

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): J. Kaler, G.J. Wassink and L.E. Green

Article Title: The inter- and intra-observer reliability of a locomotion scoring scale for sheep

Year of publication: 2009

Link to published version: <http://dx.doi.org/10.1016/j.tvjl.2007.12.028>

Publisher statement: None

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

The inter- and intra-observer reliability of a locomotion scoring scale for sheep

J. Kaler *, G. J. Wassink, L.E. Green

*Ecology and Epidemiology Group, Dept. of Biological Sciences, University of
Warwick, Coventry, CV4 7AL, UK*

*Corresponding author. Tel.: +44 2476 575860; fax: +44 2476 524619
E-mail address: j.kaler@warwick.ac.uk (J. Kaler)

18 **Abstract**

19

20 A seven point locomotion scoring scale ranging from 0 = normal locomotion
21 to 6 = unable to stand or move was developed. To test the between and within
22 observer reliability of the scale 65 movie clips of sheep with normal, and varying
23 degrees of abnormal, locomotion were made. Three observers familiar with sheep
24 locomotion were trained to read the videos. Thirty clips were randomly selected and
25 used to test the between and within observer agreement of these trained observers.
26 There was high inter- (intraclass correlation coefficient (ICC) = 0.93, weighted kappa
27 (κ_w) = 0.93) and intra-observer (intraclass correlation coefficient (ICC) = 0.90,
28 Weighted kappa (κ_w) = 0.91) reliability, with no evidence of observer bias. The main
29 differences between scores were for scores 0 (normal) and 1 (uneven posture and
30 shortened stride but no head movement). The results indicate that the locomotion
31 scoring scale using groups of defined observations for each point on the scale was
32 reliable and may be a useful research tool to identify and monitor locomotion in
33 individual sheep when used by trained observers.

34

35 *Keywords:* Sheep; Locomotion scoring; Reliability; Lameness; Trained observers

36

37 **Introduction**

38

39 Lameness is a change from normal stance or gait and is a cause of welfare
40 concern in many livestock species including cattle, sheep, pigs and poultry. In sheep,
41 lameness is associated with pain (Ley et al., 1989) and the most prevalent cause of
42 lameness, footrot, results in economic loss of £24 million per annum (Nieuwhof and
43 Bishop, 2005). Several locomotion scoring scales have been developed to monitor
44 cattle (Manson and Lever, 1988; Sprecher et al., 1997; Amory et al., 2006), pigs
45 (Main et al., 2000), poultry (Kestin et al., 1992) and sheep (Ley et al., 1989; Welsh et
46 al., 1993) to identify and quantify locomotion. Similar scales have been developed to
47 assess locomotion in horses (May and Wyn-Jones, 1987; Fuller et al., 2006; Hewetson
48 et al., 2006) and dogs (Reid and Nolan, 1991). The most frequently used approaches
49 to define locomotion include observation of stride length, duration of weight bearing
50 on both affected and unaffected limbs, body posture and joint movement (Sprecher et
51 al., 1997; Stashak, 2002).

52

53 Most of the locomotion scoring scales above have not been tested for
54 reliability and repeatability, although a few have (Kestin et al., 1992; Welsh et al.,
55 1993; Main et al., 2000; Fuller et al., 2006; Hewetson et al., 2006). Ideally, validity,
56 that is, that these scoring systems measure accurately what they are supposed to
57 measure, would be established by comparing a proposed locomotion scoring scale
58 with a gold standard (Dawson-Saunders and Trapp, 1994) which assessed the scale's
59 accuracy and objectivity. However, there is no gold standard to assess locomotion.
60 The best alternative is to investigate the reliability (Ebel, 1951; Shrout, 1998) of the
61 scale, that is, its consistency between independent measurements (Moss, 1994).

62 Reliability is at least a pre-requisite for validity since an unreliable measurement scale
63 has high variability between and within scorers and is of little use (Hewetson et al.,
64 2006).

65

66 A numerical rating scale to assess locomotion in sheep was developed in 1989
67 by Ley et al. It had categories from 0-4 (0 = normal movement, 1 = occasional
68 limping, 2 = lifting foot when standing, not lame when moving, 3 = carrying foot, but
69 lame on movement and 4 = carrying foot at all times). Observer agreement was not
70 assessed with this scoring scale. Another numerical rating scale with 'good' inter- and
71 intra-observer agreement was developed by Welsh et al. (1993) which also used a
72 scale from 0 to 4 (0 = clinically sound, 1 = barely detectable lameness, 2 = obvious
73 lameness, 3 = severe head nod and possibly resting the affected foot when standing
74 and 4 = carrying foot at the trot). The latter scale used subjective phrases e.g.
75 'obvious' lameness and neither of the scales above included all severities of
76 locomotion in sheep e.g. sheep with more than one foot affected or unable to rise are
77 not differentiated from sheep lame on one foot only.

78

79 A visual analogue scale (VAS) with good observer reliability was also
80 developed to assess locomotion in sheep by Welsh et al. (1993). This scale used a
81 straight line of 100 mm with two ends labelled 'sound' and 'could not be more lame'.
82 Although visual analogue scales are able to detect change of any size and can
83 differentiate between severities of lameness, they are highly subjective and difficult to
84 use in clinical practice (Welsh et al., 1993; Fuller et al., 2006).

85

86 In the UK, lameness in sheep has persisted over the last five decades despite
87 continued efforts to reduce its occurrence. In 2004, lameness was present in
88 approximately 97% of flocks, with a within flock prevalence of approximately 10%
89 (Kaler and Green, 2007). These estimates are from a random sample of 809 farmers
90 with no assessment of farmer ability to identify lame sheep in their sheep flocks.
91 Because of the continued high prevalence of lameness, the need to reduce the
92 subjective phrasing of scoring systems and to include the whole range of possible
93 severities, a new system was developed which provided descriptions within each
94 category of locomotion score, initially to assess locomotion in sheep in a research
95 setting. This paper presents this new scoring system together with the between and
96 within trained observer reliability.

97

98 **Materials and Methods**

99

100 *Locomotion scoring scale*

101 A seven point verbal numeric scale (0-6) was developed by a group of
102 researchers with experience of observing locomotion in lame and non-lame sheep.
103 The scale ranged from 'normal' to 'unable to stand or move' with visual descriptions
104 for locomotion for each increase in severity score (Table 1).

105

106 *Sample size estimation for reliability*

107 It was estimated that thirty observations were required to assess inter-observer
108 reliability with three observers with an expected inter-observer reliability (ρ_1) of 0.85,
109 acceptable (ρ_0) at 0.7 or higher, with $\alpha = 0.05$ and $\beta = 0.2$ (Walter et al., 1998). The
110 same 30 observations were used to assess intra-observer reliability.

111 *Locomotion scoring scale movies*

112 Movie clips of sheep rising, standing and walking were used to assess the
113 reliability of the scoring scale to ensure that there was no change in the locomotion of
114 sheep between repeated observations and that sheep position did not affect the
115 objective observation. As a consequence, 65 movies of 53 ewes, six rams and six
116 lambs with a range of locomotion scores from 0–6 (Table 1) were made, these
117 included locomotion in fore limbs, hind limbs and all four limbs per sheep.

118

119 The movies were made without disturbing sheep with sheep rising, standing
120 and walking on concrete and grass in a lateral view. The movie clips were recorded
121 with a camcorder (JVC GR-DVL 120A) and edited using Pinnacle studio 10.0
122 (Pinnacle systems, U.K.) and Video Edit Magic 4.1 (Desk share 2001- 2006). Each
123 clip was 35–50 s long (recommended by observers other than those who participated
124 in the study) with no audible sound. The 30 movie clips were randomly selected and
125 burnt onto a DVD with a 40 second lag between each clip. When more than one sheep
126 was in a clip observers were warned that the next clip contained more than one sheep
127 in the lag before the start of the clip and the sheep to be scored was circled.

128

129 *Observers*

130 Three observers were randomly selected from a group of researchers familiar
131 with observing locomotion in sheep. They were given a training session to learn to
132 score sheep locomotion based on what they saw in the clip, and using the descriptions
133 in Table 1, using ten movies with at least one of each score in the locomotion scoring
134 scale. This was followed by some test movie clips to ensure that the duration of the
135 clips and the between clip intervals were familiar to the observers. Finally, the

136 observers recorded their observations of the 30 movie clips in a room, sitting apart
137 from each other, using a copy of Table 1 and a recording form with the clip numbers
138 listed sequentially and a row of scores 0 – 6 for each clip with instructions to circle
139 one score per clip. The clips ran without a break. The forms were collected
140 immediately after the session.

141

142 To assess intra-observer repeatability the observers made a second assessment
143 of the same 30 movie clips 4 hours later. The clips were randomly reordered to reduce
144 the possibility that individual clips were recognised.

145

146 *Data analysis*

147 The data were entered in Microsoft Excel (Microsoft, 2000) and analysed
148 using SPSS 15.0 (SPSS Inc, 2006) and StatXact 7.0. The data were ordinal. The
149 percent exact agreement/ disagreement between and within observers were calculated.
150 The inter- and intra-observer reliability was assessed using intra-class correlation
151 coefficients (ICC) (Shrout and Fleiss, 1979) and weighted kappa coefficients (κ_w)
152 (Cohen, 1968). In addition, Kendall's rank correlation coefficient (τ) was used to
153 estimate between and within observer associations and Kruskal – Wallis one way
154 analysis of variance was used to investigate bias between observer ratings.

155

156 *Inter- and Intra-observer reliability*

157 a) Percent agreement

158 The percent exact agreement was estimated between observer pairs and within
159 each observer's scores. The percent of exact agreement and disagreement by one
160 point, two points and three points were calculated as:

161

162
$$\text{Percent agreement (disagreement)} = \frac{(\text{number of exact agreements (disagreements)}) * 100}{\text{Total number of observations}}$$

163

164

165 The mean percent agreement for between and within observers was also calculated.

166

167 b) Intraclass correlation coefficient (ICC)

168 The ICC was calculated with a two way random effects model (Shrout and
169 Fleiss, 1979 and McGraw and Wong, 1996) where both observers (raters) and the
170 subjects (sheep) were random effects. Absolute agreement and single measure
171 reliability were estimated. The model was specified as:

172

173
$$x_{ij} = \mu + r_i + c_j + rc_j + e_{ij}$$

174

175 where μ = population mean for all ratings, r_i = random sheep effect, c_j = random
176 observer effect, rc_j = random interaction effects and e_{ij} = residual or random error.

177 Normality of the data was checked by Shapiro-Wilk normality test. Estimates for ICC
178 were interpreted using previously recommended guidelines: 0-10% - virtually none,
179 11-40%- slight, 41-60% fair, 61-80% moderate and 81-100% substantial agreement
180 (Shrout, 1998).

181

182 c) Weighted kappa coefficients (κ_w)

183 The kappa statistic (κ) measures agreement beyond chance (Cohen, 1960).
184 Weighted kappa coefficients were calculated between observer pairs and scores of
185 each observer and as an overall average, using quadratic weights (Cohen, 1968). The
186 interpretation of kappa coefficients was made according to Landis and Koch (1977)

187 ≤ 0 = poor, .01–.20 = slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial,
188 and .81–1=almost perfect.

189

190 *Inter- and intra-observer associations*

191 The inter- and intra-observer Kendall's rank correlation was calculated by
192 comparing the scores of observer pairs and scores of each observer. An overall
193 average Kendall's rank correlation coefficient was also calculated between and within
194 observers.

195

196 *Observer bias*

197 Observer bias was assessed between observers using a Kruskal-Wallis one
198 way analysis of variance.

199

200

201 **Results**

202

203 *Inter- and Intra- observer reliability*

204 a) Percent agreement

205 The average overall exact agreement between observers and within observers
206 was 68% (range 63% -70%) and 76% (range 73%-77%) respectively. The majority of
207 disagreement between and within observers was by one point (Table 2).

208

209 b) Intraclass correlation coefficients (ICC)

210 The Sharpio- Wilk test did not reject the normality of the data at $P \leq 0.05$. The
211 ICC for inter- observer reliability was 0.93, (95% CI: 0.87-0.96) for the locomotion

212 scoring scale indicating substantial agreement between observers. The mean intra-
213 observer reliability was also substantial at 0.90 (95% CI: 0.89-0.92) with a range of
214 0.89 to 0.92 by observer (Table 2).

215

216 c) Weighted kappa coefficients (κ_w)

217 The overall average weighted kappa coefficient between observers was high:
218 $\kappa_w = 0.93$ (95% CI: 0.91-0.96) with a range of 0.92 to 0.95 between observer pairs.
219 Similarly the average weighted kappa for within observer scores was high with a
220 value 0.91 (95% CI: 0.87-0.96) that ranged from 0.89 to 0.93 by observer (Table 2).

221

222 *Inter- and intra-observer associations*

223 The overall average Kendall's rank correlations between and within observers
224 were high with $\tau = 0.87$ (range 0.75 - 0.98, $P < 0.01$) and $\tau = 0.85$ (range 0.67 - 0.98,
225 $P < 0.01$) respectively, indicating that there were very strong between and within
226 observer correlations (Table 2).

227

228 *Observer bias*

229 There was no significant difference between the mean rank scores between
230 observers ($\chi^2 = 3.58$, $df = 2$, $P = 0.16$).

231

232 The discrepancy between observers scores was mainly for scores 0 and 1 (Fig. 1).

233

234

235 **Discussion**

236 The results indicate that the locomotion scoring scale presented in this paper
237 was reproducible and repeatable between and within observers.

238

239 For an optimal design for the reliability study with a certain precision ($\alpha =$
240 0.05 and $\beta = 0.2$) the optimal combination of number of observers/number of
241 replicates per subject and the number of subjects is used for a set total number of
242 observations. For an expected reliability value, $\rho, > 0.6$ which is likely to be of use for
243 detecting high agreement the optimal design requires three replicates per subject or
244 three observers (Walter et al., 1998; Shoukri et al, 2004). Thus the choice of having
245 three observers in our study was appropriate.

246

247 Using movie clips of the sheep locomotion and posture ensured that the whole
248 locomotion scoring scale was assessed and that sheep did not alter their locomotion
249 between observations and that observers had an identical view of the sheep. Previous
250 agreement studies using movie clips have been used in horses with varying between
251 observer reliabilities ranging from moderate to good (Keegan et al., 1998; Fuller at al.,
252 2006; Hewetson et al., 2006).

253

254 Despite the objectivity of the movie clips of sheep locomotion, observers may
255 vary because of different interpretations of the locomotion scoring system because
256 they drift in scoring or because they are distracted while making a judgement, all
257 these aspects are combined and a final score is given by an observer (Uebersax,
258 2001). The overall effect is to reduce correlation of scores between and within
259 observers. This reduced correlation provides evidence that there is a random
260 (immeasurable) error and noise in observing method (Uebersax, 2001). In the study

261 presented here the greatest disagreement occurred between scores 0 and 1 between
262 observers (Figure 1). This was as hoped; score 1 is a very slight abnormal gait (Table
263 1) and provides an interim category between normal (score 0) and definitely lame
264 (score 2). This can be very useful in quantitative research. Conversely, observers may
265 have reduced the within observer variability by remembering movie clips and scoring
266 sheep identically, rather than assessing locomotion independently on the second test.
267 There were only 4 h between the tests but re-ordering and re-numbering the clips
268 should have minimised this effect. Finally, there was no statistical significant
269 evidence for bias between observers. This was useful information because the
270 presence of observer bias can affect the reliability of a scale considerably (Hewetson
271 et al., 2006).

272

273 The high inter-observer and intra-observer agreement achieved in this study is
274 comparable to the only agreement study done in sheep by Welsh et al. (1993) where
275 the locomotion in sheep was assessed using a numerical rating scale (NRS) with two
276 observers and no statistically significant difference between and within observers was
277 obtained using a Wilcoxin signed-rank test.

278

279 Our data were ordinal and so the appropriate measures of agreement were a
280 weighted kappa and intra-class correlation coefficients (Nelson et al., 1990; Morris et
281 al., 2004). When quadratic weights are applied to calculate kappa these two reliability
282 coefficients are equivalent (Fleiss and Cohen, 1973; Ludbrook, 2002), as seen in the
283 study presented in this paper. One of the limitations of these measurements is that
284 they are not comparable across populations because kappa is influenced by the
285 prevalence of a trait; this is equivalent to the dependence of ICC on between subject

286 variance (Maclure and Willet, 1987). Thus the agreement estimates can only be
287 generalised to a population with similar characteristics. In addition, both ICC and
288 kappa are affected by the number of ordinal categories in a scale. Although ICC is
289 less affected by the change in the number of categories, it tends to increase with an
290 increase in number of categories; in contrast, kappa tends to decrease with more
291 categories (Maclure and Willet, 1987). As a result we recommend that anyone
292 adopting this scale tests its reliability for their purpose since the agreement measures
293 depend on the prevalence of lameness as well as the training of personnel.

294 Modelling techniques such as log-linear models and latent trait and latent class
295 models for exploring agreement in ordinal ratings have been proposed. Log-linear
296 models can be used to estimate the amount of agreement beyond chance and also
297 agreement between two observers based on the baseline association, but they have not
298 been developed to analyse multiple observers, such as the three used in this study
299 (Agersti, 1992). Latent trait models can handle multiple observers and use the theory
300 that observed ratings are a continuous latent trait or set of latent classes (Nelson and
301 Pepe, 2000). Unlike ICC, latent trait models do not assume equal spacing of
302 categories and can provide information on all components of observer agreement
303 (Uebersax, 1993). However, both types of modelling techniques use complex
304 theoretical and statistical frameworks that are not yet widely used to assess such
305 agreements.

306

307 **Conclusion**

308

309 The scoring scale presented in this paper is objective and based on a group of
310 visual observations and a highly reliable method for trained observers to assess

311 locomotion in sheep. It may be used by trained researchers, and possibly advisers, to
312 monitor locomotion in sheep.

313

314 **6. Acknowledgements**

315 Jasmeet Kaler is supported by a Meat and Livestock Commission studentship,
316 funded by EBLEX, QMS and HCC. We thank Tristen George and Elisabeth Hawker
317 for contributing to this study.

318

319 **References**

- 320 Agresti A. Modelling patterns of agreement and disagreement. 1992 . Statistical
321 Methods in Medical Research 1, 201-228.
- 322 Amory, J.R., Kloosterman, P., Barker, Z.E., Wright, J.L., Blowey, R.W., Green, L.E.,
323 2006. Risk Factors for Reduced Locomotion in Dairy Cattle on Nineteen Farms
324 in the Netherlands. Journal of Dairy Science 89, 1509-1515.
- 325 Cohen, J., 1960. A co-efficient of agreement for nominal scales. Educational and
326 Psychological Measurement 20, 37–47.
- 327 Cohen, J., 1968. Weighted kappa; nominal scale agreement with provision for scaled
328 disagreement or partial credit. Psychological Bulletin 70, 213–220.
- 329 Ebel, R.L., 1951. Estimation of the Reliability of Ratings. Psychometrika 16, 407-
330 424.
- 331 Dawson-Saunders, B., and Trapp, R.G., 1994. Basic and Clinical Biostatistics.
332 London. Prentice Hall International.
- 333 Fuller, C.J., Bladon, B.M., Driver, A.J., Barr, A.R.S., 2006. The Intra- and Inter-
334 Assessor Reliability of Measurement of Functional Outcome by Lameness
335 Scoring in Horses. The Veterinary Journal 171, 281-286.
- 336 Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass
337 correlation coefficient as measures of reliability. Educational and Psychological
338 Measurement. 33, 613-619.
- 339 Hewetson, M., Christley, T.M., Hunt, I.D., Voute, L.C., 2006. Investigations of the
340 Reliability of the Observational Gait Analysis for the Assessment of Lameness in
341 Horses. Veterinary Record 158, 852-858.
- 342 Kaler, J., Green, L.E., 2007. Naming and recognition of six foot lesions of sheep
343 using written and pictorial information: A study of 809 English sheep farmers,
344 Prev. Vet. Med.doi:10.1016/j.prevetmed.2007.06.003
- 345 Keegan, K.G., Wilson, D.A., Wilson, D.J., Smith, B., Gaughan, E.M., Pleasant, R.S.,
346 Lillich, J.D., Kramer, J., Howard, R.D., Bacon-Miller, C., Davis, E.G., May,
347 K.A., Cheramie, H.S., Valentino, W.L., Van Harreveld, P.D., 1998. Evaluation of
348 Mild Lameness in Horses Trotting on a Treadmill by Clinicians and Interns Or
349 Residents and Correlation of their Assessments with Kinematic Gait Analysis.
350 American Journal of Veterinary Research 59, 1370-1377.

- 351 Kestin, S.C., Knowles, T.G., Tinch, A.E., Gregory, N.G., 1992. Prevalence of Leg
352 Weakness in Broiler Chickens and its Relationship with Genotype. *Veterinary*
353 *Record* 131, 190-194.
- 354 Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for
355 categorical data. *Biometrics* 33, 159–174.
- 356 Ley, S.J., Livingston, A., Waterman, A.E., 1989. The Effect of Chronic Clinical Pain
357 on Thermal and Mechanical Thresholds in Sheep. *Pain* 39, 353-357.
- 358 Ludbrook, J., 2002. Statistical Techniques For Comparing Measurers And Methods
359 Of Measurement: A Critical Review. *Clinical and Experimental Pharmacology and*
360 *Physiology* 29, 527-536.
- 361 Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the Kappa statistic.
362 *American Journal of Epidemiology* 126, pp. 161-169.
- 363 Main, D.C.J., Clegg, J., Spatz, A., Green, L.E., 2000. Repeatability of a locomotion
364 scoring system for finishing pigs. *Veterinary Record* 147, 574-576.
- 365 Manson, F.J., Leaver, J.D., 1988. The Influence of Concentrate Amount on
366 Locomotion and Clinical Lameness in Dairy Cattle. *Animal Production* 47, 185-
367 190.
- 368 May, S.A., Wyn-Jones, G., 1987. Identification of Hindleg Lameness. *Equine*
369 *Veterinary Journal* 19, 185-188.
- 370 McGraw, K.O., Wong, S.P., 1996. Forming Inferences about Some Intraclass
371 Correlation Coefficients. *Psychological Methods* 1, pp. 30-46
- 372 Morris, C., Galuppi, B.E., Rosenbaum, P.L., 2004. Reliability of Family Report for
373 the Gross Motor Function Classification System. *Developmental Medicine and*
374 *Child Neurology* 46, 455-460.
- 375 Moss, P.A., 1994. Can there be Validity without Reliability? *Educational Researcher*
376 23, 5-12.
- 377 Nelson, L.M., Longstreth Jr., W.T., Koepsell, T.D., Van Belle, G., 1990. Proxy
378 respondents in epidemiologic research. *Epidemiologic Review* 12, 71-86.
- 379 Nelson, J.C., Pepe, M.S., 2000. Statistical description of interrater variability in ordinal
380 ratings. *Statistical Methods in Medical Research* 9 (5), 475-496.
- 381 Nieuwhof, G.J., Bishop, S.C., 2005. Costs of the major endemic diseases of sheep in
382 Great Britain and the potential benefits of reduction in disease impacts. *Animal*
383 *Science* 81, 57–67.

- 384 Reid, J., Nolan, A.M., 1991. A Comparison of the Postoperative Analgesic and
385 Sedative Effects of Flunixin and Papaveretum in the Dog. *Journal of Small*
386 *Animal Practice* 32, 603-608.
- 387 Shoukri, M.M., Asyali, M.H., Donner, A., 2004. Sample size requirements for the
388 design of reliability study: Review and new results. *Statistical Methods in*
389 *Medical Research* 13, 251-271.
- 390 Shrout, P.E., Fleiss, J.L., 1979. Intraclass Correlations: Uses in Assessing Rater
391 Reliability. *Psychological Bulletin* 86, 420-428.
- 392 Shrout, P.E., 1998. Measurement Reliability and Agreement in Psychiatry. *Statistical*
393 *Methods in Medical Research* 7, 301-317.
- 394 Sprecher, D.J., Hostetler, D.E., Kaneene, J.B., 1997. A lameness scoring system that
395 uses posture and gait to predict dairy cattle reproductive performance
396 *Theriogenology* 47, 1179-1187.
- 397 Stashak, T.S., 2002. Diagnosis of lameness. In: Stashak, T.S. (Eds.), *Adams*
398 *Lameness in Horses* (Fifth edition), Lippincott, Williams & Wilkins,
399 Philadelphia. Chapter 3, 113-183.
- 400 Uebersax JS. Statistical modeling of expert ratings on medical treatment
401 appropriateness. 1993. *Journal of the American Statistical Association* 88, 421-
402 427.
- 403 Uebersax, J.S., 2001. Statistical Methods for Rater Agreement
404 <http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>. Accessed
405 November 7, 2006.
- 406 Walter, S.D., Eliasziw, M., Donner, A., 1998. Sample Size and Optimal Designs for
407 Reliability Studies. *Statistics in Medicine* 17, 101-110.
- 408 Welsh, E.M., Gettinby, G., Nolan, A.M., 1993. Comparison of a Visual Analogue
409 Scale and a Numerical Rating Scale for Assessment of Lameness, using Sheep as
410 a Model. *American Journal of Veterinary Research* 54, 976-983.
- 411

412 Table 1. The locomotion scoring scale, shaded area = all required for score
 413
 414

Scale	0	1	2	3	4	5	6
Posture and locomotion							
Bears weight evenly on all four feet							
Uneven posture, but no clear shortening of stride							
Short stride on one leg compared with others							
Visible nodding of head in time with short stride							
Excessive flicking of head, more than nodding, in time with short stride							
Not weight bearing on affected limb when standing							
Discomfort when moving							
Not weight bearing on affected limb when moving							
Extreme difficulty rising							
Reluctant to move once standing							
More than one limb affected							
Will not stand or move							

415

416 Table 2: Levels of agreement between and within observers

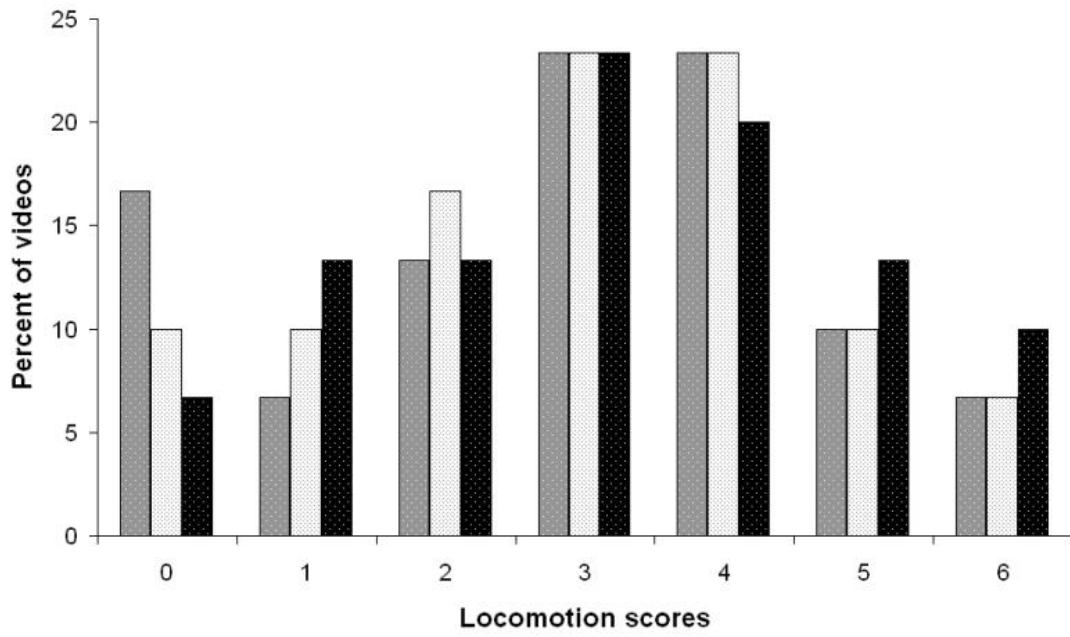
417

Observer (s)	Between observers (% , N/30)			Within observers(% , N/30)		
	1 and 2	2 and 3	1 and 3	1	2	3
Exact agreement	70 21	70 21	63 19	77 23	73 22	77 23
One point difference	30 9	27 8	33 10	20 6	23 7	17 5
Two points difference	-	3 1	3 1	-	3 1	3 1
Three points difference	-	-	-	3 1	-	3 1
Kendall's rank correlation coefficient	0.88	0.88	0.85	0.87	0.87	0.82
Intra-class correlation coefficient (95% CI ^a)		0.93 (0.87-0.96)		0.89 (0.79-0.94)	0.92 (0.84-0.96)	0.90 (0.81-0.95)
Weighted Kappa (95% CI ^a)	0.95 (0.91-0.99)	0.92 (0.86-0.98)	0.93 (0.88-0.98)	0.92 (0.82-1.00)	0.93 (0.87-0.99)	0.89 (0.78-1.00)

418

419 ^aCI = confidence interval

420 Fig 1. Distribution of locomotion scores ($n = 30$) by observers 1 – 3 as grey, white
421 and black bars



422