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

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)

Abstract

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

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

Introduction

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

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

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

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

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

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

Materials and Methods

Locomotion scoring scale

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

Sample size estimation for reliability

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

Locomotion scoring scale movies

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

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

Observers

Three observers were randomly selected from a group of researchers familiar with observing locomotion in sheep. They were given a training session to learn to score sheep locomotion based on what they saw in the clip, and using the descriptions in Table 1, using ten movies with at least one of each score in the locomotion scoring scale. This was followed by some test movie clips to ensure that the duration of the clips and the between clip intervals were familiar to the observers. Finally, the observers recorded their observations of the 30 movie clips in a room, sitting apart from each other, using a copy of Table 1 and a recording form with the clip numbers listed sequentially and a row of scores 0 – 6 for each clip with instructions to circle one score per clip. The clips ran without a break. The forms were collected immediately after the session.

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

Data analysis

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

Inter- and Intra-observer reliability

a) Percent agreement

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

$$\frac{\text{Percent agreement (disagreement)} = (\text{number of exact agreements (disagreements)} * 100}{\text{Total number of observations100}}$$

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

b) Intraclass correlation coefficient (ICC)

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

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

where μ = population mean for all ratings, $r_{i\,=}$ random sheep effect, c_j = random observer effect, rc_j = random interaction effects and e_{ij} = residual or random error. Normality of the data was checked by Shapiro-Wilk normality test. Estimates for ICC were interpreted using previously recommended guidelines: 0-10% - virtually none, 11-40%- slight, 41-60% fair, 61-80% moderate and 81-100% substantial agreement (Shrout, 1998).

c) Weighted kappa coefficients (κ_w)

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

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

Inter- and intra-observer associations

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

Observer bias

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

Results

Inter- and Intra- observer reliability

a) Percent agreement

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

b) Intraclass correlation coefficients (ICC)

The Sharpio- Wilk test did not reject the normality of the data at $P \leq 0.05$. The ICC for inter- observer reliability was 0.93, (95% CI: 0.87-0.96) for the locomotion scoring scale indicating substantial agreement between observers. The mean intra-observer reliability was also substantial at 0.90 (95% CI: 0.89-0.92) with a range of 0.89 to 0.92 by observer (Table 2).

c) Weighted kappa coefficients (κ_w)

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

Inter- and intra-observer associations

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

Observer bias

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

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

Discussion

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

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

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

Despite the objectivity of the movie clips of sheep locomotion, observers may vary because of different interpretations of the locomotion scoring system because they drift in scoring or because they are distracted while making a judgement, all these aspects are combined and a final score is given by an observer (Uebersax, 2001). The overall effect is to reduce correlation of scores between and within observers. This reduced correlation provides evidence that there is a random (immeasurable) error and noise in observing method (Uebersax, 2001). In the study presented here the greatest disagreement occurred between scores 0 and 1 between observers (Figure 1). This was as hoped; score 1 is a very slight abnormal gait (Table 1) and provides an interim category between normal (score 0) and definitely lame (score 2). This can be very useful in quantitative research. Conversely, observers may have reduced the within observer variability by remembering movie clips and scoring sheep identically, rather than assessing locomotion independently on the second test. There were only 4 h between the tests but re-ordering and re-numbering the clips should have minimised this effect. Finally, there was no statistical significant evidence for bias between observers. This was useful information because the presence of observer bias can affect the reliability of a scale considerably (Hewetson et al., 2006).

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

Our data were ordinal and so the appropriate measures of agreement were a weighted kappa and intra-class correlation coefficients (Nelson et al., 1990; Morris et al., 2004). When quadratic weights are applied to calculate kappa these two reliability coefficients are equivalent (Fleiss and Cohen, 1973; Ludbrook, 2002), as seen in the study presented in this paper. One of the limitations of these measurements is that they are not comparable across populations because kappa is influenced by the prevalence of a trait; this is equivalent to the dependence of ICC on between subject variance (Maclure and Willet, 1987). Thus the agreement estimates can only be generalised to a population with similar characteristics. In addition, both ICC and kappa are affected by the number of ordinal categories in a scale. Although ICC is less affected by the change in the number of categories, it tends to increase with an increase in number of categories; in contrast, kappa tends to decrease with more categories (Maclure and Willet, 1987). As a result we recommend that anyone adopting this scale tests its reliability for their purpose since the agreement measures depend on the prevalence of lameness as well as the training of personnel.

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

Conclusion

The scoring scale presented in this paper is objective and based on a group of visual observations and a highly reliable method for trained observers to assess locomotion in sheep. It may be used by trained researchers, and possibly advisers, to monitor locomotion in sheep.

6. Acknowledgements

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

References

Agresti A. Modelling patterns of agreement and disagreement. 1992 . Statistical Methods in Medical Research 1, 201-228.

Amory, J.R., Kloosterman, P., Barker, Z.E., Wright, J.L., Blowey, R.W., Green, L.E., 2006. Risk Factors for Reduced Locomotion in Dairy Cattle on Nineteen Farms in the Netherlands. Journal of Dairy Science 89, 1509-1515.

Cohen, J., 1960. A co-efficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–47.

Cohen, J., 1968. Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220.

Ebel, R.L., 1951. Estimation of the Reliability of Ratings. Psychometrika 16, 407-424.

Dawson-Saunders, B., and Trapp, R.G., 1994. Basic and Clinical Biostatistics. London. Prentice Hall International.

Fuller, C.J., Bladon, B.M., Driver, A.J., Barr, A.R.S., 2006. The Intra- and Inter-Assessor Reliability of Measurement of Functional Outcome by Lameness Scoring in Horses. The Veterinary Journal 171, 281-286.

Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement. 33, 613-619.

Hewetson, M., Christley, T.M., Hunt, I.D., Voute, L.C., 2006. Investigations of the Reliability of the Observational Gait Analysis for the Assessment of Lameness in Horses. Veterinary Record 158, 852-858.

Kaler, J., Green, L.E., 2007. Naming and recognition of six foot lesions of sheep using written and pictorial information: A study of 809 English sheep farmers, Prev. Vet. Med.doi:10.1016/j.prevetmed.2007.06.003

Keegan, K.G., Wilson, D.A., Wilson, D.J., Smith, B., Gaughan, E.M., Pleasant, R.S., Lillich, J.D., Kramer, J., Howard, R.D., Bacon-Miller, C., Davis, E.G., May, K.A., Cheramie, H.S., Valentino, W.L., Van Harreveld, P.D., 1998. Evaluation of Mild Lameness in Horses Trotting on a Treadmill by Clinicians and Interns Or Residents and Correlation of their Assessments with Kinematic Gait Analysis. American Journal of Veterinary Research 59, 1370-1377.

Kestin, S.C., Knowles, T.G., Tinch, A.E., Gregory, N.G., 1992. Prevalence of Leg Weakness in Broiler Chickens and its Relationship with Genotype. Veterinary Record 131, 190-194.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Ley, S.J., Livingston, A., Waterman, A.E., 1989. The Effect of Chronic Clinical Pain on Thermal and Mechanical Thresholds in Sheep. Pain 39, 353-357.

Ludbrook, J., 2002. Statistical Techniques For Comparing Measurers And Methods Of Measurement: A Critical Review. Clinical and Experimental Pharmacology and Physiology 29, 527-536.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the Kappa statistic. American Journal of Epidemiology 126, pp. 161-169.

Main, D.C.J., Clegg, J., Spatz, A., Green, L.E., 2000. Repeatability of a locomotion scoring system for finishing pigs. Veterinary Record 147, 574-576.

Manson, F.J., Leaver, J.D., 1988. The Influence of Concentrate Amount on Locomotion and Clinical Lameness in Dairy Cattle. Animal Production 47, 185-190.

May, S.A., Wyn-Jones, G., 1987. Identification of Hindleg Lameness. Equine Veterinary Journal 19, 185-188.

McGraw, K.O., Wong, S.P., 1996. Forming Inferences about Some Intraclass Correlation Coefficients. Psychological Methods 1, pp. 30-46

Morris, C., Galuppi, B.E., Rosenbaum, P.L., 2004. Reliability of Family Report for the Gross Motor Function Classification System. Developmental Medicine and Child Neurology 46, 455-460.

Moss, P.A., 1994. Can there be Validity without Reliability? Educational Researcher 23, 5-12.

Nelson, L.M., Longstreth Jr., W.T., Koepsell, T.D., Van Belle, G., 1990. Proxy respondents in epidemiologic research. Epidemiologic Review 12, 71-86.

Nelson, J.C., Pepe, M.S., 2000.Statistical description of interrater variability in ordinal ratings. Statistical Methods in Medical Research 9 (5), 475-496.

Nieuwhof, G.J., Bishop, S.C., 2005. Costs of the major endemic diseases of sheep in Great Britain and the potential benefits of reduction in disease impacts. Animal Science 81, 57–67.

Reid, J., Nolan, A.M., 1991. A Comparison of the Postoperative Analgesic and Sedative Effects of Flunixin and Papaveretum in the Dog. Journal of Small Animal Practice 32, 603-608.

Shoukri, M.M., Asyali, M.H., Donner, A., 2004. Sample size requirements for the design of reliability study: Review and new results**.** Statistical Methods in Medical Research 13, 251-271.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin 86, 420-428.

Shrout, P.E., 1998. Measurement Reliability and Agreement in Psychiatry. Statistical Methods in Medical Research 7, 301-317.

Sprecher, D.J., Hostetler, D.E., Kaneene, J.B., 1997. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance Theriogenology 47, 1179-1187.

Stashak, T.S., 2002. Diagnosis of lameness. In: Stashak, T.S. (Eds.), Adams Lameness in Horses (Fifth edition), Lippincott, Williams & Wilkins, Philadelphia.Chapter 3, 113-183.

Uebersax JS. Statistical modeling of expert ratings on medical treatment appropriateness.1993. Journal of the American Statistical Association 88, 421-427.

Uebersax, J.S., 2001. Statistical Methods for Rater Agreement http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm. Accessed November 7, 2006.

Walter, S.D., Eliasziw, M., Donner, A., 1998. Sample Size and Optimal Designs for Reliability Studies. Statistics in Medicine 17, 101-110.

Welsh, E.M., Gettinby, G., Nolan, A.M., 1993. Comparison of a Visual Analogue Scale and a Numerical Rating Scale for Assessment of Lameness, using Sheep as a Model. American Journal of Veterinary Research 54, 976-983.

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

<table>
<tr><th>Scale</th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th></tr>
<tr><td>Posture and locomotion</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Bears weight evenly on all four feet</td><td style="background:#999"></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Uneven posture, but no clear shortening of stride</td><td></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Short stride on one leg compared with others</td><td></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Visible nodding of head in time with short stride</td><td></td><td></td><td style="background:#999"></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Excessive flicking of head, more than nodding, in time with short stride</td><td></td><td></td><td></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Not weight bearing on affected limb when standing</td><td></td><td></td><td></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Discomfort when moving</td><td></td><td></td><td></td><td style="background:#999"></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Not weight bearing on affected limb when moving</td><td></td><td></td><td></td><td></td><td style="background:#999"></td><td style="background:#999"></td><td></td></tr>
<tr><td>Extreme difficulty rising</td><td></td><td></td><td></td><td></td><td></td><td style="background:#999"></td><td></td></tr>
<tr><td>Reluctant to move once standing</td><td></td><td></td><td></td><td></td><td></td><td style="background:#999"></td><td></td></tr>
<tr><td>More than one limb affected</td><td></td><td></td><td></td><td></td><td></td><td style="background:#999"></td><td></td></tr>
<tr><td>Will not stand or move</td><td></td><td></td><td></td><td></td><td></td><td></td><td style="background:#999"></td></tr>
</table>

Table 2: Levels of agreement between and within observers

	Between observers (%, N/30)			**Within observers(%, N/30)**		
Observer (s)	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)

[a]CI = confidence interval

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