The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK

Animal Welfare 2009, 18: 177-187
ISSN 0962-7286

Observer reliability for working equine welfare assessment: problems with high prevalences of certain results

CC Burn*†, JC Pritchard†‡ and HR Whay†

† University of Bristol, Department of Clinical Veterinary Science, Langford, Bristol BS40 5DU, UK
‡ The Brooke Hospital for Animals, Broadmead House, 21 Panton Street, London SW1Y 4DR, UK
* Contact for correspondence and requests for reprints: charlotte.burn@worcester.oxon.org

Abstract

Welfare issues relevant to equids working in developing countries may differ greatly to those of sport and companion equids in developed countries. In this study, we test the observer reliability of a working equine welfare assessment, demonstrating how prevalence of certain observations reduces reliability ratings. The assessment included behaviour, general health, wounds, and limb and foot pathologies. In Study 1, agreement between five observers and their trainer (the 'gold standard') was assessed using 80 horses and 80 donkeys in India. Intra-observer agreement was later tested on 40 of each species. Study 2 took place in Egypt, using nine observers, their trainer, 30 horses and 30 donkeys, adjusting some scoring systems and providing observers with more detailed guidelines than in Study 1. Percentage agreements, Fleiss kappa (with a weighted version for ordinal scores) and prevalence indices were calculated for each variable. Reliability was similar across both studies, but was significantly poorer for donkeys than horses. Age, sex, certain wounds and (for horses alone) body condition, consistently attained clinically-useful reliability. Hoof-horn quality, point-of-hock lesions, mucous membrane abnormalities, limb-tether lesions, and skin tenting showed poor reliability. Reporting the prevalence index alongside the percentage agreement showed that, for many variables, the populations were too homogenous for conclusive reliability ratings. Suggestions are made for improving scoring systems showing poor reliability, but future testing will require deliberate selection of a more diverse equine population. This could prove challenging given that, in both populations of horses and donkeys studied here, many pathologies apparently showed 90–100% prevalence.

Keywords: *animal welfare, donkeys, horses, methodology, observer agreement, prevalence*

Introduction

Until recently, the health and welfare of the estimated 40.5 million horses (*Equus caballus*) and 39 million donkeys (*Equus asinus*) working in developing countries (FAOSTAT 2005), have been little studied. The environmental challenges they face, and the work they are required to carry out, can make their health issues considerably different to those of sports and companion equids in developed countries (eg Svendsen 1997; Pritchard *et al* 2005; Tesfaye & Curran 2005). The prevalences of welfare problems in horses, mules and donkeys working in five developing countries have been described in a large-scale study (Pritchard *et al* 2005), showing that over 90% were lame (see also Maranhão *et al* 2006; Broster *et al* 2009), 70% were thin (see also Pearson & Ouassat 1996), and a high proportion had skin lesions (see also Tesfaye & Curran 2005; Burn *et al* 2008). A potentially-high proportion also suffer from heat stress due to physical exertion in hot climates (Pritchard *et al* 2006, 2008). Therefore, the appropriateness of previously-established welfare assessment methods for Western equids may be limited when applied to these working equids.

In this study, we describe a process in the development of a general welfare assessment protocol intended to underpin future research into factors affecting working horse and donkey welfare. The assessment was animal-based, rather than resource-based, ie assessing the animals' behaviour and health directly, rather than aspects of husbandry, handling or harnessing (Johnsen *et al* 2001; Whay *et al* 2003; Main *et al* 2007). As a result of the heavy reliance equine owners have on their animals in developing countries, the assessment was required to be rapid (limiting the time animals would spend away from employment) and, simple, so that relatively few errors would be possible. Also, a quick and easy, broad-brush welfare assessment could be more readily passed on as a concept to the equine owners, encouraging them to regularly check the welfare of their animals themselves. The assessment was developed for use by the veterinarians and animal health workers of an equine charity, the Brooke Hospital for Animals, and therefore practicality was essential.

The aim of the welfare assessment was to record horse and donkey body condition, disease and behaviour, including

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 1 Examples of the relationship between percentage agreement, prevalence index, and kappa values.

Percentage agreement	Maximum PI threshold		
	$k \geq 0.4$	$k \geq 0.6$	$k \geq 0.8$
95	0.91	0.86	0.70
90	0.81	0.70	0.00
85	0.69	0.49	–
80	0.58	0.00	–
75	0.40	–	–
70	0.00	–	–

k is the kappa reliability rating. For each given percentage agreement, the maximum prevalence index (PI) for which it is possible to obtain Moderate ($k \geq 0.4$), Substantial ($k \geq 0.6$), or Excellent ($k \geq 0.8$) reliability ratings are shown. The PIs are calculated as shown in Byrt *et al* (1993). As the percentage agreement increases, the degree of population imbalance that can be tolerated for the given kappa thresholds increases. For less than 80 or 90% agreement, it is not possible to obtain kappa values above 0.6 or 0.8, respectively.

response to humans and, in this paper, we report the degree of inter- and intra-observer reliability of the various scores. We used kappa statistics, with a weighted equivalent, Kendall's coefficient of concordance, for ordinal scales (Maclure & Willett 1987) to assess the degree to which the proportion of agreement was better than chance. Thus, kappa statistics are more conservative than correlations or raw percentage agreements alone (Hoehler 2000). Finding poor observer agreement in any of the variables would be useful in alerting us to scoring systems that require modification, clearer definition, or more in-depth training.

However, kappa values become ambiguous when relative prevalences in the sample population greatly exceed 50%, ie when prevalences become unbalanced. This is because the probability of agreeing purely by chance is very high in near-homogenous populations, making evidence for good observer agreement difficult or impossible to identify (Hoehler 2000; Vach 2005). To illustrate this, when a condition is near ubiquitous in a population, a high percentage agreement is no guarantee that observers would reliably identify the rare instances of the opposite condition were it presented to them; they might agree with each other purely because none of them can detect the seemingly rare condition. Low kappa values can therefore indicate either genuinely poor agreement, or that a population was too homogenous for any agreement above chance to be detected (eg Burn & Weir, submitted). This ambiguity can complicate the interpretation of low kappa values.

An alternative kappa calculation, 'PABAK', has been proposed that adjusts for prevalence and observer bias (Byrt *et al* 1993), but this has been criticised for readjusting for the very factors that kappa is designed to control for (Hoehler 2000). Aside from ignoring all variables with unbalanced prevalences (as suggested in Hoehler 2000), there is no easy way around the problem, so here we present prevalence indices and the raw percentage agreements alongside the kappa values, making the interpretation of kappa more transparent (Burn & Weir, submitted). Presenting these three factors together allows a distinction to be made between variables attaining genuinely poor agreements, versus those ambiguous variables that attain poor kappa ratings because the population was too homogenous for any above chance agreement to have been detectable. We illustrate the relationship between kappa values and prevalence indices (Byrt *et al* 1993; Sim & Wright 2005) for given percentage agreements in Table 1 (see also Burn & Weir, submitted, for more detail). Our own simulations show that Kendall's coefficient of concordance is also reduced when prevalences are imbalanced (data not shown). However, the relationship is more complex than for kappa — for example, the coefficient is reduced more when errors are made in the more common scores than in the rarer scores — so detailed exploration of this relationship is beyond the scope of this study.

This study is not intended as a validation of the welfare significance of any of the measurements taken, which would require in-depth studies of specific variables. Instead, it marks one of the first steps in developing a workable assessment protocol for a species in conditions thus far little explored. The general principles, and some of the specific results, may have relevance for welfare assessment protocols in other species or animal management systems. Two assessment methods are compared, the first in India, and the second being an adjusted version in Cairo. The results are interpreted in the light of the percentage agreements, the reliability ratings (kappa or Kendall's coefficient) and, for binary variables, the prevalence indices.

Study 1

Materials and methods

Animals and observations

In Delhi, India, the health and welfare of working horses (n = 80) and donkeys (n = 80) were assessed by six observers during the course of two days per species in August 2003. The welfare assessment was a standardised, non-invasive protocol as summarised by Pritchard *et al* (2005) and detailed in Pritchard and Whay (2003, unpublished) (available from the authors upon request). Briefly, the measures included age and sex, behavioural responses to humans and the environment, general health, the locations and severity of skin lesions, and limb and foot pathologies relevant to lameness (Table 2).

Observers 2–6 were trained by observer 1, the 'trainer', and were experienced at using the assessment protocol from previous work. The training procedure consisted of observers being given a detailed verbal explanation of each score, and provided with guidance notes and photographs. They then conducted 100 assessments, paired with the trainer. All observers received training a minimum of six

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 2 Scoring systems used in working horse and donkey welfare assessments in India (Study 1) and Egypt (Study 2).

Variable	Scoring range in Study 1	Scoring range in Study 2
General characteristics		
Age	< 5/5–15*/> 15	< 5/5–15*/>15
Sex	Stallion*/gelding/mare	Stallion/gelding/mare
Behaviour		
Attitude	Alert*/apathetic/depressed	Alert*/apathetic/depressed
Chin contact	Accepts*/avoids	Accepts*/avoids
Heat stress	Present/absent*†	Present/absent*†
Response to observer approach	Moves away/turns head away†/no response*/turns head towards*/aggressive	Moves away/turns head away*/no response/turns head towards*/aggressive
Response to observer walking down side	No interest/signs of interest*†	No interest/signs of interest*†
Tail duck (donkeys only)	No response to observer walking past rear/clamps tail down	–
General health		
Body condition	1, very thin/2, thin*/3, medium/4, fat/5, very fat	1, very thin/1.5/2*†/2.5/3/3.5/4/4.5/5, very fat
Coat condition	Healthy*/dull/poor condition	Healthy*/unhealthy
Diarrhoea	Faecal soiling present/absent*†	Faecal soiling present/absent*†
Ectoparasites	Present/absent*†	Present/absent*†
Eyes	No abnormalities*/abnormal*†	No abnormalities/abnormal*†
Hooks or edges on teeth	Present*/absent	–
Mucous membranes	Normal colour*/abnormal	Normal colour*/abnormal
Skin tent	Immediate return*/delay < 3 s/delay ≥ 3 s	Immediate return*/delayed
Teeth missing	Yes/no*†	–
Skin lesions		
Belly	0 ≤ 2 × 2 cm*†/superficial/broken skin/deep	–
Breast	As for belly lesions	0 ≤ 2 × 2 cm*†/superficial/broken skin/deep
Ears	As for belly lesions	As for breast lesions
Firing lesions	As for belly lesions	As for breast lesions
Forelegs	As for belly lesions	As for breast lesions
Girth	As for belly lesions	–
Girth and belly	–	As for breast lesions
Head	As for belly lesions	As for breast lesions
Hindlegs	As for belly lesions	As for breast lesions
Hindquarters	As for belly lesions	As for breast lesions
Knees	Lesion present*†/absent	Lesion present*†/absent
Limb-tether lesions	As for belly lesions	As for breast lesions
Lips	Lesion present/absent*†	Lesion present*/absent*†
Neck	As for belly lesions	As for breast lesions
Point of hocks	Lesion present*†/absent	Lesion present*†/absent
Ribs	As for belly lesions	As for breast lesions
Spine	As for belly lesions	–
Tail	As for belly lesions	As for breast lesions
Withers	As for belly lesions	–
Withers and spine	–	As for breast lesions
Limb and foot pathology		
Cow hocks	Yes*/no	Yes*/no
Deformed limb	None/mild*†/severe	–
Gait	Normal/abnormal*†	Normal/abnormal*†
Hoof-horn abnormality	Normal/mild*†/severe	Normal/abnormal*†
Hoof overgrown	Yes*/no	–
Hoof shape	–	Normal/abnormal*†
Hoof short	Yes*/no	–
Sole shape and structure	–	Normal/abnormal*†
Sole surface	Normal/abnormal*†/closed shoe	–
Swollen tendons and joints	None/mild*†/severe	Yes*/no

The welfare assessment was a standardised protocol as detailed in Pritchard and Whay (2003, unpublished), (available from the authors upon request). The most prevalent classification(s) observed for each variable are shown by * for horses and † for donkeys.

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

months prior to the study, and all had consolidated their experience through applying the assessment to a minimum of 100 animals in a developing country.

The animals in this study were chosen from the population working in the vicinity of Delhi. Each animal was identification marked by a harness tag and hoof brand so that intra-observer reliability could be tested at a later date, and was rested for approximately 1 h prior to being assessed. The animals stood in a row of ten standing bays, with new animals being brought in only after all ten of the previous ones had been assessed by every observer. Observers were instructed not to talk during assessments and not to discuss their assessments with the other observers. Only one observer was allowed to assess an animal at a time and, for logistical reasons, the observers moved along the row of animals from left-to-right, although each started simultaneously with a different individual.

To allow intra-observer reliability to be tested, the observers (including the trainer but missing one observer) repeated their assessments on 40 of the horses four days after finishing the first assessment. They also repeated their assessment on 40 of the donkeys, this time two days after their initial assessment.

Statistical analyses

The percentage agreement between and within observers for each variable was calculated, and those categorical variables with less than 75% agreement were considered to have insufficient agreement for clinical use. The 75% cut-off was not used for ordinal scales because expected percentage agreements decline rapidly as the numbers of possible scores increases, without necessarily jeopardising clinical relevance. Nominal variables consisting of more than two categories were separated into their binary components, so that each category was individually assessed against the remaining categories combined (Kraemer *et al* 2004).

Categorical variables were assessed using Fleiss' kappa statistics, and Kendall's coefficient of concordance was used for ordinal scales. Kappa values and Kendall's coefficients that are closer to 1.0 indicate better agreement, and the reliability rating scale used here (poor to excellent, see Table 3) was adapted from Landis and Koch (1977), taking moderate values above 0.4 to be clinically useful (Sim & Wright 2005). The trainer (observer 1) was used as the gold standard to test whether the training technique was effective. The software used was Minitab® (version 14).

For categorical variables, prevalence indices were calculated (Byrt *et al* 1993; Sim & Wright 2005) (no prevalence index is yet available for use with Kendall's coefficient of concordance). The prevalence index is the absolute difference between the agreed numbers for the two categories, divided by the total number of animals:

$$\text{Prevalence index} = \frac{|a-d|}{n}$$

Where *a* is the number of agreed upon animals in one of the categories and *d* is the number of agreed upon animals for the other category; n is the total number of possible agreements, ie the number of animals. A prevalence index of 0 indicates a completely balanced population, while an index of 1 would be a homogenous population in which only one of the categories is represented. Since our calculations were based around a gold standard, the prevalence indices were calculated pairwise between each observer and the trainer, and the mean taken for each variable.

To assess any correlation between inter- and intra-observer reliability, a regression was used that took into account the species and the prevalence index associated with each variable.

Results

Agreement between observers and the trainer

The results of the inter-observer reliability tests are shown in Table 3. Many prevalences were unbalanced, with 18 of the 30 categorical variables having prevalence indices above 0.75 for donkeys, and 13 of the 28 for horses. Only three variables in donkeys (non-response to observer approach, lesions of the point-of-hock, and overgrown hooves) and in horses (non-response to observer approach, lesions of the point-of-hock, and knee lesions) had well-balanced prevalence indices below 0.25.

Taking kappa values above 0.4 to be clinically useful (Sim & Wright 2005), all five observers exceeded criterion for seven variables in horses (sex, age, body condition and four skin lesion variables) and in donkeys (sex, age, three behaviours, and two skin lesion variables) — some of these acceptable reliability ratings were obtained despite unbalanced prevalence indices. The reliability rating of body condition was poor in donkeys, achieving only 59.3% agreement and, yet, it was substantial in horses, achieving 80.5% agreement.

Many variables with unbalanced prevalences apparently showed poor reliability as indicated by their kappa values and, yet, they had high percentage agreement values, meaning that their interpretation is unclear. On the other hand, several variables attained genuinely poor ratings (percentage agreements below 75%, and kappa or Kendall's *W*-values below 0.4), with eye abnormalities, hoof-horn quality, lesions on the point-of-hock, and rib lesions being poor for both species (Table 3).

In-depth, pair-wise analyses of each variable (data not shown) indicated that reasons for inter-observer disagreement could include four main factors. Firstly, observer opinions sometimes differed in where the cut-off points between scores lay, or when classifying borderline animals. Examples are coat condition, where observers disagreed on cut-off points between healthy, dull or poor condition; hoof-horn abnormalities, where observers disagreed on how to distinguish mild from severe; and eye abnormalities, where observers differed in what they classified as 'abnormal'. Secondly, lack of agreement could come about through some observers not using as wide a range of the scale as others. For example, when describing lesion severity in most anatomical locations, some but not all observers used score 2 (moderate); in most locations, no observers used score 3 (severe). Incorrect recollection of

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 3 Inter-observer reliability ratings of a working horse and donkey welfare assessment in India (Study 1).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Poor	PA < 75% for binary variables	Eyes (61.3%; PI = 0.32)	Eyes (60.5%; PI = 0.29)
		Hoof overgrown (74.8%; PI = 0.24)	Hooks/edges on teeth (67.5%; PI = 0.57)
		Knee lesions (65.3%; PI = 0.32)	Lip lesion (73.5%; PI = 0.51)
		Point-of-hock lesions (74.3%; PI = 0.21)	Point-of-hock lesions (74.8%; PI = 0.09)
	PA < 75% and W < 0.4 for ordinal variables	Body condition (59.3%)	Horn quality (53%)
		Deformed limbs (70.5%)	Rib lesions (58.5%)
		General attitude (60%)	Skin tent (66.5%)
		Hindleg lesions (62.9%)	Swollen tendons (56.3%)
		Horn quality (48.5%)	
		Rib lesions (68.3%)	
		Tail lesions (66.2%)	
Ambiguous	PA ≥ 75% but k < 0.40	Belly lesions (95.8%; PI = 0.96)	Cow hocks (94.5%; PI = 0.92)
		Cow hocks (97%; PI = 0.97)	Diarrhoea (83%; PI = 0.78)
		Diarrhoea (89.6%; PI = 0.9)	Ectoparasites (99.3%; PI = 0.99)
		Ectoparasites (100%; PI = 1)	Gait (96.5%; PI = 0.96)
		Firing lesions (96.5%; PI = 0.97)	General attitude (76.8%; PI = 0.77)
		Gait (97.5%; PI = 0.98)	Heat stress (77.3%; PI = 0.61)
		Heat stress (99.8%; PI = 1)	Hoof short (76.8%; PI = 0.64)
		Hoof short (76%; PI = 0.63)	Mucous membranes (79.3%; PI = 0.78)
		Hooks/edges on teeth (79.5%; PI = 0.77)	Observer approach: Aggressive (99%; PI = 0.98)
		Limb-tether regions (77.3%; PI = 0.58)	Observer approach: Moves away (95.8%; PI = 0.96)
		Lip lesion (90.1%; PI = 0.81)	Sole surface: Closed shoe (99.8%; PI = 1)
		Mucous membrane (83.3%; PI = 0.81)	Sole surface: Normal (75.3%; PI = 0.72)
		Observer approach: Aggressive (100%; PI = 1)	Teeth missing (94.8%; PI = 0.93)
		Sex: Mare (100%; PI = 1)	Walk down (79.8%; PI = 0.57)
		Sole surface (90%; PI = 0.9)	
		Teeth missing (92.8%; PI = 0.91)	
	PA ≥ 75% but W < 0.40	Breast lesions (82.5%)	Belly lesions (85.8%)
		Coat condition (75.3%)	Ear lesions (88%)
		Deformed limbs (70.5%)	
		Foreleg lesions (79%)	
		General attitude (60%)	
		Horn quality (48.5%)	
		Rib lesions (68.3%)	
		Skin tent (82.5%)	
		Swollen tendons (78.8%)	
		Tail lesions (66.2%)	

k is the kappa reliability rating, and W is Kendall's coefficient of concordance. The reliability rating scale is adapted from Landis and Koch (1977) and Sim and Wright (2005). The mean percentage agreements (PA) obtained are shown in parentheses for each variable. For categorical variables, mean prevalence imbalances are given as a prevalence index (PI) (Byrt *et al* 1993).

the scoring range is a third reason: for example, for the binary variable assessing overgrown hooves, one observer used a 'score 2', presumably to indicate severe overgrowth. Finally, notes made by observers on the original datasheets indicated that disagreement about lesion severity scores originated from uncertainty about how to label the locations of lesions at the borders between anatomical demarcations — this could be responsible for the poor reliability of the rib-lesion scores if observers disagreed on the boundaries between girth, ribs, spine and belly.

Agreement within observers

Variables that showed lower reliability between observers and the trainer, also showed significantly lower reliability within observers ($F_{1,50} = 33.0$; $P < 0.001$) (Table 4). Intra-observer reliability was above criterion in all observers for 13 variables in horses, and 12 in donkeys (age, sex [horses only, because no donkeys were female], body condition and 10 lesion sites). Conversely, several variables showed poor reliability; eye abnormalities again showed poor reliability in all observers. As a category, behaviours showed

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 3 (cont) Inter-observer reliability ratings of a working horse and donkey welfare assessment in India (Study 1).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Moderate	*k* = 0.40–0.59	Hoof overgrown (74.8%; PI = 0.24)	Chin contact (94.5%; PI = 0.87)
		Observer approach; Moves away (94.3%; PI = 0.89)	Observer approach; No response (75.8%; PI = 0.18)
		Observer approach; No response (78.3%; PI = 0.18)	Observer approach; Turns away (92.3%; PI = 0.82)
		Observer approach; Turns away (81.5%; PI = 0.3)	Observer approach; Friendly (81.3%; PI = 0.43)
		Point-of-hock lesions (74.3%; PI = 0.21)	
		Walk down (77.8%; PI = 0.39)	
	W = 0.40–0.59	Ear lesions (70%)	Coat condition (56.3%)
		Head lesions (70.5%)	Deformed limbs (53.3%)
		Hindquarter lesions (70.5%)	Foreleg lesions (66.5%)
		Neck lesions (82.5%)	Head lesions (70.8%)
		Spine lesions (77.5%)	Hindleg lesions (67.3%)
			Hindquarter lesions (74%)
			Hoof overgrown (69%)
			Limb-tether lesions (70.5%)
			Neck lesions (81.5%)
			Spine lesions (75%)
			Tail lesions (71.3%)
			Withers lesions (67.8%)
Substantial	*k* = 0.60–0.79	Chin contact (89.8%; PI = 0.61)	Body condition (80.5%)
		Observer approach: Friendly (90%; PI = 0.63)	Knee lesions (82%; PI = 0.18)
		Sex: Gelding (98.8%; PI = 0.96)	Sex: Gelding (97.8%; PI = 0.93)
		Sex: Stallion (98.5%; PI = 0.97)	
		Tail tuck (98%; PI = 0.94)	
	W = 0.60–0.79	Age (79%)	Age (73.5%)
		Girth lesions (82.5%)	Breast lesions (77.8%)
		Withers lesions (79.8%)	Firing lesions (92%)
			Girth lesions (74.8%)
Excellent	*k* = 0.80–1.00		Sex: Mare (100%; PI = 0.65)
			Sex: Stallion (97.8%; PI = 0.58)

poor or ambiguous reliability ratings across both species, with the exception of general attitude, which attained moderate reliability ratings.

Study 2

Materials and methods

On the basis of preliminary analyses of data from Study 1, a second version of the assessment was evaluated during April 2004. In an attempt to obtain different prevalence indices for some variables, the location was changed to Cairo, Egypt. Some changes were made to the scoring systems, as shown in Table 2, and the accompanying notes, diagrams and photographs were made more detailed and comprehensive (Pritchard & Whay 2004, unpublished) (available upon request from the authors).

All observers were trained just prior to the study — for most this updated their previous training, but for three observers it represented their first training. The training was classroom-based and each measure was explained in detail, illustrated with pictures and any modifications highlighted. This was followed by one practical training session in the Helwan brick kilns near Cairo, and one practical session in the Brooke clinic in Cairo, where observers were paired-up and encouraged to compare and discuss discrepancies in their observations. Finally, the observers underwent an examination consisting of pictures and multiple-choice questions to test their knowledge of the assessment criteria and their accuracy of scoring.

For the inter-observer reliability study, ten observers who passed the examination (including the trainer and four others who took part in Study 1) assessed 30 working horses on the first day and 30 donkeys on the second. Intra-observer reliability was not tested. In other respects, the procedure was similar to that in Study 1.

Statistical analyses were as before but an additional general linear model was used to compare reliability ratings across

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 4 Intra-observer reliability ratings of a working horse and donkey welfare assessment in India (Study 1).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Poor	PA < 75% for binary variables	Eyes (73.8%; PI = 0.6)	Eyes (74.5%; PI = 0.64)
		Observer approach: No response (53.8%; PI = 0.08)	Heat stress (63.5%; PI = 0.49)
		Observer approach: Turns away (60.4%; PI = 0.3)	Hoof overgrown (70%; PI = 0.41)
		Point-of-hock lesions (73.8%; PI = 0.35)	Hocks/edges on teeth (72.5%; PI = 0.57)
		Walk down (63.3%; PI = 0.31)	Observer approach: No response (63%; PI = 0.21)
			Observer approach: Friendly (68%; PI = 0.45)
	PA < 75% and *W* < 0.4 for ordinal variables	–	–
Ambiguous	PA ≥ 75% but *k* < 0.40	Belly lesions (95%; PI = 0.91)	Chin contact (87%; PI = 0.82)
		Chin contact (81.7%; PI = 0.66)	Cow hocks (95.5%; PI = 0.94)
		Diarrhoea (92.1%; PI = 0.8)	Diarrhoea (87%; PI = 0.75)
		Ectoparasites (99.6%; PI = 1)	Ear lesions (94%; PI = 0.88)
		Firing lesions (98.8%; PI = 0.97)	Ectoparasites (98%; PI = 0.98)
		Gait (100%; PI = 1)	Gait (99%; PI = 0.99)
		Heat stress (91.3%; PI = 0.91)	Hoof short (87.5%; PI = 0.73)
		Hoof overgrown (78.3%; PI = 0.43)	Lip lesion (80.5%; PI = 0.64)
		Hoof short (82.5%; PI = 0.63)	Mucous membranes (95%; PI = 0.91)
		Hooks/edges on teeth (85.4%; PI = 0.75)	Observer approach: Aggressive (97%; PI = 0.97)
		Lip lesion (90%; PI = 0.84)	Observer approach: Moves away (94.5%; PI = 0.95)
		Mucous membrane (94.6%; PI = 0.88)	Observer approach: Turns away (82.5%; PI = 0.82)
		Observer approach: Friendly (82.1%; PI = 0.75)	Sole surface: (83%; PI = 0.6)
		Observer approach: Moves away (92.1%; PI = 0.92)	Teeth missing (96.5%; PI = 0.94)
		Sex: Mare (100%; PI = 1)	Walk down (75.5%; PI = 0.71)
		Sole surface (95.4%; PI = 0.95)	
		Tail tuck (98.3%; PI = 0.98)	
		Teeth missing (95.4%; PI = 0.9)	
	PA ≥ 75% but *W* < 0.40	–	Swollen tendons (78.5%)

k is the kappa reliability rating, and *W* is Kendall's coefficient of concordance. The reliability rating scale is adapted from Landis and Koch (1977) and Sim and Wright (2005). The mean percentage agreements (PA) obtained are shown in parentheses for each variable. For categorical variables, mean prevalence imbalances are given as a prevalence index (PI) (Byrt *et al* 1993).

both studies. The model included the study location (Delhi or Cairo), the species, whether the variables were binary or ordinal, and the prevalence index; the variables themselves were included as random factors.

Results

As with Study 1, the prevalences remained unbalanced for many variables (Table 5). For most of the limb and foot pathology scores, prevalences were highly unbalanced in this study, as in the previous one, making reliability difficult to prove. All observers exceeded criterion for seven of the variables in horses and six in donkeys (age, sex, body condition [in horses], and four lesion sites). Of the behaviours, chin contact and some of the responses to observer approach showed reliability ratings of moderate or above in both studies and both species. Hoof-horn quality, limb-tether lesions, mucous membrane abnormalities, lesions on the point-of-hock, and skin tent all showed poor reliability ratings for both species.

There was no significant improvement in the reliability ratings in Study 2 compared with Study 1 ($P = 0.913$). Of the variables that were altered from Study 1, the body-condition score seemed to have improved. Its reliability for horses was substantial in both studies but, in donkeys, overall reliability increased from poor to moderate between the two studies. However, combining hoof overgrowth (moderate) and shortness (poor) in Study 1 into an overall measure of hoof shape in the current study resulted in an overall rating of poor reliability. It is notable that many observers used a more limited range of lesion scores than in Study 1, frequently resulting in binary scores.

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 4 (cont) Intra-observer reliability ratings of a working horse and donkey welfare assessment in India (Study 1).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Moderate	k = 0.40–0.59	Knee lesions (79.6%; PI = 0.41)	Point-of-hock lesions (79.5%; PI = 0.3)
		Neck lesions (88.3%; PI = 0.63)	
	W = 0.40–0.59	Coat condition (92%)	General attitude (93%)
		Cow hocks (98.3%)	Skin tent (77%)
		General attitude (66.2%)	
Substantial	k = 0.60–0.79	Sex: Gelding (99.2%; PI = 0.96)	Sex: Gelding (99%; PI = 0.95)
		Sex: Stallion (99.2%; PI = 0.96)	
	W = 0.60–0.79	Body condition (75%)	Belly lesions (90%)
		Breast lesions (90.4%)	Coat condition (67.5%)
		Deformed limbs (81.7%)	Deformed limbs (76.5%)
		Head lesions (76.3%)	Foreleg lesions (74.5%)
		Hindleg lesions (73.1%)	Head lesions (74.5%)
		Hindquarter lesions (73.8%)	Hindquarter lesions (76%)
		Horn quality (68.8%)	Horn quality (64%)
		Limb-tether lesions (81.6%)	Limb-tether lesions (72%)
		Rib lesions (78.3%)	Neck lesions (79%)
		Skin tent (84.6%)	Rib lesions (67%)
		Swollen tendons (83.8%)	Spine lesions (78%)
		Tail lesions (88.8%)	Withers lesions (69%)
Excellent	k = 0.80–1.00	–	Knee lesions (92%; PI = 0.27)
			Sex: Mare (99%; PI = 0.66)
			Sex: Stallion (98%; PI = 0.61)
	W = 0.80–1.00	Age (83.8%)	Age (78%)
		Ear lesions (80%)	Body condition (85%)
		Foreleg lesions (87.4%)	Breast lesions (79%)
		Girth lesions (82.5%)	Firing lesions (97%)
		Spine lesions (78.8%)	Girth lesions (78.5%)
		Withers lesions (82.9%)	Hindleg lesions (74%)
			Tail lesions (77%)

The general linear model showed that the welfare assessment was more reliable for horses than for donkeys ($F_{1,72}$ = 5.58; P = 0.002), and demonstrated empirically that reliability ratings decreased as prevalence indices increased ($F_{1,72}$ = 11.72; P = 0.001). The random effect of the variables themselves was also significant ($F_{42,72}$ = 5.48; P < 0.001), suggesting that their ratings showed some degree of stability across both species and both studies.

Discussion

In this study, we aimed to evaluate the inter-observer reliability of a subjective welfare assessment for working equids, quantifying the extent to which trained observers agreed with the trainer. The results were interpreted with reference to the prevalence indices for each measure because, as we have demonstrated, unbalanced prevalences reduce the chance of proving good observer reliability. For some measures, we have been able to establish whether reliability within and between observers was clinically acceptable or not. In other cases, when unbalanced variables showed poor reliability ratings, we simply remain unaware of whether inter-observer reliability really was poor, or whether the agreement expected by chance was simply so high that good reliability could not be statistically proven (Hoehler 2000; Vach 2005; Burn & Weir, submitted). In future research, a more variable population of equids will be necessary to properly assess these variables, but it will require the gold standard to artificially pre-select this sample, since working equids across several developing countries are already known to have extremely high prevalences of certain welfare problems (Pritchard *et al* 2005; Tesfaye & Curran 2005; Maranhão *et al* 2006; Broster *et al* 2009). Any effort at selection would therefore be time-consuming, and would be complicated by each working equid having multiple and variable conditions (Pritchard *et al* 2005; Tesfaye & Curran 2005; Maranhão *et al* 2006; Broster *et al* 2009).

Consistently reliable measures in the current study were age, sex, horse body condition, and certain skin lesions, particularly those on the withers, girth, and hindquarters. The specific lesions that attained high reliability ratings changed between studies and species, but most lesion scores exceeded criterion (k or $W \geq 0.4$) in most observers, suggesting that observers agreed on the general severity scale. Poor reliability ratings over lesions arose from unbal-

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 5 Inter-observer reliability ratings of a working horse and donkey welfare assessment in Cairo (Study 2).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Poor	PA < 75% for binary variables	Horn quality (73.8%; PI = 0.6)	Gait (73.8%; PI = 0.59)
		Limb-tether lesions (71.3%; PI = 0.4)	Horn quality (68%; PI = 0.36)
		Mucous membranes (62.9%; PI = 0.54)	Limb-tether lesions (67.7%; PI = 0.2)
		Point-of-hock lesions (73%; PI = 0.34)	Lip lesion (65%; PI = 0.13)
		Skin tent (66.3%; PI = 0.13)	Mucous membranes (60.4%; PI = 0.34)
		Walk down (67.5%; PI = 0.23)	Point-of-hock lesions (68.3%; PI = 0.16)
			Skin tent (63.6%; PI = 0.27)
	PA < 75% and W < 0.4 for ordinal variables	General attitude (59%)	–
Ambiguous	PA ≥ 75% but k < 0.40	Coat condition (82.9%; PI = 0.68)	Breast lesions (100%; PI = 1)
		Cow hocks (94.8%; PI = 0.95)	Coat condition (79.9%; PI = 0.76)
		Diarrhoea (76.7%; PI = 0.63)	Cow hocks (91.3%; PI = 0.91)
		Ear lesions (89.8%; PI = 0.81)	Ear lesions (100%; PI = 1)
		Ectoparasites (93.2%; PI = 0.92)	Ectoparasites (75.8%; PI = 0.7)
		Eyes (93.6%; PI = 0.94)	Eyes (89.2%; PI = 0.88)
		Gait (100%; PI = 0.92)	General attitude (92.8%; PI = 0.93)
		Heat stress (84%; PI = 0.84)	Girth and belly lesions (93.8%; PI = 0.89)
		Lip lesion (78.6%; PI = 0.39)	Head lesions (95.8%; PI = 0.92)
		Observer approach: Aggressive (100%; PI = 1)	Heat stress (85%; PI = 0.85)
		Observer approach: Friendly (96.5%; PI = 0.96)	Neck lesions (100%; PI = 1)
		Observer approach: Moves away (83.9%; PI = 0.63)	Observer approach: Aggressive (96.5%; PI = 0.96)
		Observer approach: Turns away (78.4%; PI = 0.55)	Sole surface:(82.8%; PI = 0.83)
		Swollen tendons (92.3%; PI = 0.91)	Swollen tendons (87.4%; PI = 0.8)
			Tail lesions (96.2%; PI = 0.96)
			Walk down (75.8%; PI = 0.6)
	PA ≥ 75% but W < 0.40	Neck lesions (79.3%)	–
		Rib lesions (85.8%;)	
		Sole shape (100%)	
		Tail lesions (91.9%)	

k is the kappa reliability rating, and W is Kendall's coefficient of concordance. The reliability rating scale is adapted from Landis and Koch (1977) and Sim and Wright (2005). The mean percentage agreements (PA) obtained are shown in parentheses for each variable. For categorical variables, mean prevalence imbalances are given as a prevalence index (PI) (Byrt *et al* 1993).

anced prevalence indices for some anatomical locations, from uncertainty about lesions at the boundaries between anatomical regions, and from disagreement about thresholds between different severity scores. As with any of the variables, it is also possible that order effects could have contributed to disagreements between observers because they each started by assessing different individual animals.

Overall, there was no significant improvement in reliability between the two studies, but the overall reliability for donkeys was significantly lower than for horses. While the reliability over body condition was substantial for horses, for donkeys in Study 1 it was poor. It increased to moderate for donkeys in Study 2 which could have been due to the introduction of half-scores, the more detailed descriptions provided, and/or the additional training.

Variables that consistently showed poor observer reliability ratings included hoof-horn quality, lesions on the point-of-hock, mucous membrane abnormalities, limb-tether lesions, and skin-tent duration (Tables 2 and 4). The low reliability for eye health in Study 1, may be because the 'abnormal' category was highly heterogeneous, ranging from small amounts of discharge to having an eye completely missing. In Study 2, the percentage agreements for eye health increased from 61.3 and 60.5% in donkeys and horses, respectively to 93.6 and 89.2%, but the reliability rating remained low (ambiguous). This could reflect population differences between Delhi and Cairo, or it could suggest that by providing more detailed descriptions and more photographic examples in Cairo, the observers could now reliably identify subtle eye abnormalities in most animals;

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Table 5 (cont) Inter-observer reliability ratings of a working horse and donkey welfare assessment in Cairo (Study 2).

Reliability rating	Criterion for given rating	Variable (PA obtained; PI) Donkeys	Variable (PA obtained; PI) Horses
Moderate	*k* = 0.40–0.59	Chin contact (92.2%; PI = 0.85)	Chin contact (95.4%; PI = 0.91)
		Firing lesions (81.7%; PI = 0.5)	Diarrhoea (83.3%; PI = 0.45)
		Girth and belly lesions (86.7%; PI = 0.73)	Observer approach: Friendly (85.2%; PI = 0.58)
		Hindleg lesions (80.4%; PI = 0.11)	Observer approach: No response (78%; PI = 0.09)
		Observer approach: No response (80.1%; PI = 0.15)	
		Sex: Gelding (93.6%; PI = 0.85)	
	W = 0.40–0.59	Body condition (60.7%)	Foreleg lesions (93.7%)
Substantial	*k* = 0.60–0.79	Knee lesions (84.6%; PI = 0.37)	Hindleg lesions (95.3%; PI = 0.85)
			Hindquarter lesions (98.2%; PI: 0.93)
			Knee lesions (87.4%; PI = 0.27)
			Observer approach: Moves away (99%; PI = 0.94)
			Observer approach: Turns away (90.2%; PI = 0.63)
	W = 0.60–0.79	Age (76.6%)	Age (82.6%)
		Breast lesions (89.3%)	Rib lesions (94.8%)
		Head lesions (82.1%)	
		Hindquarter lesions (83.3%)	
Excellent	*k* = 0.80–1.00	Sex: Mare (99.1%; PI = 0.33)	Firing lesions (96.7%; PI = 0.83)
		Sex: Stallion (94.5%; PI = 0.18)	Sex: Gelding (99.2%; PI = 0.88)
			Sex: Mare (100%; PI = 0.27)
		–	Sex: Stallion (99.2%; PI = 0.39)
	W = 0.80–1.00	Withers and spine lesions (90.7%)	Withers and spine lesions (93%)

thereby they may simultaneously have increased the percentage agreement and the prevalence index, meaning that the amount of agreement above chance remained low. Future versions of the system could incorporate more categories to better capture the variation that observers actually discriminate, either nominal categories (eg healthy/infected/traumatic injury/cataract), or ordinal estimates of the severity of pain or visual interference. Possible contributing factors for disagreement over skin-tent duration are covered in a related paper (Pritchard *et al* 2007), and the validity of this test for dehydration has recently been questioned (Pritchard *et al* 2008).

Gait abnormalities were usually reported to be so prevalent that ratings were ambiguous despite high percentage agreement, but when the prevalence index dropped to 0.59 for horses in Cairo, the percentage agreement fell below 75%, meaning that gait attained a poor rating (Table 5). In future studies, an ordinal scale of lameness might be more informative, especially since lameness is already known to be highly prevalent in these equine populations, varying from slight inconsistencies in gait to limbs being non-weight-bearing (Lindberg *et al* 2004; Pritchard *et al* 2005; Maranhão *et al* 2006; Broster *et al* 2009).

Another factor that could lower the reliability statistics, apart from poor observer reliability and unbalanced prevalence is, of course, whether we would expect the measure to change between observations. Behavioural responses to humans were particularly important to assess here, not just because some consisted of subjective scores, but also because the animals might actually respond differently towards different observers and across days. Chin contact, tail-tuck, and some responses to observer approach, consistently obtained moderate or above inter-observer reliability ratings (Tables 2 and 4), but they showed poor or ambiguous intra-observer reliability (Table 4). This might suggest that they changed across days, which could occur if the animals are generally inconsistent in these behaviours, or that there was an order effect, with the animals or the assessors being more familiar with the assessment situation on their second experience of it.

Reliability concerning most general health measures, and limb and foot pathologies, were difficult to assess because their prevalences were so unbalanced. Many of the general health measures were actually biased towards more positive welfare (eg virtually no ectoparasites and little evidence of diarrhoea), although the majority of animals were thin or very thin (Table 2). Conversely, most limb and foot pathologies were biased towards potentially poor welfare (eg cow hocks, abnormal gait, abnormal hooves and soles, and swollen joints and tendons).

https://doi.org/10.1017/S0962728600000324 Published online by Cambridge University Press

Overall, the high prevalences of welfare problems (Table 2) corroborate previous studies of the welfare conditions of working equids in developing countries (Svendsen 1997; Lindberg *et al* 2004; Pritchard *et al* 2005; Tesfaye & Curran 2005; Maranhão *et al* 2006). For example, the trainer's prevalences suggest that 98% of horses in Delhi had abnormal gaits, 80% were thin or very thin, 98% had swollen tendons, and most limb and foot abnormalities were ubiquitous. Lesions were prevalent in some parts of the body, especially the knees, breast, girth and withers in both species and, in donkeys, also the spine, hindquarters, and hindlegs, and lesions from limb-tethers.

Conclusion and animal welfare implications

Observer reliability tests are essential for testing the repeatability of subjective welfare and behaviour scoring, but this study illustrates the importance of interpreting reliability ratings in the light of the prevalences of the categories making up the scores. Results are ambiguous when variables attain a clinically useful percentage agreement, but their prevalence imbalance means that an adequate kappa rating cannot be achieved. For these variables, the extent of observer reliability remains unknown until they can be retested on a more balanced population. It is clear from many of these results that welfare problems are highly prevalent in these working equids, highlighting the need for an appropriate welfare assessment. This would allow scientific research to inform and evaluate interventions aiming to improve working equine welfare in the future.

Acknowledgements

This project was funded by the Brooke Hospital for Animals. We would like to thank all those observers and the animal owners who made these assessments possible.

References

Broster CE, Burn CC, Barr ARS and Whay HR 2009 The range and prevalence of pathological abnormalities associated with lameness in working horses from developing countries. *The Equine Veterinary Journal,* in press

Burn CC, Pritchard JC, Farajat M, Twaissi AAM and Whay HR 2008 Risk factors for strap-related lesions in working donkeys at the World Heritage site of Petra in Jordan. *The Veterinary Journal 178*: 261-269

Burn CC and Weir AAS Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. *The Veterinary Journal*, submitted

Byrt T, Bishop J and Carlin JB 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46*: 423-429

FAOSTAT 2005 FAO statistical database website. *Food and Agricultural Organisation of the United Nations*. http://faostat.fao.org/site/409/default.aspx. Date accessed: 7 July 2006.

Hoehler FK 2000 Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology 53*: 499-503

Johnsen PF, Johannesson T and Sandøe P 2001 Assessment of farm animal welfare at herd level: Many goals, many methods. *Acta Agriculturae Scandinavica Section A, Animal Science S30*: 26-33

Kraemer HC, Periyakoil VS and Noda A 2004 Agreement Statistics. Kappa coefficients in medical research. In: D'Agostino RB (ed) *Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies* pp 85-105. John Wiley & Sons, Ltd: Queensland, Australia

Landis JR and Koch GG 1977 The measurement of observer agreement for categorical data. *Biometrics 33*: 159-174

Lindberg AC, Leeb C, Pritchard JC, Whay HR and Main DCJ 2004 Determination of welfare problems and their perceived causes in working equines. *Animal Welfare 13*: S247

Maclure M and Willett WC 1987 Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology 126*: 161-169

Main DCJ, Whay HR, Leeb C and Webster AJF 2007 Formal animal-based welfare assessment in UK certification schemes. *Animal Welfare 16*: 233-236

Maranhão RPA, Palhares MS, Melo UP, Rezende HHC, Braga CE, Silva Filho JM and Vasconcelos MNF 2006 Most frequent pathologies of the locomotor system in equids used for wagon traction in Belo Horizonte. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia 58*: 21-27

Pearson RA and Ouassat M 1996 Estimation of the liveweight and body condition of working donkeys in Morocco. *Veterinary Record 138*: 229-233

Pritchard JC and Whay HR 2003 *Guidance notes to accompany working equine welfare assessment*. University of Bristol: Bristol, UK, unpublished

Pritchard JC and Whay HR 2004 *Guidance notes to accompany working equine welfare assessment*. University of Bristol: Bristol, UK, unpublished

Pritchard JC, Lindberg AC, Main DCJ and Whay HR 2005 Assessment of the welfare of working horses, mules and donkeys, using health and behaviour parameters. *Preventive Veterinary Medicine 69*: 265-283

Pritchard JC, Barr ARS and Whay HR 2006 Validity of a behavioural measure of heat stress and a skin tent test for dehydration in working horses and donkeys. *Equine Veterinary Journal 38*: 433-438

Pritchard JC, Barr ARS and Whay HR 2007 Repeatability of a skin tent test for dehydration in working horses and donkeys. *Animal Welfare 16*: 181-183

Pritchard JC, Burn CC, Barr ARS and Whay HR 2008 Validity of indicators of dehydration in working horses: a longitudinal study of changes in skin tent duration, mucous membrane dryness and drinking behaviour. *Equine Veterinary Journal 40*: 558-564

Sim J and Wright CC 2005 The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy 85*: 257-268

Svendsen ED 1997 *The Professional Handbook of the Donkey, Third Edition*. Whittet Books Limited: London, UK

Tesfaye A and Curran MM 2005 A longitudinal survey of market donkeys in Ethiopia. *Tropical Animal Health and Production 37*: 87-100

Vach W 2005 The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology 58*: 655-661

Whay HR, Main DCJ, Green LE and Webster AJF 2003 Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: Consensus of expert opinion. *Animal Welfare 12*: 205-217