Journal of Equine Veterinary Science 88 (2020) 102932

Contents lists available at ScienceDirect

Journal of Equine Veterinary Science

journal homepage: www.j-evs.com

Original Research

Should We Agree to Disagree? An Evaluation of the Inter-Rater Reliability of Gait Quality Traits in Franches-Montagnes Stallions

Annik Imogen Gmel [a, b, c, *], Gerhard Gmel [d], Rudolf von Niederhäusern [a], Michael Andreas Weishaupt [c], Markus Neuditschko [a]

[a] Agroscope - Swiss National Stud Farm, Les Longs-Prés, Avenches, Switzerland
[b] Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, Switzerland
[c] Equine Department, Vetsuisse Faculty, University of Zurich, Zurich, Switzerland
[d] Alcohol Treatment Centre, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 15 November 2019
Received in revised form
10 January 2020
Accepted 13 January 2020
Available online 22 January 2020

Keywords:
Horse
Breeding
Linear profiling
Scoring
Intraclass correlation coefficient

ABSTRACT

Gait quality, that is, the way horses move according to functional and aesthetic principles, englobes many traits that are scored by experts during breeding competitions. The experts can score a trait on a subjective valuating (SV) scale or on a linear profiling (LP) scale representing the biological extremes of the population. However, the reliability of the appraisal of gait quality traits has not been extensively evaluated. In this study, seven breed experts appraised the walk and trot quality of 24 Franches-Montagnes stallions presented in hand on a sand track. Inter-rater reliabilities of six traits (five SV and one LP) at the walk and eight traits (five SV and three LP) at the trot were estimated with intraclass correlation coefficients (ICCs). The inter-rater reliabilities were poor (ICC < 0.50). The scale anchoring varied between experts, and the variance of scores was low. There were no systematic differences in inter-rater reliability between LP and SV traits. Future studies should determine whether the inter-rater reliabilities may be increased by a more precise definition of the scores within each trait to improve the absolute agreement between experts, by a more uniform scale anchoring between experts, and by decreasing the number of scale items. However, considering the inherent limitations of the human eye in observing high-speed movement, the use of a field-applicable kinematic measurement system may support breeding experts in the appraisal of gait quality traits in the future.

© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The breeding programs of many horse breeds particularly focus on conformation [1,2] and gait quality [3], as these traits are most likely associated with sports performance, health, and longevity. Conformation and gait quality traits can already be observed in young horses presented in hand, whereas information on sports performance and other traits are available late in life (long generation interval [4]; peak performance at around 10 years old in many equestrian sports [5]). Most of these traits are essentially appraised visually by experts of the breed, and not measured. The visual appraisal by an expert can be carried out using either subjective valuating (SV) scores on a scale from “bad” to “good” or a linear profiling (LP) scale, also called linear description. SV scores strongly depend on the individual expert and may need to be adapted to the horses presented at each competition, so that a ranking can be

Authors' contributions: A.I.G., R.v.N., and M.A.W. contributed to conceptualization. A.I.G. contributed to data curation, formal analysis, investigation, and visualization. A.I.G., M.N., M.A.W., and R.v.N. contributed to funding acquisition. A.I.G., M.N., R.v.N., G.G., and M.A.W. contributed to methodology. A.I.G., M.N., and M.A.W. contributed to project administration. R.v.N. and M.A.W. contributed to resources. M.N. and M.A.W. did supervision. G.G. contributed to validation. A.I.G., G.G., M.N., and M.A.W. wrote, reviewed, and edited the article.

Animal welfare/ethical statement: This study was conducted under the animal experiment permit number VD 3164 complying to Swiss Federal Legislation. No animal was harmed during the experiment.

Conflicts of interest statement: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the article, or in the decision to publish the results.

* Corresponding author at: Annik Imogen Gmel, Agroscope - Swiss National Stud Farm, Les Longs-Prés, 1580 Avenches, Switzerland.
E-mail address: annik.gmel@agroscope.admin.ch (A.I. Gmel).

https://doi.org/10.1016/j.jevs.2020.102932
0737-0806/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

made at the end of the day. With LP, an expert attributes a score on a linear scale based on predefined biological extremes [6]. For example, the gait quality trait "elasticity" may range from "stiff" to "elastic." Linear profiling is considered less subjective [6], and the scores should be attributed in relation to the overall population. However, both appraisal types (SV and LP) are scored on an itemized scale and based on an expert opinion, which may be influenced by different types of biases (e.g., individual preferences, experience, and reduced concentration because of fatigue).

Many breeding associations (e.g., Swedish Warmblood [7], Belgian Warmblood [8], and Dutch Warmblood [3,9]) have included traits describing gait quality (the manner in which horses move according to aesthetic and functional principles), such as stride length, cadence, rhythm, suppleness, elasticity, and impulsion in their assessment protocols during breeding competitions. Only stallions with the highest conformation and gait quality scores will be allowed for breeding (licensing or "Körung" by a breeding federation). Furthermore, these scores will be used in breeding value estimation, which is becoming an important tool for breeders to select the best stallion for their mares. The scoring of conformation and gait quality traits, therefore, has a substantial impact on early breeding decisions, and selection based on inaccurate information may lead to unwanted genetic bottleneck effects without the intended breeding progress. Despite their importance in the selection procedure of sports horses, the reliabilities of conformation and gait quality traits by different experts have not been systematically evaluated, and there is only limited research on the reliability and consistency of conformation traits. Druml et al. [10,11] reported poor inter-rater reliabilities calculated with a kappa (κ) coefficient ($0.06 < \kappa < 0.49$) between multiple experts scoring SV conformation traits at the same time of either Lipizzan stallions [11] or mares [10]. In contrast, the repeated assessments of LP conformation traits of Pura Raza Espanol horses in different competitions showed higher inter-rater reliabilities evaluated with intraclass correlation coefficients (ICCs; $0.96 < ICC < 0.99$) [12]. Whether the substantial differences in reliabilities are because of methodology (sample size, statistics, and scaling) or the type of appraisal is unknown, as there are no direct comparisons of the reliabilities between SV and LP scores within a population using the same scaling and methodology.

The ability to reliably score gait quality traits in horses may additionally be affected by the relatively low temporal resolution capacity of the human eye (estimated by Dyson [13]) to follow, for example, the footfall patterns of faster gaits such as the trot. This limitation spurred the development of sequential photography by Eadweard Muybridge in the 19th century to prove an aerial phase at the trot [14,15]. The difficulty in assessing equine movement patterns also concerns veterinary lameness examinations, for which reliabilities were previously estimated [16–21]. These studies report poor to moderate inter-rater reliabilities in lameness scorings (minimum $\kappa = 0.17$ for 13 raters and 24 horses [17] to maximum $\kappa = 0.52$ for 131 horses and 2–5 raters on average [18]), whereas intrarater reliability was generally higher (minimum $\kappa = 0.34$ on 24 horses [17] to maximum $\kappa = 0.68$ on 19 horses [21]). The aim of the present study was to quantify the inter-rater reliabilities of gait quality traits appraised by experts of the breed in 24 Franches-Montagnes (FM) stallions presented in-hand on a sand track and to evaluate the difference in inter-rater reliability between traits appraised on an LP scale in comparison to an SV scale. We hypothesized that the inter-rater reliabilities would be in the same range as the reported inter-rater reliabilities from lameness scorings, and that LP traits would have higher inter-rater reliabilities than SV traits.

2. Materials and Methods

In this study, seven of nine official breeding judges appraised 24 stallions at the walk and trot on five SV scores—*ground coverage, overtracking, suppleness and relaxation, regularity and harmony,* and *activity and impulsion*—on a scale from score 1 ("unfavorable") to score 9 ("ideal") and on four linear traits: *walk* (score 1 "short" to score 9 "long"), *trot* (score 1 "short" to score 9 "long"), *trot: impulsion* (score 1 "little" to score 9 "much"), and *trot: elasticity* (score 1 "stiff" to score 9 "elastic"). All the traits were defined in Table S1.

The majority of the FM stallions (20) were from the Swiss National Stud Farm (SNSF), whereas four old-type FM (RRFB) stallions were from private owners (animal experiment permit number VD 3164). The 24 stallions were part of a large study describing the morphologic variation in FM stallions using the horse shape space model [22]. Our sample of 24 stallions was representative of the variation of 194 stallions born between 1992 and 2013 (oldest respectively youngest approved stallions at the SNSF at the time of the scheduled appraisal, S1 File).

The stallions were presented in hand with a bridle, walked and trotted in a delimited triangle on a sand surface by an experienced handler (Fig. S1). The stallions were first led away from the experts so that the experts saw the stallions from behind, followed by the side and finally from the front. The whole evaluation lasted one and a half hours for all 24 stallions, that is, approximately 3 minutes per stallion. The experts independently filled out the scoring sheet for each horse anonymously and received a coded designation ("#1" to "#7"). The seven experts of the FM breed had the same background training.

The scores for each trait and gait were summarized for each expert in violin plots to visualize the scale anchoring (frequency of use for each score and effective range of scores). Furthermore, a cross-correlation matrix was calculated between the experts over all traits to investigate potential similarities in appraisals using the R library Corrplot [23]. The inter-rater reliability was estimated with an ICC for ordinal data. The ICCs were computed as two-way random models, type "single measurement" and estimated for both absolute agreement (the scores are equal for absolute agreement, $ICC_{(A,1)}$) and consistency (the rater scores of the same animals are additively correlated but not equal, $ICC_{(C,1)}$) [24] using the R library irr [25]. The ICC was interpreted based on Koo and Li's publication [24], where ICC values lower than 0.5 were considered as poor, between 0.5 and 0.75 as moderate, between 0.75 and 0.90 as good, and over 0.90 as excellent agreement [24]. All ICCs were reported with their 95% confidence intervals. We also evaluated the inter-rater reliability using Fleiss' kappa for comparisons with lameness evaluation studies.

3. Results

The summary statistics of all appraisal scores over all experts are reported in Table 1. In general, there was a preference (represented by the mode) for a score of seven for most traits. The mean score ranged from 6.37 to 6.82, and the mode was seven for all traits except for *suppleness and relaxation (SV)* at the walk with a mode of six.

There were some differences in the scale anchoring depending on the experts (Fig. 1A). Experts #2, #3, #4, and #5 preferentially scored stallions on average with a seven. Expert #5 had the narrowest use of the scale with a heavy preference for the score 7 (overall median, except for two traits). Expert #7 used all scores relatively evenly across all traits. The score 1 was exclusively attributed by Expert #7. There were also individual similarities in scorings between experts (Fig. 1B). Experts #2 and #3 showed the highest correlation to one another, whereas Expert #5 had the lowest similarity to the other experts (correlations between 0.31 and 0.53).

Table 1
Descriptive statistics (mean, standard deviation [SD], median, mode, minimum, and maximum) and inter-rater reliabilities of gait quality traits (subjective valuating [SV] traits and linear profiling [LP] traits), estimated with intraclass correlation coefficients (absolute agreement $ICC_{(A,1)}$, consistency $ICC_{(C,1)}$, and their 95% confidence intervals [CIs]).

Trait	Descriptive statistics						Absolute agreement			Consistency		
	Mean	SD	Median	Mode	Minimum	Maximum	$ICC_{(A,1)}$	Lo. CI	Hi. CI	$ICC_{(C,1)}$	Lo. CI	Hi. CI
Walk												
Ground coverage (SV)	6.80	1.04	7.00	7	3	9	0.49	0.31	0.69	0.56	0.40	0.74
Overtracking (SV)	6.82	1.27	7.00	7	1	9	0.42	0.24	0.62	0.51	0.35	0.70
Suppleness and relaxation (SV)	6.50	1.34	7.00	6	2	9	0.43	0.27	0.63	0.46	0.29	0.65
Regularity and harmony (SV)	6.59	1.38	7.00	7	1	9	0.40	0.24	0.59	0.44	0.27	0.63
Activity and impulsion (SV)	6.76	1.33	7.00	7	2	9	0.39	0.22	0.60	0.50	0.34	0.68
Walk (LP)	6.79	1.20	7.00	7	2	9	0.37	0.20	0.59	0.47	0.30	0.67
Trot												
Ground coverage (SV)	6.71	1.08	7.00	7	3	9	0.45	0.27	0.65	0.55	0.38	0.73
Overtracking (SV)	6.38	1.27	6.00	7	2	9	0.37	0.21	0.57	0.44	0.28	0.64
Suppleness and relaxation (SV)	6.44	1.27	7.00	7	2	9	0.29	0.15	0.49	0.33	0.18	0.54
Regularity and harmony (SV)	6.57	1.22	7.00	7	2	9	0.30	0.15	0.50	0.38	0.22	0.58
Activity and impulsion (SV)	6.37	1.39	7.00	7	1	9	0.41	0.24	0.61	0.50	0.33	0.68
Trot (LP)	6.76	1.19	7.00	7	3	9	0.44	0.26	0.64	0.55	0.38	0.72
Trot: impulsion (LP)	6.41	1.32	6.50	7	1	9	0.36	0.20	0.56	0.42	0.26	0.62
Trot: elasticity (LP)	6.58	1.33	7.00	7	2	9	0.45	0.28	0.64	0.52	0.35	0.70

The inter-rater reliability of the traits based on the $ICC_{(A,1)}$ ranged from 0.29 to 0.49 for total agreement and from 0.33 to 0.56 for consistency (poor to moderate inter-rater reliability). The traits at the walk (Table 1, $0.37 < ICC_{(A,1)} < 0.49$) were more reliably assessed than those at the trot (Table 1, $0.29 < ICC_{(A,1)} < 0.45$). The linearly scored trait *walk (LP)* had the lowest reliability among the walk traits ($ICC_{(A,1)} = 0.37$), whereas the linearly scored traits *trot (LP)* and *trot: elasticity (LP)* had among the highest reliabilities of the traits scored at the trot ($ICC_{(A,1)} = 0.44$ and 0.45, respectively). The SV trait *ground coverage (SV)* had the highest reliabilities at the walk (ICC = 0.49) and at the trot (ICC = 0.45) but were still poorly reliable. Fleiss' kappa for more than two raters never exceeded 0.20, also indicating poor inter-rater reliability (Table S2).

4. Discussion

For all appraisals, the median and mode of the traits were between 6 and 7, with only few scores being below 5 or exceeding 8. The knowledge that the appraised horses were approved breeding stallions that were judged to be of high quality as 3-year-olds may have biased the experts to this anchoring toward higher than average scores (bias because of previous knowledge). Intuitively, this low variance should lead to higher probabilities of having the same score across multiple experts (the chance agreement increases) and, therefore, have higher reliabilities. ICCs compare the variance within each stallion with the variance across all stallions. Thereby, low ICC estimates indicate higher variance of scores within animals (between raters/experts) than between animals. However, a higher effective range of use of the scale is not a sufficient indicator for high ICCs. The ICC depends on the range of scores attributed by all experts in each trait. If only one or two experts use the whole scale (score 1 to 9 for expert #7 and 2 to 9 for expert #1), these experts will be outliers and decrease inter-rater reliabilities. In this study, the differences in scale anchoring not only had an effect on the absolute agreement ($ICC_{(A,1)}$; supported by the low κ values < 0.2) but also on the consistency ($ICC_{(C,1)}$), indicating it was not a simple offset in the values between experts. The cross-correlation matrix between experts showed that the agreement in scorings differed between pairs of experts. To increase reliabilities, the anchoring between experts needs to be standardized (same range and median) by more precisely defining the items on the score within each trait or at least the extremes and the median of the population. This should have the effect that the scale would be used more comprehensively (potential increase in variance between animals) and that the experts would differ less in absolute values.

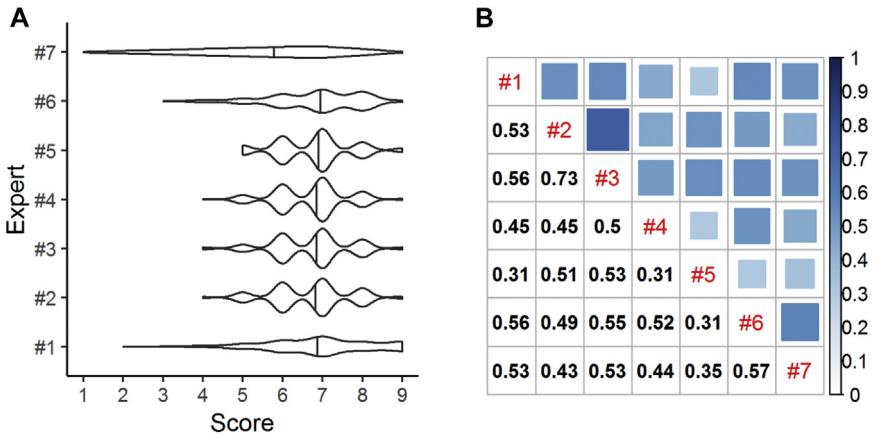

Fig. 1. (A) Violin plots of the scores given by each expert for the appraisal of 24 FM stallions at the walk and trot. The middle line represents the median. The width of the violin plot represents the number of times a score was used. (B) Cross-correlation matrix of the experts across all traits (LP and SV, at the walk and trot), with the corresponding numerical values below the diagonal.

Decreasing the scale range to seven items may also improve the distribution of scores, as with smaller scales raters tend to use the extreme items on the scale more [26].

In this study on the FM breed, the reliabilities of the few linearly scored traits, which the experts were more familiar with, were not systematically higher than the traits that received a valuating score. The reliability estimate for the linearly scored trait *walk (LP)* was lower than those estimated for the walk traits that had a valuating score, whereas the linearly scored trait *trot: elasticity* had the highest reliability estimate, together with the trait *ground coverage* (SV). Although LP traits might have the potential to be better than SV traits [6], we cannot see a substantial improvement in reliability of LP scores compared with SV scores based on our results. One potential reason for the small differences in reliabilities using an LP scale could be that the differences in scoring methods (SV vs. LP) are not sufficiently clear to the experts. In addition, the optimum scores for the LP traits for gait quality are all fixed at the score 9, further confusing the experts regarding the difference between scoring an SV trait from "bad" to "good" or an LP trait such as impulsion from "little" to "much," if "much" impulsion is also the optimum score.

In comparison to reliability studies available on other breeds, the poor inter-rater reliability of SV gait quality scores in our study was consistent with poor inter-rater reliability in valuating scores of conformation traits of 102 Lipizzan stallions appraised by a group of eight experts ($0.06 < \kappa < 0.24$) [11]. The study design was similar to the present study, that is, all experts appraised the same horses at the same time. The results from the study on the Spanish Purebred horse [12], in which experts have achieved excellent inter-rater reliability (mean ICC = 0.98) for linearly scored conformation traits of 876 horses are not directly comparable to the present study, as the reliabilities were estimated as pairwise comparisons of experts at different breeding competitions. At this time, because of the differences in study design, it is not possible to conclude whether the inter-rater reliabilities were higher in the Spanish Purebred horse because of the appraisal type (LP) in comparison to the inter-rater reliabilities of the Lipizzan horse (SV) or whether the sample size played a considerable role in the results. A large-scale study comparable to the Spanish Purebred horse is currently not possible in the FM horse, as each horse is linearly profiled only once in its life, by a pair of experts transmitting one consensus score to the FM breeding association. Further studies of reliabilities would be beneficial in understanding the quality of LP and SV scores in equine breeding schemes of different breeds. The studies could either be designed to use available data from breeding federations on the model of the Spanish Purebred horse study [12] or independently gather experts outside the breeding competition season to appraise the same number of horses at the same time (as was performed in the present study and those on the Lipizzan horse [11]).

Our study had a relatively small sample size compared with the abovementioned studies on conformation evaluations, yet has a similar sample size as reliability studies on lameness assessments ($0.17 < \kappa < 0.19$ for 13 raters and 24 horses [17], $\kappa = 0.41$ for 3 raters and 19 horses [21]). The reliabilities of gait quality traits in our study based on the κ statistics, all below 0.20, were similar or lower than those reported for lameness assessments. The scales used in lameness assessments usually have a detailed definition for each degree of lameness (e.g., [16], also discussed in [13]), which is not the case in the current scoring scales for gait quality (neither SV nor LP). To determine the degree of lameness, veterinarians are trained to observe specific aspects of locomotion, in particular, the movement of the head and pelvis [27,28], and basically assess only one trait during an examination—lameness. FM breeding experts, in contrast, have to assess several traits at the same time, and the trait definition may not be sufficient to discriminate between different degrees on a scale from 1 to 9. Although lameness assessments had a slightly higher reliability than the ones reported in our study for SV and LP traits, they remain in the range of poor ($\kappa < 0.40$) to fair ($0.41 < \kappa < 0.75$) reliability for Fleiss κ estimates. Some veterinarians have stated that lameness detection itself is very challenging [29]. The challenge is particularly high in cases of subtle lameness, for which the reliability is often lower than when considering horses with higher lameness degrees [30]. Scoring the gait quality of horses on an itemized scale using quantitative traits such as the ones presented in this study is expected to be similarly challenging to diagnosing horses with subtle lameness on an itemized scale. Although horses on both extremes of the spectrum should be easily identified, the differences between horses of similar gait quality are likely to be elusive. This is especially the case for prospective breeding stallions, which have undergone strong preselection beforehand and would likely show smaller differences in quality when being compared with one another.

Problematically, the stallion selection has potentially the largest impact on the FM population, as only few stallions are selected for breeding. A selection based on imperfect information may have large consequences on genetic diversity in the long term. Therefore, the low reliability of appraisals in our study is critical. The equine veterinary community [31,32] has largely gathered around the notion that field-applicable kinematic systems such as inertial measurement units may improve lameness assessments, especially in cases of subtle lameness, by contributing additional information that cannot be detected by the naked eye [33]. Given the low reliabilities in this study, it may be beneficial to have access to kinematic data related to gait quality of the horses presented in the field to improve decision-making in the context of breeding selections based on gait quality. However, the appropriate indicator traits (equivalents of head nod or hip hike for lameness assessments) would need to be defined in a kinematic study before the wide-scale field implementation of a quantitative gait measurement system.

5. Conclusions

In this study, the inter-rater reliability of scores for gait quality traits based on absolute values seemed to be independent from the type of scoring (linear traits did not have systematically and noticeably higher inter-rater reliabilities compared with valuating traits). Future studies would need to investigate whether expert training, more precise trait definitions, and a scoring scale using fewer items could improve the reliability of appraisal regardless of the type (SV or LP). However, considering the low reliability of lameness assessments by trained veterinarians, the improvement may turn out to be only modest. Quantitative gait measurement systems may provide new opportunities to define and measure gait quality traits in the future.

Acknowledgment

The authors thank all the judges participating in the experiments for their valuable time and insights and the private owners for providing the four RRFB stallions for the study's purpose. The authors would also like to thank Dr. Pierre-André Poncet, former director of the Swiss National Stud Farm and main instructor of the FM judges for his valuable comments with regard to the formal definitions of traits.

Financial disclosure

This study was funded by the Swiss Federal Office for Agriculture (FOAG) under contract numbers 625000469 and 627001325.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jevs.2020.102932.

References

[1] Jönsson L, Näsholm A, Roepstorff L, Egenvall A, Dalin G, Philipsson J. Conformation traits and their genetic and phenotypic associations with health status in young Swedish warmblood riding horses. Livest Sci 2014;163:12–25.
[2] Wallin L, Strandberg E, Philipsson J. Phenotypic relationship between test results of Swedish Warmblood horses as 4-year-olds and longevity. Livest Prod Sci 2001;68:97–105.
[3] Ducro B, Koenen E, Van Tartwijk J, Bovenhuis H. Genetic relations of movement and free-jumping traits with dressage and show-jumping performance in competition of Dutch Warmblood horses. Livest Sci 2007;107:227–34.
[4] Viklund Å, Näsholm A, Strandberg E, Philipsson J. Genetic trends for performance of Swedish warmblood horses. Livest Sci 2011;141:113–22.
[5] Stewart I, Woolliams J, Brotherstone S. Genetic evaluation of horses for performance in dressage competitions in Great Britain. Livest Sci 2010;128:36–45.
[6] Duensing J, Stock KF, Krieter J. Implementation and prospects of linear profiling in the warmblood horse. J Equine Vet Sci 2014;34:360–8.
[7] Viklund Å, Eriksson S. Genetic analyses of linear profiling data on 3-year-old Swedish warmblood horses. J Anim Breed Genet 2018;135:62–72.
[8] Rustin M, Janssens S, Buys N, Gengler N. Multi-trait animal model estimation of genetic parameters for linear type and gait traits in the Belgian warmblood horse. J Anim Breed Genet 2009;126:378–86.
[9] Koenen E, Van Veldhuizen A, Brascamp E. Genetic parameters of linear scored conformation traits and their relation to dressage and show-jumping performance in the Dutch warmblood riding horse population. Livest Prod Sci 1995;43:85–94.
[10] Druml T, Dobretsberger M, Brem G. The use of novel phenotyping methods for validation of equine conformation scoring results. Animal 2015;9:928–37.
[11] Druml T, Dobretsberger M, Brem G. Ratings of equine conformation–new insights provided by shape analysis using the example of Lipizzan stallions. Arch Anim Breed 2016;59:309–17.
[12] Sánchez M, Gómez M, Molina A, Valera M. Genetic analyses for linear conformation traits in Pura Raza Español horses. Livest Sci 2013;157:57–64.
[13] Dyson S. Can lameness be graded reliably? Equine Vet J 2011;43:379–82.
[14] Solnit R. River of shadows: Eadweard Muybridge and the technological wild west. Penguin; 2004.
[15] Van Weeren P. History. In: Back W, Clayton H, editors. Equine locomotion. 2nd ed 2013. p. 1–30.
[16] Hewetson M, Christley R, Hunt I, Voute L. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. Vet Rec 2006;158:852–8.
[17] Keegan K, Wilson D, Wilson D, Smith B, Gaughan E, Pleasant R, Lillich J, Kramer J, Howard R, Bacon-Miller C. Evaluation of mild lameness in horses trotting on a treadmill by clinicians and interns or residents and correlation of their assessments with kinematic gait analysis. Am J Vet Res 1998;59:1370–7.
[18] Keegan K, Dent E, Wilson D, Janicek J, Kramer J, Lacarrubba A, Walsh D, Cassells M, Esther T, Schiltz P. Repeatability of subjective evaluation of lameness in horses. Equine Vet J 2010;42:92–7.
[19] Keegan KG, Wilson DA, Kramer J, Reed SK, Yonezawa Y, Maki H, Pai PF, Lopes MA. Comparison of a body-mounted inertial sensor system–based method with subjective evaluation for detection of lameness in horses. Am J Vet Res 2013;74:17–24.
[20] Hammarberg M, Egenvall A, Pfau T, Rhodin M. Rater agreement of visual lameness assessment in horses during lungeing. Equine Vet J 2016;48:78–82.
[21] Fuller CJ, Bladon BM, Driver AJ, Barr AR. The intra-and inter-assessor reliability of measurement of functional outcome by lameness scoring in horses. Vet J 2006;171:281–6.
[22] Gmel AI, Druml T, Portele K, von Niederhäusern R, Neuditschko M. Repeatability, reproducibility and consistency of horse shape data and its association with linearly described conformation traits in Franches-Montagnes stallions. PLoS One 2018;13:e0202931.
[23] Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. Package 'corrplot'. Statistician 2017;56:316–24.
[24] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155–63.
[25] Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W. Package 'irr'. Various coefficients of interrater reliability and agreement. 2012. https://www.rdocumentation.org/packages/irr/versions/0.84.1. [Accessed 15 November 2019].
[26] Bardo J, Yeager S, Klingsporn M. Preliminary assessment of format-specific central tendency and leniency error in summated rating scales. Percept Mot Skills 1982;541:227–34.
[27] Dyson S. Recognition of lameness: man versus machine. Vet J 2014;3:245–8.
[28] Pfau T, Fiske-Jackson A, Rhodin M. Quantitative assessment of gait parameters in horses: useful for aiding clinical decision making? Equine Vet Educ 2016;28:209–15.
[29] de Mira M, Santos C, Lopes M, Marlin D. Challenges encountered by Federation Equestre Internationale (FEI) veterinarians in gait evaluation during FEI endurance competitions: an international survey. Comp Exerc Physiol 2019;15:1–8.
[30] Weishaupt M, Wiestner T, Hogg H, Jordan P, Auer J, Barrey E. Assessment of gait irregularities in the horse: eye vs. gait analysis. Equine Vet J 2001;33:135–40.
[31] Van Weeren P, Pfau T, Rhodin M, Roepstorff L, Serra Bragança F, Weishaupt M. Do we have to redefine lameness in the era of quantitative gait analysis? Equine Vet J 2017;49:567–9.
[32] Adair S, Baus M, Belknap J, Bell R, Boero M, Bussy C, Cardenas F, Casey T, Castro J, Davis W. Response to Letter to the Editor: do we have to redefine lameness in the era of quantitative gait analysis. Equine Vet J 2018;50:415–7.
[33] Bragança FS, Rhodin M, van Weeren P. On the brink of daily clinical application of objective gait analysis: what evidence do we have so far from studies using an induced lameness model? Vet J 2018;234:11–23.