GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

CiBreed Center for Integrated Breeding Research

LANDGESTÜT CELLE

8TH IWSLP ADELHEIDSDORF / CELLE

Better than random?

Assessing Inter- and Intra-Rater Reliability of Linear Scoring in Warmblood Breeding

Alexandra Weigt (Georg-August-University Göttingen, Germany)
Dr. Axel Brockmann (LGST Celle),
Prof. Dr. Jens Tetens (Georg-August-University Göttingen, Germany),
Prof. Dr. Armin O. Schmitt (Georg-August-University Göttingen, Germany)

Contents of today's presentation

- Study objectives
- Data collection
- Inter-rater reliability
- Intra-rater reliability
- Better than random?
- Practical implications

Study objectives

- Reliable phenotypes essential for successful breeding programmes and advancement of genomic applications
- Quantifying reliability of Linear Scoring
- Under standardised environmental conditions (images)
- Identifying strengths and limitations to further advance equine breeding programmes

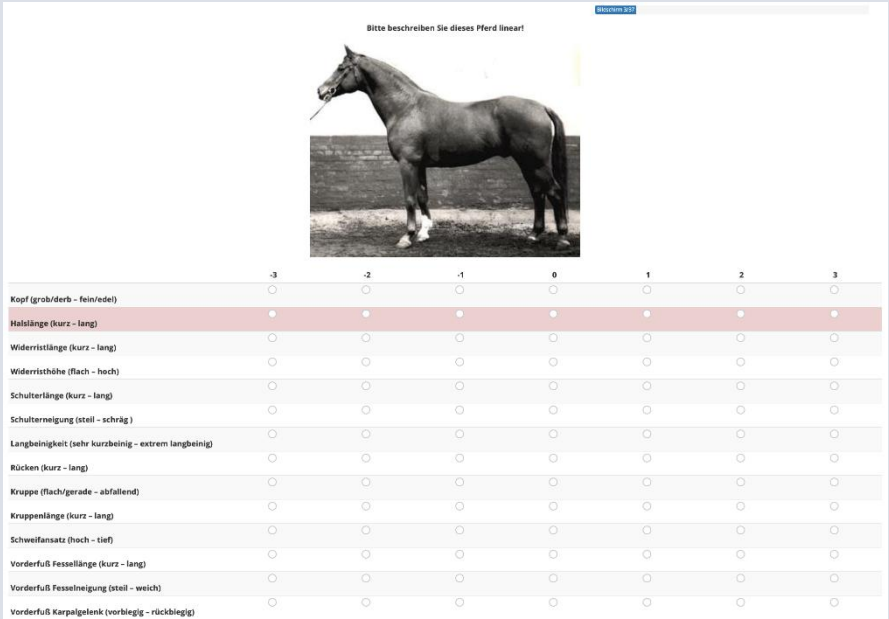



Materials & Methods

- Online questionnaire
- Programme: EyeQuestion[1]

- 34 horse images born between 1929 and 1993
- 17 linear traits

- Survey repetition after one month
- Personal information (e.g. professional experience)

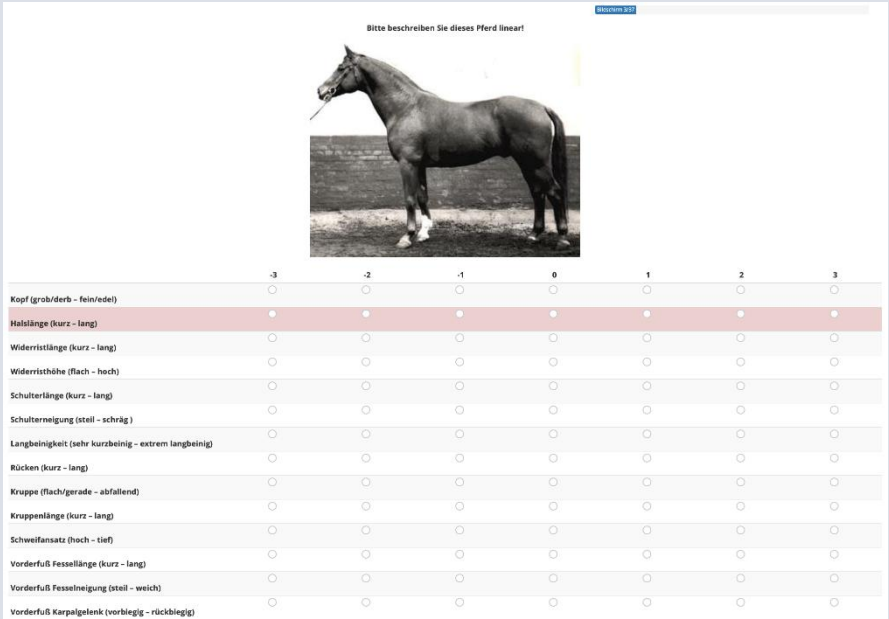


[1] EyeQuestion (Version 6.1.2.6, The Netherlands)

Materials & Methods

- Intraclass Correlation Coefficient (ICC)
- 0 (no reliability) – 1 (excellent reliability)
- irr[2] package in R[3]

- Permuation tests: Randomly shuffled horse x judge matrices within each trait

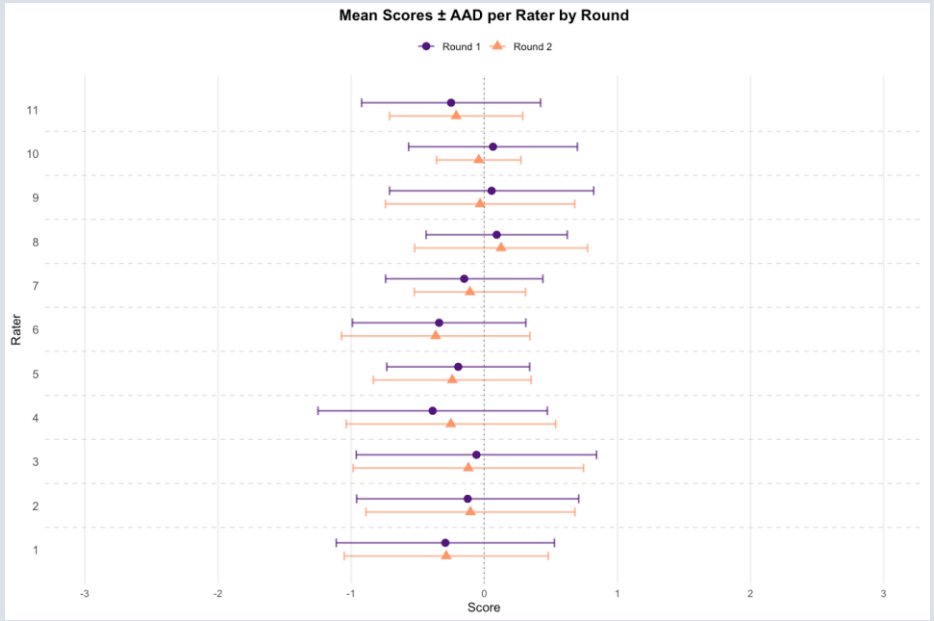


[2] Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2025) URL https://www.R-project.org/
[3] Gamer, J. Lemon, I. Fellows, P. Singh, Irr: Various Coefficients of Interrater Reliability and Agreement, r package version 0.84.1 (Jan. 2019). doi:10.32614/CRAN.package.irr.

General data analysis

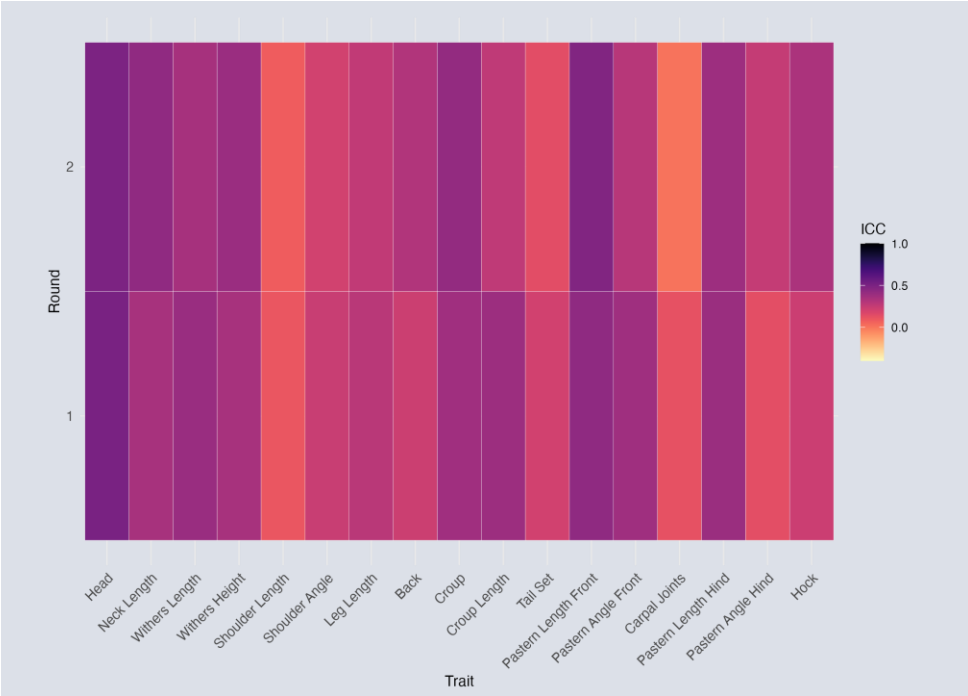
- 11 out of 24 experts completed both rounds of the questionnaire
- $34 \; x \; 17 \; x \; 11 = 6{,}358$ scores per round

- Limited usage of the -3 to +3 scale



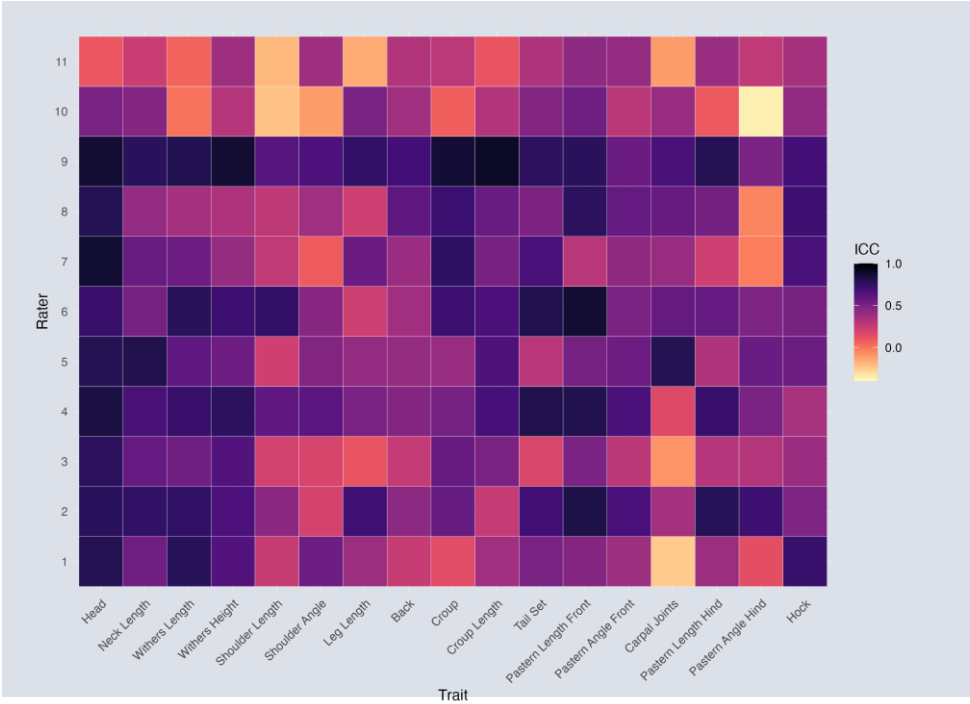
Inter-rater reliability

- Consistency between raters per round



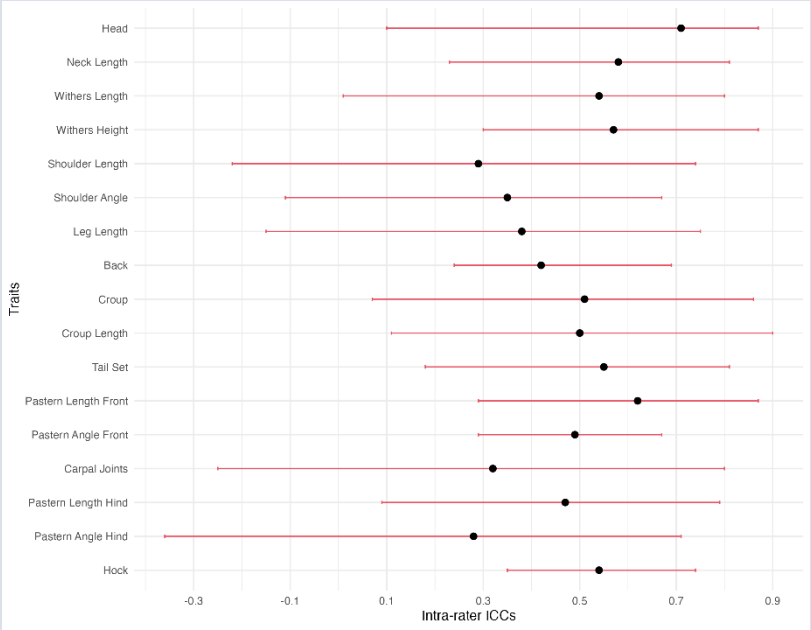
Intra-rater reliability

- Consistency between rounds per rater



Intra-rater reliability

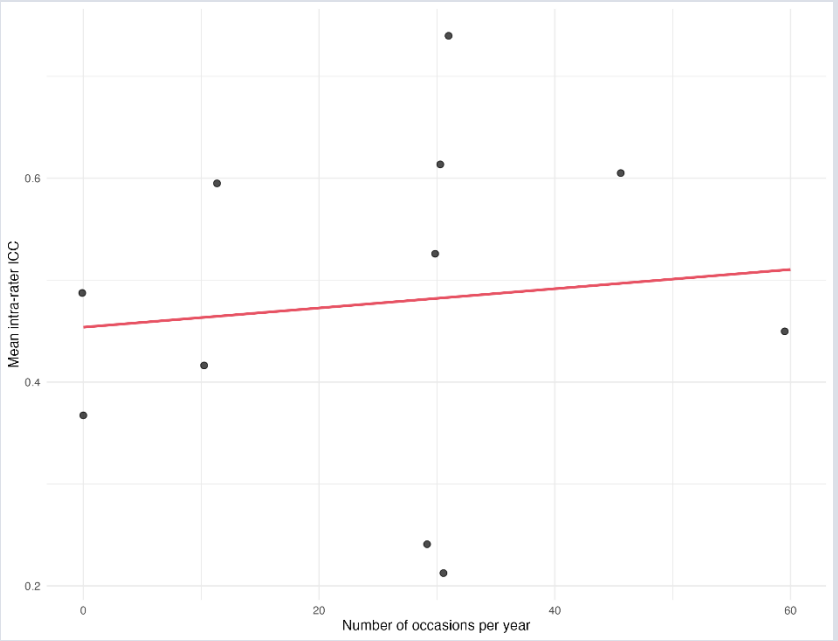
- Consistency between rounds per rater
- Mean ICCs per trait and their respective ranges



Intra-rater reliability

In the context of experience

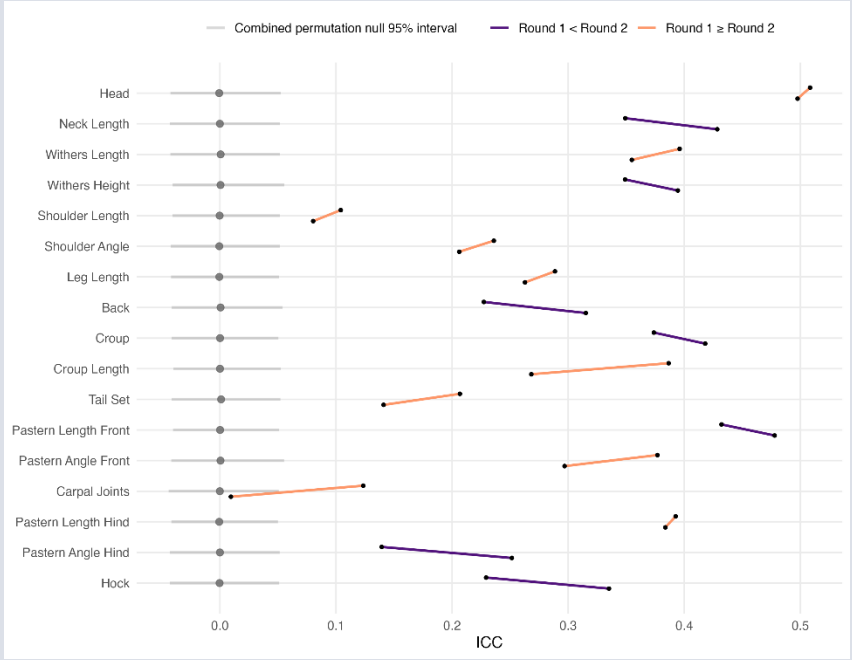
- Partipants were asked for the number of occasions they apply linear scoring per year
- No significant correlation



Better than random?

Inter-rater reliability

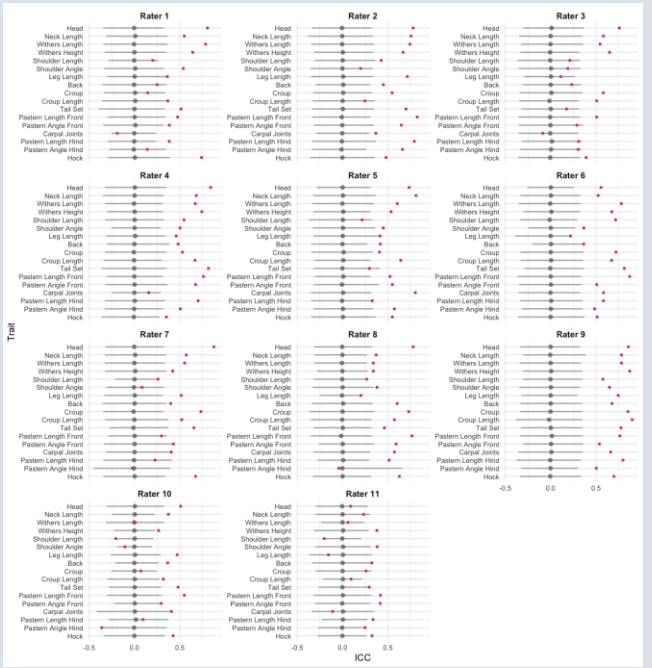
- Permuation tests: Randomly shuffled horse x judge matrices within each trait



Better than random?

Intra-rater reliability

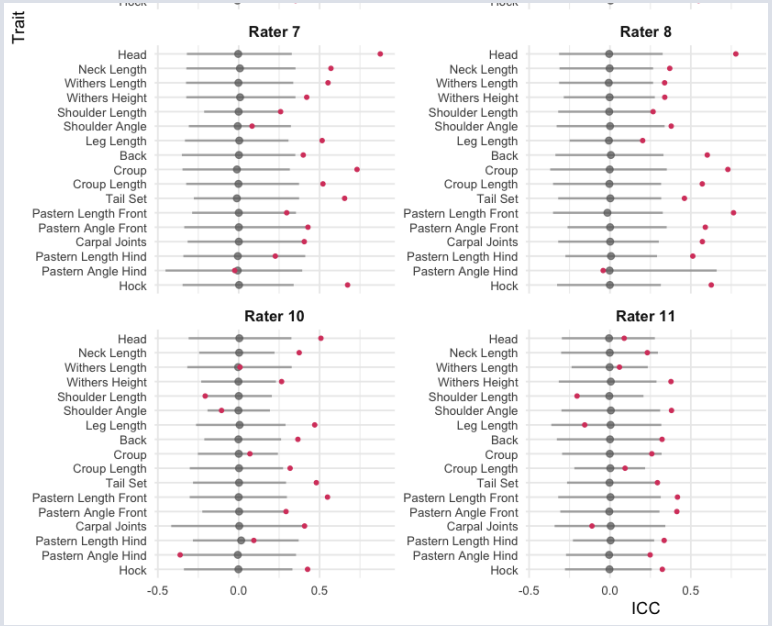
- Permuation tests: Randomly shuffled horse x judge matrices within each trait



Better than random?

Intra-rater reliability

- Permuation tests: Randomly shuffled horse x judge matrices within each trait

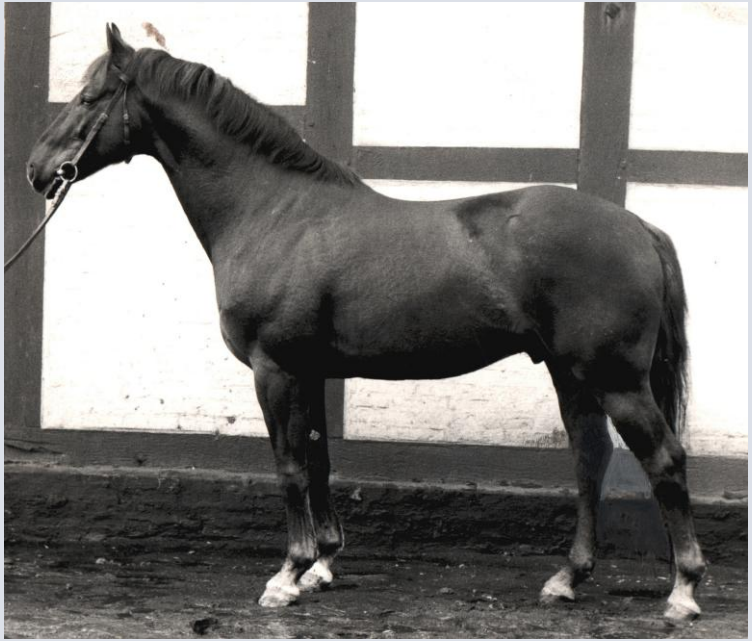


Practical implications

- Better than random? – Yes!
- But large differences between traits

- Exchange and standardised training
- Focus on underlying anatomical structures

- Research on automated phenotyping technologies assisting the human eye



Thank you for listening!

Time for questions and discussion