Animal Science Papers and Reports vol. 38 (2020) no. 4, 355-368
Institute of Genetics and Animal Biotechnology
of the Polish Academy of Sciences, Jastrzębiec, Poland

Inter-rater reliability of evaluation of breed characteristics, structural correctness, quality of movements and body condition in Konik horses – application of rank-invariant method

Magdalena Sobczyńska, Tadeusz Jezierski*

Institute of Genetics and Animal Biotechnology of Polish Academy of Sciences,
Department of Animal Behaviour and Welfare, Jastrzębiec,
Postępu 36A, 05-552 Magdalenka, Poland

(Accepted October 23, 2020)

The aim of this study was to assess which of the subjectively evaluated traits of Konik horses demonstrate the highest systematic and random disagreement between judges. The inter-rater agreement in ratings among judges assessing characteristic breed type, structural correctness, movements in walk and trot and body condition of 215 Konik horses presented at the annual National Konik Show was assessed by calculation of Cohen's and weighted kappa coefficients, and by rank invariant method. By means of an augmented ranking, the systematic and random disagreement was measured. To analyse the systematic disagreement the relative position (RP), relative concentration (RC) and relative rank variation (RV) were assessed. The kappa of agreement were 0.23, 0.26, 0.29, 0.31 and 0.39 for trot, body condition, structural correctness, breed type and walk, respectively. Weighted kappa were slightly higher ranging from 0.31 (structural correctness) to 0.46 (walk). The highest RP values were obtained for body condition (-0.18) and type (-0.14), indicating that one rater systematically more frequently classified horses into higher categories than the other. The raters disagreed concerning the cut-off points when assessing walk and trot in the same scale (RP=0.12). Movement in walk was assessed more often in the upper categories of the scale than the trot. The RC values were small (ranging from -0.02 to 0.08) and mostly non-significant. The random difference was highest for the structural correctness (0.15), whereas low, though significant, for the characteristic breed type (0.03). From the relatively high level of systematic disagreement between judges in evaluation of body condition and breed type it could be concluded that the assessment criteria for these two traits should be more precise and/or judges should discuss what is a desirable

*Corresponding authors: t.jezierski@igbzpan.pl

model of the Konik breed type and body condition. The applied rank-invariant method shows that the subjective assessment of structural correctness and quality of movements, according to the current criteria in Konik horses can be acceptable.

KEYWORDS: horse assessment / inter-rater agreement / kappa coefficient / ordered categorical data / rank invariance

In horse breeding many traits are subjectively evaluated by scoring procedures in order to identify superior individuals. Higher evaluation scores for such traits as structural correctness, quality of movements and temperament, are believed to be correlated with performance capability, whereas traits as characteristic breed type are important for preservation of horse genetic resources. Other traits, as body condition tell about care of animals and thus about their welfare.

In some cases the ratings are on a continuous scale such as the stride length or distance walked in a known time. In other cases, however, raters' judgements are presented as discrete categories. These categories may be nominal ("present" or "absent") or ordinal (range of numerals indicating the ordered categorical structure of the responses). In each case, the categories are mutually exclusive and collectively exhaustive, so that each case falls in one category only. Rating scales are widely used for measuring variables such as breed type, structural correctness, and movement quality which are usually assessed by one authorized expert when culling a young horse or at qualification to the studbook. At horse shows, however, these traits in addition to body condition, care and preparedness for the show are usually judged by several judges. Although mean scores are taken for the final evaluation of the horse, if several judges evaluate the horse, the inter-rater agreement is an important aspect of the evaluation system. Discrepancies in scores given by different judges may have several causes. First, some traits may be not precisely defined or may consist of some different sub-traits or qualities. For example a trait labelled as structural correctness may be regarded as one trait composed of correctness of particular body parts as front/hind leg and hooves, head, neck, shoulder, barrel etc. Structural correctness also includes the way in which the horse's parts are put together. Movement correctness in particular gaits (walk, trot, canter) includes such particular traits as regularity, straightness, width and length of stride, snap and flexion. Moreover, the judges may have different experience in judging in general, or in judging a particular horse breed, or may have different ideas of an ideal horse. In addition, in majority of scoring systems the judges do not use the whole scale of scores, which reduces the variation between evaluated animals and makes application of some statistical methods problematic.

Data from rating scales represent a rank order only and not numerical values in a mathematical sense, even when the assessments are numerically labelled. These non-numerical properties of data from scale assessments imply that calculation of sums and differences is meaningless [Svensson 1993, Svensson and Holm 1994]. Consequently evaluation of change in an outcome variable that is measured by assessments on a rating scale cannot be based on differences [Svensson 1998]. The kappa coefficient is widely used in assessing categorical agreement between experts in the judgements of outcome variables. However, kappa value being strongly influenced by the prevalence of the outcome, can be counter-intuitive and can depend on number of categories [Xier 2010]. The rank-invariant method proposed by Svensson [1993] provides a new way to deepen the analysis of the agreement of two raters using ordinal scales. This method makes it possible to identify and measure different kinds of disagreements: related to the group (systematic) and caused by individual variability. Systematic disagreement can be reduced or taken into account when the reason for such disagreement is identified. A high level of additional individual variation indicates that the scale categories do not fit well to the rater or that the assessments are sensitive to disturbing factors of the test situation [Svensson 1998]. The study of Stachurska and Bartyzel [2011] showed that the horse's scores were not sufficiently reliable and should be processed by checking the quality of judging before using the results both in horse breeding or assessing sport performance and in further scientific analyses. A need of a precise examination of exterior traits of horse breeds is currently a common subject of discussion in scientific studies [Koenen *et al.* 2004, Kristjansson *et al.* 2013, Duensing *et al.* 2014, Druml *et al.* 2015]. Because the classification is not purely objective there is a need of objective methods to evaluate rater assessments within the refinement of classification procedures.

Our study aimed at investigation of the variability of classification and degree of agreement in ratings among two judges – both recognized horse specialist – assessing traits of Konik polski horses presented at the annual National Konik Show in Poland. The following traits were assessed: characteristic breed type, structural correctness of body conformation, movement quality in walk and trot as well as body condition and care/preparedness for the show. Konik horse is a Polish native breed of small horses believed to derive directly from the extinct East-European wild horse called Tarpan. This breed is regarded as a "primitive" breed, meaning that it has retained many traits both exterior and behavioural, from its wild predecessors. Konik breed has never been intensively selected for high performance as working or leisure horse, though it has retained many so called functional traits as good health, high reproduction indices, low feeding and keeping requirements and ability to survive "next to nothing". Nevertheless they show satisfactory trainability and are well suitable for leisure riding or driving. The selection system and thus the judging criteria in the Konik breed are mainly aimed not at improving the sport or working performance, but rather at preserving of the characteristic type of this breed and removing of traits that would be generally unfavourable for the existing suitability for leisure riding and driving. Subjective assessments of Konik horses both at qualification for the studbook and on horse shows, are based on scales, whereas the judgements are qualitative and the measurements are not standardized. Despite of a general acceptance and application of this system, its reproducibility and validity have never been tested using more advanced methods. The present study identifies and measures systematic disagreement related to the group, when present, separately from disagreement caused by individual variability in assessments. A good agreement between raters allows judgements to be made by different raters with some confidence in their consistency, whereas a poor agreement can indicate deficiencies in classification system which may mean that there is a need for refinement of definitions, or improving training of the raters.

Material and methods

Dataset

A total of 215 Konik horses, including 85 stallions and 130 mares that had been exhibited at the annual National Konik Shows over a period of 8 years, were evaluated by a commission consisted of three recognized experts specialized in this breed. For the organizational reason two of the experts (judges) were the same persons at all 8 events and the third person kept changing in consecutive years. As the rank-invariant method is applicable for two raters only, the two judges who evaluated the horses during all the 8 National Shows were considered in this study. The horses derived from 4 state-owned studs and from 12 private breeders and were preselected for the National Konik Show to eliminate individuals of poor quality. The horses were presented to the commission and to the audience by grooms or by owners on a special place outdoors the stable, to enable a thorough evaluation both when standing still and in motion. The five main traits that were evaluated include: (1) characteristic breed type taking into account coat colour, sex differences, proportions and generally to what extend a horse is the ideal representative of the breed, (2) structural correctness including imperfections of all body parts, (3) movement in walk, (4) movement in trot including gait straightness, width, length, snap and flexion and (5) body condition, fitness, general health, care and preparedness for the show. The judges gave their individual scores independently for each of the 5 traits of a horse, using a scale of up to 10 points for each of the traits, totalling a maximum of 50 points for a horse. Usually not the full available scale of 0-10 points but rather of 5-10 points was applied.

Data analysis

Agreement between raters was assessed by calculation of the Cohen's and weighted kappa coefficients, and using the rank invariant method. The kappa coefficient was computed using the formula:

$$\kappa = \frac{P(o)-P(e)}{1-P(e)}$$

where P(o) denotes the observed percentage of agreement, and P(e) denotes the probability of expected agreement due to chance. The weighted kappa was obtained by giving weights considering disagreement. The weights (w_{ij}) were calculated by the rule suggested by Cicchetti and Allison [1971]:

$$w_{ij} = 1 - \frac{|i-j|}{k-j},$$

where k is the total number of response categories, i=1,…k and j=1,…k. Weights are assigned to each cell and their value range is $0 \leq w_{ij} \leq 1$. The cells on the diagonal (i=j) are given the maximum value, $w_{ij} = 1$. Differences in raters' marginal rates were assessed with Bhapkar test of marginal homogeneity [Uebersax 2006]. This statistics is interpreted as a chi-squared value. The proportions of cases below each category were calculated and equality of the rater 1 and rater 2 thresholds for each category was tested using the McNemar test. To obtain more information needed to assess rater proficiency in scoring categorical responses, the rank-invariant method for inter-scale comparison, described by E Svensson, was applied [Svensson 1997, Avdic and Svensson 2010]. By means of an augmented ranking approach, an observed disagreement was separated and measured in terms of systematic and random disagreement. Two measures of systematic disagreement between raters were used: relative position (RP) and relative concentration (RC). The measure of RP expresses the extent to which the marginal distribution of rater Y is shifted towards higher categories than the marginal of rater X. Therefore, RP is the difference between the probabilities P_{xy} and P_{yx}:

$RP = P_{xy} - P_{yx}$, where:

$$P_{xy} = \frac{1}{n^2}\sum_{i=1}^{m}[y_i \cdot C(X)_{i-1}]$$

$$P_{yx} = \frac{1}{n^2}\sum_{i=1}^{m}[x_i \cdot C(Y)_{i-1}]$$

m – the number of scale categories;

n – the number of individuals;

x_i and y_i – the *i*th category frequencies of marginal distributions of X and Y;

$C(X)_i$ and $C(Y)_i$ – the *i*th category cumulative frequencies.

The measure of RC expresses the extent to which the marginal distribution of Y is more concentrated to central scale categories than the marginal of X, $P(X_1 < Y_k < X_0)$

$$RC = \frac{1}{Mn^3}\sum_{i=1}^{m}\{y_i \cdot C(X)_{i-1}[n - C(X)_i] - x_i \cdot C(Y)[n - C(Y)_i]\}, \text{wh}$$

where M = minimum value of $(p_{xy} - p^2_{xy})$ and $(p_{yx} - p^2_{yx})$ provided $0<(p_{xy}$ and $p_{yx})<1$.

The values for RP and RC range from -1 to 1 and value of 0 means that there are no systematic changes, while a value of 1 or -1 means that there are systematic differences. Measures of systematic disagreement between raters are based solely on the marginal distribution and do not entirely explain the pattern of disagreement in paired assessments. Often there is an individual heterogeneity in the results in addition to the systematic disagreement in inter-rater reliability studies. To estimate the contribution of the individual variation to the pattern of disagreement, firstly an augmented mean rank procedure was used. The augmented ranking was defined by the observation in the (i,j)th cell of the contingency table according to values for rater X as:

$$\bar{R}_{ij}^{(X)} = \sum_{k=1}^{i-1}\sum_{l=1}^{m} x_{kl} + \sum_{l=1}^{j-1} x_{il} + \frac{1}{2}\left(1 + x_{ij}\right)$$

for i ≥ 1, m ≥ j where x_{ij} is the ijth cell frequency. The corresponding augmented mean ranks according to values for rater Y( ) where similarly defined. An empirical measure of random differences between two ordered categorical judgements of the same individual, called the relative rank variance (RV) was defined by:

$$RV = \frac{6}{n^3}\sum\sum x_{ij}\left[\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)}\right]^2$$

RV ($0 \leq RV < 1$) expresses the level of disagreement from a total agreement in rank ordering, given the marginals. RV < 0.1 would in general be regarded as negligible [Avdic and Svensson 2010].

Percentage agreement (PA) was calculated as the number of agreement scores divided by the total number of scores. The proportion disordered pairs out of all possible combination of pairs defines the measure of disorder (D). The level of homogeneity of individual disagreement relative to systematic disagreement is measured by the correlation of the augmented mean ranks, r_a:

$$r_a = 1 - \frac{n^3}{n^3 - n} RV$$

Standard agreement measures and results of the marginal homogeneity test are presented in Table 1. The kappa measures of agreement were satisfactory, ranging from 0.23 (movement in trot) to 0.39 (movement in walk) and the weighted kappa were slightly higher ranging from 0.31 (structural correctness) to 0.46 (movement in

Table 1. Common agreement measures and test of marginal homogeneity (MH) for different traits assessed in Konik horses

Measure/Trait	Type	Structural correctness	Body condition	Movement in walk	Movement in trot
Kappa	0.31	0.29	0.26	0.39	0.23
Weigted kappa	0.38	0.31	0.32	0.46	0.34
Percentage agreement	66	52	61	60	51
Test of MH					
Chi-square	30.3	3.7	29.0	21.1	20.1
degrees of freedom	6	4	5	6	6
p-value	0.00	0.44	0.00	0.00	0.00

walk). The proportion of perfectly agreeing pairs, i.e. the percentage agreement was highest for breed type (66%), moderate for body condition and movement in walk (about 60%) and lowest for structural correctness and movement in trot (about 50%). The test of marginal homogeneity did not reject the null hypothesis of marginal homogeneity only for the structural correctness. Therefore two raters were similar in terms of how often they use each category when rating the conformation of the same horse. Usually, only five categories from the whole rating scale of structural correctness evaluation were used.

Table 2. Measures of systematic and individual-based disagreement for different traits assessed in Konik horses (se – standard errors)

Measure/Trait		Type	Structural correctness	Body condition	Movement in walk	Movement in trot
Systematic disagreement						
in position	RP (se)	-0.14(0.03)	0.06 (0.04)	-0.18 (0.03)	0.12 (0.03)	0.12 (0.03)
in concentration	RC (se)	0.08 (0.03)	-0.02 (0.04)	-0.05 (0.04)	0.08 (0.04)	-0.04 (0.04)
Individual disagreement						
relative rank variance	RV (se)	0.03(0.01)	0.15 (0.03)	0.05 (0.02)	0.10 (0.03)	0.09 (0.02)
disorder	D	0.05	0.14	0.07	0.10	0.10
correlation of the augmented ranks	r_a	0.97	0.85	0.94	0.90	0.91

As shown in Table 2, the main part of disagreement can be explained by the systematic difference (RP). The systematic difference in position between the raters was -0.18 and -0.14 for body condition and type assessments, respectively, whereas for movement in walk /trot assessment the same value of RP was obtained (0.12). It means that the raters disagreed concerning the cut-off points. A negative RP values for type and body condition indicates that the rater X systematically more frequently classified horses into higher categories than the rater Y. On the other hand, when movement assessment was considered, the rater X classified more horses into lower categories than the rater Y. The measure of RP for structural correctness was small (0.06) and insignificant. The RC values were also small (from -0.02 to 0.08) and mostly insignificant indicating that none of the marginal distributions is concentrated to the central scale categories, so the raters have similar ideas of the cut points between categories in the middle of the scale. Apart from the systematic difference, the significant values of RV can also reflect random difference, which indicates that the scale categories do not fit well to the raters or that the assessments are sensitive to disturbing factors of the test situation. The level of random difference was highest for the structural correctness (0.15), whereas low, but significant, for the type (0.03). The measure of disorder (D) show that 14% of all possible pairs are disordered in case of the structural corectness. Corresponding measures for movements, body condition and type were 10, 7 and 5%, respectively. The correlation of the augmented mean rank was highest for type (0.97) and lowest for structural correctness (0.85).

The graphs 1-5 show cumulative proportions of cases below each rating level for each rater assessing different traits. The locations of a rater's thresholds determine how often the rater uses each rating category. Threshold locations do not differ between raters for structural correctness (Fig. 2). Rater Y has a higher threshold category 5 and 6 for breed type (Fig. 1). This corresponds to a wider definition of the lower rating categories and a narrower definition of the higher rating categories. Similar pattern was observed for body condition (Fig 3). The highest differences between raters refer to category 3 and 4- rater Y has higher threshold for these categories than rater X.

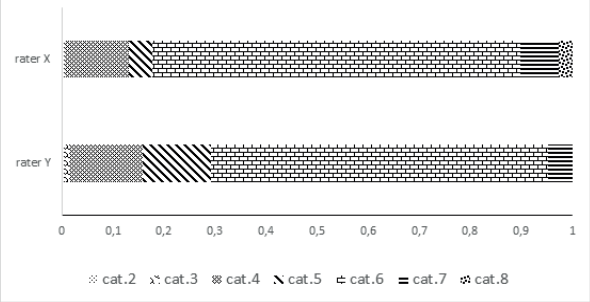

Fig.1. The cumulative proportion of cases below each category for breed type.

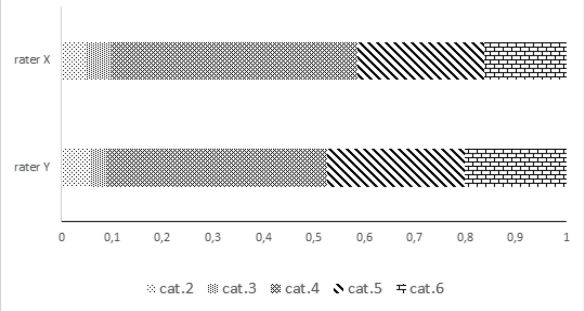

Fig.2. The cumulative proportion of cases below each category for structural correctness.

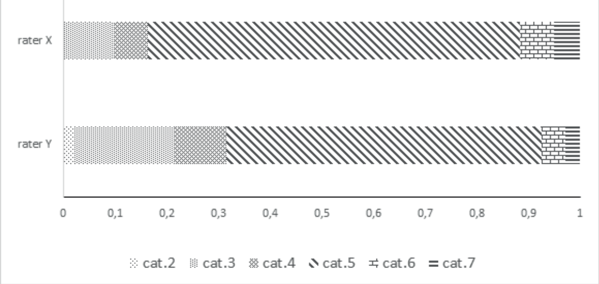

Fig. 3. The cumulative proportion of cases below each category for body condition.

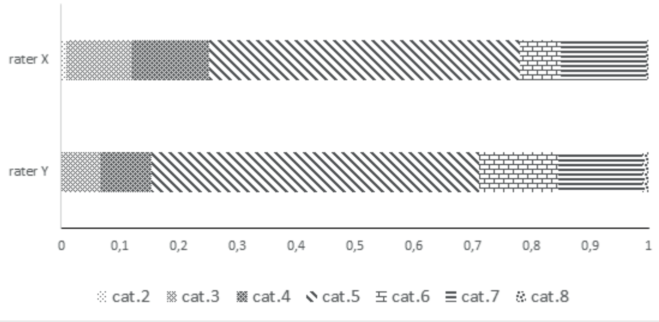

Fig. 4. The cumulative proportion of cases below each category for movement in walk.

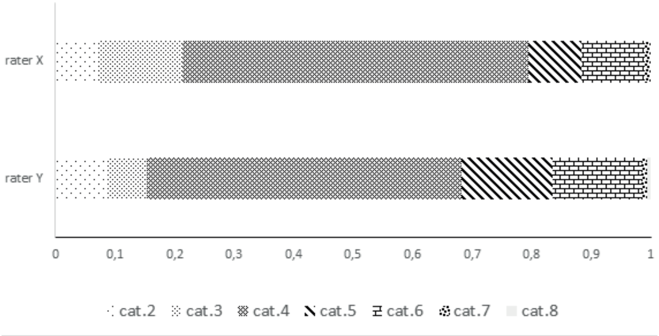

Fig. 5. The cumulative proportion of cases below each category for movement in trot.

Threshold patterns were similar for movements in walk (Fig. 4) and trot (Fig. 5), however, movement in walk was assessed more often in the upper categories of the scale than the trot.

Evaluation of body conformation, in particular the structural correctness, is an obligatory procedure for all horse breeds at qualification for stud books or at horse shows for choosing of champions. Since the creation horse breeding programs it has been assumed that there is a strong relation between body conformation and working, or racing, or sport performance of horses. The evaluation system should, however, take into account breed-specific traits and purpose of a breed. Also, the evaluation system at qualification for studbooks is not identical with that applied at horse shows or exhibitions.

For example, at the qualification for studbooks the structural correctness is evaluated separately for the main body parts (head and neck, trunk, forelegs, hind legs, hooves) and instead of the trait labelled generally as body condition, fitness, general health, care and preparedness for the show, the trait labelled as general appearance is evaluated. Separate assessment of head, trunk or legs, allows the raters to focus attention on a specific body part, which seems to provide more information about the horse conformation. However, in this case the probability of inconsistencies between the judges increases. Evaluation of an overall structural correctness can be more difficult and problematic, because it requires a holistic look at the horse. One judge may attach more importance to the assessment of the head than to the trunk, while the other judge vice versa. This may be the reason for relative high individual disagreement (RV) but low systematic disagreement (RP) for structural correctness. According to results obtained by Druml *et al.* [2015], the scale or the number of traits are not the limiting factors in equine conformation assessment and the highest impact for the validation of ranking scores on a biological scale is the consistency between ratings and raters. Druml *et al.* [2016] underline the difficulty in assessing morphological traits (neck, withers, shoulder, chest, croup and legs) and the low correlations between experts' classification rankings assessed in a test situation. On the other hand, Sanchez *et al*. [2013] showed a high level of reproducibility of analyzed conformation traits (ICC>0.9) in Andalusian horse.

In horse breeding, selection is strongly based on the type concept, but the trait „type" itself can not be easily described using a morphological or biological scale [Druml *et al.* 2015]. Type represents a complex trait including several levels of information. Results obtained by Druml *et al.* [2016] showed a higher agreement for type traits (breed and sex type, harmony) than for morphological traits (kappa=0.27 vs. 0.14, respectively). This findings may indicate that raters have an overall idea of how a horse represents its particular breed or sex, but they disagree in the assessment of the individual body parts. In our study inter-rater agreement for type was fair (according to scale proposed by Landis and Koch [1977], weighted kappa=0.38) and random part of disagreement which cannot be explained by the systematic difference showed the smallest value (RV=0.03) giving the highest correlation of the mean augmented ranks (r_a=0.97). However, different marginal distributions (RP=-0.14) are signs of some systematic disagreement between raters. When assessing the type, one rater defines the lower rating categories in a wider range.

Traditionally, the breeders evaluate the movement of horses visually. Watching the horse in motion is a standard practice for both quality of gaits and lameness examination. However, the human eye is only capable of registering images with low frequency of 20 Hz. This makes the human ability insufficient for a consistent and objective evaluation of the functioning of the horse's locomotion system, especially when lameness has to be diagnosed and to a lesser extend when predicting a horse's performance. Results obtained by Hammarberg *et al.* [2016] indicate that visual lameness assessment of horses in trot in a circle (video recordings) has moderate agreement for experienced raters (kappa=0.38), which corresponds with average kappa values found in our study for walk (0.46) and trot (0.34) quality, and poor (kappa=0.25) agreement for less experienced equine veterinarians. Disadvantage of video gait analysis is inability to change horse speed or angle of view. On the other hand, video recordings can be replayed and raters can evaluate the same sequence many times. When experienced equine practitioners have assessed lameness in horses in real time, over ground setting [Keegan *et al*. 2010], inter-rater agreement was high (kappa=0.86) in case of clear lameness (mean score >1.5 in AAEP scale)) but when the mean score was <1.5 the agreement was poor (kappa=0.23). Therefore study of Keegan *et al*. [2010] suggested that for horses with slight lameness a subjective evaluation is not very reliable. Olsen *et al*. [2014] came to similar conclusion comparing the gait assessment component in neurological examination of horses. They found that the agreement on gait grading (0-4) during ataxia examination was good overall (intra-class correlation coefficient ICC=0.74), but very poor for grades ≤1 (ICC=0.08) i.e. in normal horses or those with subtle ataxia. Similarly, there was a worse agreement between raters for their assessment of moderate to high-grade group, compared to assessment of horses with higher grades of lameness. In our study horses were preselected and probably represented a population that was free from incorrectly moving animals. In other studies [Fuller *et al*. 2006, Keegan *et al*. 2010] horses might be included in the examination because trainers had suspicion of lameness in these horses and trait prevalence in such population will be higher than in population of randomly selected horses. The consequence of population homogeneity may be a reduced range of the scores used by judges. So it might be necessary to train judges to use the full scale in order to collect the maximum level of variation from the assessed population.

In studies on agreement, two or more raters independently judge each object. Although independently evaluated such data are dependent, as each individual was assessed twice. These dependencies in data must be accounted for and particular attention must be paid to the invariance property in the case of repeated measurements, as it makes no sense of calculating differences between categorical labels [Svensson 1993, 1997]. In most studies evaluating classifier ratings of equine conformation traits the analysis of variance, the mixed models or kappa statistics were applied [Keegan *at al*. 2010, Sánchez *et al*. 2013, Olsen *et al*. 2014, Druml *et al*. 2015]. However, it has to be stated that the assumptions for these approaches in smaller datasets may be violated due to heterogeneity of variance among classifiers and due to the discrete scale of trait descriptors, where scores are not normally distributed [Druml *et al*. 2015].

One of the most commonly used measures of the inter-rater agreement is Cohen's kappa measure [Cohen 1960]. If observations are classified in more then two categories, the possibility of disagreement increases and weighted kappa measure has been proposed [Cohen 1968] with different types of weights. However, widely-used kappa statistics can be misleading in many cases, especially when prevalence and bias effects exist [Xier 2010]. When trait prevalence is high, calculation of expected chance agreement is also high and it is more difficult to achieve high agreement above the chance. Moreover kappa measure depends on the number of categories. Comparing the overall kappa values we can only know that raters have better agreement assessing walk compared to trot or body condition. But we cannot say which rater is more accurate than the other via kappa values. The values of weighted kappa depend of the choice of weights but this choice is subjective. Thus, the weighted kappas can be different in the same investigation. Weighted kappa is usually higher than unweighted kappa because disagreements are more likely to concern only one category than several categories. The weighting procedure ignores the rank-invariant properties of ordinal data. There are different interpretation of the strength of agreement when the same kappa values are obtained [Landis and Koch 1977, Altman 1991, Fleiss *et al.* 2003].

These recommendations are just rules that, however, are not based on proper scientific rationale. Therefore the question of how the magnitude of kappa should be judged is still open. So, we can claim that in the case of the breed type, structural correctness and trot examination, there was a poor agreement between raters according to Fleiss *et al.* [2003], but fair according to Landis and Koch [1977].

Some parameters used in the agreement studies are measures of concordance and association for ordered categorical variables such as Kendall's tau-b or Spearman's rank-order correlation. The augmented rank order agreement coefficient coincides with Spearman's rank correlation coefficient for untied observations. Moreover Spearman's rank correlation coefficient may be used as a reliability measure provided that there is marginal homogeneity. This strong limitation for its use as a reliability measure means that there is a high risk of inappropriate application [Svensson 1997]. A correlation coefficient measures a degree of association between two variables and does not measure the level of agreement in assessments of the same variable within or between individuals, as demonstrated by Svensson [2012]. The misuse of correlation coefficient in reliability studies could have serious consequences on conclusion and decisions, because a strong correlation runs the risk of hiding biased, unreliable assessments [Altman 1991, Svensson 2012]. Correlation of the augmented ranks measures of reliability in ordering irrespective of the marginal distribution, as the augmented ranking procedure provides adjustment for systematic disagreement. In general, measures of association are only adequate as measures of agreement if the marginal distributions are similar. Thus correlations between judges assessed the structural correctness of cold-blooded horses obtained by Polak and Lewczuk [2018] correspond well with correlation of the augmented ranks obtained in the present study (0.73-0.94 vs 0.85).

Druml *et al.* [2015] presented a novel method based on image analysis, which offers the possibility to evaluate the association of individual ratings made from classifiers who are experts in evaluation of body conformation with the shape of horses. This method can be helpful in both trait definition and rating evaluation. By augmented ranking approach in which the ranks are tied to the pairs it is possible to perform a comprehensive evaluation of the sources of disagreement in paired ordinal assessments. This approach enables to separate the inconsistency into random and systematic errors and to quantify this lack of consistency in a few measures. These give more detailed descriptions of the variability than does the kappa coefficient. The method was applied to ordinal data, but is also suitable for equidistant or continuous data. The advantage of statistical methods that do not require distributional properties of data, are that the results are reliable and valid without restrictions and may also be used for small populations. Furthermore, the possibility of separating the disagreement into both systematic and individual components is important in horse evaluating system. The one component measuring the systematic effect indicate a systematic disagreement in how the raters interpret the scale categories; the other component measuring individual effect concerns additional random variation.

The present study, conducted on Konik horses, a breed participating in the program of conservation of horse genetic resources, identifies agreements or disagreements in subjective assessing of traits that are subjected to a rather moderate stabilizing selection than an intensive directional selection. The results of our study may contribute to improvement of the traditional evaluation method of individuals of this breed. However, assessing of traits in other horse breeds may also benefit from the application of rank-invariant method.

Conclusions

from the satisfactory level of the inter-rater agreement when conformation traits and movement quality of Konik horses were assessed, it could be concluded that the two raters had similar skills or were similarly trained for evaluation of Konik horses. By using the rank-invariant method we concluded that the main reason for the lack of agreement between two raters in evaluation of horses, may consist in their interpretation of category description, especially in the case of body condition and breed type traits. This poor agreement can be considerably reduced by specifying the category description and training the raters. The poorest closeness of observations from the best possible agreement in ordering when the marginal heterogeneity is taken into account, was found for the structural correctness. It is highly recommended to standardize the definition of the body condition trait to prevent differences in interpretation between raters.

REFERENCES

1. ALTMAN D.G., 1991 – Practical statistics for medical research. Chapman and Hall, London, 403-415.
2. AVDIC A., SVENSSON E., 2010 – Svenssons method 1.1 ed. Interactive software supporting Svenssons method. http://avdic.se/svenssonsmetod.html.(accessed 5 June 2018).
3. CICCHETTI D.V., ALLISON T., 1971 – A new procedure for assessing reliability of scoring EEG sleep recordings. ***American Journal of EEG Technolo*****gy** 11, 101-109.
4. COHEN J., 1960 – A coefficient of agreement for nominal scales. ***Educational and Psychological Measurement*** 20(1), 37-46.
5. COHEN J., 1968 – Weighted kappa:nominal scale agreement with provision for scaled disagreement or partial credit. ***Psychological Bulletin*** 80(4), 213-220.
6. DRUML T., DOBRETSBERGER M., BREM G., 2015 – The use of novel phenotypic mehods for validation of equine conformation scoring results. ***Animal*** 9, 928-937.
7. DRUML T., DOBRETSBERGER M., BREM G., 2016 – Rating of equine conformation – new insights provided by shape analysis using the example of Lipizzan stallions. ***Archives Animal Breeding*** 59, 309-317.

8. DUENSING J., STOCK, K.F., KRIETER J., 2014 – Implementation and prospects of linear profiling in the Warmblood horse. ***Journal of Equine Veterinary Science*** 34, 260-368.
9. FLEISS J., LEVIN B., PAIK M., 2003 – Statistical Methods for Rates and Proportions, 3rd Edition, Wiley & Sons, New York.
10. FULLER C.J., BLADON M.B., DRIVER A.J., BARR A.R.S., 2006 – The intra- and inter-assessor reliability of measurementof functional outcome by lameness scoring in horses. ***Veterinary Journal*** 171, 281-286.
11. HAMMARBERG M., EGENVALL A., PFAU T., RHODIN M., 2016 – Rater agreement of visual lameness in horses during lungeing. ***Equine Veterinary Journal*** 48, 78-82.
12. KEEGAN K.G., DENT E.V., WILSON D.A., JANICEK J., KRAMER J., LACARRUBBA A., WALSH D.M., CASSELLS M.W., ESTHER T.M., SCHILTZ P., FREES K.E., WILHITE C.L., CLARK J.M., POLLITT C.C, SHAW R., T. NORRIS T., 2010 – Repeatability of subjective evaluation of lameness in horses. ***Equine Veterinary Journal*** 42, 92-97.
13. KOENEN E.P.C., ALDRIDGE L.I., PHILIPSON J. 2004 – An overview of breeding objective for warmblood sport horses. ***Livestock Production Science*** 88, 77-84.
14. KRISJANSSON T., BJONSDOTTIR S., SIGURDSSON A., CREVIER-DENOIX N., POURCELOT P., ARNASON T., 2013 – Objective quantification of conformation of the Icelandic horse based on 3-video morphometric measurements. ***Livestock Science*** 158, 12-23.
15. LANDIS J.R., KOCH G.G., 1977 – The measurement of observer agreement for categorical data. ***Biometrics*** 33, 59-174.
16. OLSEN E., DUNKEL B., BARKER W.H.J., FINDING E.T.J., PERKINS J.D., WITTE T. H., YATES L.J., ANDERSEN P.H., BAIKER K., PIERCE R.J., 2014 – Rater agreement on gait assessment during neurologic examination of horses. ***Journal of Veterinary Internal Medicine*** 28, 630-638.
17. POLAK G.M., LEWCZUK D., 2018 – The stability of conformation and movementtraits evaluation tested in cold-blooded horses ofdifferent endangerment status. ***Journal of Applied Animal Research*** 46, 547-551.
18. SANCHEZ M.J., GOMEZ M.D., MOLINA A., VALERA M., 2013 – Genetic analyses for linear conformation traits in Pura Raza Español horses. ***Livestock Science*** 157, 57-64.
19. STACHURSKA A., BARTYZEL K., 2011 – Judging dressage competitionsin the view of improving horse performance assessment, ***Acta Agriculturae Scandinavica, Section A – Animal Science*** 61, 92-102.
20. SVENSSON E.. HOLM S., 1994 – Separation of systematic and random differences in ordinal rating scales. ***Statistics in Medicine*** 13, 2437-53.
21. SVENSSON E., 1993 – *Analysis in systematic and random differences between paired ordinal categorical data*. Almqvist & Wiksell International, Stockholm.
22. SVENSSON E., 1997 – A coefficient of agreement adjusted for bias in paired ordered categorical data. ***Biometrical Journal*** 39, 643-657.
23. SVENSSON E., 1998 – Ordinal invariant measures for individual and group changes in ordered categorical data. ***Statistics in Medicine*** 17, 2923-2936.
24. SVENSSON E., 2012 – Different ranking approaches defining association and agreement measures of paired ordinal data. ***Statistics in Medicine*** 31, 3104-3117.
25. UEBERSAX J.S. 2006 – *User guide for the MH program (v. 1.2). Statistical Methods for Rater Agreement*website. http://john-uebersax.com/stat/mh.htm. (accessed 6 June 2018).
26. XIER L., 2010 – Kappa – a critical review. Uppsala University, Department of Statistics.
27. http://uu.diva-portal.org/smash/get/diva2:326034/FULLTEXT01.pdf (accessed 6 June 2018).