Body condition scoring of sheep: intra- and inter-observer variability

RA Corner-Thomas[a]*, AM Sewell[b], P Kemp[a], BA Wood[a], DI Gray[a], ST Morris[a], HT Blair[a] and PR Kenyon[a]

[a]*School of Agriculture and Environment, Massey University, Palmerston North, New Zealand,* [b]*Institute of Education, Massey University, Palmerston North, New Zealand*

**Corresponding author.* E-mail: r.corner@massey.ac.nz

Abstract

Body condition scoring (BCS) is a hands-on tool that farmers can use to make decisions about their animal feeding and management. BCS, however, is a subjective measure of the muscle and fat cover of the lumbar spine. Observers, therefore, may show variability in scores both across time and from other observers. This study aimed to determine the intra- and inter-observer variation of both farmers and research technicians as part of a learning exercise of a farmer-learning group based at Massey University between 2011 and 2015. Nineteen farmers and three research technicians condition scored 45 mixed-age ewes on two consecutive days. Data from both farmers and technicians were analysed to determine the intra- and inter-observer variability using a weighted kappa. The results indicate that the majority of farmers and technicians had 'excellent' agreement (21 of the 22 observers had kappa values greater than 0.75) between days. Similarly, among pairs of observers the agreement was also 'excellent' (212 of 231 comparisons had kappa values greater than 0.75). The distribution of scores that contributed to each median condition score, however, indicated that lower scores (1, 1.5 and 2) has less variability than did higher scores (2.5 or greater). These results suggest that BCS is a robust farm-management tool that can be used with a high degree of repeatability.

Keywords: body condition scoring; intra-observer variability, inter-observer variability

Introduction

Body condition scoring (BCS) of sheep is a management tool that farmers can use to aid on-farm decision making and optimise animal performance (Kenyon et al. 2014). BCS provides a subjective assessment of the subcutaneous fat and muscle of the lumbar spine (Jefferies 1961). BCS is assessed by the palpation of both the spinous and transverse processes of the lumbar vertebrae and is assessed against a five-point scale which includes either half or, sometimes, quarter scores (Jefferies 1961; Russel 1984). BCS has advantages over the assessment of live weight alone as BCS is not influenced by the stature of the animal nor by its physiological state, fleece weight or gut fill (Jefferies 1961; Sezenler et al. 2011).

Body condition score has a positive relationship with range of reproductive parameters such as ovulation rate, conception rate, pregnancy rate, fecundity, and with lamb birth weight and survival (Kenyon et al. 2014). For most production traits, such as ovulation rate and pregnancy rate, there is a curvilinear relationship among scores with the response plateauing to an optimum usually between 3 and 3.5 (Gunn et al. 1991; Kenyon et al. 2014). Due to this curvilinear relationship selecting ewes with sub-optimal BCS for additional feeding will likely result in a greater production increase for ewes with poor condition compared with greater BCS. The use of BCS is a potentially valuable farm management tool, however, its value is maximised if consistency of scoring within or between assessors can be achieved.

Due to its subjective nature, BCS has the potential to be limited by repeatability, both within and between assessors. In their review, Kenyon et al. (2014) reported that there was inconsistency in the repeatability of BCS technique between and within assessors, which had been attributed to the different experience of the assessors. In general, low reliability was found in studies with inexperienced assessors compared to studies of experienced assessors. The aim of the current study was to determine the intra- and inter-observer repeatability of a group of farmers and research technicians involved in a farmer-learning group at Massey University.

Materials and methods

From 2011 to 2015 a farmer-learning project was conducted at Massey University. The group of 26 farmers worked with an interdisciplinary group of seven University experts (three animal scientists, an agronomist, a farm-management specialist, an educationalist and a sociologist; Blair et al. 2013). The project focused on a University farmlet trial that investigated lamb finishing on herb-mix pastures (clover, chicory and plantain; Kenyon et al. 2017). The participants met four times per year at Massey University during a 24-hour period from noon to noon. Farmer participants were involved in a number of learning experiences. One of the learning experiences was focussed on the technique of sheep BCS.

Of the 26 farmer participants, 23 attended the session during which the BCS activity was conducted. The session began with a presentation by an experienced academic about BCS followed by a discussion about the technique and the potential opportunities that BCS could provide to the farmers. Participants were provided with both a one-page fact sheet on the technique created by the university experts, and a copy of a one-page handout produced by Beef+Lamb NZ (2013). Prior to the session an experienced BCS- technician selected and weighed 45 mixed-age ewes which they assessed to have a BCS between 1.5 and 5 (Jefferies 1961). The experienced technicians in this study

regularly BCS sheep as part of their research studies and had worked together for a number of years. The learning group participants (n=23 observers) and technicians experienced in BCS (n=3 observers) were then asked to manually record to 0.5 the BCS of all 45 ewes. The following day the participants and technicians recorded the BCS of the same 45 ewes. No record was made of the previous experience of the observers in this study.

Statistical analysis

All statistical analyses were conducted using SPSS 26.0 (SPSS Inc, 2006). Data from four observers (n=4 farmers) was discarded due to data being recorded for only one of the two observation days. This resulted in data being analysed for a total of 19 farmers and three experienced technicians. Across the two days of observations, nine observers recorded scores for all 45 ewes, 12 observers did not record a BCS for one or two ewes and four observers omitted between eight and ten ewes. Descriptive statistics including mode, minimum, maximum, median, and lower and upper quartiles were computed for BCS data recorded on each day and for the ΔBCS between days. Descriptive statistics were also calculated for each observer. The percentage of exact agreement between and within observers was calculated and the inter- and intra-observer variability was assessed using intra-class correlation coefficients (ICC; Shrout & Fleiss 1979) and weighted kappa coefficients (K_w) (Cohen 1968). BCS is an ordinal measurement and, therefore, the weighted κ was used as it attributed more weight to large measurement differences than to small ones. κ_w was calculated the using quadratic weights for paired intra- and inter-observer assessments. All κ_w results were interpreted according to Fleiss (1981), where values >0.75 suggested 'excellent', 0.4 to 0.75 indicated 'fair-good' and <0.4 indicated 'poor' levels of agreement.

Results

Body condition scores

For the 45 ewes that were observed, a total of 1927 individual BCS were recorded by the 22 observers during two days of observation. Kendall's coefficient of concordance was 0.887, which indicated that for each ewe there was a high level of agreement across all observers and both days of observation. A breakdown of the scores that contributed to each ewe's median BCS is shown in Table 1. The ewe that had a median BCS of 1, was scored 1 in almost 80% of the observations and scored a 1.5 by just over 20% (Table 1). In contrast, the four ewes with a median score of 1.5 were scored a 1, 1.5, 2 and 2.5. The median score with the greatest variability in scores was 3.5 which was given scores ranging from 2 to 5.

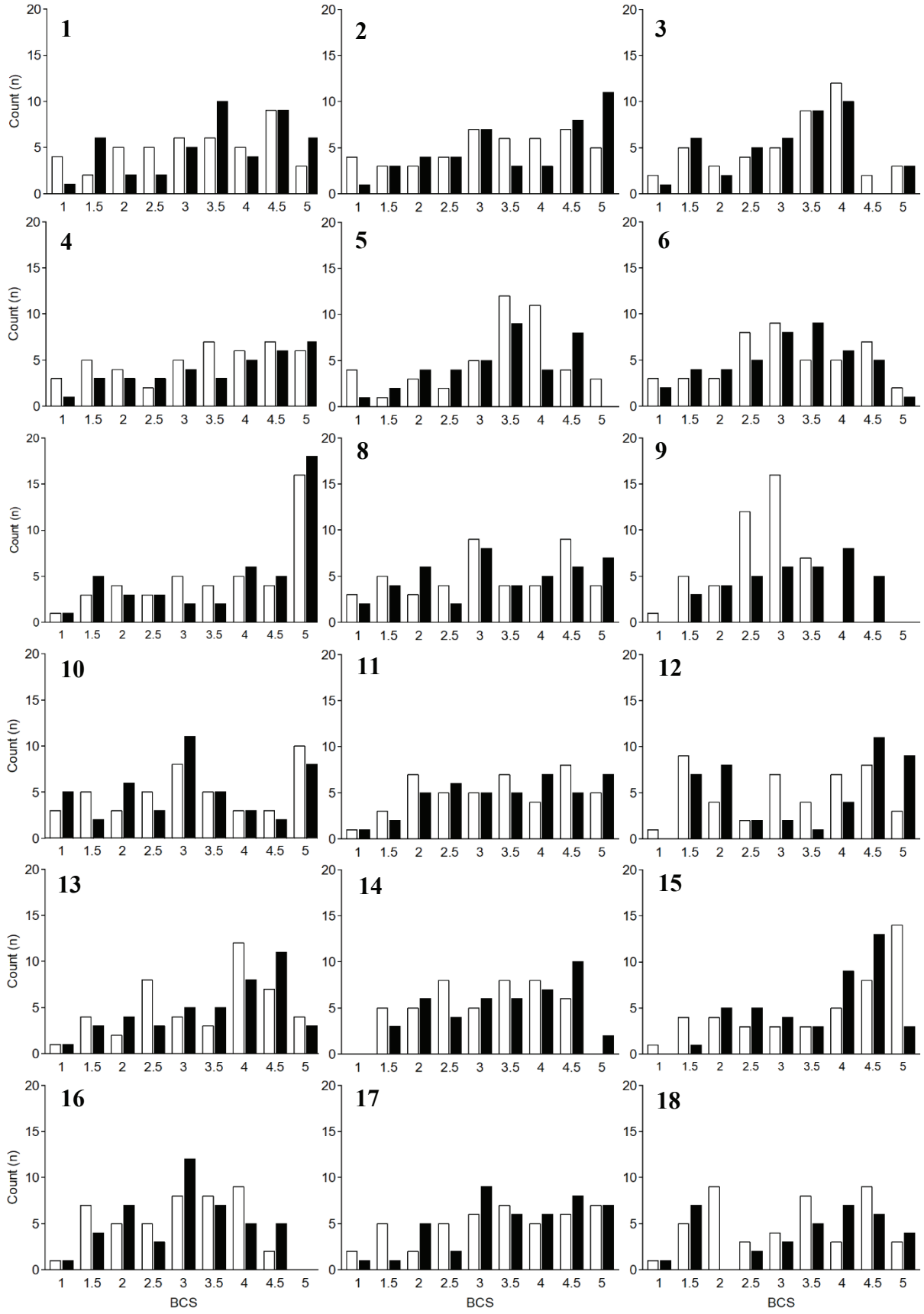
Intra-observer variability

The BCS data recorded by each observer on day 1 and 2 are shown in Figure 1. The average exact agreement between day 1 and 2 ranged from 16% (observer 9) to 71% (observer 4; Table 2). The percentage of observations that were within half a score of the previous day's observation ranged from 70 to 98% (Table 2). Similarly (the κ_w values suggest that all but one observer had 'excellent' agreement between days (Table 2). While (observer 9 had a $\kappa_w = 0.61$ which indicated 'fair-good' agreement.

Table 1 The percentage (and frequency) of each median body condition score (BCS) showing the number of ewes in each median category, their mean live weight and the distribution of scores that contributed to that score.

Median BCS	n ewes	Live weight (kg)	Distribution of individual BCS for each median value								
			1	1.5	2	2.5	3	3.5	4	4.5	5
1	1	41.5	79.1 (34)	20.9 (9)	-	-	-	-	-	-	-
1.5	4	61.5	18.0 (31)	59.9 (103)	19.8 (34)	2.3 (4)	-	-	-	-	-
1.75	1	58.5	-	50.0 (19)	45.2 (19)	4.8 (2)	-	-	-	-	-
2	3	66.8	-	26.0 (33)	54.3 (69)	19.7 (25)	-	-	-	-	-
2.5	5	69.5	-	1.4 (3)	25.9 (56)	51.9 (112)	19.4 (42)	0.9 (2)	0.5 (1)	-	-
3	6	77.8	-	-	3.1 (8)	19.4 (50)	47.7 (123)	21.7 (56)	6.2 (16)	1.9 (5)	
3.5	6	80.6	-	-	0.4 (1)	2.0 (5)	25.1 (64)	41.2 (105)	22.0 (56)	8.6 (22)	0.8 (2)
4	6	90.9	-	-	-	0.4 (1)	7.0 (18)	24.4 (63)	31.8 (82)	26.4 (68)	10.1 (26)
4.5	8	90.6	-	-	-	-	4.1 (14)	11.9 (41)	23.5 (81)	37.8 (130)	22.7 (78)
5	5	92.2	-	-	-	-	1.4 (3)	2.4 (5)	12.3 (26)	25.0 (53)	59.0 (125)

Figure 1 Frequency of body condition scores of ewes recorded by observers 1 to 22 (observer id is shown in the top left of each panel) on day 1 (open bars) and 2 (closed bars) of observation



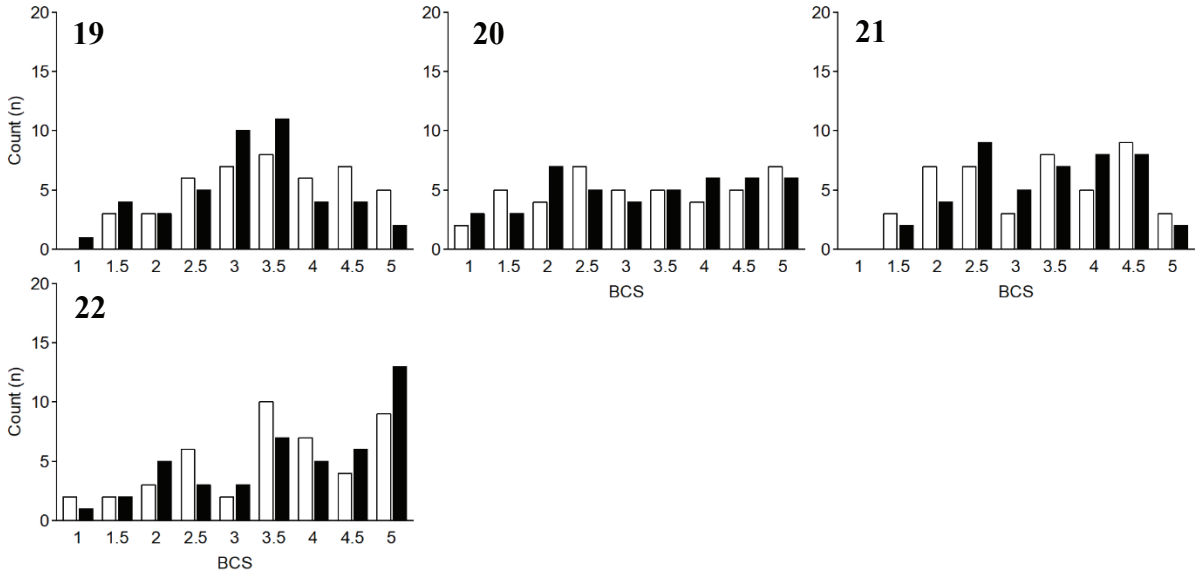


Table 2 Descriptive statistics of body condition scores (BCS) recorded by each observer (Obs) including number of individual ewe BCS observations (n) and the median BCS recorded (the exact agreement and agreement half a score between day 1 and 2 (and the intra-observer weighted kappa (κ_w)

Obs	Observer	BCS		Agreement Day 1 vs 2		Intra-observer
Id		n	Median	Exact (%)	Half score (%)	κw
1	Farmer	90	3.5	42.2	91.1	0.92 (0.87 - 0.97)
2	Farmer	89	3.5	29.5	72.7	0.85 (0.78 - 0.92)
3	Farmer	87	3.5	71.4	97.6	0.96 (0.93 - 0.99)
4	Farmer	80	3.5	51.4	94.3	0.94 (0.91 - 0.98)
5	Farmer	82	3.5	24.3	70.3	0.76 (0.64 - 0.89)
6	Farmer	89	3	50.0	93.2	0.92 (0.88 - 0.96)
7	Farmer	90	4	53.3	95.6	0.96 (0.93 - 0.98)
8	Farmer	89	3	38.6	79.5	0.88 (0.82 - 0.94)
9	Farmer	82	3	16.2	73.0	0.61 (0.45 - 0.76)
10	Farmer	90	3	51.1	86.7	0.90 (0.82 - 0.97)
11	Farmer	88	3.5	41.9	88.4	0.89 (0.82 - 0.96)
12	Farmer	89	3.5	38.6	72.7	0.84 (0.75 - 0.93)
13	Farmer	88	4	38.6	93.0	0.92 (0.88 - 0.97)
14	Farmer	89	3.5	46.5	81.8	0.87 (0.81 - 0.93)
15	Farmer	88	4	45.5	97.7	0.93 (0.91 - 0.96)
16	Farmer	89	3	34.9	88.6	0.89 (0.83 - 0.95)
17	Farmer	90	3.5	56.8	80.0	0.89 (0.83 - 0.94)
18	Farmer	80	3.5	40.0	80.0	0.87 (0.79 - 0.94)
19	Farmer	89	3.5	31.4	86.4	0.83 (0.73 - 0.93)
20	Technician	90	3	40.9	86.4	0.90 (0.85 - 0.96)
21	Technician	90	3.5	46.7	97.8	0.93 (0.90 - 0.96)
22	Technician	90	3.5	53.3	86.7	0.91 (0.86 - 0.97)

Inter-observer variability

The assessment of pairs of observers had κ_w values that ranged between 0.54 (observer 9 vs. 15) and 0.94 (observer 4 vs. 13; Table 3). Among the pairs of observers (the majority were classified as having 'excellent' agreement (>0.75 (n=210) and with the remainder 'Fair-good' (0.4 to 0.75 (n=21). The κ_w between pairs of farmers showed more variability (0.54 to 0.94) than between pairs of technicians (range 0.82 to 0.88; Table 3).

Discussion

BCS in the current study showed both intra- and inter-observer agreement that was 'fair-good' (κ_w=0.4 to 0.75) to 'excellent' (κ_w>0.75). This finding is in agreement with Teixeira et al. (1989) and Shands et al. (2009) who reported that the Pearson's correlation coefficient between experienced observers was between *r*=0.7 and 0.8. Phythian et al. (2012), however, reported lower agreement between veterinarian assessors with a κ_w of 0.4 and 0.6. It possible that the relatively small range of BCS that study (2 to 3.5) compared with the current study (1 to 5) may have influenced the repeatability.

The prior experience of the farmers in this study was not recorded. Kenyon et al. (2014) reported that

Table 3 Pairwise comparisons of body condition scores (BCS) recorded on days 1 and 2 by each observer (Obs 1 to 22) showing the Pearson's correlation coefficient (above the diagonal) and weighted Kappa (below the diagonal). White areas indicate areas of comparisons between farmer observers (light grey areas are farmer and technician observer comparisons and dark grey areas are technician observers.

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6	Obs 7	Obs 8	Obs 9	Obs 10	Obs 11	Obs 12	Obs 13	Obs 14	Obs 15	Obs 16	Obs 17	Obs 18	Obs 19	Obs 20	Obs 21	Obs 22
Obs 1	-	0.91	0.91	0.93	0.85	0.90	0.93	0.91	0.82	0.88	0.87	0.87	0.93	0.93	0.91	0.91	0.92	0.93	0.85	0.91	0.93	0.91
Obs 2	0.91	-	0.89	0.91	0.81	0.88	0.92	0.90	0.79	0.87	0.87	0.86	0.89	0.88	0.90	0.90	0.93	0.90	0.81	0.87	0.86	0.90
Obs 3	0.90	0.86	-	0.94	0.89	0.91	0.92	0.89	0.76	0.91	0.88	0.89	0.93	0.89	0.93	0.94	0.92	0.91	0.89	0.91	0.90	0.91
Obs 4	0.92	0.91	0.92	-	0.88	0.91	0.96	0.93	0.79	0.93	0.90	0.89	0.95	0.93	0.94	0.96	0.94	0.93	0.89	0.92	0.91	0.93
Obs 5	0.84	0.80	0.88	0.86	-	0.84	0.87	0.82	0.70	0.82	0.78	0.81	0.88	0.83	0.86	0.86	0.84	0.84	0.81	0.83	0.79	0.85
Obs 6	0.87	0.83	0.90	0.88	0.83	-	0.90	0.88	0.75	0.93	0.86	0.80	0.90	0.89	0.90	0.91	0.92	0.89	0.87	0.89	0.87	0.89
Obs 7	0.87	0.89	0.80	0.91	0.78	0.76	-	0.93	0.82	0.89	0.89	0.87	0.94	0.92	0.94	0.91	0.92	0.94	0.87	0.90	0.90	0.92
Obs 8	0.91	0.89	0.88	0.92	0.81	0.86	0.86	-	0.79	0.89	0.90	0.85	0.90	0.90	0.90	0.88	0.89	0.91	0.85	0.89	0.89	0.88
Obs 9	0.71	0.66	0.71	0.65	0.62	0.71	0.56	0.69	-	0.71	0.75	0.75	0.77	0.82	0.75	0.80	0.80	0.81	0.68	0.75	0.76	0.80
Obs 10	0.87	0.85	0.90	0.92	0.80	0.91	0.80	0.88	0.64	-	0.88	0.83	0.91	0.90	0.90	0.92	0.90	0.90	0.89	0.91	0.89	0.90
Obs 11	0.87	0.86	0.87	0.89	0.77	0.84	0.83	0.89	0.65	0.86	-	0.87	0.90	0.90	0.85	0.90	0.89	0.89	0.84	0.88	0.88	0.86
Obs 12	0.86	0.85	0.87	0.88	0.79	0.77	0.82	0.85	0.66	0.83	0.86	-	0.89	0.87	0.87	0.87	0.85	0.89	0.80	0.88	0.88	0.89
Obs 13	0.92	0.88	0.90	0.94	0.87	0.85	0.89	0.89	0.65	0.87	0.89	0.88	-	0.91	0.92	0.93	0.90	0.94	0.87	0.90	0.91	0.92
Obs 14	0.92	0.85	0.89	0.90	0.82	0.88	0.80	0.89	0.76	0.87	0.89	0.85	0.89	-	0.89	0.93	0.91	0.92	0.84	0.90	0.91	0.90
Obs 15	0.86	0.87	0.83	0.90	0.80	0.78	0.93	0.84	0.54	0.82	0.81	0.82	0.90	0.81	-	0.91	0.91	0.90	0.89	0.91	0.90	0.90
Obs 16	0.84	0.80	0.92	0.86	0.80	0.89	0.69	0.83	0.79	0.87	0.83	0.80	0.83	0.90	0.71	-	0.93	0.93	0.86	0.89	0.91	0.91
Obs 17	0.92	0.92	0.89	0.93	0.83	0.87	0.88	0.87	0.67	0.87	0.88	0.84	0.90	0.88	0.89	0.82	-	0.90	0.83	0.89	0.87	0.90
Obs 18	0.93	0.88	0.91	0.93	0.83	0.88	0.86	0.90	0.69	0.89	0.89	0.89	0.92	0.91	0.83	0.88	0.89	-	0.84	0.91	0.91	0.90
Obs 19	0.84	0.78	0.89	0.86	0.81	0.86	0.77	0.84	0.61	0.87	0.83	0.77	0.85	0.84	0.82	0.81	0.81	0.83	-	0.86	0.85	0.87
Obs 20	0.90	0.85	0.90	0.91	0.81	0.87	0.82	0.89	0.67	0.91	0.87	0.88	0.88	0.88	0.84	0.84	0.87	0.91	0.84	-	0.91	0.92
Obs 21	0.91	0.84	0.89	0.88	0.79	0.85	0.81	0.87	0.68	0.85	0.88	0.86	0.90	0.90	0.84	0.85	0.86	0.90	0.85	0.89	-	0.91
Obs 22	0.89	0.89	0.84	0.91	0.81	0.80	0.91	0.84	0.62	0.84	0.84	0.86	0.90	0.84	0.90	0.75	0.89	0.85	0.82	0.87	0.87	-

BCS repeatability was inconsistent among studies and hypothesised that this was due to difference in the experience of the assessors. For example Yates and Gleeson (1975) reported poor repeatability of inexperienced assessors whereas Evans (1978) and Shands et al. (2009) reported 'good' or 'excellent' agreement. Keinprecht et al. (2016) reported that the experience of six assessors with BCS of sheep and who were trained for one day prior to the evaluation had no effect of on repeatability or reproducibility of scoring. In that study kappa ranged from 0.54 to 0.80 which was lower than in the current study. The lack of difference in this study may have been a result of the small number of assessors, the quality of the training provided and the short interval between training and evaluation. The farmers in the current study may have had some experience with BCS as it is a farm-management tool used by 43% of New Zealand farmers (Corner-Thomas et al. 2015). The agreement between technicians in this study was high, which is perhaps not surprising given that all three technicians regularly conducted BCS as part of their research duties. In addition, the three technicians had received similar training and had worked closely with one another for a number of years.

In the current study, low BCS (1, 1.5 and 2) were comprised of a lesser range of scores than scores of 2.5 or greater. To these authors' knowledge an assessment of the variability of different scores has not been reported. Teixeira et al. (1989) reported that the relationship of BCS with carcass composition measures of fat of sheep showed that ewes with BCS between 1.5 and 2.5 had higher proportions of intermuscular and mesenteric, whereas for ewes with higher BCS had primarily subcutaneous and omental fat. This difference in fat depot location is likely to result in ewes with BCS less than 2.5 being more easily identified due to their lack of sub-cutaneous fat. Further research is required to determine the repeatability of different scores and the influence of subcutaneous fat deposition on BCS.

Conclusion

Body condition scoring is an important 'hands-on' assessment of the condition of sheep that can easily be learnt (Kenyon et al. 2014). The results of this study indicate that both the farmers and technicians that were involved in this study had a high level of agreement between days and each other. This suggests that BCS is a farm-management tool that farmers can use to increase the productivity of their sheep flock. Further work is required, however, to determine the effect of prior experience with the technique, on the intra- and inter-observer variability.

Acknowledgements

This research was funded by Massey University and Partnership for excellence.

References

Beef + Lamb NZ 2013. Ewe body condition scoring (BCS). Wellington, New Zealand, Beef + Lamb New Zealand. 1 p.

Blair HT, Sewell AM, Corner-Thomas RA, Kemp P, Wood BA, Gray DI, Morris ST, Greer AW, Logan CM, Ridler AL, Hickson RE, Kenyon PR 2013. Understanding how farmers learn. Proceedings of the Association for the Advancement of Animal Breeding and Genetics 20: 1-5.

Cohen J 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin 70: 213.

Corner-Thomas RA, Kenyon PR, Morris ST, Ridler AL, Hickson RE, Greer AW, Logan CM, Blair HT 2015. Influence of demographic factors on the use of farm management tools by New Zealand farmers. New Zealand Journal of Agricultural Research 58: 412-422.

Evans DG 1978. The interpretation and analysis of subjective body condition scores. Animal Science 26: 119-125.

Fleiss JL, Levin B, Paik MC 1981. The measurement of interrater agreement. Statistical methods for rates and proportions 2: 22-23.

Gunn RG, Smith WF, Senior AJ, Bartham E, Sim DA, Hunter EA 1991. Pre-mating herbage intake and the reproductive performance of north country cheviot ewes in different levels of body condition. Animal Production 52: 149-156.

Jefferies BC 1961. Body condition scoring and its use in management. Tasmanian Journal of Agriculture 32: 19-21.

Keinprecht H, Pichler M, Pothmann H, Huber J, Iwersen M, Drillich M 2016. Short term repeatability of body fat thickness measurement and body condition scoring in sheep as assessed by a relatively small number of assessors. Small Ruminant Research 139: 30-38.

Kenyon P, Morel P, Corner-Thomas R, Perez H, Somasiri S, Kemp P, Morris S 2017. Improved per hectare production in a lamb finishing system using mixtures of red and white clover with plantain and chicory compared to ryegrass and white clover. Small Ruminant Research 151: 90-97.

Kenyon PR, Maloney SK, Blache D 2014. Review of sheep body condition in relation to production characteristics. New Zealand Journal of Agricultural Research 57: 38-64.

Phythian C, Hughes D, Michalopoulou E, Cripps P, Duncan J 2012. Reliability of body condition scoring of sheep for cross-farm assessments. Small Ruminant Research 104: 156-162.

Russel AJF 1984. Means of assessing the adequacy of nutrition of pregnant ewes. Livestock Production Science 11: 429-436.

Sezenler T, Özder M, Yildirir M, Ceyhan A, Yüksel M 2011. The relationship between body weight and body condition score in some indigenous sheep breeds in turkey. Journal of Animal and Plant Science 21: 443-447.

Shands CG, Mcleod B, Lollback ML, Duddy G, Hatcher S, O'halloran WJ 2009. Comparison of manual assessments of ewe fat reserves for on-farm use. Animal Production Science 49: 630-636.

Shrout PE, Fleiss JL 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86: 420-428.

Teixeira A, Delfa R, Colomer-Rocher F 1989. Relationships between fat depots and body condition score or tail fatness in the rasa aragonesa breed. Animal Science 49: 275-280.

Yates W, Gleeson A 1975. Relationships between condition score and carcass composition of pregnant merino sheep. Australian Journal of Experimental Agriculture 15: 467-470.