UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

EUROSLA, 28–31 August 2019

ESTIMATES OF NATIVELIKENESS AMONG L2 SPEAKERS CAN'T BE INTERPRETED: THE PROBLEM AND TWO SOLUTIONS

Jan Vanhove
University of Fribourg, Switzerland

jan.vanhove@unifr.ch
janhove.github.io
@janhove

Slides, references, code, and additional material:
https://janhove.github.io/nativelike

NATIVELIKENESS AMONG L2 SPEAKERS

- Who, if anyone, can ultimately become nativelike in L2?

    ~ critical periods (e.g., Birdsong 2005; Long 2005)

- Lots of empirical studies: How many of a sample of L2 speakers perform similarly to L1 speakers on one or several tasks?
- Criticism of "nativelikeness"
    - necessary? useful?

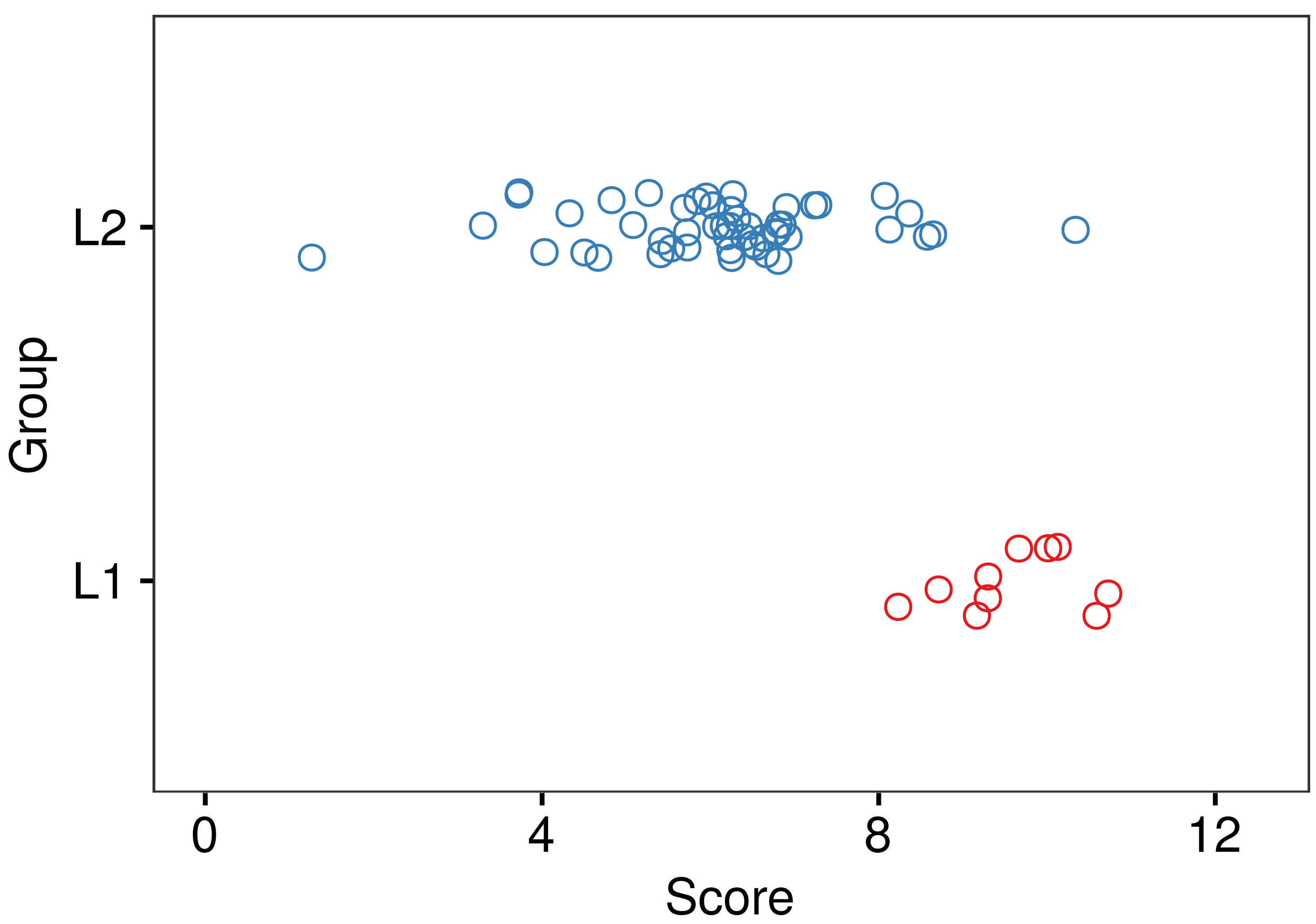
      (e.g., Birdsong & Gertken 2013, Cook 1992, Davies 2003, Grosjean 1989, Ortega 2013)
    - appropriate samples? (Andringa 2014, see also Dąbrowska 2012)
- Today: statistical problem

jan.vanhove@unifr.ch | janhove.github.io | @janhove

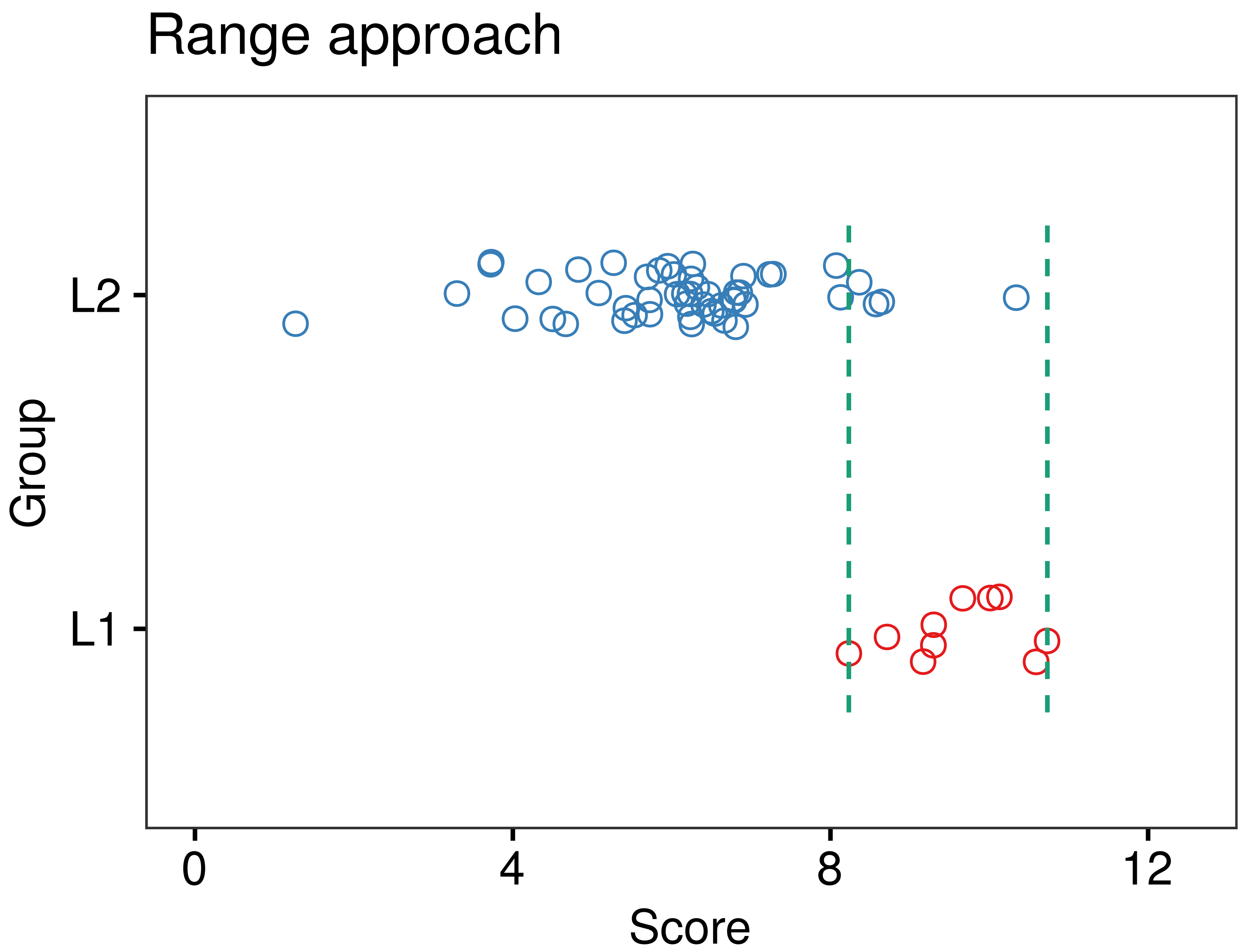
Part I

The problem

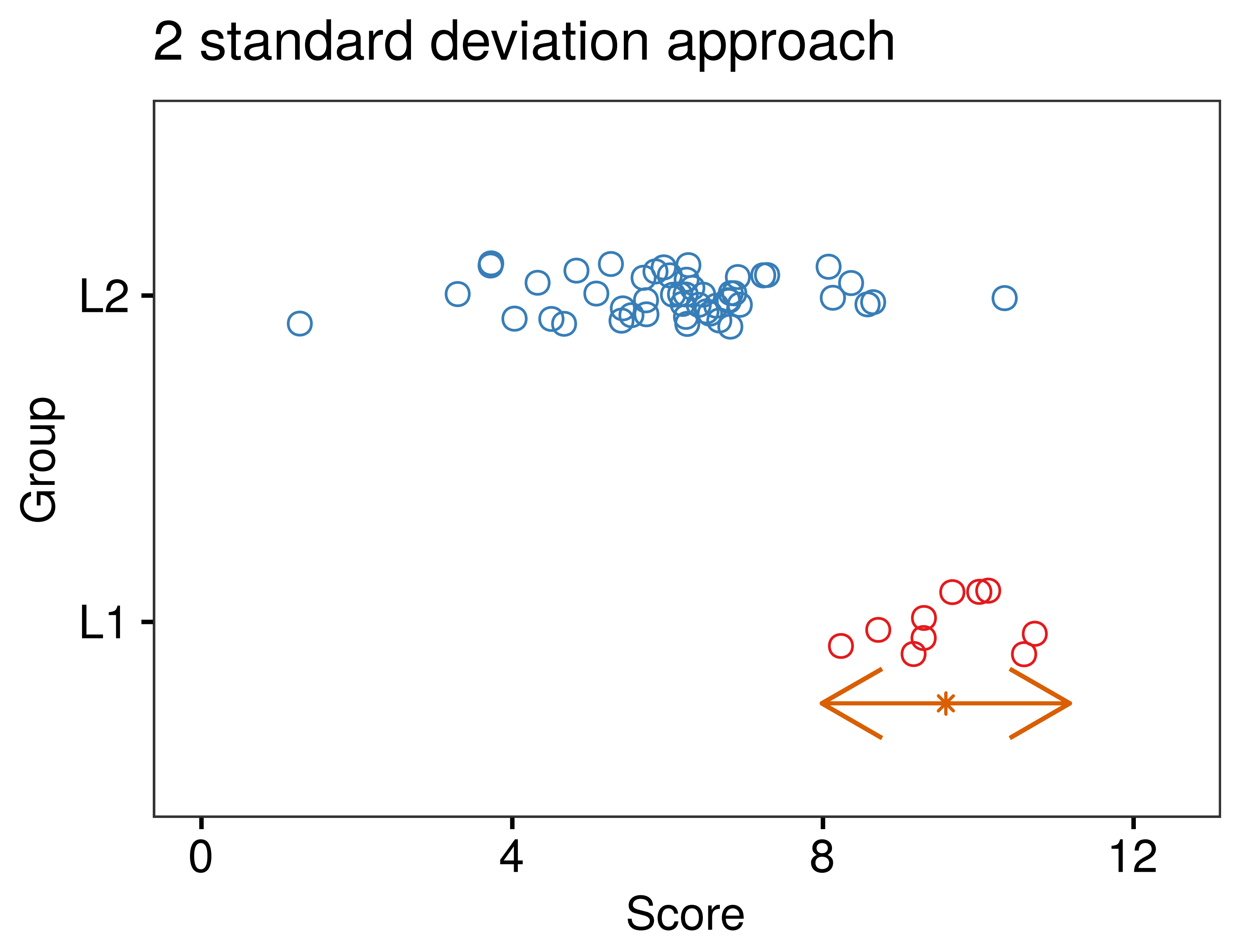
ASSESSING NATIVELIKENESS

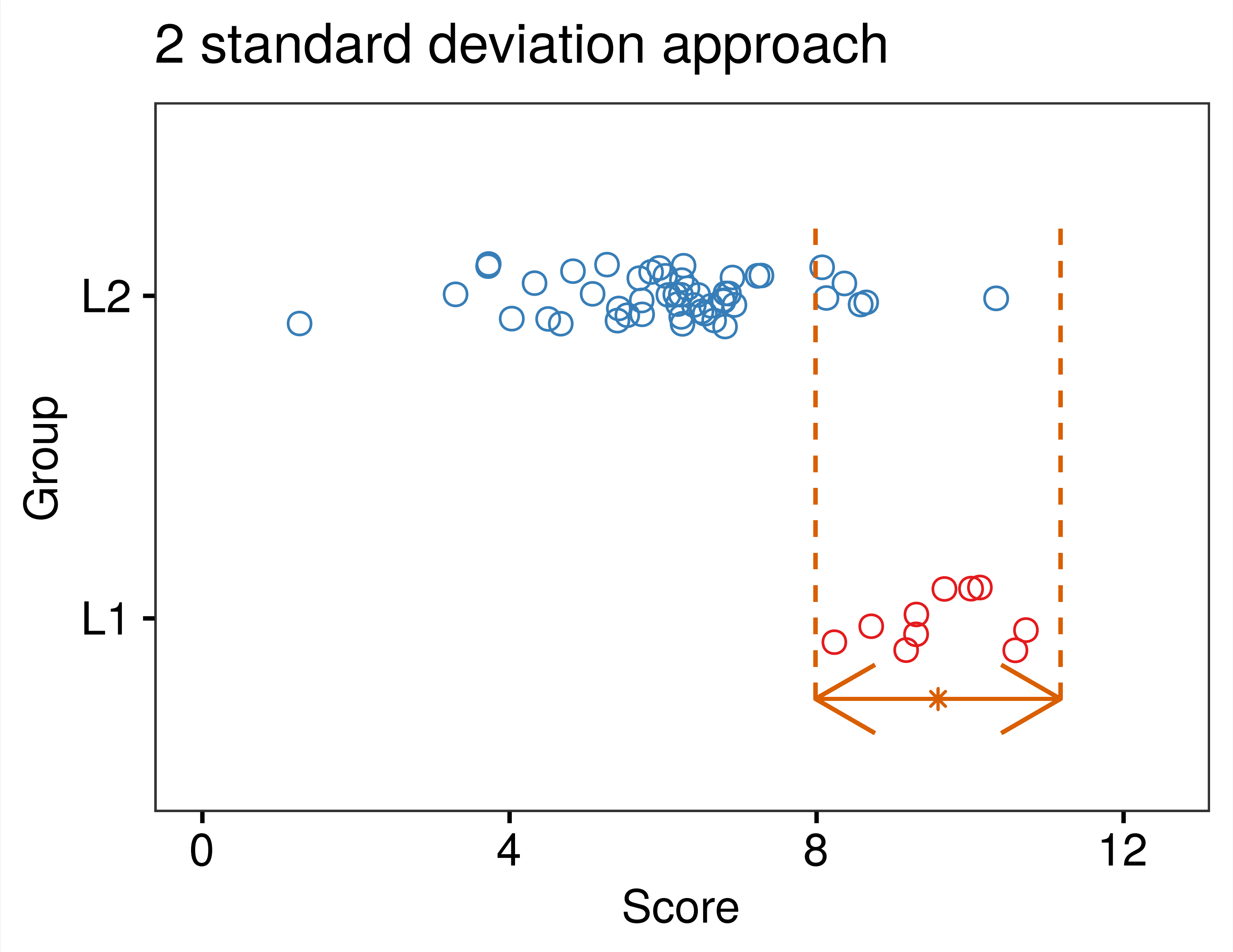


ASSESSING NATIVELIKENESS

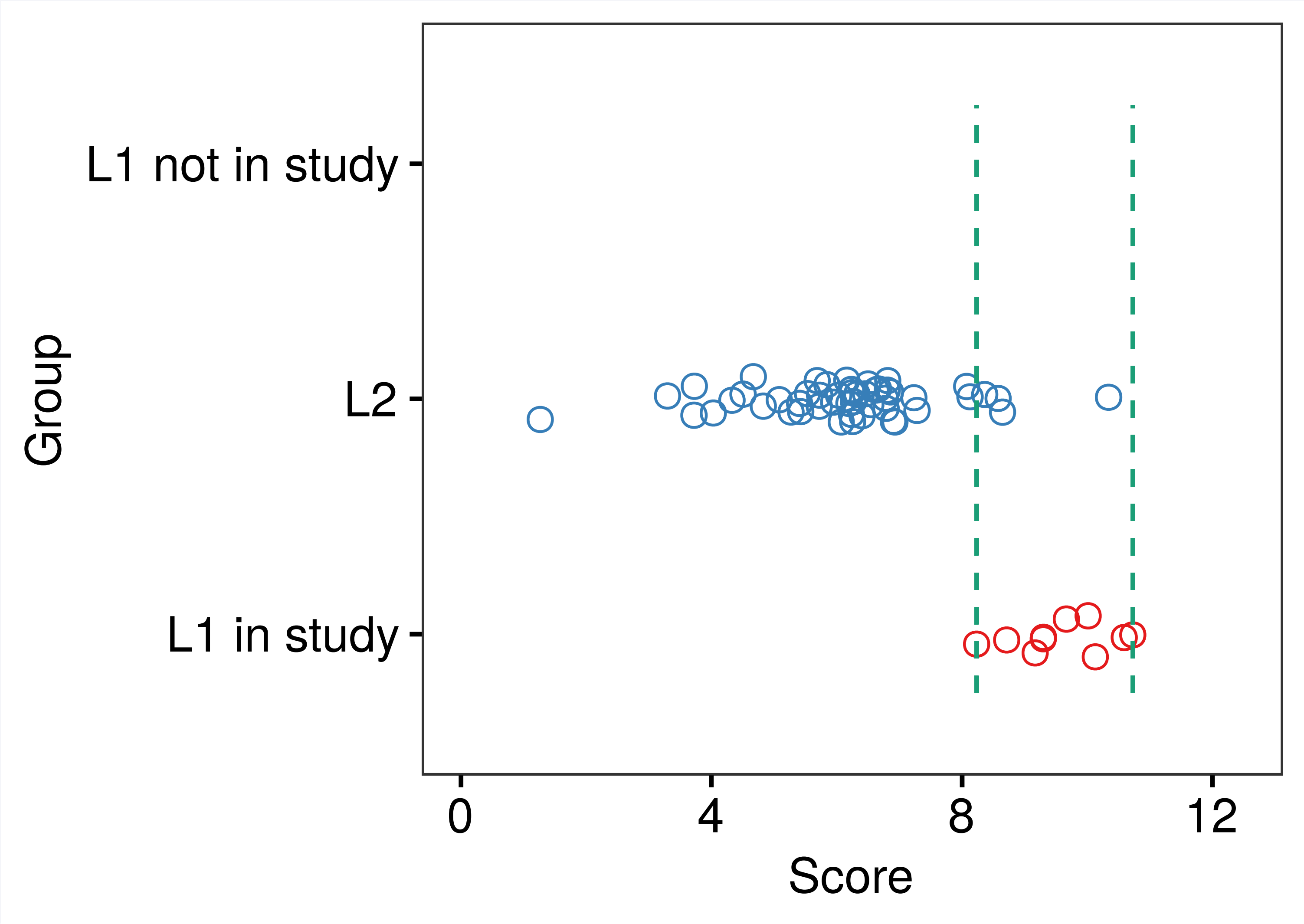


ASSESSING NATIVELIKENESS

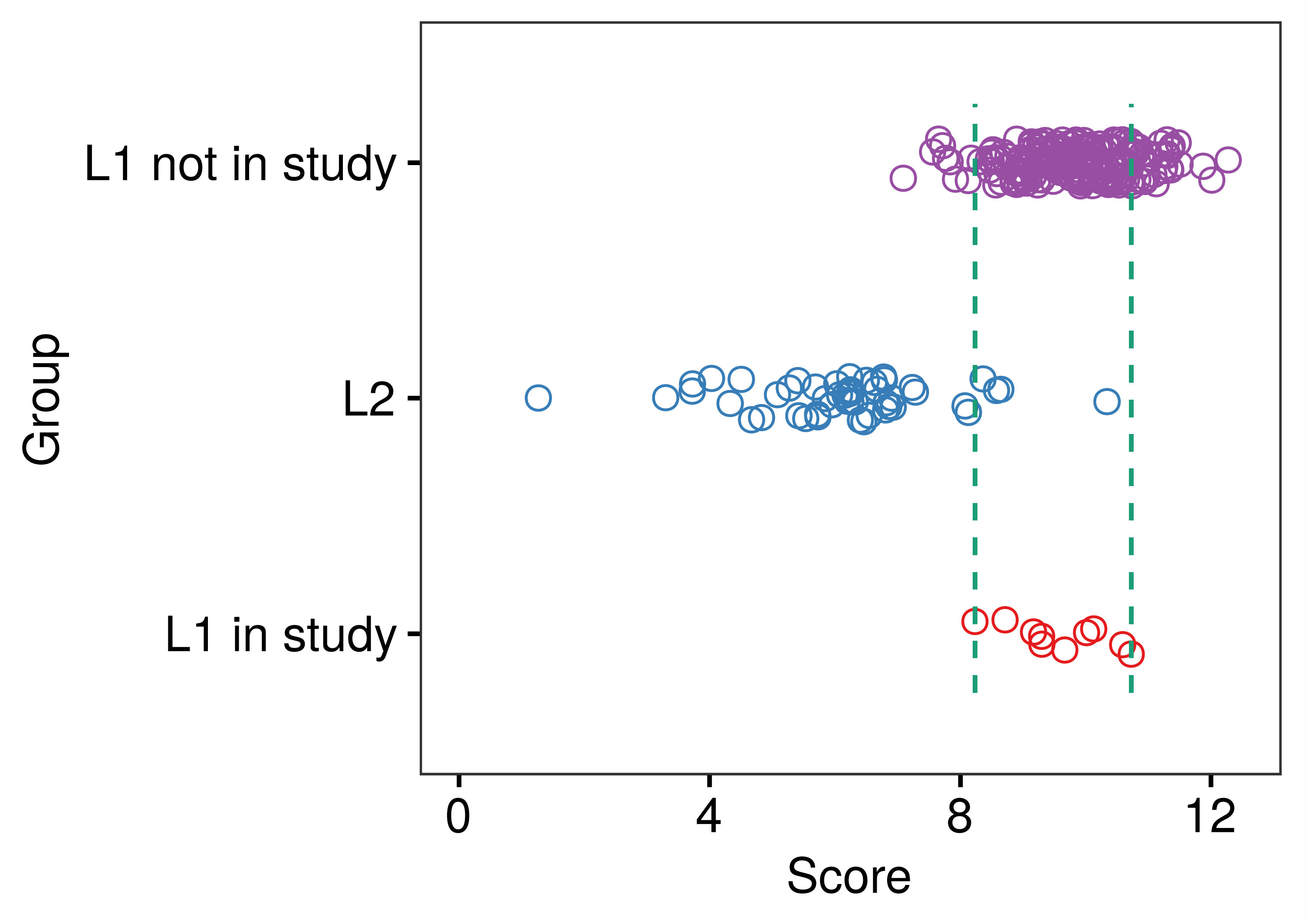



2 standard deviation approach
L2
L1
Group
0
4
8
12
Score

ADDITIONAL NATIVE SPEAKERS



NON-NATIVELIKE NATIVE SPEAKERS?



MISS RATES

Miss rates should affect interpretation of estimates of incidence of nativelikeness in L2 speakers:

- If the nativelikeness interval encompasses the whole relevant L1 population (miss rate: 0.00), then L2 speakers labelled as non-nativelike really are non-nativelike.
- If the nativelikeness interval excludes, say, 50% of the relevant L1 population (miss rate: 0.50), then many of the L2 speakers labelled as non-nativelike may yet be nativelike: the interval is too narrow.

“SCRUTINISED” NATIVELIKENESS

- Set bar for nativelikeness higher by requiring L2 speakers to perform to native standards on several tasks.

  (Abrahamsson & Hyltenstam 2009, Hyltenstam & Abrahamsson 2003, Long 2005)

- Example:
  Abrahamsson & Hyltenstam (2009) subjected 41 highly proficient L2 speakers of Swedish as well as 15 L1 speakers to a whole battery of tasks. Only ‘two, possibly three’ of the L2 speakers fell within the L1 range on all 10 tasks.

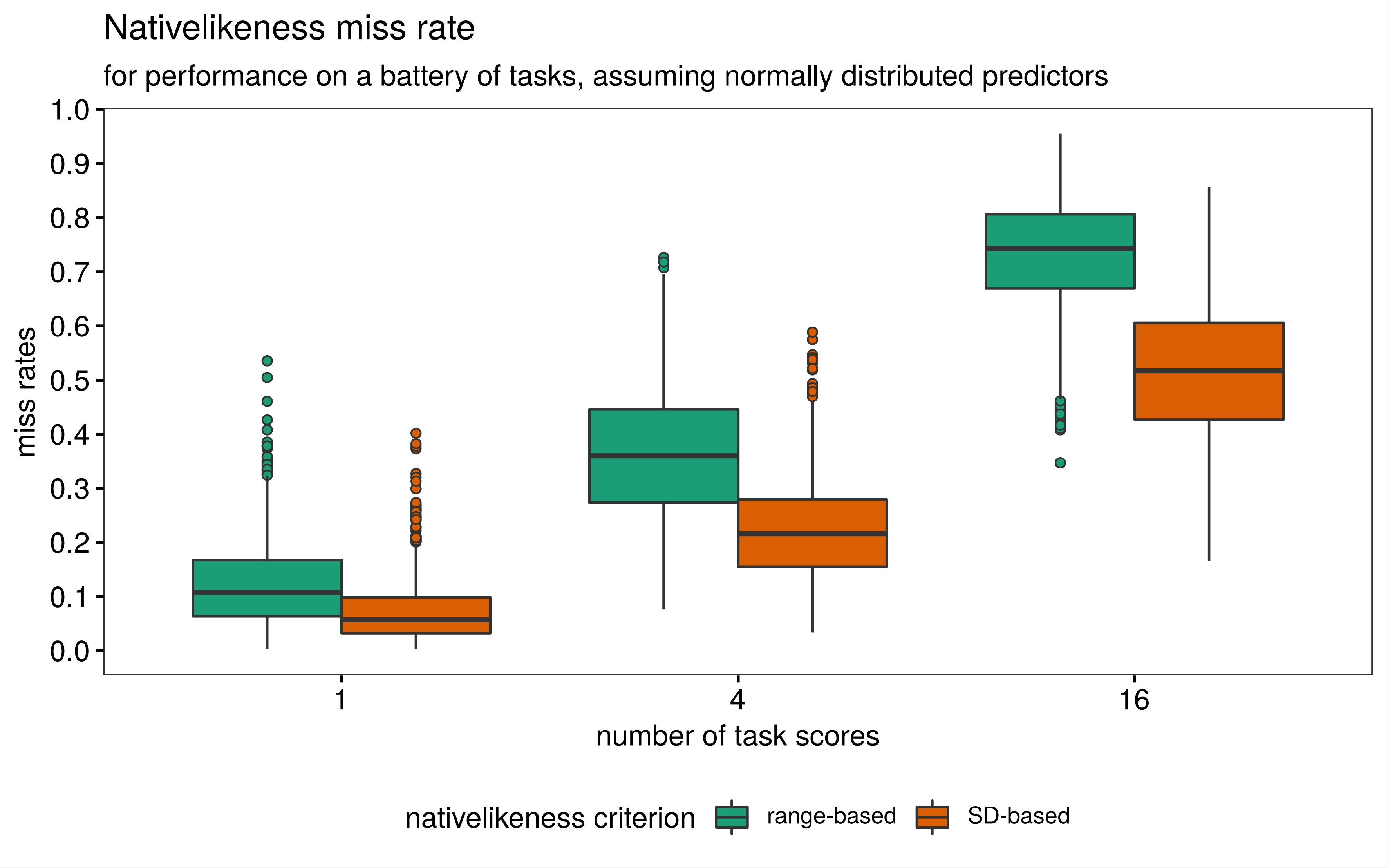
- But how often would *other* L1 speakers erroneously be categorised as nonnative if judged by the same criteria?

jan.vanhove@unifr.ch | janhove.github.io | @janhove

HOW BAD COULD IT BE?

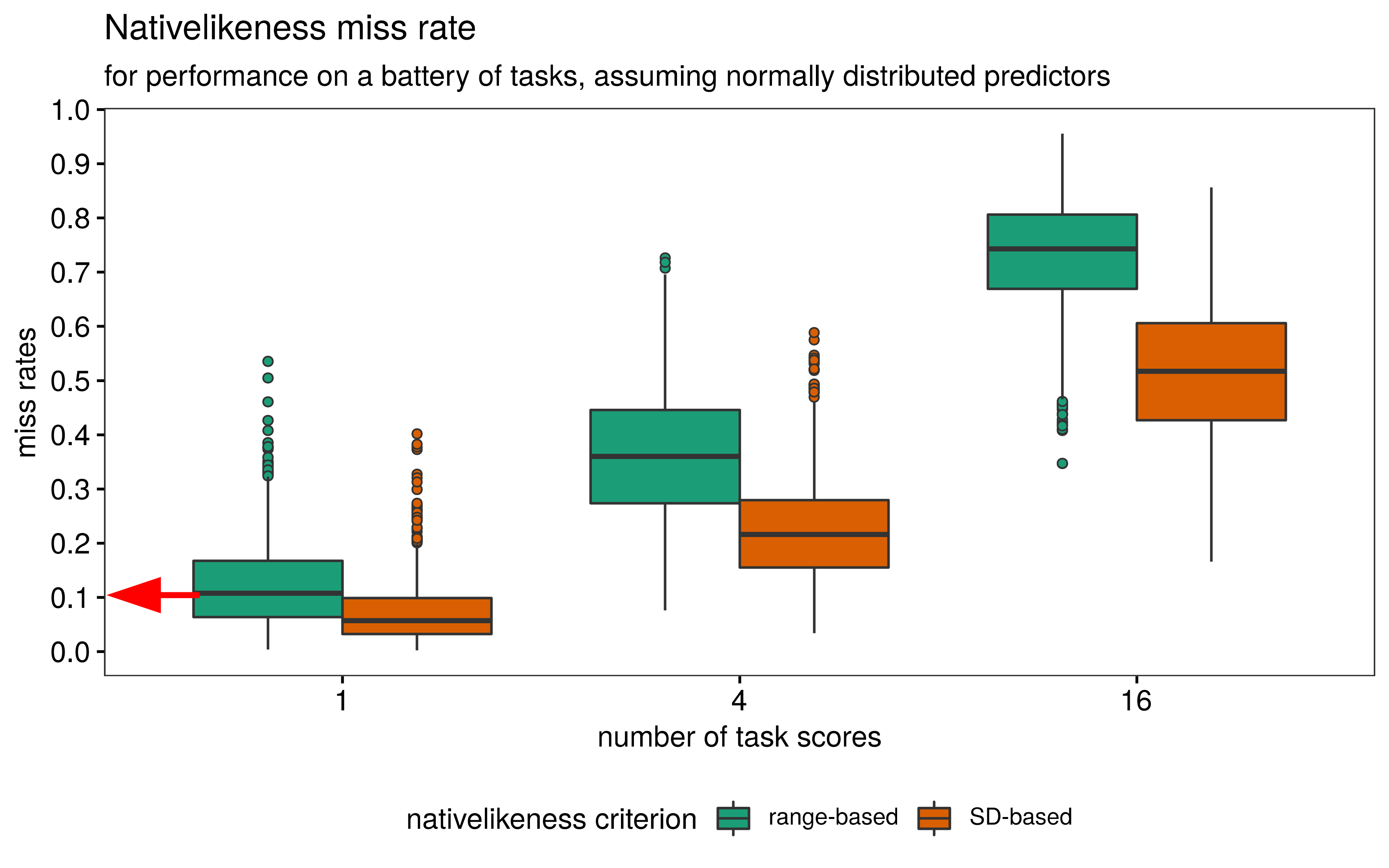
- Simulation:
  - L1 sample size: $n = 15$
  - Generate random sample ($n = 15$) from multivariate normal distributions with 1, 4 or 16 variables (= task scores).
  - Task scores are moderately correlated in population ($\rho = 0.50$).
  - Use sample to construct range- and SD-based intervals.
  - Calculate the proportion of the parent population that falls outside *any* of the range- or SD-based intervals.

HOW BAD COULD IT BE?



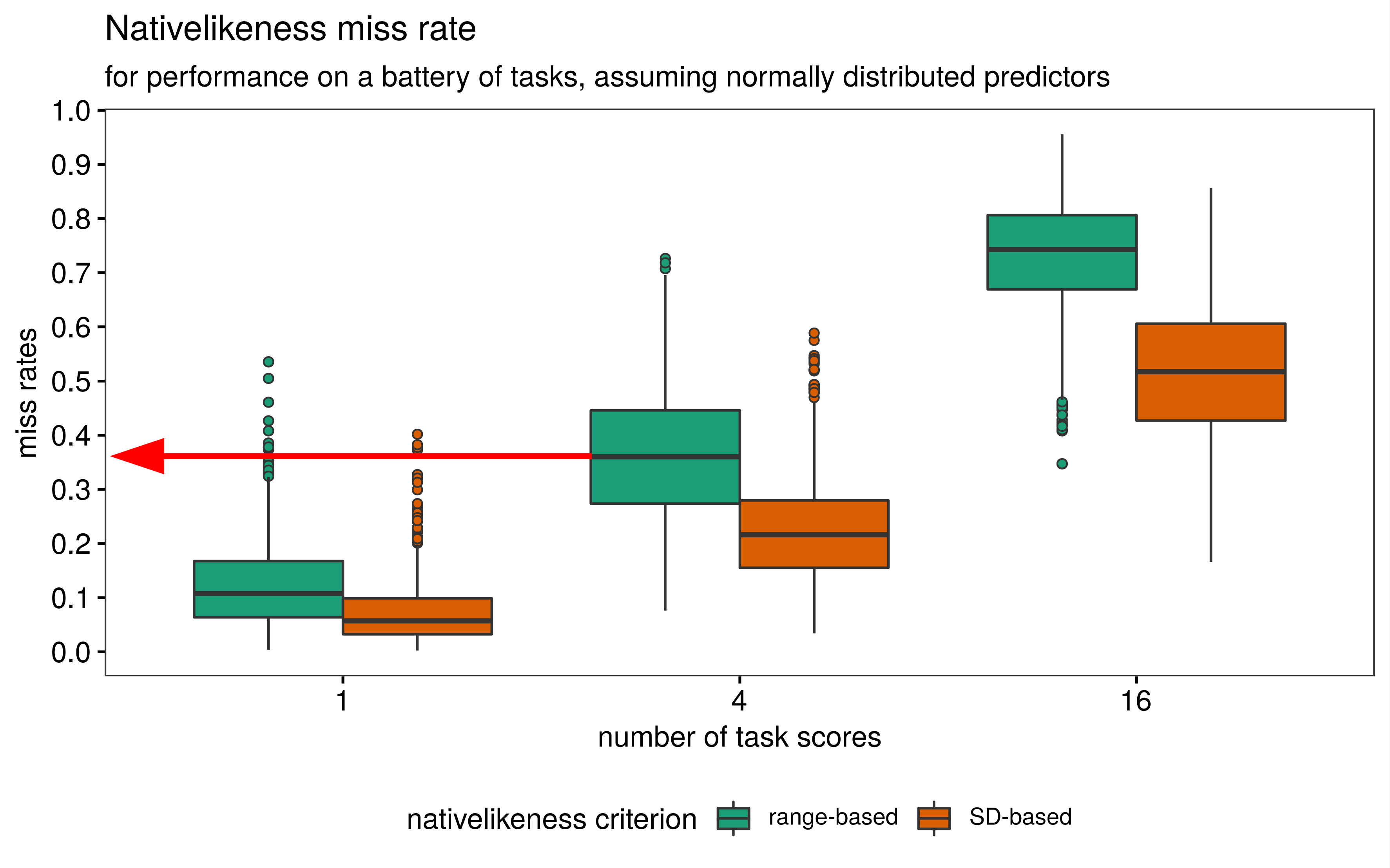
jan.vanhove@unifr.ch | janhove.github.io | @janhove

HOW BAD COULD IT BE?



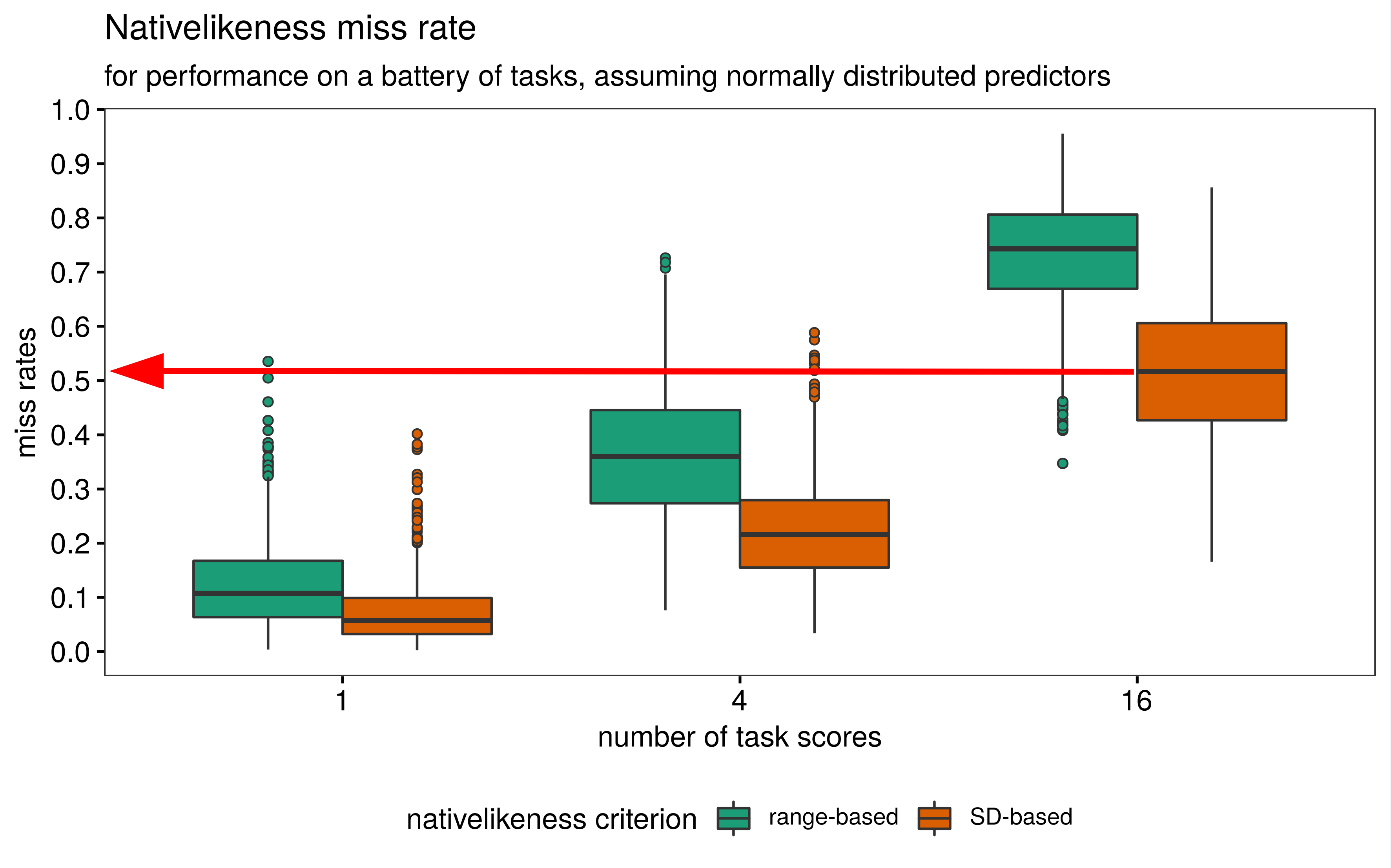
jan.vanhove@unifr.ch | janhove.github.io | @janhove

HOW BAD COULD IT BE?

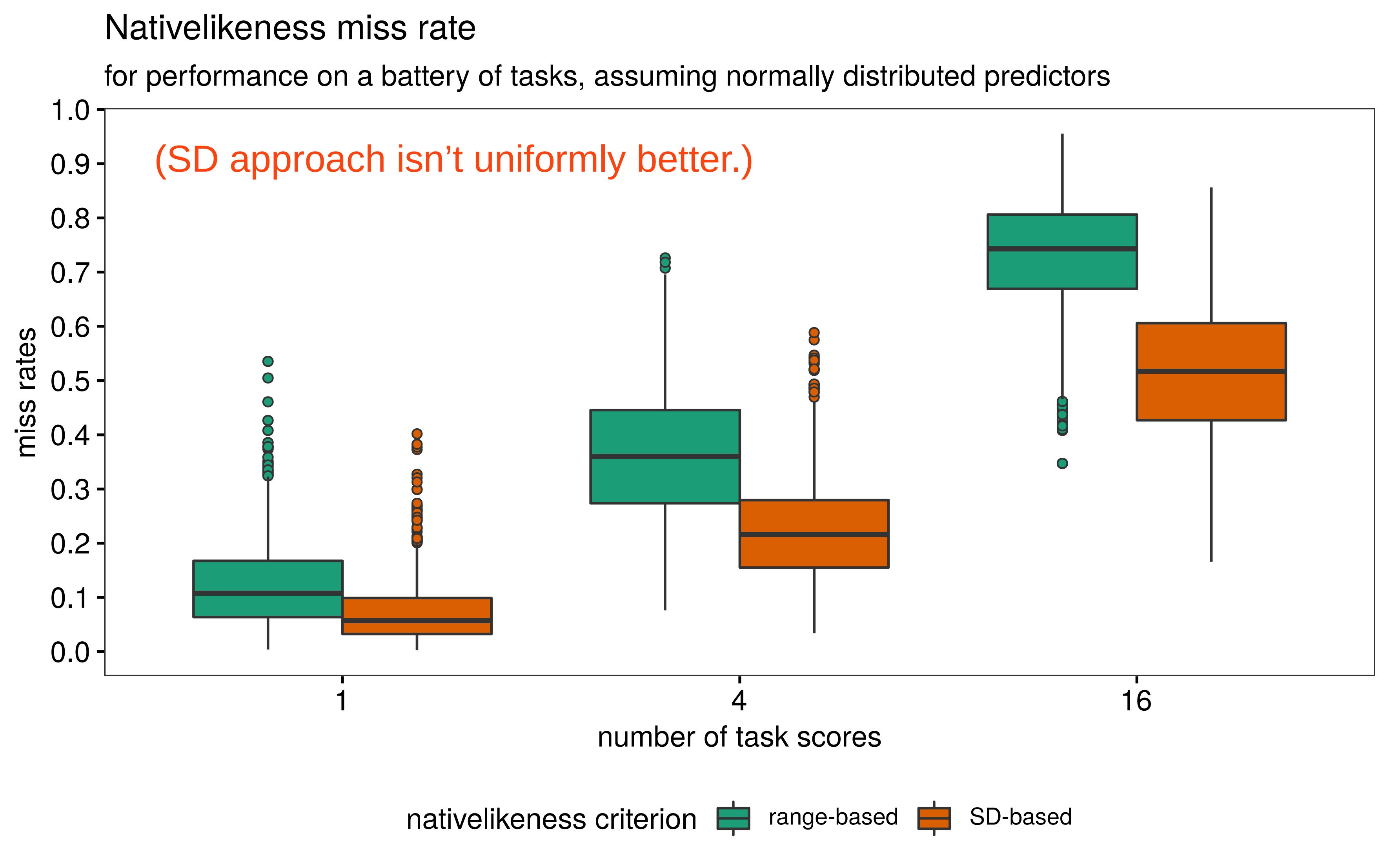


jan.vanhove@unifr.ch | janhove.github.io | @janhove

HOW BAD COULD IT BE?



HOW BAD COULD IT BE?



OVERSCRUTINISED NATIVELIKENESS?

- Ambitious attempts to scrutinise nativelikeness run the risk of setting the bar for nativelikeness so high that even many L1 speakers *sampled from the same population as the controls* may not pass it.
  - Same population = same age, linguistic background, neurological functioning, SES, motivation, blood alcohol level, ...
- But also non-negligible miss rates even for 1 task.
- Necessary to at least estimate how well/poorly calibrated these criteria are in individual studies.
- Probably better to set the criteria in a different way.

jan.vanhove@unifr.ch | janhove.github.io | @janhove

Part II

Two solutions

(well, one, really)

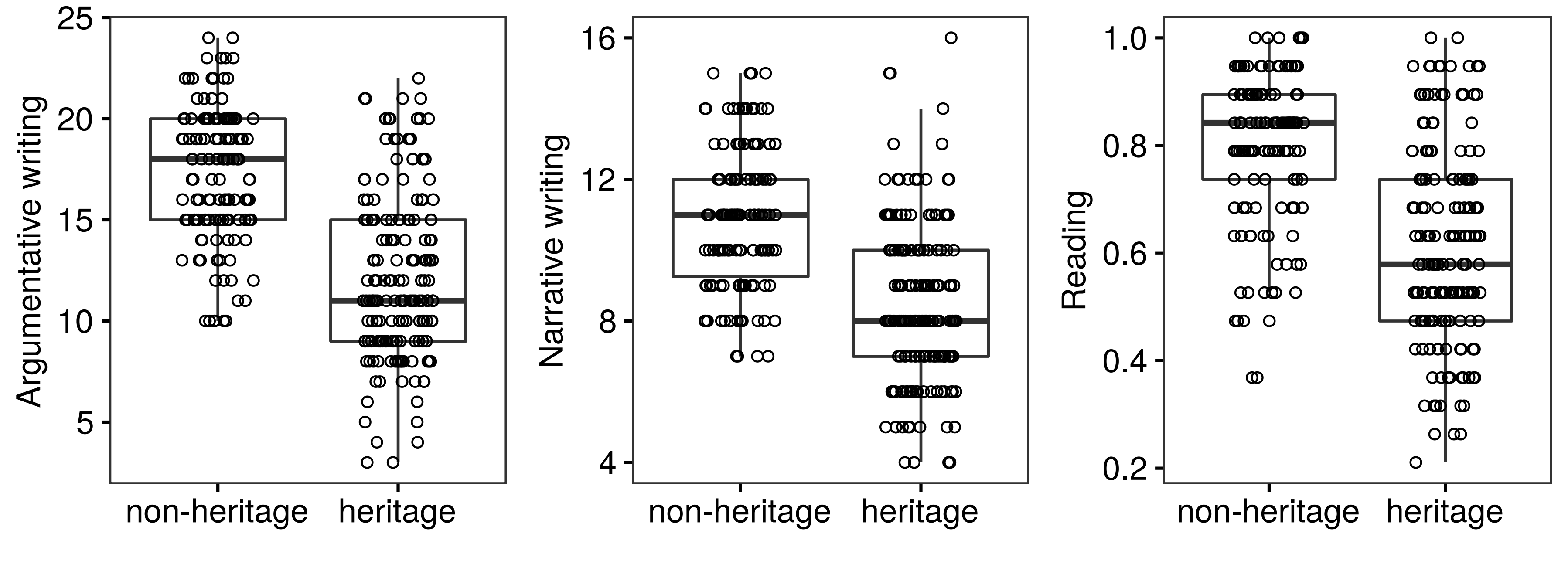
jan.vanhove@unifr.ch | janhove.github.io | @janhove

SOLUTION 2: CLASSIFICATION MODELS

- Fit the original L1 and L2 data in a classification model/algorithm with the task scores as predictors and L1/L2 group membership as the outcome.
  - e.g., logistic regression, discriminant analysis, classification trees, random forests, support vector machines etc.
- How confidently does it assign L1 group membership to which L2 speakers, and how confidently does it assign L2 group membership to which L1 speakers?
- Use cross-validation to avoid fitting the model too closely to the data at hand (see Yarkoni & Westfall, 2017, and online materials).

jan.vanhove@unifr.ch | janhove.github.io | @janhove

EXAMPLE: HERITAGE VS. NON-HERITAGE SPEAKER?

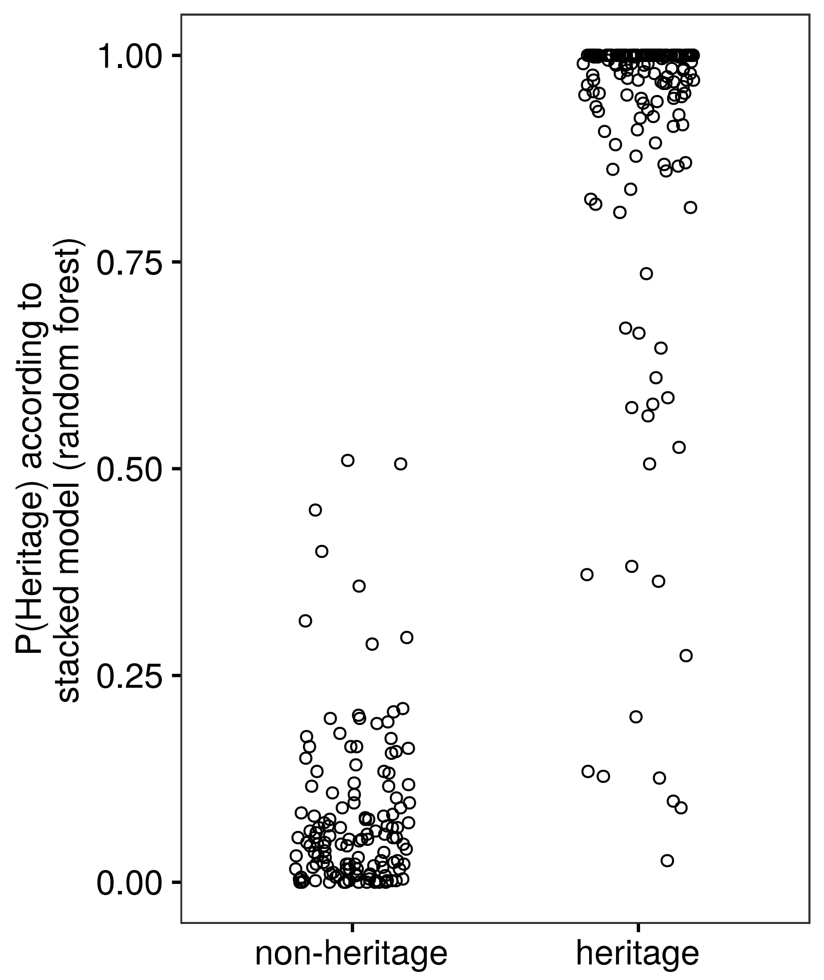


Data from Desgrippes et al. (2017), Pestana et al. (2017) and Berthele & Vanhove (2017).

jan.vanhove@unifr.ch | janhove.github.io | @janhove

EXAMPLE: HERITAGE VS. NON-HERITAGE SPEAKER?

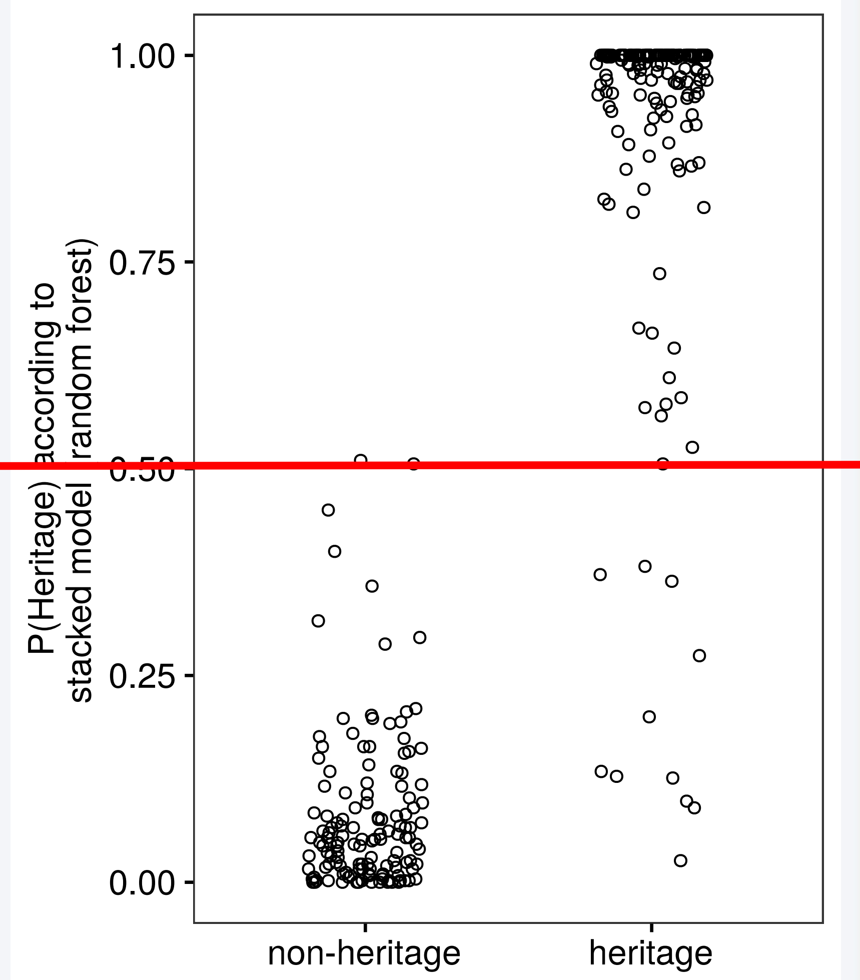
- Fit and cross-validate several models/algorithms on these data.
- Classification probabilities according to best fitting model (of the ones I've tried)



EXAMPLE: HERITAGE VS. NON-HERITAGE SPEAKER?

- Fit and cross-validate several models/algorithms on these data.
- Classification probabilities according to best fitting model (of the ones I've tried)
- 50% threshold:

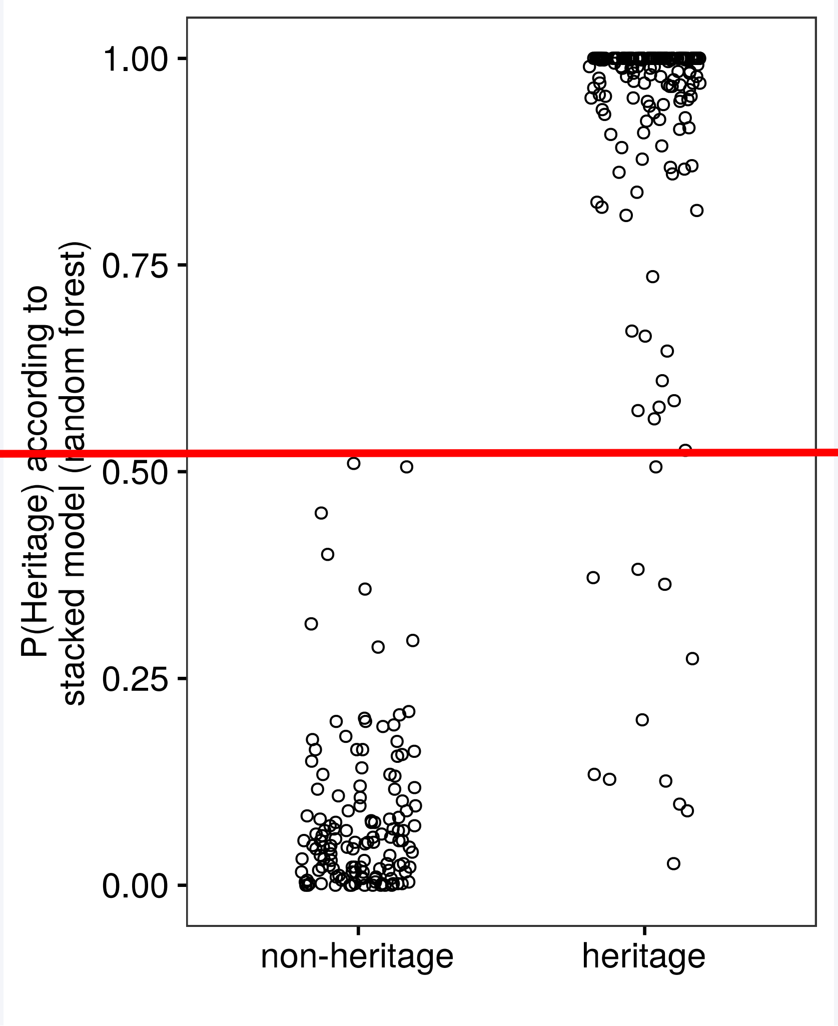
  6% of heritage speakers labelled as non-heritage speakers;
  1.5% of non-heritage speakers labelled as heritage speakers



EXAMPLE: HERITAGE VS. NON-HERITAGE SPEAKER?

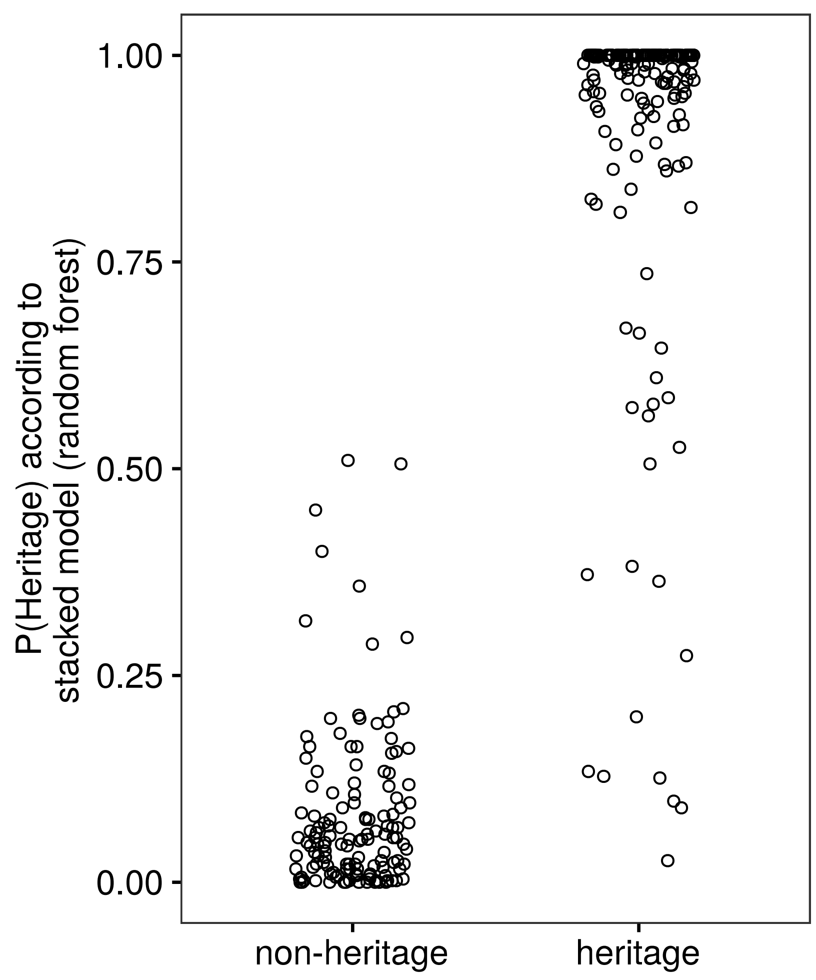
- Fit and cross-validate several models/algorithms on these data.
- Classification probabilities according to best fitting model (of the ones I've tried)
- 51% threshold:

  7% of heritage speakers labelled as non-heritage speakers;
  0% of non-heritage speakers labelled as heritage speakers



EXAMPLE: HERITAGE VS. NON-HERITAGE SPEAKER?

- Fit and cross-validate several models/algorithms on these data.
- Classification probabilities according to best fitting model (of the ones I've tried)
- Regardless of threshold:
  - no perfect separation
  - considerable 'doubt' for many speakers



SOLUTION 2: CLASSIFICATION MODELS

- Possible to reanalyse old datasets!
- Most models/algorithms can output continuous probabilities for each speaker rather than just yes/no classifications.

❢ Different models/algorithms will yield different answers.
↳ Publish data so others can check and improve on your analyses.

TAKE-HOME POINTS

- Need to estimate how often nativelikeness criteria will exclude native speakers.
- Possible to reanalyse old datasets using classification models/algorithms + cross-validation.
- Classification probabilities may show that, even if nativelikeness is conceived of as binary category, the data in any given study may not allow for a neat categorisation of all participants.

https://janhove.github.io/nativelike/

jan.vanhove@unifr.ch | janhove.github.io | @janhove

WHY DOES THIS HAPPEN?

- **Range approach:**
  Sample ranges consistently underestimate the population range.
- **Standard deviation approach:**
  Even for random samples from normal distributions,
  "**sample** mean ± 2 **sample** standard deviations"
  does not encompass 95% of the **population**.
  - Sample means estimate the population mean imperfectly.
  - Sample standard deviations estimate the population standard deviation imperfectly (and tend to underestimate it, more so for small samples).
- More difficult to pass several hurdles.

jan.vanhove@unifr.ch | janhove.github.io | @janhove

SOLUTION 1: ESTIMATE MISS RATE USING NEW L1 SAMPLE

- Take the nativelikeness interval(s) of a study.
- Subject a number of L1 speakers from the relevant population to the same task(s).
- Check which proportion of the new L1 sample falls outside the original study's interval(s) = estimate of miss rate.
- Compute an uncertainty interval (e.g., credible interval or Wilson's binomial confidence interval) around this estimate.
- Don't redefine the interval(s) using the new sample, unless you want to collect another validation sample, etc.

SOLUTION 1: ESTIMATE MISS RATE USING NEW SAMPLE

- Example:
  - Original study: Nativelikeness interval = [4, ∞]
    It doesn't matter how this interval was constructed.
  - 35 new L1 speakers
  - If 2 of them have a score below 4:
    2/35 = 6% miss rate, 95% CI: [1.6%, 19%].
  - If none of them have a score below 4:
    0/35 = 0% miss rate, 95% CI: [0%, 10%].

SOLUTION 1: ESTIMATE MISS RATE USING NEW SAMPLE

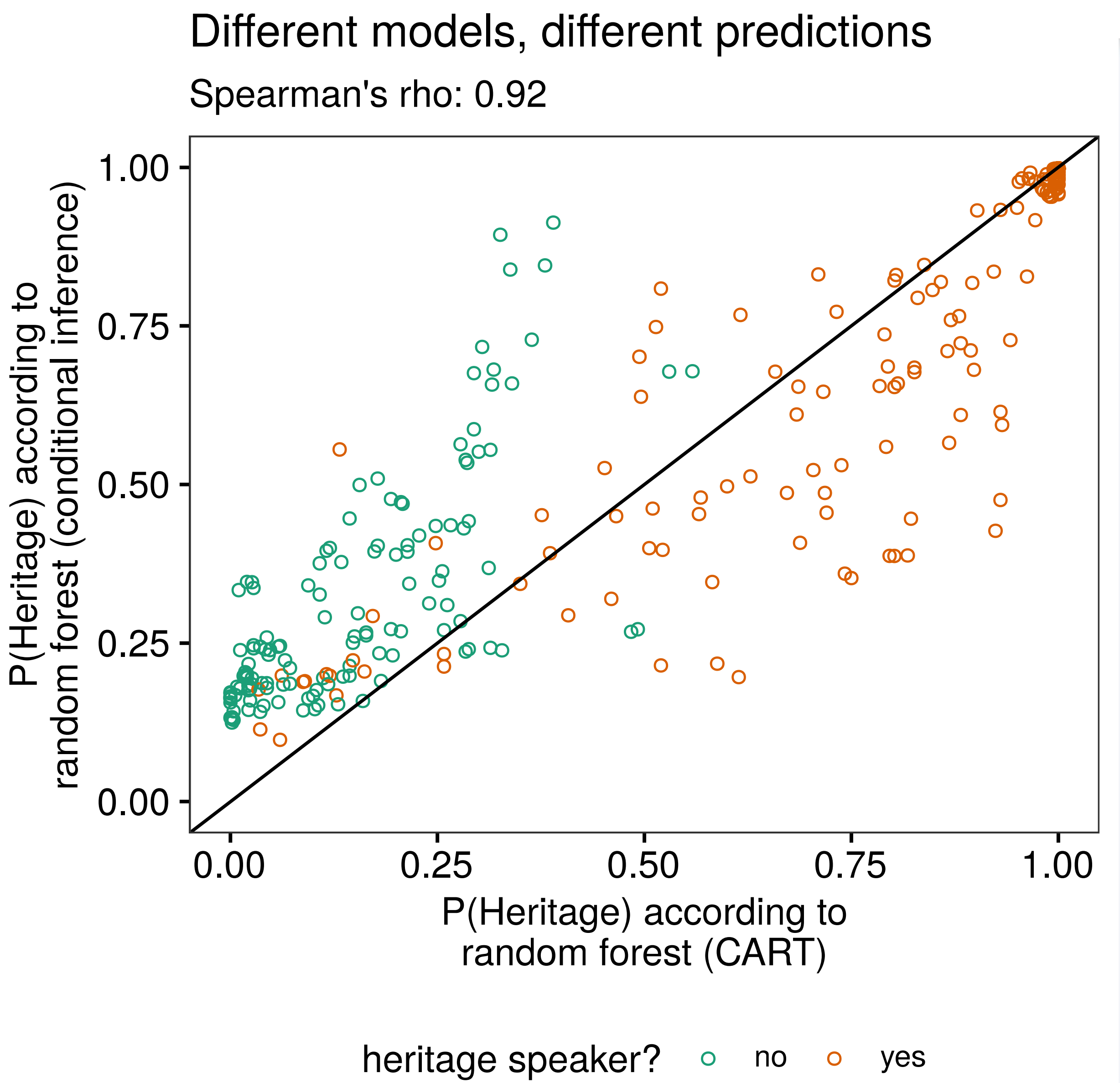
- Answers the question *How much too strict were the intervals that this study applied?*
- Conceptually pretty easy.
- But need to collect additional data.
- And you can't even redefine the criteria when it turns out that they are much too strict (unless you validate these new criteria with new data).

CROSSVALIDATION?

- **Why?** To combat overfitting (see Yarkoni & Westfall 2017)
- **What?** Fit model, but leave out part of the dataset.
- Easiest case: *leave-one-out* crossvalidation (LOOCV)

  1) Leave out the *i*th observation from the dataset.
  2) Fit model using the remaining observations.
  3) Use model to predict the class (i.e., L1 or L2) of the *i*th observation (which wasn't used for fitting the model).

  Repeat steps 1-3 until you have a predicted class for each observation based on a model that didn't 'see' this observation.


Different models, different predictions
Spearman's rho: 0.92
1.00
0.75
0.50
0.25
0.00
P(Heritage) according to
random forest (conditional inference)
0.00
0.25
0.50
0.75
1.00
P(Heritage) according to
random forest (CART)
heritage speaker?
no
yes