Arch. Anim. Breed., 59, 309–317, 2016
www.arch-anim-breed.net/59/309/2016/
doi:10.5194/aab-59-309-2016


Open Access
Archives Animal Breeding

# Ratings of equine conformation – new insights provided by shape analysis using the example of Lipizzan stallions

**Thomas Druml, Maximilian Dobretsberger, and Gottfried Brem**

Institute of animal breeding and genetics, Veterinary University Vienna, Veterinärplatz 1, 1220 Vienna, Austria

*Correspondence to:* Thomas Druml (thomas.druml@vetmeduni.ac.at)

Received: 13 January 2016 – Revised: 15 June 2016 – Accepted: 17 June 2016 – Published: 28 June 2016

**Abstract.** The quality of individual ratings of conformation traits can commonly be evaluated by calculating inter-rater correlations and repeatability coefficients. We present an approach in which we associate the individual rating scores with the underlying horse shapes derived from standardized images, performing a shape regression. Therefore, we analyzed the shape of 102 Lipizzan stallions from the Spanish Riding School in Vienna, defined by 246 shape-correlated two-dimensional coordinates using techniques from the field of image analysis and geometric morphometrics. In addition we examined the differences in the conformation classifiers' perceptions of type traits and functional traits. In this study part, the rating scores of eight conformation classifiers were tested for agreement, yielding inter-rater correlations ranging from 0.30 to 0.55 and kappa coefficients ranging from 0.08 to 0.42. From the 12 scoring traits assessed on a valuating scale, type traits with a mean kappa coefficient ($\kappa$) of 0.27 demonstrated a higher agreement than functional traits ($\kappa = 0.14$). Based on 246 two-dimensional anatomical and somatometric landmarks, the shape variation was analyzed by the use of generalized orthogonal least-squares Procrustes (generalized Procrustes analysis – GPA) procedures. Shape variables were regressed into the results from visually scored linear type trait classifications (shape regressions). From the 48 performed shape regressions (eight classifiers, six traits), 42 % resulted in a significant equation. In 58 % of the ratings, no association between scores and the phenotype of the horses was found. Phenotypic differences of model horses along significant regression curves of mean ratings and individual ratings were exemplarily visualized and compared by warped and averaged images. Finally, we demonstrated that the method of shape regression offers the possibility to evaluate the association of individual ratings from expert conformation classifiers with the shapes of horses. The detected bias in classifiers' rankings have not been considered in breeding programs, and its impact on selection procedures still needs further research.

## 1 Introduction

This paper deals explicit with the transformation process of visual perceptions of equine conformation into numerical scoring data, which is commonly titled conformation scoring. In horse breeding, different conformation traits are routinely assessed by visual scoring procedures, which allow for the prediction of future performance at an early age. Such "secondary" or "indicator" traits can be split into three groups: functional traits, type traits and locomotion traits. Functional traits are based mostly on empirically defined anatomic and orthopedic criteria. In human sport science the visual assessment of such traits is called "interpretative anatomy" (Kilgore, 2012). In contrast to this anatomical approach, type trait definitions are based on ideal imaginations. They may differ from breed to breed, from classifier to classifier and between the sexes. Type (from the Greek word týpos, figure) represents a complex trait including several levels of information. Therefore, ratings such as type, overall expression, breed type and harmony cannot be considered to be purely empirical, as they are principally based on likes or dislikes. In several European breeds which are being bred on a small scale, often embedded into conservation programs (Bodo et al., 2005), the maintenance of a specific type represents a major part of breeding objectives.

Conformation scoring is performed in two manners, either on a biological scale (linear profiling) or on a valuat-

ing scale (traditional approach). A characteristic of valuating scoring systems is the strong correlations between conformation traits, which are due to overlapping morphological segments and overlapping trait definitions (Druml et al., 2008). In order to avoid such overlaps and to overcome the problem of subjective scoring on a valuating scale, linear profiling systems were introduced from the 1990s (cf. data and references given in Duensing et al., 2014). Here breeding-goal-related traits are visually assessed relative to biological extremes. In the presented study the conformational evaluation is based on a valuating scale which is routinely used in the breeding of Lipizzan horses and which is also the standard protocol for the Austrian federal stud farm of Piber (see Table S1 in the Supplement).

In conformation scoring processes the extent of subjectivity can be measured by inter-rater agreement expressed by correlation coefficients. Although this matter has been extensively studied in other species and also in humans (Veerkamp et al., 2002; Kristensen et al., 2006; Janssens et al., 2004; Cunningham et al., 1995) and has been frequently discussed within the equine scientific community, it has not yet been sufficiently addressed in equine scientific literature. This lack of documentation may be due to the central position conformation classifiers have within the framework of equine selection procedures. In the few available studies dealing with this question, Koenen et al. (1995) report "large differences in scores between traits and classifiers" without giving further numerical information, and Grundler and Pirchner (1991) and Sanchez et al. (2013) calculated repeatabilities and/or reproducibility scores for single raters and rater pairs.

In a previous study based on conformational evaluations of 44 Lipizzan mares (Druml et al., 2015), we analyzed agreement between six raters and found moderate to low values for inter-rater correlations, which also differed between type traits and functional traits. For this reason we expanded the analysis of inter-rater agreement to 102 Lipizzan stallions from the Spanish Riding School in Vienna and to the ratings of eight professional classifiers.

The aim of this study was to analyze the visual process of conformation scoring of the standing horse model. Based on the standard scoring protocol of the federal stud farm of Piber we quantified the inter-rater agreement and evaluated each individual type-related rating. This evaluation of individual ratings of conformation traits can be done by shape regressions. The individual horse shape, encoded in unified, centered and Procrustes rotated coordinate data derived from standardized digital images, is regressed into the scores given by the individual classifiers. Significant regressions account for a consistent and correct rating and non-significant regressions indicate inconsistencies and a poor association of individual visual perceptions with the horses' shapes. Further, we applied different visualization techniques in order to depict significant differences of individual ratings.

**Table 1.** Mean values, minimum and maximum values, standard deviations, and range of 12 conformation traits evaluated by eight classifiers using a scale of 10 points with one unit increases for 102 Lipizzan stallions.

| Variable | *N* | Mean | Min. | Max. | SD | Range |
|---|---|---|---|---|---|---|
| Type | 816 | 7.83 | 6 | 9 | 0.65 | 3 |
| Breed type | 816 | 7.93 | 6 | 9 | 0.70 | 3 |
| Sex type | 816 | 7.67 | 6 | 9 | 0.74 | 3 |
| Harmony | 816 | 7.65 | 5 | 9 | 0.62 | 4 |
| Head | 816 | 7.72 | 6 | 9 | 0.69 | 3 |
| Neck | 816 | 7.65 | 5 | 9 | 0.68 | 4 |
| Withers | 816 | 7.43 | 6 | 9 | 0.58 | 3 |
| Shoulder | 816 | 7.46 | 6 | 9 | 0.57 | 3 |
| Chest | 816 | 7.58 | 6 | 9 | 0.55 | 3 |
| Back | 836 | 7.42 | 5 | 9 | 0.59 | 4 |
| Croup | 836 | 7.53 | 6 | 9 | 0.59 | 3 |
| Legs | 835 | 7.08 | 5 | 9 | 0.60 | 4 |

## 2 Material and methods

### 2.1 Conformation scoring, anatomical measurements and imaging

For this study 102 Lipizzan stallions from the Spanish Riding School in Vienna were classified for, in total, 12 conformation traits which are recorded regularly in the Austrian Lipizzan breeding program on a basis of a valuating score sheet. Scores were given on a scale from 1 to 10 with increments of one for the following traits: type, breed type, sex type, harmony, head, neck, withers, shoulder, chest, back, croup and legs. The two sub-traits breed type and sex type were added to the score sheet in order to study associated shape differences. For further description of the score sheet see Table S1. Traits and their mean values and standard deviations are presented in Table 1. The assessment of conformation was carried out separately by three directors (EU-approved classifier) of Lipizzan state stud farms, four EU-approved classifiers from private Lipizzan breeding organizations, and one member of the riding staff of the Spanish Riding School. In connection with the assessments of conformation, morphological measurements (height at withers measured by stick (WI), chest circumference (CC) and front cannon bone circumference (CBC) measured by tape) and standardized pictures were taken from every single horse. The caliber index (CI) (Druml et al., 2008) was calculated from three body measurements according to following formula:

$$CI = ((CC \times CBC)/WI) \times 1000.$$

The resulting photographs were used for the definition of the horse shapes via the digitization of coordinates (landmarks) related to the horses' anatomy. Digital images were taken by the main researcher using a DSLR camera with the following specifications: distance between horse and camera – 18 m; focal width of camera lens – 100 mm; camera focus

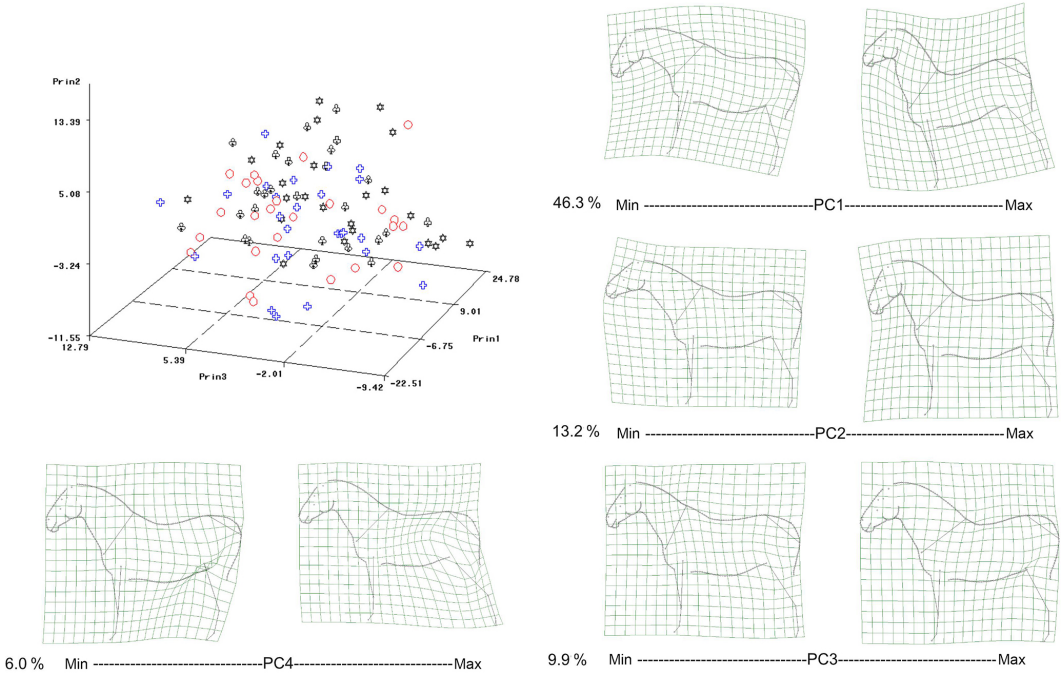


**Figure 1.** Shape space of Lipizzan stallions; principal component analysis (PCA) of partial warp scores derived from the bending energy matrix of thin plate splines based on the data pool of 102 Lipizzan stallions. Shape changes are depicted along the PCA axes 1, 2, 3 and 4, which account for 75.4 % of shape variation.

set to the center of gravity of the horse (height of camera set to 110 cm at a 90° angle to the approximate position of the animal's heart). The horses were presented by a groom in so-called open posture: left foreleg standing vertically; hoof of the right foreleg one to two hoof lengths behind the left foreleg; cannon bone of the right hind leg nearly vertical; hoof of the right hind leg located two to three hoof lengths in front of the left hind leg. The presentation of horses in open posture matches the standard protocol for the evaluation of conformation during stud book registrations. The imaging process was repeated several times and a minimum of two line-ups was performed per horse. For the selection of the pictures finally used an optimal-fit criterion according to previously mentioned guidelines was applied (also see Druml et al., 2015).

The definition of landmarks, i.e. the $x$ and $y$ coordinates describing anatomical and shape features, is essential for developing a horse model suited to study biological shape variation. Morphological landmarks should fulfill five criteria: (a) homologous position; (b) consistent somatometric position (replicable relative topological positions along outlines of a morphological feature); (c) adequate coverage of landmarks on a morphological structure; (d) chosen landmarks should be reliable and repeatable; (e) all landmarks should be coplanar (Zelditch et al., 2004). For the description of equine shape, we chose a horse model combining the outline and 17 anatomical landmarks (Fig. 1) (also see Druml et al., 2015). The outlines were transformed to single landmarks and further defined as sliding semilandmarks in order to minimize the bending energy and result in homologous-like points along a curve (Gunz and Mitteroecker, 2013). The horse model finally comprised 246 landmarks (17 anatomical landmarks, 14 somatometric landmarks and 215 sliding semilandmarks).

In total, 25 092 two-dimensional landmark coordinates were extracted from 102 standardized digital images using the software packages tpsDig version 2.17 (Rohlf, 2005) and tpsUtils version 1.58 (Rohlf, 2004).

## 2.2 Statistical analysis of shape variation and ratings

A major task for the statistical analyses of image data is the establishment of a reference system in which one position on a specimen can be assigned exactly to the homologue position on another specimen. In this case we generated the reference system by a generalized Procrustes superimposition (generalized Procrustes analysis – GPA), which scales, rotates and centers every single specimen according to the mean configuration of the sample.

GPA eliminates non-shape variation in configurations of landmarks by superimposing landmark configurations using least-squares estimates for translation and rotation parameters. First, the centroid of each configuration is translated to the origin, and configurations are scaled to a common unit size. The centroid is defined as the mean of all $x$ and $y$ values from all coordinates. Each shape is characterized by its

size, called the centroid size. This variable is the square root of the sum of all squared distances from each landmark to the centroid. If all $x$ and $y$ coordinates of the landmarks are divided through centroid size, the resulting specimens are of unit size 1. Finally, the configurations are optimally rotated to minimize the squared differences between corresponding landmarks, which is also called Procrustes distance $d$ (Rohlf and Slice, 1990). The process is iterated to compute the mean shape. At the end of this process, the original coordinate data have been replaced by substitute Cartesian coordinates (shape coordinates; Bookstein, 1991), as they vary around their own sample mean, and are corrected for effects of scale (centroid size), for effects of orientation and for effects of location of the original specimens (Fig. 1). The 488 resulting shape coordinates were used for the determination of the horse shape space, i.e. the shape variation within our sample of horses, by principal component analysis. The axes of principal components (relative warp axes) can be described, quantified and visualized by thin-plate splines (linear combinations of shape coordinates) of horse shapes moving along each PC axis.

For the evaluation of each classifier's rankings we applied shape regressions, where shape coordinates are regressed into the scores given by the classifiers in each trait. The following traits were analyzed: type, breed type, sex type, harmony, the overall mean and the mean of the four type traits (type, breed type, sex type, harmony). Further we recalculated the trait-score-associated Lipizzan horse shapes along the regression curve for the mean configuration, an unfavorable configuration (score of 5) and a favorable configuration (score of 10) in order to visualize the aspects of shape encoded in ratings. For further information and details, see Druml et al. (2015), Mitterröcker and Gunz (2009), Slice (2007) and Bookstein (1991).

For assessing the agreement among classifiers and because the analyzed conformation traits are represented by ordinal scores, the $\kappa$ statistics according to Fleiss (1971) were calculated using the SAS macro mkappa (Chen et al., 2005). In order to compare the rater consistency with results cited in most of the scientific literature, we also calculated the inter-class correlation (ICC) adjusted to the formula of Spearman. The differences between the single classifiers' mean ratings were analyzed applying a generalized linear model with classifier and horse as fixed effects. Multiple pairwise comparisons of means were adjusted according to the Tukey and Kramer.

Repeatability ($r$) of image data was analyzed according to following procedure: because after a GPA, relative coordinate data are available, the level of repeatability can be calculated on the basis of relative warp scores (principal components of shape coordinates) or on the Procrustes distance $d$ (Arnqvist and Martensson, 1998). In this study we analyzed the repeatability of images captured in two or three different sessions and afterwards performed image selection according to the optimal-fit (stance) criterion for each horse per session. We evaluated the repeatability of different horse image data estimating the variance components from relative warp scores with the SAS Varcomp procedure (SAS Inc., 2009) using the following mixed model:

$$y_{ijkl} = y + \mathrm{rep}_i + \mathrm{animal}_{jk} + \mathrm{error}_{ijkl}$$

In study 1, we imaged 15 animals three times and in study 2 we imaged 55 animals two times. The repeatability coefficients were calculated as follows:

$$r = \sigma^2\mathrm{animal}_{jk}/\sigma^2\mathrm{animal}_{jk} + \sigma^2\mathrm{error}_{ijkl}.$$

All statistical analyses were performed using the SAS software packages version 9.1 (SAS Inc., 2009), and morphometric analyzes and shape graphs were performed using tpsRelw v1.53 (Rohlf, 2003), tpsRegr v1.40 (Rohlf, 2005) and tpsSuper v2.00 (Rohlf, 2013).

## 3 Results and discussion

### 3.1 Agreement among classifiers

Mean values and standard deviations from the 102 Lipizzan stallions scored by eight raters are presented in Table 1. In Table 2 least squares means for each rater and trait from a linear model in which the fixed effects classifier and horse accounted for 34–55 % of variance ($R^2$) within the scoring data are presented. Comparisons of means adjusted according to Tukey and Kramer illustrate the diversity between the single classifier mean ratings of the following traits:

- type, breed type, sex type, harmony, head (type traits);
- neck, withers, shoulder, chest, back, croup and legs (functional traits).

Concerning the influence of the animals' age on the conformation scores, we could observe a significant age effect for the traits withers ($p < 0.001$), back ($p < 0.001$) and legs ($p < 0.04$). Regarding the agreement among raters, the $\kappa$ coefficients in Table 2 showed similar values to those previously reported for 44 Lipizzan mares. In general, the coefficients for stallions are at a lower level. The highest agreement was found in the traits type, sex type, neck, breed type, harmony and head. The following traits were characterized by the lowest $\kappa$ coefficients: legs, croup, shoulder and withers. Taking the results of Table 2 into account, the concordance among raters was low and it was lowest in functional traits. When sorting the different classifiers into groups, all raters achieved a general $\kappa$ value of 0.21 (0.33 for mares), stud directors a value of 0.27 (0.39 for mares) and private breeders showed a value of 0.18 (0.29 for mares). The highest $\kappa$ values and Spearman rank correlations ($r$) for pairwise calculations were achieved by the rater pairs 1–6, 6–2, 1–2, 5–9 and 2–8 ($\kappa$ from 0.42 to 0.33, $r$ from 0.55 to 0.49) and lowest $\kappa$ values were observed for the rater pairs 2–7, 7–8, 1–7, 6–7 and 4–7, with a $\kappa$ ranging from 0.13 to 0.08 ($r$ from 0.44 to 0.30).

**Table 2.** Multiple comparisons of means by classifier per trait, $\kappa$ value and standard error for stallions and mares*.

| Trait/rater | 9[3] | 4[1] | 1[1] | 6[2] | 7[2] | 3[1] | 8[2] | 5[2] | $\kappa$ stallions | SE | $\kappa$ mares* | SE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | $7.90^{a}$ | 7.81 | $7.87^{a}$ | 7.82 | $7.63^{b}$ | $7.85^{a}$ | $7.98^{a}$ | 7.79 | 0.24 | 0.013 | 0.49 | 0.028 |
| Breed type | $7.92^{ab}$ | $7.79^{a}$ | $7.92^{ab}$ | $7.76^{a}$ | $7.83^{a}$ | $8.12^{cb}$ | $8.16^{b}$ | 7.97 | 0.21 | 0.013 | 0.38 | 0.027 |
| Sex type | $7.59^{ac}$ | 7.68 | $7.52^{a}$ | $7.59^{ac}$ | $7.60^{ac}$ | 7.73 | $7.80^{bc}$ | $7.84^{b}$ | 0.23 | 0.013 | 0.28 | 0.025 |
| Harmony | 7.63 | 7.67 | $7.71^{bc}$ | $7.61^{ac}$ | $7.48^{a}$ | 7.63 | $7.83^{b}$ | 7.65 | 0.19 | 0.014 | 0.38 | 0.026 |
| Head | $7.84^{ac}$ | $7.97^{c}$ | 7.77 | $7.56^{b}$ | $7.61^{b}$ | $7.67^{ab}$ | $7.58^{b}$ | 7.76 | 0.18 | 0.013 | 0.28 | 0.025 |
| Neck | $7.47^{a}$ | $7.69^{bc}$ | $7.82^{b}$ | $7.74^{bc}$ | $7.55^{ac}$ | $7.73^{bc}$ | 7.61 | 7.64 | 0.22 | 0.014 | 0.30 | 0.026 |
| Withers | $7.54^{a}$ | $7.48^{ac}$ | $7.42^{ac}$ | $7.55^{ab}$ | $7.10^{bc}$ | $7.33^{c}$ | $7.53^{ac}$ | $7.49^{ac}$ | 0.13 | 0.016 | 0.16 | 0.030 |
| Shoulder | $7.49^{ac}$ | $7.48^{ac}$ | $7.49^{ac}$ | $7.58^{a}$ | $7.09^{b}$ | $7.66^{a}$ | $7.34^{c}$ | $7.52^{ac}$ | 0.09 | 0.016 | 0.27 | 0.030 |
| Chest | $7.47^{d}$ | $7.78^{a}$ | $7.61^{acd}$ | $7.74^{ac}$ | $7.17^{b}$ | $7.57^{cd}$ | $7.59^{cd}$ | $7.71^{ac}$ | 0.15 | 0.016 | 0.27 | 0.033 |
| Back | $7.49^{ab}$ | $7.50^{ab}$ | $7.66^{b}$ | $7.51^{ab}$ | $7.21^{c}$ | $7.34^{ac}$ | $7.32^{ac}$ | $7.35^{ac}$ | 0.17 | 0.016 | 0.32 | 0.028 |
| Croup | $7.55^{ab}$ | $7.70^{a}$ | $7.53^{ab}$ | $7.65^{a}$ | $7.21^{c}$ | $7.54^{ab}$ | $7.62^{ab}$ | $7.43^{b}$ | 0.08 | 0.016 | 0.29 | 0.027 |
| Legs | $7.08^{a}$ | $7.45^{b}$ | $7.09^{a}$ | $7.23^{ab}$ | $6.54^{c}$ | $7.02^{a}$ | $7.32^{b}$ | $6.97^{a}$ | 0.06 | 0.013 | 0.24 | 0.036 |

Different superscripts indicate significant differences ($p < 0.05$) after correction for multiple testing. [1] Stud farm directors. [2] Private classifiers. [3] Member of the riding staff.
* See Druml et al. (2015).

Overall, the agreement among the eight raters can be judged as low, even lower than it had been the case in the conformational evaluation of 44 Lipizzan mares. Nevertheless the results regarding the structure of different traits and the relationships between the classifiers could be repeated, and these results underline the difficulty in assessing functional traits and the low correlations between expert classifier rankings assessed in a test situation.

### 3.2 The Lipizzan horse shape space and encoded ratings

The shape space, i.e. the distribution of shapes within a sample, can be reconstructed and visualized using principal component analysis of the shape coordinates (e.g. partial warp scores and uniform components) (see Fig. 1). The first four principal components (PCs) accounted for 75.40 % of total shape variation within the sample of 102 Lipizzan stallions and gave a similar picture as in the shape space for Lipizzan mares. PC1 (46.3 %) accounted for neck posture variation, PC2 (13.2 %) for conformational differences, PC3 (9.9 %) for sex-related traits, shoulder and chest conformation, and PC4 (6 %) for variation due to differences in stance.

Repeatabilities were calculated for the number of relative warps, explaining 95 % of total shape variation. In study 1 (15 horses imaged and tracked three times), repeatabilities for 12 relative warps ranged from 0.24 to 0.72, with a mean of 0.50 and a standard deviation of 0.14. In study 2 (55 horses imaged and tracked two times) the repeatabilities for 16 relative warps varied between 0.23 and 0.59, with a mean of 0.42 and a standard deviation of 0.12. In comparison Kristjansson et al. (2013) presented repeatabilities for standardized image-based anatomical measurements within a range of 0.34 to 0.99. In this study 20 animals were measured two times per frame, resulting in an $r$ between 0.56 and 0.99, representing only the operator error. In a second case study, the authors analyzed the repeatability of traits measured on the left and right side of 10 horses. The values for $r$ varied here between 0.34 and 0.96. Zechner et al. (2001) reported repeatability coefficients for 37 body measurements taken two times from 362–368 Lipizzan horses. The $r$ values varied between 0.24 and 0.95, with a mean repeatability of 0.68. In relation to these results, our data, which comprise multiple anatomical features at the same time, expressed by shape coordinates or relative warp scores, fits within the range of values reported. As the individual stance of the horse and its excitement during imaging is a significant source of variation in shape data, we suggest to improve the data quality by selecting horses according to their individual level of repeatability of shape data.

In order to test which of the ratings of the four type traits (type, breed type, sex type, harmony) account for specific variation within the sample, we applied multiple linear regressions of shape coordinates into the single rankings of each classifier and four classifier groups. In total 57 % of the 54 regression equations achieved a non-significant coefficient. Considering the single ratings of the eight classifiers, 50 % achieved a significant ranking of animals within the traits type, breed type, sex type and harmony (Table 3). The amount of shape variation explained by the significant ratings ($p < 0.05$) was small and ranged from 1.99 % (rater 4 for breed type) to 0.83 % (rater 6 for sex type). In comparison the mean ratings in mares for breed type explained 3.2 %, those for harmony 5.9 % and those for type 3.0 % of the shape variation. When these values are compared to shape regressions in humans, dominance ratings explained 8 %, masculinity 7.3 %, attractiveness 3.3 % (Windhager et al., 2011), trustworthiness 8.6 % and eye color 5.5 % (Kleisner et al., 2013) of human face shape variation.

**Table 3.** Significance levels of 48 shape regressions for eight raters and for the mean score of eight raters within the overall mean and type-related traits.

| Rater | Mean of all traits | Mean of type traits | Type | Breed type | Sex type | Harmony |
|---|---|---|---|---|---|---|
| 6 | n.s. | n.s. | n.s. | n.s. | 0.0033 | 0.0005 |
| 1 | n.s. | n.s. | n.s. | n.s. | n.s. | n.s. |
| 3 | n.s. | n.s. | n.s. | n.s. | n.s. | n.s. |
| 5 | n.s. | n.s. | 0.0005 | 0.0005 | n.s. | n.s. |
| 4 | n.s. | 0.0005 | 0.0329 | 0.0005 | 0.0005 | 0.0005 |
| 7 | n.s. | 0.0005 | 0.0005 | 0.0005 | n.s. | 0.0005 |
| 8 | n.s. | 0.0033 | 0.0005 | n.s. | 0.0005 | 0.0005 |
| 9 | n.s. | 0.0206 | 0.0005 | 0.0005 | n.s. | n.s. |
| All | n.s. | 0.0116 | n.s. | 0.0005 | 0.0005 | n.s. |

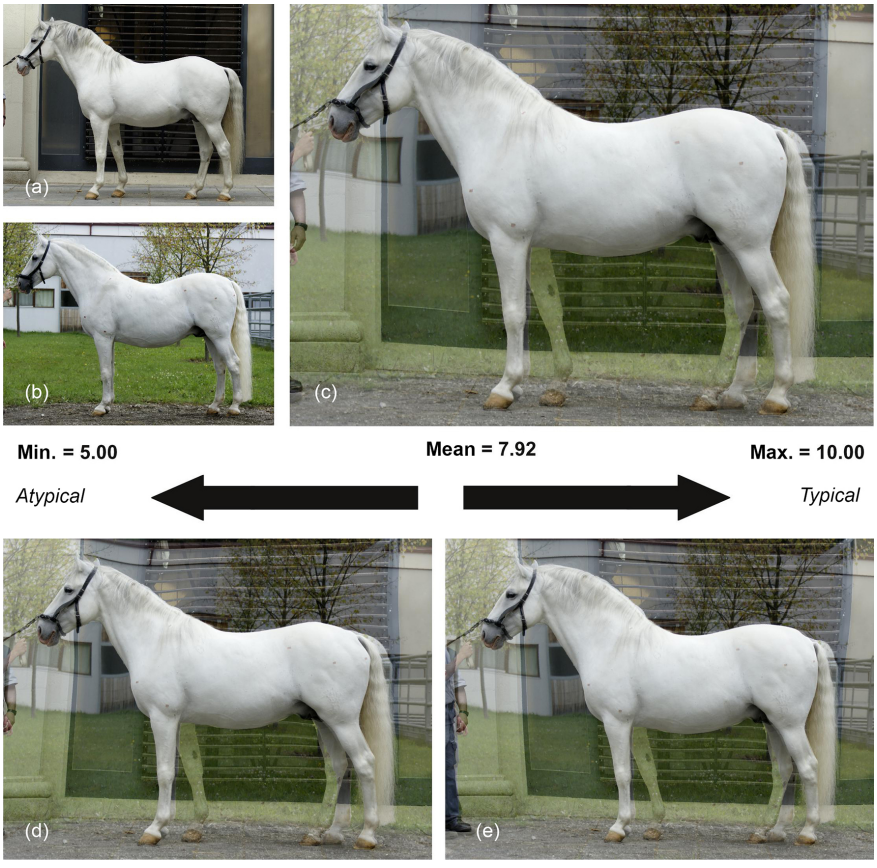


**Figure 2.** Lipizzan horse shape changes according to the valuating score sheet of the trait breed type based on the full horse model; scores from classifier group 3. The consensus **(c)** is created by image averaging from the pictures of the two Lipizzan stallions Conversano Kitty II **(a)** and Neapolitano Allegra **(b)**, which are warped onto the mean configuration of the sample ($n = 102$). The averaged "model horse" **(c)** is deformed to a target configuration score of 5 **(d)** and of 10 points **(e)** following the regression curve of the shape regression. The resulting favorable and unfavorable model horses represent the horse shapes which would be expected according to the minimum and maximums score given by classifier group 3 for the trait breed type (illustration and photos by T. Druml).

Shape regressions for breed type (based on the mean assessments of raters whose ratings resulted in a significant shape regression) are displayed by two model horses that were averaged and warped onto the estimates of the trait mean, the minimum score of 5 points and the maximum score of 10 points in Fig. 2.

In Figs. 3 and 4 the significant differences of individual ratings within the trait type between single classifiers (rater 4,

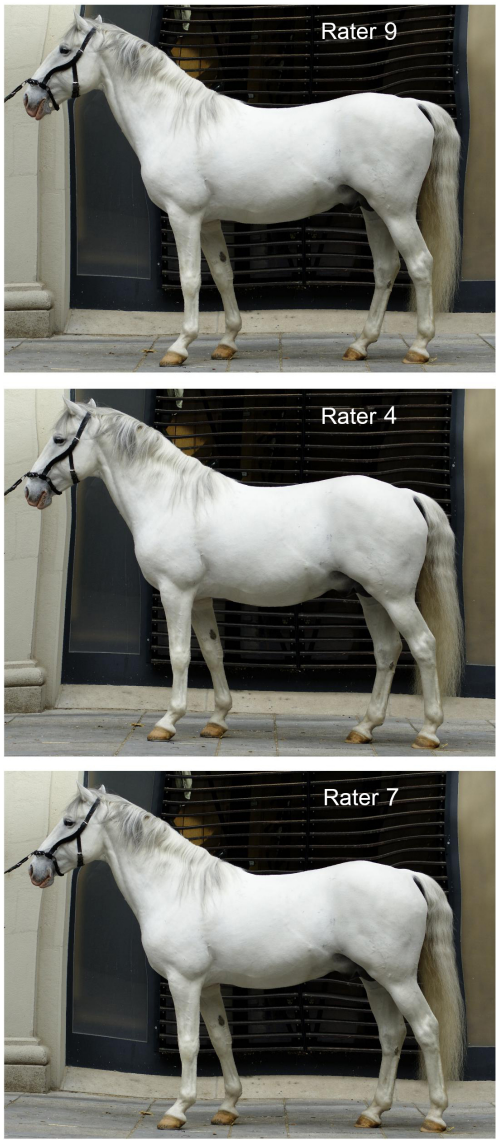


**Figure 3.** Differences of individual ratings: unfavorable Lipizzan model horses representing the score 5 in the trait type for rater 9, rater 4 and rater 7. Images were recalculated using the configurations of each rater along the regression curves (illustration and photos by T. Druml).

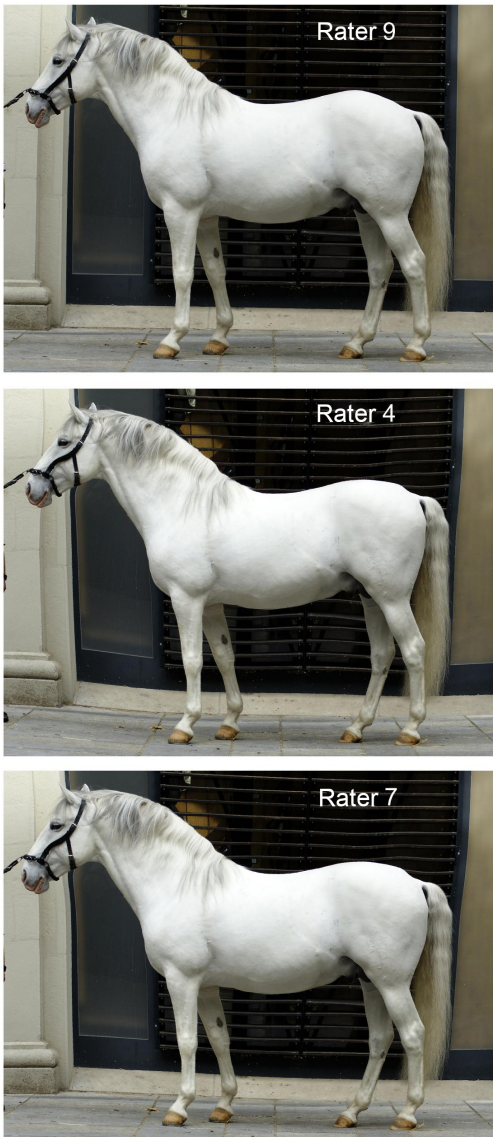


**Figure 4.** Differences of individual ratings: favorable Lipizzan model horses representing the score 10 in the trait type for rater 9, rater 4 and rater 7. Images were recalculated using the configurations of each rater along the regression curves (illustration and photos by T. Druml).

7 and 9) are shown, which illustrate the subjective nature of equine conformation ratings. Within the warped model horses based on individual ratings, clearly visible subjective differences in perception existed (Figs. 3 and 4). For example, within the trait type, a riding staff member (rater 9) focused on the back and head morphology, whereas rater 4 (stud director) clearly emphasized the croup and neck conformation and rater 7 (private breeder) favored the opposite head conformation to rater 9 and attached more value to sex-related properties. In general, the phenotypes that scored favorably (score approximating 10 along the regression curves) for the four type traits under study were all characterized by a muscular horse of quadratic format, long neck and shoulder, deep chest, and relatively short back. Within the phenotypes that scored unfavorably (score approximating 5 along the regression curves), model horses and graphical representations (thin plate splines) were characterized by a longer back, a relatively short and weak-muscled neck, a flat chest and a globally androgynous appearance.

The central goal of this paper was to study the "black-box" problem which arises in the evaluation of equine conformation based on valuating scales, where individual scores

only imply a difference between what is favorable and not favorable. Although visual perception of equine conformation is regarded as an empirical method, we found out in our experiment that the agreement in expert ratings of interpretative anatomy (correlations ranging from 0.36 to 0.56) did not differ from ratings derived from several samples of people regarding attractivity in humans. Human attractivity ratings are typically correlated with a coefficient ranging from 0.30 to 0.50. Also the test for the association of rating scores with phenotypic variability of horse shapes showed that the scores given by classifiers did not correspond with shape variation in 43 % of cases. We conclude that the ratings are inconsistent. Although the inconsistency of ratings is the subject of scientific interest in equestrian dressage competitions (Stachurska et al., 2005; Stachurska and Bartyzel, 2011; Wolframm et al., 2013), this question is poorly addressed in studies concerning equine conformation. Further, we demonstrated that significant differences in visual perceptions result in different phenotype rankings and phenotypic averages (model images).

## 4 Conclusions

From the results of this study we conclude that inconsistency of ratings (a major reason for non-significant shape regressions) and low rater agreement (subjectivity of raters) need to be further described and studied, as such bias in classifiers' rankings is also present in genomic and genetic analyses of conformation data. Further, the detected error sources have not been considered in equine breeding programs and selection procedures. Therefore, the methods and tools presented in this paper offer the possibility to evaluate conformation classifiers in a quantitative (reliability analysis, shape regressions) and qualitative (model horses from shape regressions) way. These methods can further be used to assist in harmonization processes of scoring protocols and in the training of equine conformation classifiers, especially in small populations with a limited number of offspring.

**The Supplement related to this article is available online at doi:10.5194/aab-2-309-2016-supplement.**

**Author contributions.** Thomas Druml and Gottfried Brem designed the experiments and Thomas Druml and Maximilian Dobretsberger carried them out. Thomas Druml analyzed the data and prepared the manuscript.

**Acknowledgements.** The authors wish to thank the Austrian Research Promotion Agency (FFG) and Xenogenetik for financial support and the Spanish Riding School in Vienna and the International Lipizzan Federation (LIF) for assistance and cooperation.

Edited by: M. Mielenz
Reviewed by: two anonymous referees

## References

Arnqvist, G. and Martensson T.: Measurement error in geometric morphometrics: empirical strategies to assess and reduce its impact on measures of shape, Acta zoological academia scientiarum hungaricae, 44, 73–96, 1998.

Bodo, I., Alderson, L., and Langlois, B.: Conservation genetics of endangered horse breeds, in: EAAP publication Nr. 116, Wageningen, the Netherlands, 2005.

Bookstein, F.: Morphometric tools for landmark data: geometry and biology, Cambridge University Press, Cambridge, UK, 1991.

Chen, B., Zaebst, D., and Seel, L.: A macro to calculate Kappa statistics for categorizations by multiple raters, in: Proceedings of the 30th annual SAS User Group International conference, 10–13 April 2005, Philadelphia, Pennsylvania, USA, 2005.

Cunningham, M. R., Roberts, A., Barbee, A. P., Druen, B. P., and Wu, C. H.: Their ideas of beauty are, on the whole, the same as ours: consistency and variability in the cross-cultural perception of female physical attractiveness, J. Pers. Soc. Psychol., 68, 261–279, 1995.

Druml, T., Baumung, R., and Sölkner, J.: Morphological analysis and effect of selection for conformation in the Noriker draught horse population, Livest. Sci., 115, 118–129, 2008.

Druml, T., Dobretsberger, M., and Brem, G.: The use of novel phenotyping methods for validation of equine scoring results, Animal 9, 928–937, 2015.

Duensing, J., Stock, K. F., and Krieter, J.: Implementation and Prospects of Linear Profiling in the Warmblood Horse, J. Equine Vet. Sci., 34, 360–368, 2014.

Fleiss, J. L.: Measuring nominal scale agreement among many raters, Psychol. Bull., 76, 378–382, 1971.

Grundler, C. and Pirchner, F.: Wiederholbarkeit der Beurteilung von Exterieurmerkmalen und Reiteigenschaften, Züchtungskunde, 63, 273–281, 1991.

Gunz, P. and Mitteroecker, P.: Semilandmarks: a method for quantifying curves and surfaces, Hystrix, 24, 103–109, 2013.

Janssens, S., Winandy, D., Tylleman, A., Delmotte, C. H., Van Moesecke, W., and Vandebitte, W.: The linear assessment scheme for sheep in Belgium: breed averages and assessor quality, Small Ruminant Res., 51, 85–95, 2004.

Kilgore, L.: Anthropometric variance in humans: Assessing Renaissance concepts in modern applications, Anthropological Notebooks, 18, 13–23, 2012.

Kleisner, K., Priplatova, L., Frost, P., and Flegr, J.: Trustworthy-Looking Face Meets Brown Eyes, PLoS ONE, 8, e53285. doi:10.1371/journal.pone.0053285, 2013.

Koenen, E. P. C., Van Veldhuizen, A. E., and Brascamp, E. W.: 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., 43, 85–94, 1995.

Kramer, A.: Implementation of an adjusted program of stallion selection, Bachelor thesis, Van Hall Larenstein University of Applied Science Wageningen, the Netherlands, 2012.

Kristensen, E., Dueholm, L., Vink, D., Andersen, J. E., Jakobsen, E. B., Illum-Nielsen, S., Petersen, F. A., and Enevoldsen, C.: Within- and across-person uniformity of body condition scoring in Danish Holstein cattle, J. Dairy Sci., 89, 3721–3728, 2006.

Kristjansson, T., Bjornsdottir, S., Sigurdsson, A., Crevier-Denoix, N., Pourcelot, P., and Arnason, T.: Objective quantification of conformation of the Icelandic horse based on 3-D video morphometric measurements, Livest. Sci., 158, 12–23, 2013.

Mitterröcker, P. and Gunz, P.: Advances in geometric morphometrics, Evol. Biol., 36, 235–247, 2009.

Rohlf, F. J.: tpsRegr, shape regression, version 1.28, Department of Ecology and Evolution, State University of New York at Stony Brook, New York, USA, 2003.

Rohlf, F. J.: tpsUtil, file utility program, version 1.26. Department of Ecology and Evolution, State University of New York at Stony Brook, New York, USA, 2004.

Rohlf, F. J.: tpsDig, digitize landmarks and outlines, version 2.05, Department of Ecology and Evolution, State University of New York at Stony Brook, New York, USA, 2005.

Rohlf, F. J.: tpsSuper, version 2.13, Department of Ecology and Evolution, State University of New York at Stony Brook, New York, USA, 2013.

Rohlf, F. J. and Slice, D. E.: Extensions of the Procrustes method for the optimal superimposition of landmarks, Syst. Zool., 39, 40–59, 1990.

Sánchez, M. J., Gomez, M. D., Molina, A., and Valera, M.: Genetic analyses for linear conformation traits in Pura Raza Español horses, Livest. Sci., 157, 57–64, 2013.

SAS Institute: SAS version 9.1, SAS Institute Inc., Cary, NC, USA, 2009.

Slice, D. E.: Geometric morphometrics, Annu. Rev. Anthropol., 36, 261–281, 2007.

Stachurska, A. and Bartyzel, K.: Judging dressage competitions in the view of improving horse performance assessment, Acta Agric. Scan. Section A – Anim. Sci., 61, 92–102, 2011.

Stachurska, A., Niewczas, J. and Markowski, M.: An Estimation of Reliability of Judging the Horse Dressage Competitions, in: 56th Annual Meeting of the European Association for Animal Production, 5–8 June, 2005, Uppsala, Sweden, 2005.

Veerkamp, R. F., Gerritsen, C. L. M., Koenen, E. P. C., Hamoen, A., and De Jong, G.: Evaluation of classifiers that uses linear type traits and body condition score using common sires, J. Dairy Sci., 85, 976–983, 2002.

Windhager, S., Schäfer, K., and Fink, B.: Geometric morphometrics of male facial shape in relation to physical strength and perceived attractiveness, dominance and masculinity, Am. J. Hum. Biol., 23, 805–814, 2011.

Wolframm, I. A., Schiffers, H., and Wallenborn, A.: Visual attention in Grand Prix dressage judges, J Vet. Behav., 8, p. e25, 2013.

Zechner, P., Zohman, F., Sölkner, J., Bodo, I., Habe, F., Marti, E., and Brem, G.: Morphological description of the Lipizzan horse population, Livest. Prod. Sci., 69, 163–177, 2001.

Zelditch, M., Swiderski, D., and Sheets, H. W. L. F.: Geometric morphometrics for biologists. A primer, Elsevier, San Diego, USA, 2004.