J. Dairy Sci. 100:3086–3090
https://doi.org/10.3168/jds.2016-11872

Short communication: An observational study investigating inter-observer agreement for variation over time of body condition score in dairy cows

P.-A. Morin,*[1] Y. Chorfi,† J. Dubuc,* J.-P. Roy,* D. Santschi,‡ and S. Dufour§
*Faculté de Médecine Vétérinaire, Département des Sciences Cliniques, and
†Faculté de Médecine Vétérinaire, Département de Biomédecine Vétérinaire, Université de Montréal, 3200 rue Sicotte, CP 5000, Saint-Hyacinthe, QC, Canada, J2S 7C6
‡Valacta, Ste-Anne-de-Bellevue, QC, Canada, H9X 3R4
§Faculté de Médecine Vétérinaire, Département de Pathologie et de Microbiologie, Université de Montréal, 3200 rue Sicotte, CP 5000, Saint-Hyacinthe, QC, Canada, J2S 7C6

ABSTRACT

Body condition score (BCS) is strongly correlated with energy reserves. The ease, rapidity of scoring, and high intra- and inter-observer repeatability make it a widely used herd management tool in bovine practice and in scientific studies. Loss or gain of BCS, rather than a single BCS measurement, is frequently used to monitor energy balance in dairy cows. It is unknown if the difference between 2 BCS measures taken at different moments (ΔBCS) would demonstrate inter-observer agreement similar to that of a single BCS measurement. The objective of this study was to compare inter-observer agreement of BCS and ΔBCS in dairy cows when multiple observers perform data collection. An observational study was conducted between April and September 2015; 3 observers independently assessed BCS of 73 Holstein cows from 1 commercial dairy herd. Body condition score assessments of the animals were performed between 1 and 20 d in milk (early lactation; exam 1) and again between 41 and 60 d in milk (peak of milk production; exam 2). Quadratic weighted kappa (κ_w) was computed to quantify agreement between observers for single BCS measurements and ΔBCS. For single BCS measurements, κ_w of 0.79 (95% CI: 0.69, 0.85) and 0.84 (95% CI: 0.77, 0.89) were obtained for exam 1 and exam 2, respectively. Such values would be interpreted as strong agreement and are consistent with the available literature on BCS repeatability. When computing agreement for ΔBCS, a κ_w value of 0.49 (95% CI: 0.32, 0.63) was obtained, suggesting moderate agreement between observers. These findings suggest that studies investigating single BCS measures could use many observers with a high degree of accuracy in the results. When ΔBCS is the parameter of interest, more reliable results would be obtained if one observer conducts all assessments.
Key words: dairy cow, body condition score, inter-observer agreement, kappa

Short Communication

In the early 1980s, Wildman (1982) developed a scale from 1 to 5 to evaluate body condition of dairy cattle at any stage of lactation regardless of BW and size. The strong association between BCS of dairy cows and body energy reserves (Wright and Russel, 1984) and its ease of implementation as a herd management tool, led to its adoption in dairy herd management, as well as in scientific studies (Roche et al., 2009). To optimize data quality, most professionals prefer to have only one evaluator conducting all the BCS measurements (Kristensen et al., 2006). In large field trials conducted over a long period of time, it is not always possible to have only one observer assessing all BCS. The relatively high inter-observer agreement reported among experienced observers in previous studies could suggest that the use of more than one observer would have little effect on the accuracy of BCS measures (Ferguson et al., 1994; Kristensen et al., 2006). Measurement of BCS is usually performed during the peri-partum transition period to qualify loss or gain of energy reserves. For that purpose, difference of BCS (**ΔBCS**) is used and calculated by subtracting the most recent BCS value from its previous measurement for the same cow. Despite the wide use of ΔBCS for monitoring energy balance, no studies have reported the inter-observer repeatability of this measurement. The hypothesis is that, if disagreement between observers is systematic when conducting single BCS measurement, then ΔBCS could yield higher agreement than single BCS. On the other hand, if the difference between observers is random instead of systematic, then the ΔBCS, requiring 2 BCS measurements for computation, could potentially yield lower agreement. The objective of the current study is to

Received August 15, 2016.
Accepted December 21, 2016.
[1]Corresponding author: a.morin@umontreal.ca

evaluate inter-observer agreement of ΔBCS computed by subtracting a BCS measure observed at the peak of milk production from another BCS measure observed during the early lactation, and to compare it with inter-observer agreement of single BCS measures.

The study protocol was presented to the Research Ethical Committee of the Université de Montréal (Saint-Hyacinthe, QC, Canada). Because animals were only observed, the committee approved the current research without the need for a certificate. An observational repeatability study was conducted from April to September 2015 in a single commercial dairy herd of 240 lactating Holstein cows conveniently selected and located in the vicinity of the Bovine Ambulatory Clinic of the Faculté de Médecine Vétérinaire of the Université de Montréal (St-Hyacinthe, QC, Canada). Cows in this herd were housed in a free stall barn, fed a TMR, and milked by milking robots. Mean annual herd milk production was 11,000 kg per cow per year.

Sample size calculation was based on Rotondi and Donner (2012) using the package kappaSize developed for the R software 3.2.3 (The R Project for Statistical Computing, Vienna, Austria). For this analysis, the following parameters were used: an expected quadratic weighted kappa statistic (κ_w) value of 0.86 (Kristensen et al., 2006); a lower bound of 0.7 corresponding to the median of κ_w in the strong agreement category (Landis and Koch, 1977b); a higher bound set to "Not Available," allowing the procedure to generate the number of required subjects for a 1-sided confidence interval; 3 observers; a desired type 1 error rate of 0.05; a normal distribution of BCS and ΔBCS in the population; and finally, 5 possible categories of BCS (Wildman, 1982) with respective prevalence of 0.01, 0.3, 0.38, 0.3, and 0.01. With those settings, a total of 38 cows were estimated. The R package used for the current sample size estimation can only accommodate measurements in less than or equal to 5 categories. Authors were aware that the BCS scale they used included 9 categories (Ferguson et al., 1994). The power of a study generally increases as the number of categories for the variable evaluated increases, thus requiring a smaller sample size (Cohen, 1983). The sample size estimated for the current study was higher than the sample size truly needed. Nevertheless, the estimate of 38 cows was considered a minimum sample size due to the ease of collecting BCS measurements.

Before farm sampling began, the observers, one veterinarian (observer 3) and 2 animal health technicians (observers 1 and 2), reviewed the BCS chart of Elanco (Greenfield, IN) based on the works of Wildman (1982) and Ferguson et al. (1994). Observers 1, 2, and 3 had, respectively, 15, 2, and 9 yr of experience at BCS scoring of dairy cows. Furthermore, observer 3 initially trained observer 2 on BCS scoring. Cows were systematically enrolled as they calved and examined simultaneously and independently by the 3 observers at 2 different moments. A first evaluation (exam 1) was conducted in the first 3 wk of lactation, and a second evaluation (exam 2) was performed between 6 and 8 wk of lactation. Observers were blinded to BCS values reported at first observation, and to the current and previous values from other observers. A minimum interval of 28 d between evaluations was enforced. Sampling cows in early lactation was chosen because this period is critical in term of energy deficiency, lipolysis, and weight loss (Smith and McNamara, 1990; Renaville et al., 2002; Lucy et al., 2009). Measures of BCS, and consequently subcutaneous fat, are generally at their nadir between 40 and 100 DIM (McNamara, 1991; Pedron et al., 1993; Gillund et al., 2001), which generally reflects equality in energy inputs and outputs.

For each cow, the difference between BCS measurements obtained at exam 1 and exam 2 (ΔBCS) was computed for each observer using the 2 collected measurements. Descriptive statistics (mode, minimum, maximum, median, lower, and upper quartiles) were computed for BCS measures obtained at exam 1, at exam 2, and for ΔBCS using the MEANS procedure of SAS (version 9.4, SAS Institute Inc., Cary, NC). Descriptive statistics were also computed for each observer. Scatter plots (SGPLOT procedure in SAS) comparing results across observers were built for single BCS measures observed at exam 1 and exam 2, as well as for ΔBCS, to visually compare agreement between all possible pairs of observers and with the equality line corresponding to perfect agreement (Dohoo et al., 2003).

Body condition score is a qualitative ordinal measurement (Wildman, 1982; Ferguson et al., 1994). The κ_w statistic (Sim and Wright, 2005) was chosen to report agreement beyond chance so that more weight is attributed to large measurement differences than to small ones. For each pair of observers, agreements at exam 1, at exam 2, and for ΔBCS were estimated using κ_w from the FREQ procedure of the SAS software (version 9.4, SAS Institute Inc., Cary, NC). The Bowker's test of symmetry (Bowker, 1948), testing for equal κ_w coefficients for multiple strata with multiple categories, was used to assess heterogeneity between pairs. When this test was statistically nonsignificant, an overall κ_w was produced (Kristensen et al., 2006). Based on Fleiss (1971) and Landis and Koch (1977a), a κ_w value comparing the 3 observers together is equivalent to the weighted average of the individual pairs. These overall κ_w values (Barnhart et al., 2002) were calculated using a macro created by Carrasco et al. (2013) and developed for the SAS software (version 9.4, SAS Institute

Inc., Cary, NC). This macro can be used to calculate the concordance correlation coefficient, which has been shown to be equivalent to the quadratic κ_w when applied to ordinal categorical data (King and Chinchilli, 2001). The interpretations of κ and κ_w values suggested by (Landis and Koch, 1977b) were used for interpretation.

A total of 73 cows were initially enrolled in the study. Nine cows could not be observed by one of the observers at the first exam and were excluded from the single BCS measurement exam 1 and from the ΔBCS analyses. These cows were not excluded from the single BCS exam 2 analyses. Eight cows were culled after their first exam and were excluded from the single BCS exam 2 and ΔBCS analyses. Therefore, a total of 64, 65, and 57 cows were included in the single BCS exam 1, single BCS exam 2, and ΔBCS analyses, respectively.

Descriptive statistics for single BCS measurements and for ΔBCS are presented in Table 1, whereas their distributions can be visualized in Figure 1. Inter-observer agreements for exam 1, exam 2, and ΔBCS are also presented in Figure 1. Examination of the scatter plots suggests a slightly larger disagreement between observers 1 and 2 and between observers 1 and 3 when evaluating single BCS and ΔBCS compared with observers 2 and 3. When visually appraising agreements between observers for single BCS measures compared with agreements for ΔBCS measures, ΔBCS measures seemed to produce larger disagreement as indicated by the larger shift in slope observed on the scatter plots. The maximum difference between observers for single BCS measures was 0.75 point. Disagreements $\leq$0.50 and 0.25 point were observed in 99 and 93% of observations, respectively. Perfect agreement was observed in 48% of observations.

Regarding ΔBCS measures, the maximum difference between observers was 0.75 point. Disagreements $\leq$0.5 and 0.25 point were observed for 97.7 and 83.6% of observations, respectively. Perfect agreement was observed in 33.3% of observations.

The *P*-values from the Bowker's test of symmetry for exam 1, exam 2, and ΔBCS were all greater or equal to 0.96, indicating that heterogeneity was not a problem. Overall quadratic κ_w values of 0.79 (95% CI: 0.69, 0.85) for single BCS measures at exam 1 and of 0.84 (95% CI: 0.77, 0.89) at exam 2 were obtained. These results suggest strong to almost perfect agreement between observers for single BCS measurements. For ΔBCS, this parameter was 0.49 (95% CI: 0.32, 0.63), suggesting moderate agreement between observers.

Results suggest that multiple experienced observers could be used to collect single BCS measures for monitoring or epidemiological studies with almost perfect accuracy. When evaluating ΔBCS, moderate agreement between observers would be obtained. Depending on the objective pursued, the number of observers, and the inter-observer homogeneity between BCS ratings, the ΔBCS data obtained using different observers may still be acceptable. The authors also evaluated agreement of ΔBCS between pair of observers when exam 1 was conducted by one observer and exam 2 by another observer. In such case, a quadratic κ_w value of 0.52 (95% CI: 0.40, 0.63) was observed. This value is within the confidence interval of the overall quadratic κ_w value obtained from the ΔBCS assessed by the same observer at a different time. In a situation in which more than one observer cannot be avoided for data collection, the agreement between observers would not necessarily be affected if cows were to be rated by different observers on the 2 different occasions. The authors would, therefore, suggest that ΔBCS values should be computed, whenever possible, using readings collected by one observer. In this study, however, an important homogeneity between the ratings from the 3 observers was noted. Observer 3 scored many more 3.00 (exam 1: 31 and exam 2: 31) values compared with observer 1

Table 1. Descriptive statistics of BCS measured by 3 different observers on a cohort of 73 cows from 1 commercial dairy

Time	Observer	N[1]	Minimum	Lower quartile	Median	Upper quartile	Maximum
Exam 1[2]	1	64	2.00	2.75	3.25	3.25	4.00
Exam 1[2]	2	64	2.25	3.00	3.25	3.25	4.00
Exam 1[2]	3	64	2.00	3.00	3.00	3.25	4.00
Exam 2[3]	1	65	2.00	2.75	3.00	3.25	4.00
Exam 2[3]	2	65	2.25	2.75	3.00	3.00	3.75
Exam 2[3]	3	65	2.25	2.75	3.00	3.25	3.75
ΔBCS[4]	1	57	−0.75	0.00	0.25	0.50	1.00
ΔBCS[4]	2	57	−0.75	0.00	0.25	0.50	0.75
ΔBCS[4]	3	57	−0.75	0.00	0.25	0.50	1.00

[1]Number of cows.

[2]Evaluation between 1 and 20 DIM.

[3]Evaluation between 41 and 60 DIM.

[4]Difference in BCS (BCS at exam 1 minus BCS at exam 2).

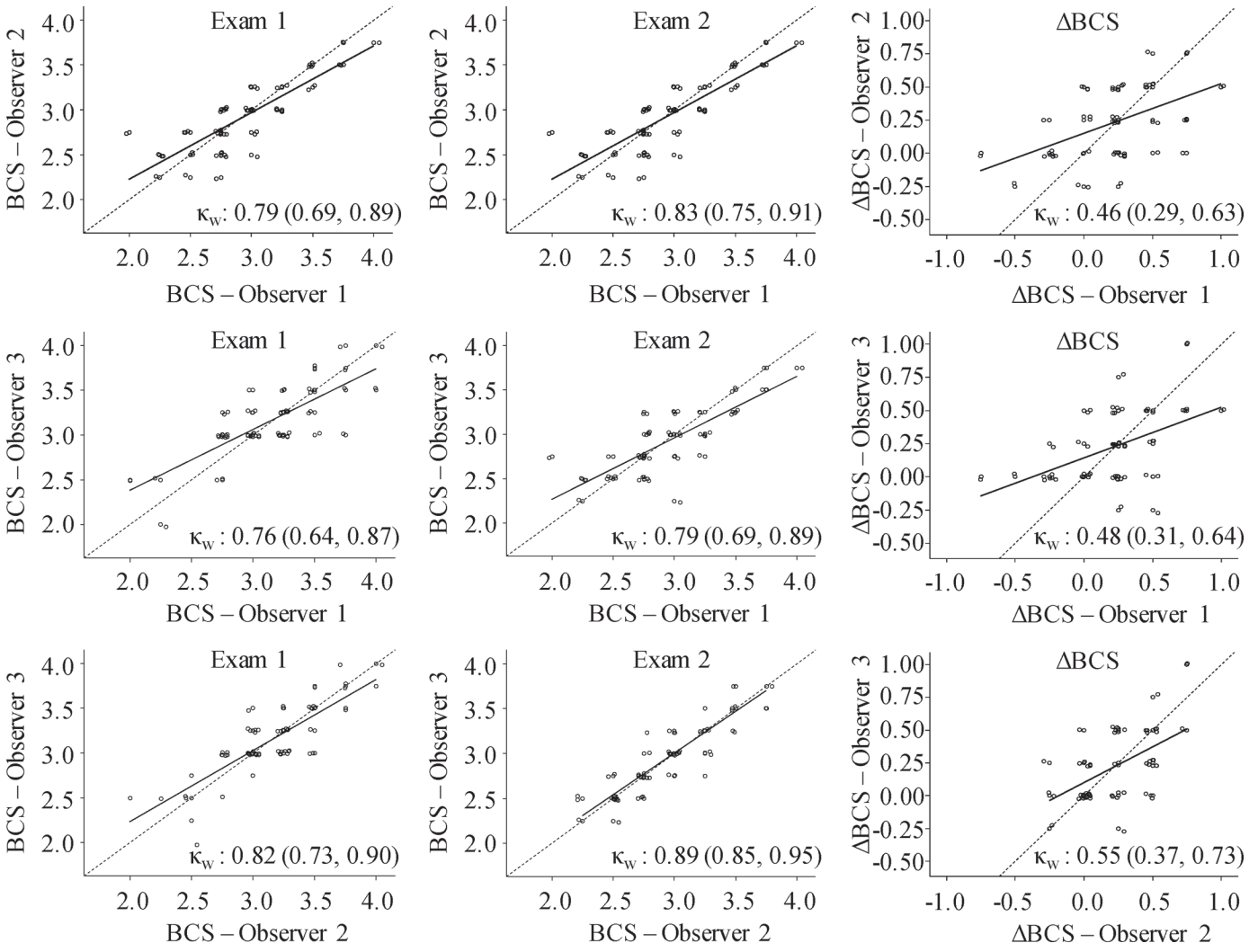

Figure 1. Scatter plots of BCS at exam 1 (1–20 DIM), exam 2 (41–60 DIM), and for difference in BCS between exam 1 and exam 2 (ΔBCS), for each pair of observers, with quadratic weighted kappa (κ_w) and its 95% CI; BCS were measured by 3 different observers on a cohort of 73 cows from 1 commercial dairy.

(exam 1: 15 and exam 2: 15) and observer 2 (exam 1: 21 and exam 2: 19), and fewer 2.75 values (exam 1: 0 and exam 2: 1) compared with observer 1 (exam 1: 13 and exam 2: 12) and observer 2 (exam 1: 7 and exam 2: 5). All the other categories had similar numbers of observations. When a low homogeneity between observers is expected, the quality of ΔBCS measurements could possibly suffer from using different observers at first and second exams. In the current study, no extreme BCS values (i.e., BCS of 1 or 5) were observed and practically all values ranged between 2.5 and 3.5. This could be explained by the good management of the farm selected for this study, which may not be representative of less well managed herds. The 3 observers who rated the cows worked together on a daily basis as part of their veterinary clinical appointments. Their BCS assessments were, therefore, possibly more homogeneous than what would be expected when comparing various independent observers. In the current study, errors between observers when conducting single BCS measurement appears to have been generated by a random, rather than a systematic process. In the authors' opinion, this observation is probably representative of the reality and systematic error in single BCS measurement between observers is unlikely to occur. Because BCS is scaled in many categories, it would be unlikely that an observer shifted all his classifications on the lean side or on the fat side. However, if this happened a new training could calibrate the observer to suppress this systematic error.

For practical reasons, and because the aim of the current study was inter-observer agreement, repeated

BCS measurements within the same observers were not collected and intra-observer agreement could not be computed. As suggested by Garcia et al. (2015), before computing inter-observer agreement, the first step in a validation process should be the assessment of the intra-observer agreement to evaluate if observers agree at least with themselves. The latter study gave an example where 2 observers using a categorical scale have intra-observer agreement of 0.80. In such case, the expected inter-observer agreement would be 0.64 (0.80 times 0.80) illustrating the fact that a low intra-observer agreement for even just one rater would inevitably lead to a lower inter-observer agreement. Although intra-observer agreement could not be estimated in the current study, the findings of Garcia et al. (2015) suggest that it would be greater than the observed inter-observer agreement value. The high inter-observer agreements observed in both exams of the present study suggest higher intra-observer agreement.

To conclude, when ΔBCS is the parameter of interest, more reliable results would be obtained if one observer conducts all assessments. When single BCS measurements are of interest, more than one observer could be used with a high degree of accuracy in the results.

ACKNOWLEDGMENTS

This research was funded by a grant to S. Dufour from the National Sciences and Engineering Research Council of Canada Discovery Grant (number RGPIN/435637-2013).

REFERENCES

Barnhart, H. X., M. Haber, and J. Song. 2002. Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 58:1020–1027.

Bowker, A. H. 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43:572–574.

Carrasco, J. L., B. R. Phillips, J. Puig-Martinez, T. S. King, and V. M. Chinchilli. 2013. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. Comput. Methods Programs Biomed. 109:293–304. https://doi.org/10.1016/j.cmpb.2012.09.002.

Cohen, J. 1983. The cost of dichotomization. Appl. Psychol. Meas. 7:249–253.

Dohoo, I. R., S. W. Martin, and H. Stryhn. 2003. Veterinary Epidemiologic Research. 1st ed. AVC Inc., Charlottetown, PEI, Canada.

Ferguson, J. D., D. T. Galligan, and N. Thomsen. 1994. Principal descriptors of body condition score in Holstein cows. J. Dairy Sci. 77:2695–2703. https://doi.org/10.3168/jds.S0022-0302(94)77212-X.

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

Garcia, E., K. Konig, B. H. Allesen-Holm, I. C. Klaas, J. M. Amigo, R. Bro, and C. Enevoldsen. 2015. Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. J. Dairy Sci. 98:4560–4571. https://doi.org/10.3168/jds.2014-9266.

Gillund, P., O. Reksen, Y. T. Grohn, and K. Karlberg. 2001. Body condition related to ketosis and reproductive performance in Norwegian dairy cows. J. Dairy Sci. 84:1390–1396. https://doi.org/10.3168/jds.S0022-0302(01)70170-1.

King, T. S., and V. M. Chinchilli. 2001. A generalized concordance correlation coefficient for continuous and categorical data. Stat. Med. 20:2131–2147. https://doi.org/10.1002/sim.845.

Kristensen, E., L. Dueholm, D. Vink, J. E. Andersen, E. B. Jakobsen, S. Illum-Nielsen, F. A. Petersen, and C. Enevoldsen. 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. J. Dairy Sci. 89:3721–3728. https://doi.org/10.3168/jds.S0022-0302(06)72413-4.

Landis, J. R., and G. G. Koch. 1977a. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33:363–374.

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

Lucy, M. C., G. A. Verkerk, B. E. Whyte, K. A. Macdonald, L. Burton, R. T. Cursons, J. R. Roche, and C. W. Holmes. 2009. Somatotropic axis components and nutrient partitioning in genetically diverse dairy cows managed under different feed allowances in a pasture system. J. Dairy Sci. 92:526–539. https://doi.org/10.3168/jds.2008-1421.

McNamara, J. P. 1991. Regulation of adipose tissue metabolism in support of lactation. J. Dairy Sci. 74:706–719. https://doi.org/10.3168/jds.S0022-0302(91)78217-9.

Pedron, O., F. Cheli, E. Senatore, D. Baroli, and R. Rizzi. 1993. Effect of body condition score at calving on performance, some blood parameters, and milk fatty acid composition in dairy cows. J. Dairy Sci. 76:2528–2535. https://doi.org/10.3168/jds.S0022-0302(93)77588-8.

Renaville, R., M. Hammadi, and D. Portetelle. 2002. Role of the somatotropic axis in the mammalian metabolism. Domest. Anim. Endocrinol. 23:351–360.

Roche, J. R., N. C. Friggens, J. K. Kay, M. W. Fisher, K. J. Stafford, and D. P. Berry. 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. J. Dairy Sci. 92:5769–5801. https://doi.org/10.3168/jds.2009-2431.

Rotondi, M. A., and A. Donner. 2012. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. J. Clin. Epidemiol. 65:778–784. https://doi.org/10.1016/j.jclinepi.2011.10.019.

Sim, J., and C. C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys. Ther. 85:257–268.

Smith, T. R., and J. P. McNamara. 1990. Regulation of bovine adipose tissue metabolism during lactation. 6. Cellularity and hormone-sensitive lipase activity as affected by genetic merit and energy intake. J. Dairy Sci. 73:772–783.

Wildman, E. E., G. M. Jones, P. E. Wagner, R. L. Boman, H. F. Troutt, and T. N. Lesch. 1982. A dairy cow body condition scoring system and its relationship to selected production characteristics. J. Dairy Sci. 65:495–497.

Wright, I. A., and A. J. F. Russel. 1984. Partition of fat, body-composition and body condition score in mature cows. Anim. Prod. 38:23–32.