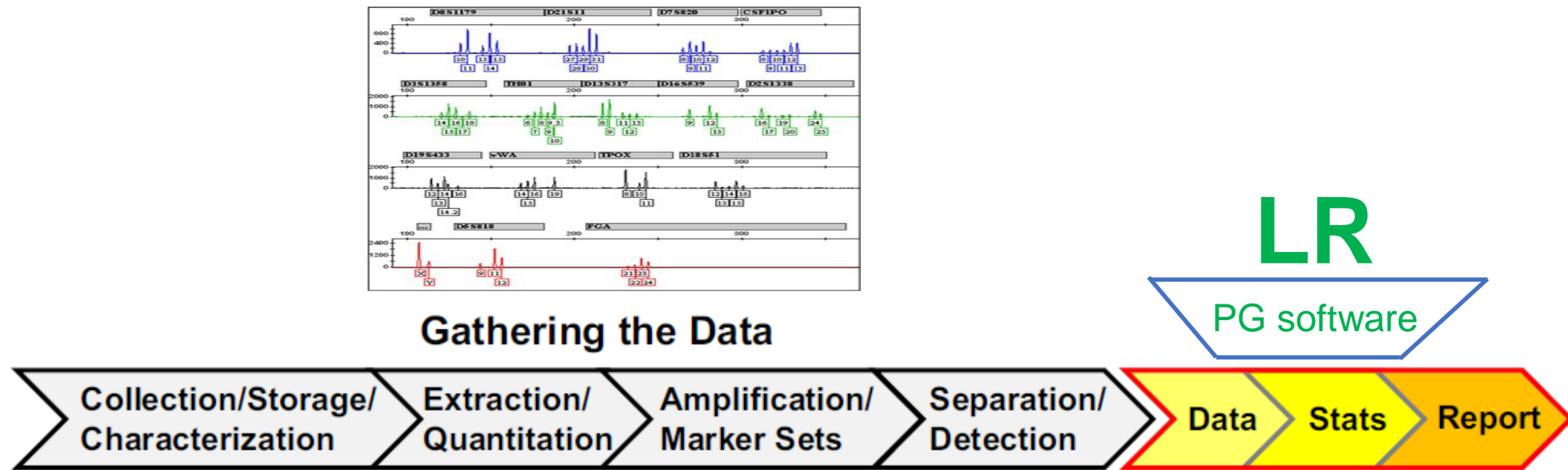


**FIGURE 1.1** Steps involved in the overall process of forensic DNA typing. This book focuses on understanding the data through data interpretation and statistical interpretation.

JOHN M. BUTLER  
National Institute of Standards and Technology  
Gaithersburg, Maryland, USA

*Advanced Topics in Forensic DNA Typing: Interpretation*  
<http://dx.doi.org/10.1016/B978-0-12-405213-0.00001-4>

# DNA: Measurement & Interpretation System

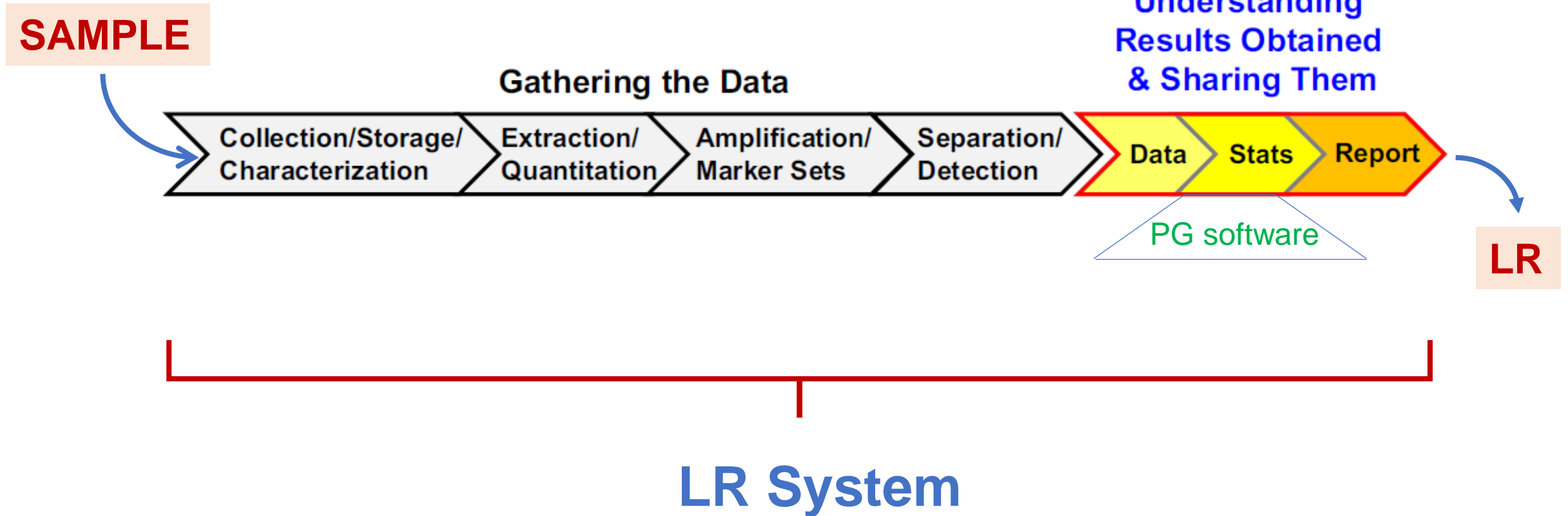


**FIGURE 1.1** Steps involved in the overall process of forensic DNA typing. This book focuses on understanding the data through data interpretation and statistical interpretation.

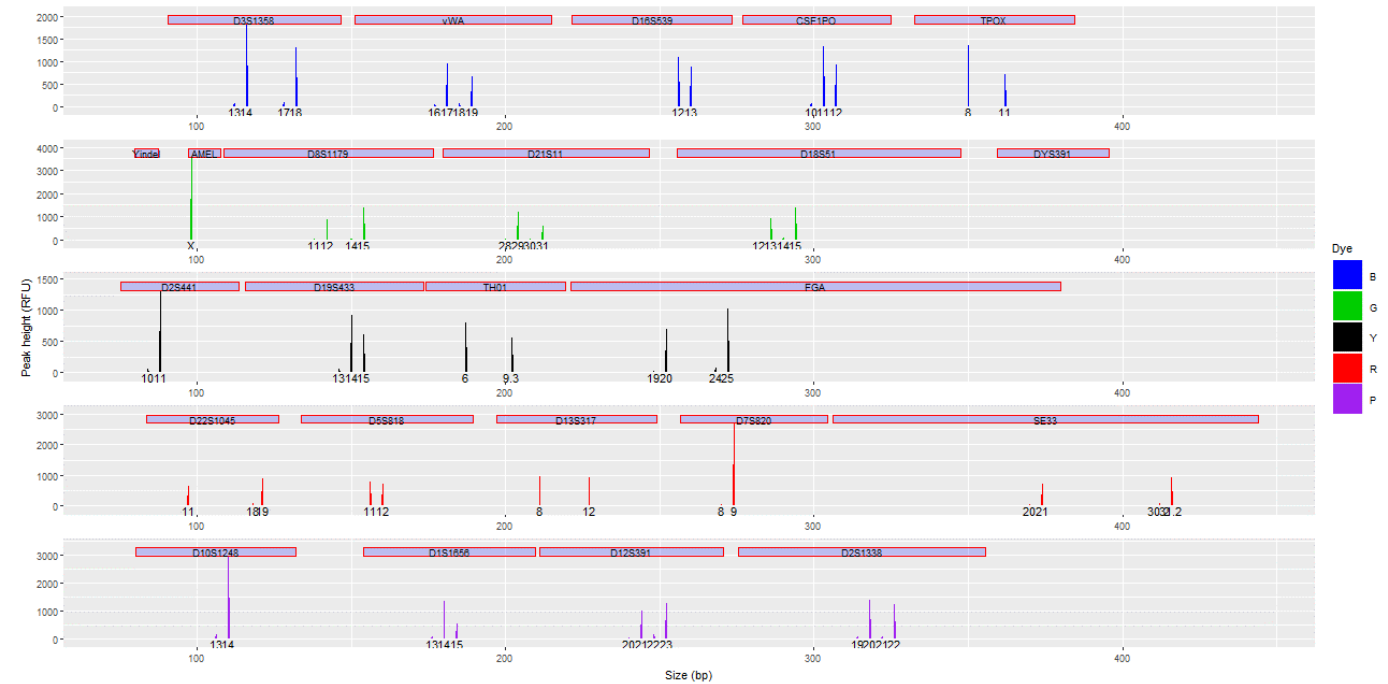
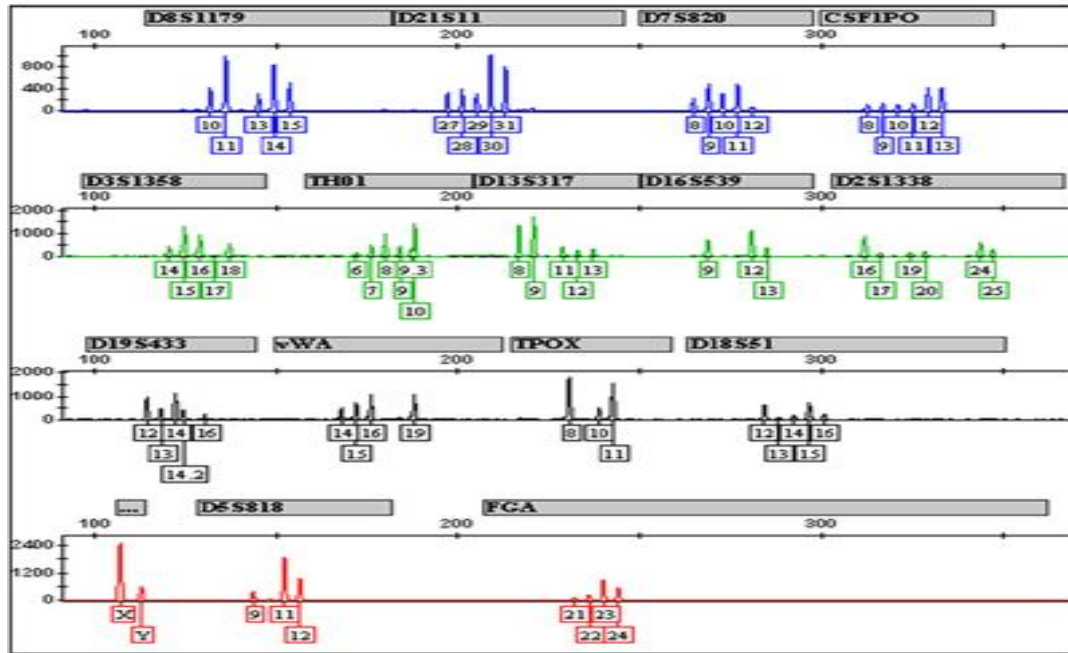
JOHN M. BUTLER  
National Institute of Standards and Technology  
Gaithersburg, Maryland, USA

*Advanced Topics in Forensic DNA Typing: Interpretation*  
<http://dx.doi.org/10.1016/B978-0-12-405213-0.00001-4>

# DNA: Measurement & Interpretation System



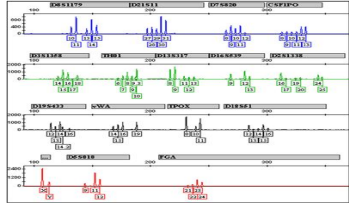
# Propositions



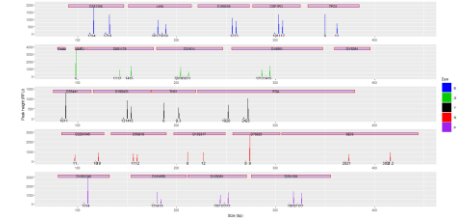
$H_p$ : DNA from POI is in the sample

$H_d$ : DNA from POI is not in the sample

# Likelihood Ratio



$$LR = \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)}$$



$H_p$ : DNA from POI is in the sample

$H_d$ : DNA from POI is not in the sample

$I$  = Background Information prior to examining crime sample

# Empirical Assessment of LR Systems

There are two aspects to judging the reliability of an LR system for assessing value of forensic DNA evidence

## 1. Discrimination power

Ability to discriminate between Hp-true situations from Hd-true situations

## 2. Calibration Accuracy

Accuracy of weight of evidence assessment

# Discrimination Power

The ability of an LR system to discriminate between  $H_p$  and  $H_d$  depends on

1. How much of the discriminating information in the sample is extracted and measured?

(e.g., CE vs NGS)

2. Does the interpretation make effective use of such information?

(e.g., model fidelity)

# Empirical Assessment of Performance

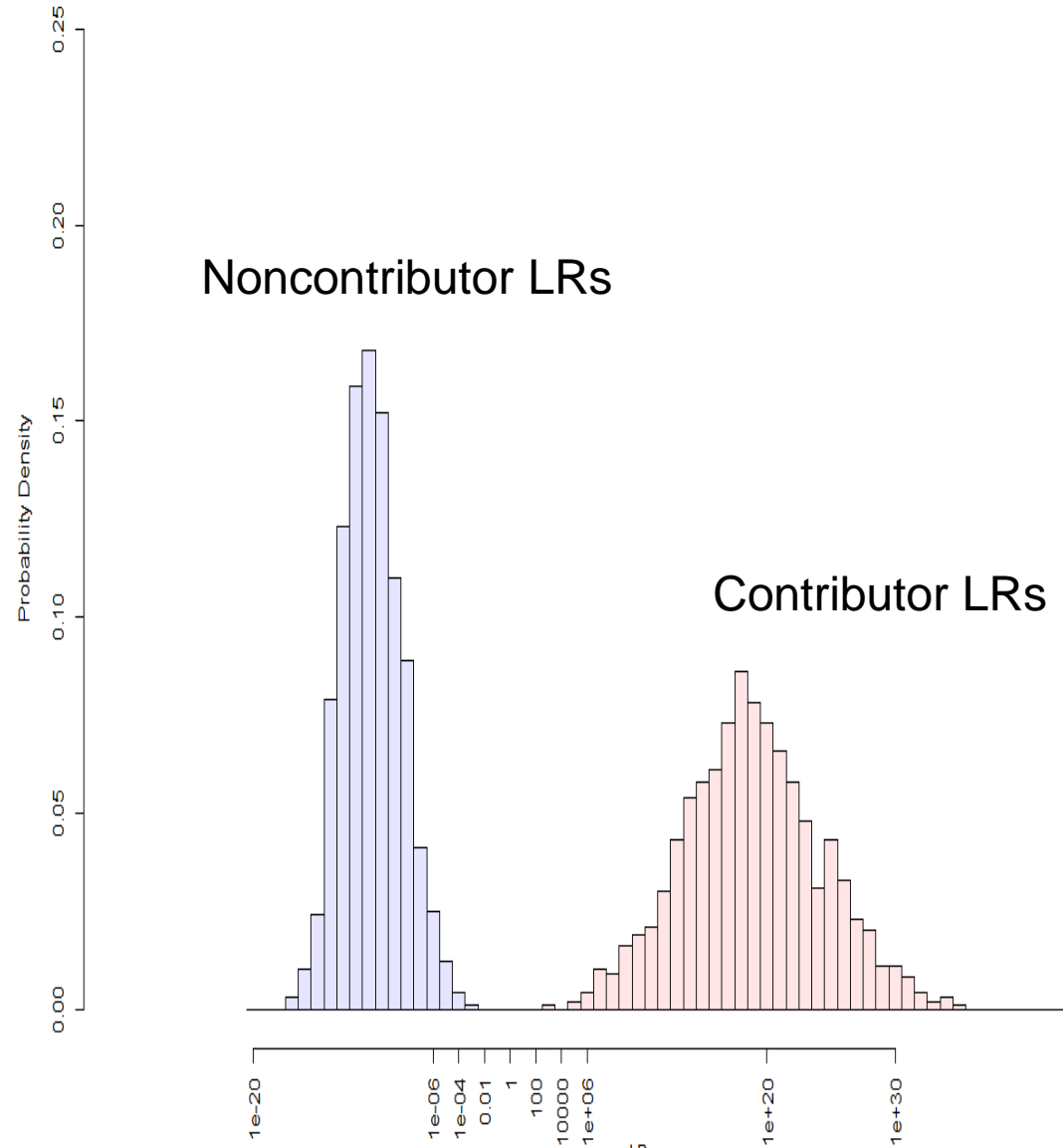
- Suppose we have a large collection of ground truth known DNA samples representing different scenarios (degradation, number of contributors, template amounts, mixture ratios) we expect to encounter in case work
- For each sample, select a known contributor profile or a known noncontributor profile (say by coin toss) and **send them through the LR pipeline, from analysis to interpretation.** (blinded)
- Record the value of LR obtained along with whether it is for an  $H_p$  true case or for an  $H_d$  true case.
- At the end of this exercise we will have a pool of  $H_p$  true LR values and a pool of  $H_d$  true LR values.



# Ground Truth Known Tests

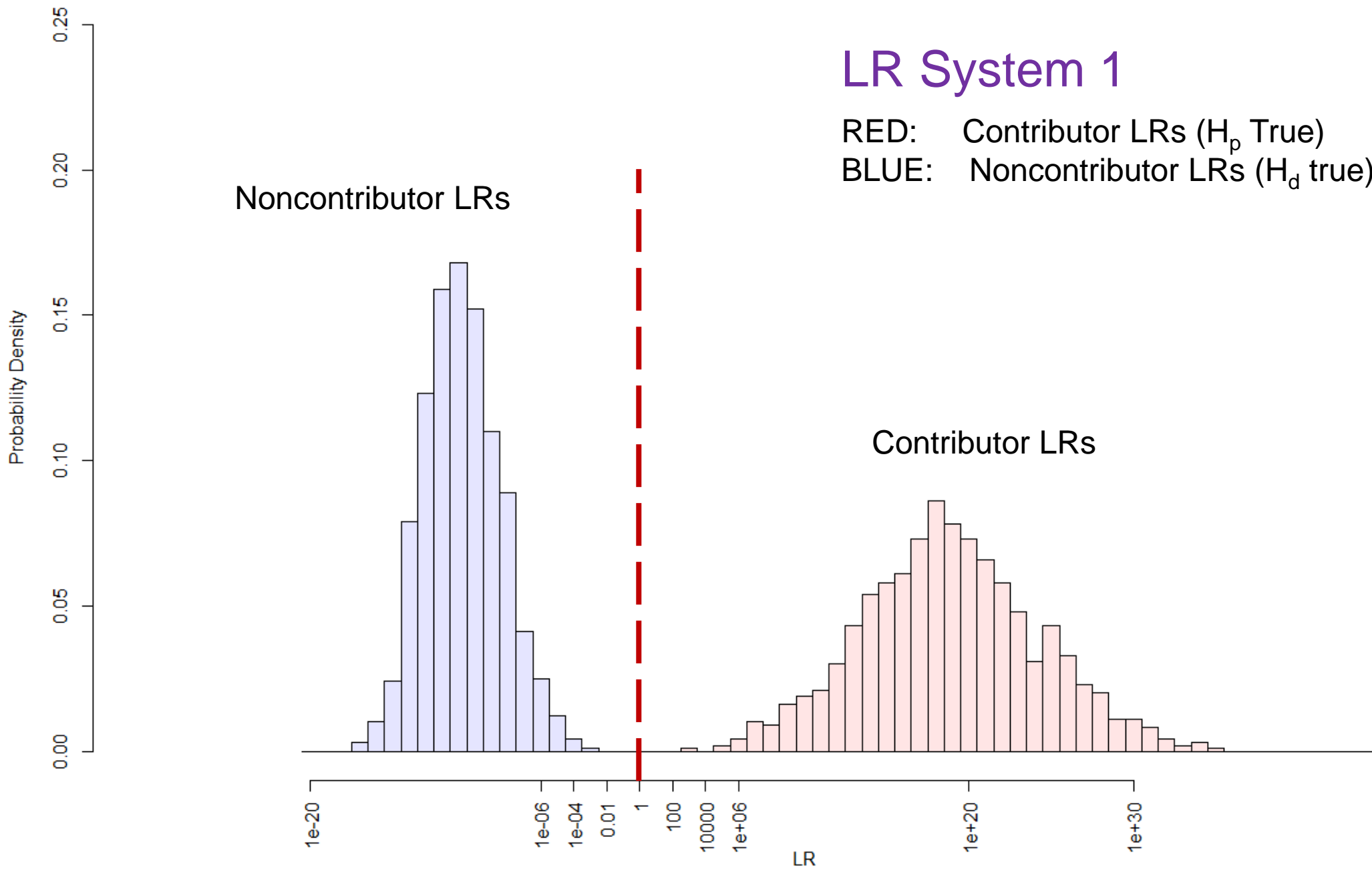
Representative of  
casework

A	B	C
Sample Details	Noncontributor LRs	Contributor LRs
2P, 1:1, degraded, 100 pg, high allele overlap	0.00E+00	3.68E+08
2P, 9:1, degraded, high allele overlap	6.69E-03	2.10E+07
	1.48E-03	7.34E+10
	1.60E-03	1.26E+09
	1.04E+00	1.45E+08
	0.00E+00	3.87E+10
	1.32E-01	3.12E+07
	3.98E-03	1.71E+06
	1.12E-02	6.56E+10
	1.85E-06	1.95E+08
	1.56E-01	1.61E+06
	5.48E-09	4.13E+10
	3.97E-04	1.87E+08
	0.00E+00	1.11E+06
	6.07E-13	5.18E+09
	5.03E-04	2.99E+07
	7.10E-03	1.87E+05
	0.00E+00	1.86E+09
	0.00E+00	8.08E+08
	5.81E-01	7.17E+17
	8.81E-08	5.81E+13
	1.32E-01	2.76E+09
	2.26E-14	3.18E+17
	2.12E-01	4.66E+13
	2.78E-01	4.78E+07
	<b>ETC</b>	
	1.21E+00	1.01E+17
	1.09E-03	1.16E+12
	2.09E-13	1.41E+06
	0.00E+00	9.87E+16
	0.00E+00	2.61E+10
	2.60E-01	2.34E+03
	1.37E-04	1.05E+16

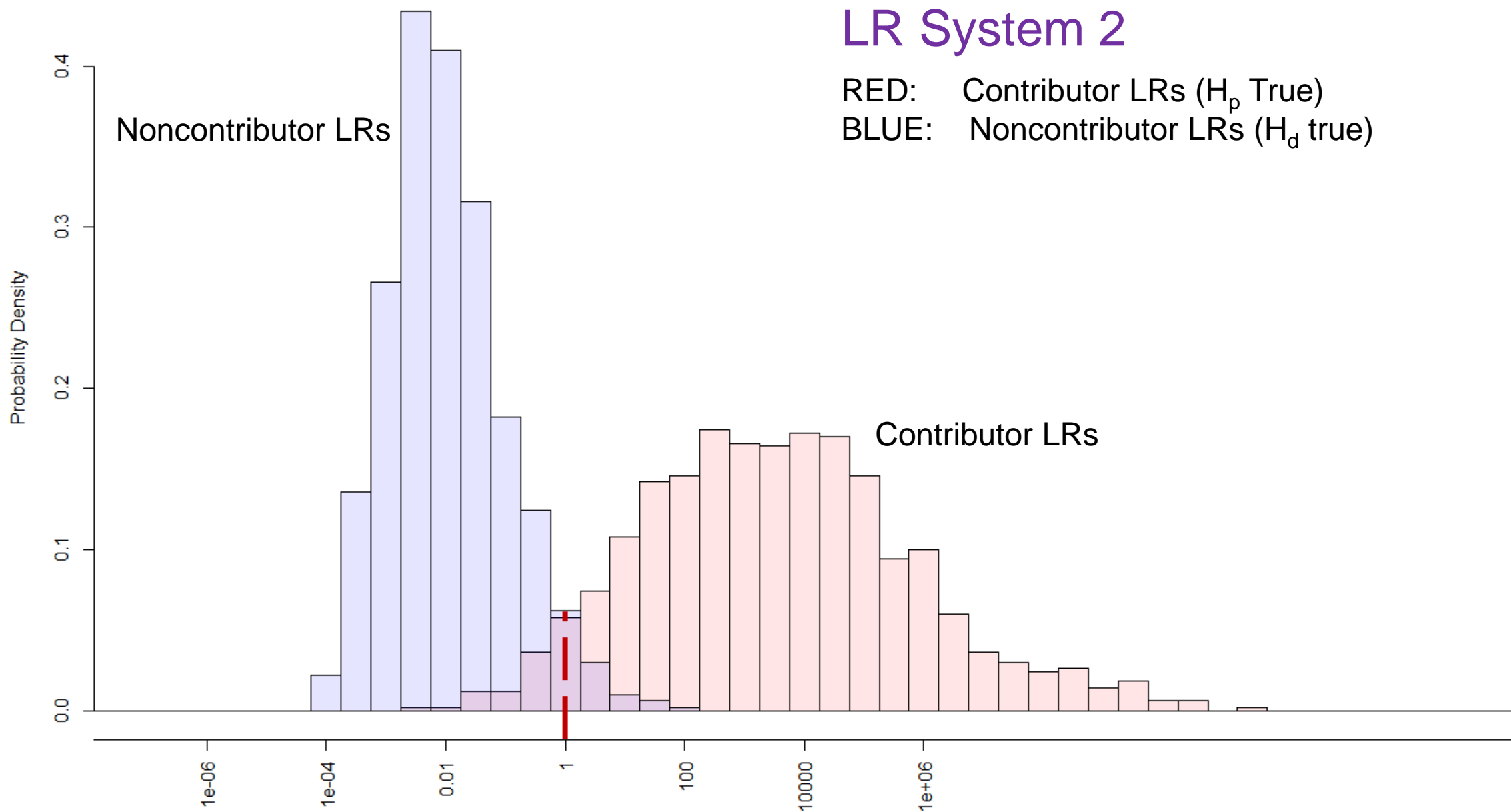


**Disclaimer:** This is only a thought experiment. Actual assessment will require a well thought out experimental design.

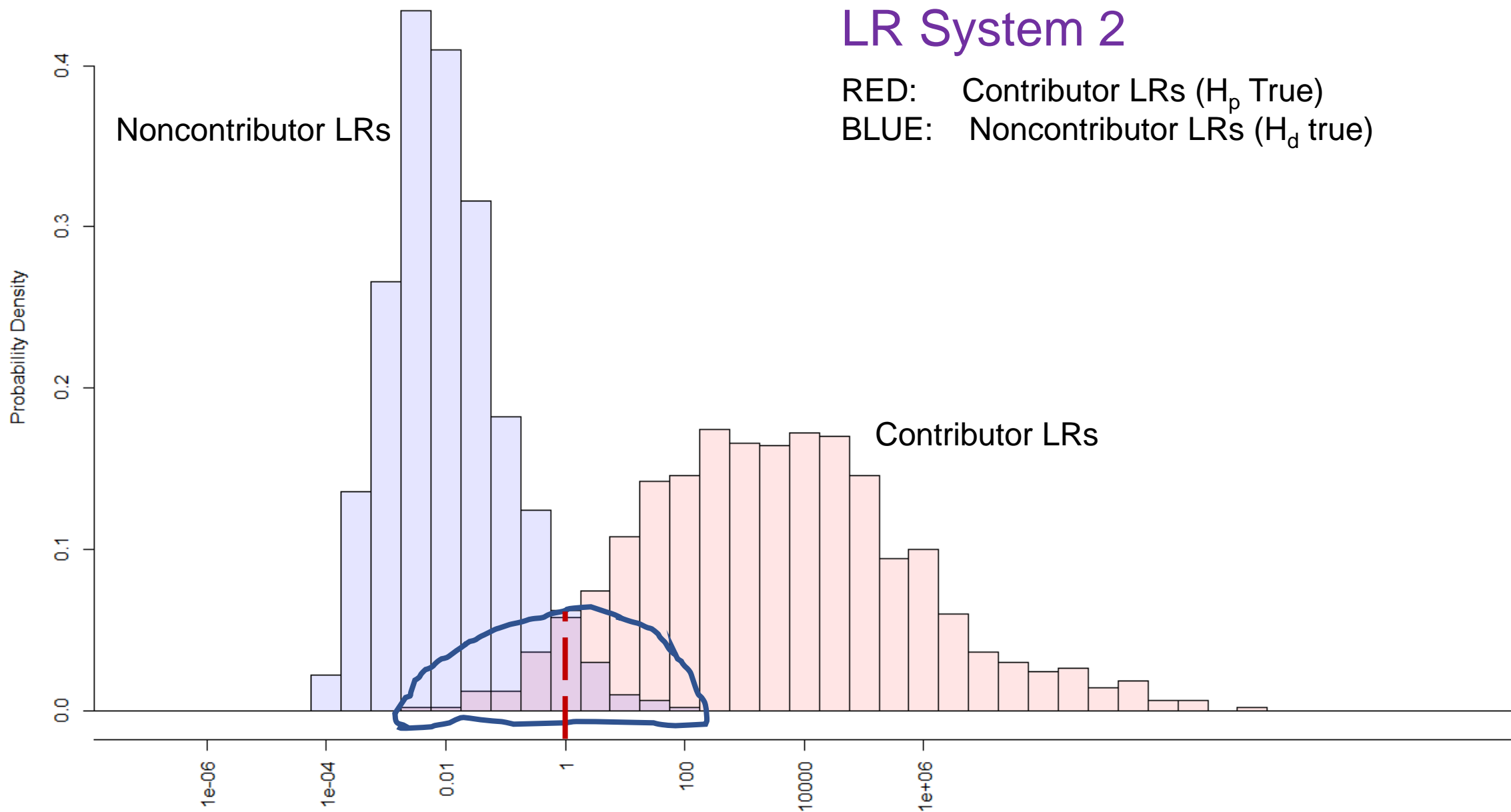
# Well Separated Hp-true & Hd-true LR Distributions



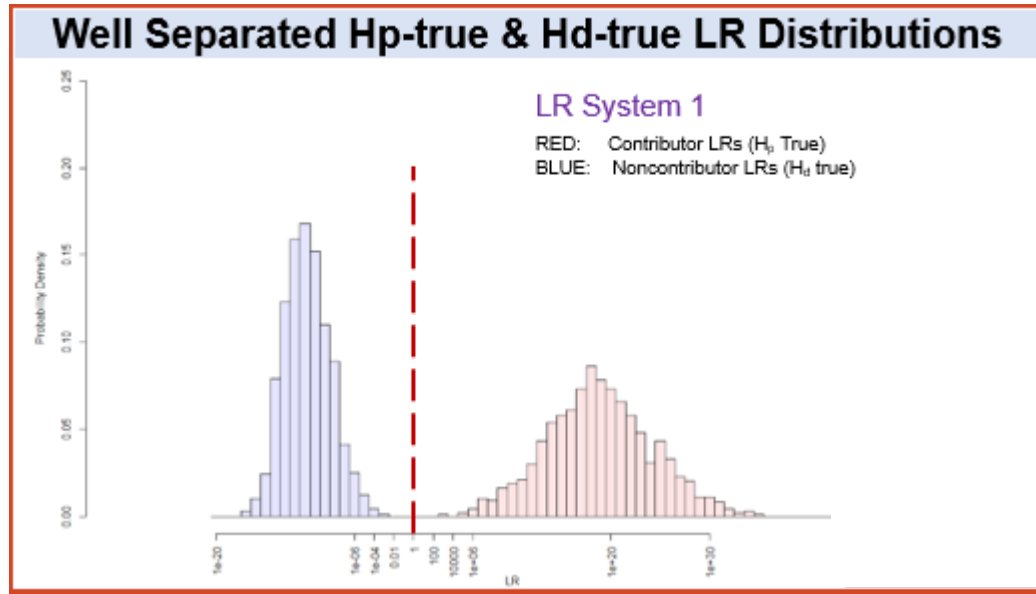
# Overlapping Hp-true & Hd-true LR Distributions



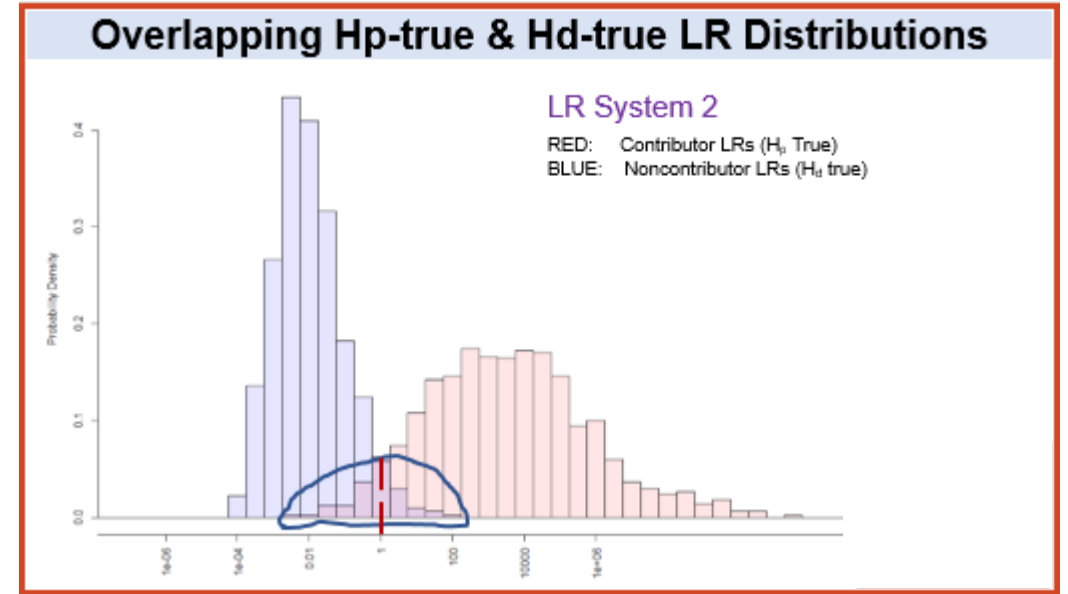
# Overlapping Hp-true & Hd-true LR Distributions



# Discrimination Power



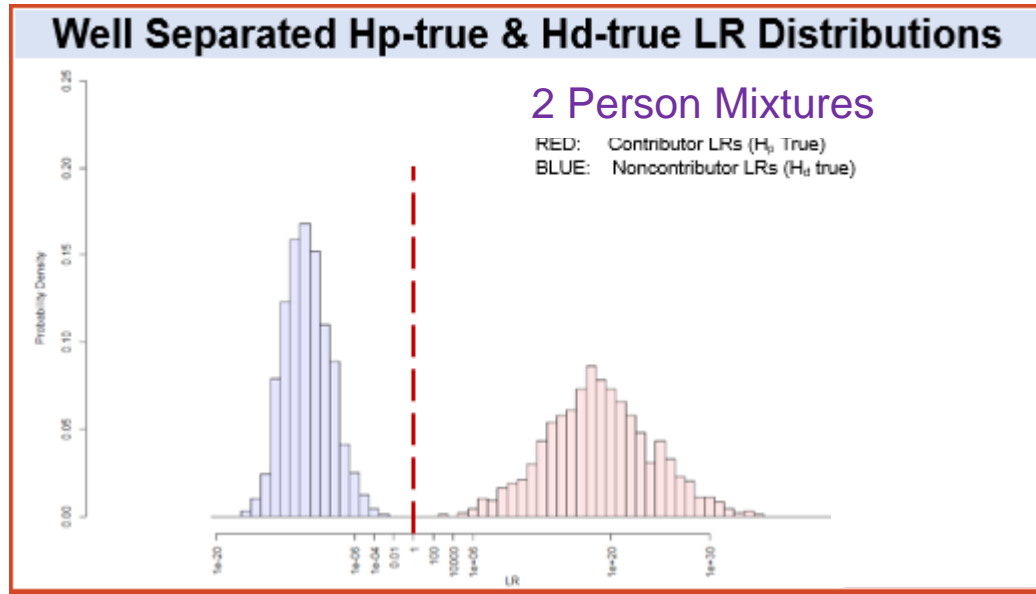
LR System 1



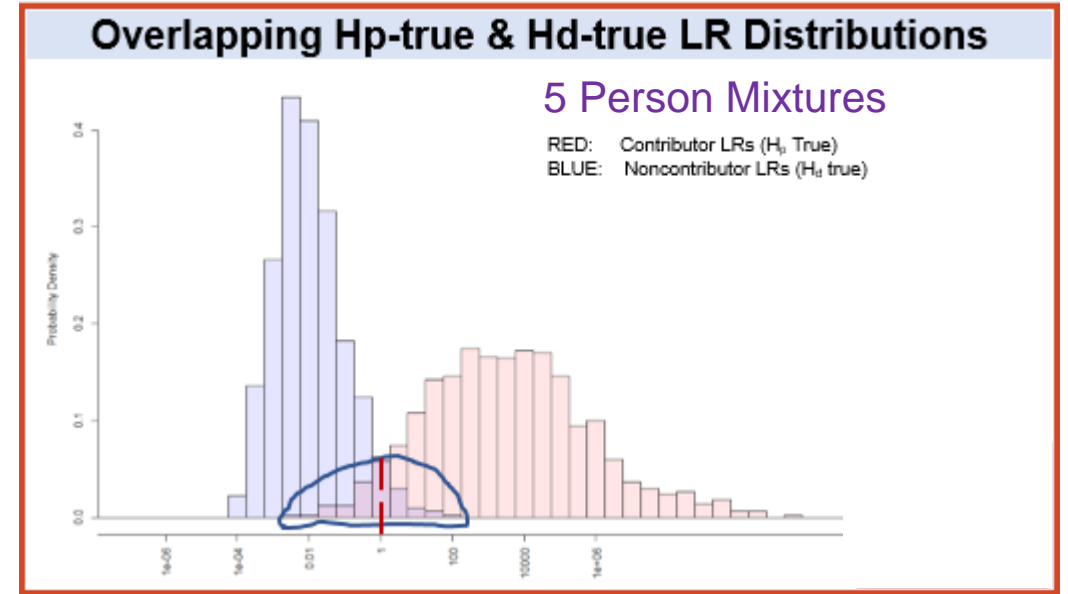
LR System 2

LR System 1 is more discriminating between  $H_p$  and  $H_d$  than LR system 2

# Discrimination Power



Performance on 2 Person Mixtures



Performance on 5 Person Mixtures

The same LR System is more discriminating for 2 person mixtures than for 5 person mixtures.

# Calibration Accuracy

## ACCURACY of Strength of Evidence Assessment

If the model used correctly describes the underlying process:

- LR value of 1 is equally likely under  $H_p$  as it is under  $H_d$
- LR value of 10 is 10 times more likely to occur under  $H_p$  than it is under  $H_d$ .
- LR value of 100 is 100 times more likely under  $H_p$  than it is under  $H_d$ .
- LR value of 0.1 is 10 times more likely under  $H_d$  than it is under  $H_p$ .

LR value of  $x$  is  $x$  times more likely to occur under  $H_p$  than under  $H_d$ .

# Calibration Accuracy

LR value of  $x$  is  $x$  times more likely to occur under  $H_p$  than under  $H_d$ . ( **LR of LR is LR** )

..... the likelihood ratio of the likelihood ratio is the likelihood ratio. That is

$$l[l(e_k)] = \frac{P_1[l(e_k) | h_1]}{P_2[l(e_k) | h_2]} = l(e_k) \quad (1.32)$$

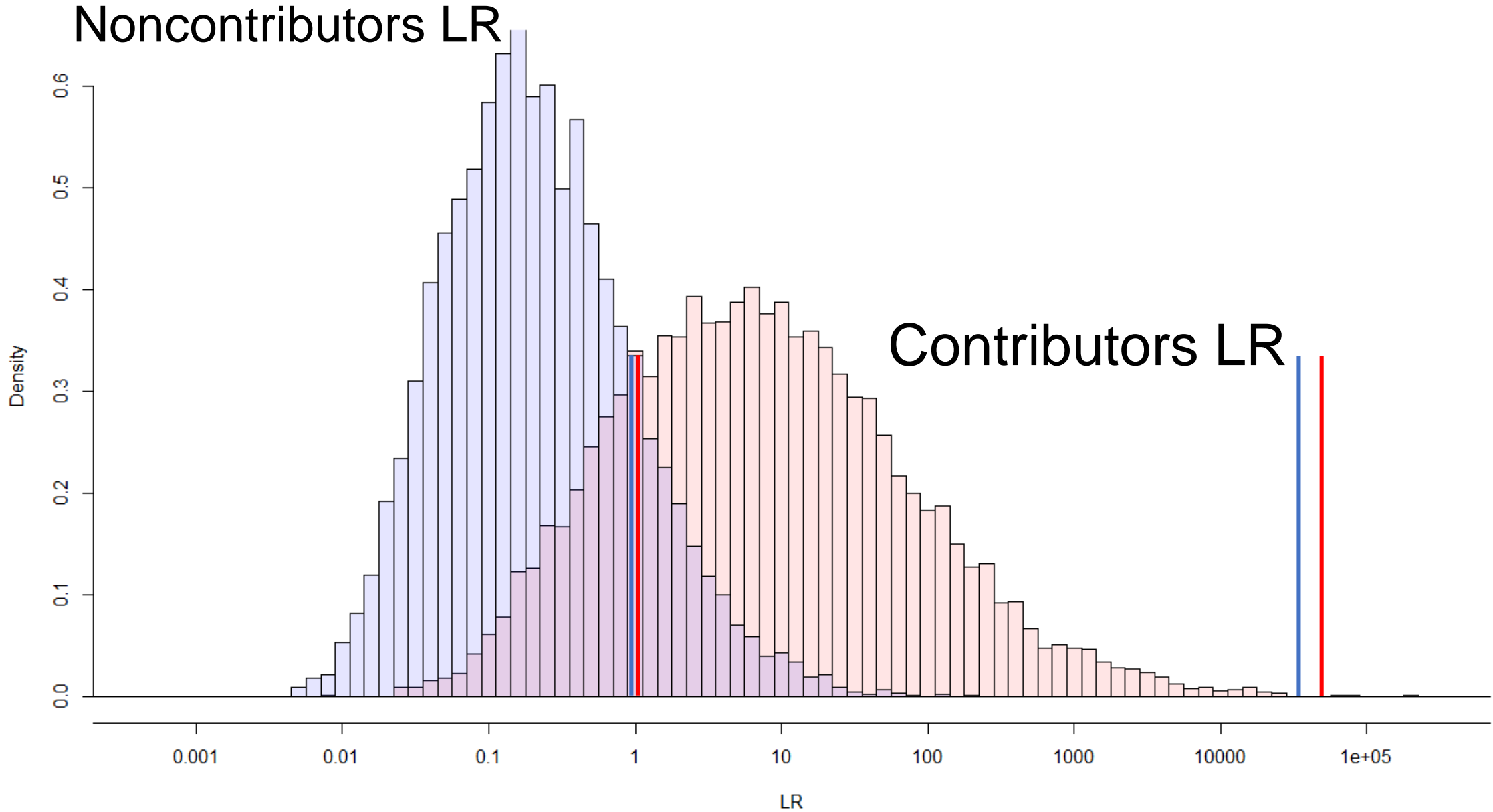
for all events  $e_k$ .

Green and Swets, 1966, page 26, section 1.8, equation (1.32)

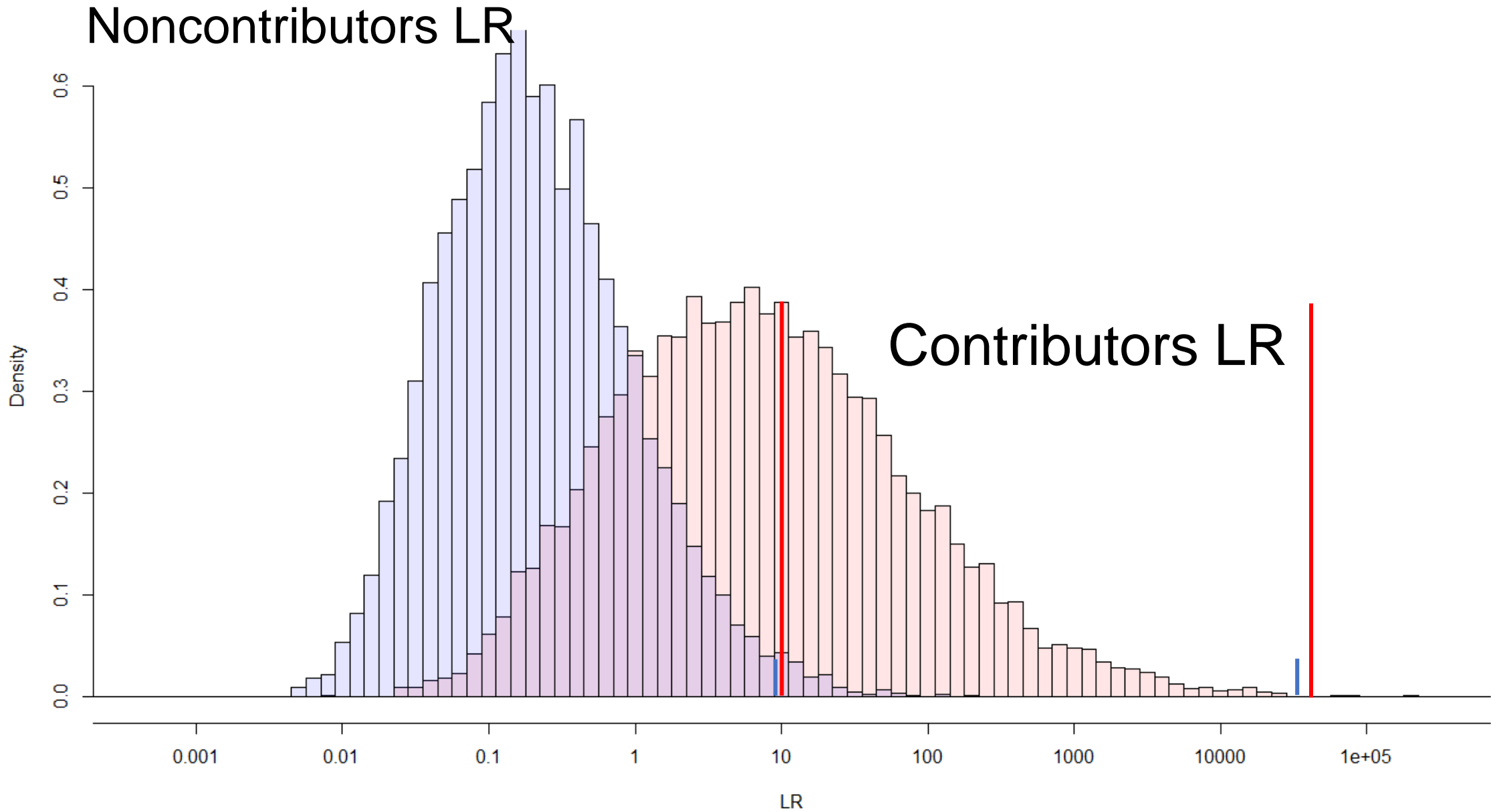
**In principle**, this property can be empirically tested



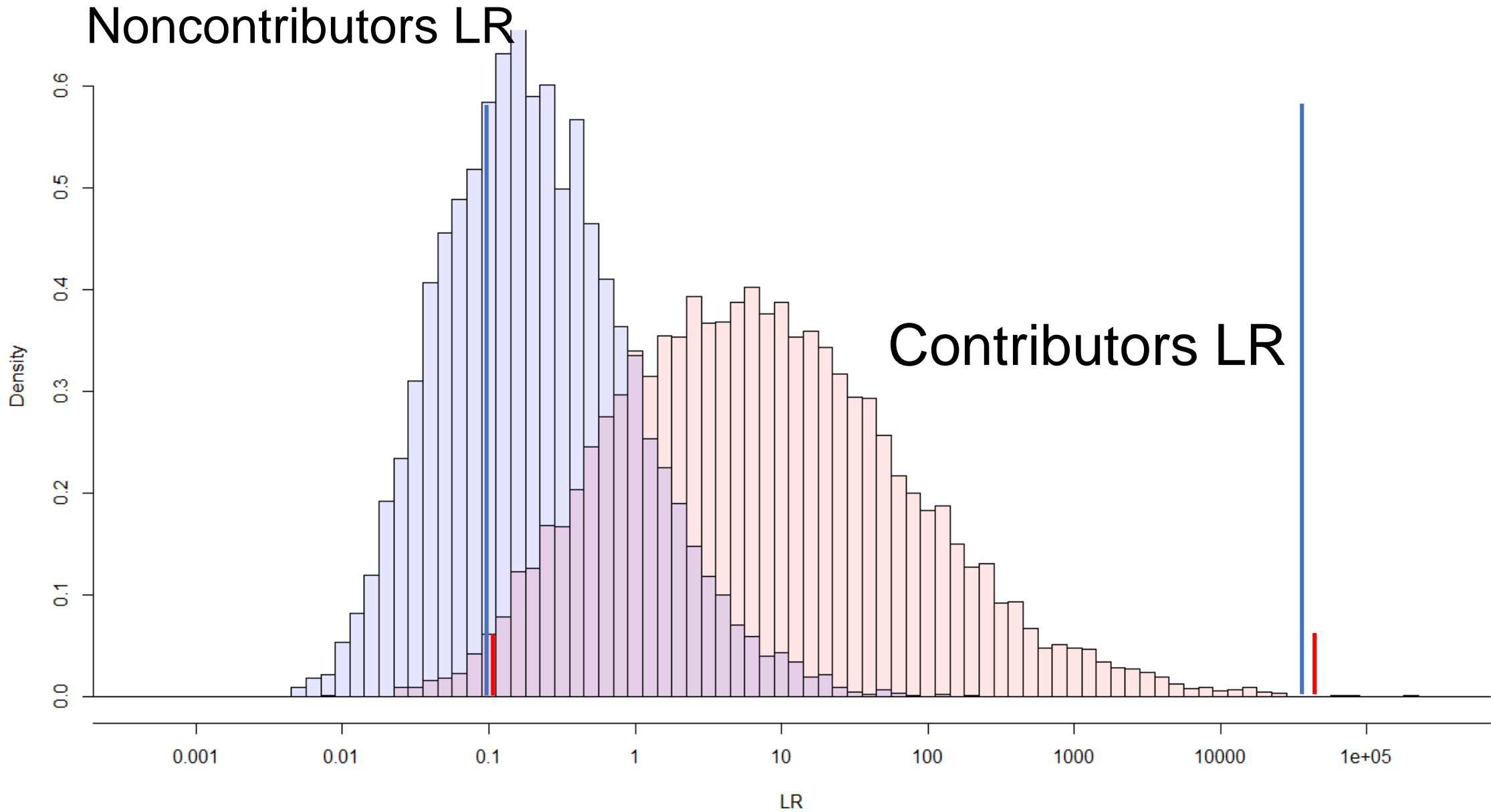
# Calibration Accuracy: Empirical Assessment



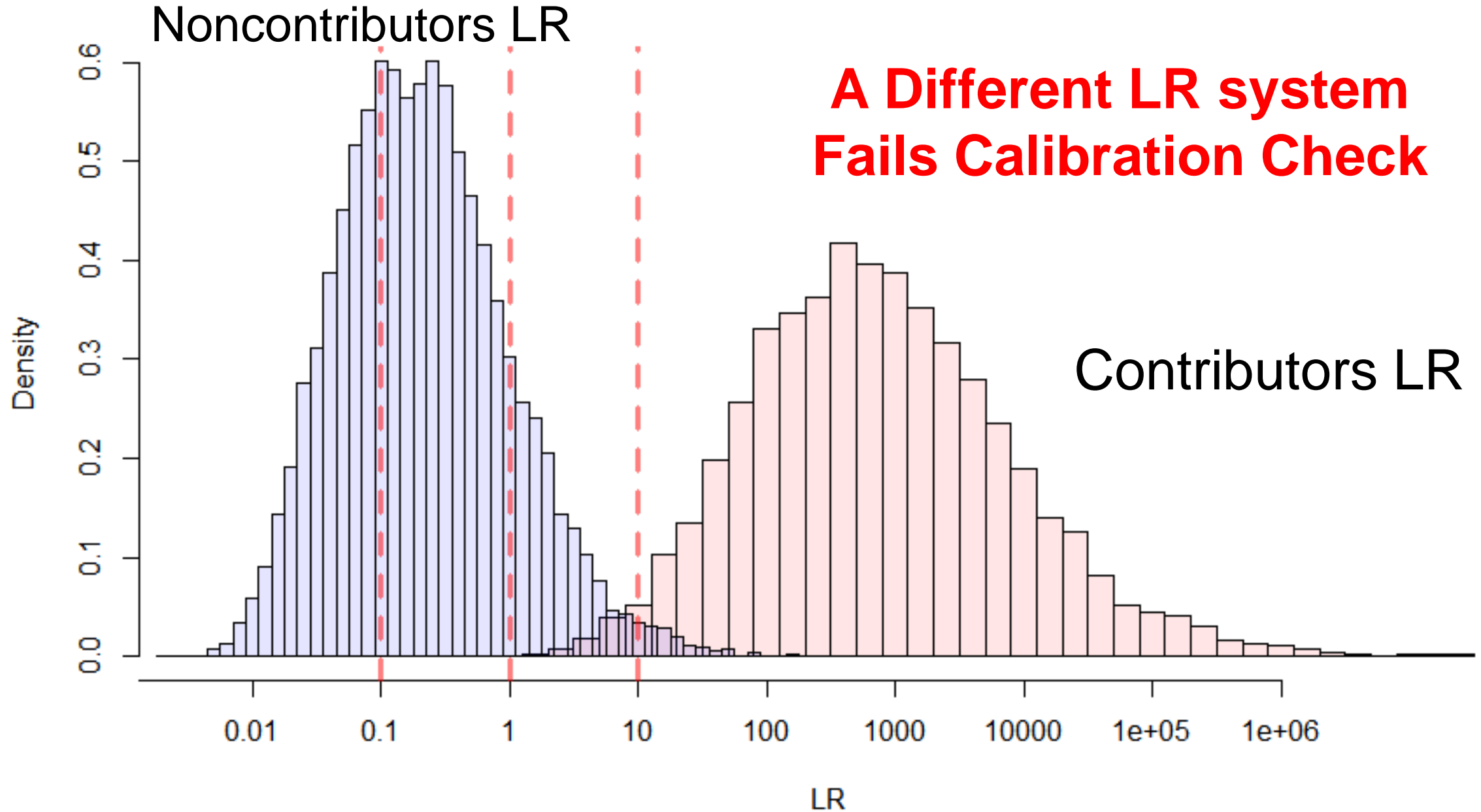
# Calibration Accuracy: Empirical Assessment



# Calibration Accuracy: Empirical Assessment



# Calibration Accuracy: Empirical Assessment



# Some Factors That May Affect Reliability of an LR System

1. Sample
  - a) Sample amount (contributor template amounts)
  - b) Sample quality (degradation level)
2. Labs
  - a) Kits used
  - b) Equipment Used
  - c) Number of PCR cycles
  - d) Analyst
  - e) Choice of Analytical Threshold (AT)
3. Probabilistic Genotyping (PG) Model
  - a) Choice of model
  - b) Choice of laboratory specific parameters for use in the PG model
  - c) Propositions Chosen ( $H_p$  and  $H_d$ )**
4. Software Implementing the PG Model
  - a) Choice of numerical methods for computing LR (MCMC, Numerical Integration)
  - b) Choice of number of iterations OR numerical integration parameters (e.g. grid size)

**FACTOR  
SPACE**

# Reproducibility is not Reliability

Degree of agreement among a group of labs by itself does not characterize degree of reliability

but

Degree of substantial disagreement among labs (or methods) makes it difficult to discern the degree of reliability of results provided by any particular laboratory.

Such judgements will have to be based on internal validation data from the laboratory providing the analysis and report in any given case.

# Reproducibility: An Interlab Study

Forensic Science International: Genetics 35 (2018) 156–163



Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



Research paper

## GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: Results and evaluation



P.A. Barrio<sup>a,b,c</sup>, M. Crespillo<sup>a,c,\*</sup>, J.A. Luque<sup>a,c</sup>, M. Aler<sup>d</sup>, C. Baeza-Richer<sup>e</sup>, L. Baldassarri<sup>f</sup>,  
E. Carnevali<sup>g</sup>, P. Coufalova<sup>h</sup>, I. Flores<sup>i</sup>, O. García<sup>j</sup>, M.A. García<sup>k</sup>, R. González<sup>l</sup>, A. Hernández<sup>m</sup>,  
V. Inglés<sup>n</sup>, G.M. Luque<sup>b</sup>, A. Mosquera-Miguel<sup>o</sup>, S. Pedrosa<sup>p</sup>, M.L. Pontes<sup>q</sup>, M.J. Porto<sup>r</sup>, Y. Posada<sup>s</sup>,  
M.I. Ramella<sup>t</sup>, T. Ribeiro<sup>u</sup>, E. Riego<sup>v</sup>, A. Sala<sup>w</sup>, V.G. Saragoni<sup>x</sup>, A. Serrano<sup>c</sup>, S. Vannelli<sup>y</sup>

Participants were provided with the thresholds values used/employed: analytical threshold of 50 RFUs, stochastic threshold of 150 RFUs, and stutter threshold for each of the markers/kits according to the manufacturer's specifications.

# GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06).

## Reporting conclusions: Results and evaluation

**Table 1**

Hypothesis and LR values obtained by each of the participating laboratories. All laboratories used the *LRmixStudio* software, except those marked as \* (*EuroForMix*) and \*\* (*DNAMIX*). Legend: V (Victim), S (Suspect), P (Regular partner), U (Unknown).

Labs	LR value	Hypothesis	Other evaluations	
			LR value	Hypothesis
GHEPMIX_08*	1.7200E + 02	V + S + P/V + U + P		
GHEPMIX_23	2.6000E + 03	V + S + P/V + U + P		
GHEPMIX_26	6.1640E + 03	V + S + P/V + U + P		
GHEPMIX_17	6.5565E + 04	V + S + P/V + U + P		
GHEPMIX_07	6.8487E + 04	V + S + P/V + U + P		
GHEPMIX_05	1.4800E + 05	V + S + P/V + U + P		
GHEPMIX_22	2.8776E + 05	V + S + P/V + U + P		
GHEPMIX_06	3.2224E + 05	V + S + P/V + U + P		
GHEPMIX_16	4.3423E + 05	V + S + P/V + U + P		
GHEPMIX_18	1.3900E + 06	V + S + P/V + U + P		
GHEPMIX_03	1.8200E + 06	V + S + P/V + U + P		
GHEPMIX_02	2.7323E + 06	V + S + P/V + U + P		
GHEPMIX_20	5.5183E + 06	V + S + P/V + U + P		
GHEPMIX_15	1.9820E + 07	V + S + P/V + U + P		
GHEPMIX_27	1.3587E + 08	V + S + P/V + U + P	7.4048E + 19	P/U
GHEPMIX_13**	2.7300E + 10	V + S + P/V + U + P		
GHEPMIX_10	3.2032E + 14	V + S + P/V + U + P	1.1551E + 07	V + S + P/V + U1 + U2
GHEPMIX_24			1.3400E + 19	V + P/V + U



# Reproducibility: Comparison of PG Software

Forensic Science International: Genetics 37 (2018) 143–150

Contents lists available at ScienceDirect



Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



DNA mixtures interpretation – A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples



E. Alladio<sup>a,b,\*</sup>, M. Omedei<sup>b</sup>, S. Cisana<sup>b</sup>, G. D'Amico<sup>b</sup>, D. Caneparo<sup>b</sup>, M. Vincenti<sup>a,b</sup>, P. Garofano<sup>b,c</sup>

<sup>a</sup> Dipartimento di Chimica, Università degli Studi di Torino, Via P. Giuria 7, 10125, Torino, Italy

<sup>b</sup> Centro Regionale Antidoping e di Tossicologia "A. Bertinaria", Regione Gonzole 10/1, 10043, Orbassano, Torino, Italy

<sup>c</sup> Accademia Italiana di Scienze Forensi, Viale Regina Margherita 9/D, 42124, Reggio Emilia, Italy

Lab Retriever  
LRmix Studio

DNA•VIEW<sup>®</sup>,  
EuroForMix and  
STRmix

Page 145

Furthermore, log(LR) results provided by fully-continuous models proved similar and convergent to one another, with slightly higher within-software differences (i.e. **approximately 3–4 degrees of magnitude**).

**A factor of 1000 to 10000 ?**

# Potential Impact of LR Differences - Illustration

**Effect of 3 to 4 orders of magnitude:**

Suppose prior odds = 1: 1000000 = (1/1,000,000)

(Crime occurred in the city of New York, say)

LR1 = 50000 (Strong evidence)

LR2 = 50000000 (Very Strong Evidence) [ a factor of 1000 higher than LR1 ]

Posterior Probability 1 = 0.048 = 4.8%

Posterior Probability 2 = 0.98 = 98%

**Posterior Odds = Prior Odds x LR**

$$\text{Posterior Probability} = \frac{(\text{LR} \times \text{prior odds})}{1 + (\text{LR} \times \text{prior odds})}$$

# Summary

1. What is meant by “Reliability”?
2. System Reliability vs Component Reliability
3. The need for empirical testing of models
4. Main requirements for reliability: Discrimination power and Calibration Accuracy
5. Discussion illustrating the concepts of discrimination power and calibration accuracy with data from validation studies
6. Factor Space
7. Reproducibility is not Reliability
8. Impact of LR differences between systems in casework



**COMING  
SOON**

**In Module 3  
John will talk about  
Validation Plans &  
Experimental Design**



**ISHI 2020 Validation Workshop**  
**Friday September 18th, 2020 // 9:00 am - 12:30 pm**

**Validation Principles, Practices, Parameters,  
Performance Evaluations, and Protocols**

**Validation Plans &  
Experimental Design**

**Module 3**

**John M. Butler**

National Institute of Standards and Technology



# Disclaimers

**Points of view are those of the presenter** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

## **Identification does not imply endorsement**

Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

## Module 3 (John)

- Review Input Received
- Creating a Validation Plan
- Considering Experimental Design
  - Types of studies and numbers of samples depend on what you decide is fit for purpose
  - Factor space coverage for DNA mixture interpretation
  - Review what has been done in some published PGS studies



# Some Specific Input Received for This Workshop

- **Teresa Cheromcha** (Colorado Bureau of Investigation-Grand Junction)
  - Assistant TL for CBI system with 5 laboratories
- **Kristy Kadash** (Jefferson County Regional Crime Laboratory, Colorado)
  - Member of SWGDAM and OSAC and former TL
- **Kate Philpott** (Adjunct Faculty/Research Analyst, VCU Forensic Science Program)
  - Legal and scientific consultant; recently co-authored the June 2020 *Gissantaner* amicus brief
- **Janel Smith** (Phoenix Police Department)
  - DNA Technical Leader for a large city laboratory; member of OSAC

*I reached out to each of them and asked for **ideas of things we should cover** to best assist DNA analysts and TLs and specifically **what information on the topic of validation would be most helpful** to them in their work*



# Thoughts from Kristy Kadash (1)

CODIS Admin, Jefferson Co. Colorado; member of SWGDAM & OSAC

What would be most helpful:

We spoke by phone for about an hour

- **How to design validation studies**

- Review purpose of each study and discuss appropriate experiments to test the system

- **How to analyze the data**

- Going beyond calculating averages and standard deviations, how to display and graph information, how to assess differences from previous systems, how to state results (want to avoid repetitive explanations in summaries)

- **How to report and communicate results**

- Without being too brief or too verbose, how to convey what you have done and why studies were performed

# Thoughts from Kristy Kadash (2)

CODIS Admin, Jefferson Co. Colorado; member of SWGDAM & OSAC

What would be most helpful:

- **How to assist auditors in deciding what is an appropriate validation study**
  - Often if auditors see the right key words and headings following QAS or SWGDAM, then they may view the study as good enough and not necessarily consider how effective or complete the validation studies are
- **How much testing is needed to verify that specific parts of probabilistic genotyping software are working properly**
  - With software version changes, it can be challenging to do function testing. What are the most important tests?
  - How you use the software dictates how you would validate it
  - When do you have to do validation vs. verification vs. performance check
- **Provide a reminder that validation and proficiency tests are an important part of doing quality work**

# Thoughts from Janel Smith

DNA Technical Leader, Phoenix PD; member of OSAC

What would be most helpful:

She provided an email response to my questions

- **How to thoroughly test and define limitations**, especially with PGS
  - ESR has provided some excellent resources for validation and implementation
  - Potential area of concern: **the ability to interpret mixtures of related contributors**
  - She commented that it would be beneficial to develop mixtures in-house where you can know the ground truth of the contributors and the ratios so you can see the output files from the PGS you are using to interpret
  - **“When is enough, enough...knowing you can’t test everything?”**
- **How many people should be involved** in the validation studies
  - Desire to have sufficient people to provide a greater depth of knowledge
  - But laboratory management wants to minimize the impact to casework production while still completing the validation in a timely manner

# Thoughts from Teresa Cheromcha (1)

Assistant DNA Technical Leader, Colorado Bureau of Investigation (CBI)

- Teresa provided her thoughts to me on her validation experience in a four-page single-spaced outline, then we talked for about 90 minutes the next day
- Some of my favorite quotes from my conversation with Teresa:
  - “We all want to do the best science”
  - “Don’t be afraid to ask your peers for help”
  - “It’s okay if you don’t know everything”
- From August 2017 to September 2018, Teresa organized and conducted STRmix validation studies and brought PGS online for their lab system
  - ESR provided a four-day training course and a one-day follow-up was received a year later
  - An **8-member committee** she chaired (including representatives from each of their 5 laboratories at the time, TL, & QM) met regularly and used Trello for project planning and tracking assignments
  - ESR (STRmix provider) supplied a proposal on studies to meet SWGDAM 2015 PGS guidelines
  - Two committee members designed mixtures (used DNA from staff members for unrelated individuals and a family reunion to collect related individuals); examined number of contributors, allele sharing (related), ratios, template amounts
  - Mixture samples were created after carefully quantifying DNA samples; replicates were run; tested samples on all 9 CE instruments across their 5 laboratories (now down to 4 laboratories)
  - ESR crunched their data and wrote up the CBI validation summary

# Thoughts from Teresa Cheromcha (2)

Assistant DNA Technical Leader, Colorado Bureau of Investigation (CBI)

Some additional thoughts and information:

- Validation should explore the edges, the challenging samples – committee members provided ideas on the types of samples that would be representative of casework seen in their laboratories or samples they had previously seen that were challenging
  - CBI has purchased a software upgrade, will conduct another internal validation study, and hopes to move up to 5-person mixtures after conducting more experiments
- Struggles with “analysis paralysis” -- **when do I have enough data, or did I over do it?**
- To follow up on issues seen, CBI holds a monthly TL meeting with all analysts

## Continuing Education through Reading the Literature

- Each analyst selects articles to read (8 is the minimum per year, 2 are summarized and shared)

# Thoughts from Teresa Cheromcha (3)

Assistant DNA Technical Leader, Colorado Bureau of Investigation (CBI)

Before implementation of a new method:

- A training plan was developed which included study questions, terms, readings, and tasks
- Analysts at CBI are expected to read and know the validation summary results and to understand the limitations
  - Analysts would not likely examine the original data used to generate the validation summary
- Competency testing
  - Developed a training plan which included study questions, terms, readings and required tasks
  - Written exam: 10-12 questions
  - Practical exam: single source to 4p mixtures (e.g., redefine an OL allele as stutter)
  - Oral exam: mock trial assessment by TL and assistant TL before going to court

# Thoughts from Kate Philpott (1)

Adjunct Faculty/Research Analyst, VCU Forensic Science Program

We spoke by phone for about an hour following a presentation that I gave a few weeks ago entitled “DNA Mixtures: Where We Were and Where We Are Now” for the National Association of Criminal Defense Lawyers (NACDL) National Forensic College DNA Day

1. She was surprised that the **ISFG DNA Commissions** (2006, 2012, 2016, 2018, 2020) **commented years ago on issues faced today with probabilistic genotyping**
2. She has observed that **labs are using STRmix in casework on much more challenging mixtures than are tested during validation**; standard operating procedures do not provide guidance to analysts as to what kinds of samples go beyond the scope of the lab’s validation.

## Thoughts from Kate Philpott (2)

Adjunct Faculty/Research Analyst, VCU Forensic Science Program

Comments continued:

- 3. Validation summaries**, which are often the product of a template supplied by the PGS developer, **do not provide enough information to allow an external reviewer to connect the dots** (i.e., correspondence between samples tested and results obtained). While the full set of validation data would presumably supply the needed information, **labs largely resist efforts to access this information** (even when requested in discovery or pursuant to public records laws), and **there is an unfortunate dearth of requirements expressly related to validation data accessibility**.



## Thoughts from Kate Philpott (3)

Adjunct Faculty/Research Analyst, VCU Forensic Science Program

Comments continued:

4. While ASB Standard 020 requires investigation of mixtures with low and high degrees of allele sharing, SWGDAM does not expressly require this and **many labs have either not investigated the impact of allele sharing at all or have done so in a cursory manner**. Kinship studies – where mixtures are comprised of multiple related individuals, and are tested both against true contributors, and *related* non-contributors – are rarely included in validation studies despite the fact that **scenarios involving multiple related individuals as potential contributors are not uncommon in casework**.

# **Creating a Validation Plan (Internal Validation)**

# Preliminary Work Requested

by the SWGDAM 2015 PGS Validation Guidelines

- “**Prior to validating a probabilistic genotyping system**, the laboratory should ensure that [DNA analysts possess] the appropriate foundational knowledge in the calculation and interpretation of likelihood ratios.” (p. 3)
- “Laboratories should also be aware of the features and limitations of various probabilistic genotyping programs and the impact that those items will have on the validation process.” (p. 3)
- “...prerequisite studies may be required to, for example, establish parameters for allele drop-out and drop-in, stutter expectations, peak height variation, and the number of contributors to a mixture.” (p. 3)

# Preliminary Work Requested

by the SWGDAM 2015 PGS Validation Guidelines

- “**Each laboratory** seeking to evaluate a probabilistic genotyping system **must determine which validation studies are relevant** to the methodology, in the context of its application, **to demonstrate the reliability of the system and any potential limitations.**” (p. 3)
- “The laboratory **must determine the number of samples required to satisfy each guideline** and may determine that a study is not necessary.” (p. 3)

- Don't treat your validation plan as a checklist of tasks
- Think about why each experiment is to be performed and what you hope to learn from it

# FBI Quality Assurance Standards Section 8 on Validation

<https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf/view>

July 1, 2020

8.8.2 New software or new modules of existing software that are used as a component of instrumentation, for the analysis and/or interpretation of DNA data, or for statistical calculations **shall be subject to internal validation specific to the laboratory's intended use prior to implementation in forensic DNA analysis.**

8.8.2.1 Internal software validation studies for new software or new modules of existing software used as a component of instrumentation shall include functional testing and reliability testing.

8.8.2.2 Internal software validation studies for new software or new modules of existing software **for the analysis and/or interpretation of DNA data shall include functional testing, reliability testing, and, as applicable, precision and accuracy studies, sensitivity, and specificity studies.**

8.8.2.3 Internal software validation studies for new software or new modules of existing software **for statistical calculations** shall include functional testing, reliability testing, and, as applicable, precision and accuracy studies.

8.8.2.4 Software that does not impact the analytical process, interpretation, or statistical calculations shall require at a minimum, a functional test.

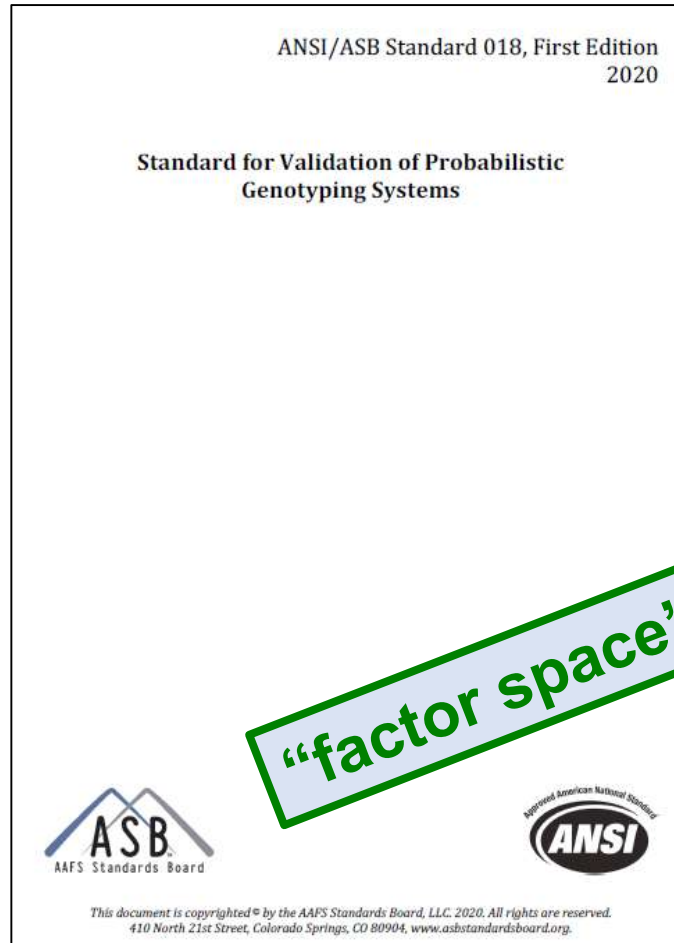
**Applies  
to PGS**

# ASB/OSAC PGS Validation Standard (2020)

(4.1.2) Developmental validation studies shall address accuracy, sensitivity, specificity, and precision and include case-type profiles of known composition

(4.1.3) Internal validation studies shall address...

- **accuracy**
  - establish that PGS calculations are correctly executed
- **sensitivity (with  $H_p$  true,  $LR > 1$ )**
  - assess the ability of PGS to support the presence of a true known contributor
- **specificity (with  $H_d$  true,  $LR < 1$ )**
  - assess the ability of PGS to support the absence of true non-contributors
- **precision**
  - evaluate variation in LRs calculated from repeated analyses of same input data using the same set of conditions/parameters



July 2020

**Case-type profiles:** data exhibiting features that are representative of a plausible range of casework conditions... [including] masked/shared alleles and stutter, degradation (including different degradation levels for different contributors to a mixture), allele and locus drop-out, and PCR inhibition

# Developing an Internal Validation Plan and Testing Samples

ISFG DNA Commission (Coble et al. 2016)

## Recommendation #10:

Before initiating the validation of a software program, the laboratory should **develop a documented validation plan**. The software should have a completed and up to date developmental validation along with other supporting materials such as publications describing the models, propositions and parameters used by the software and a user's manual.

## Recommendation #11:

The laboratory should **test the software on representative data generated in-house** with the reagents, detection instrumentation, and analysis software, used for casework. If a laboratory employs variable DNA typing conditions (e.g., within variation in the amplification and/or electrophoresis conditions to increase or decrease the sensitivity of detection of alleles and/or artifacts), then these types of profiles should also be tested as part of the internal validation plan.

## Recommendation #12:

The laboratory should **consider the range of samples expected to be analyzed in casework to define the scope of application of the software**.

Internal validation should address (1) true donors and non-donors and/or (2) related and unrelated individuals across a range of situations **that span or exceed the complexity of the cases likely to be encountered** in casework.

# Developing an Internal Validation Plan and Testing Samples

ISFG DNA Commission (Coble et al. 2016)

## Recommendation #13:

The laboratory should **determine whether the results produced by the software are consistent with the laboratory's previously validated interpretation procedure** if the data and/or method exist.

*JMB Comments:* Comparing new results back to results obtained with previous manual or software-aided interpretation is valuable to any validation study

- **To assist in this comparison, have previously used DNA samples and data accessible and in a format that permits this comparison**



# Developing a Validation Plan

1. **Define** what aspects of DNA testing process you would like to address in your validation study (e.g., bringing a PGS system online for complex DNA mixtures)
2. **Learn** from previous work
  - Examine available published articles describing developmental validation studies, PGS models and parameters
  - Examine available internal validation studies and talk to others who have performed similar validation studies to learn challenges faced
3. **Decide** on the scope of what “factor space” you want to cover
  - SWGDAM 2016 Validation Guidelines: (4.4) Mixed DNA samples that are **representative of those typically encountered by the testing laboratory should be evaluated**
4. **Design** experiments to cover this factor space
  - Decide on specific DNA samples and conditions to test

How do you define **what is “representative” of casework** encountered in your laboratory?

# Considerations with DNA Samples Used for Testing

**Remember that the goal is to represent the range and difficulty of casework samples in validation studies performed → *sample selection is key***

- Ideally, you want to have sufficient quantities of stable samples to enable testing over time and across software versions as updates are adopted in the future
1. Use of staff DNA samples?
    - May require Institutional Review Board (IRB) approval for human subjects testing
    - Potential privacy concerns for the staff with their genotypes being part of validation data that can be shared (**ideally, you want to be able to share your data for independent review**)
  2. Use of common control samples, such as 9947A and 9948?
    - Limited genotype combinations leading to narrow coverage of your desired factor space; discussed in J.M. Butler (2015) *Adv. Topics in Forensic DNA Typing: Interpretation*, pp. 164-165
    - Harder to effectively measure allele drop-out across STR loci because many of the loci are homozygous, which also limits heterozygote balance studies
  3. Purchase of anonymous blood samples from a blood bank?
    - Will require extraction and preliminary testing to determine STR genotypes
    - An important benefit is that large quantities are available for future studies

# **Experimental Design and Factor Space Coverage**

# How to Perform Validation Studies from an Analytical Chemistry Perspective

- Decide on analytical requirements
  - Sensitivity, resolution, precision, etc.
- **Plan a suite of experiments**
- **Carry out experiments**
- Use data to **assess fitness for purpose**
- Produce a statement of validation
  - Scope of the method

# Assumptions When Performing Validation

- The equipment on which the work is being done is broadly suited to the application. It is clean, well-maintained and **within calibration**.
- The staff carrying out the validation are **competent** in the type of work involved.
- There are **no unusual fluctuations in laboratory conditions** and there is no work being carried out in the immediate vicinity that is likely to cause interferences.
- The **samples being used** in the validation study **are known to be sufficiently stable**.

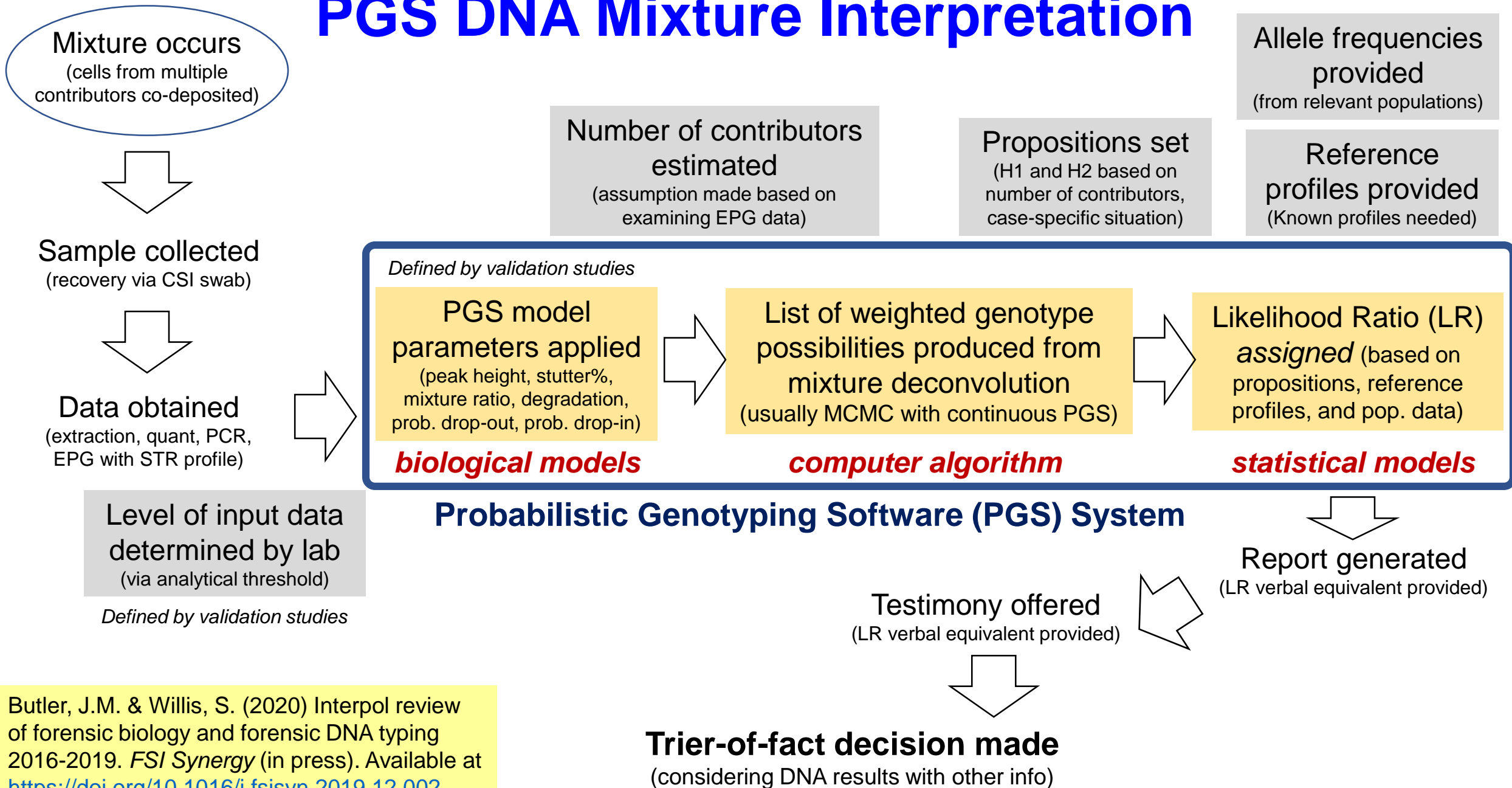
# Tools of Method Validation

- Standard samples
  - positive controls
  - NIST SRMs
- Blanks
- Reference materials prepared in-house and spikes
- Existing samples
- Statistics
- **Common sense**

# Some Thoughts on Experimental Design

- **Purpose and Scope:** Consider the question you are asking and decide what you are going to evaluate
- **Parameters:** Consider carefully the parameters you would like to study and how you can isolate the variables you are trying to examine
- **Coverage:** Explore the “factor space” needed (e.g., to understand the limitations of a method, you will need to go the “edges” and beyond)
- **Replication:** Repeatability (under similar conditions) and reproducibility (under different conditions) need to be understood

# PGS DNA Mixture Interpretation



Butler, J.M. & Willis, S. (2020) Interpol review of forensic biology and forensic DNA typing 2016-2019. *FSI Synergy* (in press). Available at <https://doi.org/10.1016/j.fsisyn.2019.12.002>



# Factors Influencing LR Values Determined by PGS Systems

Input	By Who	Impact/Example
Modeling choices	PGS system architect(s)	Peak height ratio variance allowed, how potential degradation is modeled, etc.
Data input choices	DNA analyst	Defining alleles (setting analytical threshold), categorizing artifacts from alleles (e.g., stutter)
Proposition choices and assumptions	DNA analyst	Use of unrelated individuals vs. relatives or conditioning on a victim's profile with an intimate sample
Population database choices	DNA analyst/ laboratory policy	Different allele frequency values will influence LR values
Reporting statistic choices	DNA analyst/ laboratory policy	Handling sampling variation (e.g., HPD*)

\*HPD=highest posterior density-defines interval most likely to contain the true value

# Some Factors That May Affect Reliability of an LR System

1. Sample
  - a) Sample amount (contributor template amounts)
  - b) Sample quality (degradation level)
2. Labs
  - a) Kits used
  - b) Equipment Used
  - c) Number of PCR cycles
  - d) Analyst
  - e) Choice of Analytical Threshold (AT)
3. Probabilistic Genotyping (PG) Model
  - a) Choice of model
  - b) Choice of laboratory specific parameters for use in the PG model
  - c) Propositions Chosen ( $H_p$  and  $H_d$ )**
4. Software Implementing the PG Model
  - a) Choice of numerical methods for computing LR (MCMC, Numerical Integration)
  - b) Choice of number of iterations OR numerical integration parameters (e.g. grid size)

**Slide from Hari's  
Module 2 Presentation**

**FACTOR  
SPACE**

# “Factor Space” in DNA Mixture Studies

- 1. Total DNA amount** (e.g., 1 ng or 100 pg)
  - Consider lowest amount of DNA in a minor contributor (be informed by sensitivity studies)
- 2. Sample quality** (DNA degradation or PCR inhibition)
- 3. Number of contributors**
  - *Factor space expands rapidly as the number of contributors increases\**
  - Sample types can differ, e.g., 2-person [sexual assault] or >4-person [touch evidence]
- 4. Degree of allele overlap** across mixture components
  - Minor contributor alleles in stutter positions of major contributor alleles
  - Mixtures involved multiple related individuals are expected to possess high allele sharing
  - *Rarely discussed in published studies or sample design (yet known to impact deconvolution)*
- 5. Contributor component ratios** (e.g., 10:1 or 1:1:1)
  - Rarely is interpretation performed beyond a 10:1 or 20:1 mixture
  - General kinds: balanced ( $\approx 1:1:1$ ), major/minor ( $\approx 7:2:1$ ), extreme ( $\approx >20:1:1$ )

\*Lynch & Cotton (2018) Determination of the possible number of genotypes which can contribute to DNA mixtures... *FSI Genetics* 37: 235-240

# An Example Experimental Plan for Internal Validation

provided by Bright & Coble in their new book

34 amplifications, if done in duplicate, then 68 samples would be generated

Number of Contributors	Range of Mixture Ratios	Total Template Amplified	DNA Amount of Smallest Contributor	Total Number of Mixtures Examined
2	1:1, 5:1, 10:1, 20:1, 100:1	1.0 & 0.5 ng	6.25 pg	10
3	1:1:1, 10:5:1, 3:2:1, 20:5:1	1.0 & 0.5 ng	6.25 pg	8
4	1:1:1:1, 10:5:2:1, 4:3:2:1, 8:4:1:1	1.0 & 0.5 ng	6.25 pg	8
5	1:1:1:1:1, 10:5:3:2:1, 6:3:2:1:1, 5:4:3:2:1	1.0 & 0.5 ng	6.25 pg	8

*This testing plan does not consider the degree of allele sharing, alleles in stutter positions, degradation/inhibition/allele drop-out, or mixtures with relatives*



Contents lists available at ScienceDirect

## Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



Research paper

### An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors



Corina C.G. Benschop\*, Alwart Nijveld, Francisca E. Duijs, Titia Sijen

*Netherlands Forensic Institute, Division of Biological Traces, Laan van Ypenburg 6, 2497GB The Hague, the Netherlands*

Also discussed in Chapter 9 “Validation” (pp. 277-308) of Peter Gill, Øyvind Bleka, Oskar Hansson, Corina Benschop and Hinda Haned (2020) *Forensic Practitioner’s Guide to the Interpretation of Complex DNA Profiles* (Elsevier Academic Press, San Diego)

# Multiple Donor Combinations Used to Create Different Degrees of Allele Sharing

**Table 1**

Overview of the six donor combinations used for mixture preparation.

Dataset number	Type of dataset	Number of contributors			
		2	3	4	5
		Donor combinations per dataset			
1	High allele sharing	a:b	a:b:c	a:b:c:d	a:b:c:d:e
2	Low allele sharing	f:g	f:g:h	f:g:h:i	f:g:h:i:j
3	Random	k:l	k:l:k	k:l:k:n	k:l:m:n:o
4	Random	p:q	p:q:r	p:q:r:s	p:q:r:s:t
5	Random	u:v	u:v:w	u:v:w:x	u:v:w:x:y
6	Random	z:aa	z:aa:ab	z:aa:ab:ac	z:aa:ab:ac:ad

**Specific genotypes can be kept anonymous and still differentiate various degrees of allele sharing**

# Different Categories of Mixture Types Were Studied in Exploring the DNA Mixture Factor Space

**Table 2**  
Mixture proportions and amounts of DNA used per donor to create a total of 20 different mixtures per dataset.

Mixture Type	Number of contributors			
	2	3	4	5
	Picograms DNA per contributor			
A: major 2x more than any minor	300:150	300:150:150	300:150:150:150	300:150:150:150:150
B: major 10x more than any minor	300:30	300:30:30	300:30:30:30	300:30:30:30:30
C: 2 majors with equal amount	150:150	150:150:60	150:150:60:60	150:150:60:60:60
D: major 5 to 2.5x more than minors	150:30	150:30:60	150:30:60:30	150:30:60:30:30
E: major 20 to 10x more than minors	600:30	600:30:60	600:30:60:30	600:30:60:30:30
Number of mixtures	5	5	5	5

**“Factor Space”  
in DNA Mixture  
Studies**

1. **Total DNA amount**
2. **Sample quality**
3. **Number of contributors**
4. **Degree of allele overlap**
5. **Contributor component ratios**

PGS System (Version)	# of Samples	# of Contributors	# of Replicates	DNA Amount (pg)	Mixture Ratio
EuroForMix (Various: v1.9.1 up to v1.11.4)	5 HAS, 5 LAS, 20 RAS	2	3	300:150	2:1
				300:30	10:1
				150:150	1:1
				150:30	5:1
				600:30	20:1
				300:150:150	2:1:1
	5 HAS, 5 LAS, 20 RAS	3	3	300:30:30	10:1:1
				150:150:60	2.5:2.5:1
				150:30:60	5:1:2
				600:30:60	20:1:2
				300:150:150:150	2:1:1:1
	5 HAS, 5 LAS, 20 RAS	4	3	300:30:30:30	10:1:1:1
				150:150:60:60	2.5:2.5:1:1
				150:30:60:30	5:1:2:1
				600:30:60:30	20:1:2:1
	5 HAS, 5 LAS, 20 RAS	5	3	300:150:150:150:150	2:1:1:1:1
300:30:30:30:30				10:1:1:1:1	
150:150:60:60:60				2.5:2.5:1:1:1	
150:30:60:30:30				5:1:2:1:1	
			600:30:60:30:30	20:1:2:1:1	

**NIST**  
**Summary of Factor Space Coverage**  
 from this Netherlands Forensic Institute study

Allele sharing levels  
**HAS:** high allele sharing  
**LAS:** low allele sharing  
**RAS:** random allele sharing

Data available from their studies:  
<http://www.euroformix.com/data>

Benschop et al. (2019) An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors. *Forensic Sci. Int. Genet.* 42: 31-38.

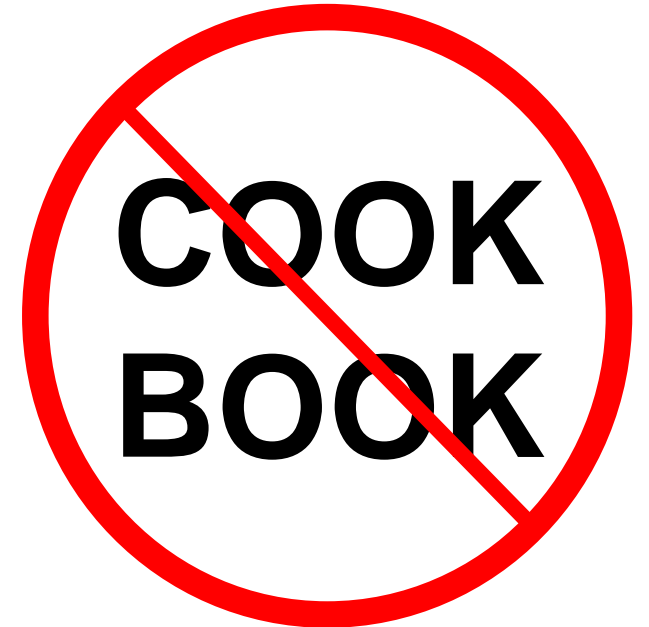


# Our Goal for This Workshop

To Review Important Principles to  
Aid Understanding of Validation...

## Key Aspects of Validation:

- How to **Design** Validation Studies
- How to **Perform** Validation Studies
- How to **Describe** Validation Studies
- How to **Utilize** Validation Data



In Module 4, Hari will examine some data examples for reliability assessment of LR results produced by PGS

# Thank you for your attention!

**Points of view are the presenters** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

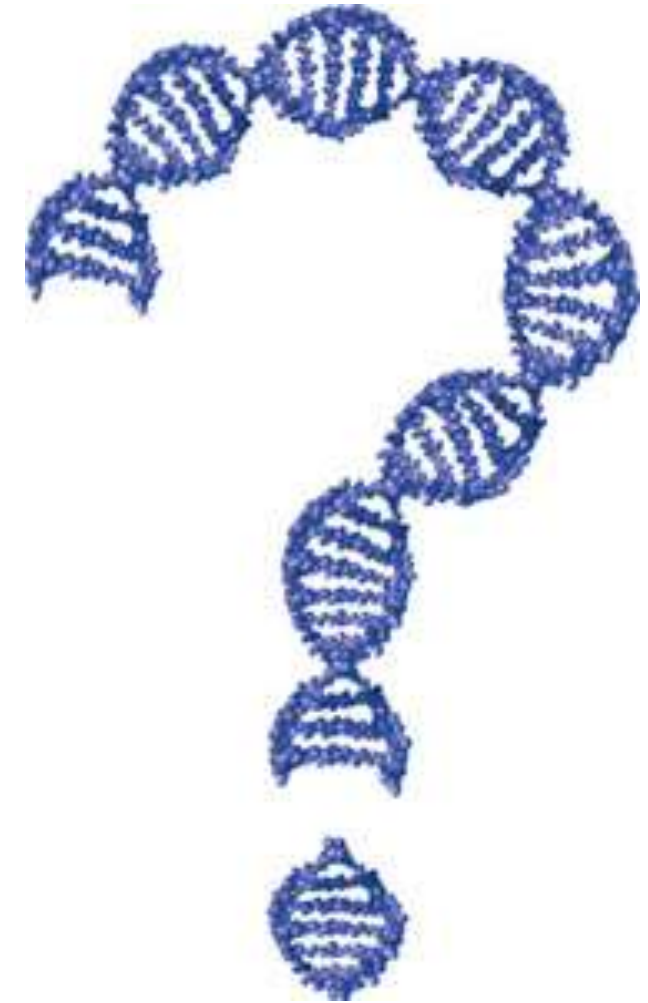
## Contact Information

**John M. Butler**

[john.butler@nist.gov](mailto:john.butler@nist.gov)

**Hari K. Iyer**

[hariharan.iyer@nist.gov](mailto:hariharan.iyer@nist.gov)



RESEARCH. STANDARDS. FOUNDATIONS.



**ISHI 2020 Validation Workshop**  
**Friday September 18th, 2020 // 9:00 am - 12:30 pm**

Validation Principles, Practices, Parameters,  
Performance Evaluations, and Protocols

# Reliability Assessment of LR Systems: Data Examples

## Module 4

**Hari K. Iyer**

National Institute of Standards and Technology



# Acknowledgements and Disclaimers

I would like to thank Steve Lund, William Guthrie, Antonio Possolo, Adam Pintar, Jan Hannig, and other NIST colleagues, for many ongoing, meaningful discussions on foundational concepts in statistics.

I wish to also thank John Butler, Katherine Gettings, Niki Osborne, Rich Press, Sarah Riman, Melissa Taylor, Pete Vallone, and Sheila Willis – and the DNA Mixture Resource Group, for educating me on DNA mixture interpretation and/or related issues.

**Points of view are of the presenters** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

# Topics for Discussion

- Validation studies for assessing LR system reliability
- Expected behavior of reliable LR systems
- Diagnostic checks of LR system performance
- Statistical tools for assessing discrimination power of LR systems
- Statistical tools for assessing calibration accuracy of LR systems
- Study design and sample size issues (briefly)
- Conclusions

# Validation Studies

Before using an LR system in casework, labs conduct validation studies to assess LR system reliability.

The LR System includes:

- Measurement step that produces an EPG
- Analyst interpretation of the EPG for preparing input to the software
- The PG model and the software that implements the model calculations
- Deciding if the results make sense and what LR to report.

**Does the system produce results that are consistent with what one would expect (since ground truth is known)?**

**What are these expectations?**

# Expected Behavior of LR Systems

- We expect LR for known contributors to be  $> 1$ .

If a known contributor LR is less than 1 we say that this is a **misleading LR**.

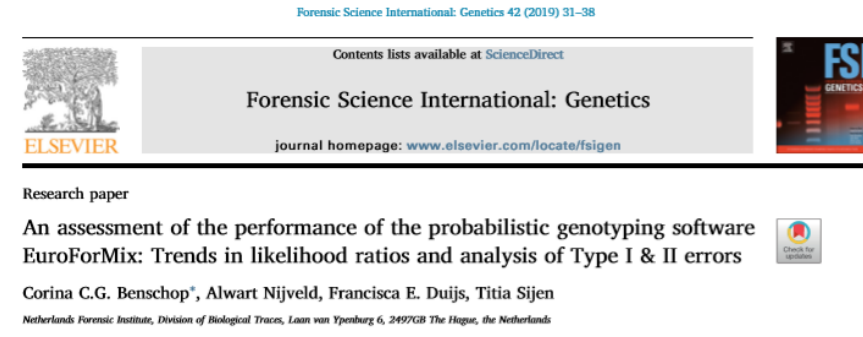
Sometimes labeled **type-I error** (Benschop, et. al, 2019)

- We expect LR for known non-contributors to be  $< 1$ .

If a known non-contributor LR is greater than 1 we say that this is a **misleading LR**.

Sometimes labeled **type-II error** (Benschop, et. al, 2019)

Well designed validation studies can provide information that can help assess the chances of obtaining misleading LRs in casework.



# Expected Behavior of LR Systems

- As information content increases, larger LR values are expected for true contributors and smaller LR values are expected for non-contributors.

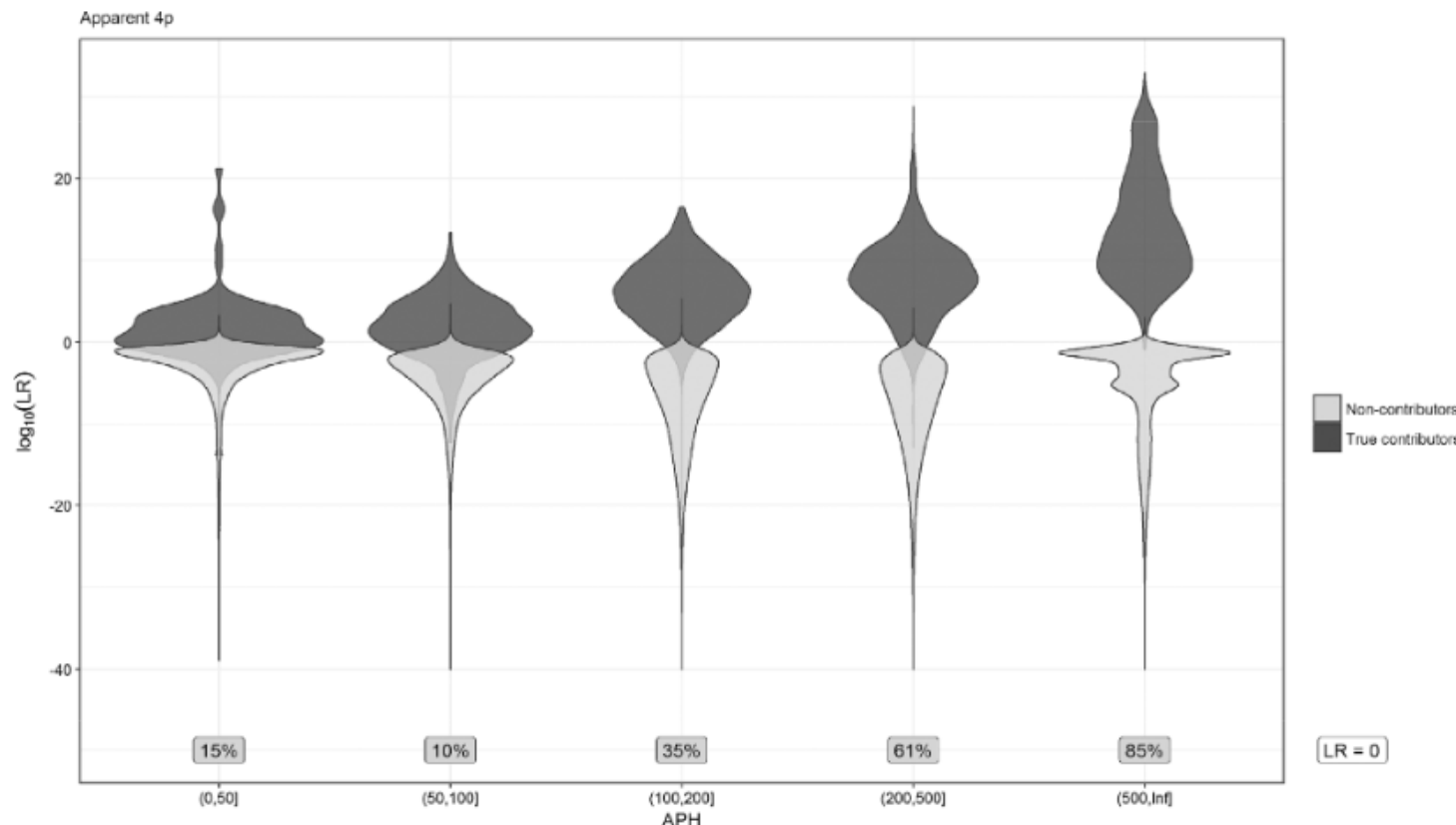


Fig. 5. Violin plot of  $\log_{10}(LR)$  versus APH for apparent four contributor mixtures.

Forensic Science International: Genetics 34 (2018) 11–24



ELSEVIER

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



Research paper

Internal validation of STRmix™ – A multi laboratory response to PCAST

Jo-Anne Bright<sup>a,\*</sup>, Rebecca Richards<sup>a</sup>, Maarten Kruijver<sup>a</sup>, Hannah Kelly<sup>a</sup>, Catherine McGovern<sup>a</sup>, Alan Magee<sup>b</sup>, Andrew McWhorter<sup>c</sup>, Anne Cieccko<sup>d</sup>, Brian Peck<sup>e</sup>, Chase Baumgartner<sup>f</sup>, Christina Buettner<sup>g</sup>, Scott McWilliams<sup>g</sup>, Claire McKenna<sup>h</sup>, Colin Gallacher<sup>i</sup>, Ben Mallinder<sup>j</sup>, Darren Wright<sup>j</sup>, Deven Johnson<sup>k</sup>, Dorothy Catella<sup>l</sup>, Eugene Lien<sup>m</sup>, Craig O'Connor<sup>m</sup>, George Duncan<sup>n</sup>, Jason Bundy<sup>o</sup>, Jillian Echard<sup>p</sup>, John Lowe<sup>q</sup>, Joshua Stewart<sup>r</sup>, Kathleen Corrado<sup>s</sup>, Sheila Gentile<sup>s</sup>, Marla Kaplan<sup>t</sup>, Michelle Hassler<sup>u</sup>, Naomi McDonald<sup>v</sup>, Paul Hulme<sup>w</sup>, Rachel H. Oefelein<sup>x</sup>, Shawn Montpetit<sup>y</sup>, Melissa Strong<sup>y</sup>, Sarah Noël<sup>z</sup>, Simon Malsom<sup>a</sup>, Steven Myers<sup>b</sup>, Susan Welti<sup>c</sup>, Tamyra Moretti<sup>d</sup>, Teresa McMahon<sup>e</sup>, Thomas Grill<sup>f</sup>, Tim Kalafut<sup>g</sup>, MaryMargaret Greer-Ritzheimer<sup>h</sup>, Vickie Beamer<sup>i</sup>, Duncan A. Taylor<sup>j,k</sup>, John S. Buckleton<sup>a,l</sup>





# Expected Behavior of LR Systems

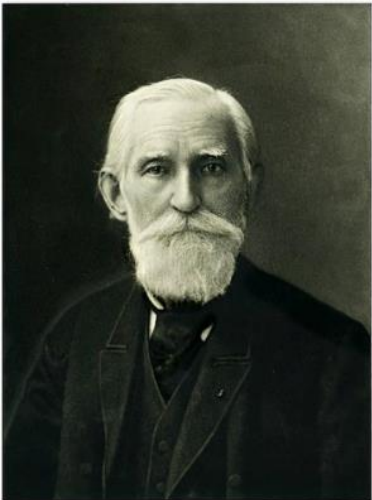
- **If the model is correct**

(a) Average of non-contributor LRs is expected to be 1. (Often attributed to Alan Turing).

(b) The chance of a non-contributor giving an  $LR=x$  or greater should be less than or equal to  $1/x$ .

(Markov-Chebyshev Inequality; sometimes also credited to Alan Turing).

Pafnuty Chebyshev



1821-1894

Andrei Andreyevich Markov



1856-1922

If  $N$  non-contributor tests are conducted, we expect the number of LRs that equal or exceed  $x$  to be at most  $N/x$ .

In  $N=10,000$  non-contributor tests, we expect the number of LRs that equal or exceed 10,000 to be at most 1; the number of LRs that equal or exceed 1000 to be at most  $N/1000 = 10000/1000 = 10$ .

# Conditions Necessary But Not Sufficient

- If the empirical results are not consistent with these expectations one might conclude that the model needs to be improved.
- If the empirical results ARE consistent with these expectations **one CANNOT conclude that the model is correct.** That requires more work.

# Conditions Necessary But Not Sufficient - Example

If you multiply two odd integers the resulting integer will also be odd. This observation can help check accuracy of calculations.

- $709463783 \times 184592267 = 130\ 961\ 528\ 058\ 366\ 162$  (is wrong)
- $709463783 \times 184592267 = 130\ 761\ 528\ 058\ 366\ 061$  (is this correct ?)
- $709463783 \times 184592267 = 130\ 961\ 528\ 058\ 366\ 061$  (is this correct ?)

“Passing” the Turing test is NECESSARY (but not SUFFICIENT)

“Passing” the Turing test DOES NOT demonstrate RELIABILITY

However, some individuals may be convinced of system reliability based on simple diagnostic checks. Others may not be convinced without more rigorous testing.

# Main Criteria for Reliability

- Distribution of true contributor LRs and the distribution of non-contributor LRs should be well-separated.  
(**Discrimination power**)
- Reported LRs should be consistent with empirically observed behavior of frequencies of contributor and noncontributor LRs.  
(**Calibration accuracy**)

# A Data Example

## PROVEDIt data

- 4P mixtures
- 1:1:1:1
- Total DNA amount < 125 pg
- Degraded
- Total of 63 mixtures
- 63 Known Contributor Tests & 63 Non-contributor Tests

When reporting deductions made using the PROVEDIt data, the following paper describing the database should be cited:

Alfonse, L.E., Garrett, A.D., Lun, D.S., Duffy, K.R. & Grgicak, C.M. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Sci. Int. Genetics* 32, 62-70. DOI:10.1016/j.fsigen.2017.10.006

PROVEDIt Database Naming Convention & Laboratory Methods

PROVEDIt\_1-5-Person CSVs Filtered.zip

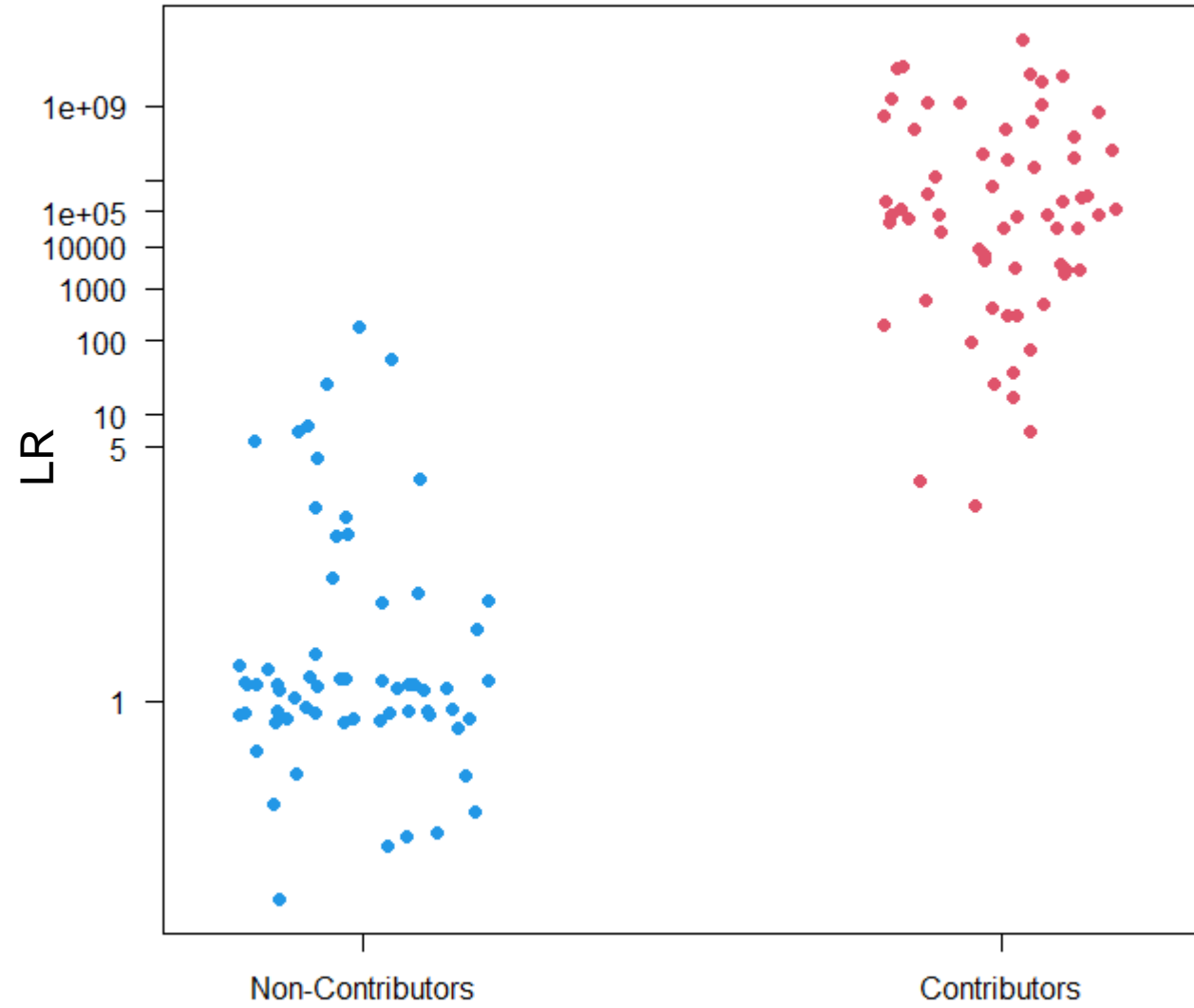
2017-10-12 09:15 22M

3500xL, GlobalFiler, 29 cycles, 15 sec Injection time  
Used 'filtered' samples (artifacts already removed)

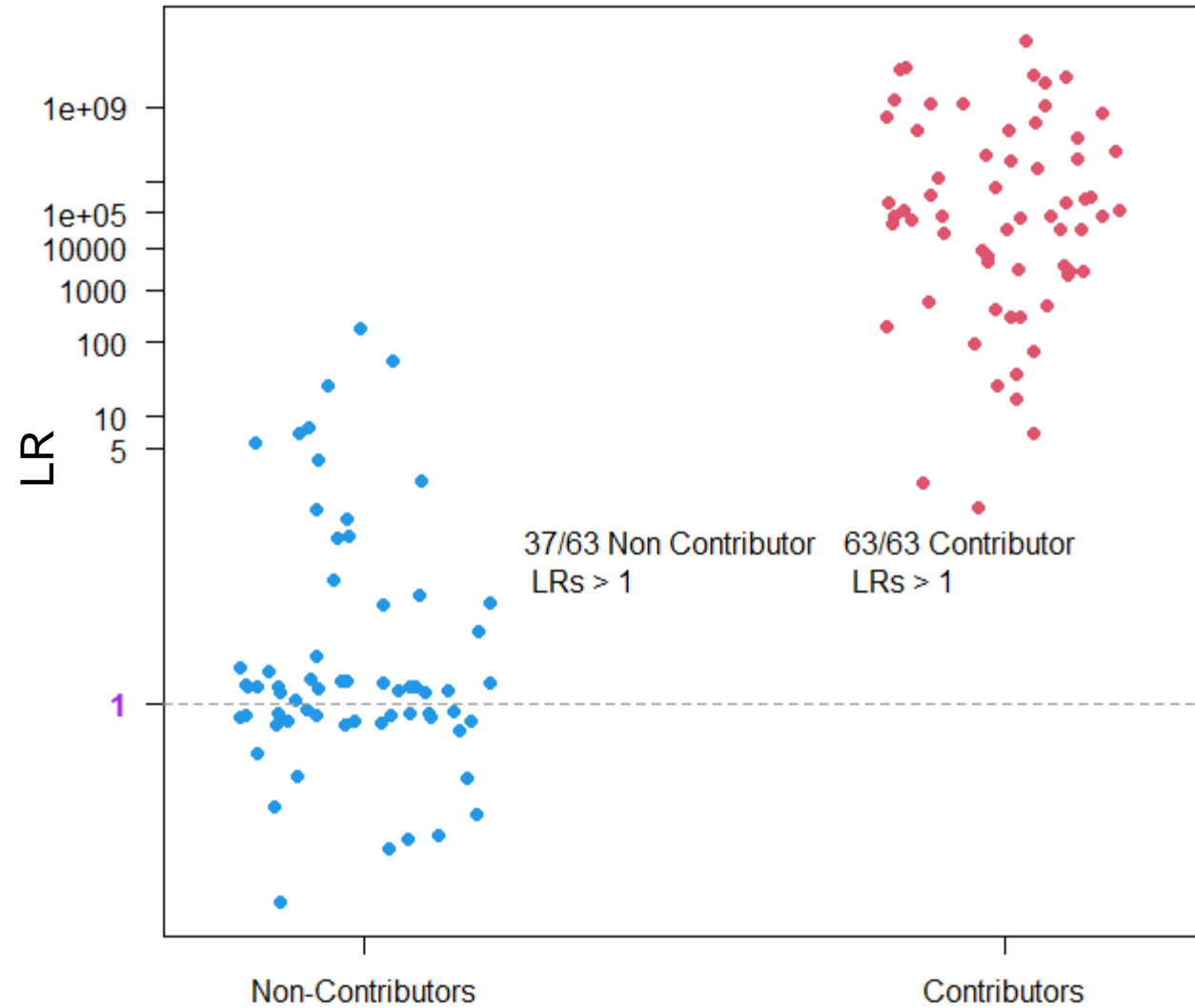
Noncontributor profiles randomly selected from the NIST 1036 sample database for each of the 63 samples

This resulted in 63 true contributor LRs & 63 non-contributor LRs

PROVEDIt Data, 4P Mixtures  
Degraded, Total Amount < 125 pg, Ratio 1:1:1:1

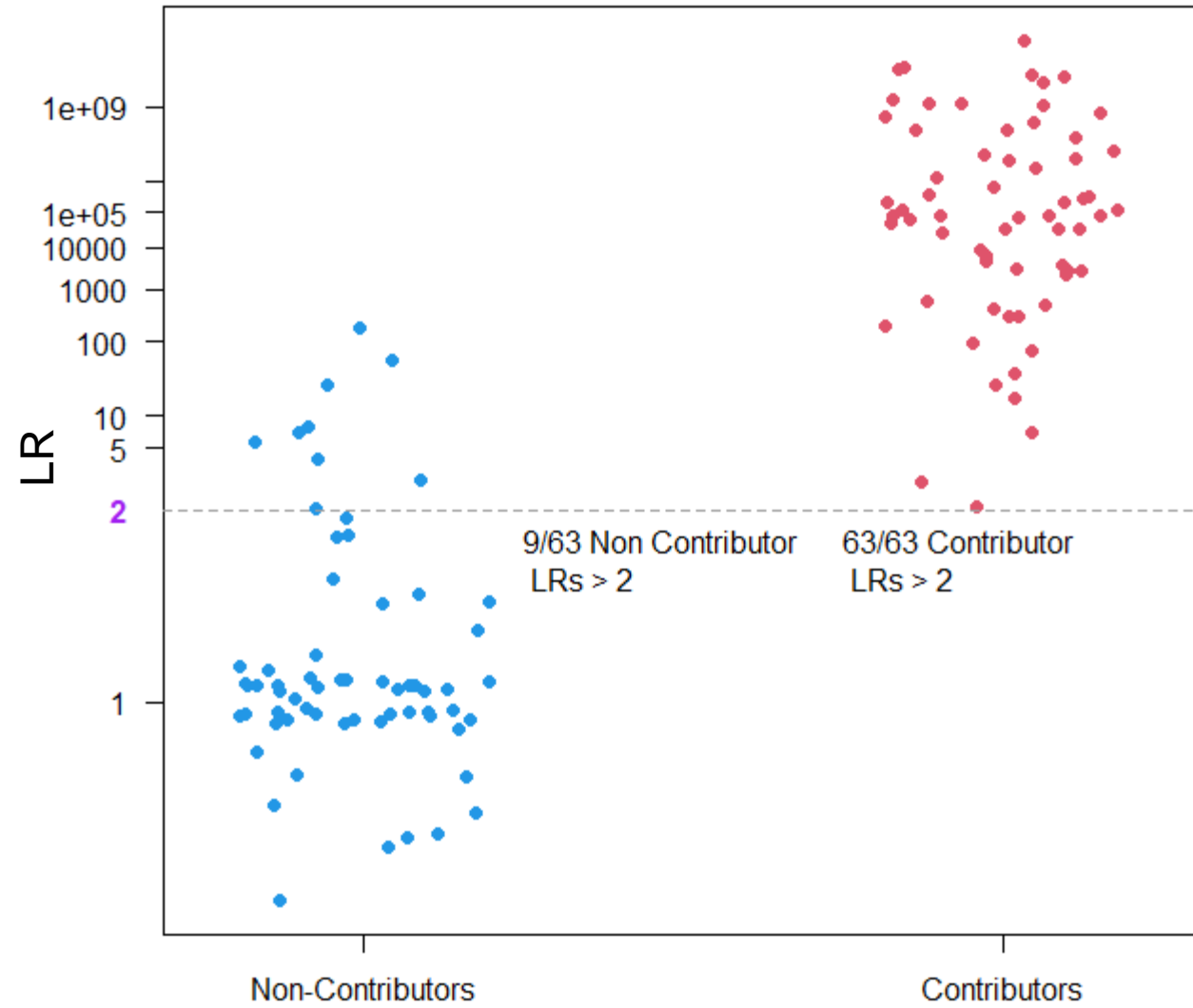


PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2			
5			
10			
100			
1000			

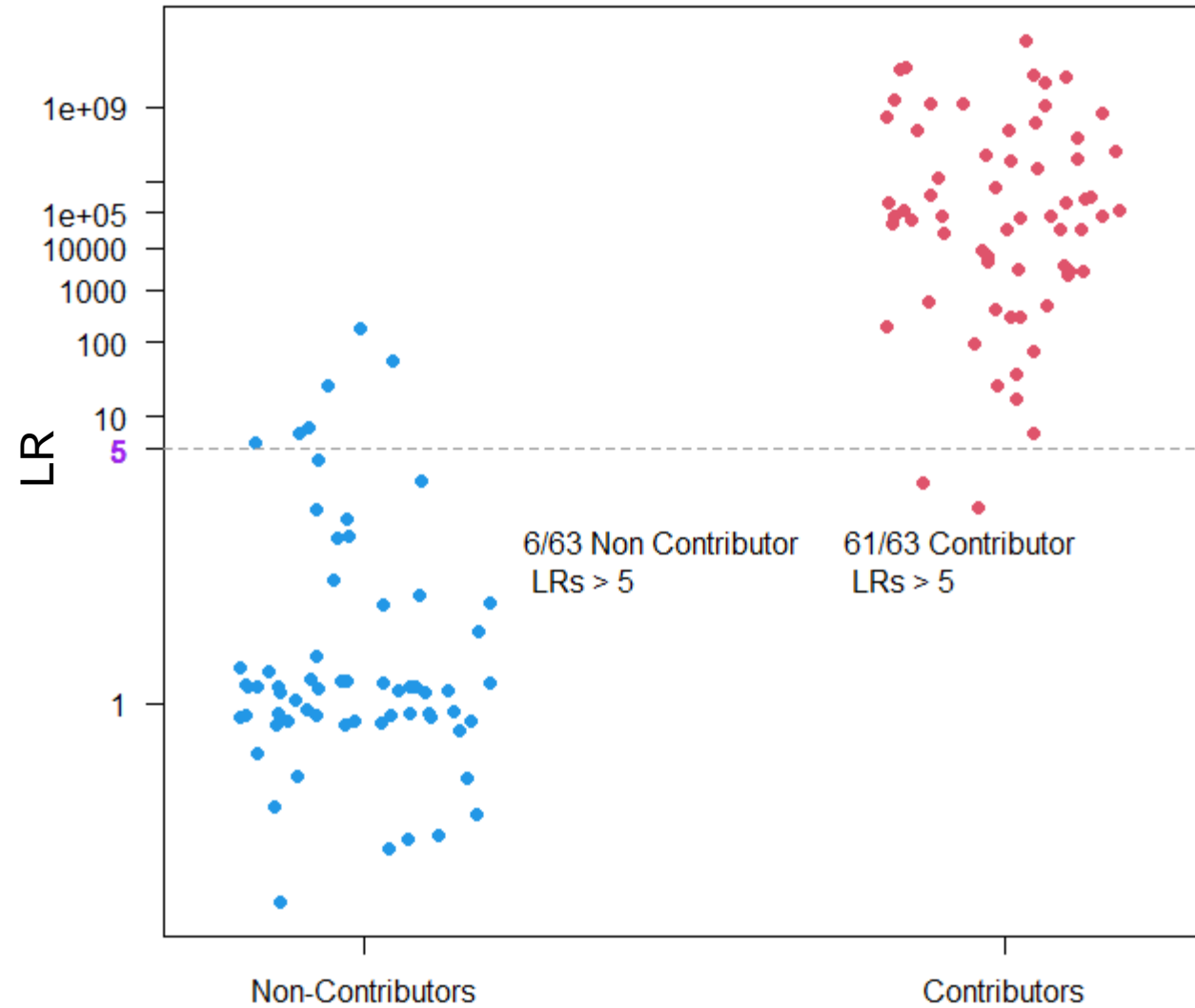
PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5			
10			
100			
1000			

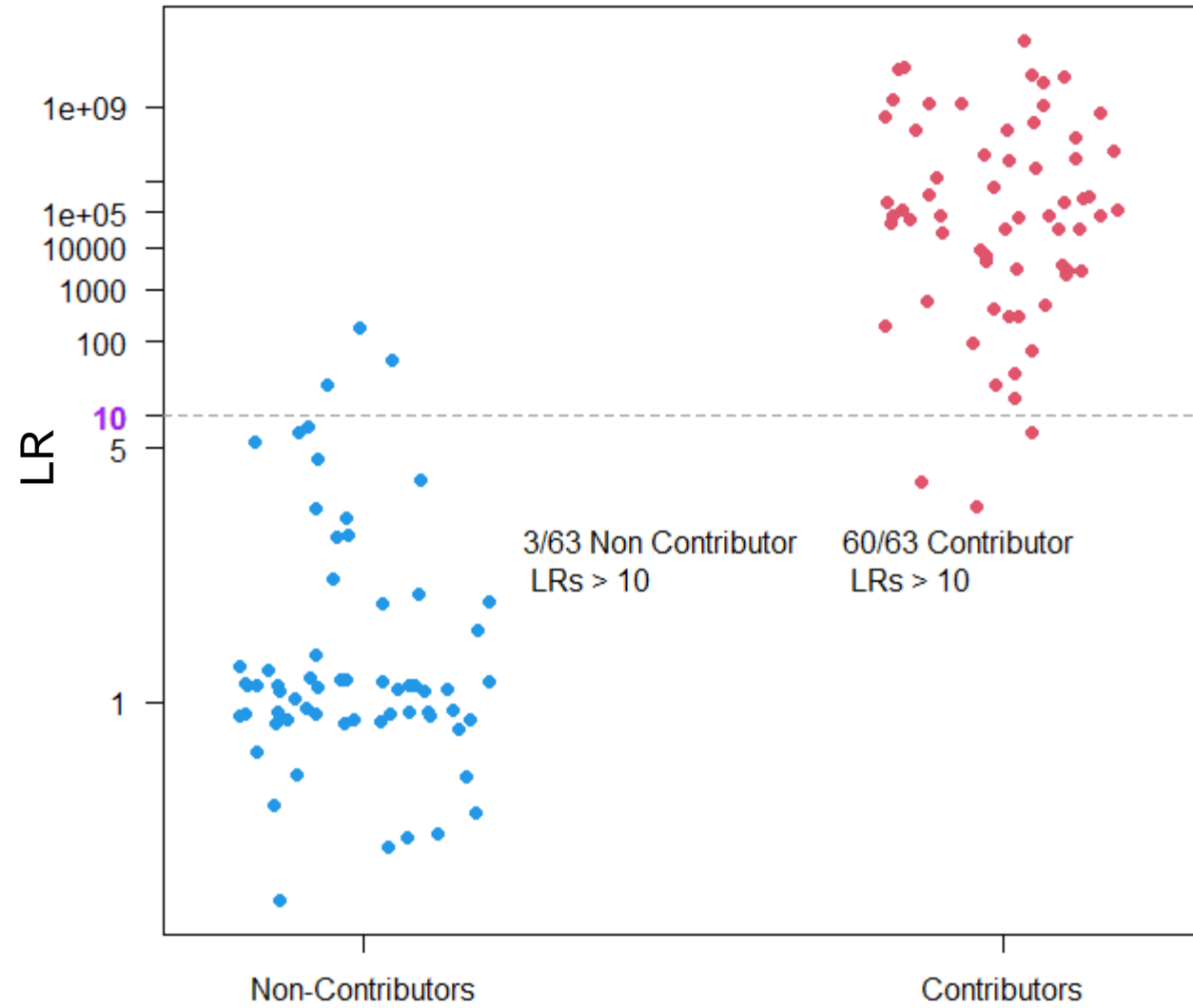


PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



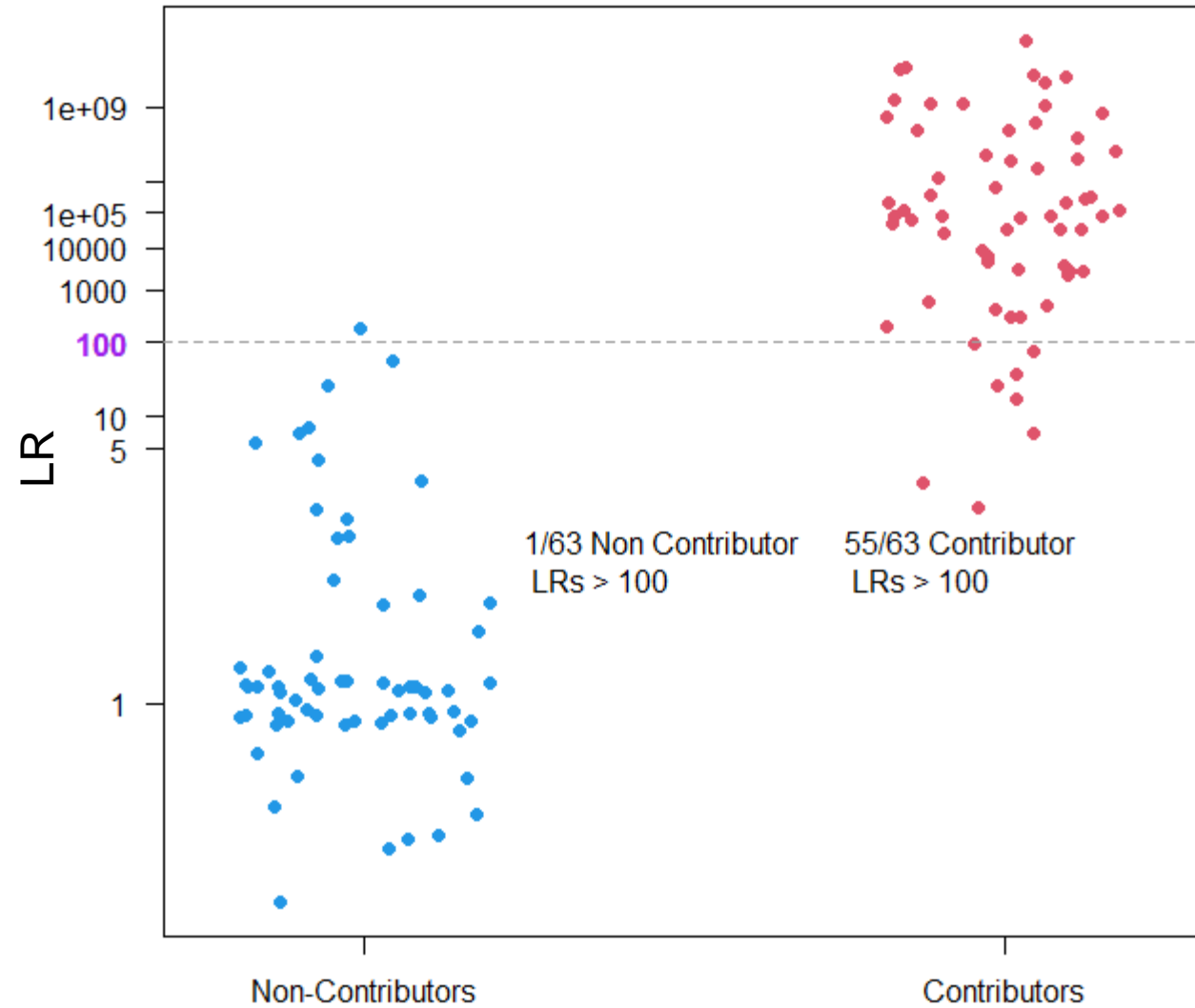
	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10			
100			
1000			

PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



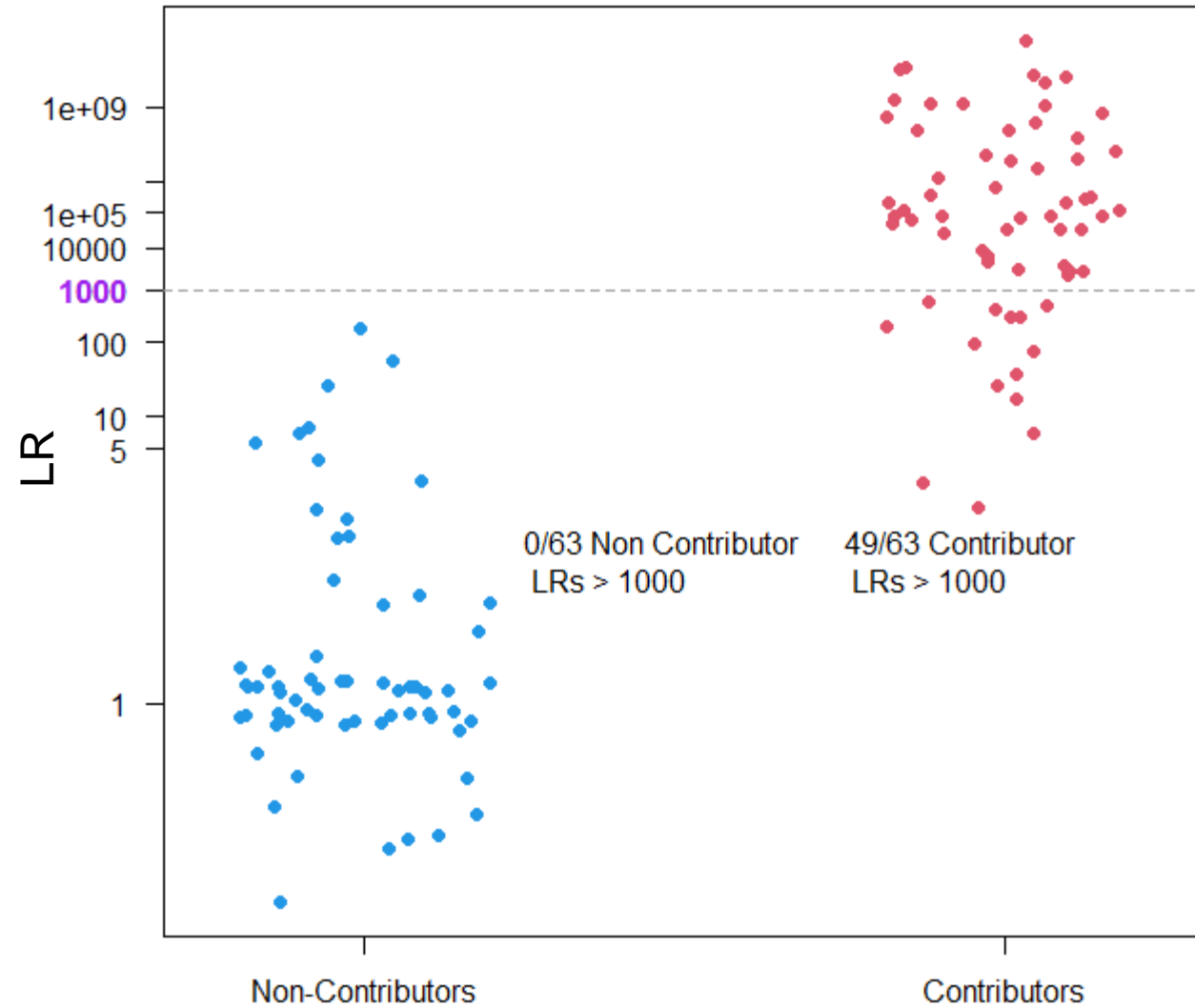
	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100			
1000			

PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



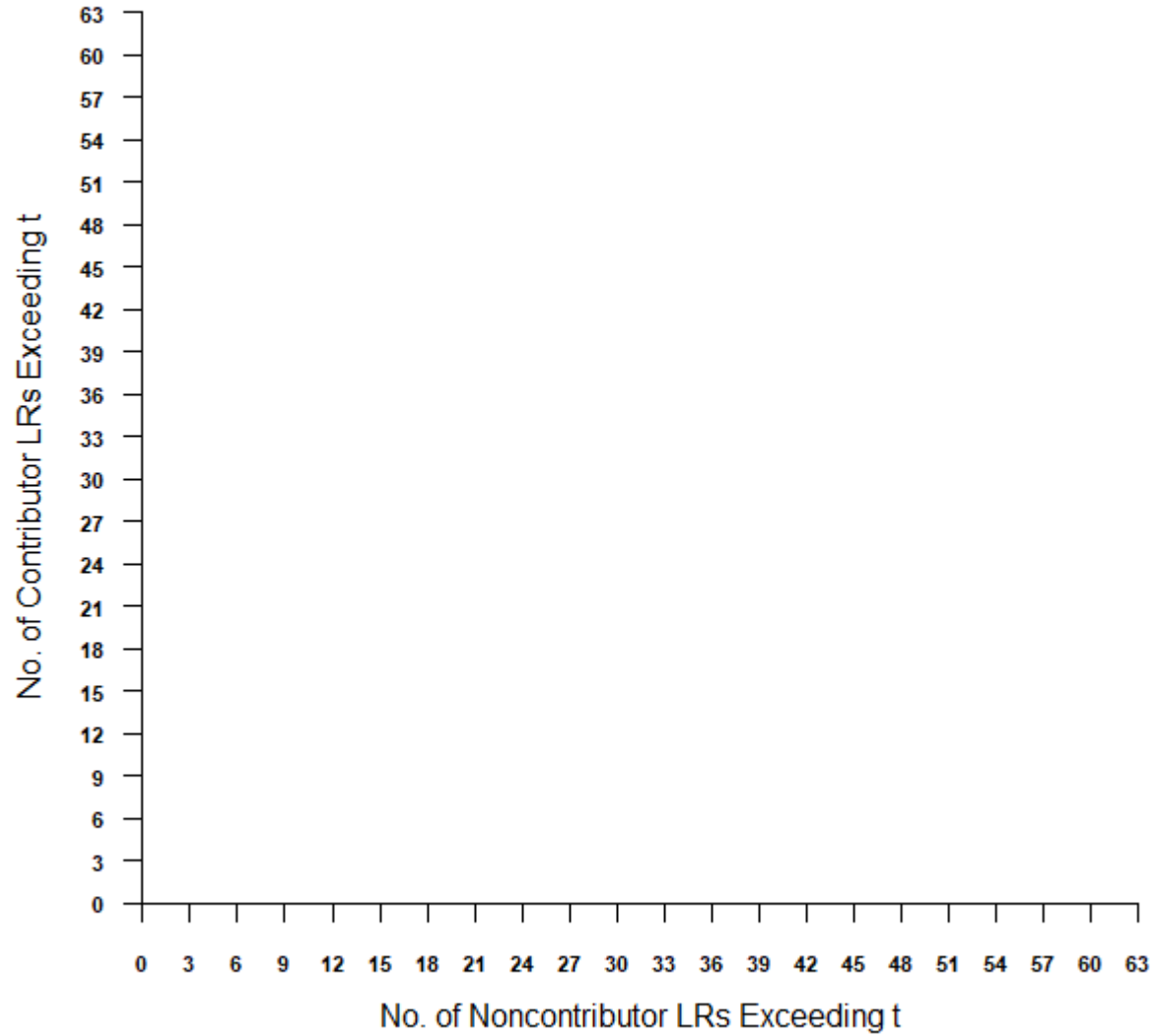
	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000			

PROVEDIt Data, 4P Mixtures  
 Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



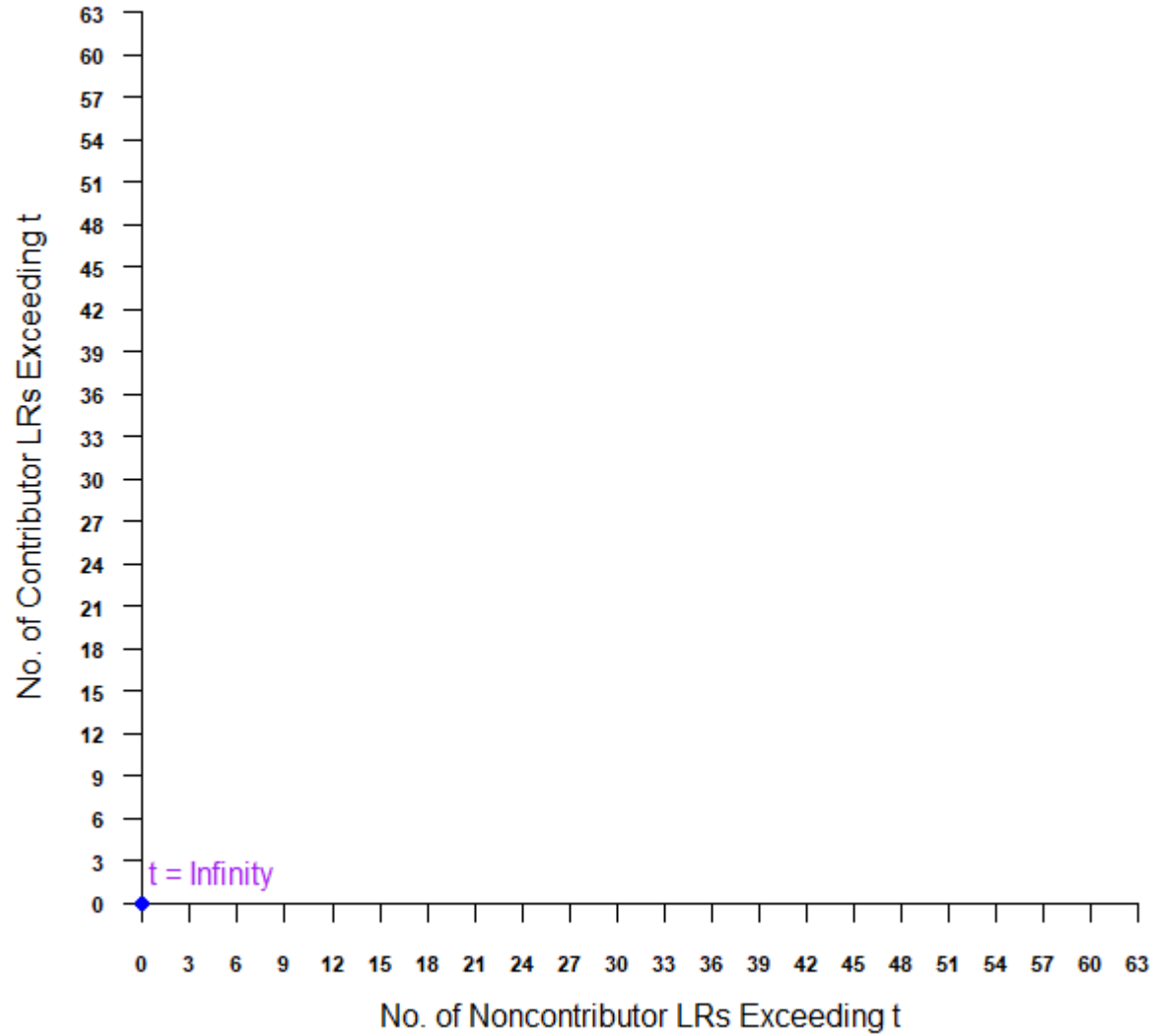
	LR > t		LR < t
t equal to	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14

No. of LRs Exceeding LR = t



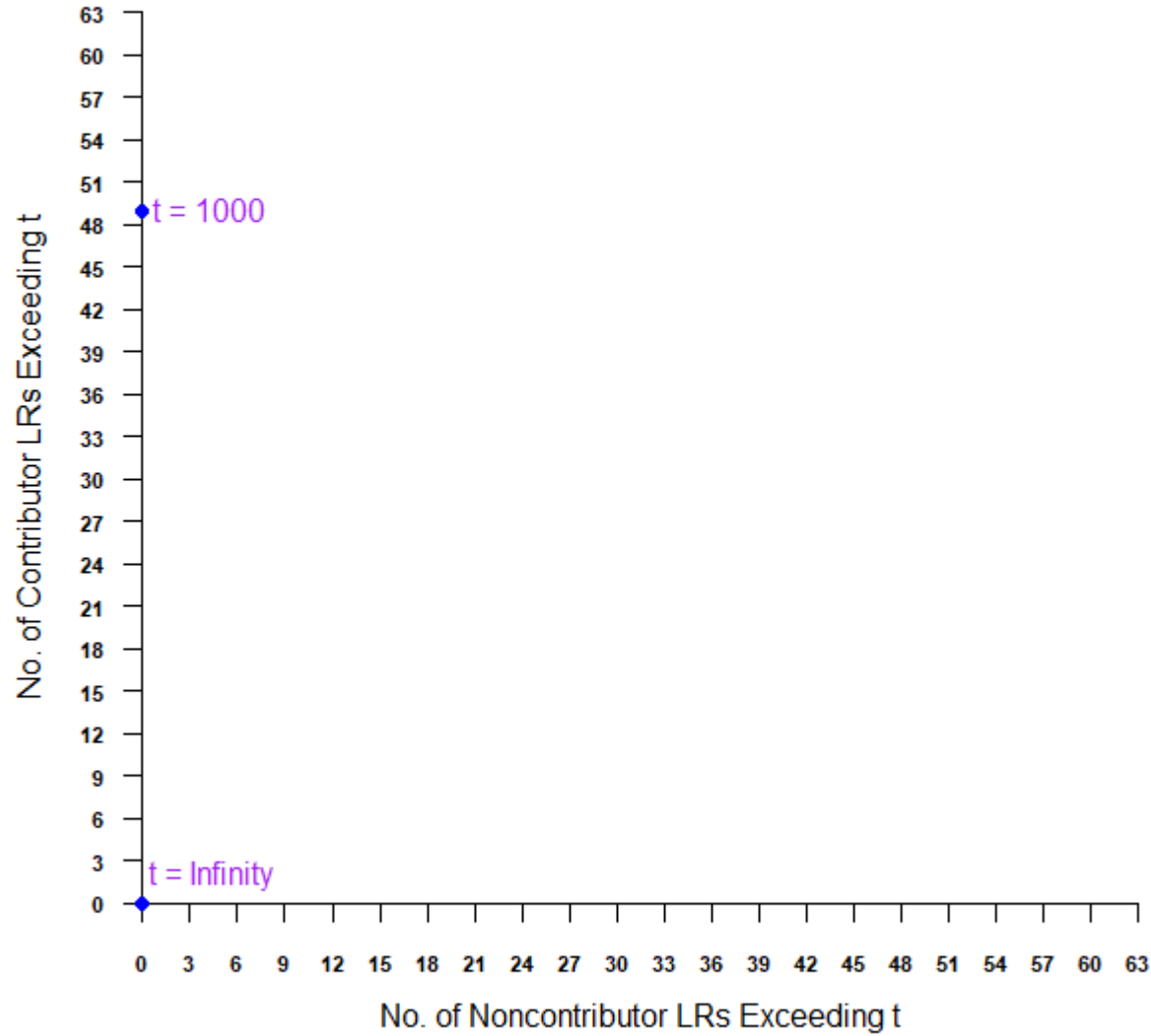
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LRs Exceeding LR = t



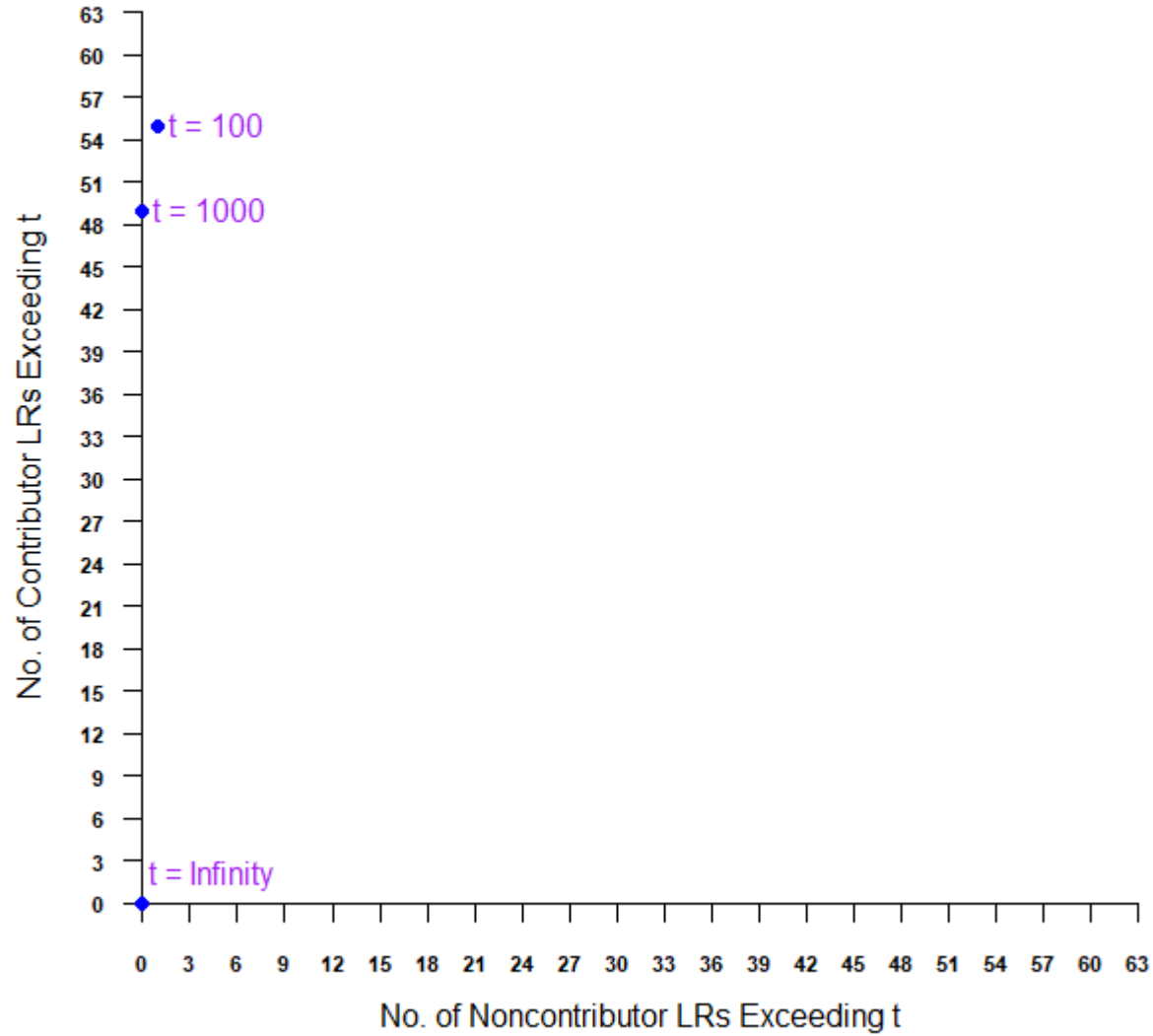
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LR Exceeding LR = t



t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

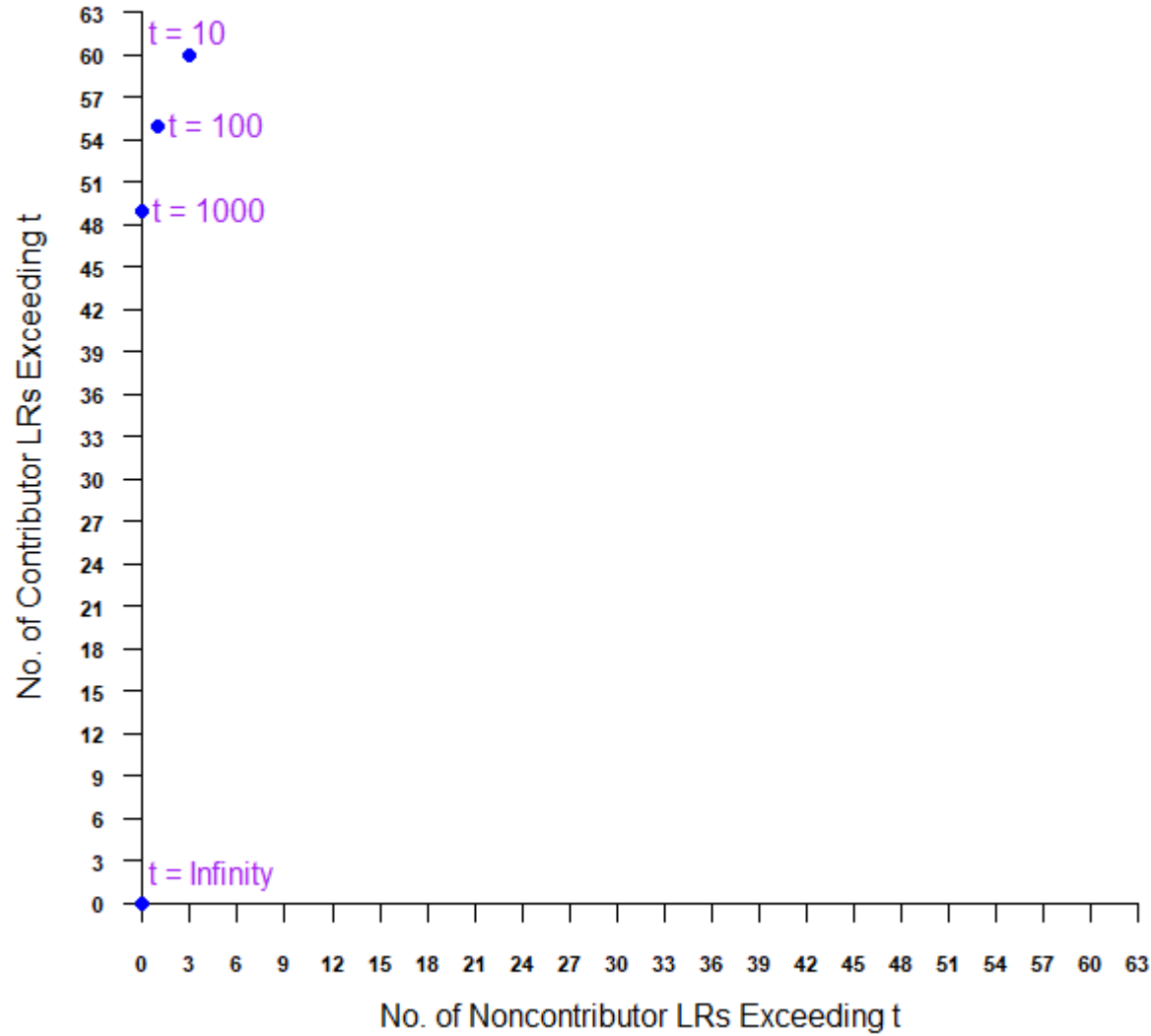
### No. of LR Exceeding LR = t



t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

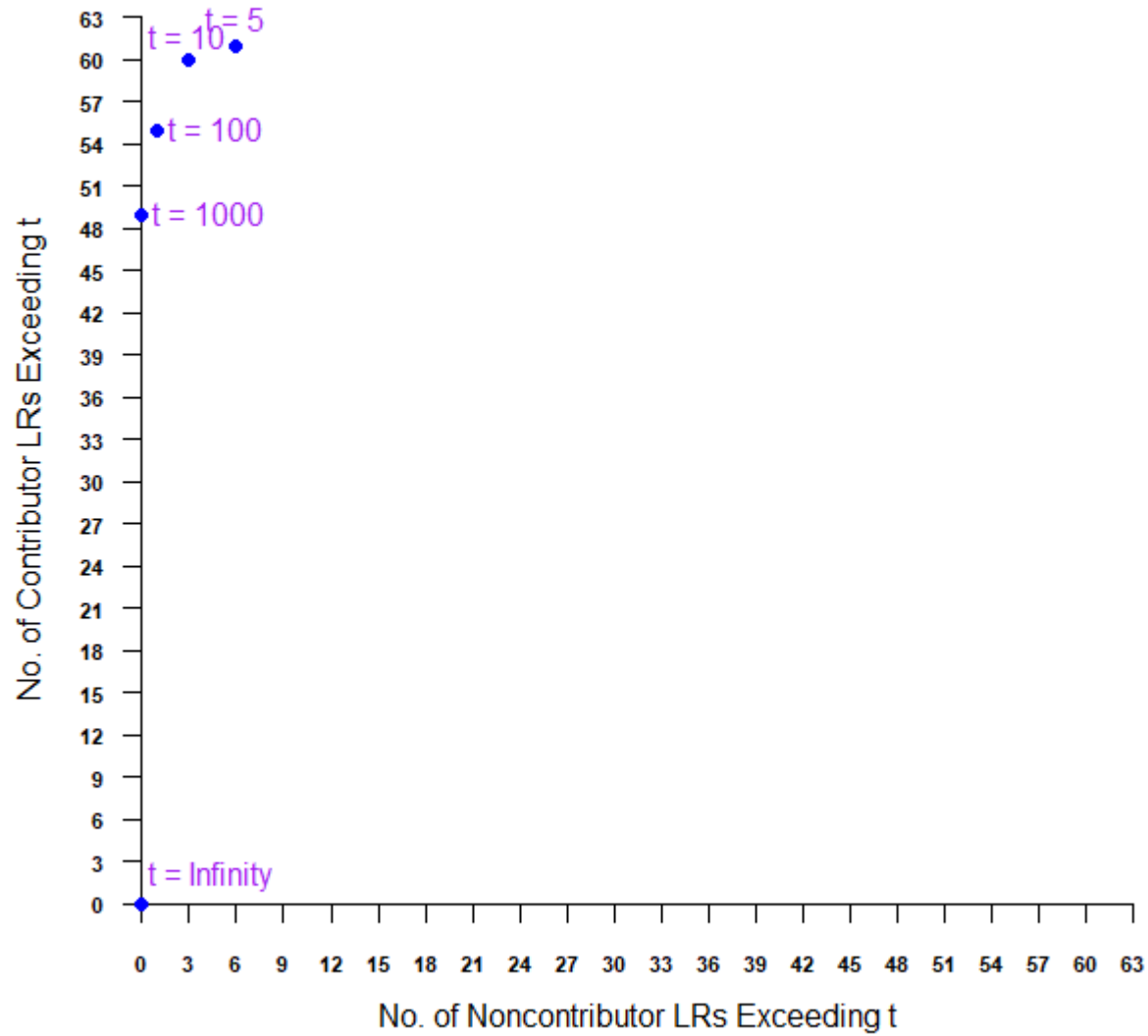


### No. of LRs Exceeding LR = t



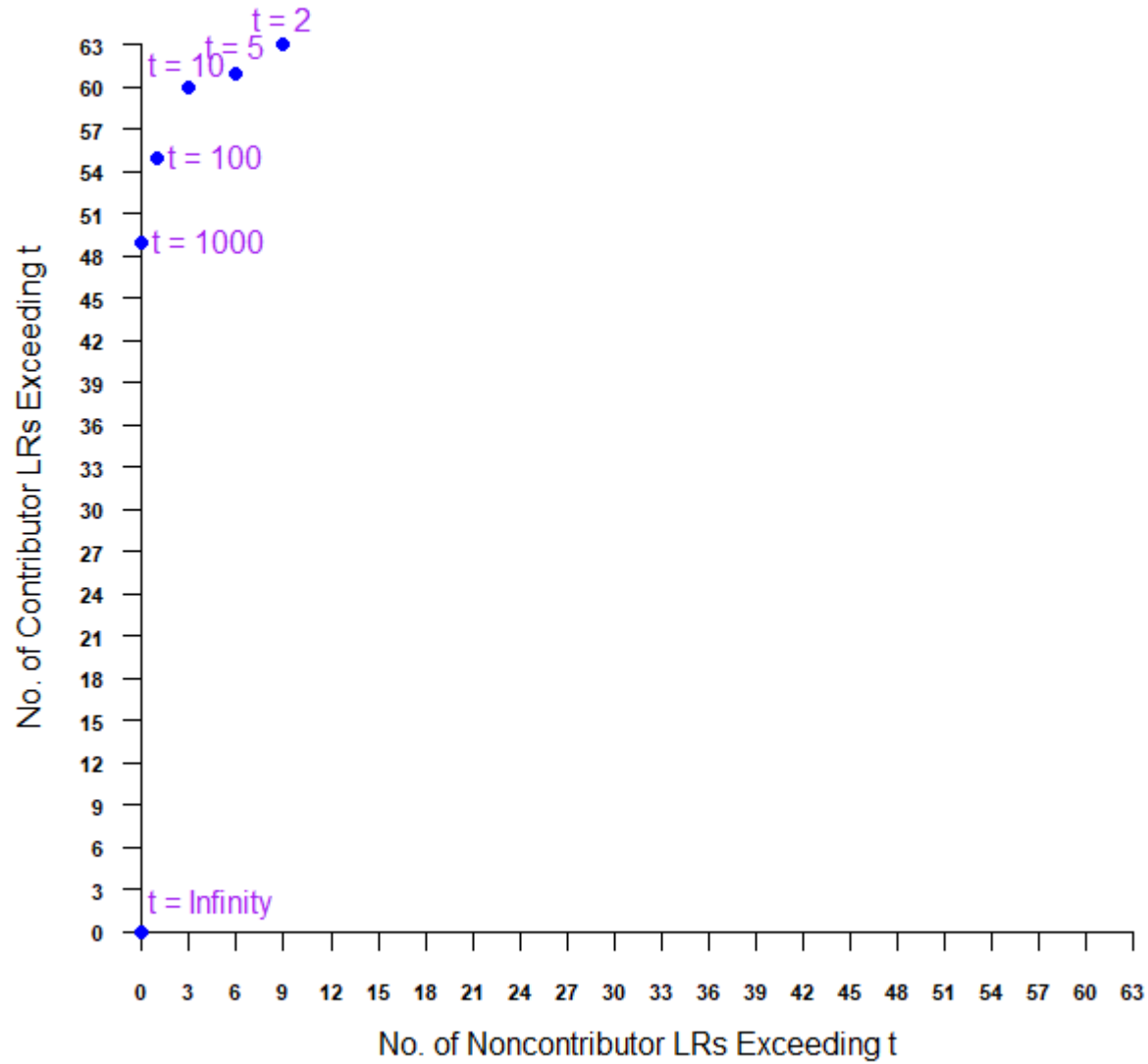
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LR Exceeding LR = t



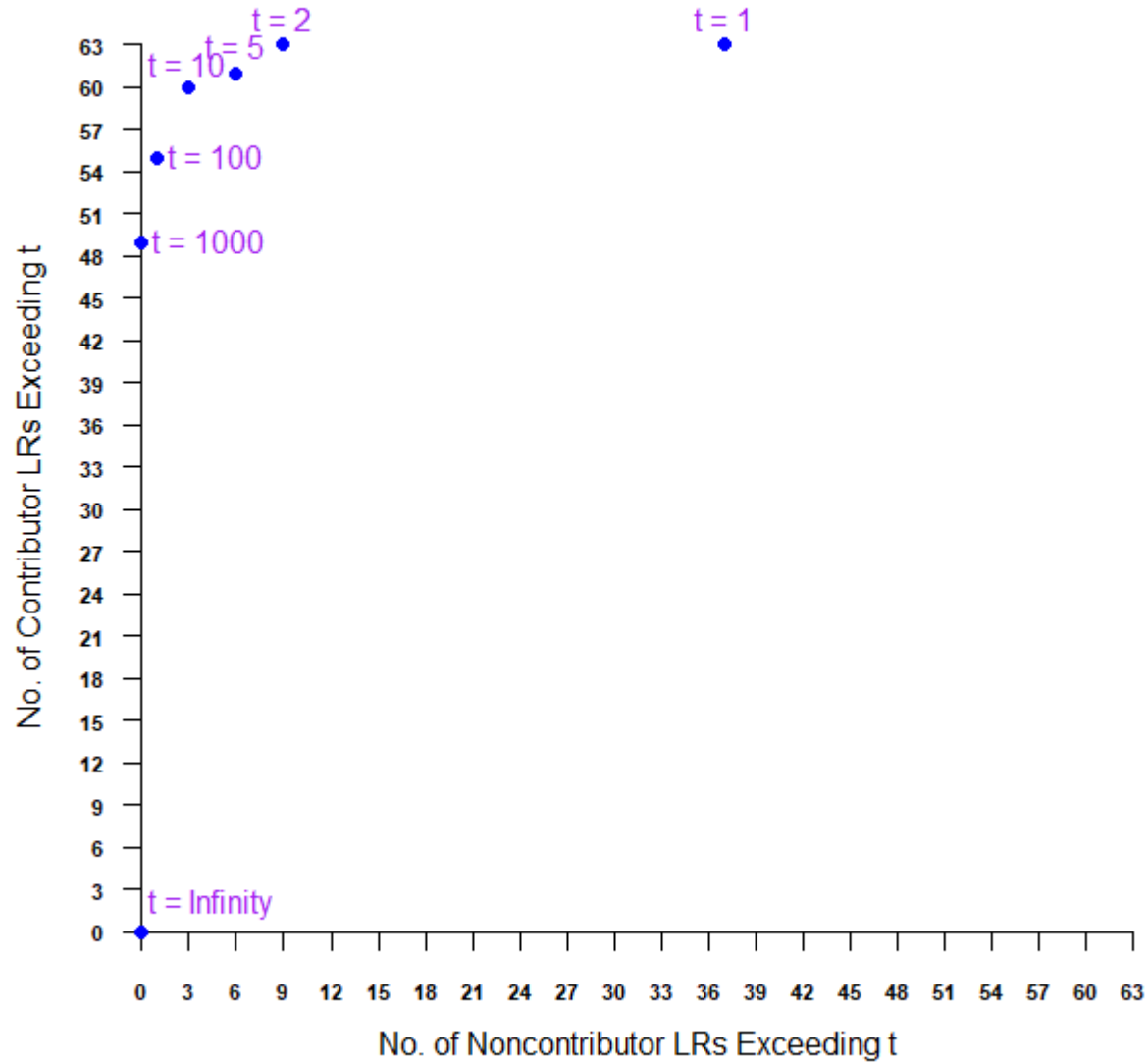
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LR Exceeding LR = t



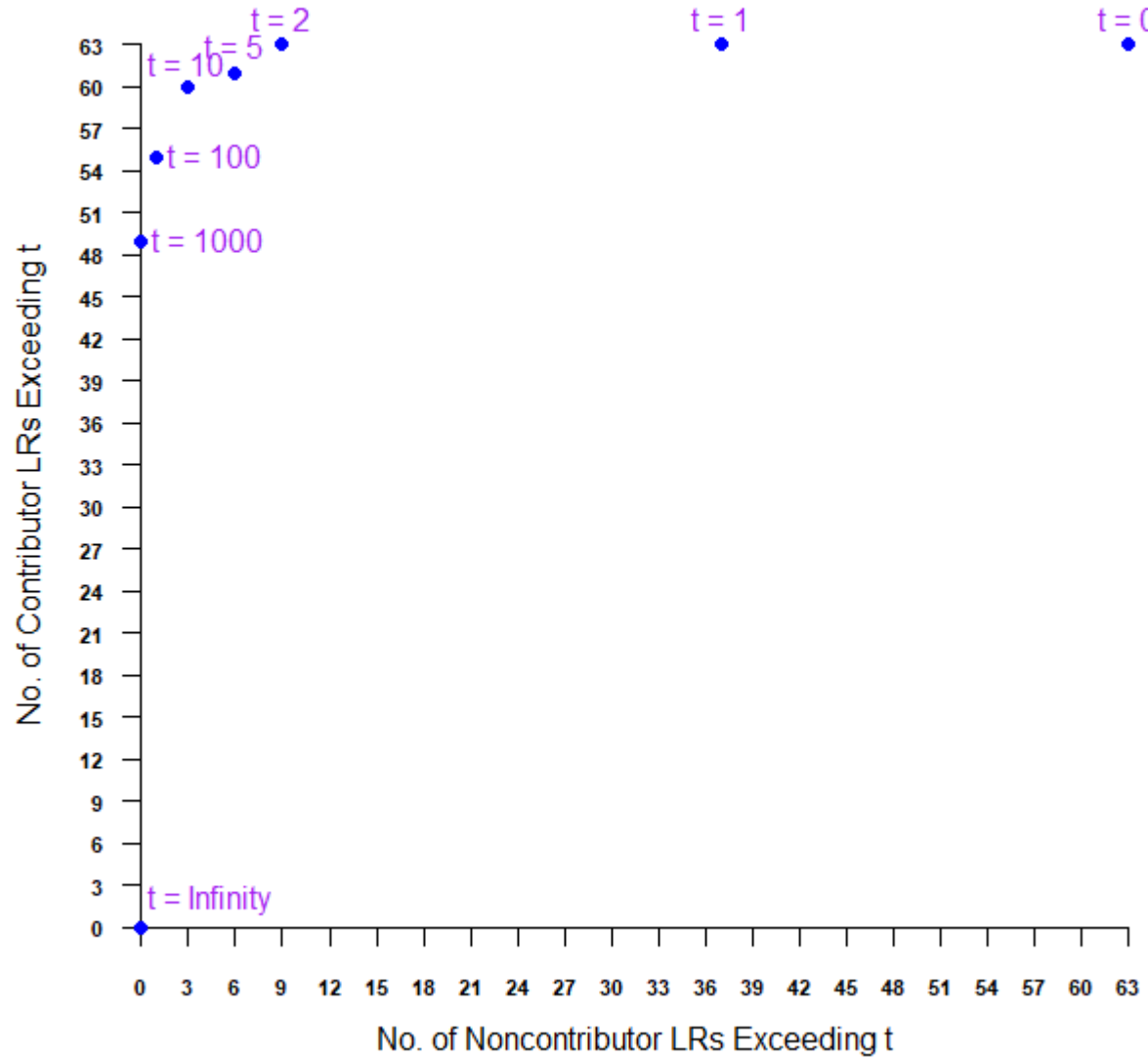
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LR Exceeding LR = t

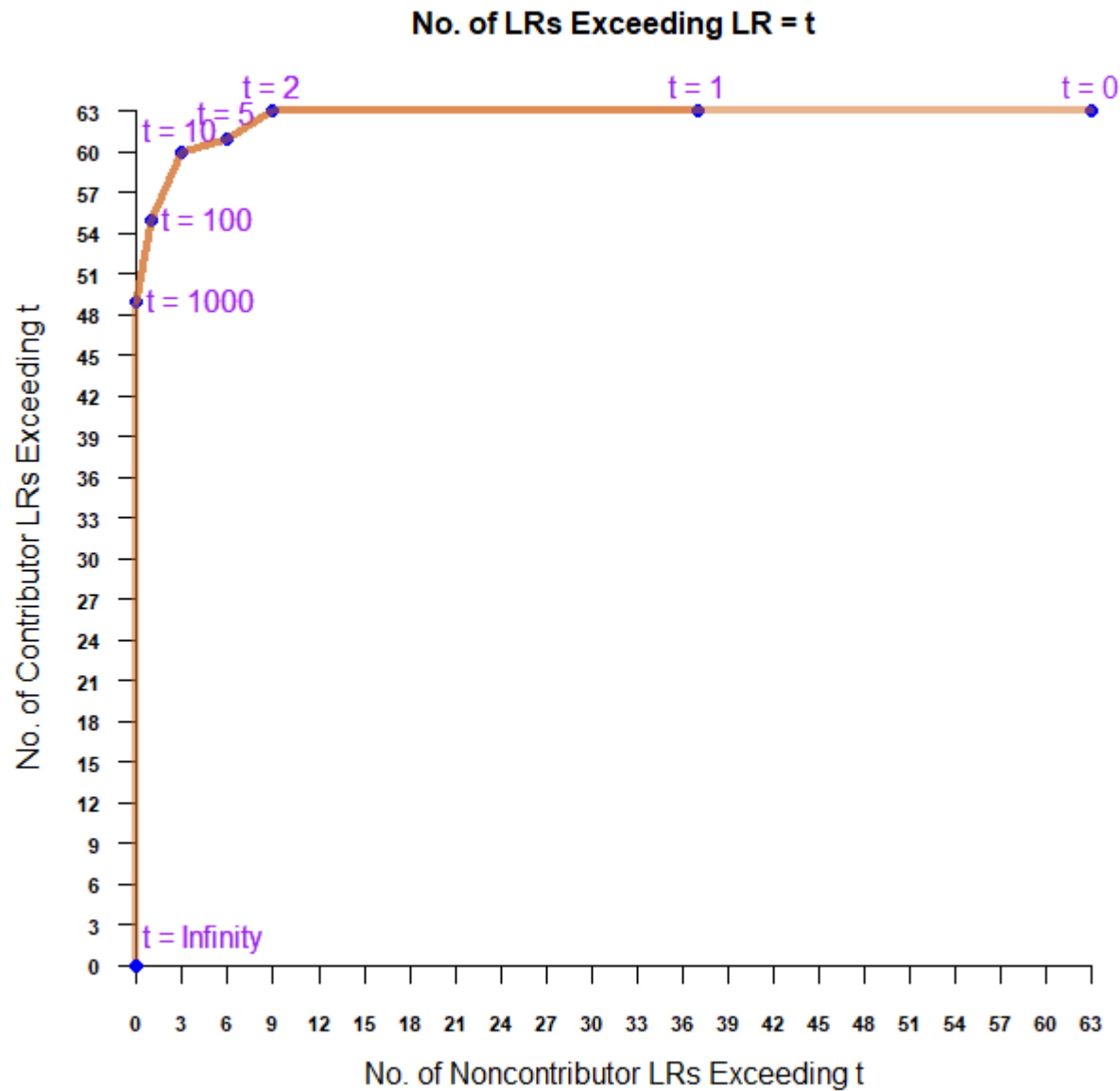


t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

### No. of LR Exceeding LR = t

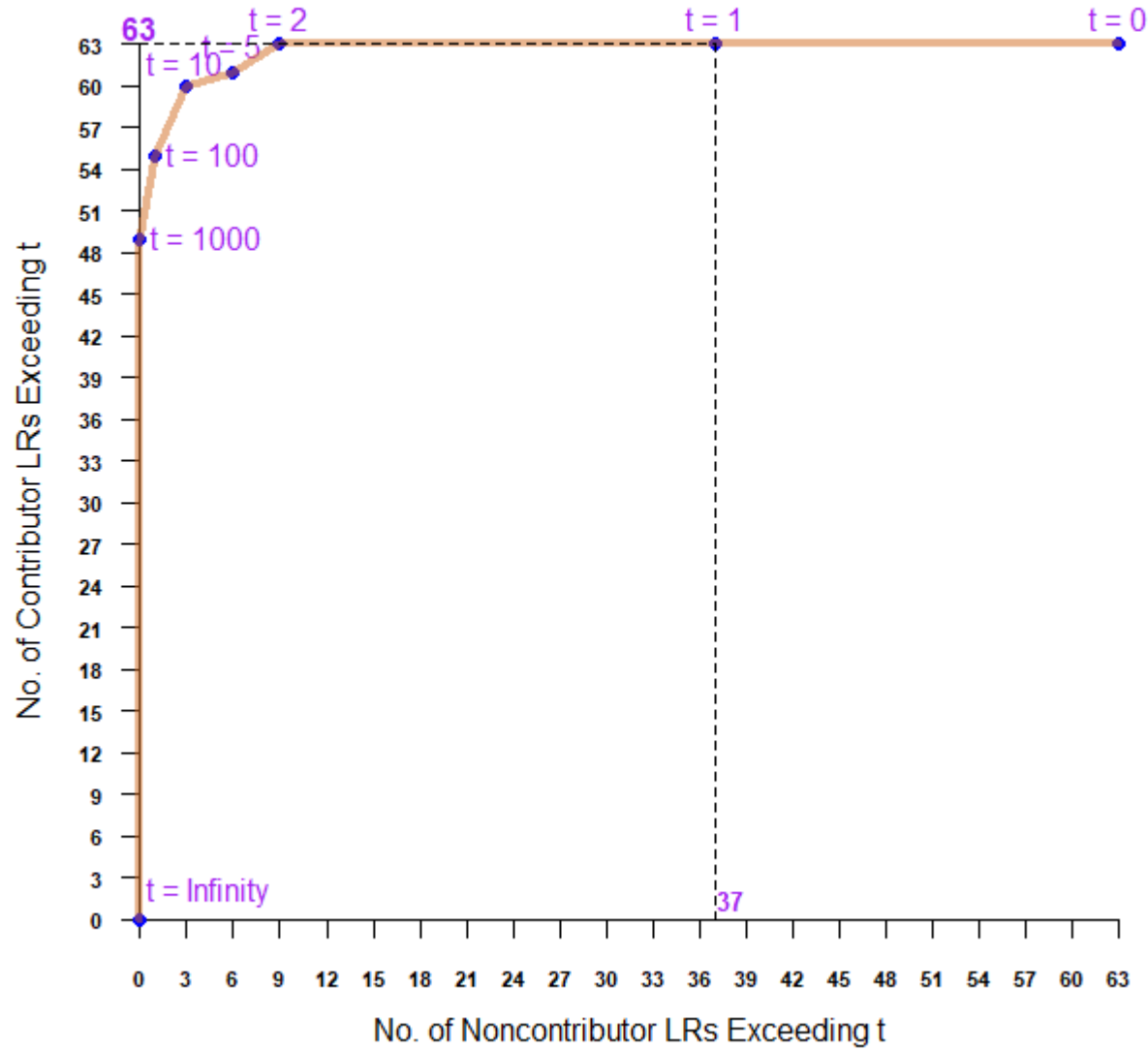


t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	63	63	0
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63



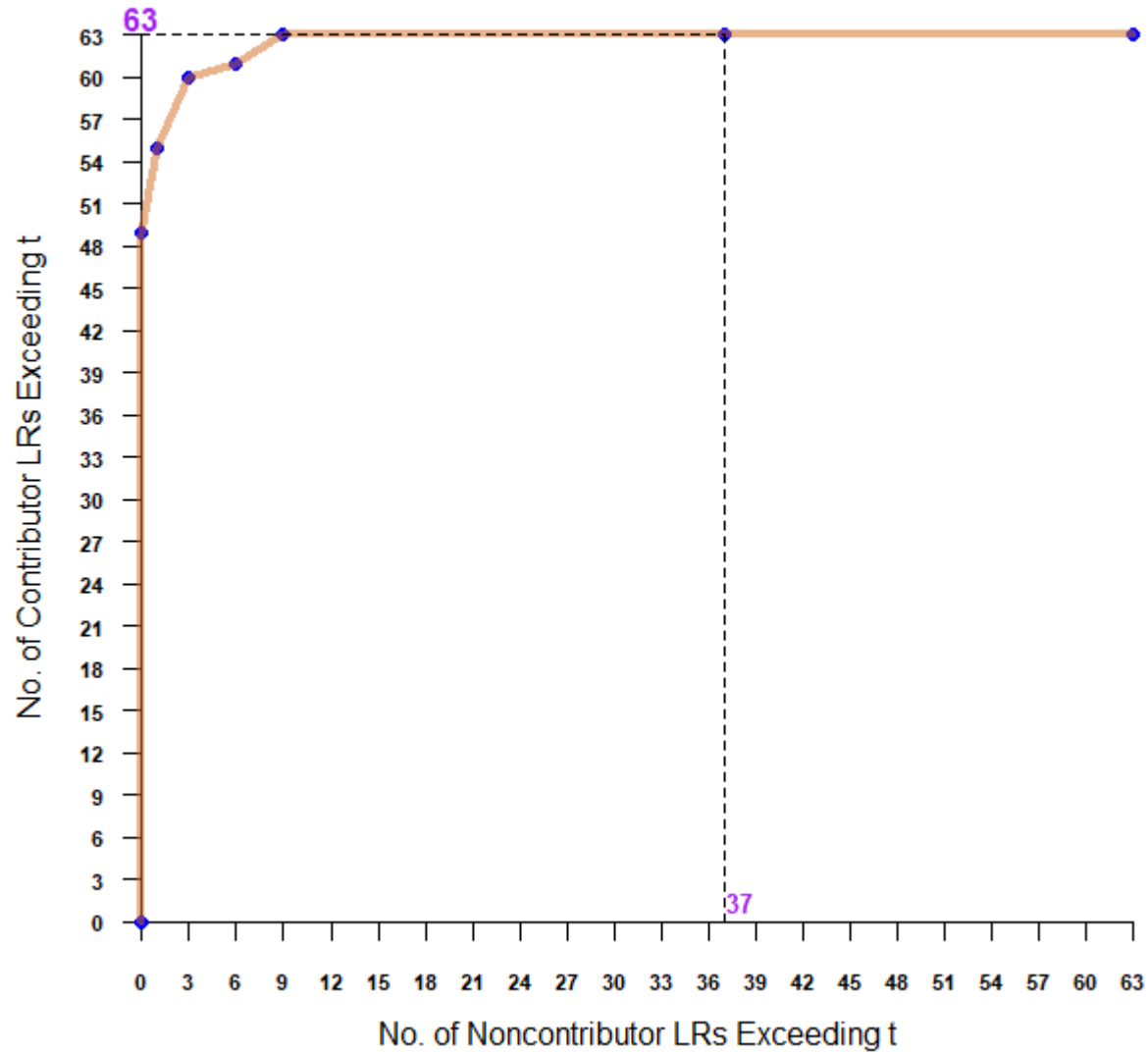
t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	63	63	0
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

No. of LR Exceeding LR = t



t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	63	63	0
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

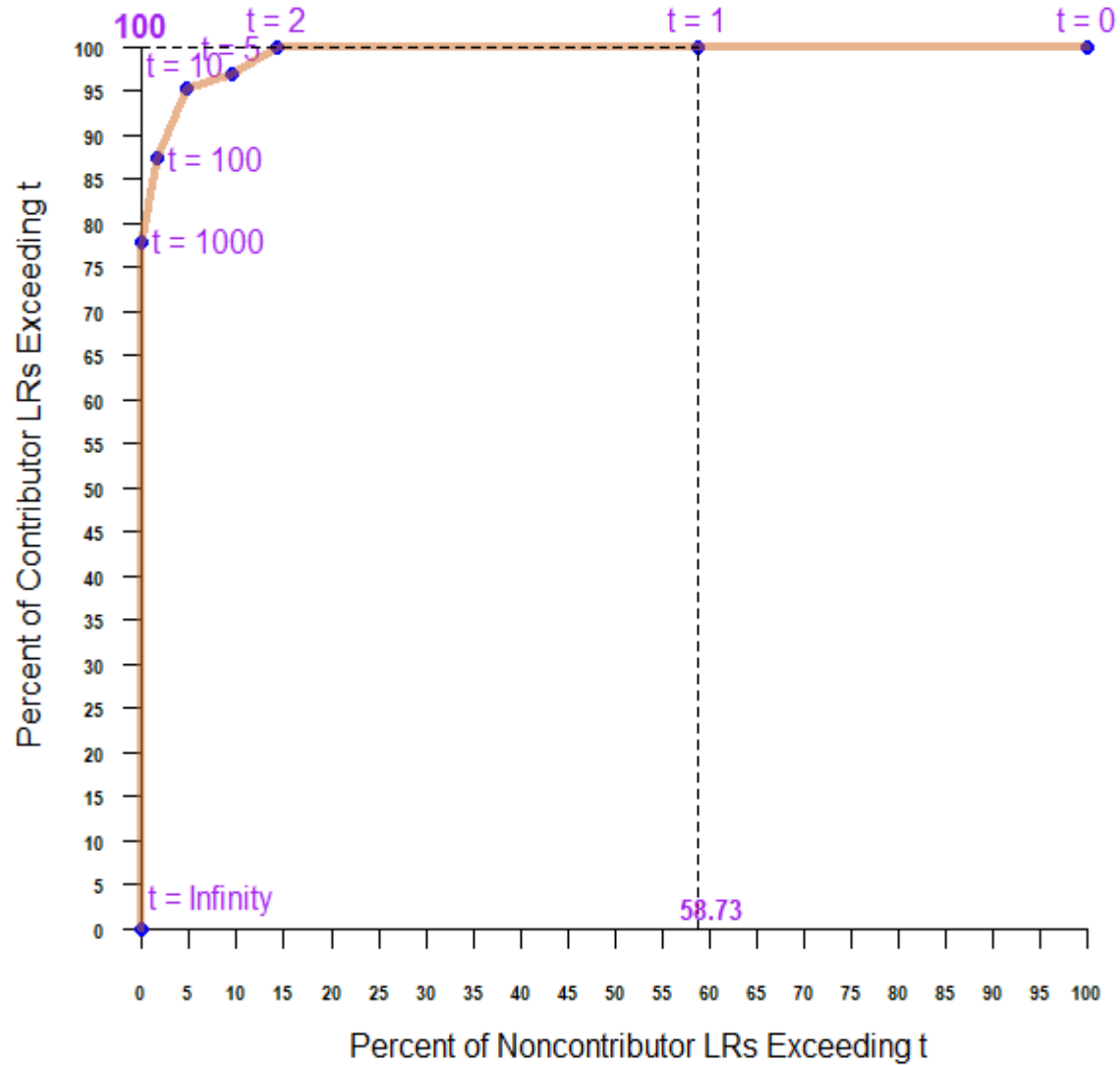
No. of LR's Exceeding LR = t



t equal to	LR > t		LR < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	63	63	0
1	37	63	0
2	9	63	0
5	6	61	2
10	3	60	3
100	1	55	8
1000	0	49	14
Infinity	0	0	63

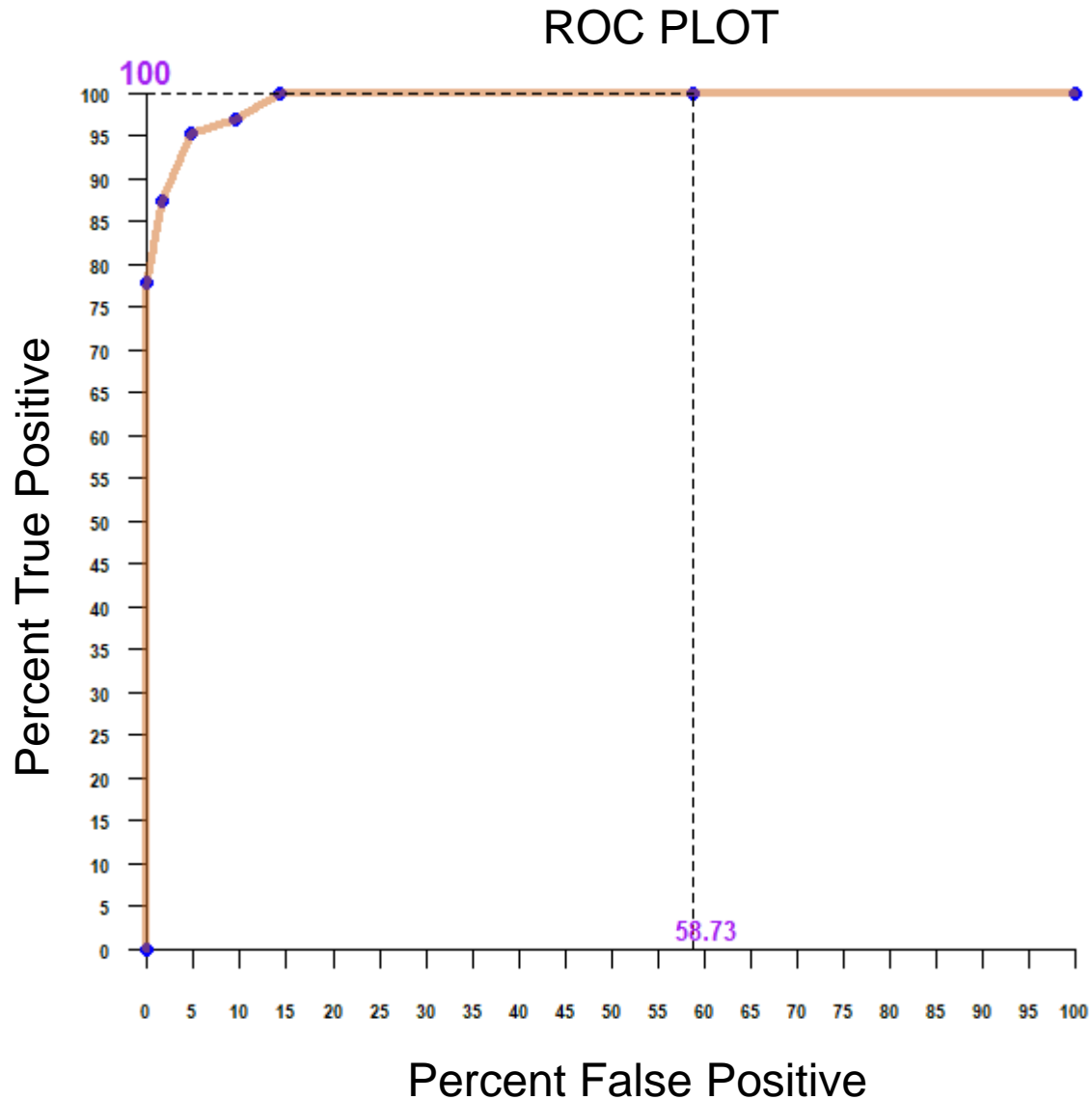


Percent of LRs Exceeding LR = t



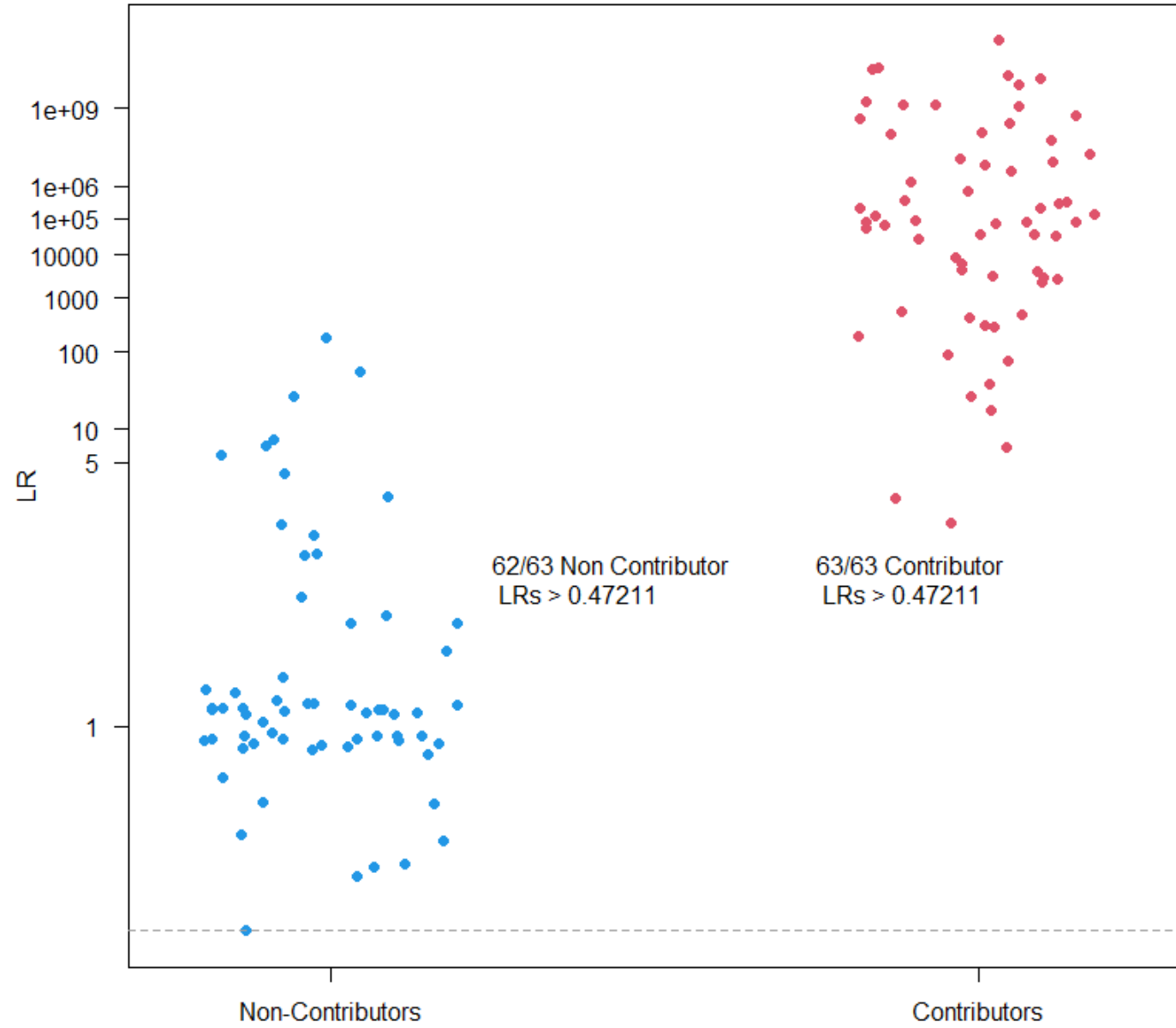
t equal to	Percent LRs > t		Percent LRs < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	100.0	100.0	0.0
1	58.7	100.0	0.0
2	14.3	100.0	0.0
5	9.5	96.8	3.2
10	4.8	95.2	4.8
100	1.6	87.3	12.7
1000	0.0	77.8	22.2
Infinity	0.0	0.0	100.0

# Receiver Operating Characteristic (ROC) Plot



t equal to	Percent LRs > t		Percent LRs < t
	Non Contributors (out of 63)	Contributors (out of 63)	Contributors (out of 63)
0	100.0	100.0	0.0
1	58.7	100.0	0.0
2	14.3	100.0	0.0
5	9.5	96.8	3.2
10	4.8	95.2	4.8
100	1.6	87.3	12.7
1000	0.0	77.8	22.2
Infinity	0.0	0.0	100.0

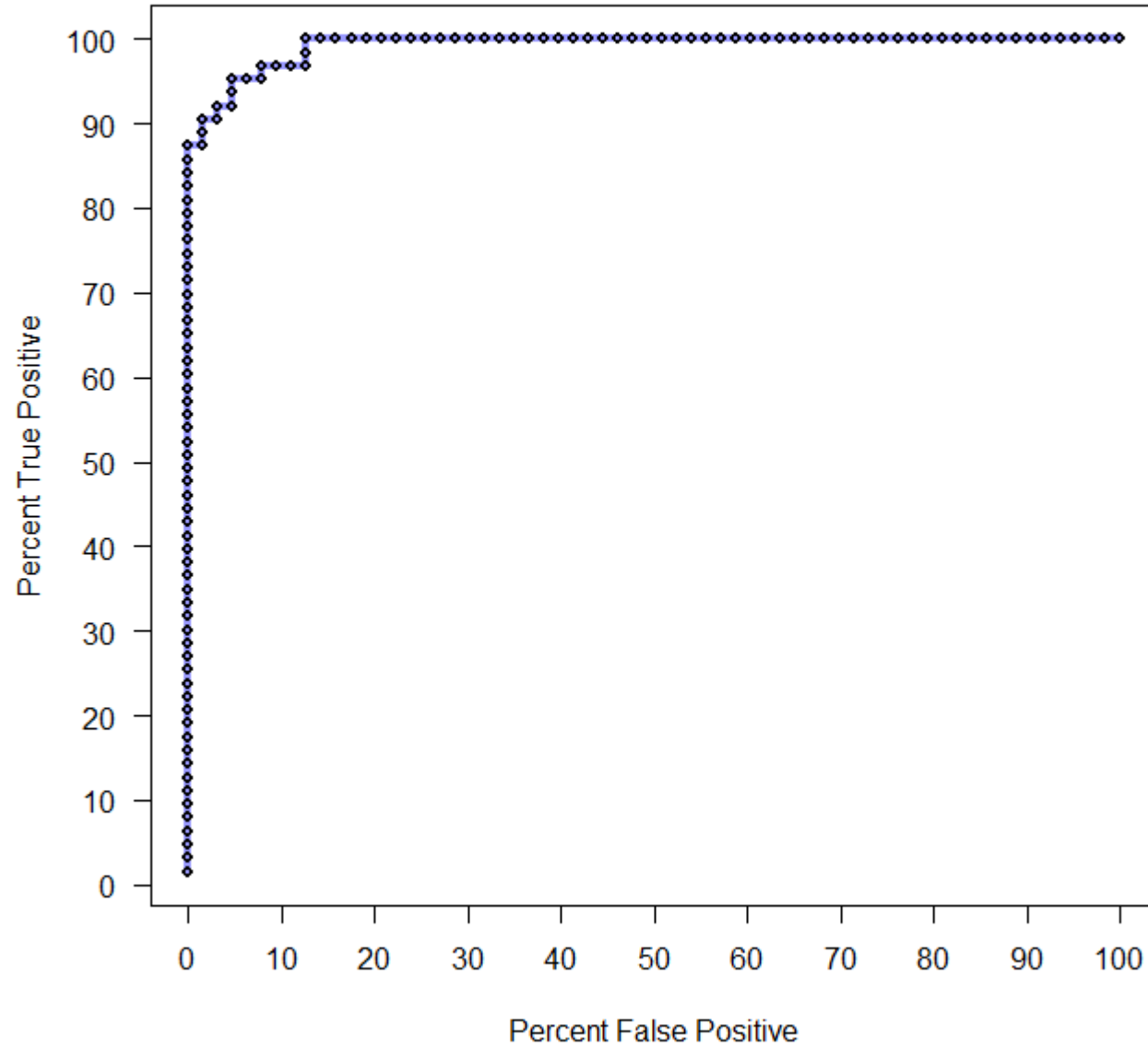
ProvedIT Data, 4P Mixtures  
Degraded, Total Amt < 125 pg



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t		Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t		Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t		Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t
12.2951	0	0		4.90887	0	31		0.897801756	3	60		9.35E-05	32	63
10.9148	0	1		4.90689	0	32		0.835598625	4	60		9.27E-05	33	63
10.8273	0	2		4.89104	0	33		0.828823886	5	60		6.10E-05	34	63
10.4955	0	3		4.8487	0	34		0.75574734	5	61		5.79E-05	35	63
10.3897	0	4		4.79175	0	35		0.612581796	6	61		2.30E-06	36	63
10.086	0	5		4.69132	0	36		0.463910598	7	61		-1.16E-05	37	63
9.321	0	6		4.53982	0	37		0.455335944	8	61		-2.59E-05	38	63
9.1639	0	7		4.51793	0	38		0.319121484	8	62		-3.24E-05	39	63
9.1602	0	8		4.4857	0	39		0.311130659	8	63		-3.29E-05	40	63
9.0723	0	9		4.41681	0	40		0.268140397	9	63		-6.13E-05	42	63
8.7013	0	10		3.91569	0	41		0.197190215	10	63		-8.30E-05	43	63
8.5595	0	11		3.77646	0	42		0.19091703	11	63		-8.49E-05	44	63
8.3297	0	12		3.63206	0	43		0.081249353	12	63		-9.37E-05	45	63
8.0076	0	13		3.57291	0	44		0.052903712	13	63		-0.000118524	46	63
7.942	0	14		3.47726	0	45		0.042156949	14	63		-0.000196661	47	63
7.6696	0	15		3.42774	0	46		0.041763772	15	63		-0.00021304	48	63
7.1663	0	16		3.42493	0	47		0.016129872	16	63		-0.000232778	49	63
6.9638	0	17		3.34671	0	48		0.004586252	17	63		-0.00031962	50	63
6.8575	0	18		2.73565	0	49		0.001954468	18	63		-0.000381701	51	63
6.7685	0	19		2.66982	0	50		0.001498548	19	63		-0.000456537	52	63
6.5161	0	20		2.59525	0	51		0.000654617	20	63		-0.00081453	53	63
6.1529	0	21		2.45627	0	52		0.000478231	21	63		-0.005131138	54	63
5.8388	0	22		2.4328	0	53		0.00047727	22	63		-0.016789575	55	63
5.5667	0	23		2.26008	0	54		0.000349308	23	63		-0.017560971	56	63
5.5074	0	24		2.22324	0	55		0.000345022	24	63		-0.047185307	57	63
5.4608	0	25		1.96222	1	55		0.000260663	25	63		-0.05773551	58	63
5.3111	0	26		1.85723	1	56		0.000224222	26	63		-0.09835191	59	63
5.3023	0	27		1.70915	1	57		0.000213759	27	63		-0.105078069	60	63
5.101	0	28		1.5207	2	57		0.000196883	28	63		-0.129165378	61	63
5.0815	0	29		1.36772	2	58		0.000173029	29	63		-0.325956532	62	63
4.9299	0	30		1.36249	3	58		0.000173028	30	63		-1.325956532	63	63
				1.19941	3	59		0.000152703	31	63				

# PROVEDIt Data, 4P Mixtures

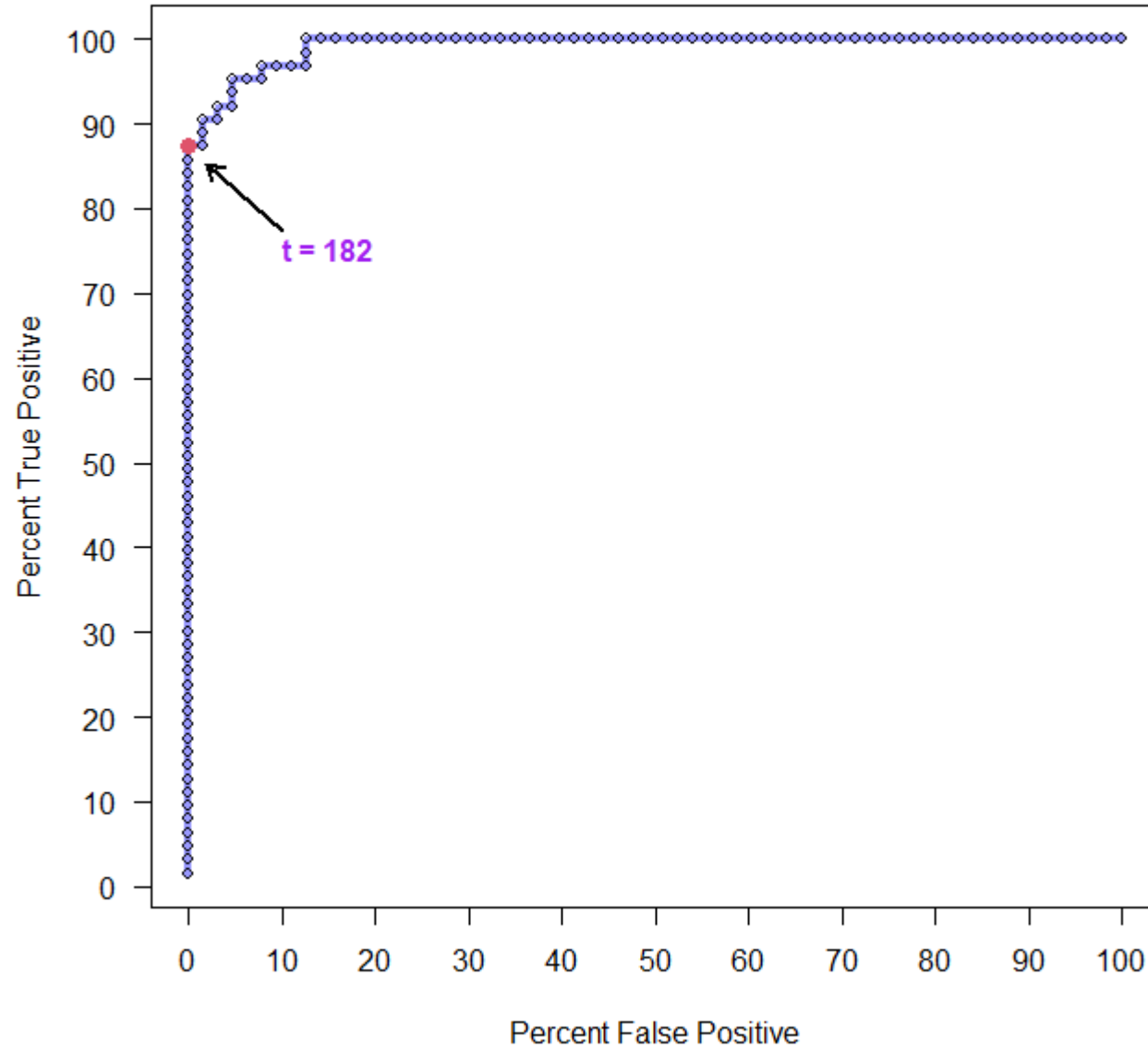
## Degraded, Total Amount < 125 pg, Ratio 1:1:1:1



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t
12.2951	0	0	4.90887	0	33	0.897801756	3	60	9.35E-05	32	63			
10.9148	0	1	4.90689	0	32	0.835598625	4	60	9.27E-05	33	63			
10.8273	0	2	4.89104	0	33	0.828823886	5	60	6.10E-05	34	63			
10.4955	0	3	4.8487	0	34	0.75574734	5	61	5.79E-05	35	63			
10.3897	0	4	4.79175	0	35	0.612581796	6	61	2.30E-06	36	63			
10.086	0	5	4.69132	0	36	0.463910598	7	61	-1.16E-05	37	63			
9.321	0	6	4.53982	0	37	0.455335944	8	61	-2.59E-05	38	63			
9.1639	0	7	4.51793	0	38	0.319121484	8	62	-3.24E-05	39	63			
9.1602	0	8	4.4857	0	39	0.311330659	8	63	-3.29E-05	40	63			
9.0723	0	9	4.41681	0	40	0.268140397	9	63	-6.13E-05	42	63			
8.7013	0	10	3.91569	0	41	0.197190215	10	63	-8.30E-05	43	63			
8.5595	0	11	3.77646	0	42	0.19091703	11	63	-8.49E-05	44	63			
8.3297	0	12	3.63206	0	43	0.081249353	12	63	-9.37E-05	45	63			
8.0076	0	13	3.57291	0	44	0.052903712	13	63	-0.000118524	46	63			
7.942	0	14	3.47726	0	45	0.042156949	14	63	-0.000196661	47	63			
7.6696	0	15	3.42774	0	46	0.041763772	15	63	-0.00021304	48	63			
7.1663	0	16	3.40493	0	47	0.016129872	16	63	-0.00023278	49	63			
6.9638	0	17	3.34671	0	48	0.004586252	17	63	-0.00031962	50	63			
6.8575	0	18	2.73565	0	49	0.001954468	18	63	-0.000381701	51	63			
6.7685	0	19	2.66982	0	50	0.001498548	19	63	-0.000456537	52	63			
6.5161	0	20	2.59525	0	51	0.000654617	20	63	-0.00081453	53	63			
6.1529	0	21	2.45627	0	52	0.000478231	21	63	-0.005131138	54	63			
5.8388	0	22	2.4328	0	53	0.00047727	22	63	-0.016789575	55	63			
5.5667	0	23	2.28008	0	54	0.000349308	23	63	-0.017569971	56	63			
5.5074	0	24	2.22324	0	55	0.000345022	24	63	-0.047185307	57	63			
5.4608	0	25	1.96222	1	55	0.000260663	25	63	-0.05773551	58	63			
5.3111	0	26	1.85723	1	56	0.000224222	26	63	-0.09835191	59	63			
5.3023	0	27	1.70915	2	57	0.000213759	27	63	-0.105078069	60	63			
5.101	0	28	1.5207	2	57	0.000196883	28	63	-0.129165378	61	63			
5.0815	0	29	1.36772	2	58	0.000173029	29	63	-0.325956532	62	63			
4.9299	0	30	1.36249	3	58	0.000173028	30	63	-1.325956532	63	63			
			1.19941	3	59	0.000152703	31	63						

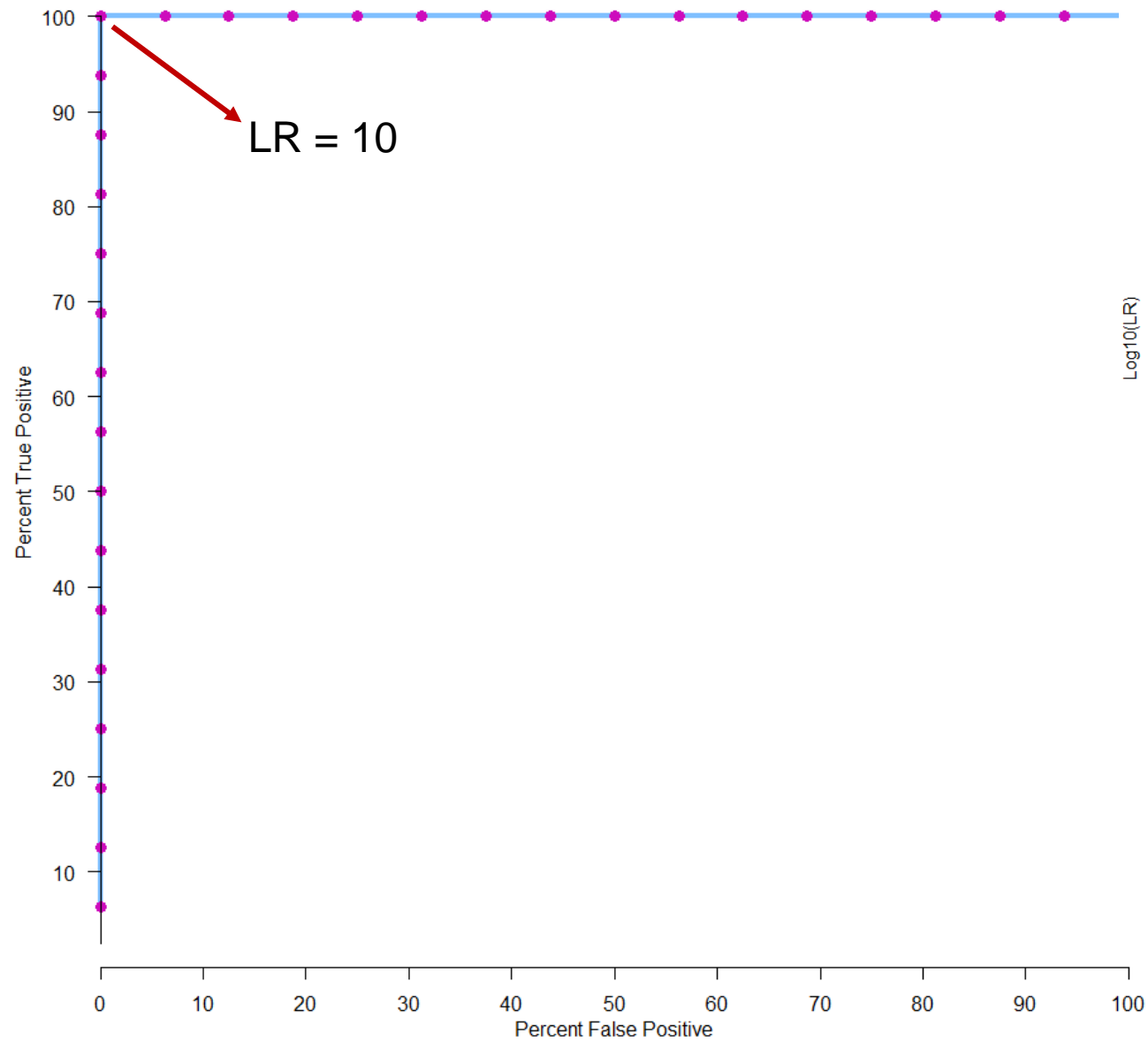
# PROVEDIt Data, 4P Mixtures

## Degraded, Total Amount < 125 pg, Ratio 1:1:1:1

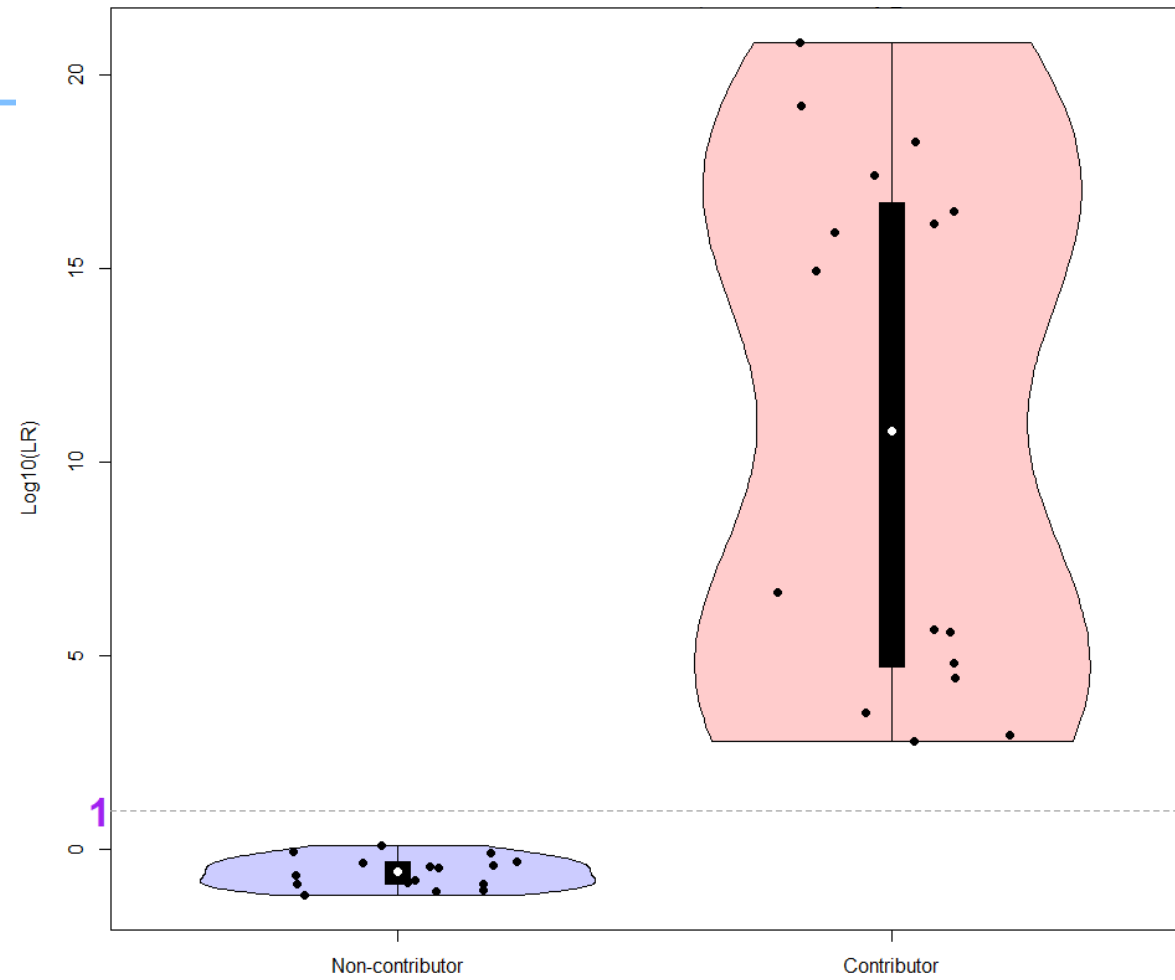


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t	Log10(LR) Threshold = t	Non Contributor Log10(LRs) > t	Contributor Log10(LRs) > t
12.2951	0	0	4.90887	0	33	0.897801756	3	60	3.35E-05	32	63			
10.9148	0	1	4.90689	0	32	0.835598625	4	60	9.27E-05	33	63			
10.8273	0	2	4.89104	0	33	0.828823886	5	60	6.10E-05	34	63			
10.4955	0	3	4.8487	0	34	0.75574734	5	61	5.79E-05	35	63			
10.3897	0	4	4.79175	0	35	0.612581796	6	61	2.30E-06	36	63			
10.086	0	5	4.69132	0	36	0.463910598	7	61	-1.16E-05	37	63			
9.321	0	6	4.53982	0	37	0.455335944	8	61	-2.59E-05	38	63			
9.1639	0	7	4.51793	0	38	0.319121484	8	62	-3.24E-05	39	63			
9.1602	0	8	4.4857	0	39	0.311330659	8	63	-3.29E-05	40	63			
9.0723	0	9	4.41681	0	40	0.268140397	9	63	-6.13E-05	42	63			
8.7013	0	10	3.91569	0	41	0.197190215	10	63	-8.30E-05	43	63			
8.5595	0	11	3.77646	0	42	0.19091703	11	63	-8.49E-05	44	63			
8.3297	0	12	3.63206	0	43	0.081249353	12	63	-9.37E-05	45	63			
8.0076	0	13	3.57291	0	44	0.052903712	13	63	-0.000118524	46	63			
7.942	0	14	3.47726	0	45	0.042156949	14	63	-0.000196661	47	63			
7.6696	0	15	3.42774	0	46	0.041763772	15	63	-0.00021304	48	63			
7.1663	0	16	3.40493	0	47	0.016125872	16	63	-0.00023278	49	63			
6.9638	0	17	3.34671	0	48	0.004586252	17	63	-0.00031962	50	63			
6.8575	0	18	2.73565	0	49	0.001954468	18	63	-0.000381701	51	63			
6.7685	0	19	2.66982	0	50	0.001498548	19	63	-0.000456537	52	63			
6.5161	0	20	2.59525	0	51	0.000654617	20	63	-0.00081453	53	63			
6.1529	0	21	2.45627	0	52	0.000478231	21	63	-0.005131138	54	63			
5.8388	0	22	2.4328	0	53	0.00047727	22	63	-0.016789575	55	63			
5.5667	0	23	2.28008	0	54	0.000349308	23	63	-0.017569971	56	63			
5.5074	0	24	2.22324	0	55	0.000345022	24	63	-0.047185307	57	63			
5.4608	0	25	1.96222	1	55	0.000260663	25	63	-0.05773551	58	63			
5.3111	0	26	1.85723	1	56	0.000224222	26	63	-0.09835191	59	63			
5.3023	0	27	1.70915	2	57	0.000213759	27	63	-0.105078069	60	63			
5.101	0	28	1.5207	2	57	0.000196883	28	63	-0.129165378	61	63			
5.0815	0	29	1.36772	2	58	0.000173029	29	63	-0.325956532	62	63			
4.9299	0	30	1.36249	3	58	0.000173028	30	63	-1.325956532	63	63			
			1.19941	3	59	0.000152703	31	63						

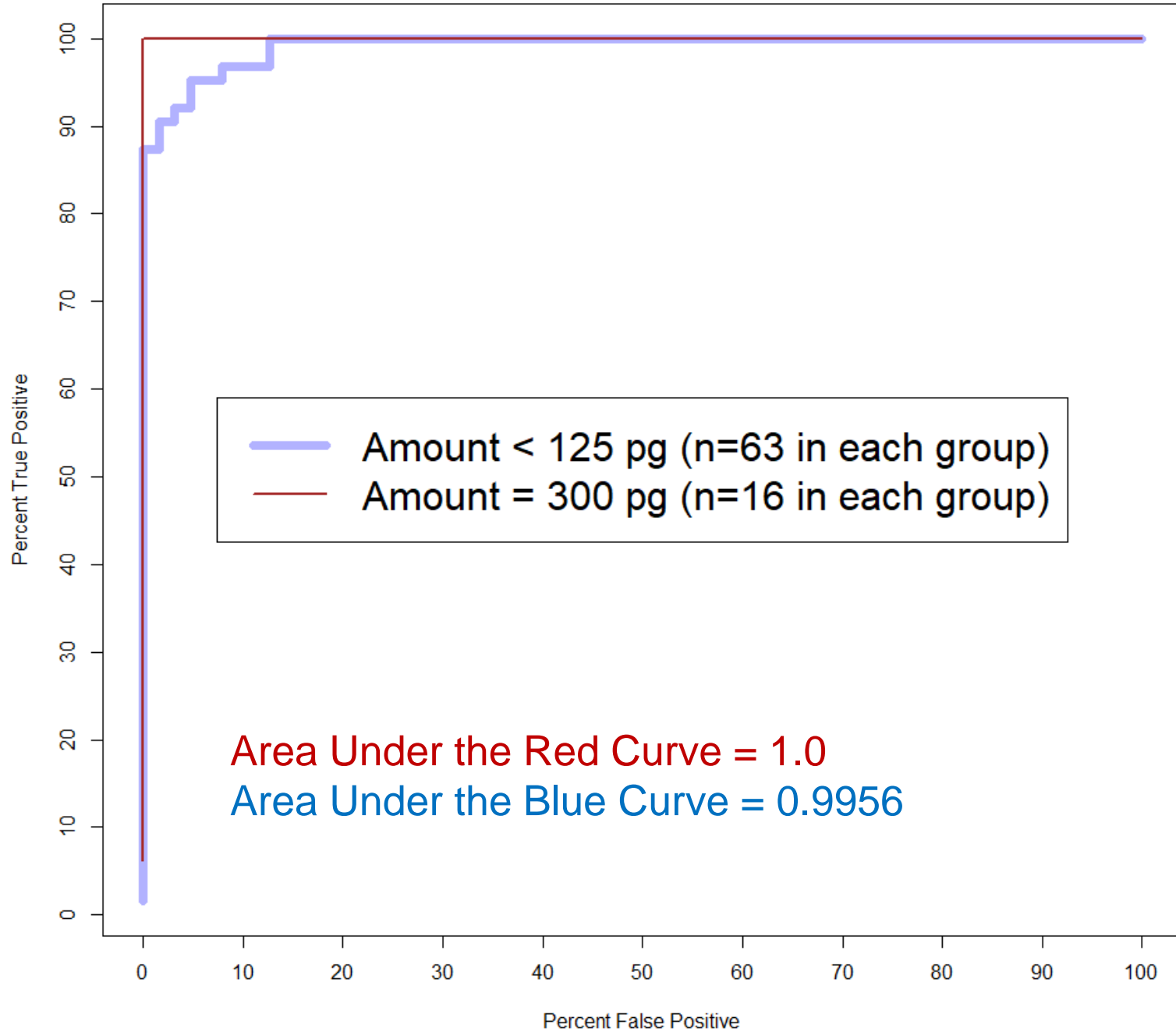
# PROVEDIt Data, 4P Mixtures Total DNA Amount = 300 pg



## Distribution of Log<sub>10</sub>(LR) PROVEDIt Data, 4P Mixtures, Amount = 300 pg



## Comparison of ROC Curves



LR System appears to **better discriminate** between Hp and Hd for **samples with 300 pg total DNA** **than for** **samples with less than 125 pg total DNA.**





ELSEVIER

Research paper

A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles

Øyvind Bleka<sup>a,b,\*</sup>, Corina C.G. Benschop<sup>c</sup>, Geir Storvik<sup>b</sup>, Peter Gill<sup>a,d</sup>

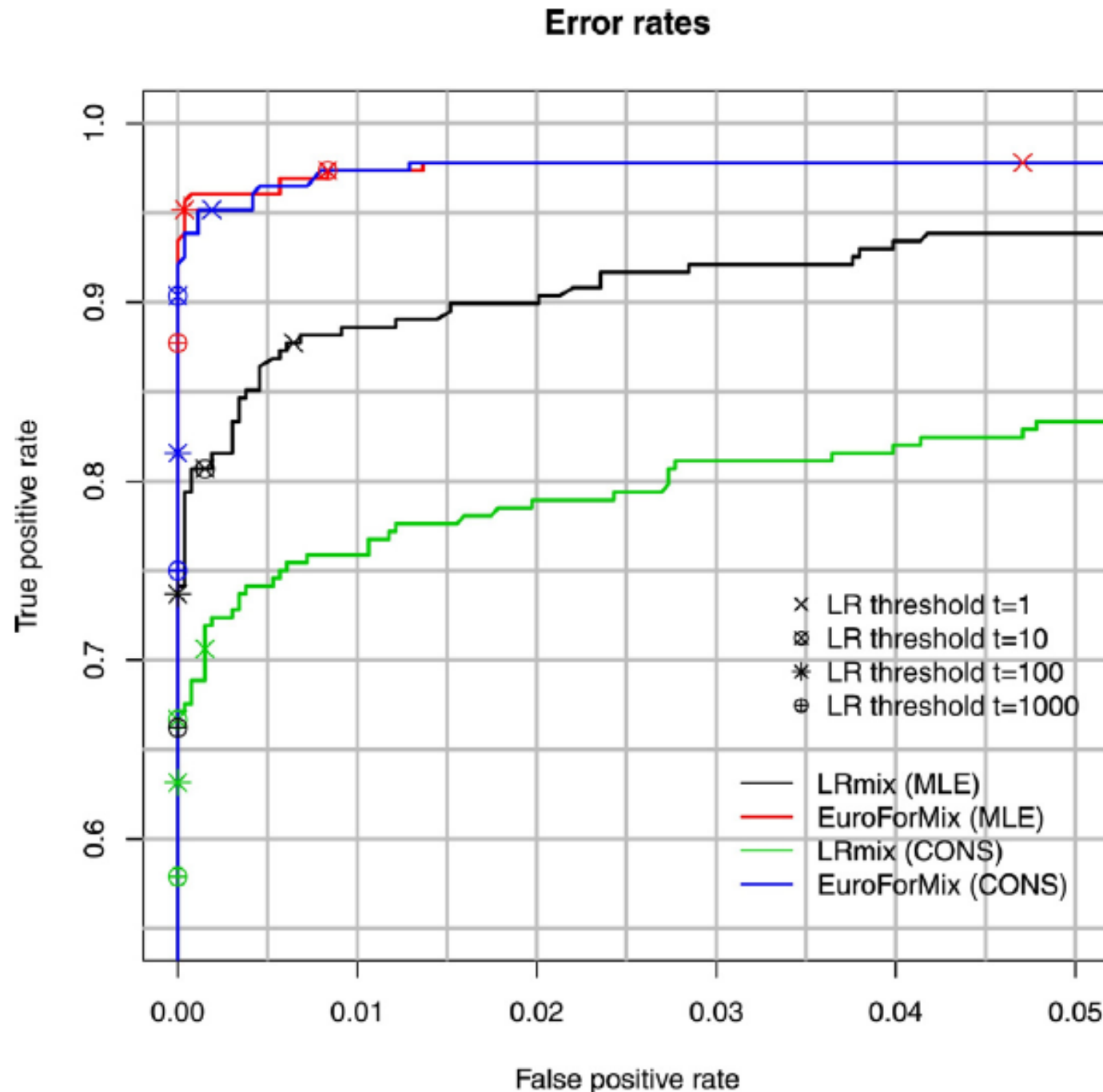
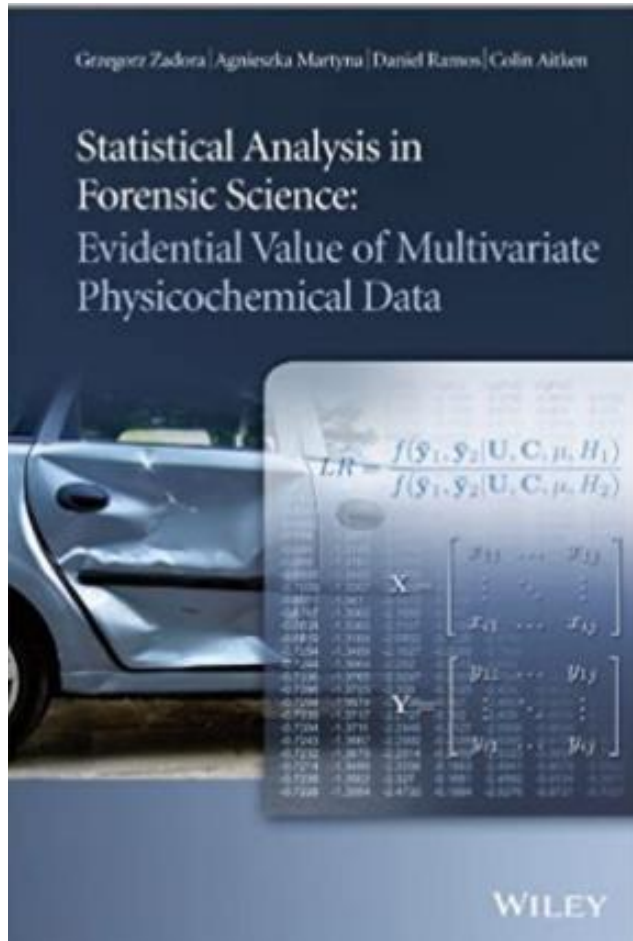


Fig. 2. Receiver operating characteristic (ROC) plot where the rate of false positives (FP) (along horizontal axis) and true positives (TP) (along vertical axis) are plotted as a function of LR thresholds. The plot shows the results for the maximum likelihood estimation method (MLE) and the conservative method (CONS) for both LRmix and EuroForMix. The points on the curves show the FP and TP rates for different LR thresholds.

# Calibration Accuracy



## 6 Performance of likelihood ratio methods

6.2 Empirical measurement of the performance of likelihood ratios

6.5 Accuracy equals discriminating power plus calibration:  
Empirical cross-entropy plots

- the discriminating power is poor. This means that the validation set of  $LR$  values is poor at separating  $LR$  values for which  $H_1$  is true from  $LR$  values for which  $H_2$  is true.
- the calibration is poor. This means that the  $LR$  values provide poor probabilistic measures of the value of the evidence. Even if the  $LR$  values have high discriminating power, poor calibration can degrade the accuracy considerably.

An R-package called **comparison** can be used to apply their method

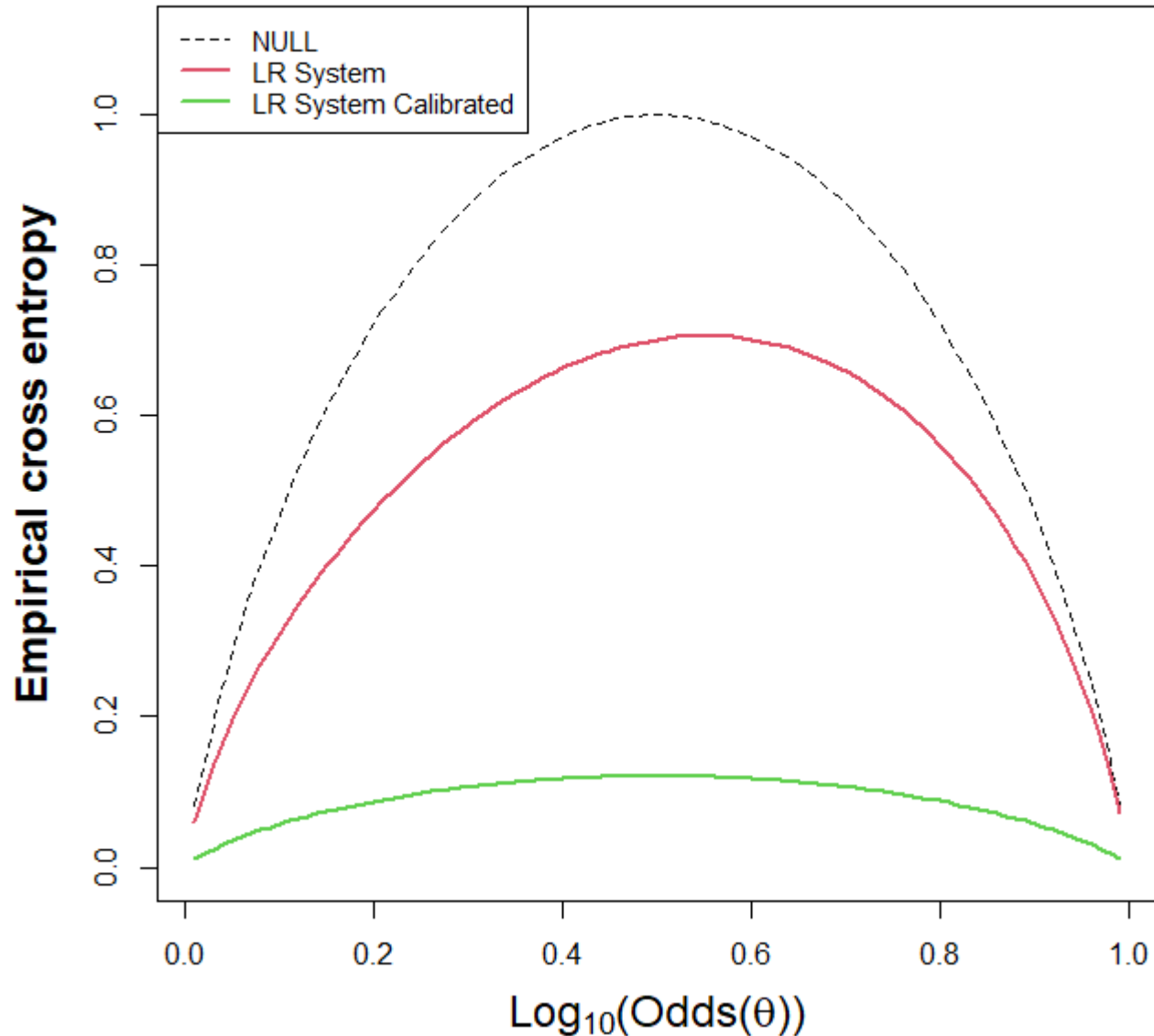
David Lucy, James Curran, Agnieszka Martyna  
1964–2018

Grzegorz Zadora, Agnieszka Martyna,  
Daniel Ramos, Colin Aitken

# Calibration Accuracy

ECE Plot for Example Data

PROVEDIt, 1:1:1:1, degraded, Amount < 125 pg (n = 63 in each group)



Inputs to the R function:

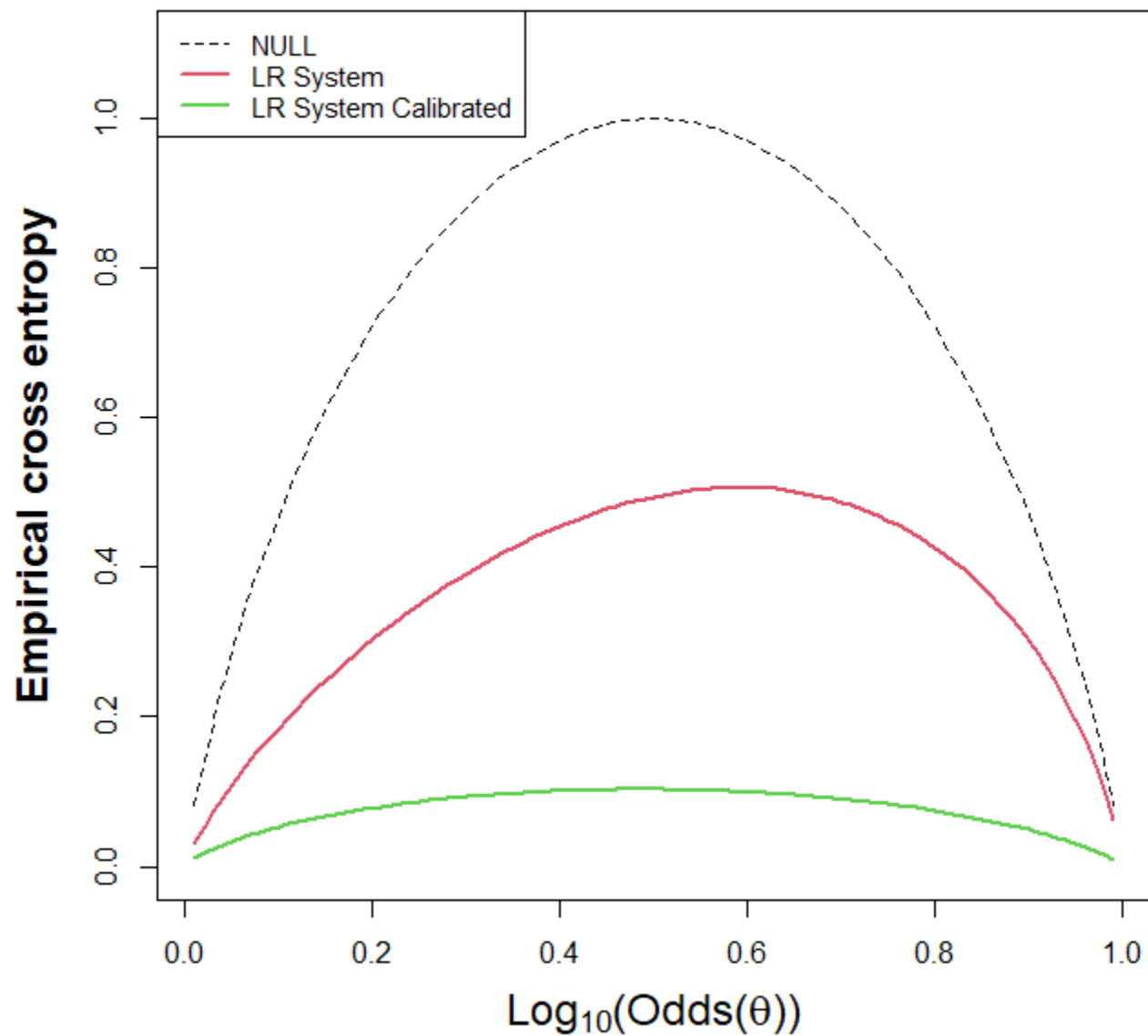
True Contributor LRs  
Non-contributor LRs

Output:

Empirical cross entropy plot

# Calibration Accuracy

ECE Plot for PROVEDIt Data  
All 4P Mixtures (n = 263 in each group)



# Interval Specific Calibration Discrepancy Plot



Are reported likelihood ratios well calibrated?

Jan Hannig<sup>a,c,\*</sup>, Sarah Riman<sup>b</sup>, Hari Iyer<sup>a</sup>, Peter M. Vallone<sup>b</sup>

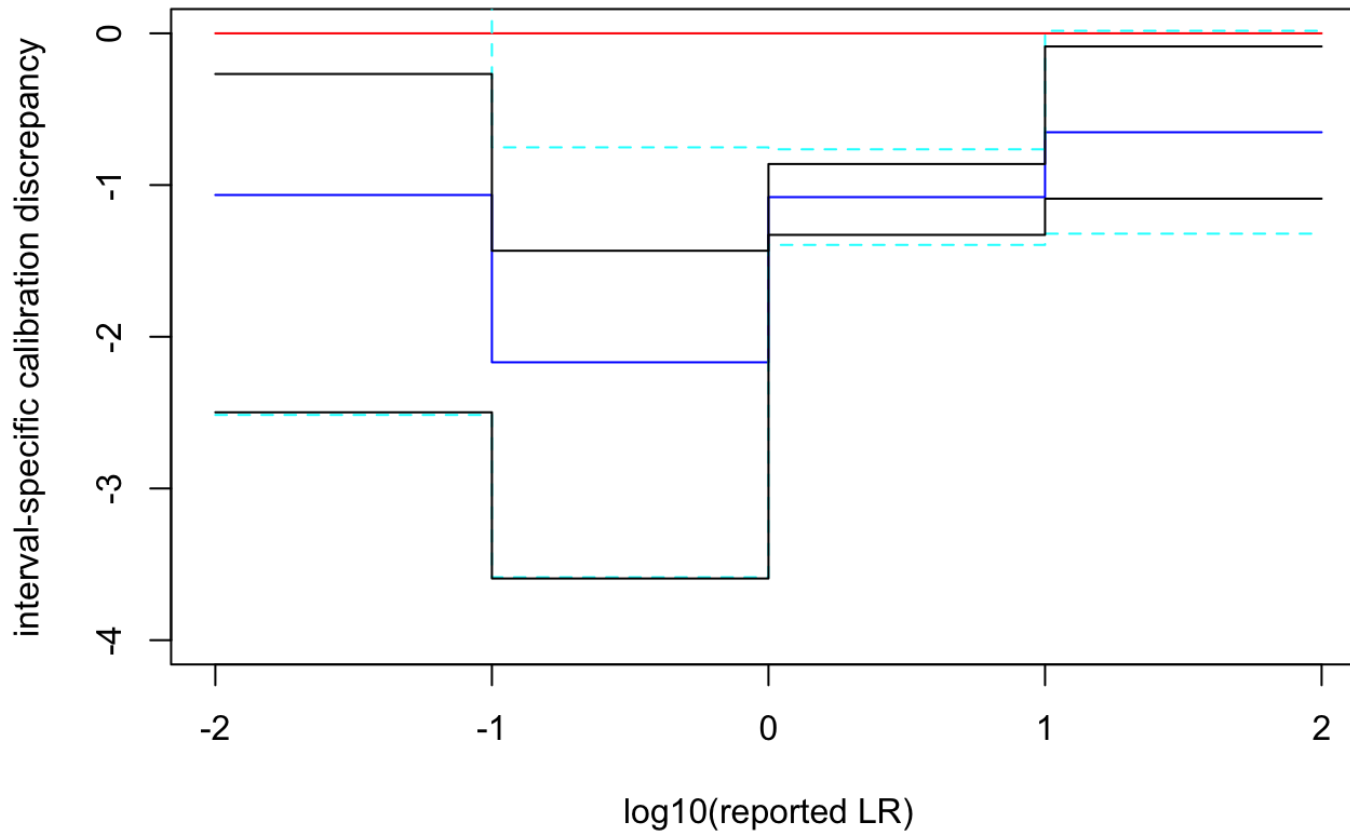
<sup>a</sup> Statistical Design, Analysis, and Modeling Group, ITL/NIST, United States

<sup>b</sup> Applied Genetics Group, National Institute of Standards and Technology, United States

<sup>c</sup> Department of Statistics and Operations Research, UNC-Chapel Hill, United States

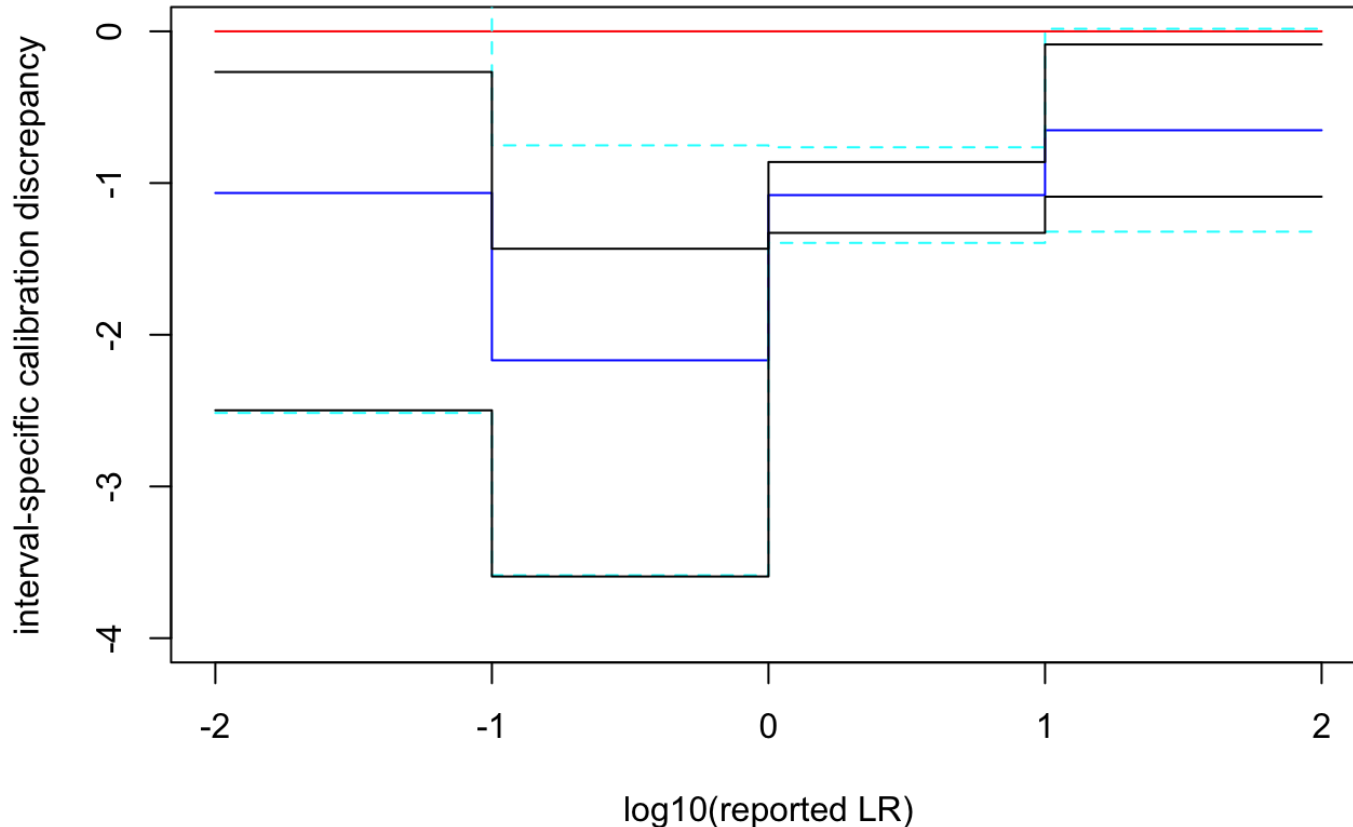


## PROVEDIt Data: All 4P Mixtures 263 Noncontributor LR<sub>s</sub>, 263 Contributor LR<sub>s</sub>



# Interval Specific Calibration Discrepancy Plot

## PROVEDIt Data: All 4P Mixtures 263 Noncontributor LRs, 263 Contributor LRs



Are reported likelihood ratios well calibrated?

Jan Hannig<sup>a,c,\*</sup>, Sarah Riman<sup>b</sup>, Hari Iyer<sup>a</sup>, Peter M. Vallone<sup>b</sup>

<sup>a</sup> Statistical Design, Analysis, and Modeling Group, ITL/NIST, United States

<sup>b</sup> Applied Genetics Group, National Institute of Standards and Technology, United States

<sup>c</sup> Department of Statistics and Operations Research, UNC-Chapel Hill, United States



**Calibration of STRmix LRs following the method of Hannig *et al.***

John Buckleton<sup>1,2</sup>, Maarten Kruijver<sup>2</sup>, James Curran<sup>1</sup>, and Jo-Anne Bright<sup>2</sup>

[https://figshare.com/articles/Calibration\\_of\\_STRmix\\_LRs\\_following\\_the\\_method\\_of\\_Hannig\\_et\\_al\\_/12324011/1](https://figshare.com/articles/Calibration_of_STRmix_LRs_following_the_method_of_Hannig_et_al_/12324011/1)

# ISO/IEC 19795-1

*Sufficient samples shall be collected per test subject so that the total number of attempts exceeds that required by the **Rule of 3** or **Rule of 30** as appropriate*

- What is the **RULE OF 3** and how is it applied when determining sample sizes?
- What is the **RULE OF 30** and how is it applied when determining sample sizes?

# Rule of 3

Suppose  $p$  = probability of an event of interest.

In  $N$  independent trials, the event of interest never occurred.

**Then we can be 95% confident that the value of  $p$  is at most  $3/N$ .**

## Illustration:

Event of interest: Non-contributor LR exceeding 5,000

Suppose no value of LR exceeded 5,000 in 1000 independent non-contributor tests. (So  $N=1000$ )

We can be 95% confident that the chances of a noncontributor test resulting in an  $LR > 5000$  will not exceed  $3/N = 3/1000 = 0.3 \%$

Turing's theorem says this probability should be less than or equal to  $1/5000 = 0.02\%$



# Rule of 30

Doddington et. al. (2000), *Speech Communication* (31), 225-254

The NIST speaker recognition evaluation – Overview,  
methodology, systems, results, perspective

George R. Doddington<sup>a,b</sup>, Mark A. Przybocki<sup>b</sup>, Alvin F. Martin<sup>b,\*</sup>,  
Douglas A. Reynolds<sup>c</sup>

2.5.2.2. *The rule of 30*. In determining the required size of a corpus, a helpful rule is what might be called “*the rule of 30*”. This comes directly from the binomial distribution, assuming independent trials. Here is the rule:

To be 90% confident that the true error rate is within  $\pm 30\%$  of the observed error rate, there must be at least 30 errors.

# Summary

- 1. Expected behavior of LR systems (when model is correct)**
- 2. Comparing validation study results with expectations – diagnostic checks**
- 3. Diagnostic checks are NECESSARY to demonstrate reliability but may not be sufficient**
- 4. Use of ROC (Receiver Operating Characteristic) plots to examine discrimination power and to compare discrimination power between two or more conditions (or two or more systems)**
- 5. Main requirements for reliability: Discrimination power and Calibration Accuracy**
- 6. Empirical Cross Entropy Plots and Interval Specific Calibration Discrepancy Plots**
- 7. Rule of 3 and Rule of 30 (ISO 19795-1)**



**In Module 5  
John will talk about  
Summarizing, Using, &  
Communicating  
Validation Data**



**ISHI 2020 Validation Workshop**  
Friday September 18th, 2020 // 9:00 am - 12:30 pm

Validation Principles, Practices, Parameters,  
Performance Evaluations, and Protocols  
**Summarizing, Using, &  
Communicating Validation Data**

**Module 5**

**John M. Butler**

National Institute of Standards and Technology



# Disclaimers

**Points of view are those of the presenter** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

## **Identification does not imply endorsement**

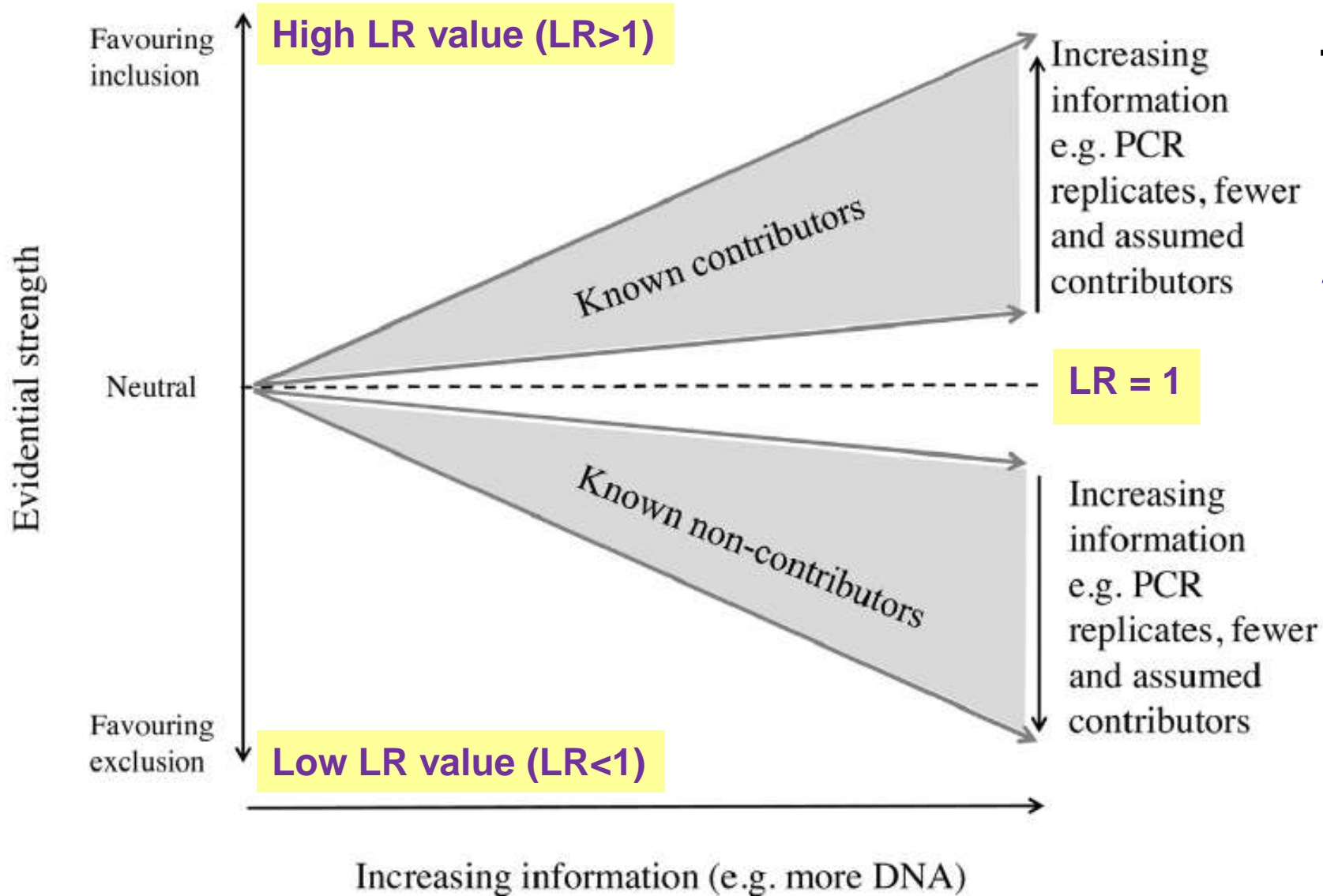
Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

# Module 5 (John)

- Summarizing Validation Data
  - Considering layout and what data to share to enable independent review of PGS data
- Using Validation Data to Inform Your Protocols
  - Establishing limits and a complexity threshold
- Communicating Validation Data and Meaning
  - Considering what questions are you answering with your data
  - Looking beyond PGS to larger issues with DNA mixture interpretation
- Some Final Thoughts

# Summmarizing Validation Data

# Desired Performance with a Mixture Interpretation Method



## Desirable Features

1. **Discrimination capacity**  
(separation of known contributors from known non-contributors)
2. **Calibration accuracy**  
(accuracy of a specific LR value)

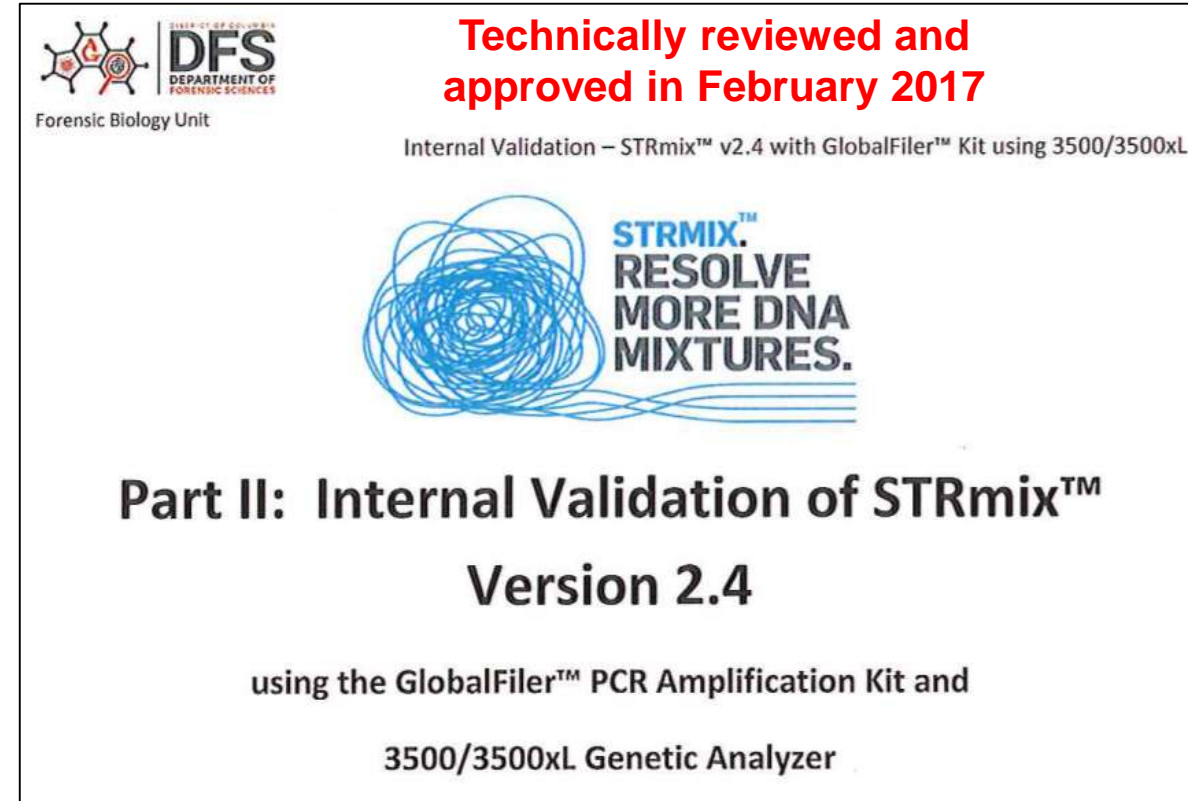
LR values vary based on amount of information available – **with less information, a lower LR value is obtained with a well-calibrated system**



# A Publicly Available PGS Internal Validation Summary

From page 11 of the summary report:

- “At high template STRmix correctly and reliably gave a high LR for true contributors and a low LR for false contributors.”
- “At low template or high contributor number STRmix correctly and reliably reported that the analysis of the sample **tends towards uninformative or inconclusive.**”



The image shows the cover of a technical report. At the top left is the logo for the Forensic Biology Unit, Department of Forensic Sciences (DFS), University of Columbia. To the right, in red text, it says "Technically reviewed and approved in February 2017". Below that, in smaller black text, it says "Internal Validation – STRmix™ v2.4 with GlobalFiler™ Kit using 3500/3500xL". In the center is a blue graphic of a tangled DNA strand with the text "STRMIX™ RESOLVE MORE DNA MIXTURES." to its right. Below the graphic, the title "Part II: Internal Validation of STRmix™" is written in large black font, followed by "Version 2.4" in a slightly smaller font. At the bottom, it says "using the GlobalFiler™ PCR Amplification Kit and 3500/3500xL Genetic Analyzer".

Forensic Biology Unit

DFS  
DEPARTMENT OF  
FORENSIC SCIENCES

Technically reviewed and  
approved in February 2017

Internal Validation – STRmix™ v2.4 with GlobalFiler™ Kit using 3500/3500xL

STRMIX™  
RESOLVE  
MORE DNA  
MIXTURES.

Part II: Internal Validation of STRmix™  
Version 2.4

using the GlobalFiler™ PCR Amplification Kit and  
3500/3500xL Genetic Analyzer

**If this is all we have, do these statements and any provided data summaries assist in understanding limitations of the system and where potential risks may exist?**

Appendix 2: Cross reference for document sections and SWGDAM recommendations

Standard	Text	Refer section
4.1	Test the system using representative data	Preamble
4.1.1	Specimens with known contributors	Preamble
4.1.2	Hypothesis testing with contributors and non-contributors	D
4.1.2.1	More than one set of hypotheses	E
4.1.3	Variable DNA typing conditions	Preamble
4.1.4	Allelic peak height, to include off-scale peaks	B
4.1.5	Single-source specimens	A
4.1.6	Mixed specimens	D
4.1.6.1	Various contributor ratios	D
4.1.6.2	Various total DNA template quantities	D
4.1.6.3	Various numbers of contributors	D
4.1.6.4	Both correct and incorrect number of contributors (i.e., over- and under-estimating)	F
4.1.6.5	Sharing of alleles among contributors	D
4.1.7	Partial profiles	D
4.1.7.1	Allele and locus drop-out	D
4.1.7.2	DNA degradation	L
4.1.7.3	Inhibition	L
4.1.8	Allele drop-in	G
4.1.9	Forward and reverse stutter	H
4.1.10	Intra-locus peak height variance	I
4.1.11	Inter-locus peak height variance	J
4.1.12	In-house parameters	Preamble
4.1.13	Sensitivity, specificity and precision	D and M
4.1.14	Additional challenge testing	K
4.2	Compare the results of probabilistic genotyping and of manual interpretation	L
4.2.1	Intuitive and consistent with expectations	L
4.2.1.1	Known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping	L
4.2.1.2	Concordance of single-source specimens with high quality results	A
4.2.1.3	Generally, as the analyst's ability to deconvolute a complex mixture decreases, so does the weighting of a genotype set determined by the software	C

## Correlation between Internal Validation Summary Topics and SWGDAM 2015 PGS Validation Guidelines

Showing where to find relevant information in an internal validation summary is helpful

- Analysts and auditors should **avoid using this as a checklist** and **seek to understand how performance metrics have been demonstrated**

## An Example of Information Provided in an Internal Validation Summary

Page 7 of 43: “These profiles represent typical profiles encountered by the laboratory. The **profiles are of varying DNA quantity and mixture proportions**. The contributors include homozygote and heterozygote alleles and **there is varying amounts of allele sharing across the different loci** ([SWGDM 2015 guidelines] standard 4.1.6.5). Given the template amounts, allele and/or locus dropout was expected to occur within the profiles containing the lower DNA amounts ([SWGDM 2015] standard 4.1.7.1).

Page 32 of 43: “Section D and E results demonstrate that **there may be overlap in likelihood ratios between true contributors and non-contributors below LR=100** (i.e., low true inclusions and high false inclusions) for three, four, and five person mixtures. Based on this information, LRs between 1 and 100 will be designated “Uninformative” for casework samples in the Forensic Biology unit at DFS.”

# An Analysis of Factor Space Coverage for an Internal Validation

DC-DFS, STRmix v2.4, GlobalFiler (29 cycles), ABI 3500

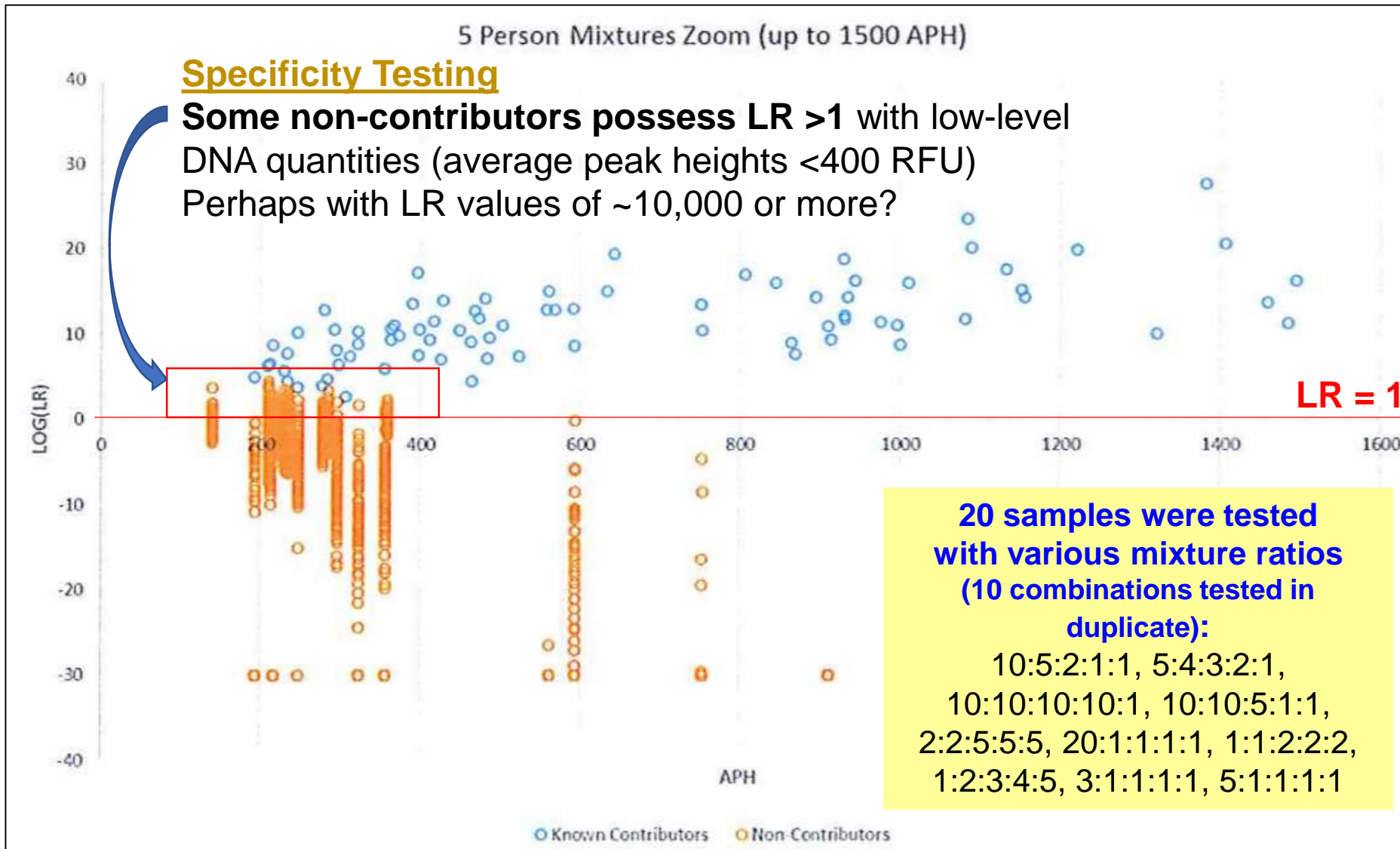
# contributors	# samples	DNA template amounts (pg)	Mixture ratios <i>(deciphered from Appendix 3)</i>	Degree of allele sharing
single-source	32	high amount of DNA (3000 pg), 250, 188, 125, 94, 63, 47, 31, 23, 15, 12, 6 pg	N/A	<i>No information</i>
2	42	<i>Not apparent from Appendix 3</i>	25:1, 20:1, 15:1, 10:1, 7:1, 5:1, 3:1, 2:1, 1:1 1:1, 1:2, 1:3, 1:5, 1:7, 1:10, 1:15, 1:20, 1:25	<i>No information</i>
3	20	<i>Not apparent from Appendix 3</i>	3:1:1, 1:10:20, 1:2:3, 10:5:1, 3:1:1, 1:1:5, 20:10:1, 3:2:1, 1:2:10, 1:5:10	<i>No information</i>
4	20	<i>Not apparent from Appendix 3</i>	2:2:2:1, 20:5:2:1, 5:1:1:1, 5:2:1:1, 5:5:5:1, 1:2:3:4, 3:3:2:1, 1:3:5:10, 2:2:1:1, 20:10:1:1, 1:1:1:3, 1:1:1:5, 1:1:1:7	<i>No information</i>
5	20	<i>Not apparent from Appendix 3</i>	10:5:2:1:1, 5:4:3:2:1, 10:10:10:10:1, 10:10:5:1:1, 2:2:5:5:5, 20:1:1:1:1, 1:1:2:2:2, 1:2:3:4:5, 3:1:1:1:1, 5:1:1:1:1	<i>No information</i>
Various	10			<i>No information</i>

## NIST Summary of Factor Space Coverage from this Internal Validation Summary

Summary page 7 of 43:

“Each profile was interpreted in STRmix and compared to the known contributors and **134 known non-contributors...**”

# 5 Person Mixture Plot of Average Peak Height vs Log(LR)



*No correlation between data points and samples used to generate them making it challenging to understand what aspect of the factor space is being covered*

**Blue circles** = LR assigned with known true contributors

- 20x5 tested?

**Orange circles** = LR assigned with known non-contributors

- 134x100 tested?

# What is Needed to Enable an Independent Review?

**A. LR values** (PGS LR assignments given specific propositions) for each data point

**B. Factor space coverage details**

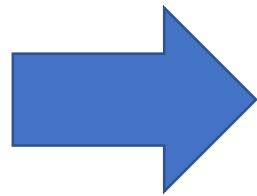
1. Sample ID
2. Sample Number (if a replicate)
3. Number of Contributors
4. Target Template Amounts
5. Degradation Status
6. NOC used for Analysis
7.  $H_1$  ( $H_p$ ) True? Yes/No
8. POI position (if  $H_1$  True)
- 9. Reported  $\log_{10}(\text{LR})$  by PGS system**
10. Mixture EPG results\*
11. POI profile\*
12. Known Contributor-A profile\*
13. Known Contributor-B profile\*
14. Etc. for additional known contributors\*

\* if privacy of the profile genotypes is a concern, then alleles in an algebraic format could be used as described previously ([Gill et al. 1998 FSI 91:41-53](#)). For example, the letters A, B, C, D, etc. can be used in place of actual alleles at the various loci

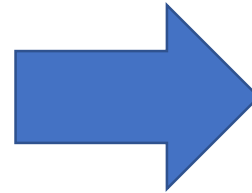
# Using Validation Data to Inform Your Protocols

# Validation Data Should Inform Laboratory Protocols

**Generate**  
Validation  
Data



**Create**  
Protocols

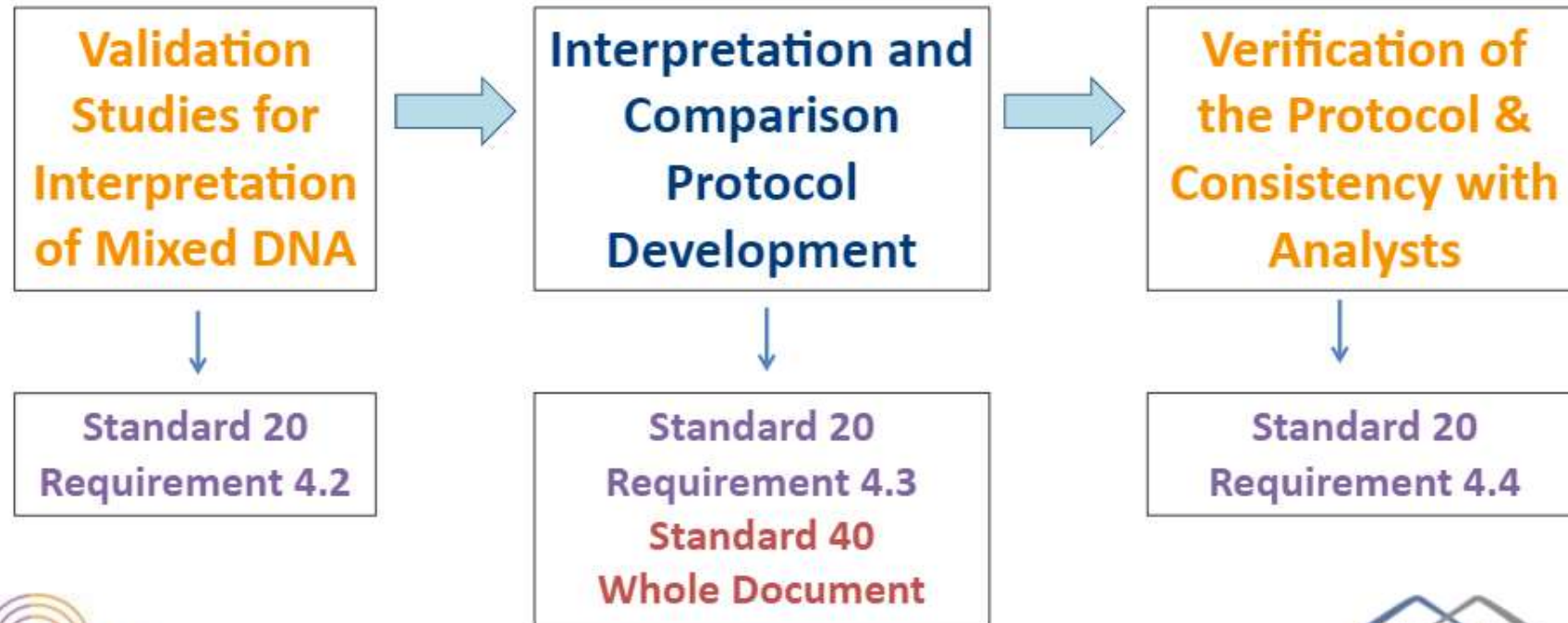


**Verify Protocols**  
*Additional Testing*  
across a range of  
sample types



# From Charlotte Word's August 5, 2020 Webinar

## STANDARDS 20 & 40 STRUCTURE



These ASB Standards are available at

<http://www.asbstandardsboard.org/published-documents/dna-published-documents/>

# ANAB Accreditation Requirements Related to ISO/IEC 17025:2017 (AR 3125)

## 7.2.2 Validation of methods

### 7.2.2.1.1

The laboratory shall have a procedure for method validation that:

- a) includes the associated data analysis and interpretation;
- b) establishes the data required to report a result, opinion, or interpretation; and
- c) identifies limitations of the method, reported results, opinions, and interpretations.

### 7.2.2.2

**NOTE** Changes to associated data analysis and interpretation are considered changes to a validated method.

<https://anab.qualtraxcloud.com/ShowDocument.aspx?ID=12371>

# Setting Limits → A Complexity Threshold

## A COMPLEXITY THRESHOLD?

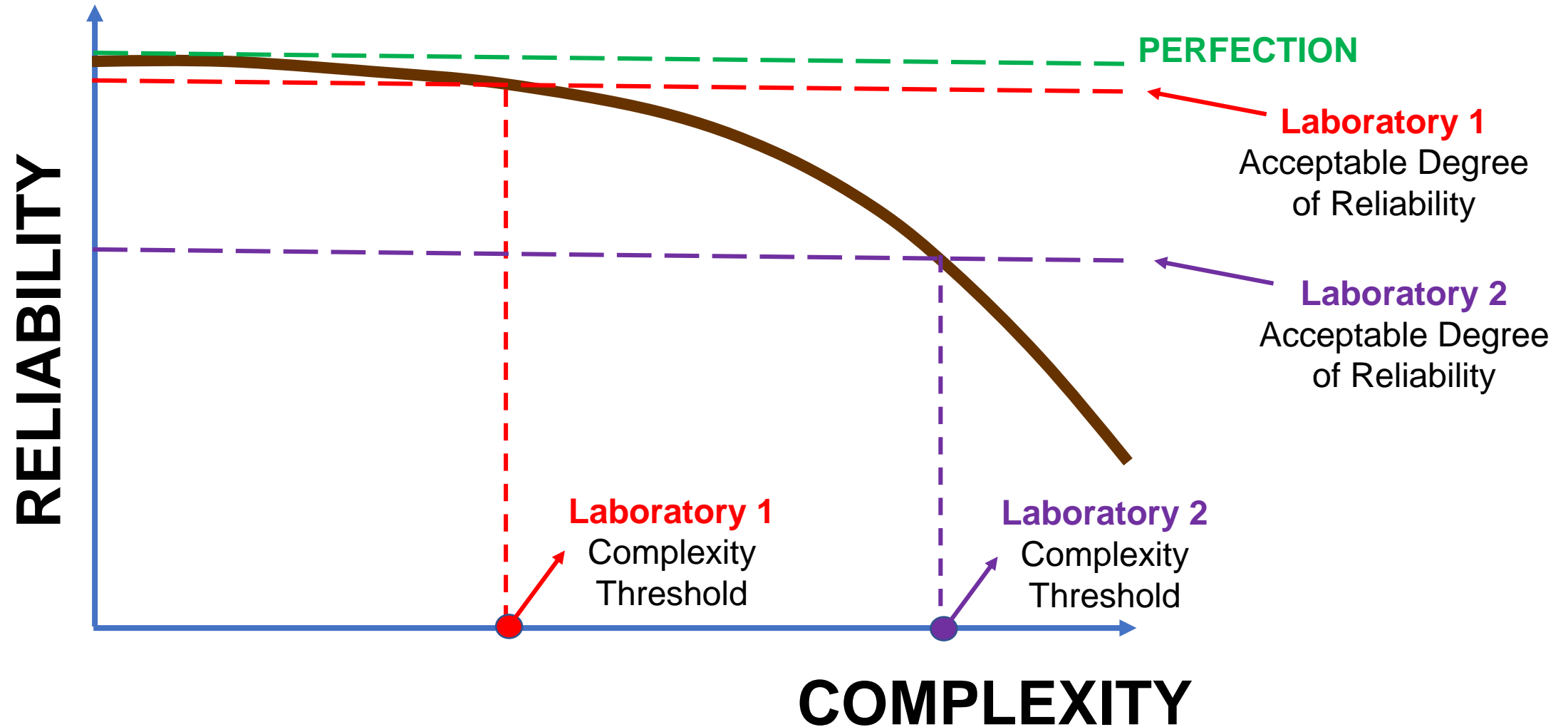
---

Some DNA mixtures will be too complex to solve. Laboratories may benefit from developing criteria for when to stop working on a sample or on a case based on a preliminary analysis of samples received. This might be termed a “complexity threshold” (Rudin & Inman 2012). One idea for creating a complexity threshold is the use of receiver operator characteristics (ROC) curves that correlate the number of false positives and false negatives under certain conditions (Gordon 2012, Grgicak 2012). For example, simulations can be run and visualized via ROC curves to determine how many non-concordant results (i.e. missing alleles in the evidence sample) are permitted before there is a chosen probability of falsely including an innocent person (Gordon 2012).

In one of their complex mixture studies, NFI proposed to develop criteria for assessing the peak heights, position of allele calls (such as in potential stutter positions), the consistency of allele calls among replicates, and a maximum number of allele drop-outs that could be considered for non-concordance (Benschop et al. 2012). Presumably studying the variability of these parameters in validation studies with known mixture contributors could lead to an effective complexity threshold.

In April 2012, an international conference was held in Rome, Italy, entitled “The hidden side of DNA profiles: artifacts, errors and uncertain evidence” (Pascali & Prinz 2012). Peter Schneider, a forensic DNA researcher from Cologne, Germany, shared his thoughts on what to do when evidence becomes too complex to reliably interpret: “If you cannot explain your evidence to someone that is not from the field (like a judge) – and you need a lot of technical excuses to report something – then the result is not good. You should leave it on your desk and not take it to court. This is a very common sense approach to this problem” (Rome 2012).

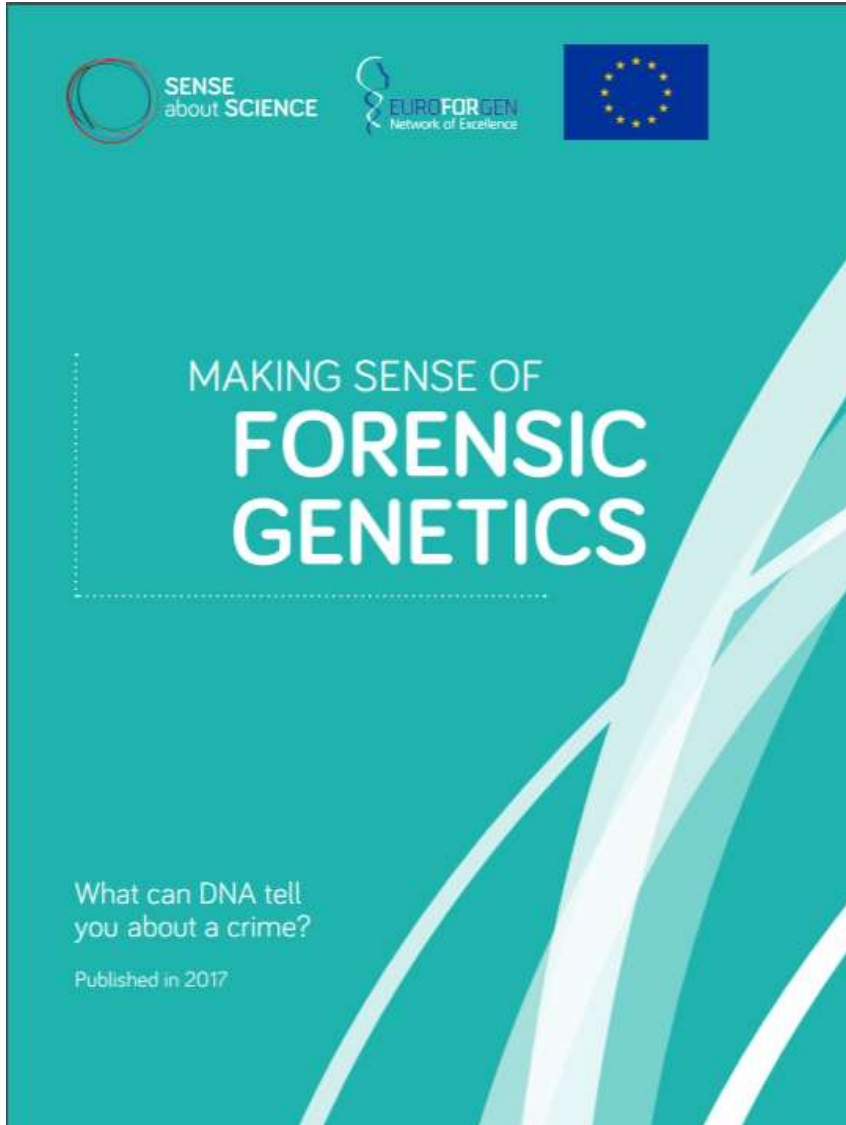
# Your Complexity Threshold is Related to Your Acceptable Degree of Reliability



# Communicating Validation Data

# Making Sense of Forensic Genetics (2017)

concepts clearly explained in 40 pages



- Developed by European Forensic Genetics Network of Excellence (EuroForGen-NoE) and published with Sense about Science
- **Free PDF file** available for download  
<https://senseaboutscience.org/wp-content/uploads/2017/01/making-sense-of-forensic-genetics.pdf>
- *Final point made:* “As DNA profiling continues to grow more sensitive, and it is used in more investigations, **the need for accurate communication between scientists and nonscientists only grows** - both **to ensure that their expectations of the technology are realistic, and its limits are properly understood...**”

# Know What Question You Are Trying to Answer



**David Balding**  
University of Melbourne  
Professor of Mathematics  
and Statistics

“...**Focus on the relevant question.**  
Many misleading statistical  
approaches [turn] out to be providing  
valid answers to the wrong  
questions.”

- David Balding, Interpreting DNA evidence: can probability theory help? In J.L. Gastwirth (ed.) *Statistical Science in the Courtroom* (pp. 51-70) New York: Springer, 2000

# Recent ISFG DNA Commission Articles

## *Forensic Sci. Int. Genet.* (2018) 36: 189-202

DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence - Guidelines highlighting the importance of propositions

Part I: evaluation of DNA profiling comparisons given (sub-) source propositions

Peter Gill<sup>a,b,\*,1</sup>, Tacha Hicks<sup>c,d,\*,1</sup>, John M. Butler<sup>e</sup>, Ed Connolly<sup>f</sup>, Leonor Gusmão<sup>g,h,i</sup>, Bas Kokshoorn<sup>j</sup>, Niels Morling<sup>k</sup>, Roland A.H. van Oorschot<sup>l,m</sup>, Walther Parson<sup>n,o</sup>, Mechthild Prinz<sup>p</sup>, Peter M. Schneider<sup>q</sup>, Titia Sijen<sup>j</sup>, Duncan Taylor<sup>r,s</sup>

## *Forensic Sci. Int. Genet.* (2020) 44: 102186

DNA commission of the International society for forensic genetics: Assessing the value of forensic biological evidence - Guidelines highlighting the importance of propositions. Part II: Evaluation of biological traces considering activity level propositions

Peter Gill<sup>a,b,\*,1</sup>, Tacha Hicks<sup>c,d,1</sup>, John M. Butler<sup>e</sup>, Ed Connolly<sup>f</sup>, Leonor Gusmão<sup>g,h,i</sup>, Bas Kokshoorn<sup>j</sup>, Niels Morling<sup>k</sup>, Roland A.H. van Oorschot<sup>l,m</sup>, Walther Parson<sup>n,o</sup>, Mechthild Prinz<sup>p</sup>, Peter M. Schneider<sup>q</sup>, Titia Sijen<sup>j</sup>, Duncan Taylor<sup>r,s</sup>

## 2018

- Difference between **investigative and evaluative reporting** is explained
- Common pitfalls of **formulating propositions** are discussed
- **Challenges of low-level mixtures** are discussed

## 2020

- Why, when and how to carry out evaluation given **activity level propositions** are addressed with examples
- Distinguishing between **results, propositions and explanations**



Levels in Hierarchy of Propositions	Purpose	Issues & Questions Addressed	Results Used	Factors Considered
<b>Sub-source</b>	Investigation	Who could be the source of the DNA?	DNA profile	Occurrence of DNA profile genotypes in the relevant population; variability of results (e.g., presence or absence of alleles) assuming the DNA came from the POI
	Evaluation	<b>Is the DNA from the person of interest (POI)?</b>		
<b>Source</b>	Investigation	Who could be the source of the biological fluid?	DNA profile; biological fluid presumptive tests	(Sub-source factors) + presumptive test false positive/ false negative rates (e.g., cross-reactivity, etc.)
	Evaluation	<b>Is the biological fluid from the POI?</b>		
<b>Activity</b>	Evaluation	<b>Did the POI perform the given activity?</b>	DNA profile; biological fluid presumptive tests; relative quantity of DNA; where DNA was recovered; existence of multiple samples	(Source factors) + <b>DNA transfer, persistence, and recovery; DNA present for unknown reasons (i.e., background DNA)</b>

**sub-sub-source if only a portion of a DNA mixture is considered**

See Taylor et al. (2018) Evaluation of forensic genetics findings given activity level propositions: A review. *Forensic Sci Int Genet.* 2018;36:34-49.

# Catalog of Research on DNA Transfer Studies

Forensic Science International: Genetics 40 (2019) 24–36



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



Review article

On DNA transfer: The lack and difficulty of systematic research and how to do it better



Annica Gosch, Cornelius Courts\*

*Institute of Forensic Medicine, University Hospital of Schleswig-Holstein, Arnold-Heller-Strasse 12, 24105 Kiel, Germany*

## ARTICLE INFO

### Keywords:

Forensic genetics  
DNA transfer  
Touch DNA  
Trace DNA

## ABSTRACT

Since DNA from touched items and surfaces (“touch DNA”) can successfully and reliably be analyzed, the question as to how a particular DNA containing sample came to be from where it was recovered is of increasing forensic interest and expert witnesses in court are increasingly challenged to assess for instance whether an incriminatory DNA sample matching to a suspect could have been transferred to the crime scene in an innocent manner and to guess at the probability of such an occurrence. The latter however will frequently entail expressing a subjective probability i.e. simply making a best guess from experience.

There is, to the present date, an extensive and complex body of literature on primary, secondary, tertiary and even higher order DNA transfer, its possibility, plausibility, dependency on an array of variables and factors and vast numbers of permutations thereof. However, from our point of view there is a lack of systematic data on DNA transfer with existing research widely varying in quality and relevance.

This German group developed an open resource and Microsoft Access database of **published research on DNA transfer** (called “DNA-TrAC”)

– see Appendix A of their article



Article in the September 2020 issue

Forensic Science International: Genetics

journal homepage: [www.elsevier.com/locate/fsigen](http://www.elsevier.com/locate/fsigen)



**Examined DNA mixtures from skin contact traces of DNA recovered from three surfaces of two types of firearms handled in four realistic, casework-relevant handling scenarios**

Research paper

## DNA transfer to firearms in alternative realistic handling scenarios

Annica Gosch, Jan Euteneuer, Johanna Preuß-Wössner, Cornelius Courts\*

*Institute of Forensic Medicine, University Medical Center Schleswig-Holstein, Kiel, Germany*

### ARTICLE INFO

**Keywords:**

DNA transfer

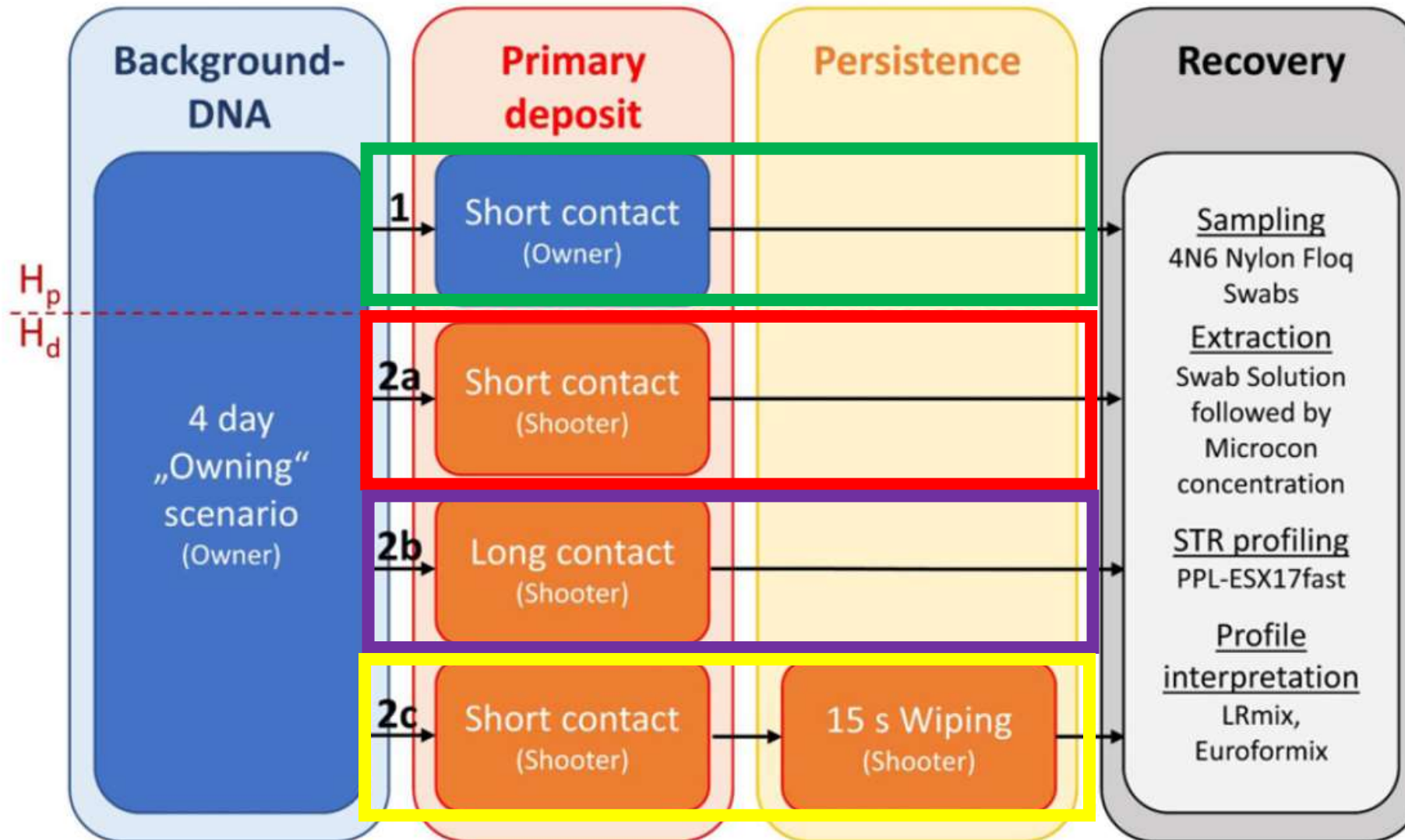
Firearms

Touch DNA

### ABSTRACT

Firearms are the most relevant items of evidence in gun-related crimes, likely bearing various traces facilitating an objective reconstruction of the crime. Trace DNA recovered from firearm surfaces might help to identify individual(s) having handled the firearm and thereby possibly to link the firearm and the corresponding shooter, however, the interpretation of DNA traces on handled items can be challenging and requires a detailed understanding of various factors impacting DNA prevalence, transfer, persistence and recovery. Herein, we aimed at improving our understanding of factors affecting the variability of trace DNA characteristics recovered from firearms handled in gun-related crimes: Skin contact traces were recovered from various outer surfaces of two types of firearms handled in four realistic, casework-relevant handling scenarios and the corresponding trace characteristics (DNA yield, number of contributors, relative profile contribution for known and unknown contributors, LRs) were compared. Trace DNA characteristics differed distinctly between handling conditions, firearm and surface types as well as handling individuals and intraindividual deposits emphasizing the variability and complexity of trace DNA profile composition expected to be recovered from firearms after realistic handling scenarios. The obtained results can provide useful insights for forensic experts evaluating alternative activity level propositions in gun-related crimes.

# First Research Study of DNA Transfer on Firearms with Casework-Relevant Alternative Handling Scenarios



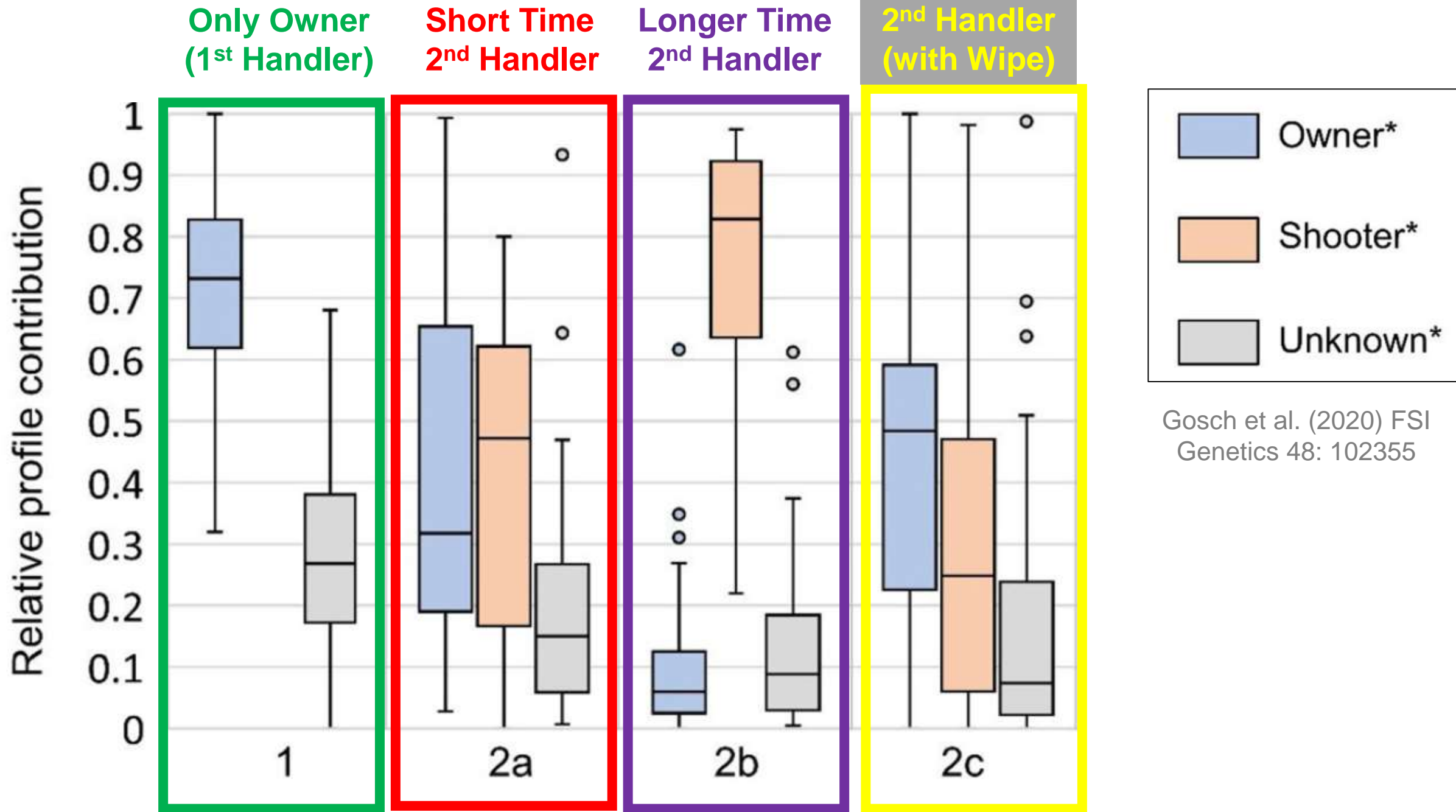
Only Owner  
(1<sup>st</sup> Handler)

Short Time  
2<sup>nd</sup> Handler

Longer Time  
2<sup>nd</sup> Handler

Short Time  
2<sup>nd</sup> Handler (with Wipe)

*Each repeated three times with two different owner/shooter pairs*



# Some Final Thoughts

# A Public Repository of Example Data is Desirable

ISFG DNA Commission (Coble et al. 2016)

## Recommendation #16:

The DNA Commission encourages the forensic community to establish a public repository of typing results from adjudicated casework covering a wide range of kinship cases and mixture samples including different challenging scenarios like low-level mixtures and related contributors. The data need to be in a universal, useful file format. The repository should be governed by a neutral organization providing equal access to all interested international parties.

- ...Meta-data associated with the submitted profiles should include relevant information such as the kit used, PCR cycle conditions, the separation polymer used, the CE system electrophoretic injection parameters, and any other relevant information about the sample.

An example is the **PROVEDIt data set** (<https://lftdi.camden.rutgers.edu/provedit/files/>):

Alfonse, L.E., Garrett, A.D., Lun, D.S., Duffy, K.R. & Grgicak, C.M. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Sci. Int. Genetics* 32: 62-70.

# Working Towards A Collaborative Validation Approach

Forensic Science International: Synergy 2 (2020) 230–237

Contents lists available at [ScienceDirect](#)

 Forensic Science International: Synergy

journal homepage: <https://www.journals.elsevier.com/forensic-science-international-synergy/>



---

Collaborative versus traditional method validation approach:  
Discussion and business case

Ray Wickenheiser <sup>a,\*</sup>, Laurel Farrell <sup>b</sup>

<sup>a</sup> *New York State Police Crime Laboratory System, Albany, NY, USA*  
<sup>b</sup> *ANSI National Accreditation Board, Milwaukee, WI, USA*

[Open Access]  
<https://doi.org/10.1016/j.fsisy.2020.08.003>



“Utilization of published validation data increases efficiency through shared experiences...”



# Learn from Previous Work (Internal Validation Studies)

Unfortunately, there are a limited number of **PGS internal validation study summaries that are publicly available\***

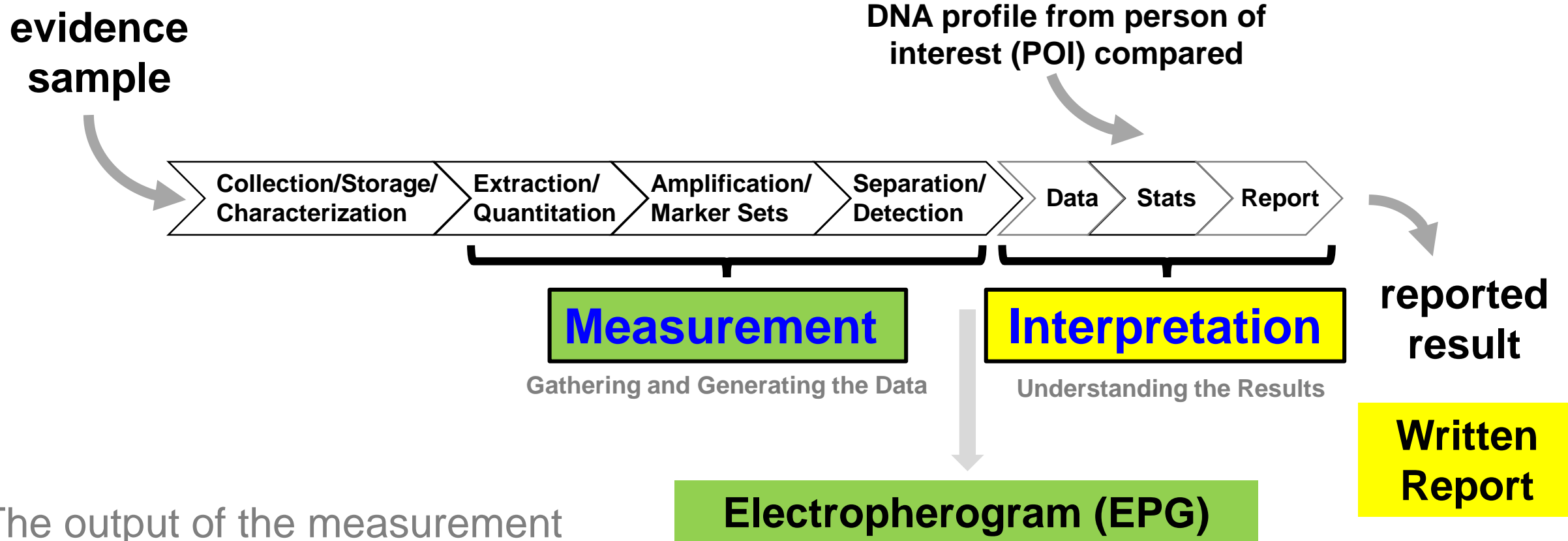
Forensic Laboratory	Information Available and Website
California Department of Justice DNA Laboratory	STRmix v2.06 (Identifiler Plus, ABI 3130/3500) <a href="https://epic.org/state-policy/foia/dna-software/EPIC-16-02-02-CalDOJ-FOIA-20160219-STRmix-V2.0.6-Validation-Summaries.pdf">https://epic.org/state-policy/foia/dna-software/EPIC-16-02-02-CalDOJ-FOIA-20160219-STRmix-V2.0.6-Validation-Summaries.pdf</a>
Erie County Central Police Services Forensic Laboratory (Buffalo, NY)	STRmix v2.3 (PowerPlex Fusion, ABI 3500) <a href="https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementation-and-internal-validation-erie-fusion.pdf">https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementation-and-internal-validation-erie-fusion.pdf</a> STRmix v2.3 (Identifiler Plus, ABI 3500) <a href="https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementation-and-internal-validation-erie-id-plus.pdf">https://johnbuckleton.files.wordpress.com/2016/09/strmix-implementation-and-internal-validation-erie-id-plus.pdf</a>
Michigan State Police	STRmix v2.3.07 (PowerPlex Fusion, ABI 3500/3500xl) <a href="https://johnbuckleton.files.wordpress.com/2016/09/strmix-summary-msp.pdf">https://johnbuckleton.files.wordpress.com/2016/09/strmix-summary-msp.pdf</a>
NYC OCME Forensic Biology Laboratory	STRmix v2.4 (Fusion, ABI 3130xl) <a href="https://www1.nyc.gov/site/ocme/services/validation-summary.page">https://www1.nyc.gov/site/ocme/services/validation-summary.page</a>
Palm Beach County (FL) Sheriff's Office	STRmix v2.4 (PowerPlex Fusion, ABI 3500xl) <a href="http://www.pbso.org/qualtrax/QTDocuments/4228.PDF">http://www.pbso.org/qualtrax/QTDocuments/4228.PDF</a>
San Diego (CA) Police Department	STRmix (GlobalFiler, ABI 3500), STRmix v2.3.07; STRmix v2.4.06 <a href="https://www.sandiego.gov/police/services/crime-laboratory-documents">https://www.sandiego.gov/police/services/crime-laboratory-documents</a>
Virginia Department of Forensic Science	TrueAllele Casework (PowerPlex 16, ABI 3130xl) <a href="https://epic.org/state-policy/foia/dna-software/EPIC-15-10-13-VA-FOIA-20151104-Production-Pt2.pdf">https://epic.org/state-policy/foia/dna-software/EPIC-15-10-13-VA-FOIA-20151104-Production-Pt2.pdf</a>
Washington DC Department of Forensic Sciences	STRmix v2.4 parameters & validation report (GlobalFiler, ABI 3500) <a href="https://dfs.dc.gov/page/fbu-validation-studiesperformance-checks">https://dfs.dc.gov/page/fbu-validation-studiesperformance-checks</a>

*\*based on Google searches performed March 23, 2020*

Validation summaries (not data) from:

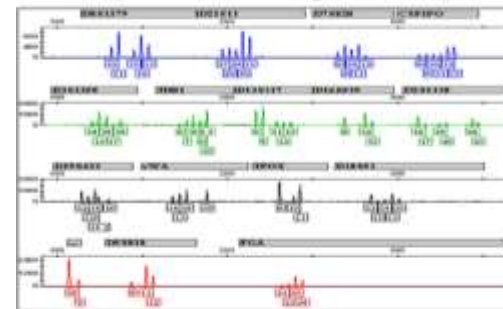
- 8 laboratories
- 8 STRmix
- 1 TrueAllele

# Steps involved in Processing an Evidence Sample containing DNA (either single-source or mixture)



The output of the measurement steps is an electropherogram

The output of interpretation is a reported result in a written report



“The origins of crime scene stains are not known with certainty, although these stains may match samples from specific people. The language of probability is designed to allow numerical statements about uncertainty, and we need to recognize that *probabilities are assigned by people rather than being inherent physical quantities*” (Evetts & Weir 1998, p. 21, *emphasis added*).

Evetts, I.W. and Weir, B.S. (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates: Sunderland, MA.

**AN IMPORTANT KEY TAKEAWAY:** Generating a DNA profile involves measuring the inherent physical properties of the sample. Interpreting a DNA profile involves judgments made by the DNA analyst assigning values that are not inherent to the sample based on other factors including case context and their own training and experience.

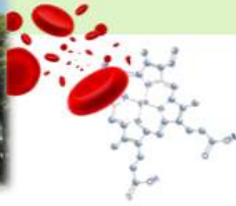
**Validation must address both measurement and interpretation**

# **For More Information, Come to ISFG 2021...**

August 23-27, 2021  
Washington, DC

# International Society for Forensic Genetics (ISFG)

[https://www.isfg.org/files/ISFG\\_50Years\\_Brochure.pdf](https://www.isfg.org/files/ISFG_50Years_Brochure.pdf)



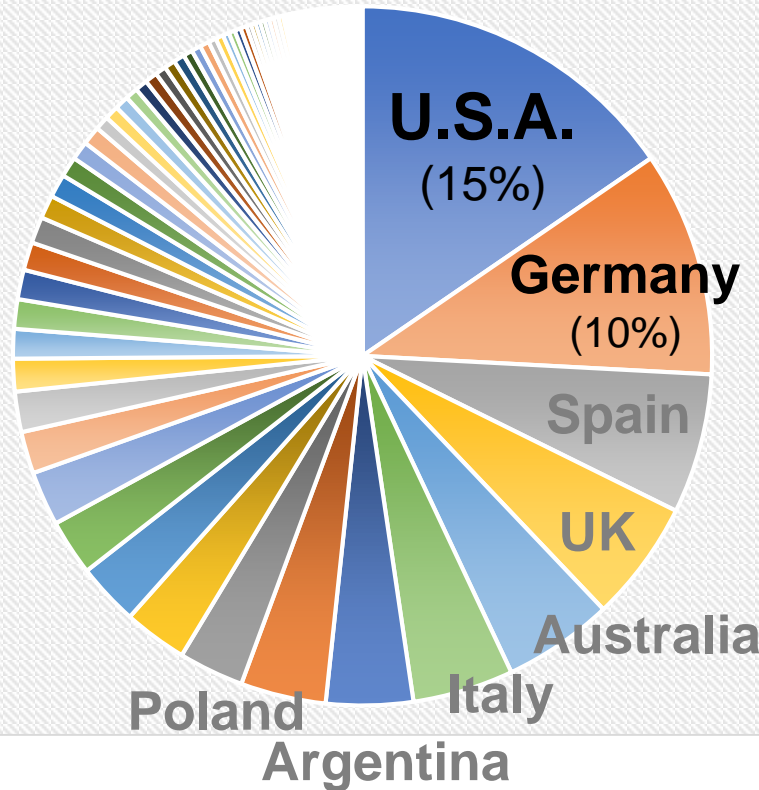
**50 Years**

**International Society  
for Forensic Genetics**

**1968 - 2018**



**1393 members  
from 84 countries**



**12 Working  
Groups**

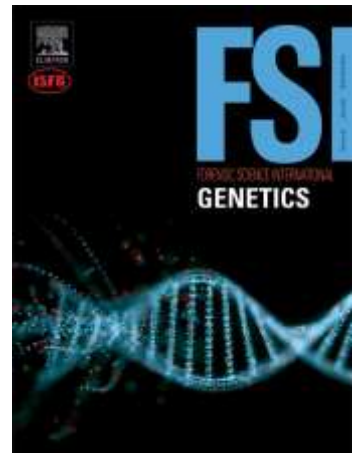
- ☑ German
- ☑ English
- ☑ French
- ☑ Italian
- ☑ Spanish and Portuguese
- ☑ Chinese
- ☑ Korean
- ☑ Arabian Speaking
- ☑ CaDNAP
- ☑ DNA Commission
- ☑ EDNAP
- ☑ Polish

**Biennial Meetings**



Prague (2019)

**#1 Journal on  
Forensic DNA**



**President:** John M. Butler, Gaithersburg • **Vice President:** Walther Parson, Innsbruck • **Secretary:** Peter M. Schneider, Cologne  
**Treasurer:** Marielle Vennemann, Münster • **Representative of the Working Groups:** Leonor Gusmão, Rio de Janeiro



# The Next ISFG Meeting is in the U.S.

<https://www.isfg2021.org/>

***Once in a Lifetime Opportunity*** – The best scientific meeting in the field with top researchers in forensic genetics coming to the United States for the first time in the 21<sup>st</sup> Century

## 16 Pre-Congress Workshops

To be held August 23-24, 2021

**DNA Mixtures (Basic)**

**DNA Mixtures (Advanced)**

Kinship Analysis

Y-STRs

**Court Testimony**

NGS Bioinformatics 101

NGS Methods | mtDNA Casework

NGS STR Markers | Phenotyping

**DNA Transfer | Evaluative Reporting**

**Probability and Statistics | Validation**

Biogeographical Ancestry | Publication

<https://www.isfg.org/Meeting>



*Previous Meetings:* Münster (2001), Archacon (2003), San Miguel, Azores (2005), Copenhagen (2007), Buenos Aires (2009), Vienna (2011), Melbourne (2013), Krakow (2015), Seoul (2017), Prague (2019)

# Thank you for your attention!

**Points of view are the presenters** and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

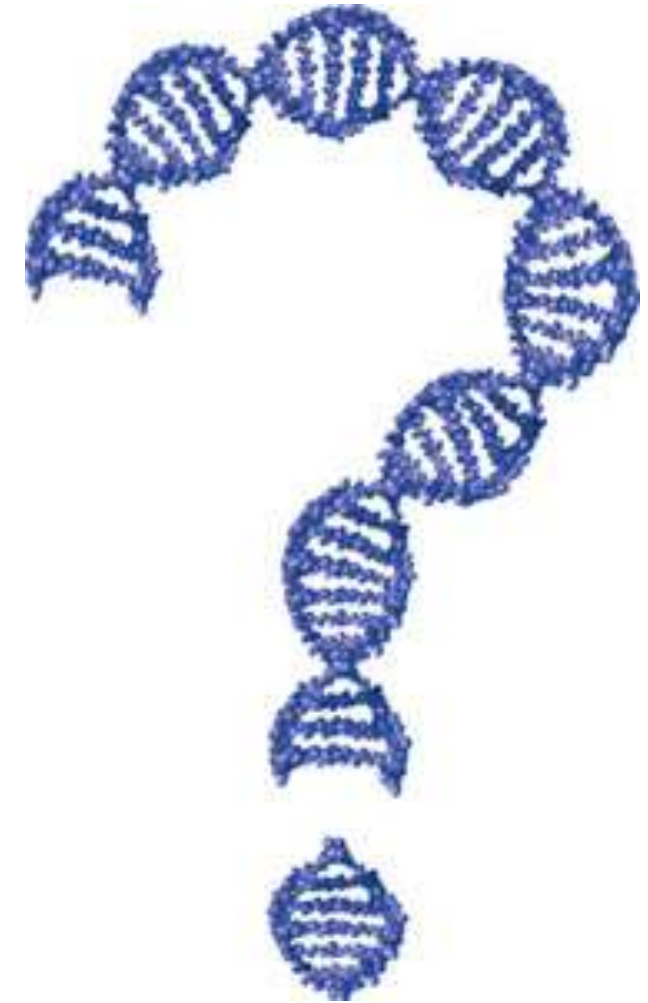
## Contact Information

**John M. Butler**

[john.butler@nist.gov](mailto:john.butler@nist.gov)

**Hari K. Iyer**

[hariharan.iyer@nist.gov](mailto:hariharan.iyer@nist.gov)



RESEARCH. STANDARDS. FOUNDATIONS.