Communicating Forensic Findings:

Framing ~~the~~ Some Issues

Steve Lund

Evidential Statistics Focus Area Lead

Statistical Engineering Division

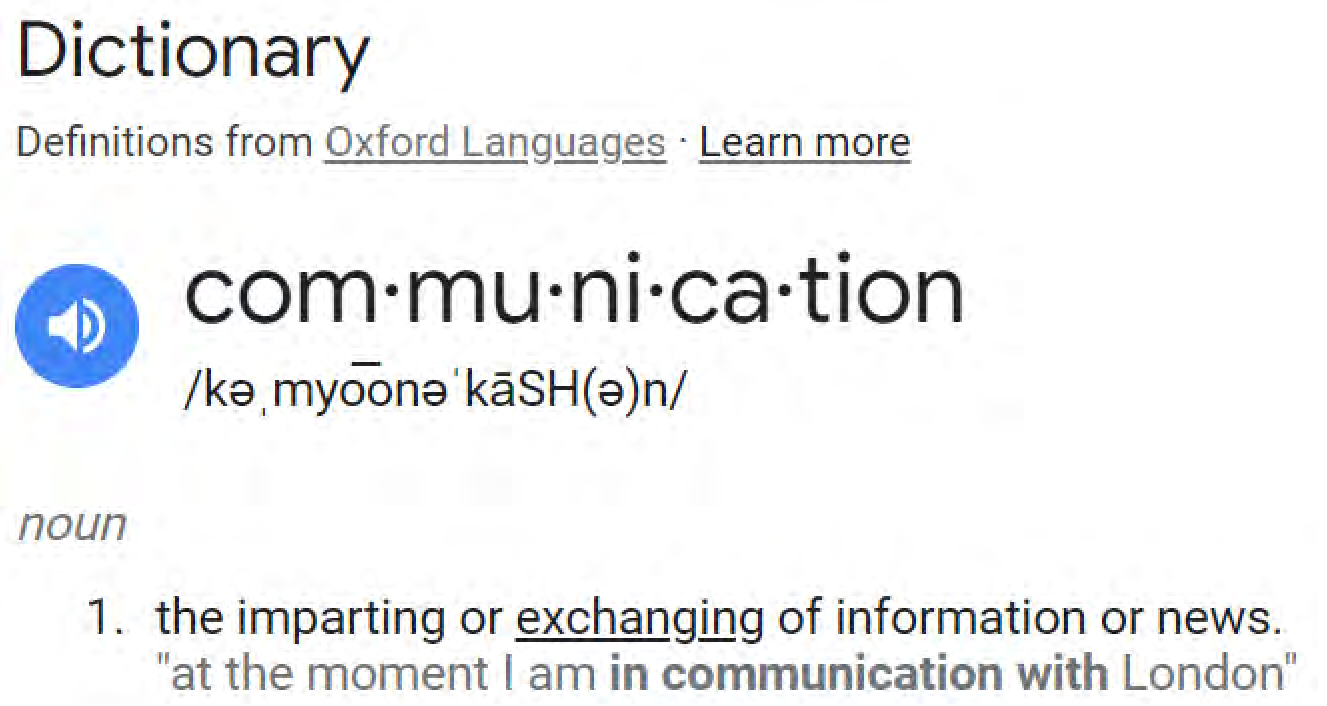
Acknowledgement

- Collaboration with colleagues Hari Iyer and Will Guthrie

Disclaimer

- Viewpoints expressed are our own and are not intended to reflect those of anyone else at NIST

Communicating Forensic Findings (CFF)



Dictionary

Definitions from Oxford Languages · Learn more

com·mu·ni·ca·tion

/kəˌmyo͞onəˈkāSH(ə)n/

noun

1. the imparting or exchanging of information or news.
   "at the moment I am **in communication with** London"

CFF: Experts imparting information to other parties in the judicial system.

What type of information?

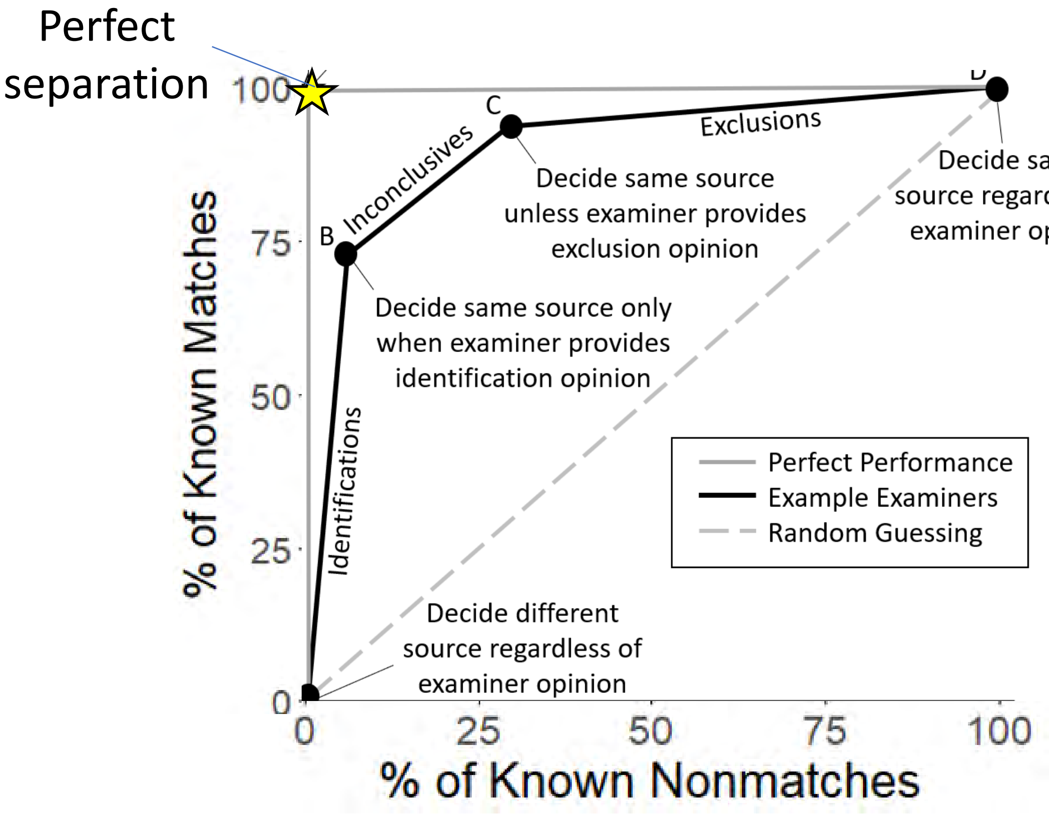
- **Observations about the evidence**
  - Descriptive, demonstrable
  - Often high-dimensional or complicated

- **Opinions of the expert(s)**
  - Interpretive, personal, some variability expected – “range of opinions”
  - Typically simpler than observations

CFF Goal: Help others make better decisions

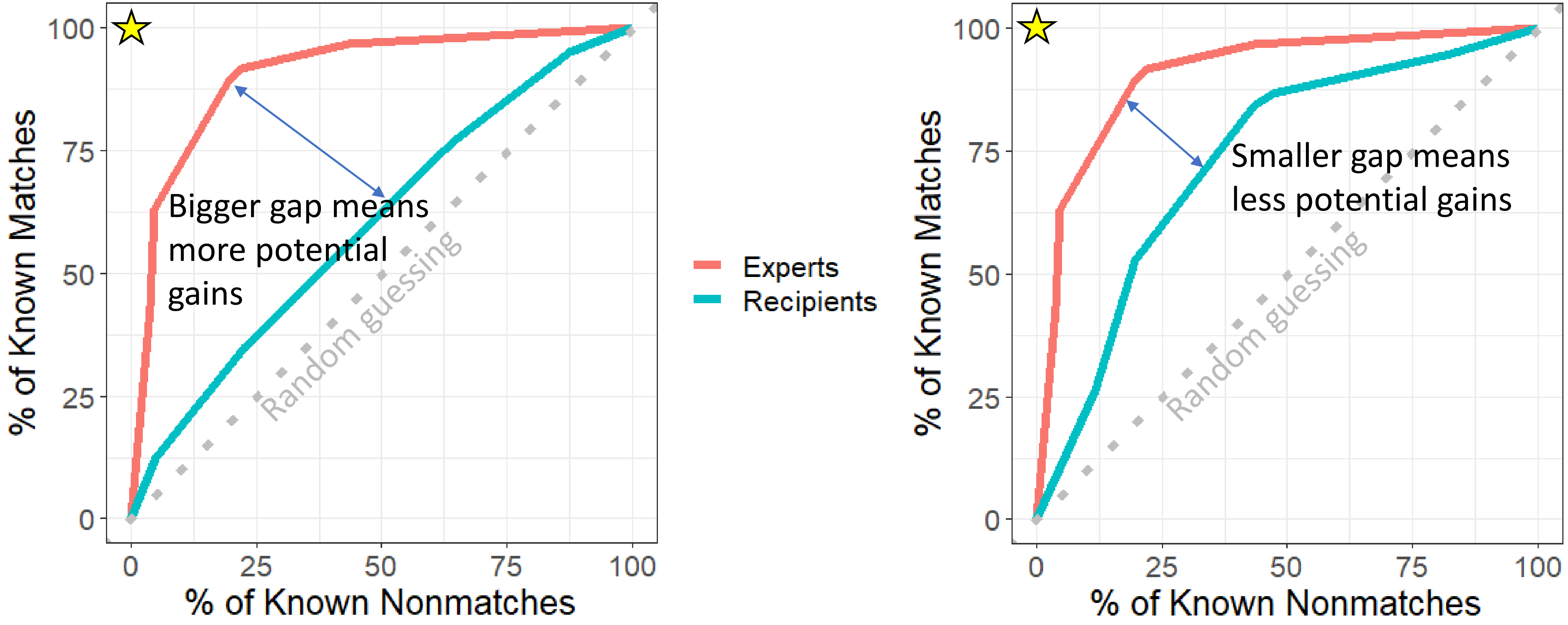
How to Measure/Grade CFF?

- What’s the greatest potential gain from the expert?
  - Among ground-truth-known tests, could compare experts’ ability to distinguish between propositions of interest with recipients’ to measure the gap



How to Measure/Grade CFF?

- What’s the greatest potential gain from the expert?
  - Among ground-truth-known tests, could compare experts’ ability to distinguish between propositions of interest with recipients’ to measure the gap
    - No gap would imply no meaningful information to communicate. Typically expect a gap

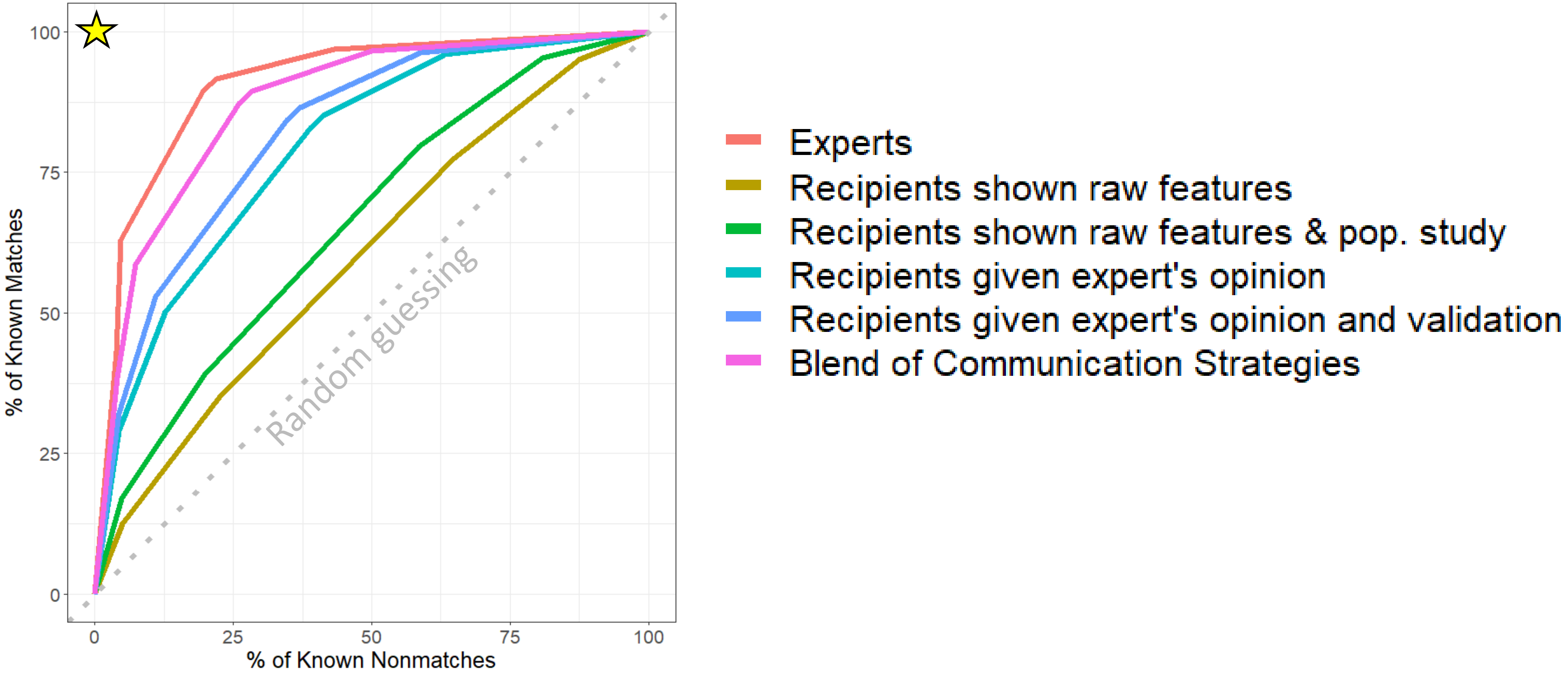


Bigger gap means more potential gains

Smaller gap means less potential gains

See how well CFF approaches close the gap

- Consider multiple approaches:
  - Presenting observations vs presenting opinions
  - Accompanying supporting information (e.g., population study summaries, theoretical explanations)
  - Attempts to educate decision makers vs. attempts to instruct the decision makers



How to Measure/Grade CFF?

- This approach may encourage suggesting recipients adopt expert’s sentiment as their own (since then they’d have the same discrimination power as the expert)
  - Ignores range of opinions / treats personal and subjective interpretation as communal fact
  - What to do with disagreements among equally competent experts?
  - What about uncharted territory?
- Blindly accepting an expert’s opinion opens a doorway for junk science or pushing boundaries too far (extrapolation)
- Focusing on validation data could help close the door
  - Recognize overconfidence or unsupported claims
- Reliable communication is critical, including validation details

Important Caveat

- Judicial outcomes relying on forensic science provide less observable feedback than real world outcomes relying on other applications of science. E.g.,
    - Building remains standing or collapses (e.g., Champlain Towers South)
    - Side effects of drug released for public consumption (e.g., Vioxx with ~30,000 adverse cardiac events)
    - Most forensic casework applications are like rockets disappearing immediately after launch
- More difficult to recognize real world successes and failures for forensic applications
    - Allowed overconfident performance conjecture unsupported by empirical testing (e.g., to the exclusion of all other sources, error-free method, etc.)
    - Prior to DNA, no obvious signs of trouble means these claims largely avoided scientific scrutiny
    - Following public errors and work of the Innocence Project, legal and scientific communities increase demand for empirical studies

Validation…

- … is even more important to assessing reliability of forensic methods than it is for most applied sciences
- … has a critical role in…
  - Labs deciding whether to use a method in a particular case
  - Recipients deciding how much weight to give a method’s result in a particular case
    - High-stake decisions made by peers rather than specialists
- … is an important component in CFF

So how do we talk about validation?

- “Validated”

- “Error rate”

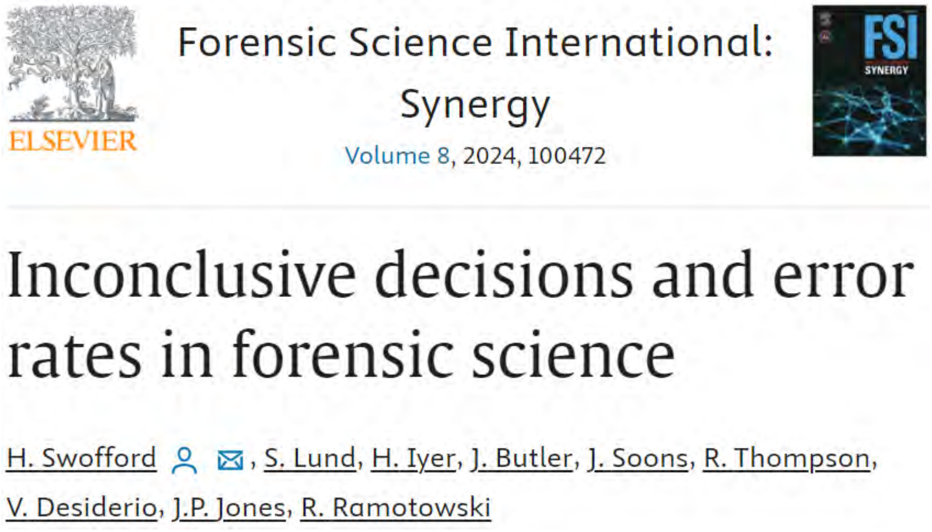
(Google Gemini result for “generate an image for the word unsatisfactory”)

“Validated”

- Falsely implies there’s a checklist that, once completed, renders uncertainty regarding method performance inconsequential
  - “How many samples do I need?”
  - Overlooks benefit to collecting additional validation data
- Suggests performance is one-size-fits-all
- Masks subjectivity of chosen validation criteria as consequence of statistics and science, making it harder to question

Error rates

- Biggest Positive: Brings attention to empirical performance studies
- Biggest Drawback: Requires oversimplifying to label each opinion/conclusion as either correct or incorrect
  - Most opinion/conclusion scales are on a more refined spectrum
  - Throws away relevant information
  - Leads to many proposals for handling inconclusive conclusions, some of which can be misleading



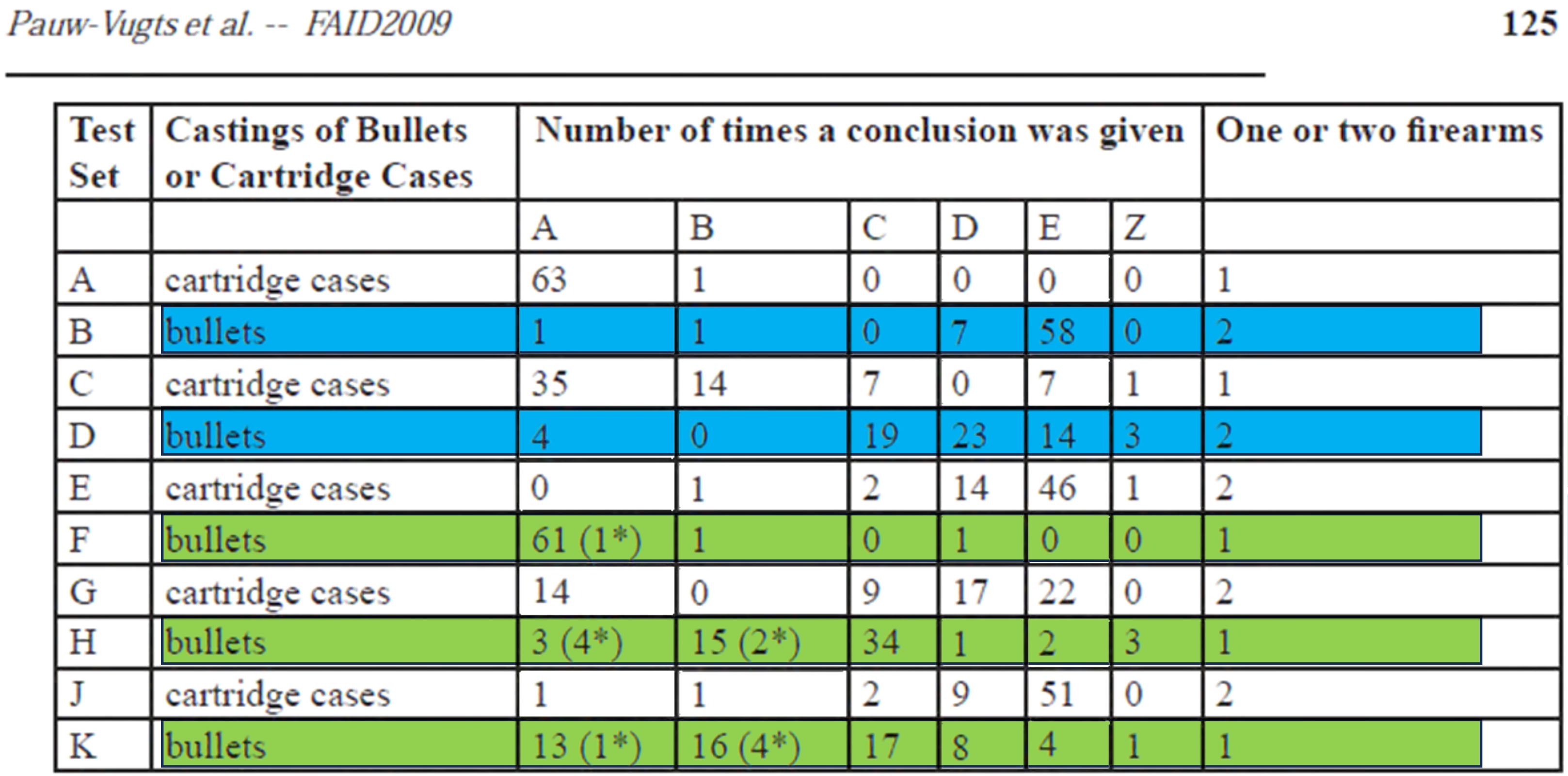
Forensic Science International: Synergy
Volume 8, 2024, 100472

Inconclusive decisions and error rates in forensic science

H. Swofford, S. Lund, H. Iyer, J. Butler, J. Soons, R. Thompson, V. Desiderio, J.P. Jones, R. Ramotowski

Example

Pauw-Vugts, P., Walters, A., Øren, L., & Pfoser, L. (2013). FAID2009: proficiency test and workshop. *AFTE Journal*, *45*(2).



Pauw-Vugts et al. -- FAID2009 **125**

Test Set	Castings of Bullets or Cartridge Cases	Number of times a conclusion was given						One or two firearms
		A	B	C	D	E	Z	
A	cartridge cases	63	1	0	0	0	0	1
B	bullets	1	1	0	7	58	0	2
C	cartridge cases	35	14	7	0	7	1	1
D	bullets	4	0	19	23	14	3	2
E	cartridge cases	0	1	2	14	46	1	2
F	bullets	61 (1*)	1	0	1	0	0	1
G	cartridge cases	14	0	9	17	22	0	2
H	bullets	3 (4*)	15 (2*)	34	1	2	3	1
J	cartridge cases	1	1	2	9	51	0	2
K	bullets	13 (1*)	16 (4*)	17	8	4	1	1

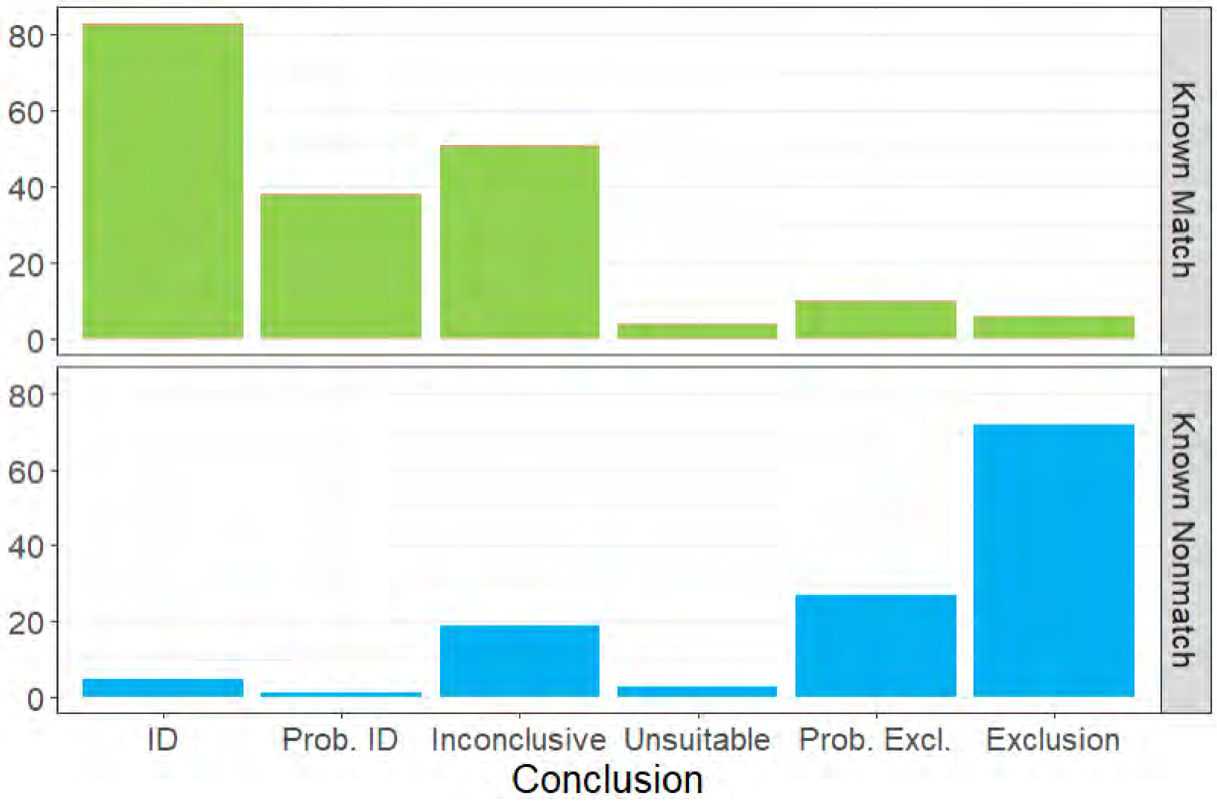
Known Matches

Known Nonmatches

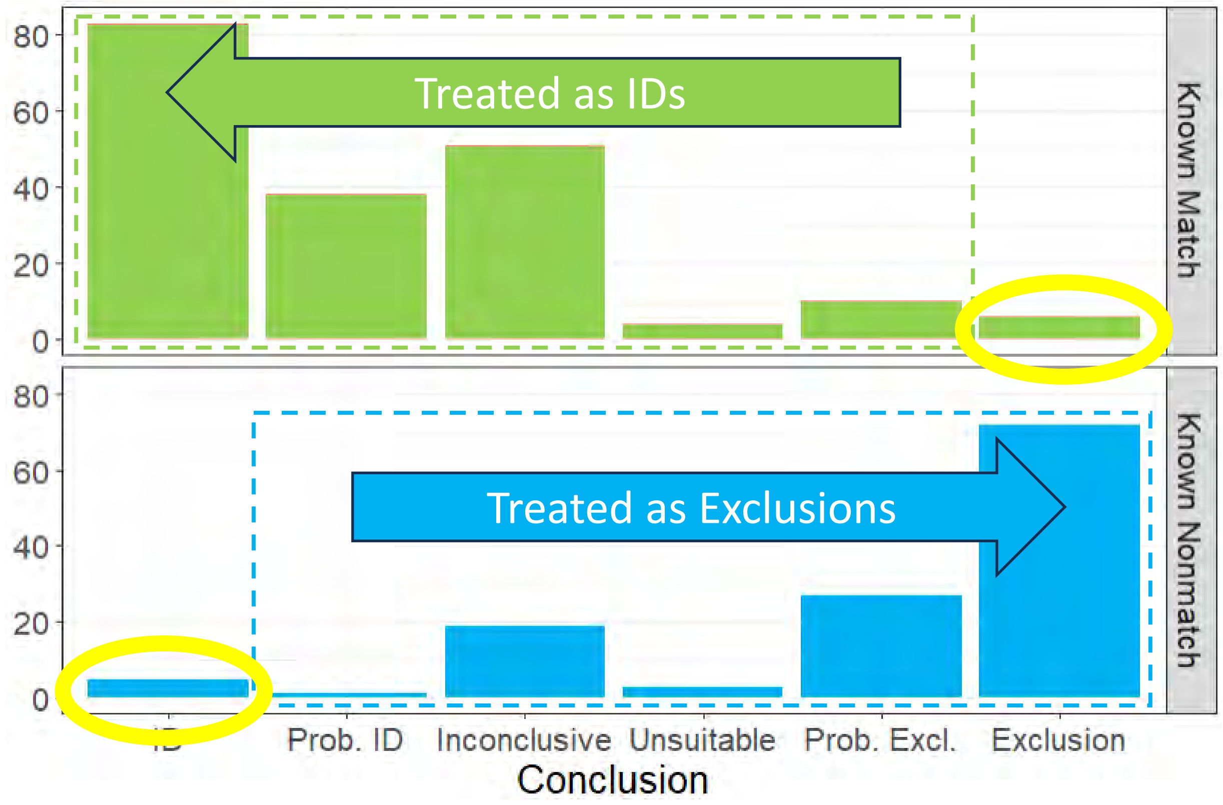
	Identification	Probable ID	Inconclusive	Probable Ex	Exclusion	Unsuitable
Known Matches	83	38	51	10	6	4
Known Nonmatches	5	1	19	27	72	3

Example

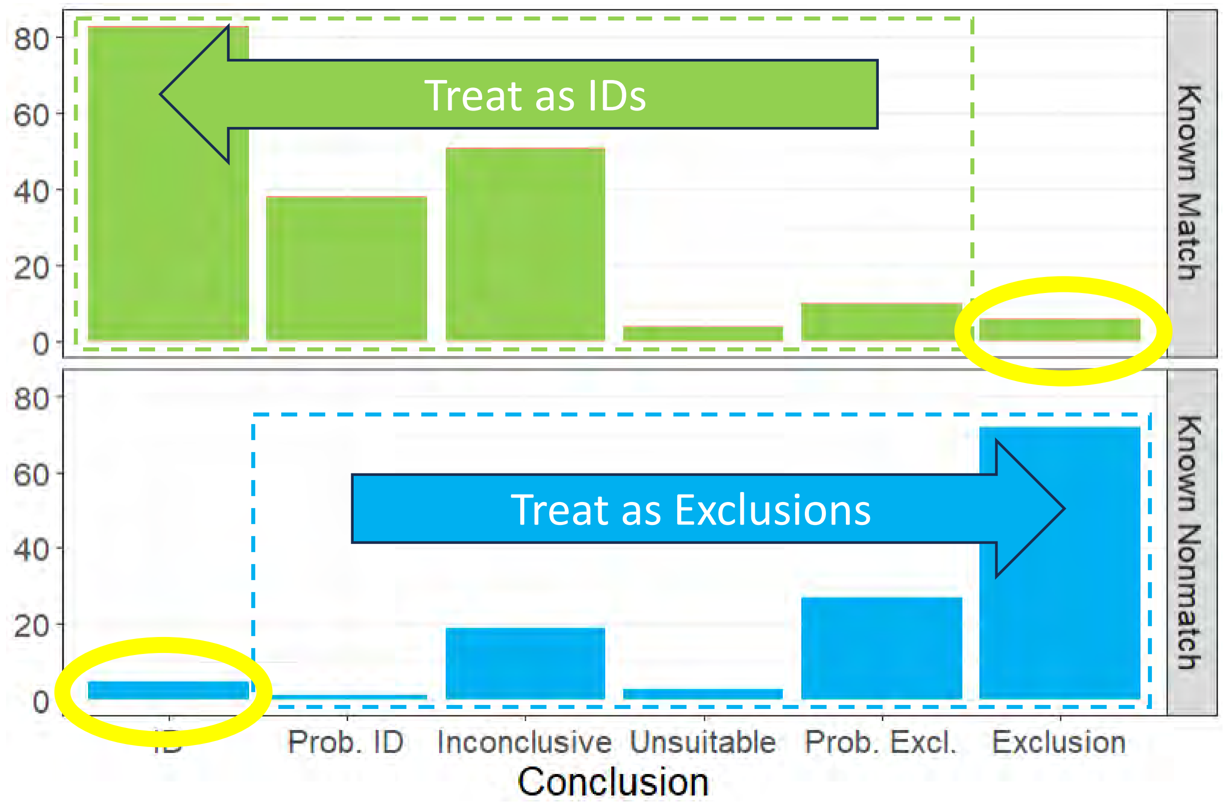
Pauw-Vugts, P., Walters, A., Øren, L., & Pfoser, L. (2013). FAID2009: proficiency test and workshop. *AFTE Journal*, *45*(2).



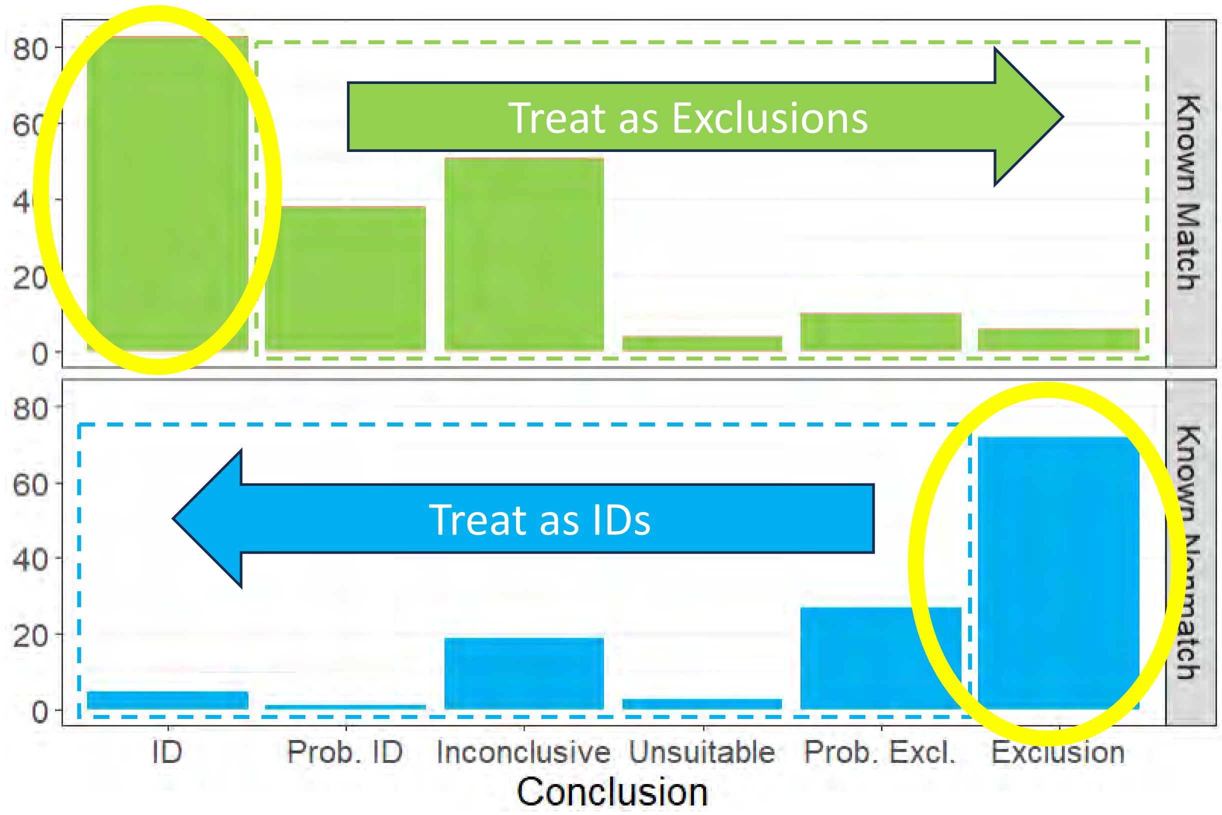
Quote Part 1: “Scientifically, an inconclusive result has to be automatically incorrect: a comparison is either from a same-source or a different-source. AFTE rules allow inconclusives to be counted as both identifications and eliminations, and therefore artificially decrease error rates.”



Quote Part 2: “If we focus on a correct source decisions only, the percentage of correct decisions can be as low as 49%, leaving at least 51% of the decisions as errors (correct source identification rate taken from bullet comparisons in Pauw-Vugts et al. (2013)).”

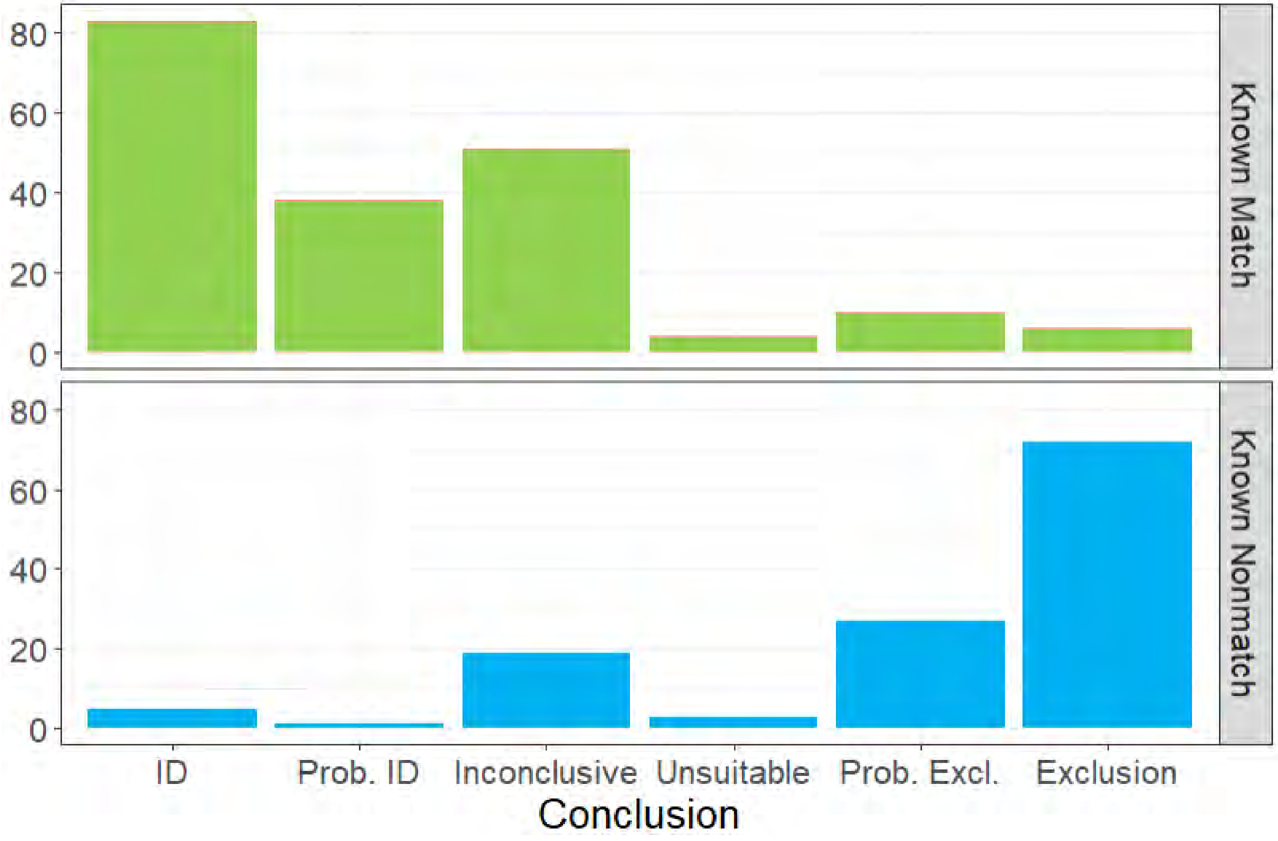


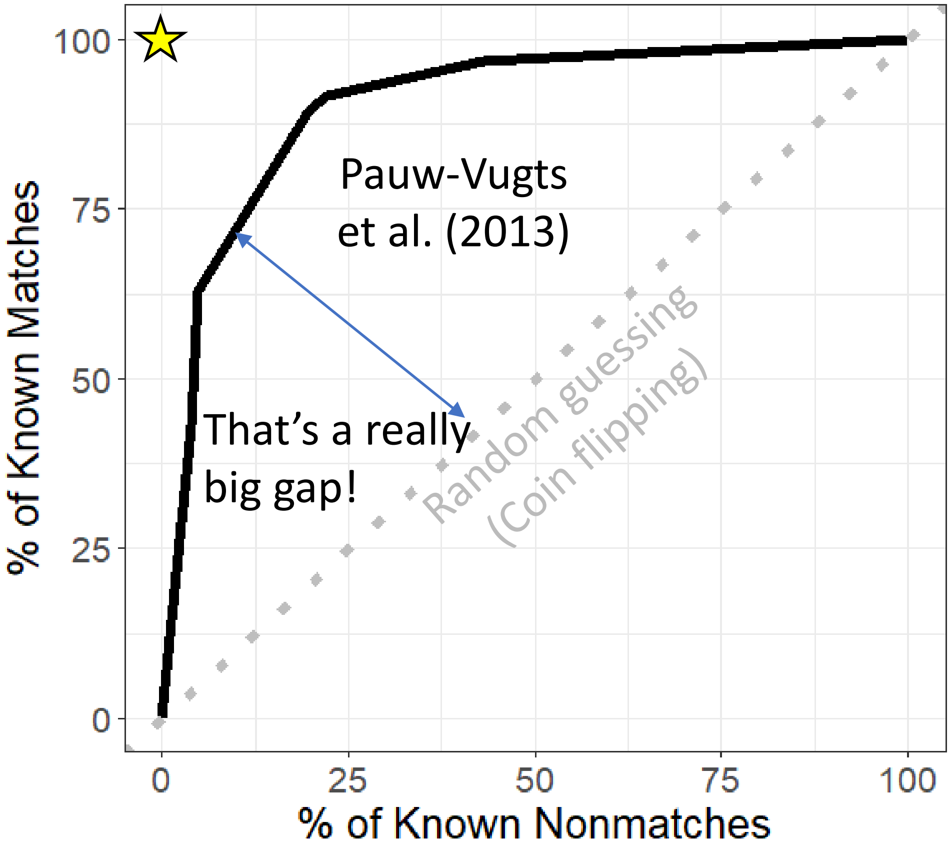
AFTE Treatment (Common)



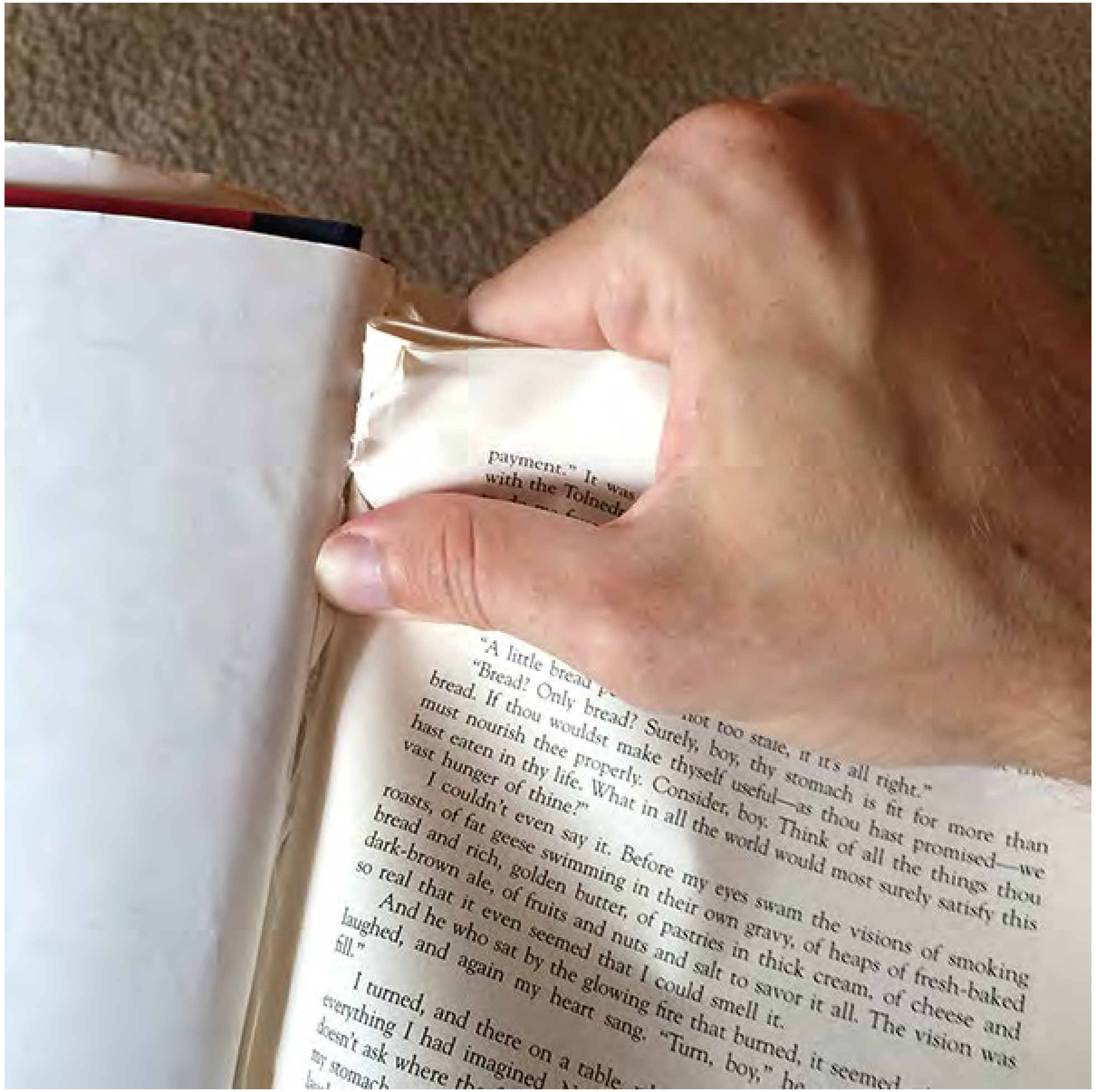
Suggestion from Authors (Statisticians)

Full quote: "Scientifically, an inconclusive result has to be automatically incorrect: a comparison is either from a same-source or a different-source. AFTE rules allow inconclusives to be counted as both identifications and eliminations, and therefore artificially decrease error rates. If we focus on a correct source decisions only, the percentage of correct decisions can be as low as 49%, leaving at least 51% of the decisions as errors (correct source identification rate taken from bullet comparisons in Pauw-Vugts et al. (2013)). **This is statistically worse than random chance - that is, examiners would perform about as well if they were flipping a coin to make the decision!**"





That's a really big gap!



payment." It was
with the Tolned

"A little bread
"Bread? Only bread? not too stale, if it's all right."
bread. If thou wouldst make Surely, boy, thy stomach is fit for more than
must nourish thee properly. thyself useful—as thou hast promised—we
hast eaten in thy life. What Consider, boy. Think of all the things thou
vast hunger of thine?" in all the world would most surely satisfy this

I couldn't even say it. Before my eyes swam the visions of smoking roasts, of fat geese swimming in their own gravy, of heaps of fresh-baked bread and rich, golden butter, of pastries in thick cream, of cheese and dark-brown ale, of fruits and nuts and salt to savor it all. The vision was so real that it even seemed that I could smell it.

And he who sat by the glowing fire that burned, it seemed laughed, and again my heart sang. "Turn, boy," he fill."

I turned, and there on a table
everything I had imagined
doesn't ask where
my stomach

Credit: https://craftbits.com/project/diy-collage-of-pages-bookcase/

Validation Nuances

- Attempt to assign weight to an opinion in a particular case
- Efficacy expected to vary across case types
    - E.g., expect mostly IDs and Exclusions when comparing two exemplars, expect mostly inconclusives for very low-quality questioned impressions
- Some factors describing case type may allow us to predict changes in examiner performance
    - What are these factors? What are their effects?

- Available data is not ideal
    - Fewer tests than we'd like (cost-benefit analysis)
    - Few, if any, tests match circumstances of current case (e.g., different quality sample(s), different lab or expert, awareness of being tested, etc.)
    - Departures from ideal statistical sampling approaches: volunteer participants, convenience sample materials, not all tests are answered
    - Important details that changes or adds uncertainty to the meaning of the data
- Despite limitations, available data can be (are) informative
    - E.g., demonstrate that some experts perform well in some scenarios (i.e., not coin-tossers)
    - How informative will depend on subjective reactions to limitations

- How to reasonably summarize or present available validation information?

Key Points

- Validation testing remains the primary means by which society can understand the efficacy of forensic science methods (more so than many other areas of science)

- Forensic science relies more on general population (e.g., fact finders) to carry out its mission than do other scientific applications
  - Don’t take 12 random people to approve space shuttle launch or decide whether open heart surgery will be performed

- We need to improve how we communicate about validation
  - “Error rates” and “validated” oversimplify in potentially misleading ways

- Looking forward to hearing your thoughts and perspectives on these, and other, important CFF topics!