Fingerprint Examination Workshop: Reliability of the ACE-V Process

Glenn Langenburg

A Framework for Hypothesis Testing

- H_0 = Individual in question left mark
- H_1 = Individual in question did not leave mark

- **A**nalysis: Feature selection, value (utility) assessment, assessment of distortion
- **C**omparison: Assessment of corresponding and discordant features (generates LR)

A Framework for Hypothesis Testing

- H_0 = Individual in question left mark
- H_1 = Individual in question did not leave mark

- **E**valuation: a decision is made
  - Identification (H_0)
  - Exclusion (H_1)
  - Inconclusive (neither H is chosen)

- **V**erification: a quality assurance step

Error Rates

- Koehler, Thompson, Taroni, etc.

- Koehler, J. Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter. Hastings Law Journal 59 (5), 2008, 1077-1100.

From Koehler (2008)

	Ground Truth of Latent Print	
Examiner Decision	**Same Source**	**Different Source**
Identification	Correct ID	
Exclusion		Correct Exclusion

From Koehler (2008)

Examiner Decision	Ground Truth of Latent Print: Same Source	Ground Truth of Latent Print: Different Source
Identification	Correct ID	Erroneous ID
Exclusion	Erroneous Exclusion	Correct Exclusion

From Koehler (2008)

	Ground Truth of Latent Print	
Examiner Decision	**Same Source**	**Different Source**
Identification	Correct ID	False +
Exclusion	False -	Correct Exclusion

From Koehler (2008)

	Ground Truth of Latent Print	
Examiner Decision	**Same Source**	**Different Source**
Identification	Sensitivity	False +
Exclusion	False -	Selectivity

From Koehler (2008)

- False + Rate = $\frac{B}{B + D}$
- False – Rate = $\frac{C}{A + C}$

::

Examiner Decision	Ground Truth of Latent Print: Same Source	Ground Truth of Latent Print: Different Source
Identification	Correct ID **A**	False + **B**
Exclusion	False - **C**	Correct Exclusion **D**

Method Performance Error Rates (Langenburg, 2009)

	Ground Truth	
Reported Conclusion	Same Source	Different Source
Identified	268 (True Positive)	1 (False Positive)
Excluded	3 (False Negative)	880 (True Negative)

- False Positive Rate: 0.1% ACE condition
- False Negative Rate: 1.1% ACE condition

Method Performance Error Rates (Langenburg, 2009)

	Ground Truth	
Reported Conclusion	Same Source	Different Source
Identified	271 (True Positive)	0 (False Positive)
Excluded	6 (False Negative)	960 (True Negative)

- False Positive Rate: < 0.1% ACE-V condition
- False Negative Rate: 2.2% ACE-V condition

Other Studies

- Gutowski (2007): used CTS data for Aussie experts; 0 false negatives, 2 false positives, 782 decisions
- Wertheim, Langenburg, Moenssens (2006): 2 false positives, ~6000 decisions
- Most recently: “Informing Judgments Study”

and FBI “Black Box Study”

Study Results-All Groups

Examiner Decision	Ground Truth of Latent Print: Same Source	Ground Truth of Latent Print: Different Source	Totals
Identification	840	23	863
Inconclusive	322	92	414
Exclusion	70	765	835
Totals	1232	880	2112

	Ground Truth of Latent Print		
Examiner Decision	Same Source	Different Source	Totals
Identification	840	23	863
Inconclusive	322	92	414
Exclusion	70	765	835
Totals	1232	880	2112

- False Positive Rates:
- 23/880 = 2.6% (keep INC decision in totals)
- 23/788 = 2.9% (do not count INC decisions)
- 115/880 = 13% (count INC decision as error)

	Ground Truth of Latent Print		
Examiner Decision	Same Source	Different Source	Totals
Identification	840	23	863
Inconclusive	322	92	414
Exclusion	70	765	835
Totals	1232	880	2112

- False Negative Rates:
- 70/1232 = 5.7% (keep INC decision in totals)
- 70/910 = 7.7% (do not count INC decisions)
- 392/1232 = 32% (count INC decision as error)

What is a False Positive Rate?

- Pr [“ID” | Not Source]
  - Therefore we must debate how to handle “Inconclusive” decisions
- Per Koehler, Thompson, etc. suggest “False Discovery Rates” (1 – “Predictive Rate”)
- Pr [Not Source | “ID”]

	Ground Truth of Latent Print		
Examiner Decision	Same Source	Different Source	Totals
Identification	840	23	863
Inconclusive	322	92	414
Exclusion	70	765	835
Totals	1232	880	2112

- False Positive Discovery Rate:
  23/863 = 2.7% Pr [Not Source | “ID”]

- False Negative Discovery Rate:
  70/835 = 8.4% Pr [Source | “Exc”]

Reproducibility

- When the same sample is given to different instruments, how consistent are the results?
- When the same fingerprint comparison is given to different analysts to work independently, how consistent are the results?

Let’s Look at the Data

- Published and Peer Reviewed:
  - Osterburg (1964)
  - Evett and Williams (1995)

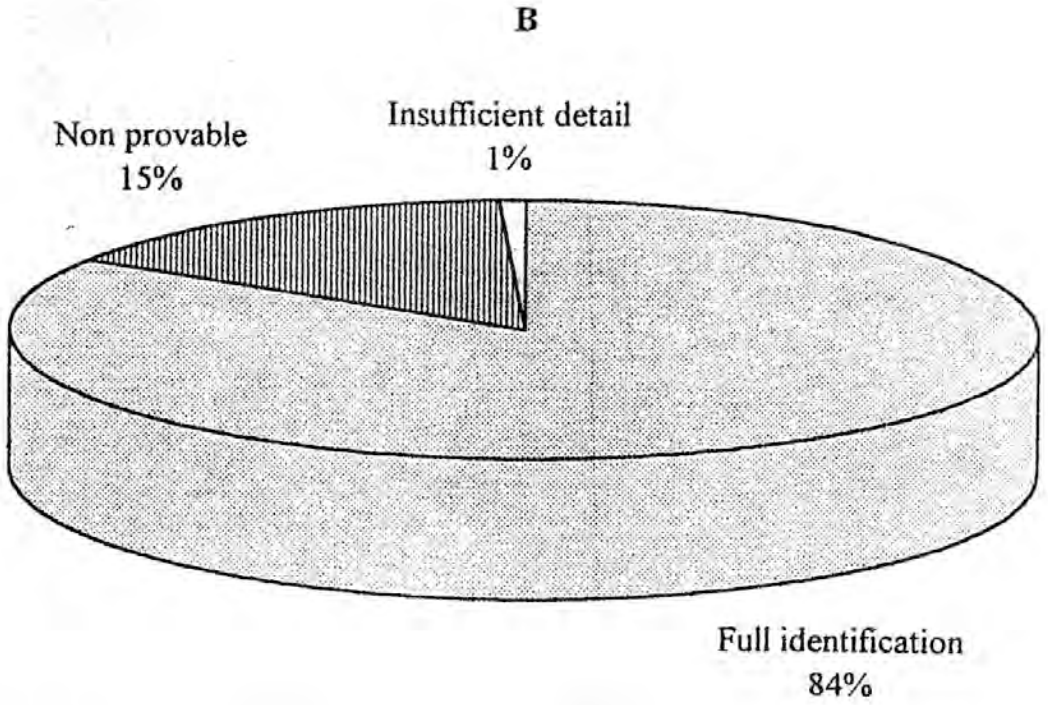
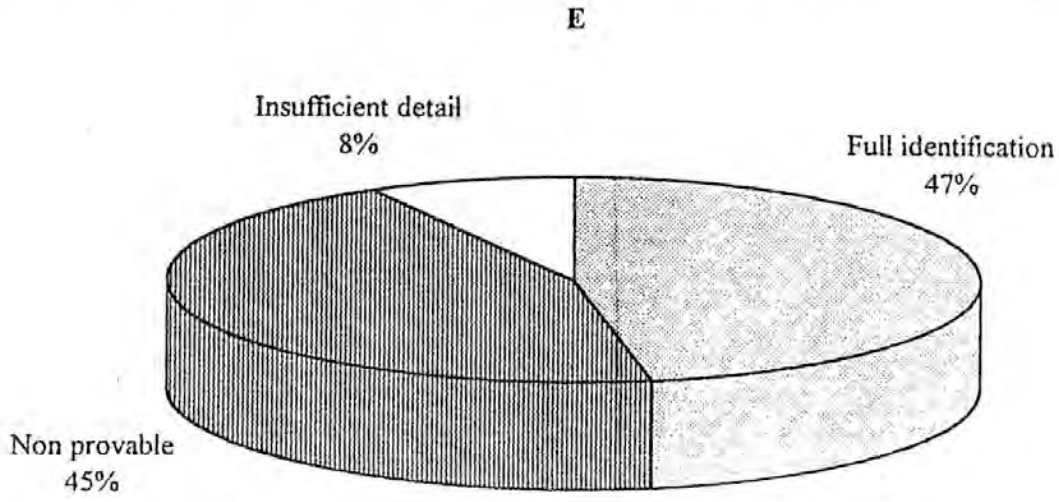

B
Non provable
15%
Insufficient detail
1%
Full identification
84%

FIGURE 5.2


E
Insufficient detail
8%
Full identification
47%
Non provable
45%

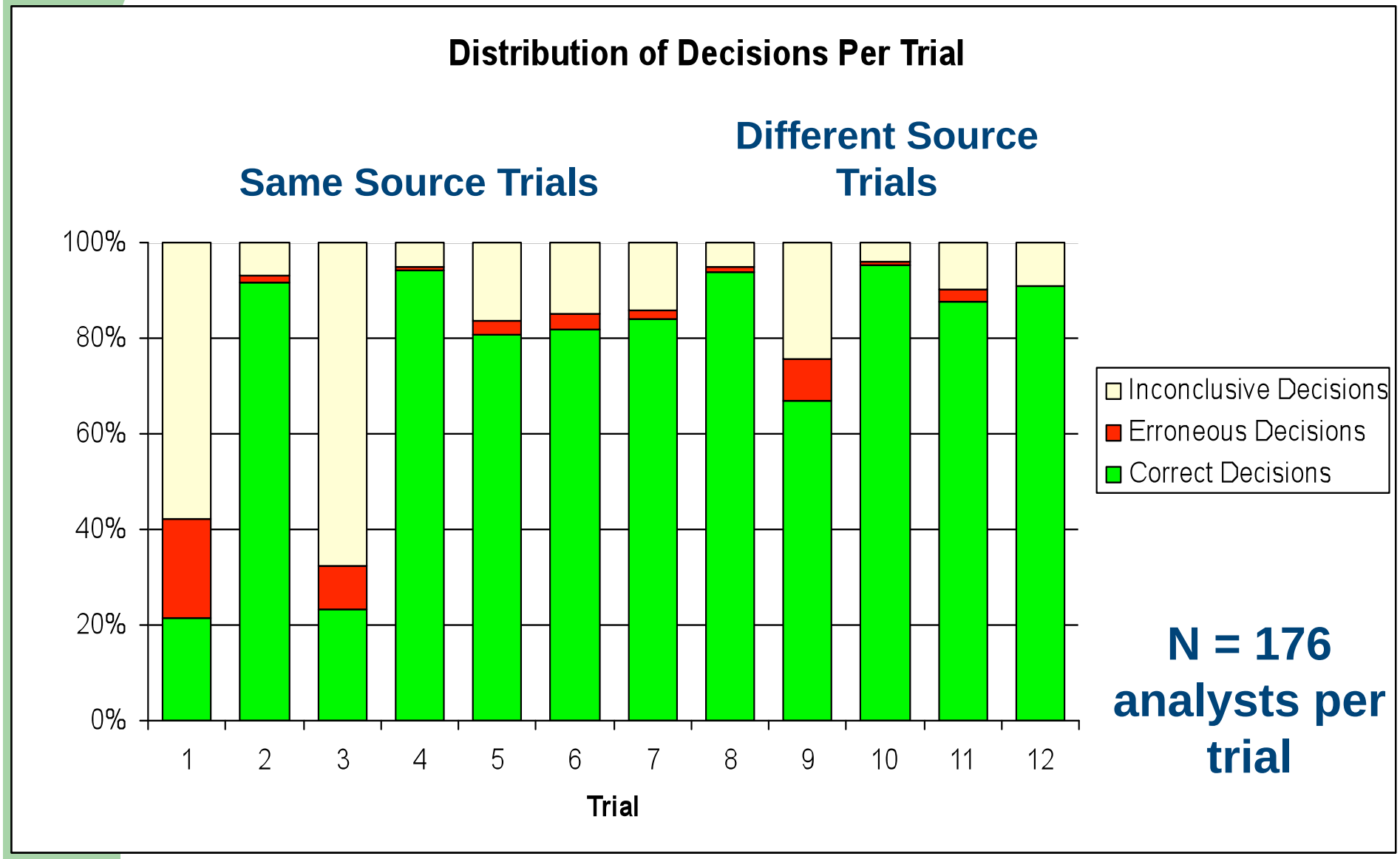
Let’s Look at the Data

- Published and Peer Reviewed:
  - Langenburg (2009)
    - ACE: 85%
    - ACE-V: 94%
  - “I can live with it”
  - 98% were sufficiency differences

Recent Research

- FBI's Black Box study
- Langenburg/Champod "Informing Expert Judgments Study" (2010)
  - Summary report of data available

Informing Judgments (2010)



Repeatability

- When the same sample is given to the same instrument on different occasions, how consistent are the results at each testing time.
- When the same fingerprint comparison is given to the same analyst (assuming the analyst does not recall the earlier trial), how consistent are the results?

Repeatability

- Approx. 95% in trials where the comparisons were moderate to easy.
- Approx. 50% in trials where the comparisons were difficult.
- For additional information, see Langenburg. “A Performance Study of the ACE-V Process...”, JFI 59 (2), 219-257.

Repeatability, continued

- Eight inconclusive trials
- Re-presented: 50% now gave definitive opinion

	Trial	Initial Conclusion	Conclusion After Re-presentation	Were the Conclusions Consistent
Inconclusive Conclusions	A-021	Inconclusive-NBK-ONE	Inconclusive	Yes
	A-025	Inconclusive-NBK-ONE	Identification	No
	A-030	Inconclusive-NBK-ONE	Inconclusive	Yes
	A-053	Inconclusive-NBK-ONE	Identification	No
	B-009	Inconclusive-NBK-ONE	Inconclusive	Yes
	B-022	Inconclusive-NBK-ONE	Exclusion	No
	C-016	Inconclusive-NBK-ONE	Inconclusive	Yes
	C-019	Inconclusive-NBK-ONE	Identification	No
	A-005	Inconclusive-NBK-ALL	Identification	
	A-007	Inconclusive-NBK-ALL	Identification	

Dror, Charlton, Peron (2006)
"Why Experts Make Errors"

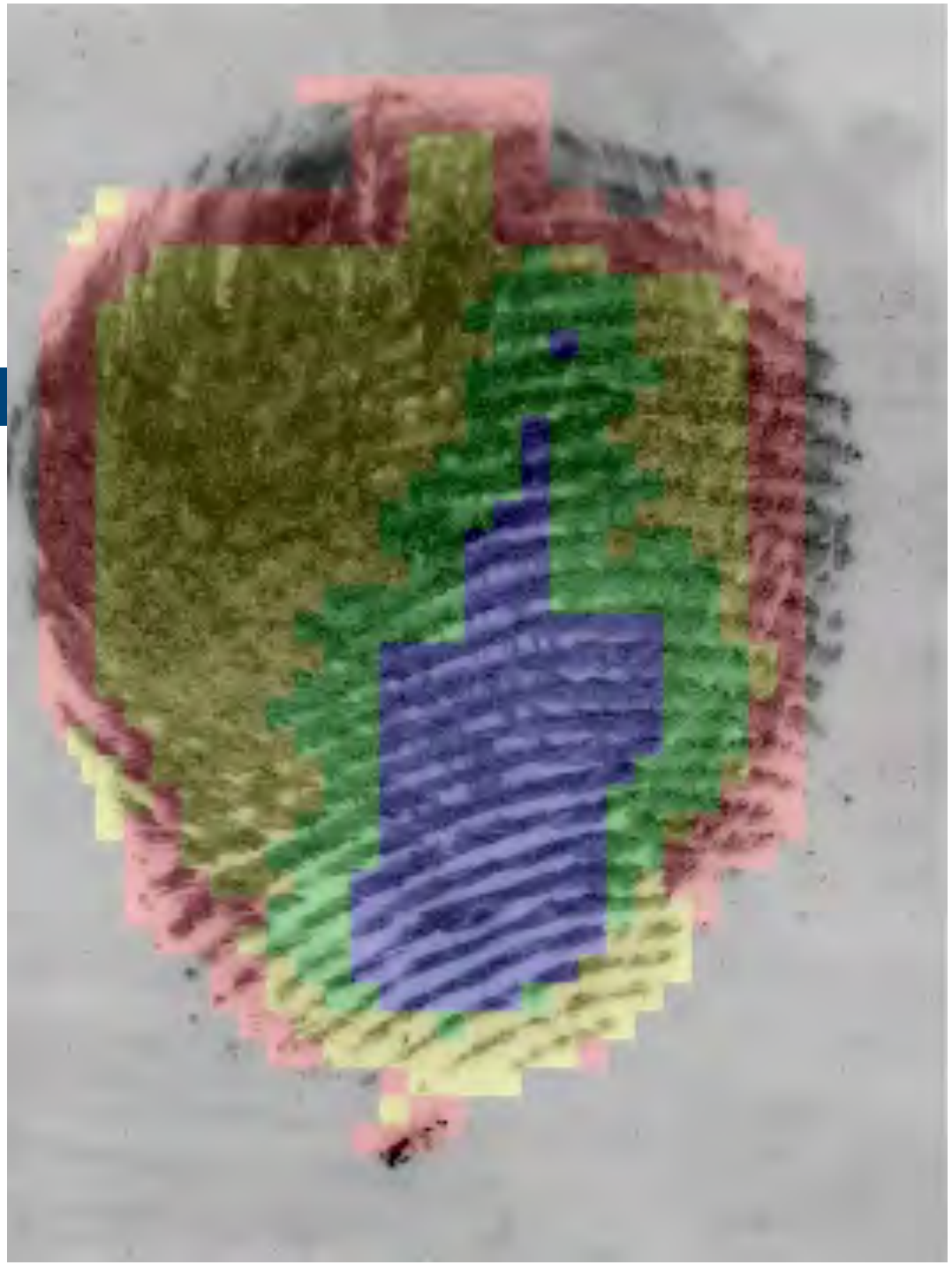
	1	2	3	4	5	6	7	8
Past Decision	individualization	individualization	individualization	individualization	exclusion	exclusion	exclusion	exclusion
Level of Difficulty	difficult	difficult	not difficult	not difficult	difficult	difficult	not difficult	not difficult
Contextual Information	none	suggest exclusion	none	suggest exclusion	none	suggest individualization	none	suggest individualization
Expert A	consistent	consistent	consistent	consistent	consistent	consistent	consistent	consistent
Expert B	change to exclusion	consistent	consistent	consistent	consistent	consistent	consistent	consistent
Expert C	consistent	change to exclusion	consistent	consistent	consistent	consistent	consistent	consistent
Expert D	consistent	change to exclusion	consistent	change to exclusion	change to individualization	consistent	consistent	consistent
Expert E	consistent	change to cannot decide	consistent	consistent	consistent	consistent	consistent	consistent
Expert F	consistent	consistent	consistent	consistent	consistent	consistent	consistent	consistent

Reliability

- We can make measurements that are indicia of reliability:
  - Error rates, false discovery rates
  - Reproducibility
  - Repeatability
- We can continue to improve the process with more measurements:
  - Measurement of quality
  - LR tools

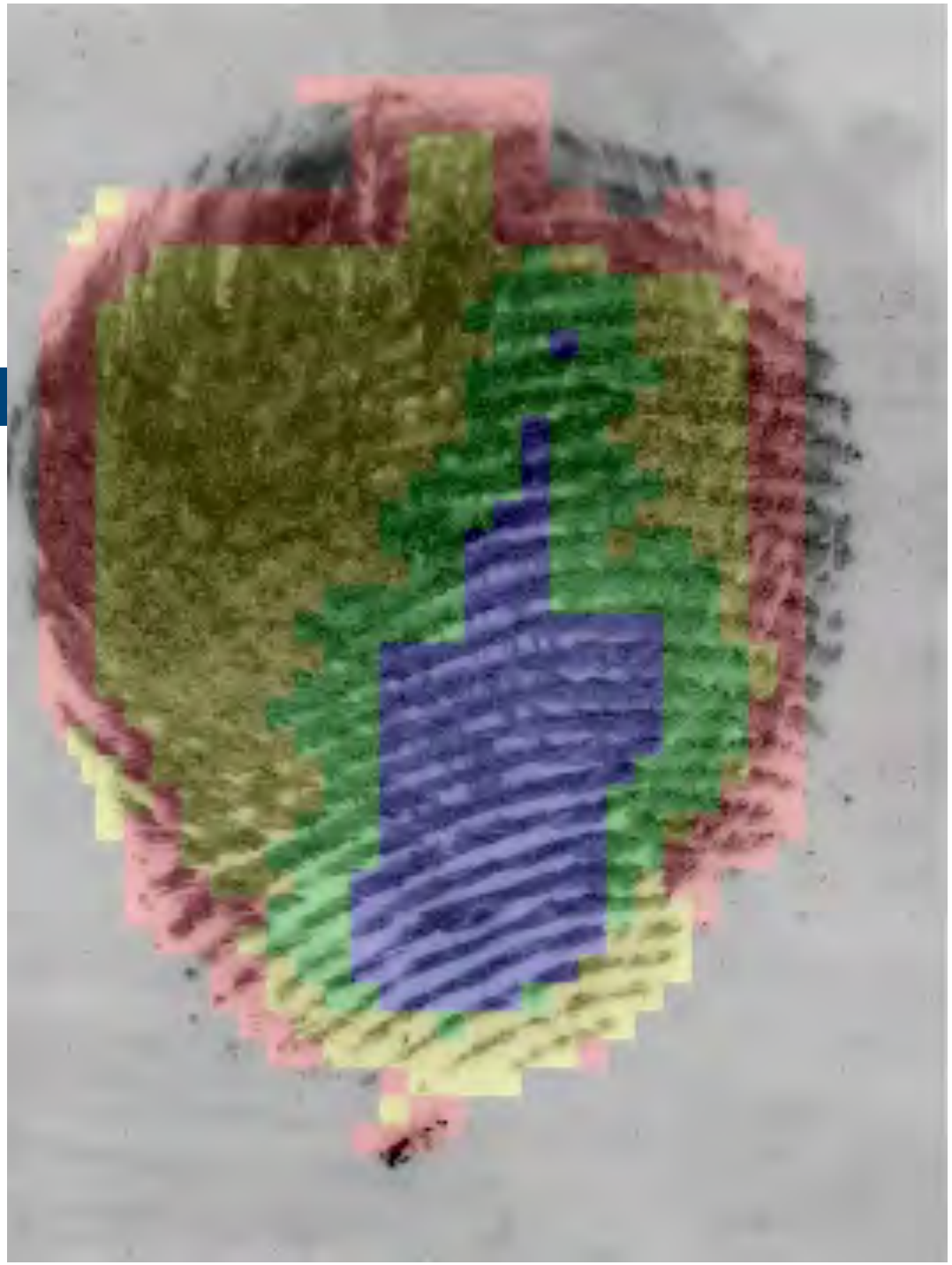
Quality Tools

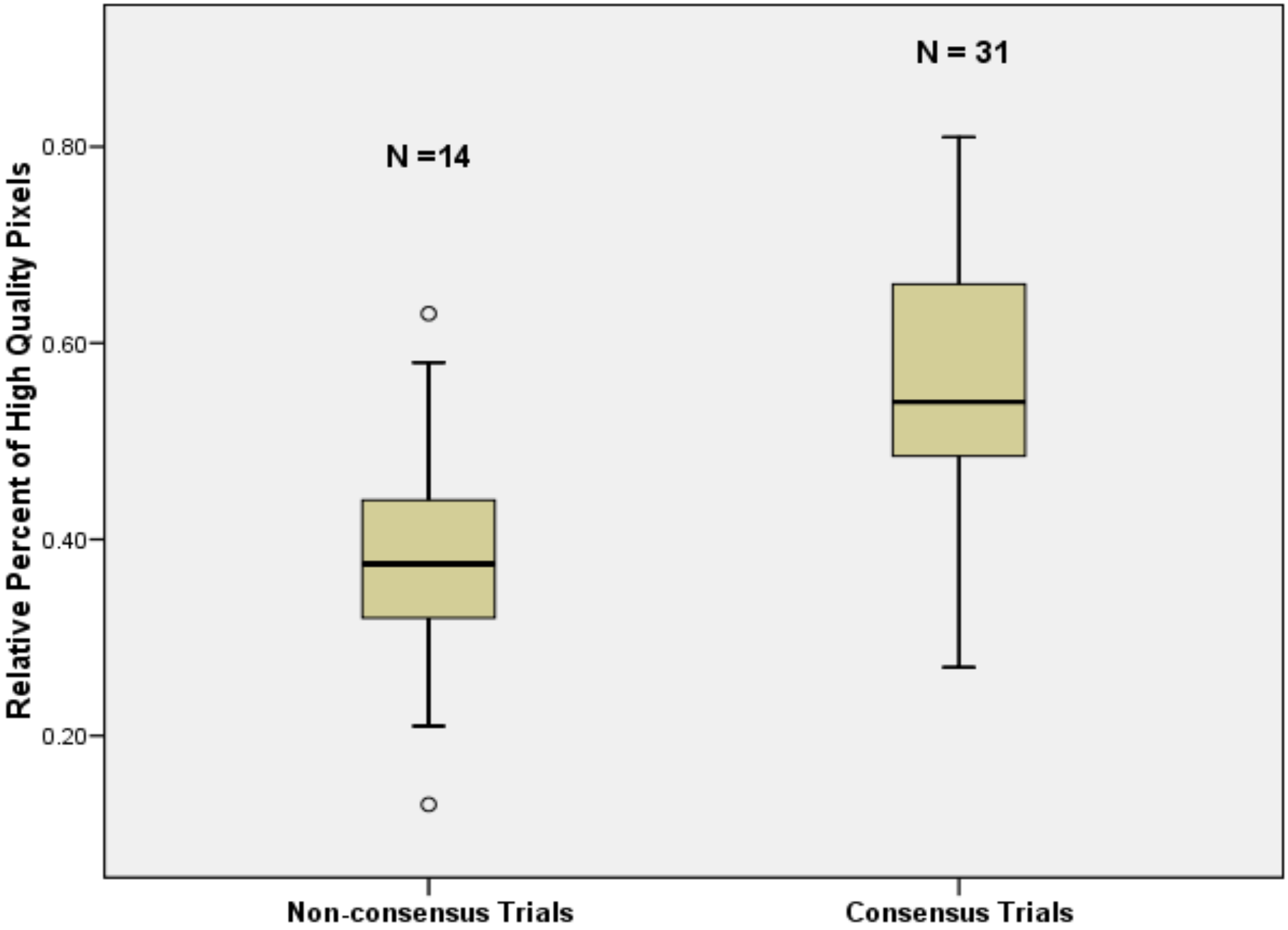
- Noblis, Inc. and FBI, ULW beta
- G & B high quality areas
- Y medium quality
- R low quality

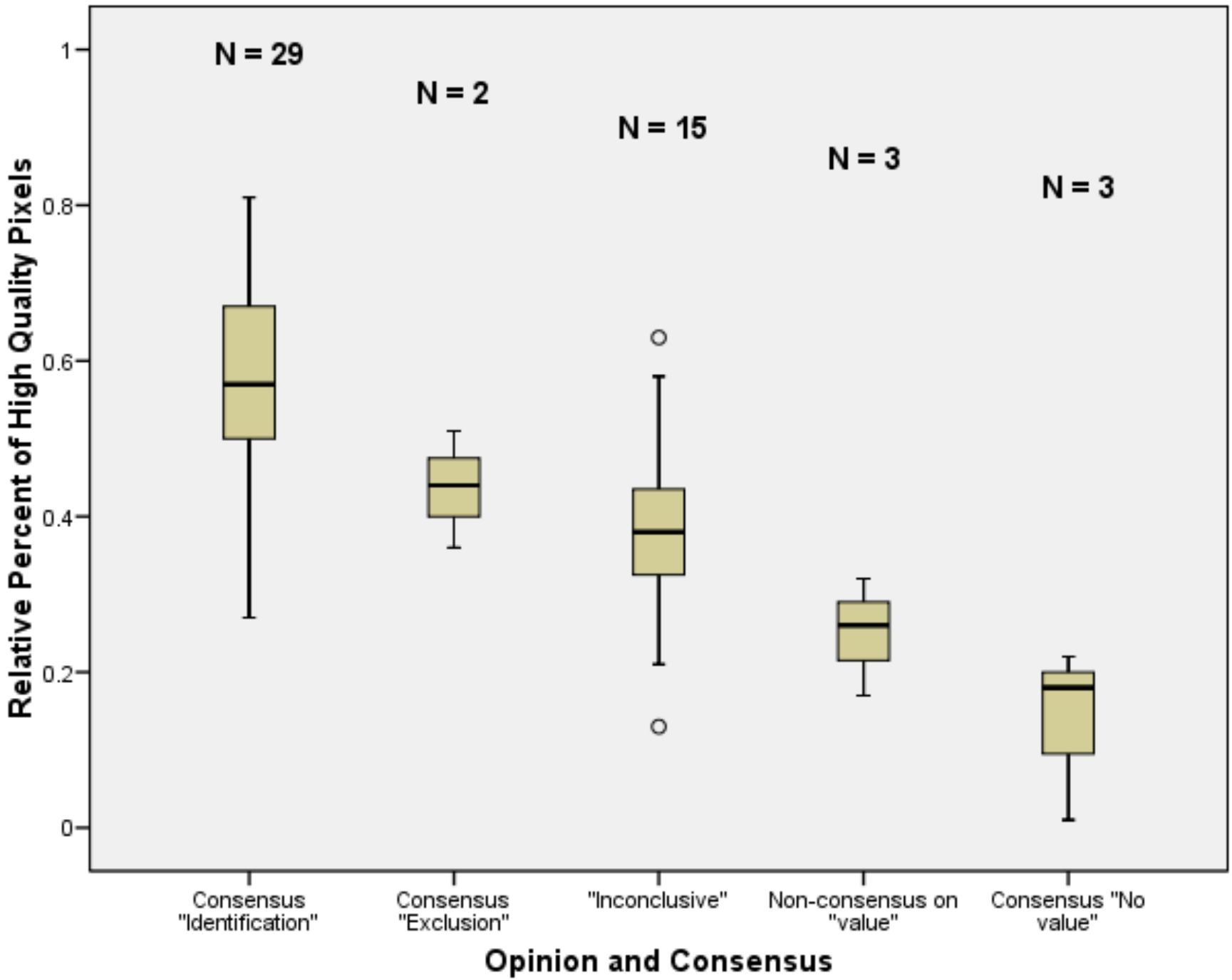


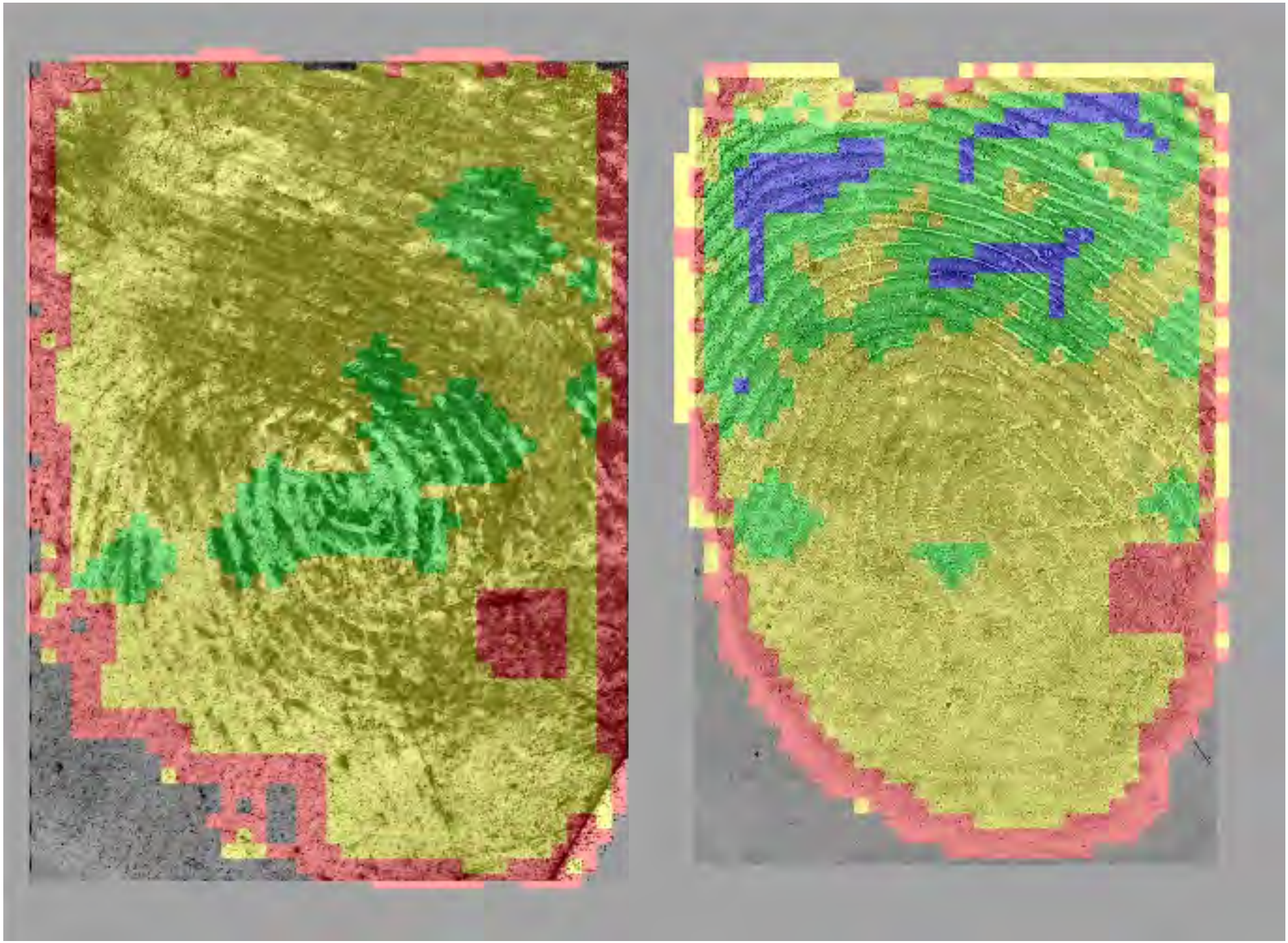
Quality Tools

- Ratio of high quality pixels to low/med quality
- 36%
  - 3 “ID”
  - 2 “Inc”
  - 1 “No value”

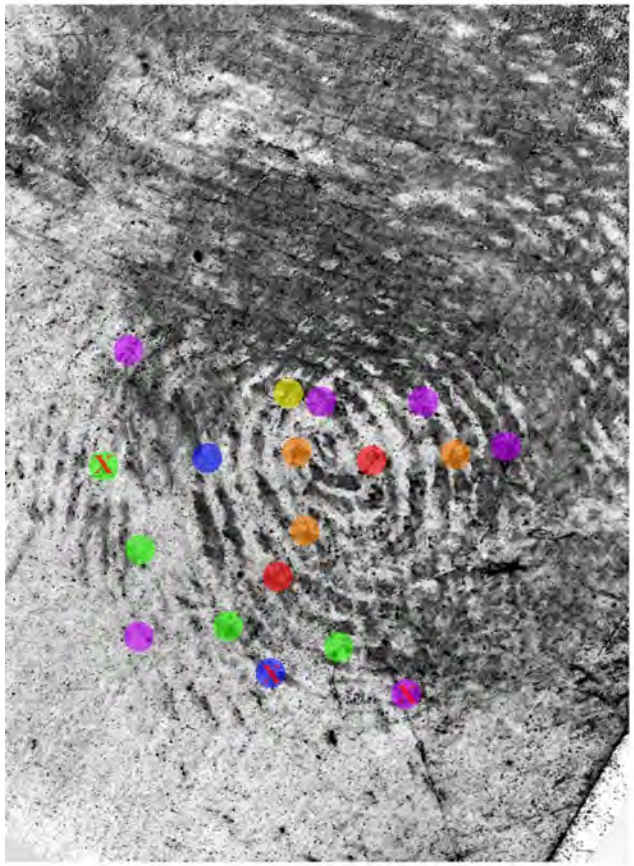



N = 31
N =14
0.80
0.60
0.40
0.20
Relative Percent of High Quality Pixels
Non-consensus Trials
Consensus Trials

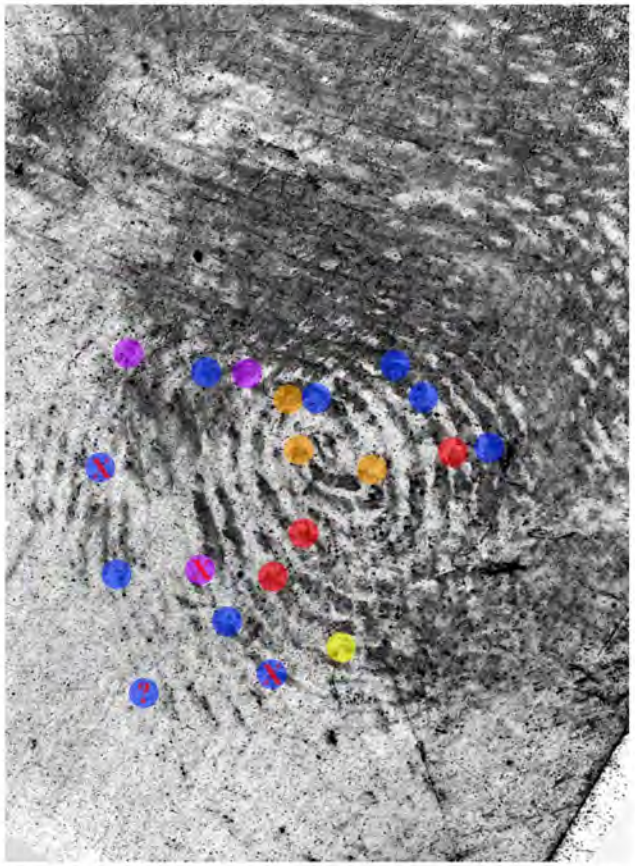

N = 29
N = 2
N = 15
N = 3
N = 3
1
0.8
0.6
0.4
0.2
0
Relative Percent of High Quality Pixels
Consensus "Identification"
Consensus "Exclusion"
"Inconclusive"
Non-consensus on "value"
Consensus "No value"
Opinion and Consensus



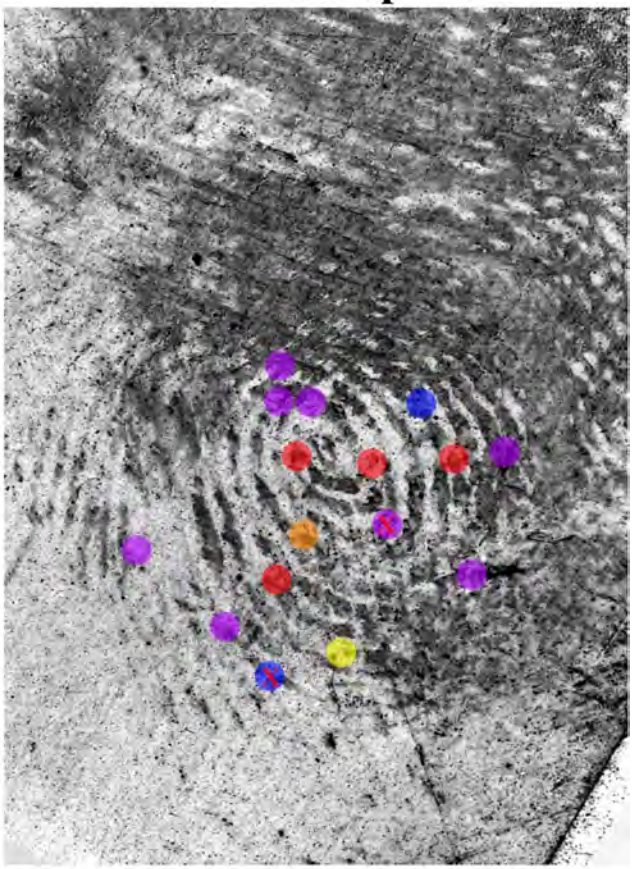
U.S. Mixed Class



U.S. Single Agency Class



Dutch Experts

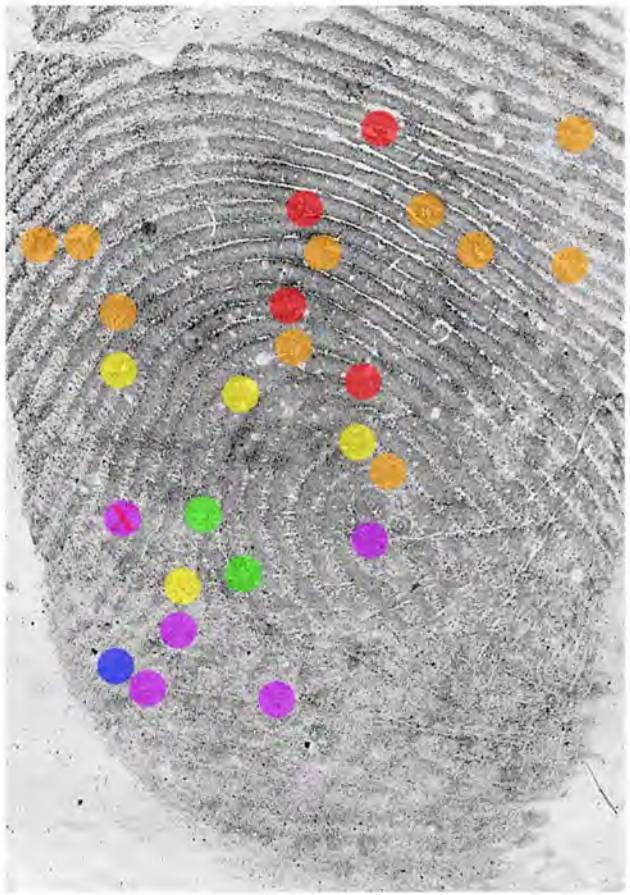


Legend:

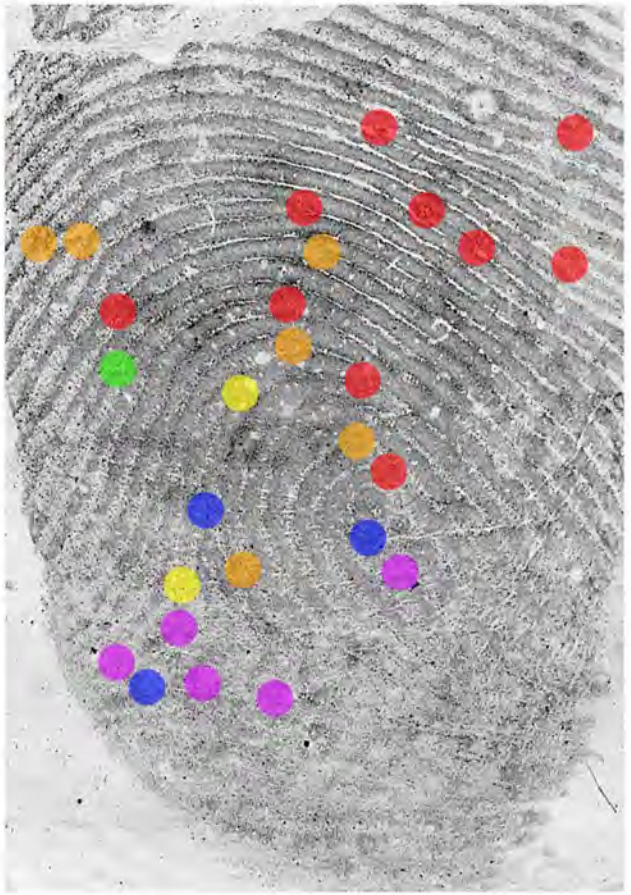
Red = 95% or higher
Orange = 80% - 94%
Yellow = 65% - 79%
Green = 50% - 64%
Blue = 35% - 49%
Violet = 20% - 34%

Less than 20% not indicated; an "X" indicates a 'false' minutia; a "?" indicates an 'indeterminate' minutia

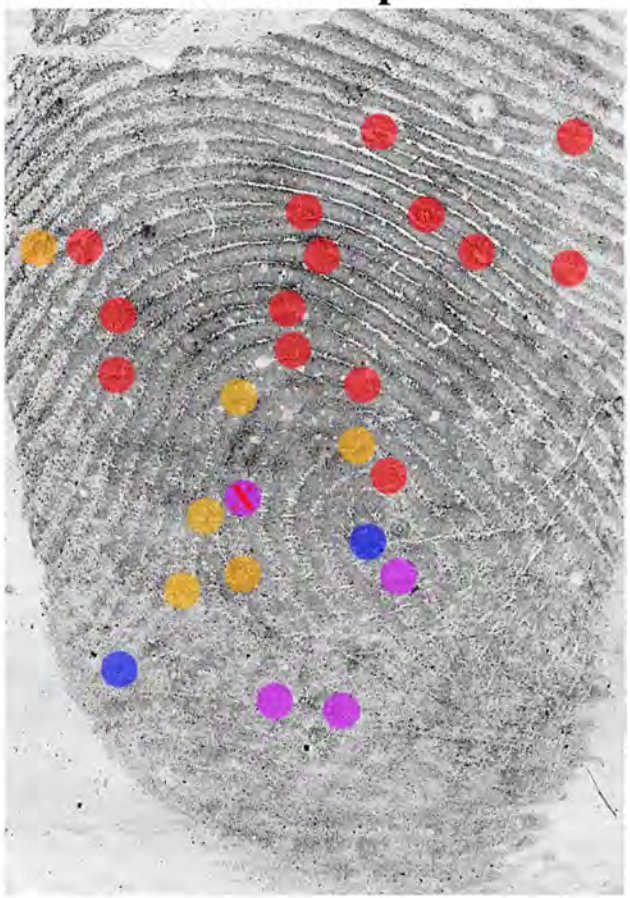
U.S. Mixed Class



U.S. Single Agency Class



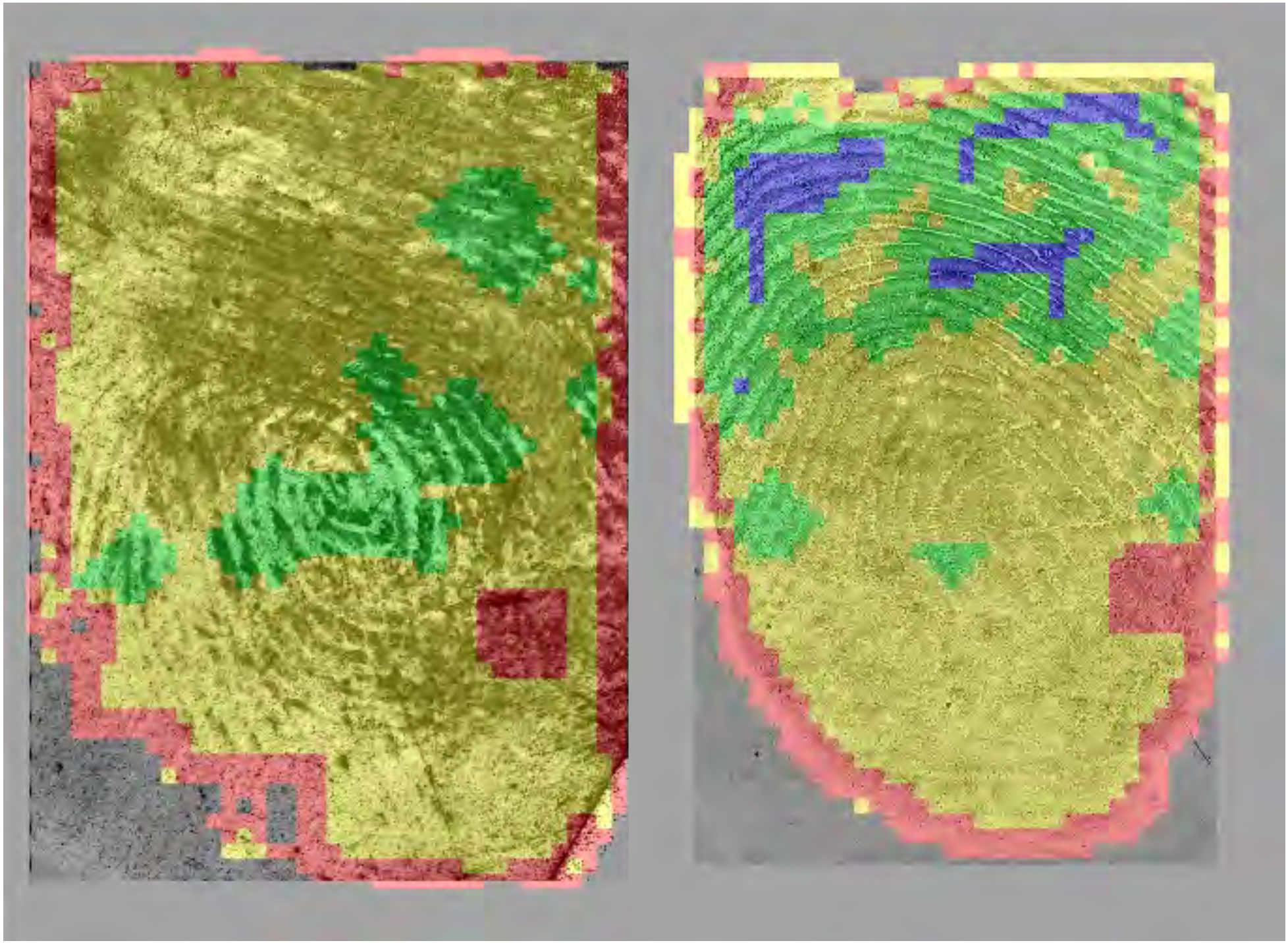
Dutch Experts



Legend:

Red = 95% or higher	Green = 50% - 64%
Orange = 80% - 94%	Blue = 35% - 49%
Yellow = 65% - 79%	Violet = 20% - 34%

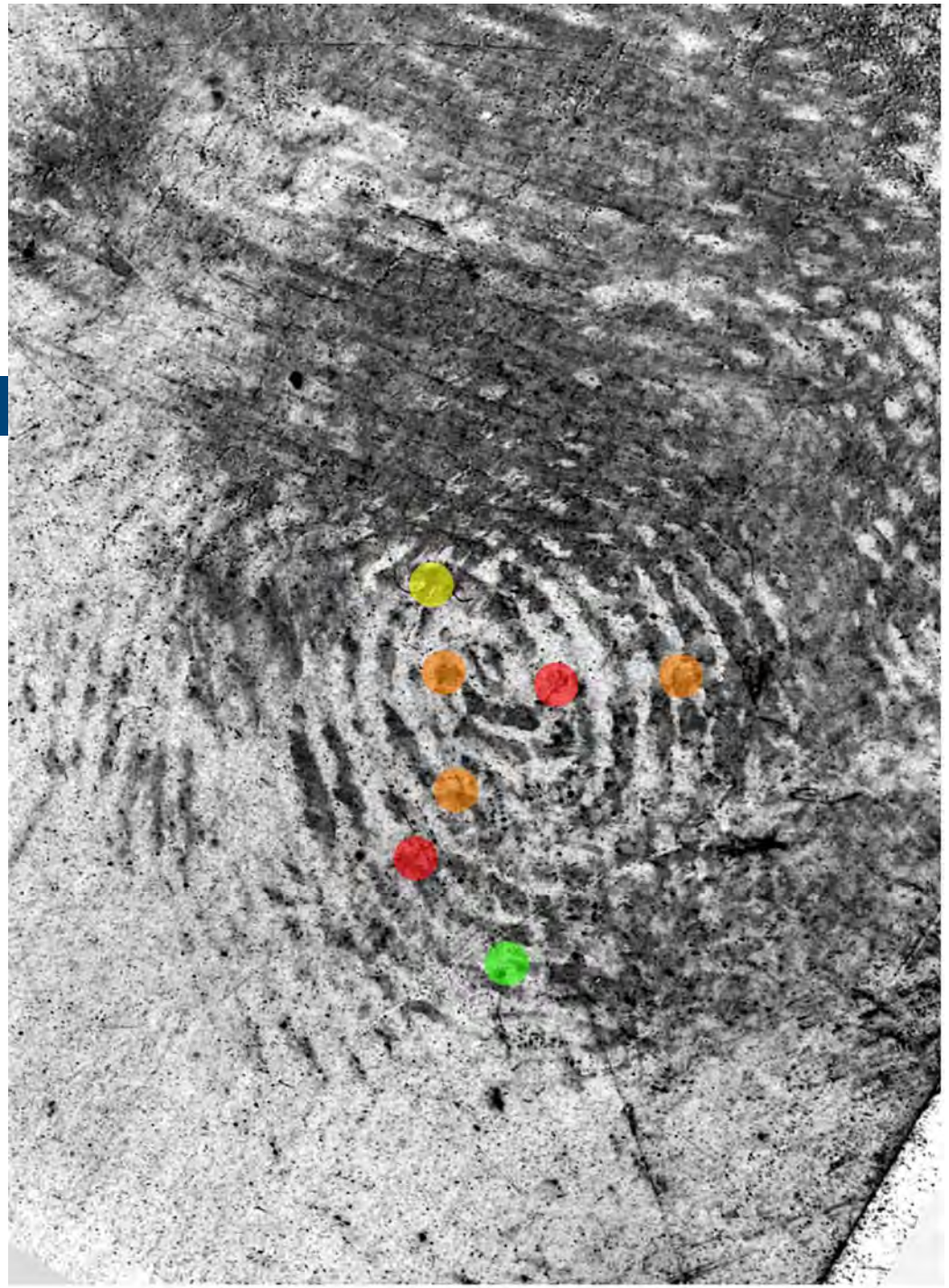
Less than 20% not indicated; an "X" indicates a 'false' minutia

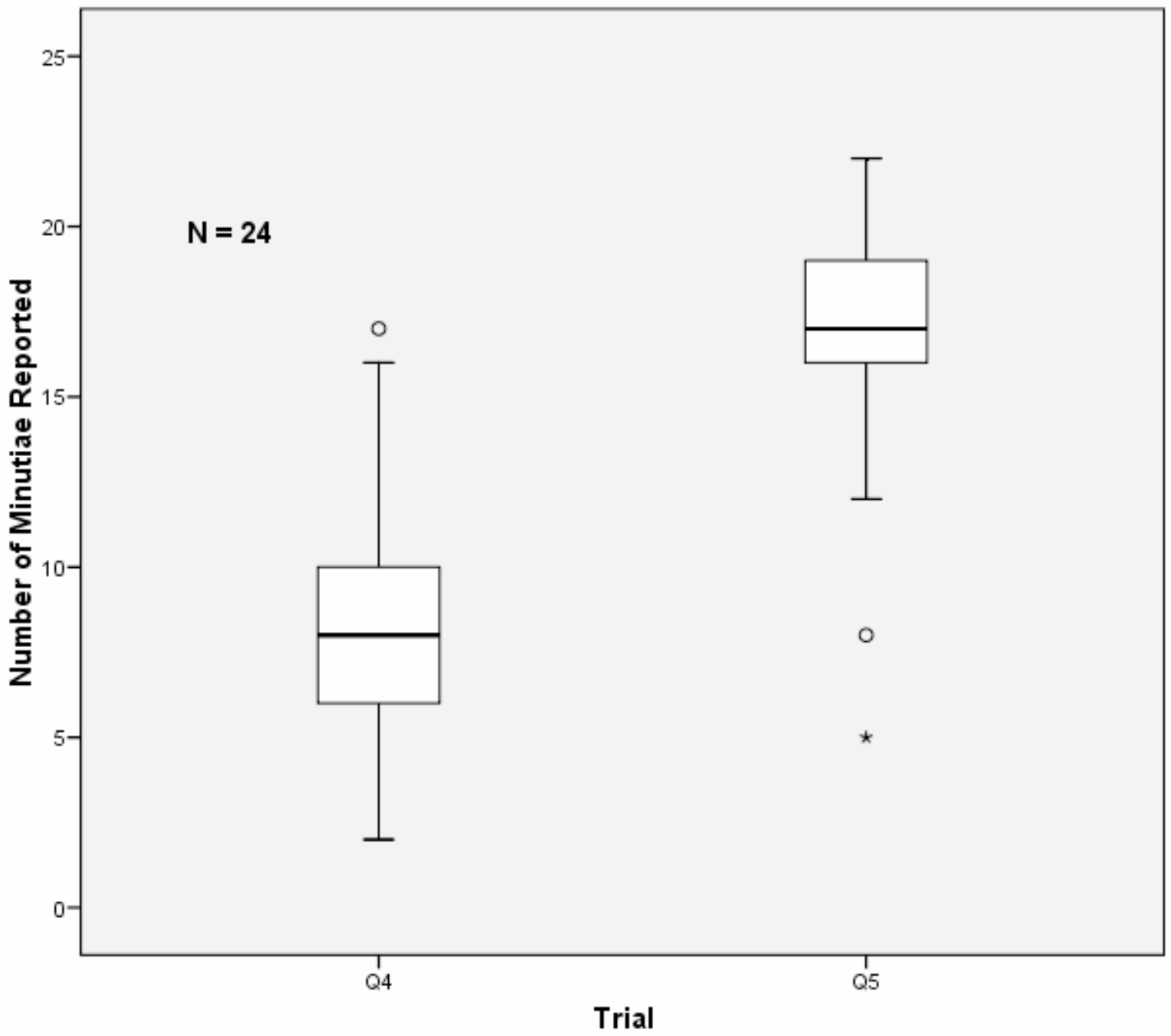


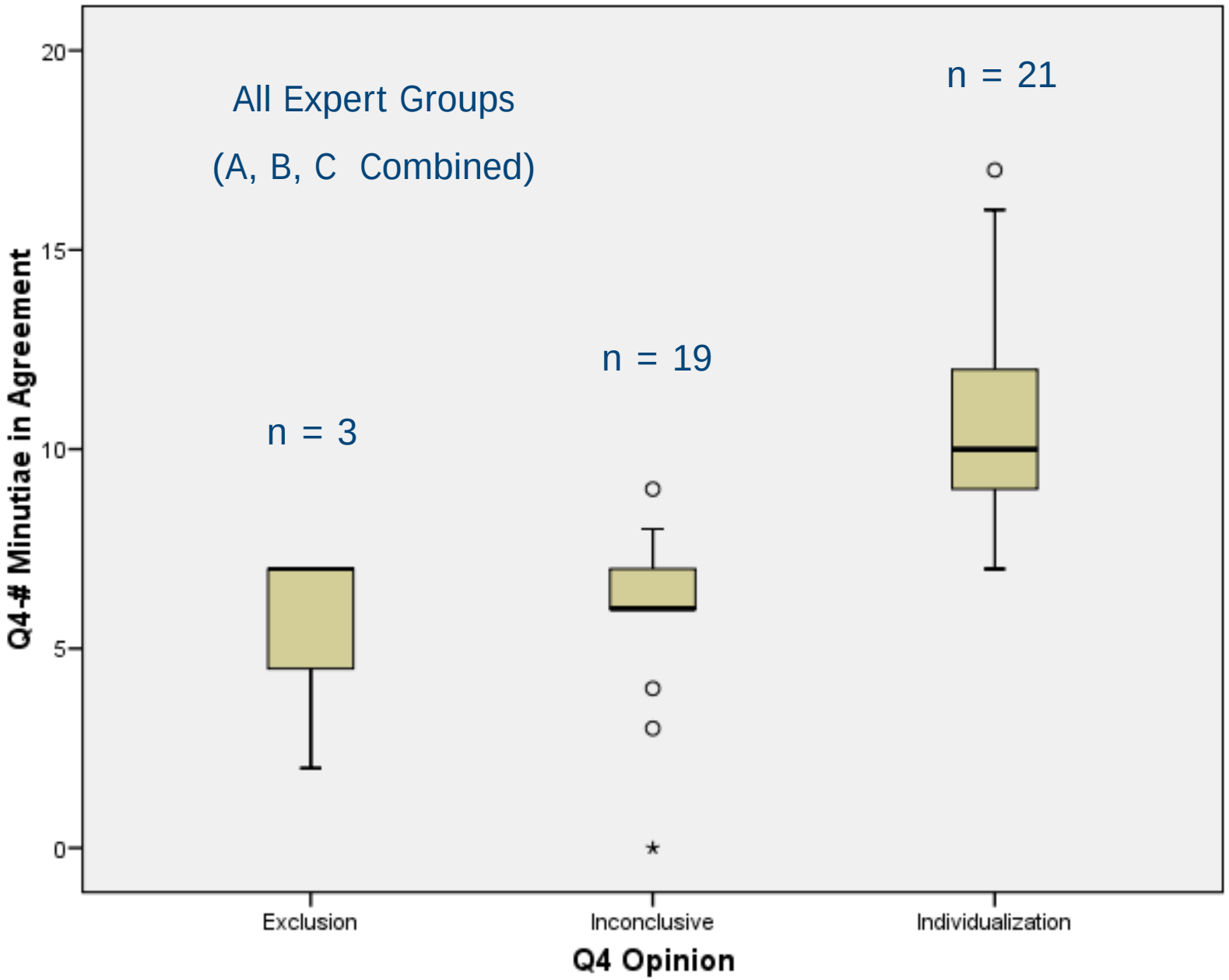
Consensus Features

- Seven consensus features

- Use these as the foundation/basis of decision.




N = 24
Number of Minutiae Reported
25
20
15
10
5
0
Q4
Q5
Trial

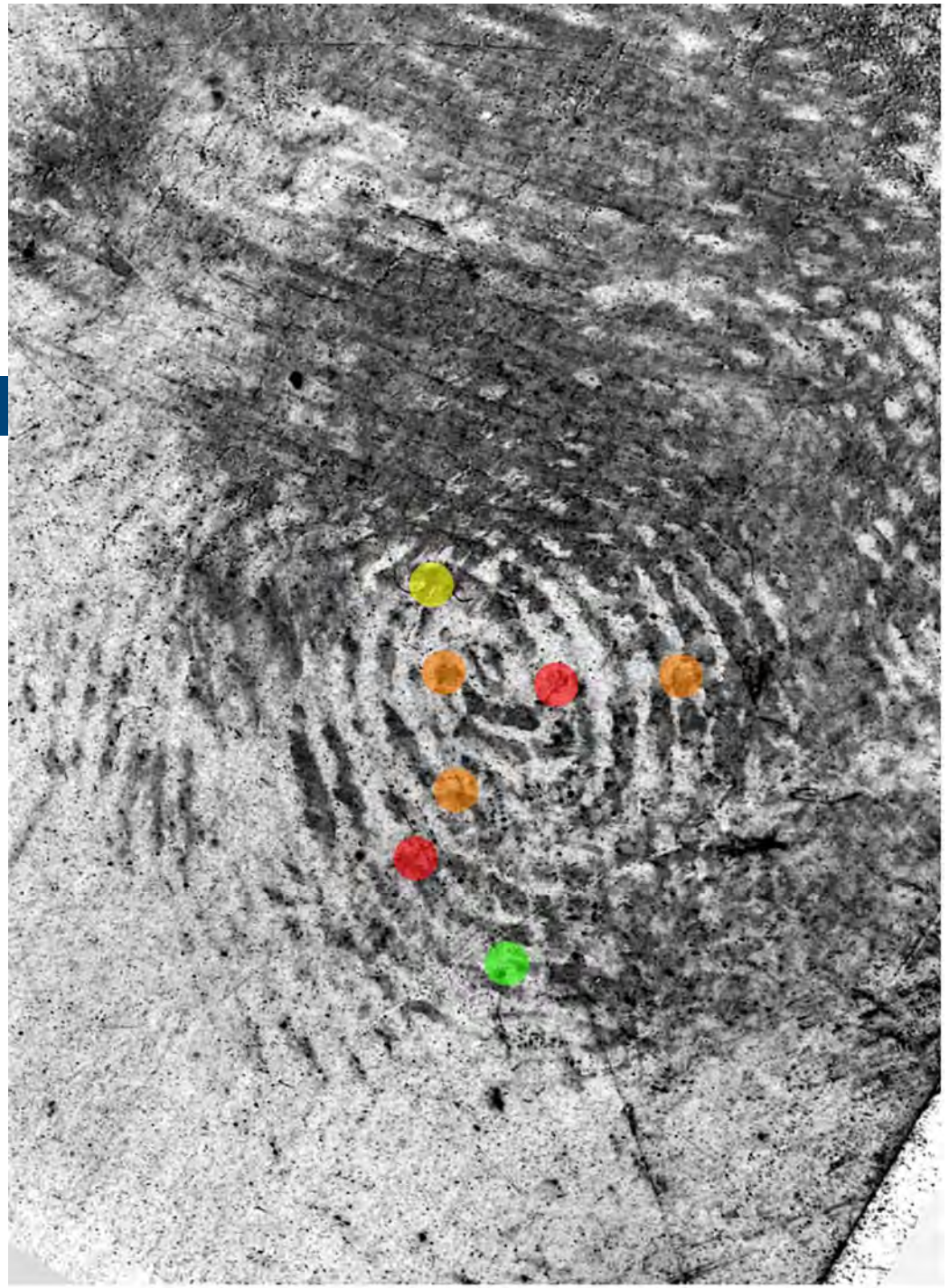

All Expert Groups
(A, B, C Combined)
n = 3
n = 19
n = 21
Q4-# Minutiae in Agreement
20
15
10
5
0
Exclusion
Inconclusive
Individualization
Q4 Opinion

p < .001, K-W test

Consensus Features

- Seven consensus features

- Use these as the foundation/basis of decision.



Summary

- FP examinations have high degree of accuracy
- But concerns about reliability as quantity and quality decrease
  - Lower reproducibility
  - Lower repeatability
- Need for tools such as:
  - Quality mapping
  - Expert consensus
  - LR tools