Order effects in pairwise voice similarity and sameness evaluations

Thomas Kettig (York University)
Vincent Hughes, Carmen Llamas (University of York)

March 27, 2026

MOThQ 2026: Québec, QC

Order effects

- Participants should generally perform better later in experiments: learning is the expected outcome of repeated exposure (Gibson 1963, Norris, McQueen, & Cutler 2003, Clarke & Garrett 2004, Samuel & Kraljic 2009)
- This is why we randomize and/or counterbalance in our experiments!
- How do we explain a shift in responses that seems not to be driven by improvement through exposure?

Humans and Machines project

- Forensic interest in performance of lay listeners at speaker recognition
- Limited amount of work on recognition of unfamiliar voices
- *Humans and Machines* project aims:
  - To compare and combine human and automatic speaker recognition judgements and evaluate them on the same scales
  - To explore how human judgements are affected by cognitive biases related to criminal trials

Humans and Machines project

- Previous work has often elicited binary/Likert scale responses, which aren't comparable with e.g. automatic systems
- Likelihood ratios = established way to evaluate forensic evidence
  - Requires responses in terms of both **similarity** and **typicality**

- Here, we elicit data through a bespoke game-like tool where participants are immersed in a 'jury of the future' context

Method

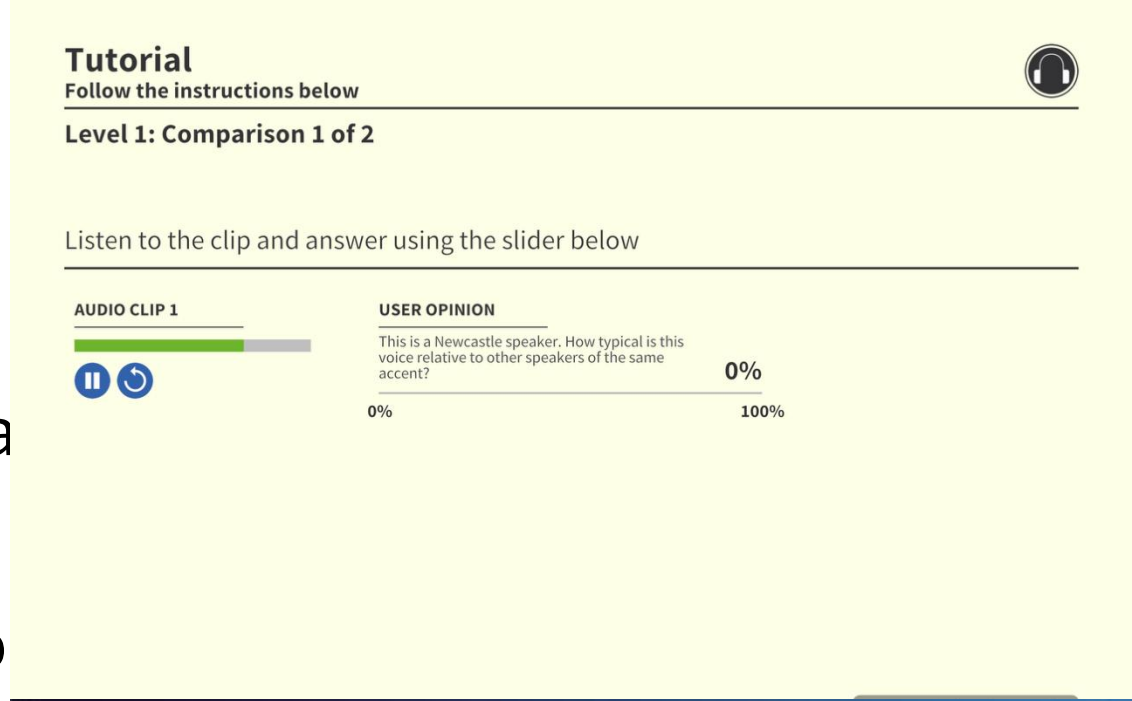
- Pairs of voices presented, listeners judged:
  - **Typicality** of the ‘offender’ voice
  - **Similarity** of the two voices
  - Whether they thought the speakers were the **same** or not
- 0–100 scale for each response
- Two conditions
  - Immersive jury game (n = 1,505)
  - Qualtrics (n = 301)

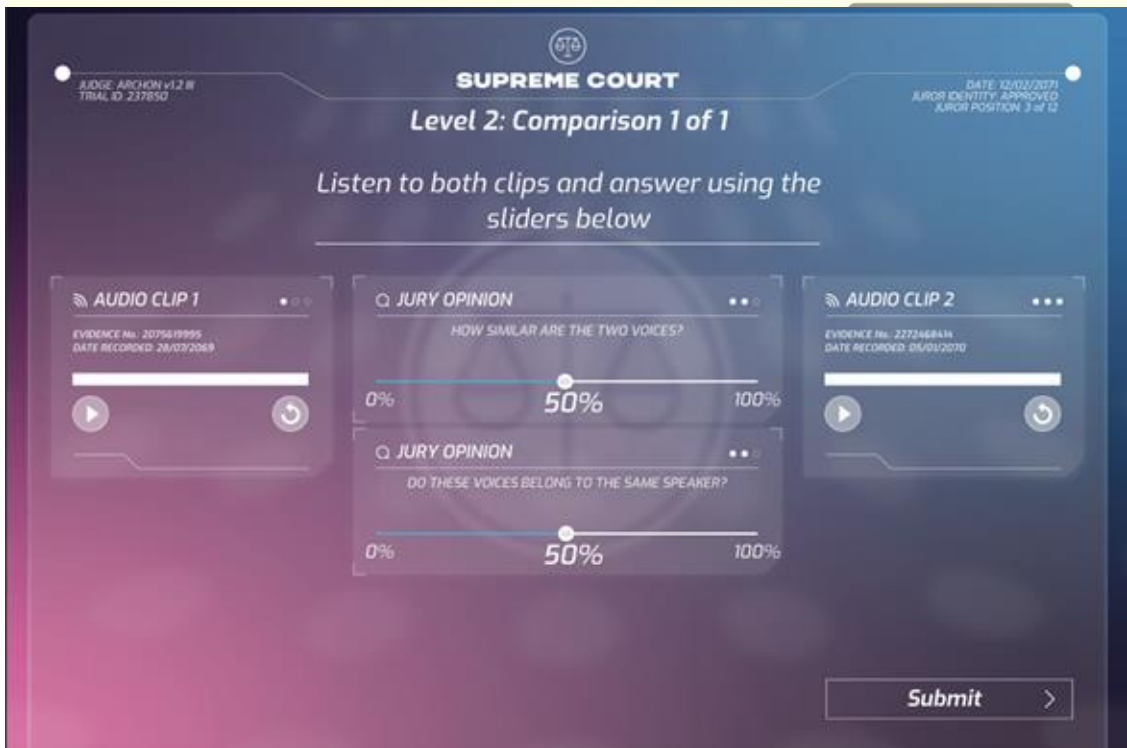
Method: Immersive jury ga

- Level 1: Tutorial (Qualtrics style) (8 pa
- Level 2: Immersive jury level (8 pairs)
- Level 3: Immersive jury level + additio evidence (8 pairs)

  or....

- Level 3: Immersive jury level + expert conclusion (8 pairs)





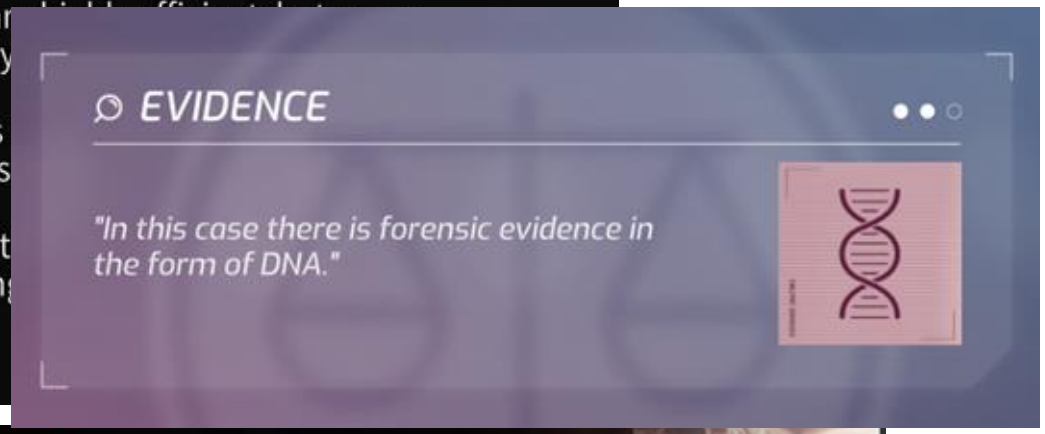
JURY OF THE FUTURE

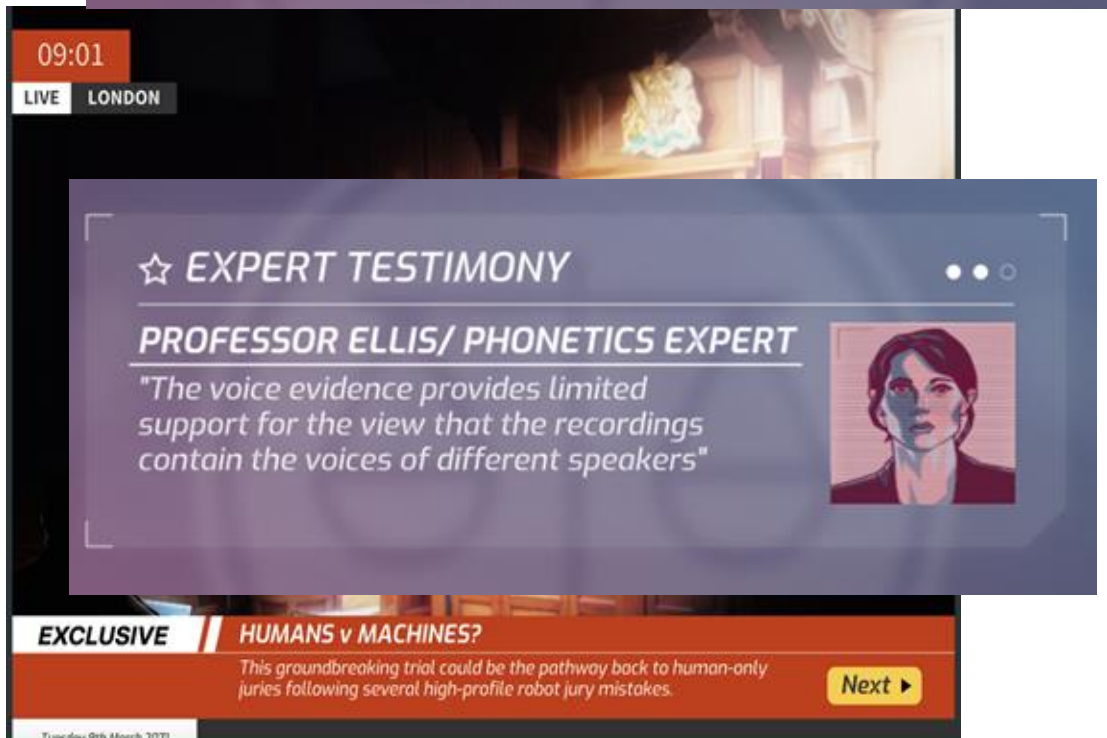
It's 2071. The justice system is now delivered by machines.

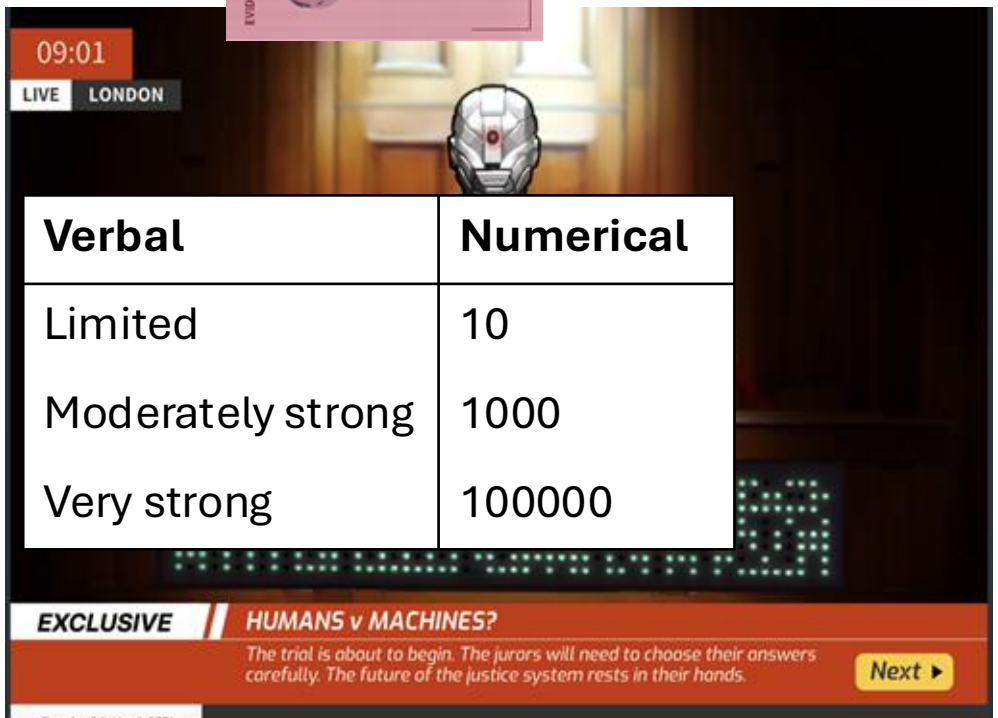
The robots ar
have recently

A new trial is
human juries

Can you beat
justice belon



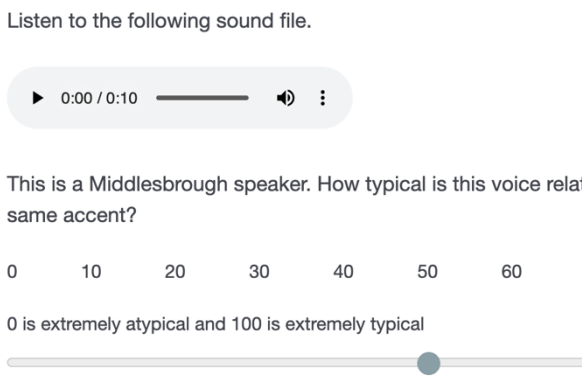


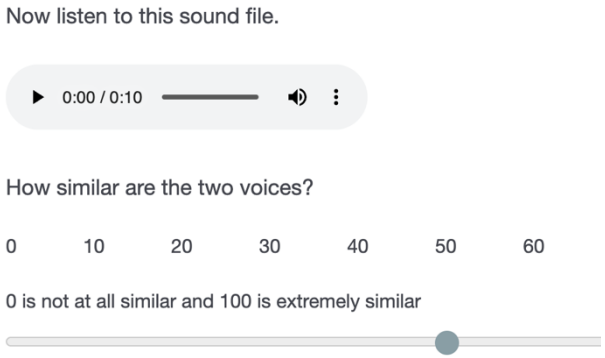


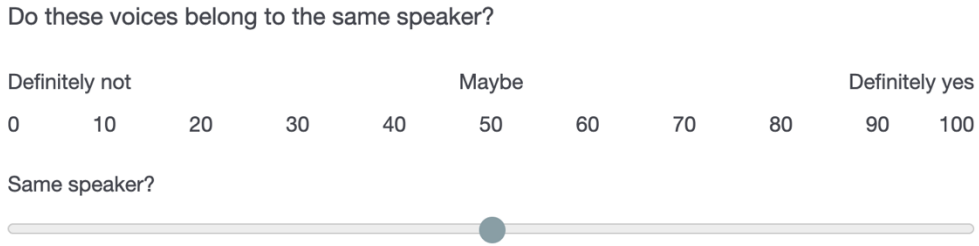
Verbal	Numerical
Limited	10
Moderately strong	1000
Very strong	100000

Method: Qualtrics

- Level 1: Boring Qualtrics interface (8 pairs)
- Level 2: Boring Qualtrics interface (8 pairs)
- Level 3: Boring Qualtrics interface (8 pairs)







Stimuli

- Samples from **Standard Southern British English**, **Newcastle** and **Middlesbrough**
- All voices from male speakers
- Forensically-realistic quality
  - First sample = landline phone quality (actual or noise/filter added)
  - Second sample = high quality, taken from mock police interviews
  - Short (10-11 seconds)
- Normed for guilt/suspiciousness of sample content

ap 1. The North East of England (from Buchstaller et a
11:3, based on two outline images: UK and Ireland

Presentation of pairs

- 120 pairs created
  - 30 SSBE pairs (15 DS, 15 SS)
  - 30 Middlesbrough pairs (15 DS, 15 SS)
  - 30 Newcastle pairs (15 DS, 15 SS)
  - 30 mixed Middlesbrough/Newcastle pairs (30 DS)
- Distributed into 15 blocks containing 8 pairs each (5 DS, 3 SS)
- Stimuli within blocks internally randomized
- 1 block presented per level, counterbalanced
- Feedback provided after each level
  - “Good job! You got 7 of the 8 correct”
  - Not possible to improve based on pair-by-pair feedback

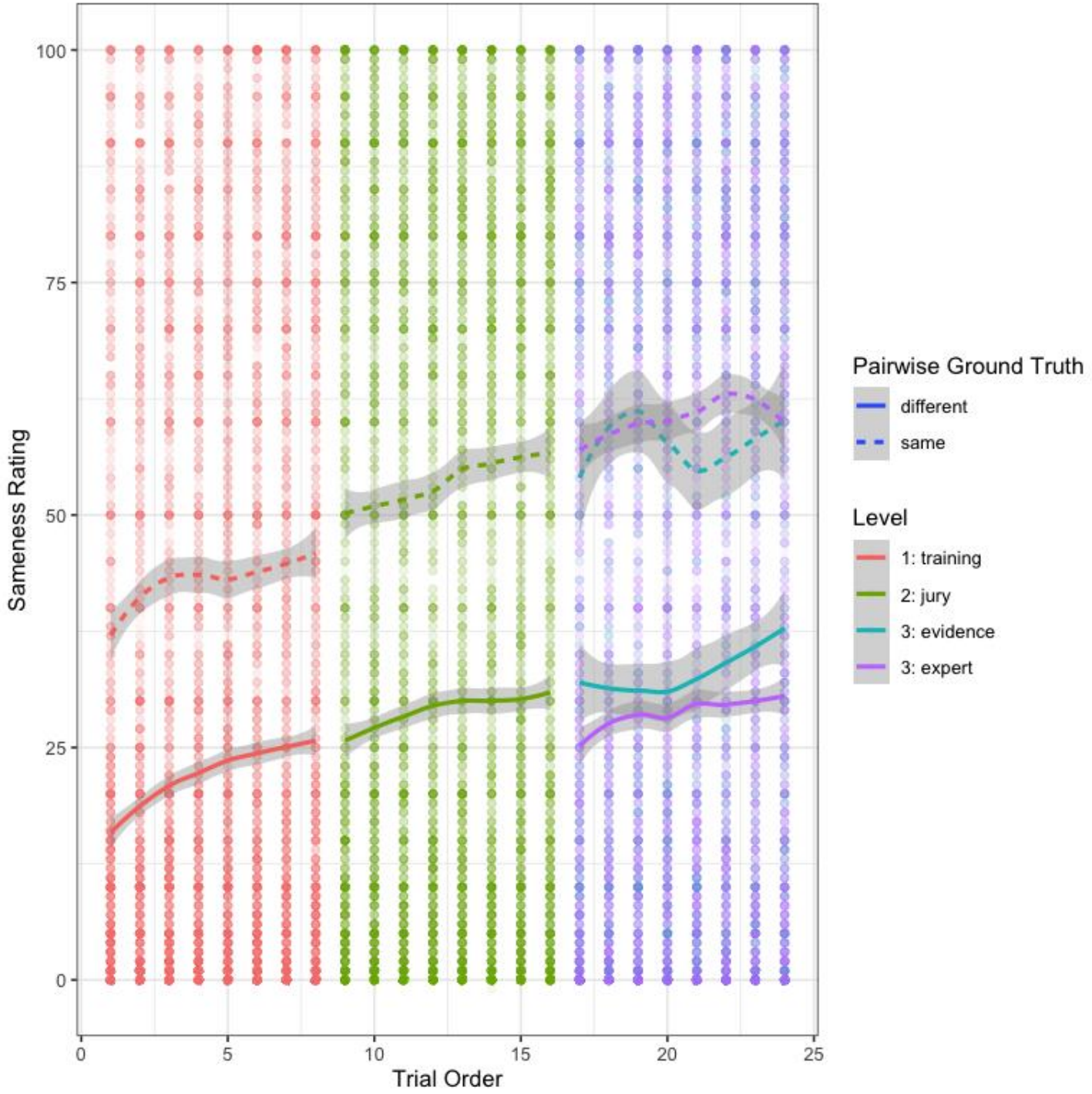
Order effect

- **Sameness** in **game**

SameAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, different-speaker

<u>Fixed effects</u>
Order β = 1.31, p < 0.001
Order:GroundTruthSame *n.s.*
Order:Level2 β = -0.56, p = 0.003
Order:Level3evidence β = -0.74, p = 0.02
Order:Level3expert β = -0.59, p = 0.003
Order 3-way interactions *all n.s.*



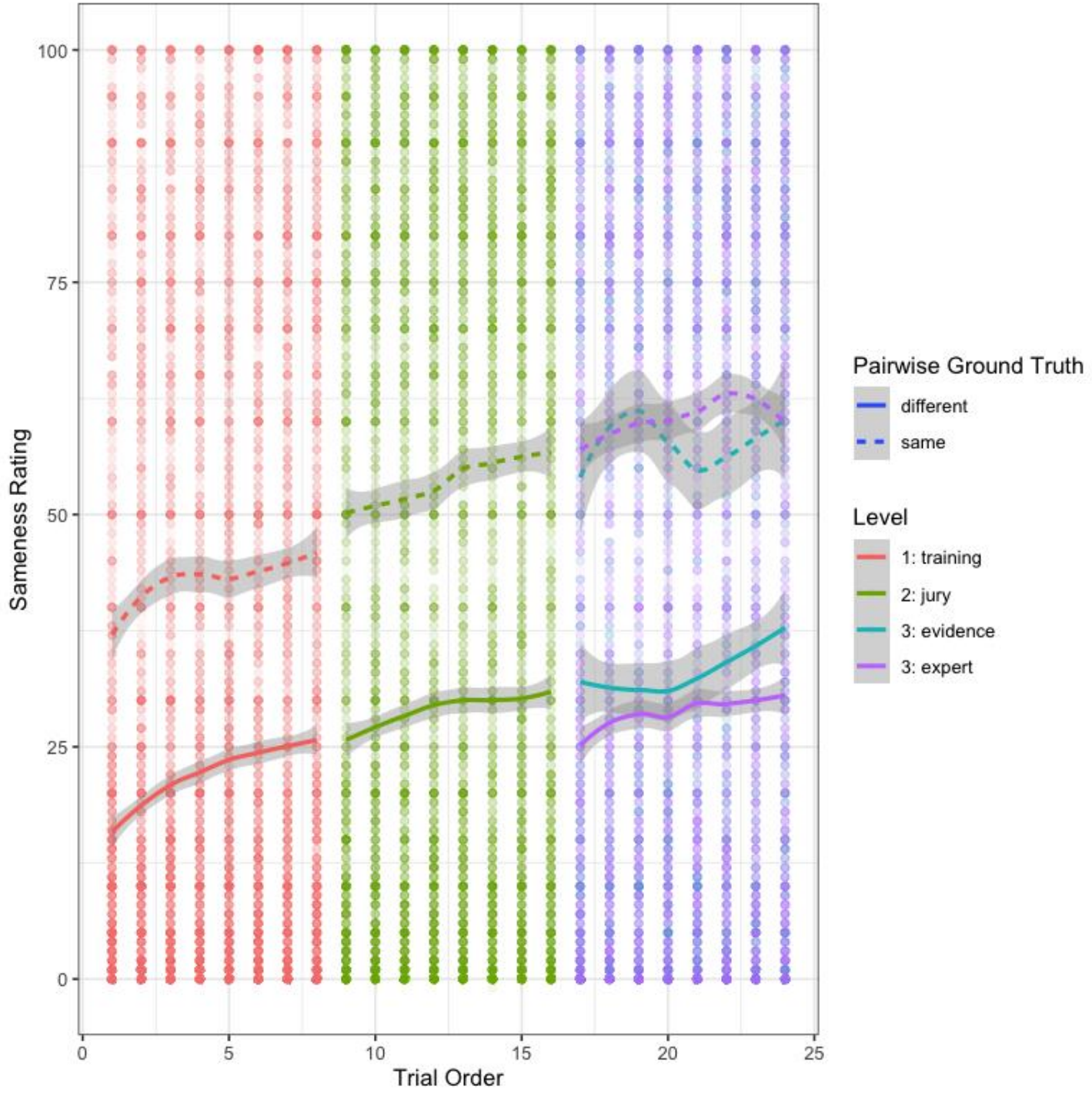
Order effect

- **Sameness** in **game**

SameAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, same-speaker

<u>Fixed effects</u>
Order β = 0.91, p < 0.001
Order:GroundTruthSame *n.s.*
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert β = -0.16, p = 0.002
Order 3-way interactions *all n.s.*



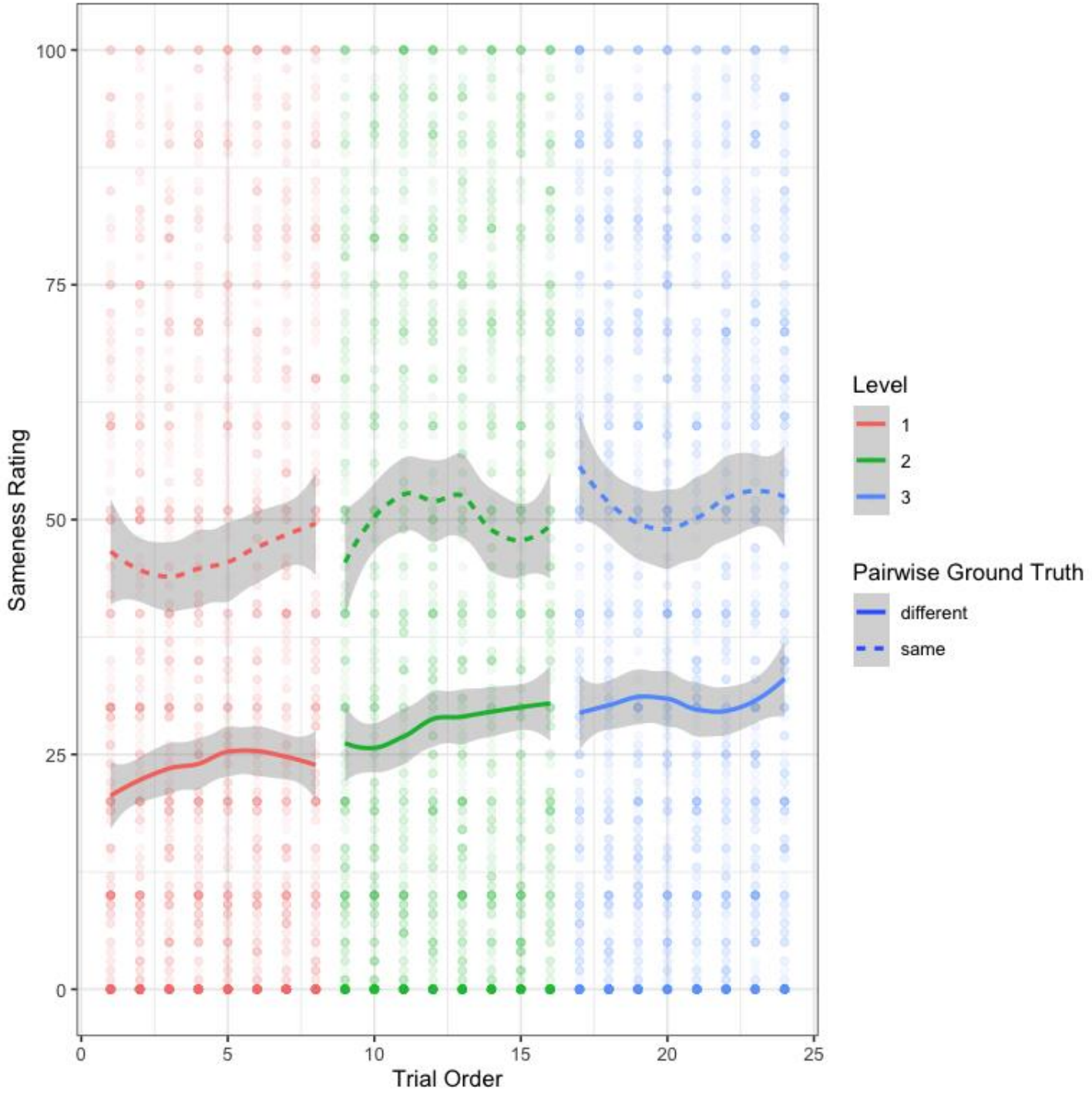
Order effect

- **Sameness** in **Qualtrics**

SameAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, different-speaker

<u>Fixed effects</u>
Order *n.s.*
Order:GroundTruthSame *n.s.*
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*



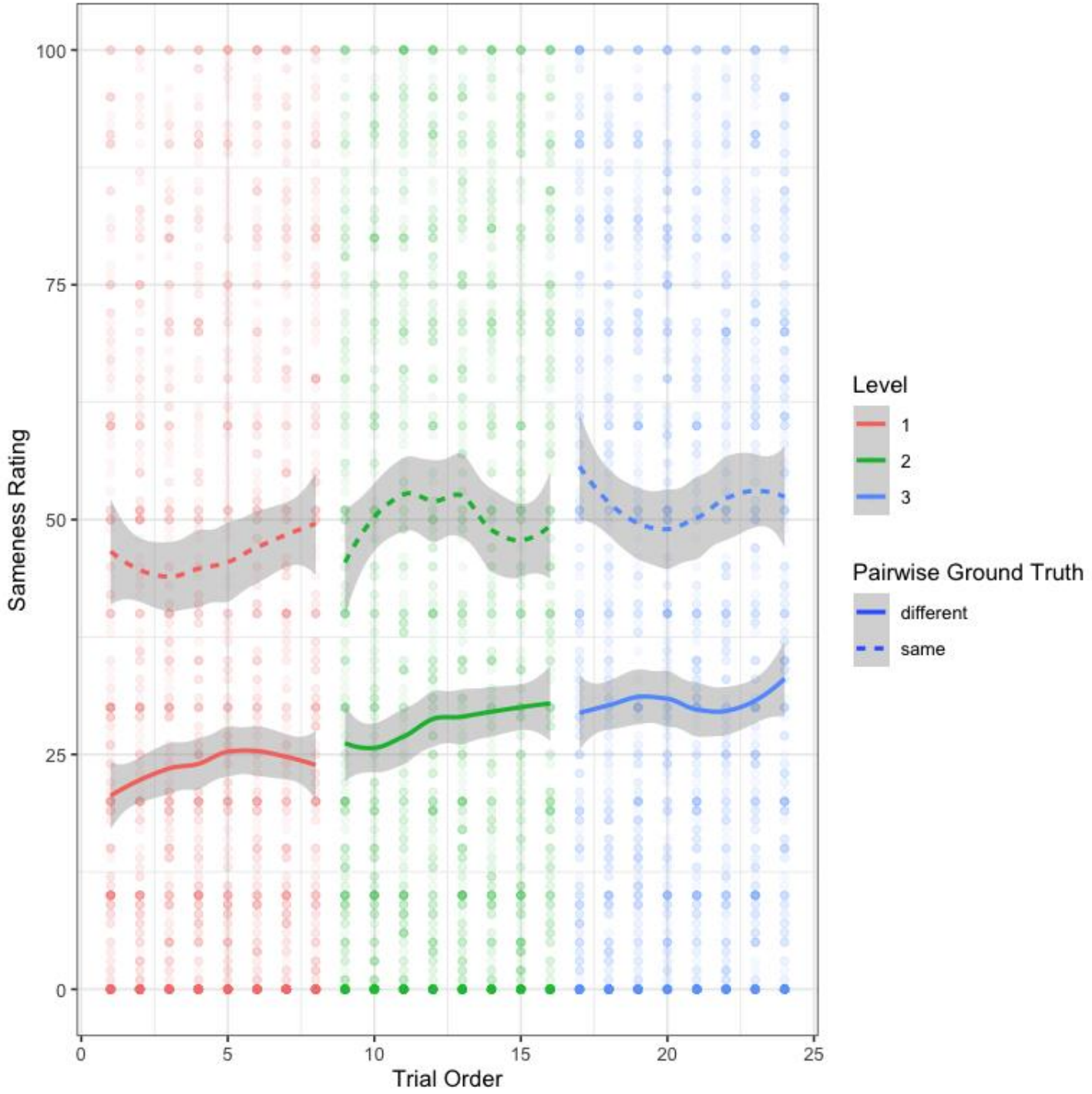
Order effect

- **Sameness** in **Qualtrics**

SameAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, same-speaker

<u>Fixed effects</u>
Order *n.s.*
Order:GroundTruthSame *n.s.*
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*



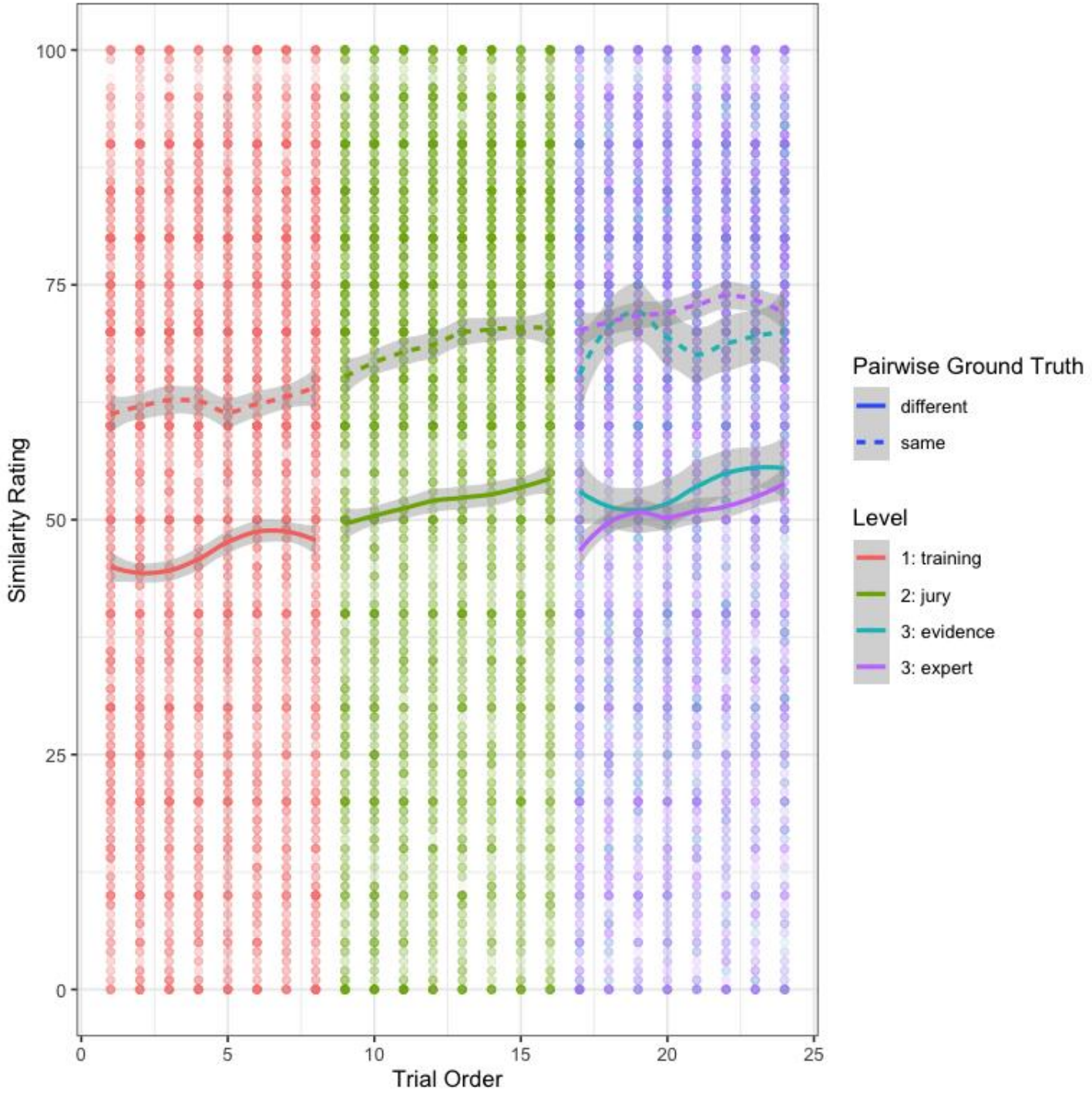
Order effect

- **Similarity** in **game**

SimilarityAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, different-speaker

<u>Fixed effects</u>
Order β = 0.69, p < 0.001
Order:GroundTruthSame β = -0.49, p = 0.01
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*



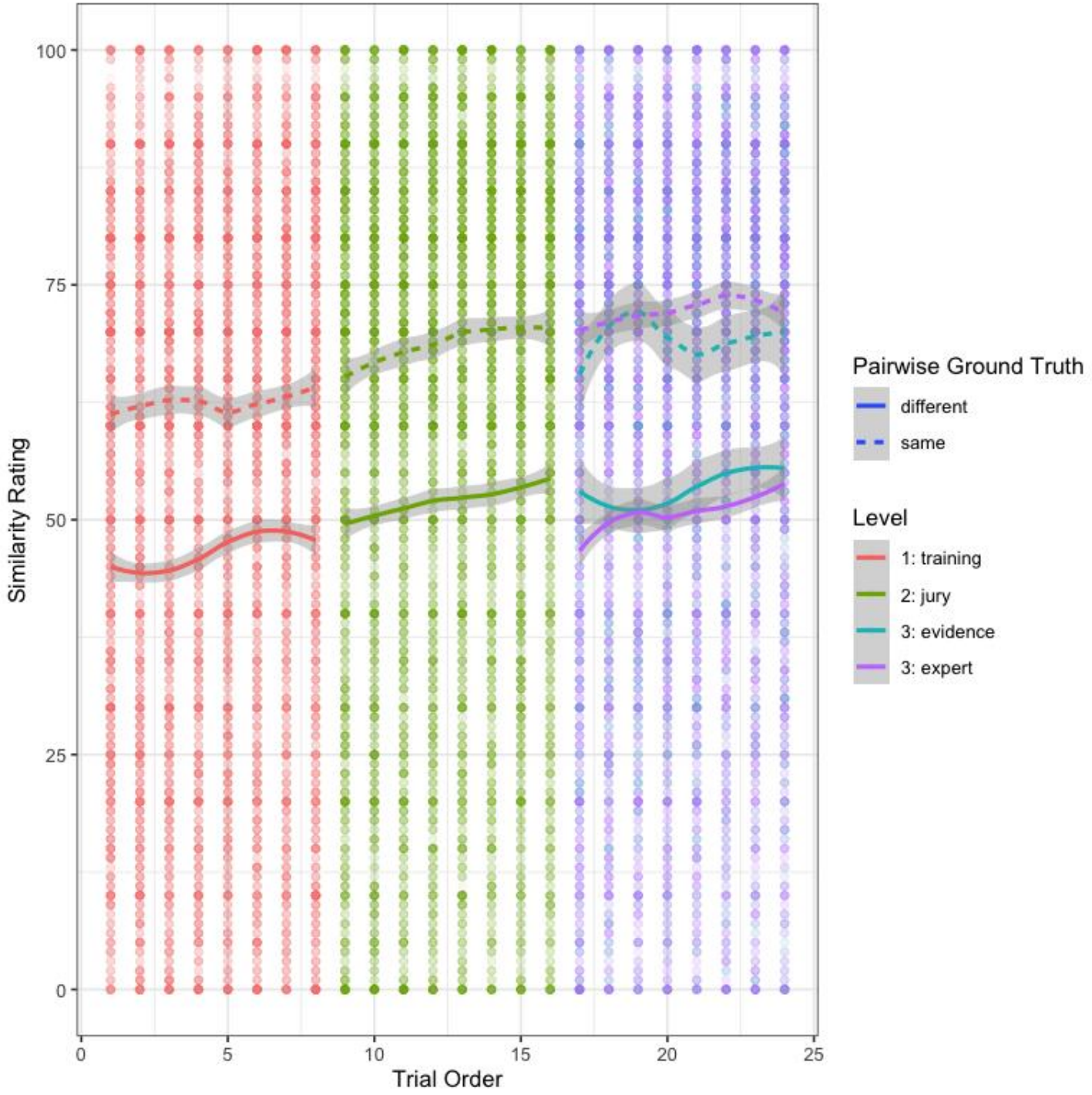
Order effect

- **Similarity** in **game**

SimilarityAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, same-speaker

<u>Fixed effects</u>
Order *n.s.*
Order:GroundTruthSame β = -0.49, p = 0.01
Order:Level2 β = 0.43, p = 0.04
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*



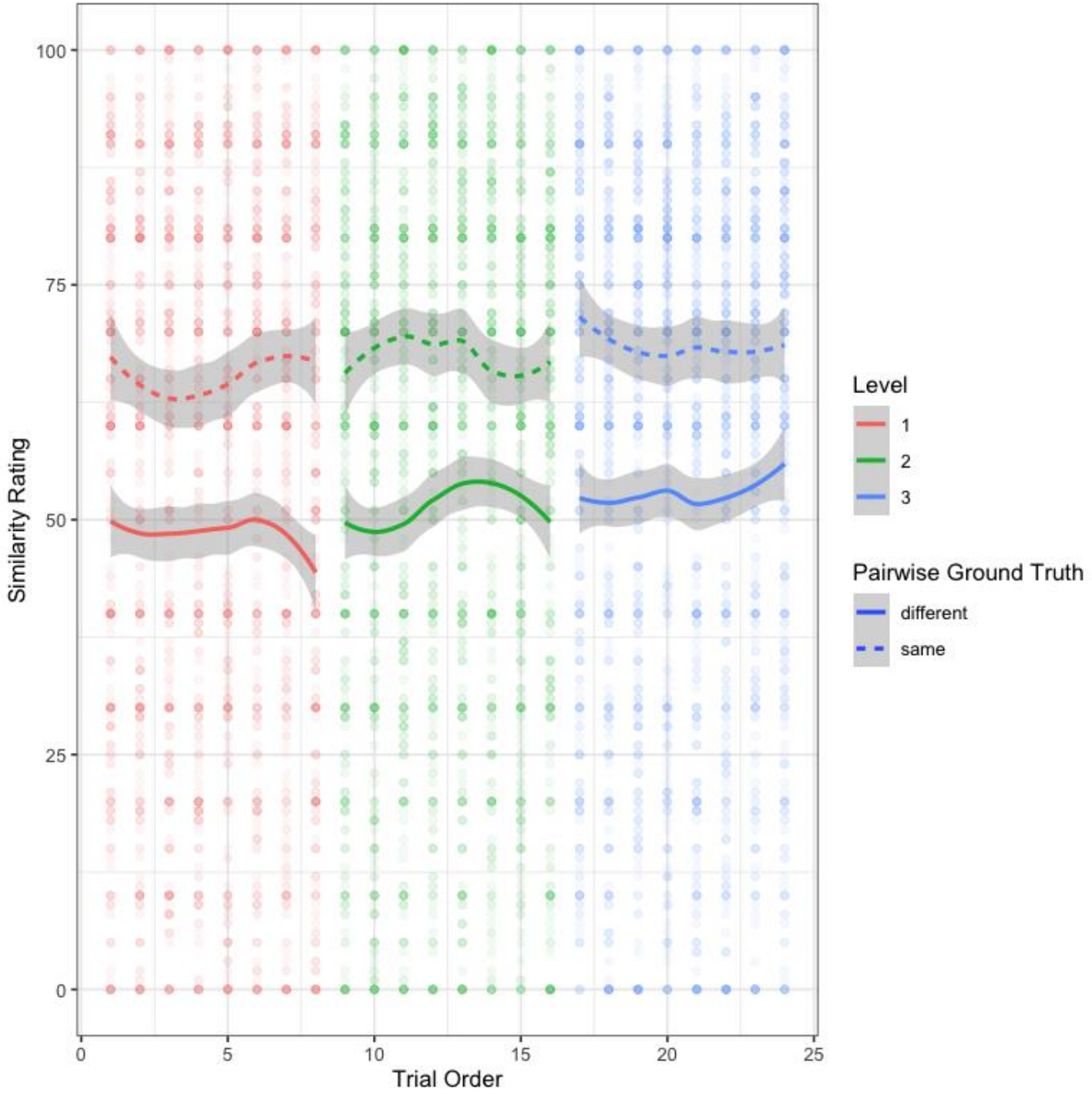
Order effect

- **Similarity** in **Qualtrics**

SimilarityAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, different-speaker

<u>Fixed effects</u>
Order *n.s.*
Order:GroundTruthSame *n.s.*
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*



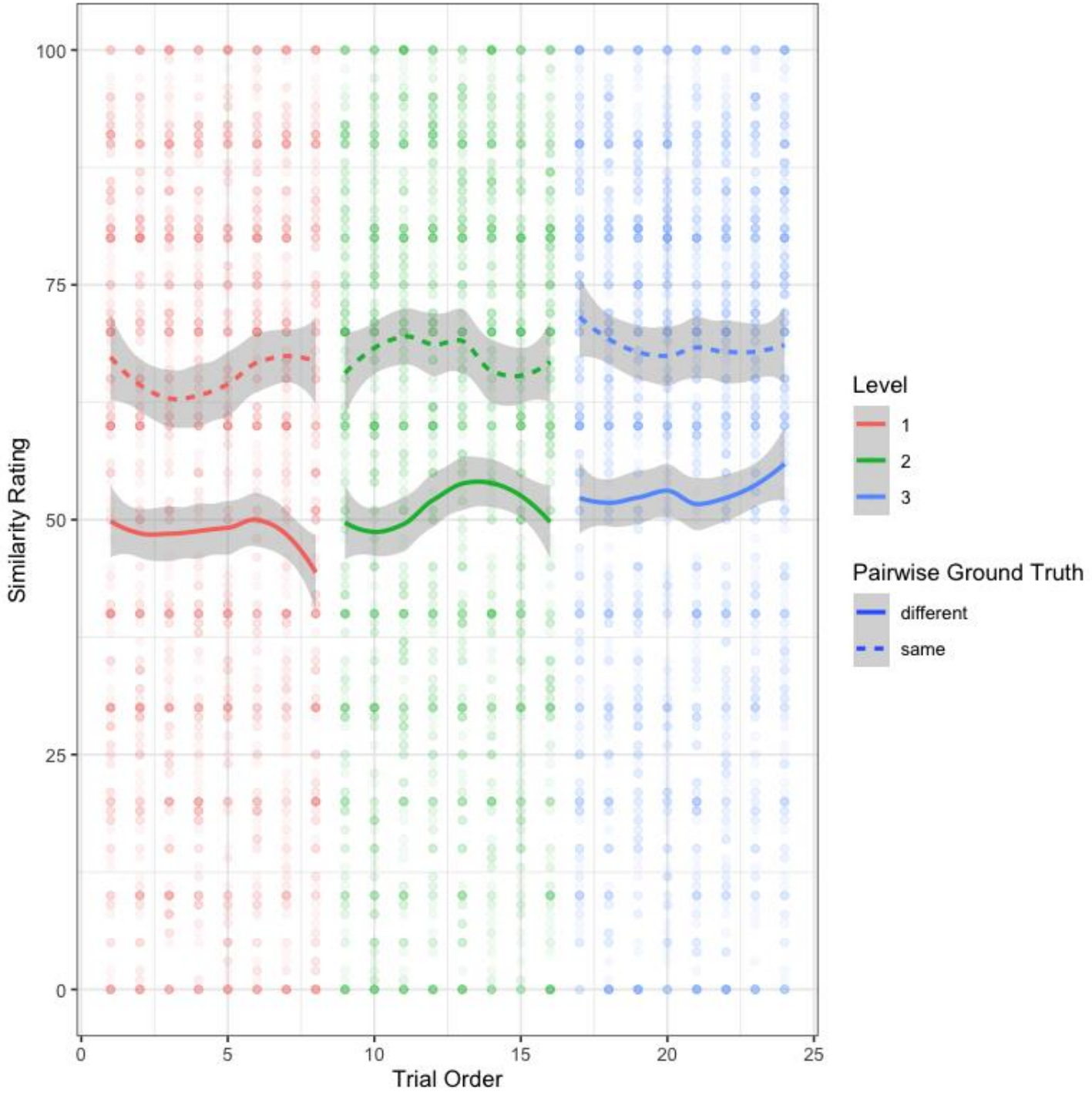
Order effect

- **Similarity** in **Qualtrics**

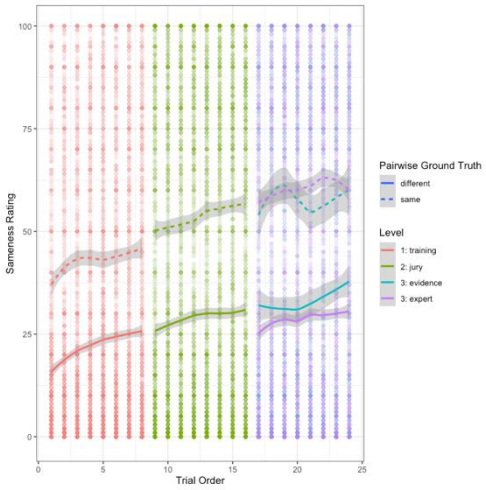
SimilarityAnswer ~ Order24 * GroundTruth * Level + (1|ParticipantID) + (1|PairName)

Reference levels: Level 1, same-speaker

<u>Fixed effects</u>
Order *n.s.*
Order:GroundTruthSame *n.s.*
Order:Level2 *n.s.*
Order:Level3evidence *n.s.*
Order:Level3expert *n.s.*
Order 3-way interactions *all n.s.*

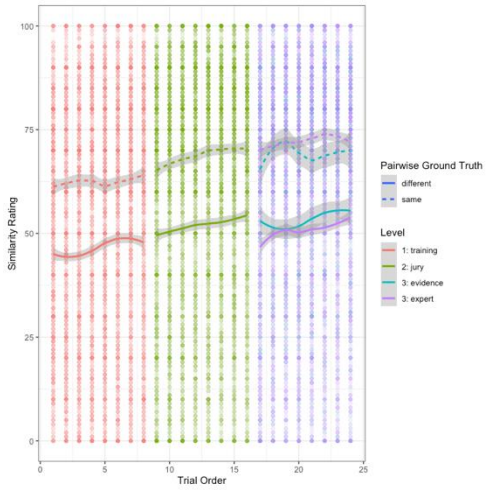


Order effects summary



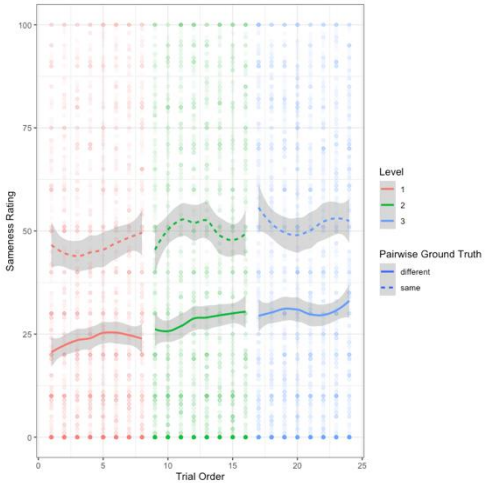
Game, sameness

- β = 1.31 for DS pairs L1
- β = 0.91 for SS pairs L1
- Weaker effect in higher Ls (mainly for DS pairs)



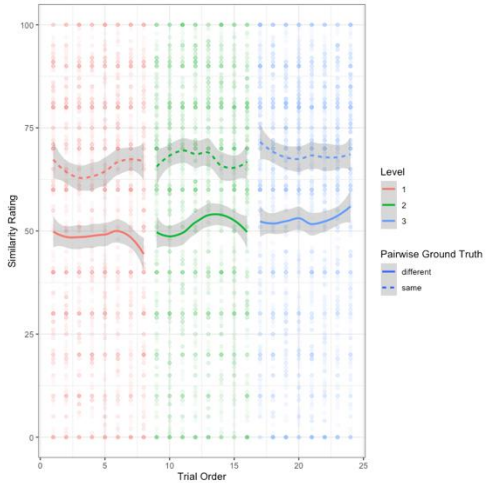
Game, similarity

- β = 0.69 for DS pairs L1
- sig. order effect only in L2 for SS pairs



Qualtrics, sameness

- No order effects

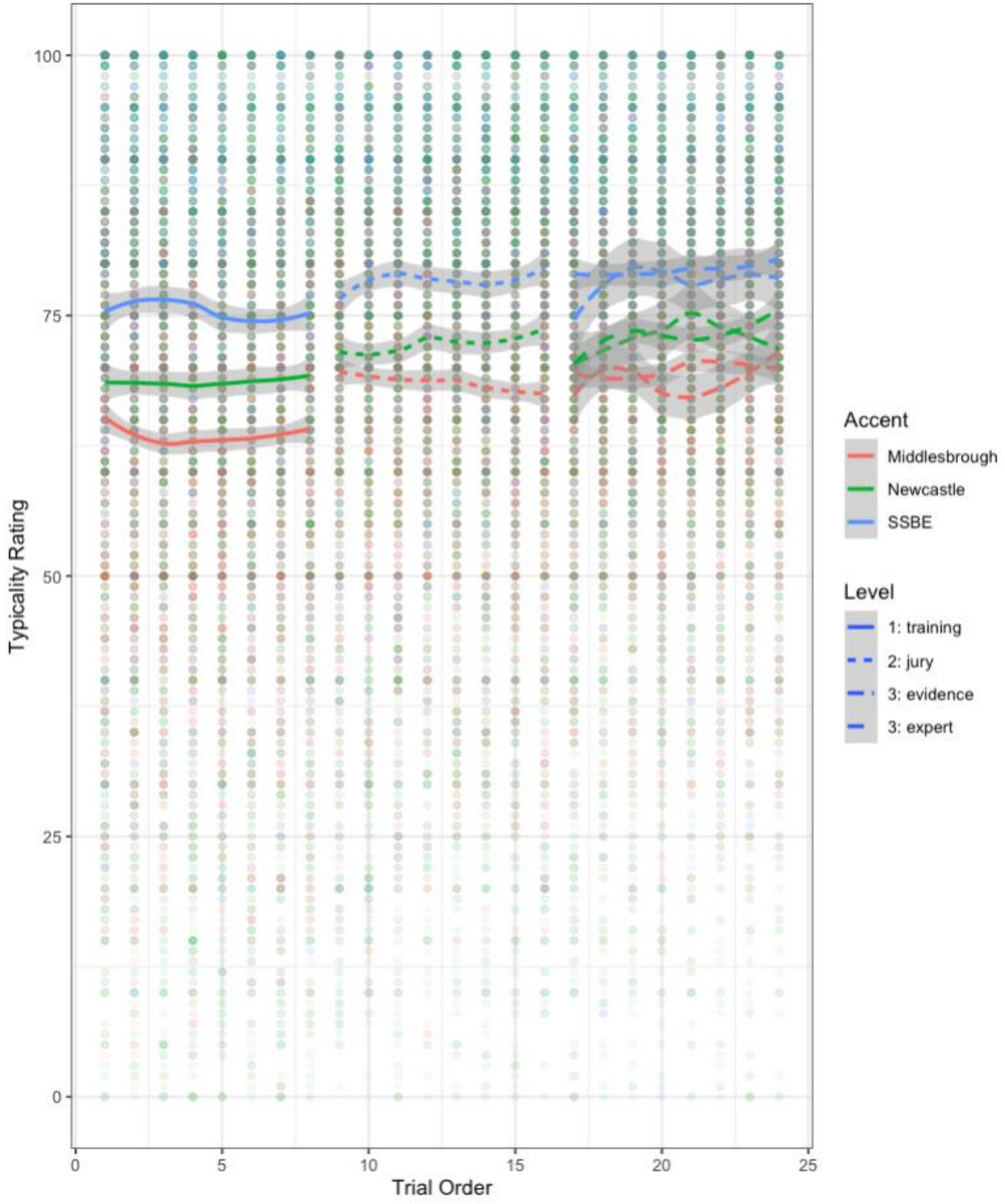


Qualtrics, similarity

- No order effects

And with typicality

- In particular, Middlesbrough voices seem to be rated more typical after Level 1
- Whatever order effects exist don't reach the same magnitude as the sameness order effect



Discussion

- Typicality effect for Middlesbrough could be a learning effect – over the course of an experiment people become more exposed to the accent, so it sounds increasingly typical
  - Middlesbrough is fairly obscure compared to other accents, less familiarity to begin with
- But for sameness/similarity, SS/DS pairs help us disentangle response bias/scale drift from learning effect
  - Because we'd expect SS to get higher, DS to get lower, if learning occuring

Discussion

- Order effect seems to occur in game, but not Qualtrics
  - Both visual analogue scales, so why is the order effect only in the game?
  - Higher engagement?
- Pressure to find speakers guilty?
  - But we have strongest order effect in Level 1, which has no jury context!
- Bartle & Dellwo (2015): When unsure, phoneticians tend to say two utterances come from different speakers, whereas naive listeners tend to say two utterances come from the same speaker
  - But start out biased towards hearing as different, only later shift to right of scale
- Ran more listeners in game – should we run hundreds more in Qualtrics to see if we end up with similar effect?

Discussion

- Order effect seems strongest in DS pairs over SS pairs
  - This indicates that listeners are getting *worse* at the comparison!
  - If they were learning/improving, we'd expect *negative* order effect for DS pairs, not the positive effect we observe
- Krueger & Chignell (1985): According to the **missing-feature principle**, distinguishing features are not fully processed at the beginning of perception, so differences are often missed, leading to false "same" responses. As more processing time becomes available, these features are resolved and discrimination improves, reducing this bias
  - So… opposite of what we find
- Still searching for answers…

Thank you!

Questions? Feedback?