Calibrating the Perceived Strength of Evidence of Forensic Testimony Statements

Tom Busey
Morgan Klutzke
Indiana University, Bloomington, IN
busey@iu.edu

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

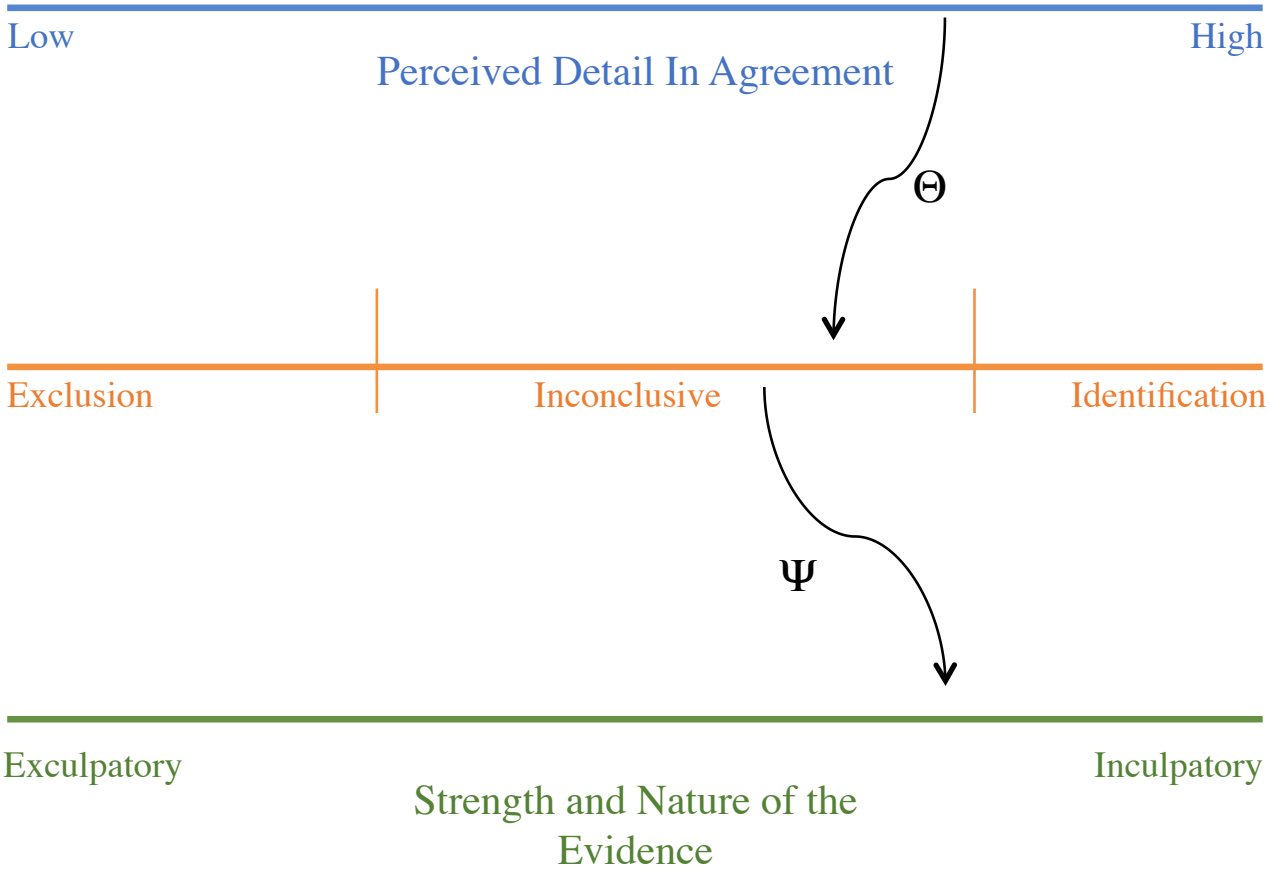
History of this Project

- Role of OSAC
- Role of psychology research
- Role of this talk (and papers)

Articulation Language

- How are different phrases interpreted by laypersons?
- This interpretation is part of the calibration of forensic conclusions!

Calibrating Forensic Conclusions



Internal Evidence Scale

Examiner's Conclusion

Consumer's Interpretation

Prior Work

- Thompson, Grady, Lai and Stern (2018) presented pairs of statements to members of the public (Amazon Mechanical Turk workers)
- Asked the participants: "Which of the following two conclusions would seem STRONGER if you heard it, meaning more convincing to you that the suspect is the source of the print?"

W.C. Thompson, R.H. Grady, E. Lai, H.S. Stern, *Perceived strength of forensic scientists' reporting statements about source conclusions*, Law, Probability and Risk, 17 (2018) 133-155.

Prior Work Conclusions

- They caution against the term ‘match’, and noted the potential misinterpretation of RMPs.
- The study found that categorical conclusions tended to be interpreted as providing strong support, which the authors found concerning.
- A downside to this approach is that it presents each statement in isolation, rather than as part of a complete scale.

How do we measure perceived strength of support?

- Measuring small numbers is hard.
- Paper and pencil activity…

1 1 Billion

Draw a line where 1 Million falls on this line.
Don't think, just guess (like a jury might do).

Let’s calibrate the articulation language

- Test both fingerprint examiners and laypersons using an online interface.

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

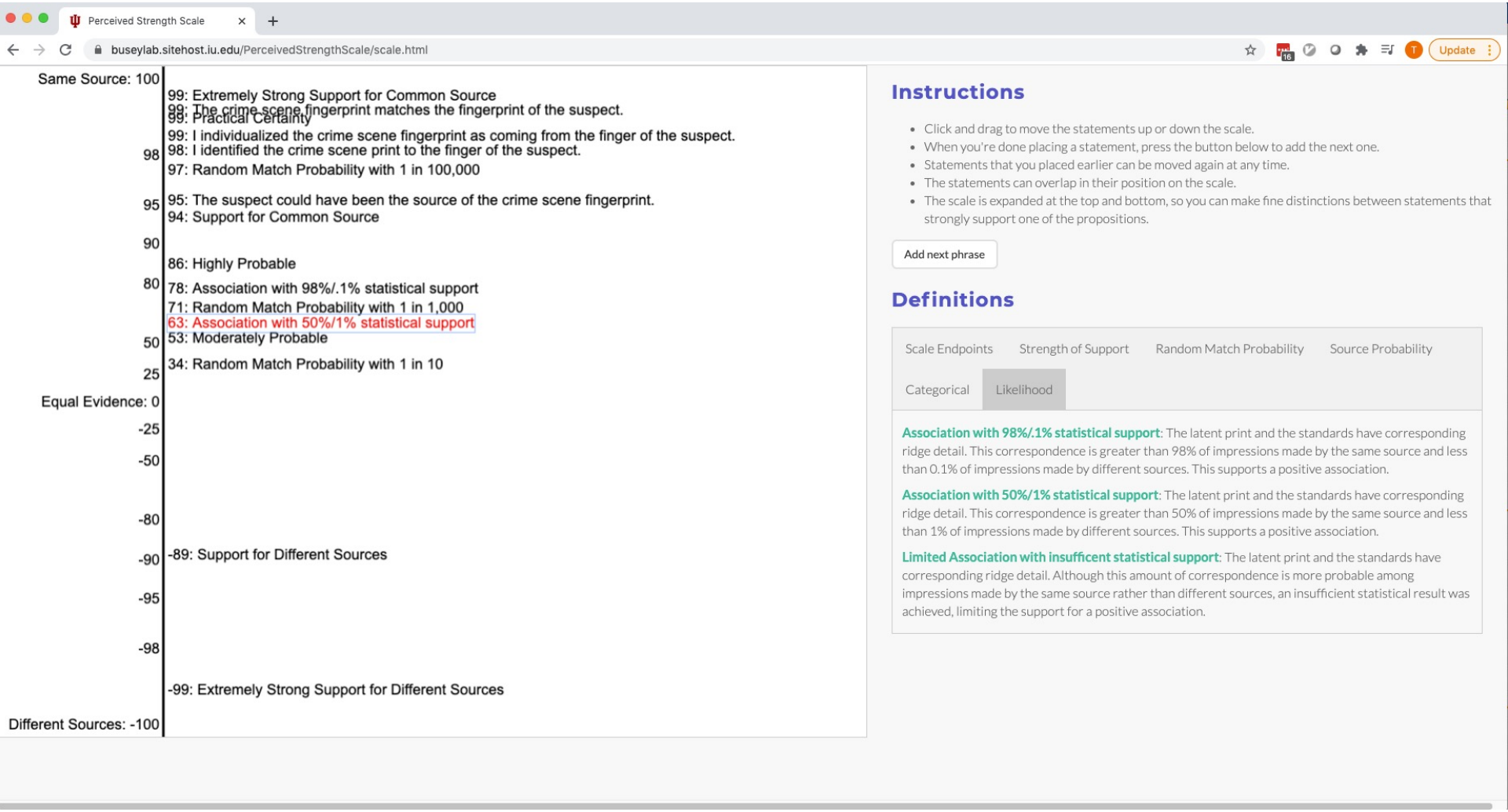
-or-

https://go.iu.edu/48KU

(the ‘7’ hotkey skips the video)

Notable features

- Subjects had to demonstrate that they understood each scale by sorting the phrases properly. That also made them read the statements enough to properly sort them.
- Note that the scale is not linear- expanded at the endpoints.


Perceived Strength Scale
buseylab.sitehost.iu.edu/PerceivedStrengthScale/scale.html
Update
Same Source: 100
99: Extremely Strong Support for Common Source
99: The crime scene fingerprint matches the fingerprint of the suspect.
99: Practical Certainty
99: I individualized the crime scene fingerprint as coming from the finger of the suspect.
98
98: I identified the crime scene print to the finger of the suspect.
97: Random Match Probability with 1 in 100,000
95
95: The suspect could have been the source of the crime scene fingerprint.
94: Support for Common Source
90
86: Highly Probable
80
78: Association with 98%/.1% statistical support
71: Random Match Probability with 1 in 1,000
63: Association with 50%/1% statistical support
50
53: Moderately Probable
25
34: Random Match Probability with 1 in 10
Equal Evidence: 0
-25
-50
-80
-90
-89: Support for Different Sources
-95
-98
-99: Extremely Strong Support for Different Sources
Different Sources: -100
Instructions
Click and drag to move the statements up or down the scale.
When you're done placing a statement, press the button below to add the next one.
Statements that you placed earlier can be moved again at any time.
The statements can overlap in their position on the scale.
The scale is expanded at the top and bottom, so you can make fine distinctions between statements that strongly support one of the propositions.
Add next phrase
Definitions
Scale Endpoints
Strength of Support
Random Match Probability
Source Probability
Categorical
Likelihood
Association with 98%/.1% statistical support: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 98% of impressions made by the same source and less than 0.1% of impressions made by different sources. This supports a positive association.
Association with 50%/1% statistical support: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 50% of impressions made by the same source and less than 1% of impressions made by different sources. This supports a positive association.
Limited Association with insufficent statistical support: The latent print and the standards have corresponding ridge detail. Although this amount of correspondence is more probable among impressions made by the same source rather than different sources, an insufficient statistical result was achieved, limiting the support for a positive association.

Traditional

These terms have been traditionally used in the pattern comparison disciplines such as fingerprint comparison, and describe a conclusion made by the examiner.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

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.

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

Exclusion is the conclusion that two friction ridge impressions did not originate from the same source. Exclusion is reached when in the examiner's opinion, considering the observed data, the probability that the two impressions came from the same source is considered negligible.

Check

Strength of Support

These terms are designed to express the strength of support for one of the two propositions.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

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.

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.

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.

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.

Likelihood

These statements are used where statistical software can provide support for some conclusions. Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

Association with 98%/.1% statistical support: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 98% of impressions made by the same source and less than 0.1% of impressions made by different sources. This supports a positive association.

Association with 50%/1% statistical support: The latent print and the standards have corresponding ridge detail. This correspondence is greater than 50% of impressions made by the same source and less than 1% of impressions made by different sources. This supports a positive association.

Limited Association with insufficent statistical support: The latent print and the standards have corresponding ridge detail. Although this amount of correspondence is more probable among impressions made by the same source rather than different sources, an insufficient statistical result was achieved, limiting the support for a positive association.

Check

Random Match Probability

A random match probability is an expression of the chance of a coincidental match of a given set of features in a population. It describes how many people in the population would have fingerprints that are similar to the present one.

Higher numbers are associated with unique or rare features in the fingerprint, such as an unusual whorl or pattern.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

Random Match Probability with 1 in 100,000 Given the size and quality of the crime scene print I would expect about one person in 100,000 to have a fingerprint similar enough to be indistinguishable from it.

Random Match Probability with 1 in 1,000 Given the size and quality of the crime scene print I would expect about one person in 1000 to have a fingerprint similar enough to be indistinguishable from it.

Random Match Probability with 1 in 10 Given the size and quality of the crime scene print I would expect about one person in 10 to have a fingerprint similar enough to be indistinguishable from it.

Check

Source Probability

Source probability statements provide the probability of the same-source proposition.

Please sort the statements by most evidence for Same Source (at the top) to most evidence for Different Sources (at the bottom).

Practical Certainty: Given the size and quality of the crime scene print, it is a practical certainty that the suspect is the person who made the crime scene print.

Highly Probable: Given the size and quality of the crime scene print, it is highly probable that the suspect is the person who made the crime scene print.

Moderately Probable: Given the size and quality of the crime scene print, it is moderately probable that the suspect is the person who made the crime scene print.

Check

Categorical

These are statements that are sometimes used in some jurisdictions to describe the conclusion of the examiner.

Unlike other scales, there is no clear ordering of these statements but you should read and interpret each sentence.

The suspect could have been the source of the crime scene fingerprint.

I individualized the crime scene fingerprint as coming from the finger of the suspect.

I identified the crime scene print to the finger of the suspect.

The crime scene fingerprint matches the fingerprint of the suspect.

Check

Participants

- LPEs recruited through prior mailing lists, CLPex, snowball recruitment
  - Unpaid
  - Qualified to testify on fingerprint evidence in the United States
  - 18 years or older
  - Jury-eligible in the United States

Participants

- Trusted Novices- Personally-recruited members of the Bloomington Community, family and friends, church and community members, former students, and close associates
  - Unpaid
  - 18 years or older
  - Jury-eligible in the United States

Participants

- Amazon's Mechanical Turk
  - paid $2 for a 15- minute experiment.
  - 18 years old and jury-eligible in the United States.
  - HIT approval rate of greater than 97
  - Number of HITs approved above 5000
  - Location in the United States.

Number of Participants

- LPEs- 126
- Bloomington Community- 45
- Amazon’s Mechanical Turk- 143

Age Demographics

Group	18-24	25-34	35-44	45-54	55-64	65-74	75+	Decline
Bloomington Community	12	3	5	7	8	1	2	0
Fingerprint Examiners	1	28	49	29	11	3	0	1
Mechanical Turk	2	27	29	19	9	6	0	0

Education

Group	Decline	Bachelor's	College Student	High School	Masters	PhD	Professional	Some College
Bloomington Community	0	10	9	1	9	6	1	2
Fingerprint Examiners	1	64	0	2	47	0	0	8
Mechanical Turk	1	41	2	13	9	1	2	23

Exclusion Criteria

- Minimum time between statements < 2 seconds
- Expect Identification to be placed above Inconclusive, and Inconclusive placed above Exclusion. Same for Extremely Strong Support for Common Source, Support for Common Source, Support for Different Sources, and Extremely Strong Support for Different Sources
- Any violation of this ordering was grounds for exclusion.

Exclusion Criteria

Also noted deviations for:

- Likelihood
- Random Match Probability
- Source Probability

However, we did not exclude participants based on these violations as these are confusing scales.

	Number of Total Violations (Unique Participants)					
Subject Type	Traditional (ID, Inc, Ex)	Strength of Support	Likelihood	Random Match Probability	Source Probability	Minimum Time Too Fast
Fingerprint Examiners	0(0)	3(3)	7(7)	13(10)	9(7)	1
Mechanical Turk	45(30)	111(37)	53(31)	99(47)	54(35)	4
Bloomington Community	2(1)	3(2)	0(0)	8(4)	5(2)	4

Participants Excluded

- This screening resulted in the exclusion of:
  - 4 of the 126 Fingerprint Examiners
  - 7 of the 45 Bloomington Community members
  - 51 out of the 143 Mechanical Turk participants

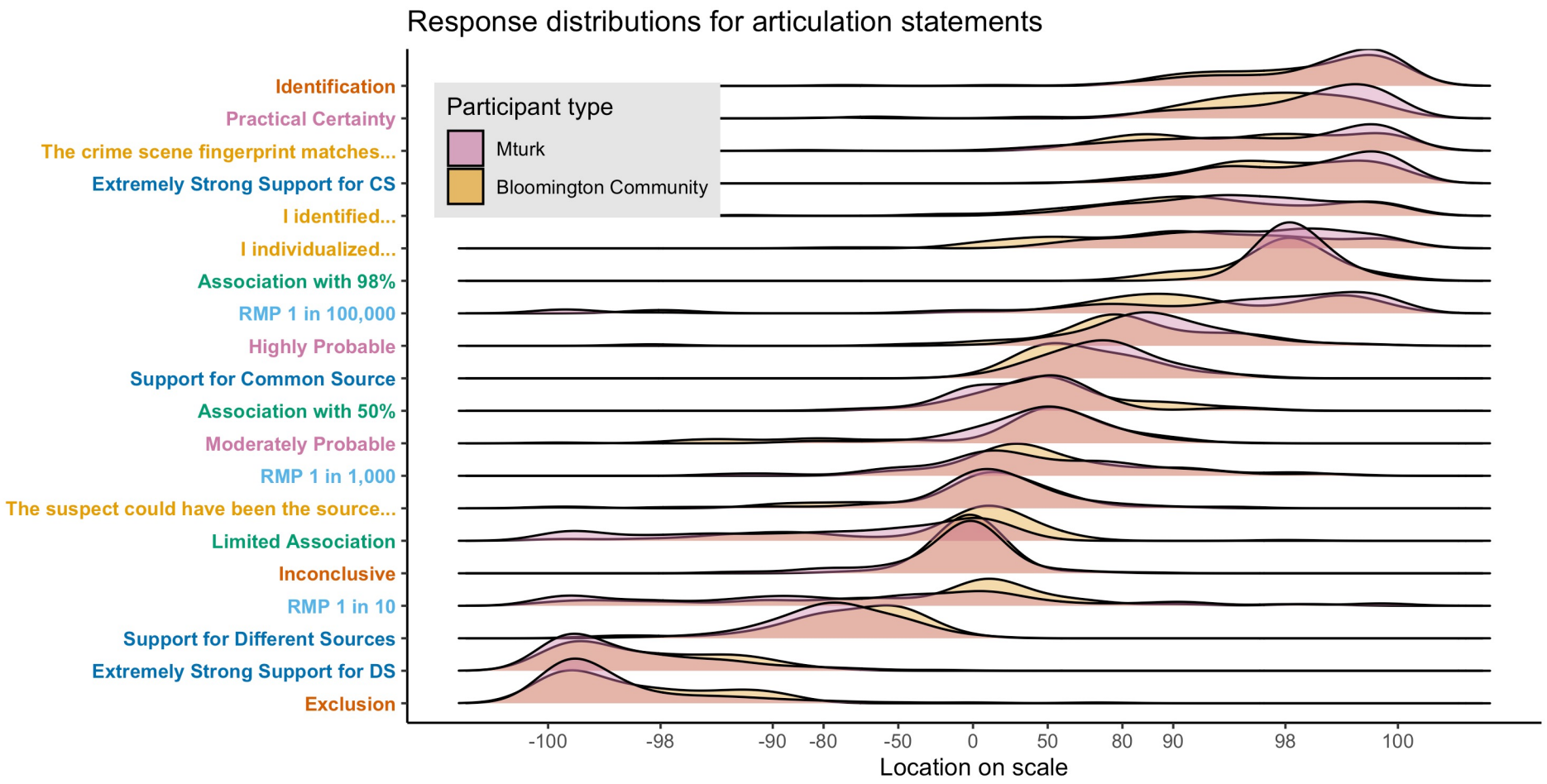
Participants

- Bottom line: some Mturk subjects don't take the task seriously.
- Having the trusted novices from the Bloomington Community helps address any concerns about the remaining Mturk subjects.
- Very similar results between the remaining Mturk and the Bloomington Community.

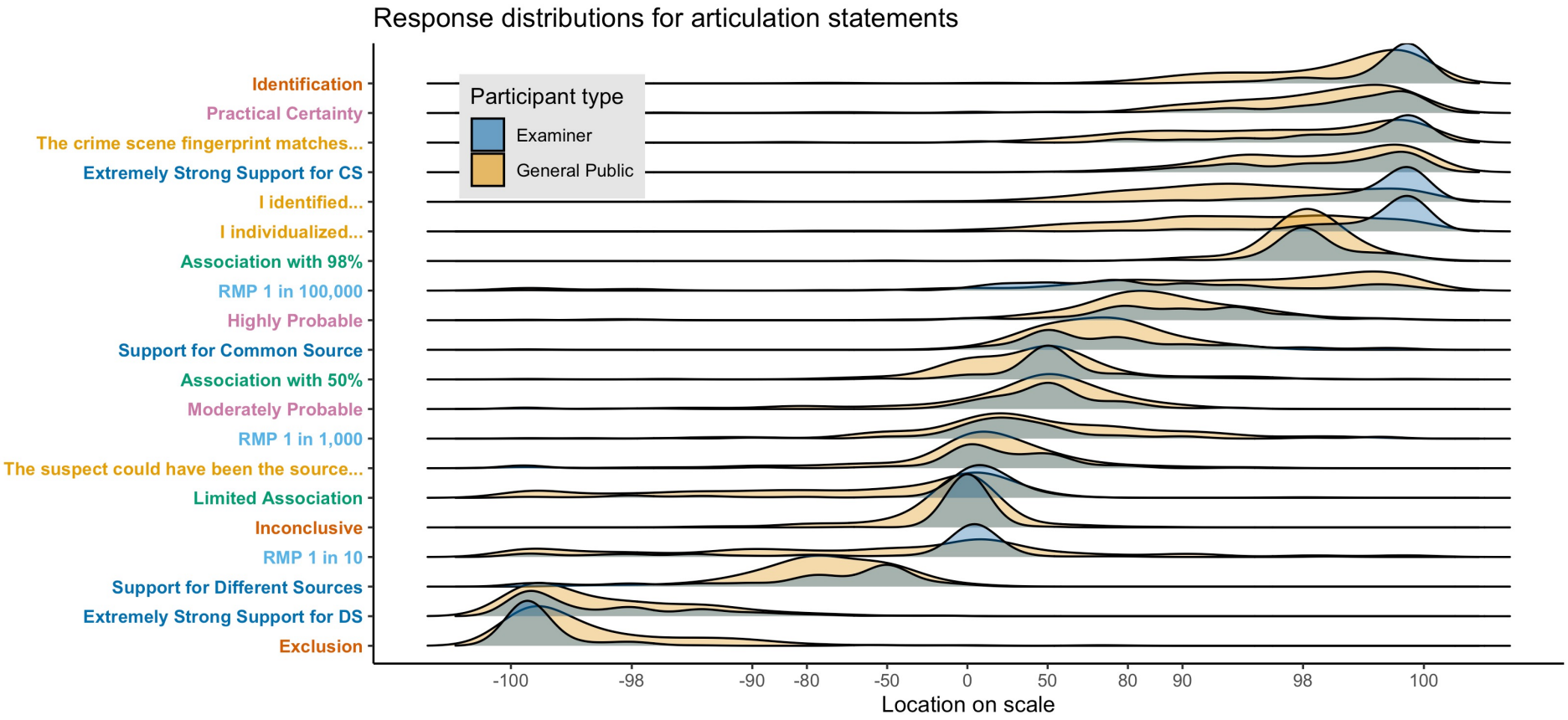
Data Analysis

- Raw scale values (visualized either transformed or untransformed)
- Use only ordinal relations and fit linear model similar to Thompson paper

Raw values (log scale)



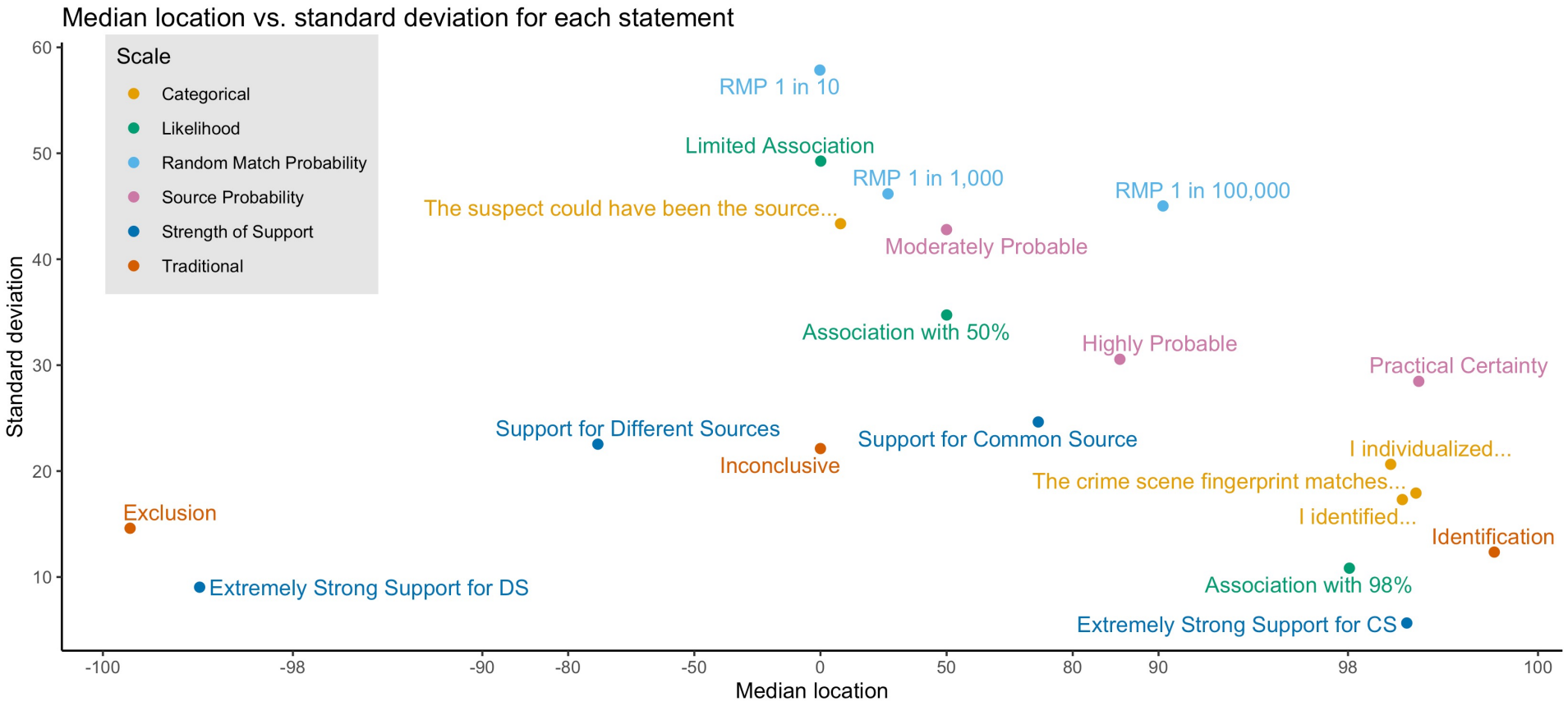
Raw values (log scale)



Summary of Raw Scores

- Not many differences between the Bloomington Community and Mturk participants
- Difficult to parse ridge plots because of variation among participants
- Look at variation for all subjects

Variation (All Participants)



Summary of Raw Scores

- RMPs have high variability (difficult for participants to interpret)
- USACIL language is interpreted literally: 98% is right at 98 on our scale with low variability. 50% is placed at 50 with somewhat higher variability.

Analysis on Ordinal Relations

- Different participants may interpret our numerical scale differently.
- What matters is *whether* a statement is placed above another, not by *how much*.
- Analyze the data using the Thurstone–Mosteller model used by Thompson et al.

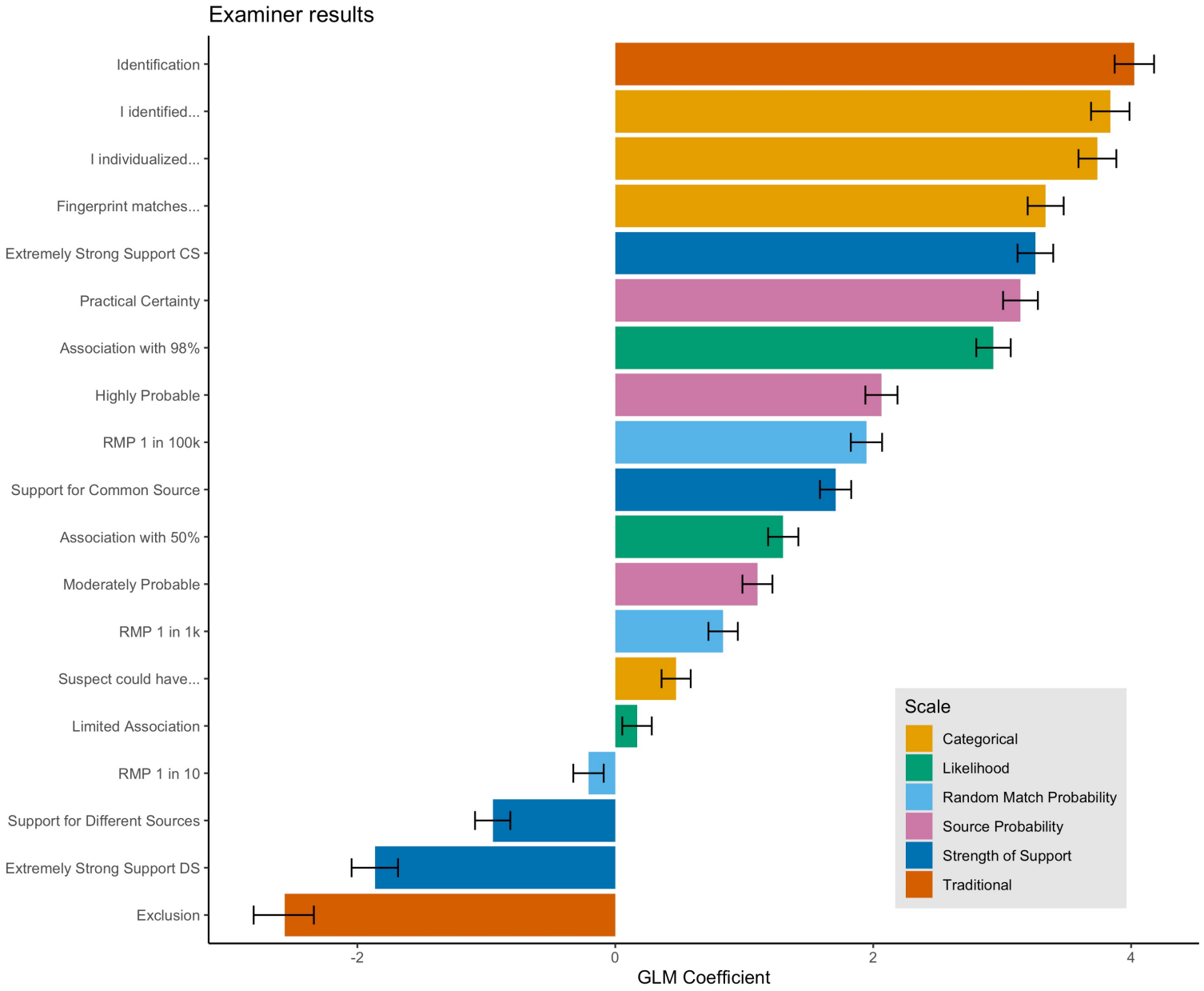
Dominance Matrix

- Create a Dominance Matrix across all participants in a group
- This computes the number of times each statement was placed above each other statement
- Diagonal is empty
- Identification dominated ESSCS 156 times
- ESSCS dominated Identification 73 times

Thurstone–Mosteller model

- Converts the ordinal relations to a ratio-scale response metric
- This model produces a parameter estimate for each statement that corresponds to the overall strength of evidence inferred from the dominance matrix for that statement
- Inconclusive is treated as the zero point

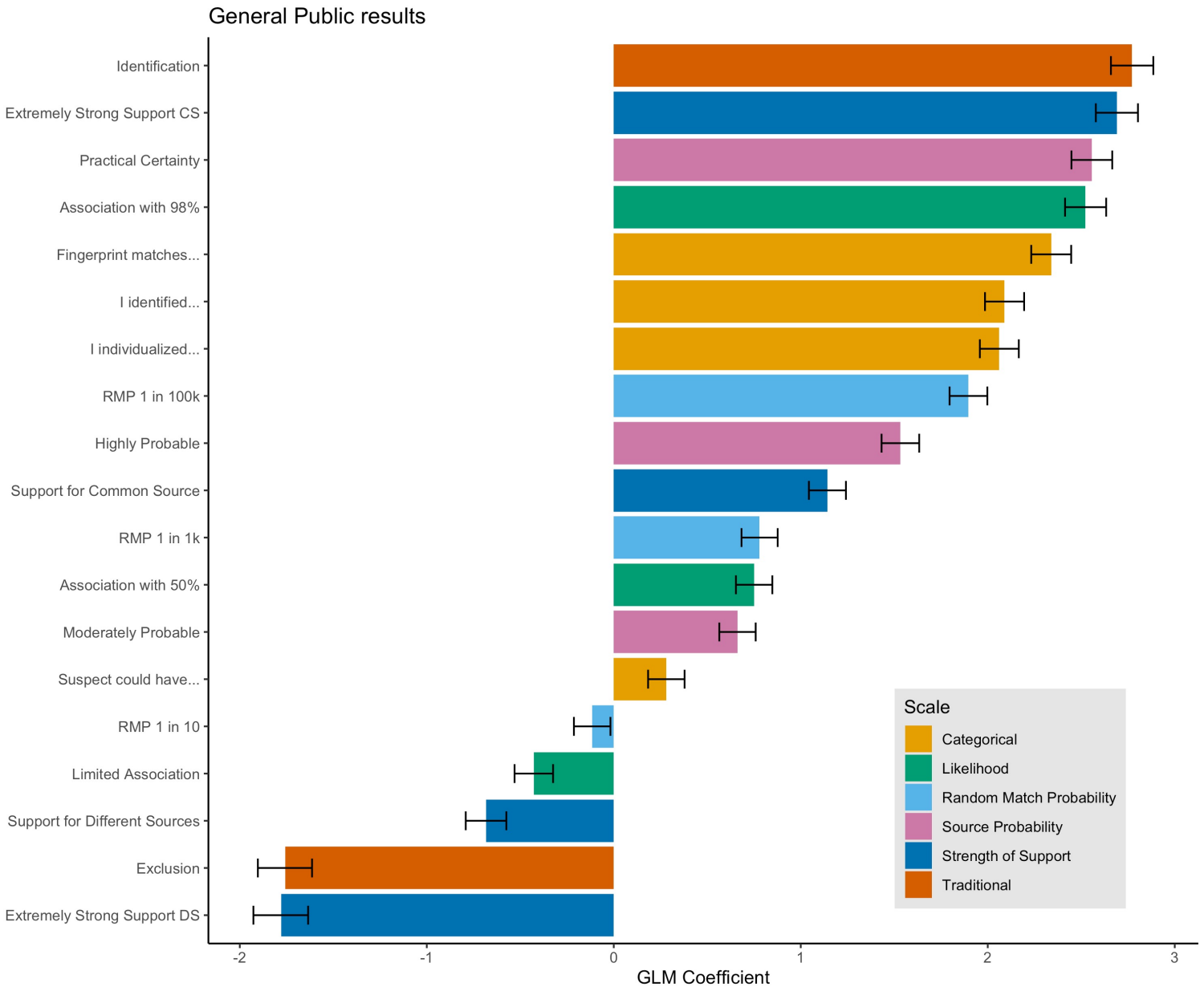
Ordinal-transformed Values



Examiner GLM Results

- Difference between Identification and Extremely Strong Support for Common Source
- Examiners make a distinction between these two statements- Identification is seen as providing more support for the common source proposition

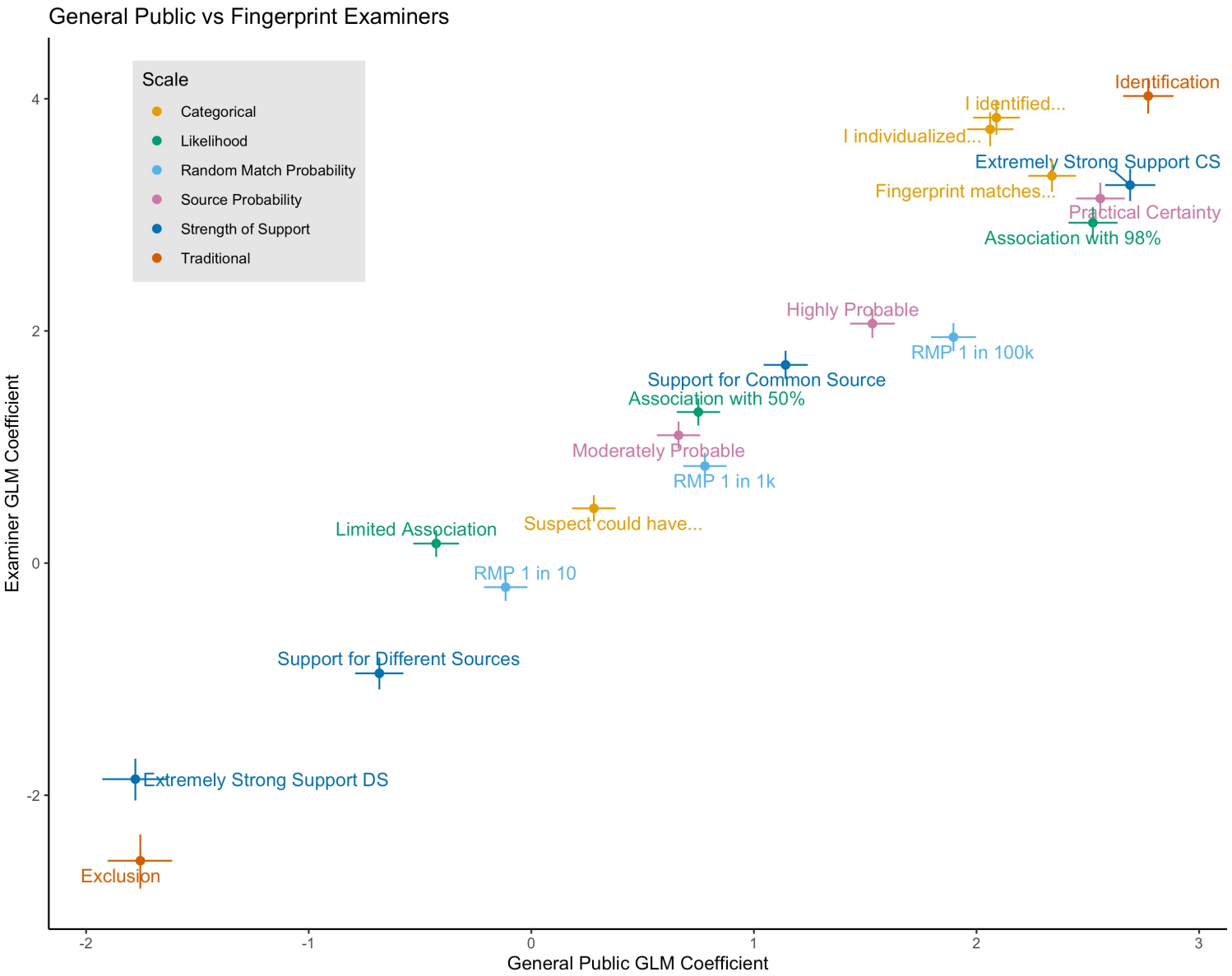
Ordinal-transformed Values



General Public GLM Results

- No difference between Identification and Extremely Strong Support for Common Source
- Members of the General Public do not distinguish between these two statements in terms of the support for the common source proposition

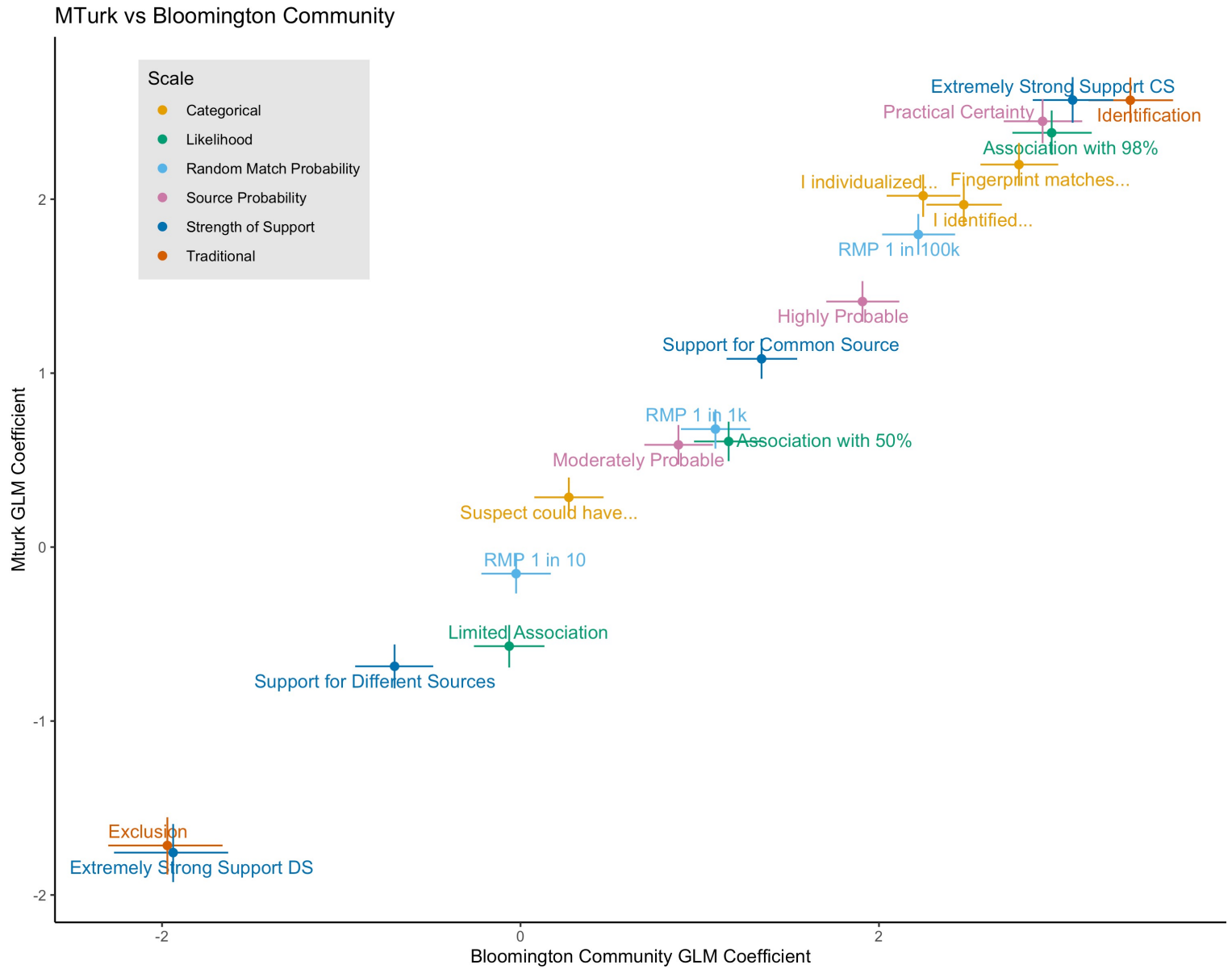
Ordinal-transformed Values



General Public GLM Results

- No difference between Identification and Extremely Strong Support for Common Source

Ordinal-transformed Values



General Public GLM Results

- No apparent differences between Mturk and Bloomington Community members

Summary of Results

- Experts distinguish between Identification and Extremely Strong Support for Common Source
- Members of the General Public do not

Summary of Results

- Experts viewed “Identification”, “I Identified”, and “I Individualized” as equivalent statements.
- Members of the General Public put “I Identified”, and “I Individualized” below “Identification”.
- Statements were part of different scales.
- “I Identified”, and “I Individualized”, and “Fingerprint Matches” were not required to be sorted.

Summary of Results

- Speculation: Member of the General Public used the sorting task to determine which item should be near 100 on our scale. If it was on top on the sorting task, it should be on the top of the scale.
- This would account for Extremely Strong Support being interpreted as equivalent to Identification.
- Exceptions: Statements with numerical values. RMP 1 in 100k is below the top cluster of statements for both groups.

DFSC/USACIL Language

- With our task, both sets of participants tended to place “Association with 98%/.1% statistical support” at a value of 98, and “Association with 50%/1% statistical support” at a value of 50.
- This may be an artifact of using a -100 to 100 scale.
- What scale do juries use during deliberations? How do they interpret these statements?

Suggestions

- Before replacing Identification with Extremely Strong Support for Common Source, ask how members of the general public interpret the new statement
- Consider the role of the examiner in the process. Do they provide evidence (support for propositions) or do they make conclusions? Who is the trier of fact?
- I will return to the DFSC/USACIL language in talk 2…

Thank You

Papers:

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

Visualization:

https://go.iu.edu/48KU

busey@iu.edu