Validating Strength of Support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions

Tom Busey
Morgan Klutzke
Alyssa Nuzzi
John Vanderkolk
Indiana University, Bloomington, IN
busey@iu.edu

Supported by NIJ Grant 2018-DU-BX-0212 to Tom Busey and Indiana University

Central Question

If we change articulation language, what happens to examiner behavior?

Traditional Categorical Conclusions

- Fingerprints: ID, Inconclusive, Exclusion
- Footwear is a variant of this, but includes statements about degree of association
- Toolmarks is a variant of this, but includes statements about ‘insufficiency’

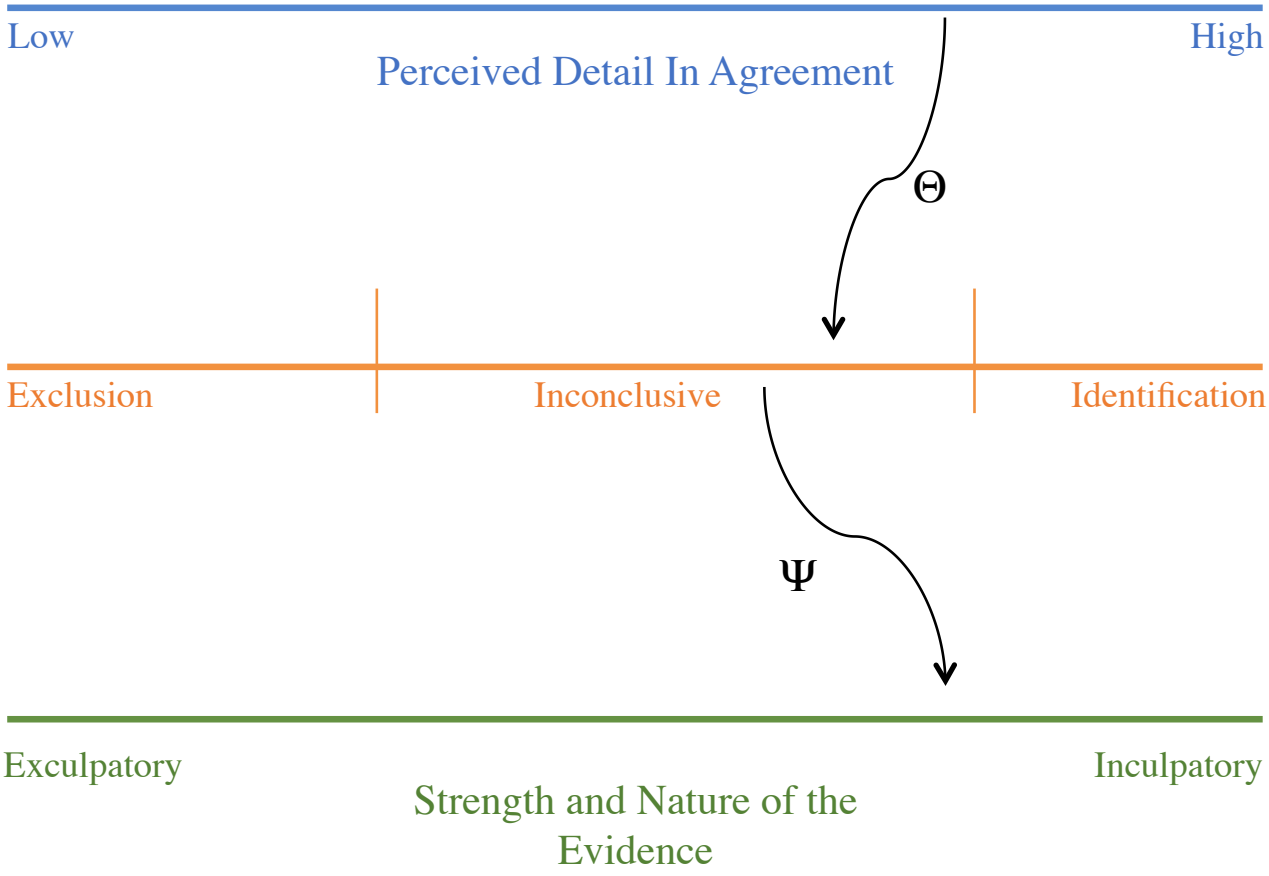
Problems with Categorical Conclusions

- Categorical conclusions tend to be overinterpreted.
- Categorical conclusions are *conclusions*. They are posteriors, and therefore must consider the priors (i.e. how good your detectives are).
- Are forensic examiners the trier of fact? Should they be making conclusions or describing the observations given two propositions?

Strength of Support Scales

- Strength of Support scales are designed to express the probability of the observations *given* two propositions (same source and different sources). ”Strong support for common source.”
- The language must be interpreted properly by both the examiner and the jury (or prosecutor).
- How would moving to strength of support scales affect examiner behavior?

Calibrating Forensic Conclusions



Internal Evidence Scale

Examiner's Conclusion

Consumer's Interpretation

Prior work: Testing an expanded conclusion scale

- We tested expanded conclusion scales with casework-like fingerprint comparisons and fingerprint examiners
- 27 examiners each did 60 comparisons, half mated and half non-mated.
- Each trial was limited to a max of 3 minutes
- Could rotate, mark up, translate the images.

- Three choices
  - Exclusion
  - Inconclusive
  - Identification
- Five choices
  - Exclusion
  - Support for different sources
  - Inconclusive
  - Support for common source
  - Identification

Pen size:

Drawing mode
(press [Space] to toggle)

29:43

Pause

Extremely strong support for different sources | Support for different sources | Inconclusive | Support for common source | Extremely strong support for common source

Testing an expanded conclusion scale

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2020, Vol. 65, No. 4
doi: 10.1111/1556-4029.14298
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Kelly E. Carter,[1] *B.A.; Macgregor D. Vogelsang*,[1] *B.S.; John Vanderkolk,*[2] *B.A.; and Thomas Busey*,[1] *Ph.D.*

The Utility of Expanded Conclusion Scales During Latent Print Examinations

Prior Work with Expanded Conclusion Scales

- When using an expanded scale with Support for Common Source, fingerprint examiners *redefined* definition of Identification, using it less often.
- Before you propose a change to your conclusion scale, test it first!

Validating Strength of Support Scales

- How would examiner behavior change if we moved to strength of support conclusion scales?
- Depends on the scale (chasing a moving target as OSAC and ASB make new proposed scales)

Fingerprint Strength of Support

Extremely Strong Support for Common Source: Extremely Strong Support for Common Source is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and weak or no support for the proposition that the impressions originated from different sources. This conclusion is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source.

Fingerprint Strength of Support

Support for Common Source: Support for Common Source is the conclusion that the observations provide more support for the proposition that the impressions originated from the same source rather than different sources.

Fingerprint Strength of Support

Inconclusive: The observed characteristics of the items are insufficient to support any of the other conclusions.

Fingerprint Strength of Support

Support for Different Sources: Support for Different Sources is the conclusion that the observations provide more support for the proposition that the impressions originated from different sources rather than the same source.

Fingerprint Strength of Support

Extremely Strong Support for Different Sources: Extremely Strong Support for Different Sources is the conclusion that the observations provide much more support for the proposition that the impressions originated from different sources and weak or no support for the proposition that the two items originated from the same source.

Fingerprint Traditional Scale Definitions

Two different definitions:

- Traditional vs. Strength of Support
  - Examiners asked to use whatever definitions they currently use in casework for the three conclusions.
- Traditional vs Expanded Traditional
  - Examiners were given explicit definitions that mirrored the Strength of Support language.

Example: Fingerprint Traditional

Identification: Identification is the strongest degree of association between two friction ridge impressions. It is the conclusion that the observations provide extremely strong support for the proposition that the impressions originated from the same source and extremely weak support for the proposition that the impressions originated from different sources. Identification is reached when the friction ridge impressions have corresponding ridge detail and the examiner would not expect to see the same arrangement of details repeated in an impression that came from a different source.

Fingerprint Traditional

These are asking slightly different questions, but we can compare them statistically.

Footwear Strength of Support

- Similar scale, but 6 levels, including Extremely Strong Support, Strong Support, Support, Indeterminate With Respect to Source.
- Includes language about randomly acquired characteristics and class characteristics.

Toolmark Strength of Support

- Similar scale to the fingerprint strength of support scale (5 levels).
- Includes language about randomly acquired characteristics and class characteristics.

Project Goal

- Compare Traditional Conclusion Scales to Strength of Support Scales in Fingerprint, Footwear, and Toolmark disciplines
- Use casework-like stimuli and an online interface to conduct the comparisons.

Fingerprint Stimuli

- Collected in our lab to simulate casework-like impressions.
- Similar in difficulty to FBI/Noblis Black Box Study
  - Sensitivity (d') is 2.39 in our study, 2.64 in Ulery (2011).
- We are comparing scales, so the exact difficulty is less important.

Footwear Stimuli

- Collected in our lab to simulate casework-like impressions.
- Heel impressions taken in chocolate ice cream to simulate blood and slapped on white paper.
- Test impressions with petroleum jelly and black powder on lifting substrate.
- Photographs of the heel.
- All images available on the github site.

Toolmark Stimuli

- Collected in our lab.
- 15 screwdrivers and 15 chisels.
- Marks made on heavy-duty aluminum foil.
- Photographed with oblique lighting.
- 20° and 35° angles used.
- All images available on github site.

Toolmark Stimuli

Study Design

- Each participant does 60 trials
  - Half are with the traditional scale, and half are with the strength of support scale
  - They know on each trial which scale they are using
- Of the trials, half are mated and half are nonmated
  - 15 mated traditional trials, 15 nonmated traditional trials, 15 mated strength of support trials, 15 nonmated strength of support trials
- 30 minute timer, which could be paused (the images were hidden during the pause)

Pen size:

Drawing mode
(press [Space] to toggle)

29:43

Pause

Extremely strong support for different sources | Support for different sources | Inconclusive | Support for common source | Extremely strong support for common source

Choose a color for the Questioned impression: Black

Hide Questioned impression

Choose a color for the Known impression: Red

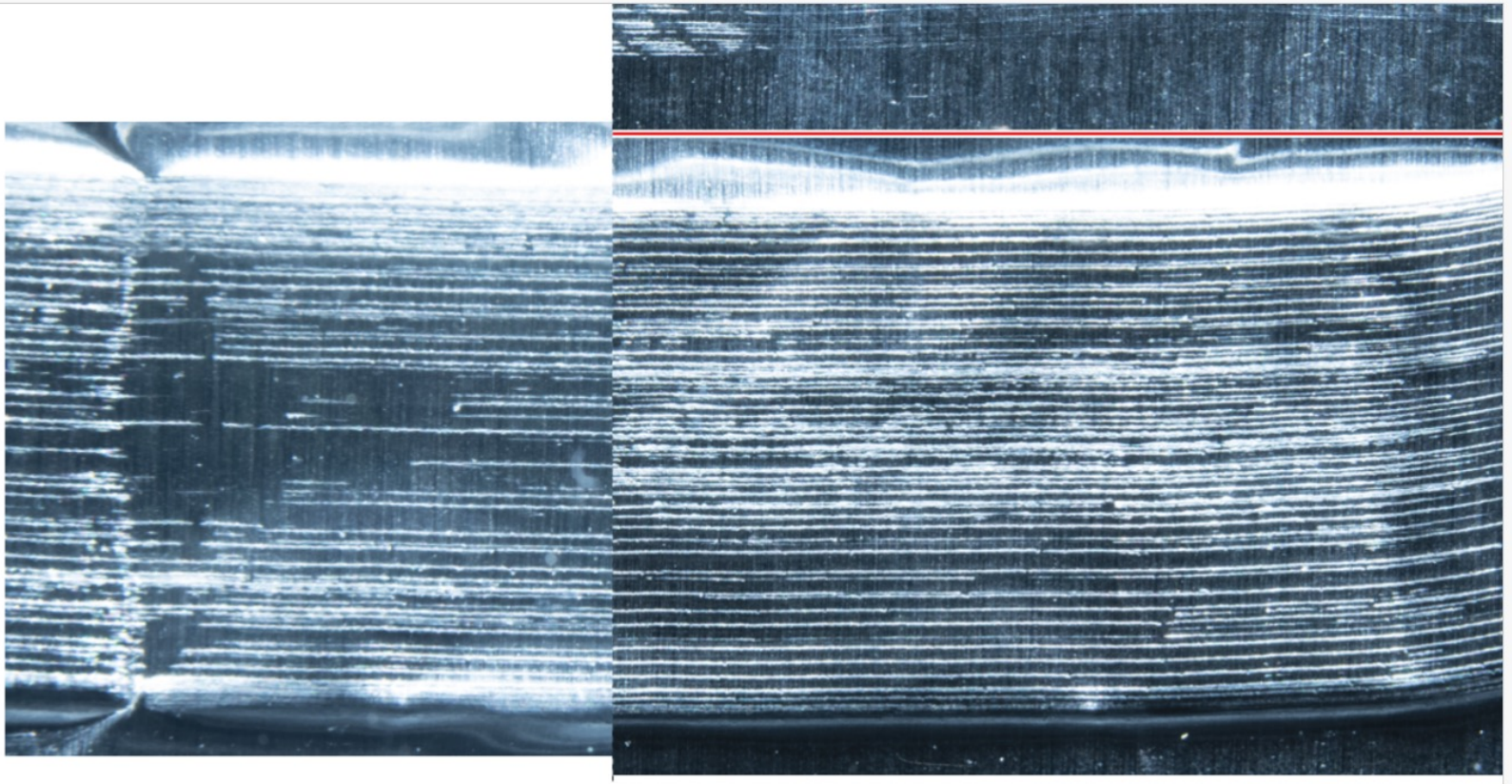
Hide Known impression

Reset

Exclusion | Indications of Non-association | Inconclusive | Limited Association | High Degree of Association | Identification

29:40

Pause



Elimination | Insufficient for Elimination | Inconclusive | Insufficient for Identification | Identification

29:31

Pause

Reset

Study Design- Strength of Support

- The images are randomly assigned to one of the two scales for each participant
- Rerandomizing for each participant means that any differences in difficulty average out across participants
- If we find that, on average, examiners use the highest conclusion less often in one scale relative to the other, this can't be because that scale had harder images.

Participants

- Participants had to be qualified to testify in their discipline in the US.
- Fingerprints: Traditional vs. Strength of Support
  - 34 examiners
- Fingerprints: Traditional vs. Expanded Traditional
  - 32 examiners
- Footwear
  - 32 examiners
- Tool marks
  - 20 examiners

Data Collection

- Data was collected via a custom javascript-based interface specific to each discipline
- Password protected, data saved remotely on our server
- The study was self-paced, and trials could be paused

Results

- Present raw frequencies of responses
- Fit the proportions of responses using a variant of signal detection theory, which estimates examiner ability
- Data fit to all participants combined (prior study also fit the data from individual subjects and found essentially equivalent results as fitting group data).

Results: Fingerprints

Traditional Scale					
Ground Truth	Exclusion		Inconclusive		Identification
Nonmated	265	NA	254	NA	1
Mated	26	NA	244	NA	250
Strength of Support Scale					
Ground Truth	Extremely Strong Support for Different Sources	Support for Different Sources	Inconclusive	Support for Common Source	Extremely Strong Support for Common Source
Nonmated	195	127	185	9	1
Mated	8	37	179	117	180

Results- Strength of Support

- The frequency of Identification (250) dropped for Extremely Strong Support for Common Source (180).
- Support for Common Source (117) absorbed the remainder, along with some of the inconclusives, which dropped from 244 to 179.

Results: Fingerprints

Traditional Scale					
Ground Truth	Exclusion		Inconclusive		Identification
Nonmated	272	NA	207	NA	2
Mated	30	NA	220	NA	231

Expanded Traditional Scale					
Ground Truth	Exclusion	Support for Different Sources	Inconclusive	Support for Common Source	Identification
Nonmated	225	92	154	6	2
Mated	22	33	131	93	199

Results- Expanded Traditional

- The frequency of Identification (231) dropped to 199 when Support for Common Source was added to the scale.
- Support for Common Source (93) absorbed the remainder, along with some of the inconclusives, which dropped from 220 to 131.
  - Absorbed some of the weaker IDs.

Results: Footwear

	Definitive Conclusions (Traditional) Scale					
Ground Truth	Exclusion	Indications of Non-Association	Inconclusive	Limited Association	High Degree of Association	Identification
Nonmated	260	126	11	65	15	3
Mated	25	23	17	117	146	151
	Strength of Support Scale					
Ground Truth	Extremely Strong Support for Different Sources	Support for Different Sources	Indeterminate with Respect to Source	Support for Common Source	Strong Support for Common Source	Extremely Strong Support for Common Source
Nonmated	258	123	33	51	15	2
Mated	18	29	23	147	103	160

Results- Footwear

- Examiners used the two scales fairly consistently.
- Perhaps this is a function of the fact that both the traditional and the strength of support have 6 categories.

Results: Tool Marks

Traditional Scale					
Ground Truth	Elimination	Insufficient for Elimination	Inconclusive	Insufficient for Identification	Identification
Nonmated	84	68	113	24	2
Mated	8	20	57	47	178

Strength of Support Scale					
Ground Truth	Extremely Strong Support for Different Sources	Support for Different Sources	Inconclusive	Support for Common Source	Extremely Strong Support for Common Source
Nonmated	45	119	110	13	2
Mated	11	19	61	84	136

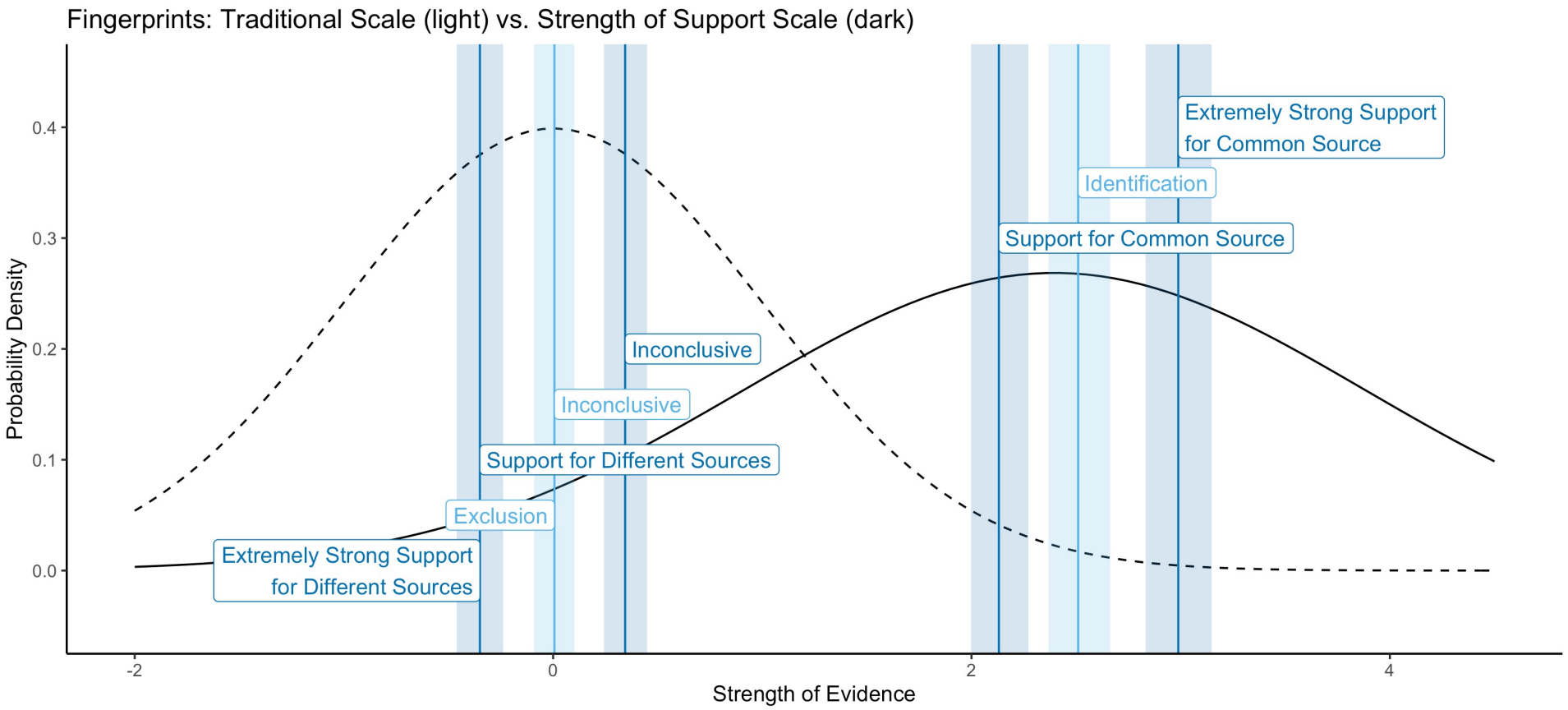
Results- Strength of Support

- The frequency of Identification (173) dropped for Extremely Strong Support for Common Source (136).
- Support for Common Source (84) increased over Insufficient for Identification (47). Support for Common Source is absorbing the conclusions that might have been IDs but examiners are now reluctant to apply the Extremely Strong Support for Common Source label.

Fitting Signal Detection Theory Models

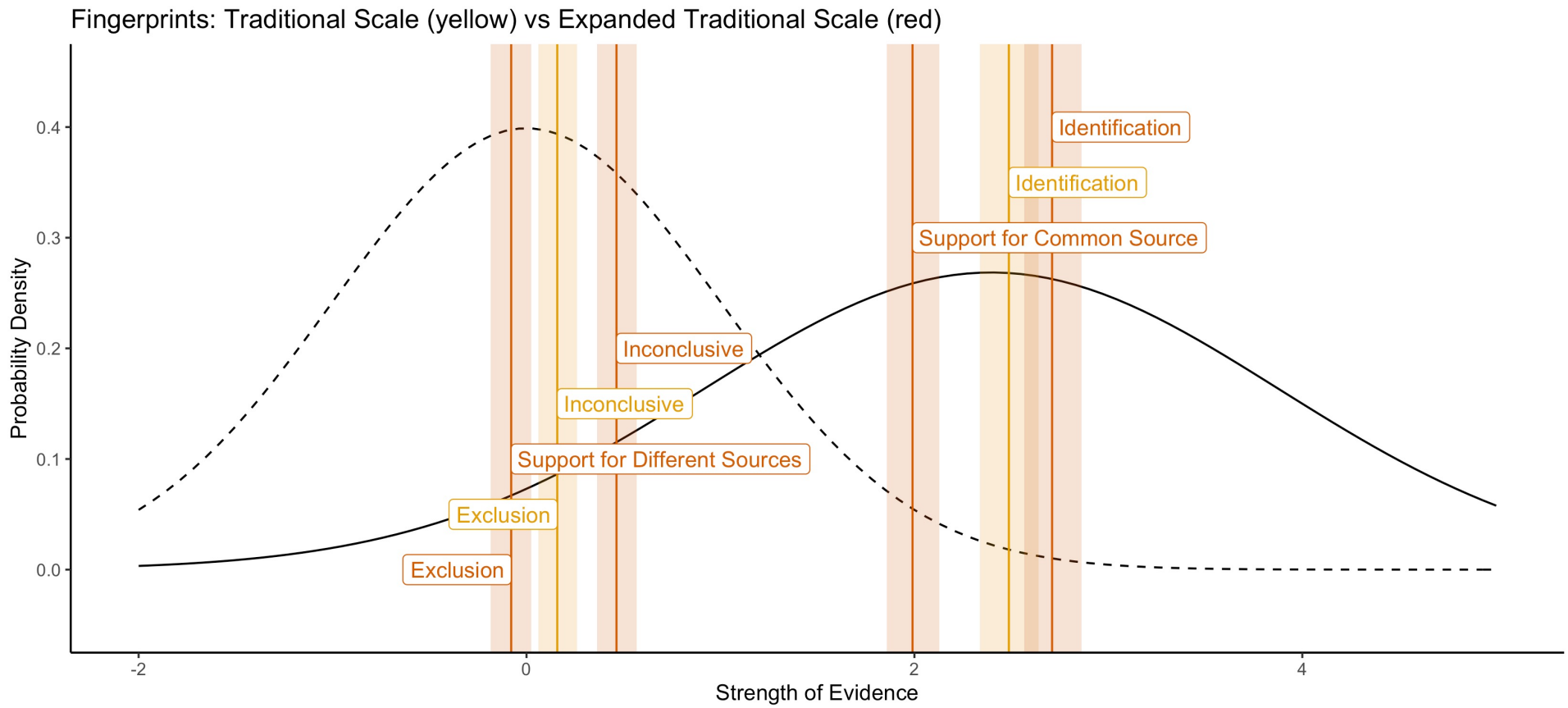
- Allows an easy estimation of the decision criteria, along with error bars around the decision criteria.
- Assume equal sensitivity for the traditional and strength of support scales.
- The frequencies are converted to proportions, which are then fit in R using the GLM procedure brm from the brms library.

Fingerprints Strength of Support



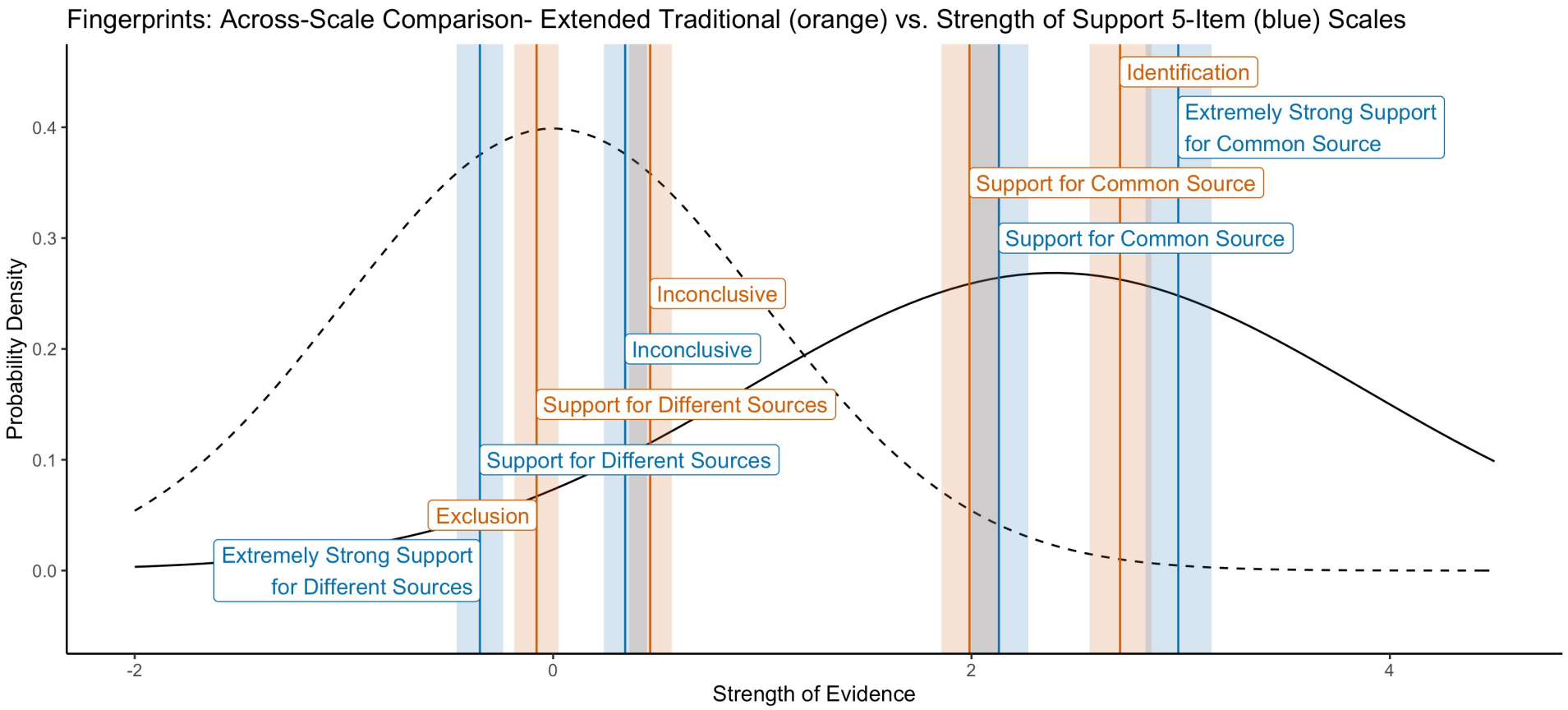
Examiners become more risk-averse with Extremely Strong Support for Common Source

Fingerprints Expanded Traditional



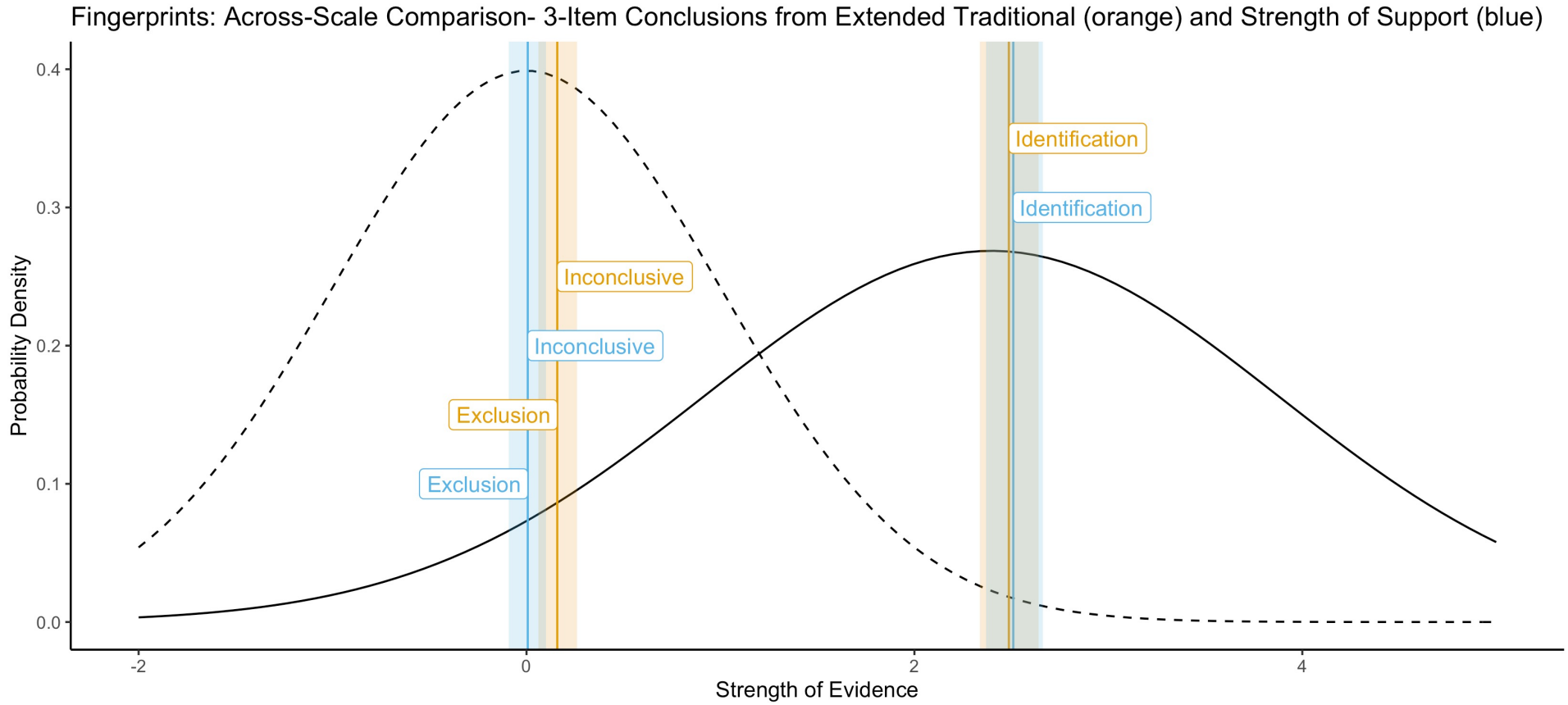
Examiners become slightly more risk-averse with an expanded traditional scale.

Compare the Two Scales



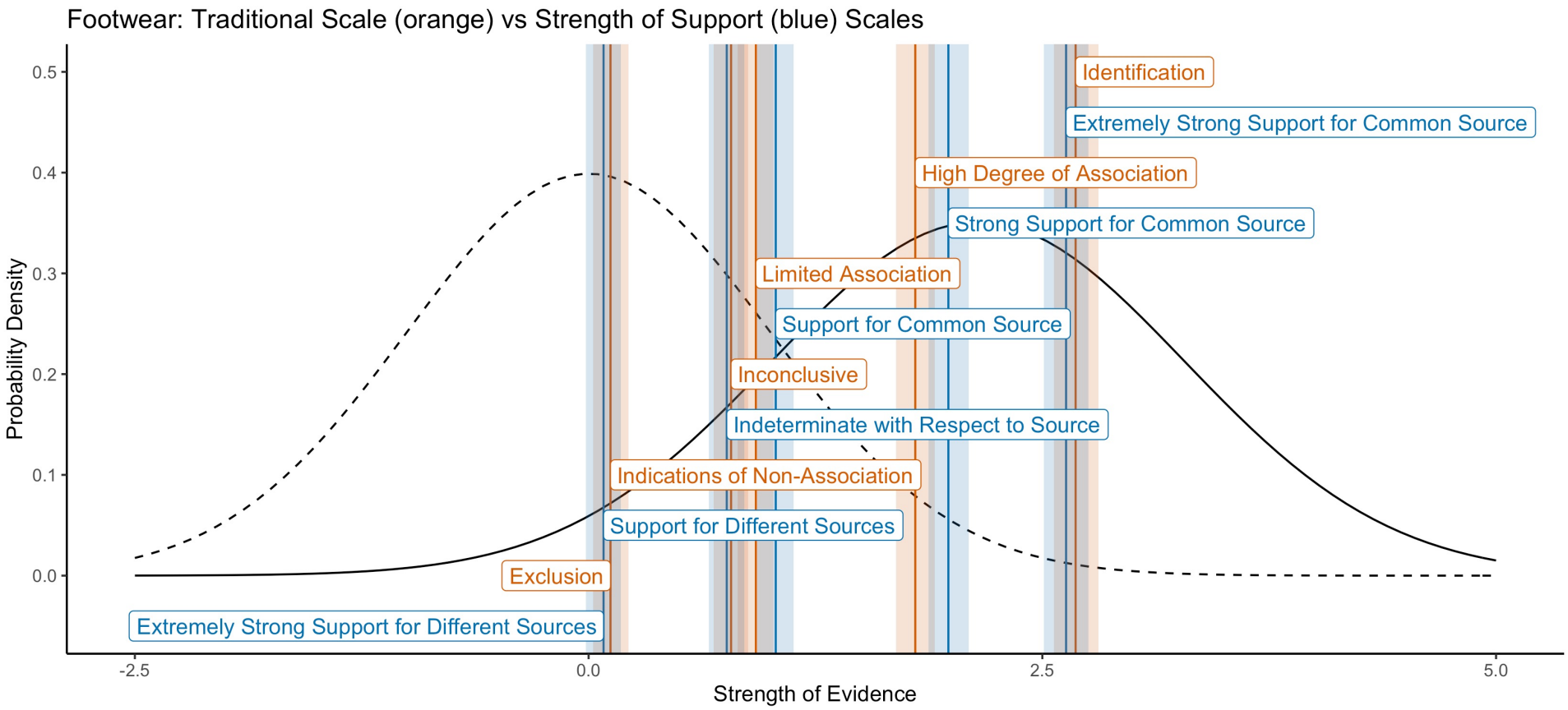
Difference between Identification and Extremely Strong Support for Common Source is evident in this comparison.

Fingerprints Traditional Scales



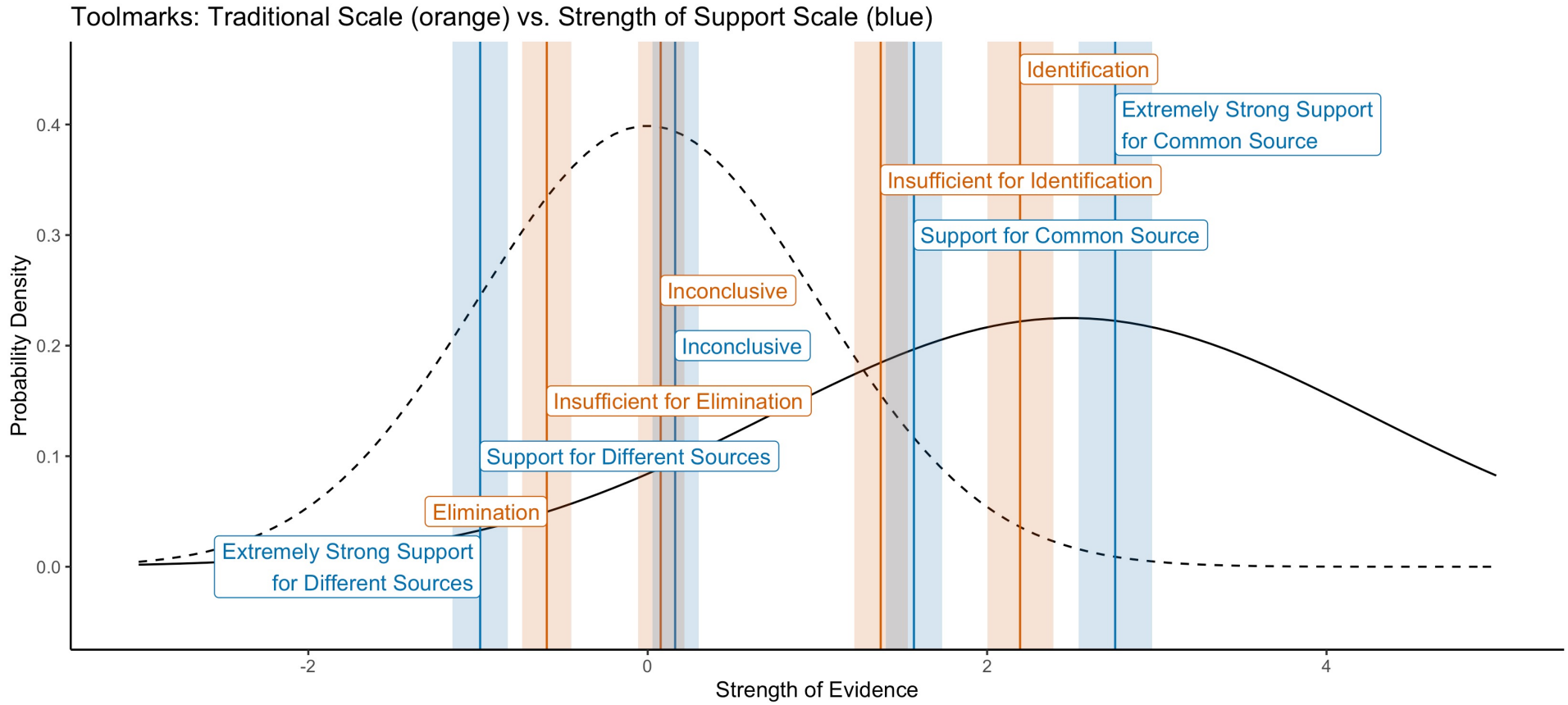
Slight task differences don't affect how fingerprint examiners treat the traditional scales in the two variants.

Footwear



Footwear examiners treat the two scales fairly equivalently, with the possible exception of High Degree of Association/Strong Support for Common Source

Toolmarks



Examiners become more risk-averse with Extremely Strong Support for Common Source

Data Summary

- Switching from a traditional scale to a strength of support scale made fingerprint examiners and toolmark examiners more risk averse.
- Expanding the traditional scale resulted in a redefinition of the Identification category.
- Footwear examiners are fairly stable across scales.

What does it all mean?

Paradox:

- Fingerprint examiners believe that Extremely Strong Support for Common Source provides less support for the Common Source Proposition than Identification Does (previous talk).
- Yet when given a chance to actually use this scale in practice, they are less likely to use it than Identification.

Why are examiners more risk averse with a strength of support scale?

- Examiners could be less familiar with the term.
- ”Extremely” is too extreme?
- Is the language calibrated against the strength of the discipline?
- How is “Extremely” treated in DNA?

Verbal Equivalents

Proposed verbal scale for reporting the value of the scientific observations (translated from French).

Verbal communication	LR
The results support the proposition that… rather than the proposition that…. This support is qualified as *extremely strong*.	>10,000
The results support the proposition that… rather than the proposition that…. This support is qualified as *very strong*.	>1000–10,000
The results support the proposition that… rather than the proposition that…. This support is qualified as *strong*.	>100–1000
The results support the proposition that… rather than the proposition that…. This support is qualified as *moderate*.	>10–100
The results support the proposition that… rather than the proposition that…. This support is qualified as *weak or limited*.	>1–10
The results support neither propositions. This support is qualified as *null*.	1

Verbal Equivalents

LIKELIHOOD RATIO	VERBAL EQUIVALENT
< 0.01 (1/100)	Exclusion
0.01 (1/100) – < (1/2)	Limited support for exclusion
1-2	Uninformative
2 – < 100	Limited support for inclusion
100 – < 10,000	Moderate support for inclusion
10,000 - < 1,000,000	Strong support for inclusion
≥ 1,000,000	Very strong support for inclusion

SWGDAM verbal equivalent scale for reporting DNA LRs

Suggestions…

- ”Extremely” seems complicated. There is a disconnect between how examiners interpret it and actually use it.
- Novices treat it like Identification
- Examiners become risk-averse when it is available (maybe appropriately so).
- Is it ok to use such a phrase where the discipline error rate is .1%?

DFSC/USACIL approach

Strengths:

- Grounded in the physical evidence
- Repeatable
- Explicitly states the strength of support for two mutually exclusive and exhaustive propositions
- Laypersons do not overinterpret the strength of the evidence

DFSC/USACIL approach

Weaknesses:

- Laypersons (and examiners!) don't know how to interpret the numbers
- Probably an underestimate of the true strength of the evidence

Final Thoughts

- Laypersons are bad at gauging absolute strength of evidence, but are better if you tell them what you *could* have said but didn't.
- Tell them the whole scale.
- A similar comparative approach might work for quantitative approaches. What values could you have given but didn't?
- Verbal equivalents are bad, but maybe the least bad option…

Thank You

Papers:

https://buseylab.sitehost.iu.edu/

busey@iu.edu