Within- and Across-Person Uniformity of Body Condition Scoring in Danish Holstein Cattle

E. Kristensen,* L. Dueholm,* D. Vink,* J. E. Andersen,* E. B. Jakobsen,* S. Illum-Nielsen,* F. A. Petersen,* and C. Enevoldsen†¹

*KoNet-Praxis Aps, Solsortevej 36, DK-8640 Ans By, Denmark

†Department of Large Animal Sciences, The Royal Veterinary and Agricultural University, Grønnegaardsvej 2, DK-1870 Frederiksberg C, Copenhagen, Denmark

ABSTRACT

Body condition scores (BCS) are very useful for dairy herd management and breeding programs, but the consistency and quality of recordings made by consultants in the field are unknown. The objectives of this study were 1) to estimate the agreement in BCS within and among practicing dairy veterinarians and 2) to provide an indication of the effects of training and the value of calibration, and of what efforts need to be made to obtain a validity and precision in BCS adequate for management purposes. A total of 2,230 scores were recorded by 51 practicing dairy veterinarians and 6 highly trained instructors. The 6 instructors were cross-trained to validate calibration consistency in assigning BCS. Each individual scored approximately 20 cows twice, with the second scoring occurring approximately 2.5 h after the first. Between the 2 recordings, the respective instructors conducted a training session for the practicing veterinarians using other cows. A weighted kappa coefficient was used to assess agreement among and within classifiers. Excellent agreement ($\kappa \geq 0.86$) was documented between repeated BCS recorded for the same cows by the highly trained instructors. In addition, the BCS provided by multiple classifiers from the instructor team appeared to be comparable across herds and classifiers. This legitimizes the use of BCS for benchmarking at both the cow and the herd level. The within-classifier and between-classifier kappa values were in the ranges of 0.22 to 0.75 and 0.17 to 0.78, respectively, in the group of practicing dairy veterinarians. Many of the veterinarians provided estimates of average BCS that differed considerably from the BCS recorded by the instructors. Between-classifier comparisons of herd BCS are not warranted unless a validation has been performed. If scores are collected by multiple classifiers with varying experience, a valid but imprecise estimate of the true population mean of BCS may

be obtained if classifiers are inexperienced. The limited training effort used in this study seemed to have brought about substantial improvement in the validity and precision of the BCS determined by practicing veterinarians, compared with the BCS recorded on the same cows by highly trained classifiers.

Key words: interobserver variation, intraobserver variation, body condition score, herd health management

INTRODUCTION

The utility of BCS for dairy herd management is well documented (e.g., Markusfeld et al., 1997) and, judging subjectively from the large number of recent publications in journals and magazines, the interest in using BCS appears to be growing rapidly, especially among dairy veterinarians. Several scales for BCS are described in the literature, but studies are lacking that evaluate the extent to which each of those scales is used. However, the visual 5-point scale with 0.25-unit increments described by Ferguson et al. (1994) is apparently widely used.

Substantial interest also exists in using BCS for breeding purposes. The issue of the validity of visual or physical inspections to classify traits, including linear type traits and carcass conformations that are included in selection indices, is addressed intensively in breeding programs. In those programs, such classifications can be of very high quality because relatively few classifiers can be trained intensively to calibrate their observations. In addition, the classifiers work with several herds and inspect the offspring of numerous bulls. Consequently, it is possible to estimate and eliminate systematic classifier effects from breeding value estimates. Veerkamp et al. (2002) suggested an approach to improving the precision and accuracy based on estimates of heritability, sire and residual variance, and genetic correlations. They also discussed the problems and advantages associated with this indirect validation. Dechow et al. (2003) added other validation criteria to the method applied by Veerkamp et al. (2002). However,

Received August 17, 2005.

Accepted March 26, 2006.

¹Corresponding author: ce@kvl.dk

in an analysis of large data files, they found that the average estimates of the precision and accuracy of BCS differed very little among groups of classifiers at different skill levels, according to their grouping criteria.

For management purposes, there is a need to issue recommendations based on BCS levels at various stages of lactation. Ideally, precision should be sufficient to allow recommendations to be issued at the cow level. For example, a consultant may recommend targets for BCS at drying off or at first breeding. The use of targets based on comparisons of herds is also an important component of herd management (benchmarking). Such recommendations obviously are invalid if the BCS scale is used differently or inconsistently by farmers and consultants. Because each farmer and consultant works with a limited number of herds and cows, the validation tools applied in breeding programs are not directly applicable to the management context. Ferguson et al. (1994) conducted a validation study based on 255 cows that were examined independently, once each by 4 classifiers. The classifiers agreed with the absolute score 58% of the time, and deviated by 0.25 units 33% of the time. Their study indicated substantial variation among classifiers because of differences in experience. The study by Ferguson et al. (1994) may be the only published formal validation of BCS of dairy cows based on repeated assessments of individual cows. Consequently, there is a need to provide more estimates of the validity and precision of BCS among consultants working in the field.

The objectives of the current study were 1) to estimate the agreement between BCS within and among practicing dairy veterinarians and 2) to provide an indication of the effects of training and calibration, and of what efforts need to be made to obtain validity and precision in BCS that is adequate for management purposes.

MATERIALS AND METHODS

Data Collection

Data were recorded during 3 identical teaching workshops by 51 dairy veterinarians and 6 instructors (subsequently labeled students and instructors, respectively) in 3 well-managed dairy herds (10,000 to 11,000 kg of ECM per 365 cow-days). The workshops were at least 60 d apart. At each workshop, the students were divided into 2 groups who worked in 2 herds. Each group included 7 to 10 practicing veterinarians as students, and the numbers of workshop sessions per herd were 1, 2, or 3. Some of the workshop students had several years of experience in using BCS, whereas others had none. (Unfortunately, precise records of this factor are not available.) All 6 instructors had worked closely together in a formal network for at least 3 yr.

In the course of at least 4 meetings and numerous mail discussions regarding the BCS scale and its validity and precision, the 6 instructors had calibrated their BCS.

The measurement protocol was defined by the instructors. This was a 5-point scale with 0.25-unit increments. It was based on components from the scale of Ferguson et al. (1994) and from another 5-point scale developed by Kristensen (1986). Evaluations involved both visual inspection and palpation. As judged in meetings with veterinarians from other countries, the BCS obtained in the current work were similar to the BCS obtained from the scale of Ferguson et al. (1994). However, because the focus was not on estimating levels of BCS, the issue of comparability with other scales is of limited importance for the current study. Specific regions on the cows included the spinous and transverse processes of the lumbar vertebrae, the ileal and ischial tuberosities, and the iliosacral and ischiococcygeal ligaments.

Each of the 3 workshops was conducted as follows. Theoretical material relating to the BCS was presented during a 2-h lecture. Subsequently, the students were divided into 2 groups and, together with 2 instructors, were transported to 2 separate Danish Holstein dairies. Upon arrival, each participant, together with 1 of the instructors, scored the body condition of the same group of approximately 20 randomly selected cows. No attempts were made to balance parities because there was no evidence that assessments of BCS would differ among parities. The classifiers recorded the scores on a signed sheet of paper that was handed to the second ("chief") instructor. All cows were locked in head-gates by the dairy manager before the arrival of workshop participants. In the training area, only one classifier scored any given cow at a time to ensure that the recording activities of classifiers were independent. Recordings were supervised by the chief instructors. Afterward, students participated in a hands-on workshop on BCS lasting approximately 2.5 h, in which the measurement protocol was discussed. The workshop did not include scoring of the 20 trial cows.

Before departing, all participants repeated their BCS assessments with the same 20 cows and handed their recordings to the chief instructor. It is possible that some cows were included in more than one workshop, but such details were not available. However, because more than 60 d had elapsed between workshops, cows' BCS were expected to have changed during that time. Therefore, it seems justified to regard the cows evaluated on different occasions as independent observational units.

Statistical Analyses

Most studies of BCS have applied ANOVA or some kind of mixed model (e.g., Evans, 1978; Nicholson and

Sayers, 1987; Audigé et al., 1998; Veerkamp et al., 2002; Dechow et al., 2003). However, the assumptions for using mixed models or ANOVA appear to have been violated either because the BCS was recorded with discrete values, heterogeneity of variance among classifiers could be expected (Veerkamp et al., 2002), or the numbers of observations in each subgroup of that study were rather small. For the same reasons, significance tests have not been used except for aggregating kappa values. The Bland–Altman test to explore agreement between quantitative variables (described by Ersbøll et al., 2004) also has been omitted for these reasons. Standard deviations have been calculated in several instances to allow readers to compare current findings with other studies in which standard deviations were used to describe the data. Although the ideal tool to study this research question may be Bayesian threshold modeling (Baadsgaard and Jørgensen, 2003), the number of measurements within each stratum appears to have been too scarce to apply this method. Finally, the correlation coefficients derived from mixed models (repeatability estimates) have been measures of linear associations, rather than agreement (Ersbøll et al., 2004).

In this study, PROC UNIVARIATE was used for descriptive analyses. Because BCS is an ordinal score, the weighted kappa coefficient (Ersbøll et al., 2004; PROC FREQ, pp. 227–230, SAS Institute, 1996) was used to assess agreement among classifiers that exceeded agreement by chance. Values 1 and 0 represent perfect agreement and no agreement, respectively. Values of 0.4 or above are arbitrarily taken to indicate at least moderate agreement, and values above 0.8 are assumed to indicate excellent agreement (suggested thresholds as summarized by Ersbøll et al., 2004). The weights were based on numerical BCS values (1 to 5). A sensitivity analysis was done to estimate the effect of changing the weights. However, even substantial changes in both directions had only marginal effects. The default Cicchetti–Alison weight type was applied. In PROC FREQ, an alternative Fleiss–Cohen weight type can be applied. This type generally yielded higher kappa values (typically a 0.2-unit increase) but did not change the conclusions. Where the available test of equality of strata was statistically nonsignificant, the kappa values of the individual classifiers could be collapsed into an overall kappa value (pp. 227–230, SAS Institute, 1996).

In addition to being a generally accepted tool to assess agreement, the kappa index was chosen for this study for the following reasons: First, the analytical setup with cross-tabulation is intuitively simple for the end users. Second, end users can easily calculate the index in their own context and compare their results with our findings. Consequently, the current study supplied useful information for benchmarking in the field.

The terms “validity” and “precision” are used to define the quality of the recordings. The validity of a measurement tool is defined as the average ability of the tool to provide the correct estimate of the condition of interest. Precision is defined as the degree of repeatability (or reproducibility) of the measurements on the same observational units.

RESULTS AND DISCUSSION

BCS Data and Distribution of BCS

The mean, standard deviation, range, lower quartile, and upper quartile of BCS across classifiers and herds were 3.17, 0.48, 2.25 to 4.50, 2.75, and 3.50, respectively. The range of BCS in the current study was slightly narrower than the 1.5 to 4.5 range in the study by Ferguson et al. (1994). However, Ferguson et al. did express reservations toward using the scale below 2.50 units. It was not an objective of this study to estimate the distribution of BCS in Danish Holstein cows. Nor is it certain how the 5-point scale used in the current study relates to other BCS described in the literature. However, based on meetings with fellow consultants in Denmark and other countries, the 5-point scales applied seemed to be similar. There is a genuine need for tools that can be used to calibrate body condition scoring protocols both nationally and internationally, as demonstrated by Roche et al. (2004).

The Kolmogorov–Smirnov tests for normality (Ersbøll et al., 2004) of the crude BCS showed that the null hypothesis should be rejected ($P < 0.01$). Deviation from the normal distribution was the result, mainly because there were too few observations in the lower range. This indicates that a one-sample *t*-test to compare a sample of BCS with a given target (which is relevant for benchmarking) may not be valid.

Means and standard deviations of the recordings in the 2 classifier groups, within herd and occasion, were numerically very similar on both the first and second body condition scoring. Table 1 shows that the distribution of the means and standard deviations of the BCS within classifiers and occasions in the student and instructor groups were numerically very similar. Within classifier groups, the variation was due to herd, cow, and classifier. The standard deviations of all classifier means were 0.08 and 0.07 BCS units on the first and second scoring, respectively (data not shown).

Table 2 shows the distribution of average changes in BCS from the first to second scoring, within classifiers stratified by students and instructors. Among the instructors, the range was 0.05 BCS units, whereas it was 0.36 among the students. This shows that the instructors were able to score the cows much more consistently than the students. Table 2 also shows the distri-

Table 1. Distributions of means and standard deviations of BCS estimated within classifiers and occasions within cows

Classifier	n	Variable	Mean	SD	Minimum	Percentile ¹		Maximum
						25	75	
Student	51	Avg. BCS 1 ²	3.17	0.08	3.00	3.13	3.23	3.35
		SD BCS 1 ³	0.50	0.06	0.33	0.46	0.55	0.66
		Avg. BCS 2	3.17	0.07	3.05	3.11	3.23	3.32
		SD BCS 2	0.48	0.06	0.37	0.44	0.51	0.65
Instructor	6	Avg. BCS 1	3.16	0.08	3.03	3.13	3.21	3.26
		SD BCS 1	0.47	0.06	0.38	0.41	0.51	0.53
		Avg. BCS 2	3.14	0.09	3.01	3.09	3.20	3.25
		SD BCS 2	0.45	0.05	0.37	0.42	0.48	0.53

¹Twenty-fifth and 75th percentile, respectively.

²Average BCS on the first and second scoring, respectively. The first and second scorings occurred approximately 2.5 h apart.

³Standard deviation of BCS on the first and second scoring, respectively.

bution of the average differences between BCS recorded by students and instructors, calculated within classifier and within occasion. Each of the 51 students was compared with just 1 instructor within herd and training session. The range of averages of the differences between students and instructors was approximately 0.2 BCS units. Because the means of the BCS estimated by instructors were quite precise (Table 2) and the mean difference between instructors and students was very close to zero, it seems justified to expect that the mean BCS estimated from records obtained by a larger group of classifiers is a valid estimate of the population average, even if those classifiers have only theoretical knowledge about the applied scale, as did some of the students in this study. Consequently, if a valid estimate of the population average is needed and exact knowledge about the quality of available data is limited, it is probably preferable to use BCS from multiple classifiers. Thomsen et al. (2002) reported a similar conclusion from the field of human orthopedics. A comparison of the veterinarians in this study with medical doctors may be valid because the clinical training builds on general principles. However, the means derived for in-

dividual veterinarians may differ considerably from the true mean. Consequently, the means from individual veterinarians or similar classifiers are not valid for benchmarking without validation.

Agreement Among and Within Classifiers

The instructor classifiers' scores on the first and second occasion agreed 72 to 95% of the time (data not shown). Across classifiers, the agreement was 83%, compared with the 58% agreement among classifiers found by Ferguson et al. (1994). Only 1 of the 117 instructor classifications differed by 0.5 units (data not shown). Overall, the student classifiers' scores on the first and second occasion agreed 30% of the time, but 21% of the scores differed by 0.5 units or more (data not shown).

Table 3 shows distributions of weighted kappa values calculated within classifiers and stratified by students and instructors. The values were, on average, 0.36 units higher among instructors than among students. The standard errors of the weighted kappa values (data not shown) were about twice as high among students as

Table 2. Distribution of average BCS changes within classifiers, and differences between instructor (I) and student (S) classifiers on the first and second scoring

Variable	n	Mean	SD	Minimum	Percentile ¹		Maximum
					25	75	
Change ² I	6	-0.02	0.02	-0.04	-0.04	-0.01	0.01
Change S	51	0.00	0.06	-0.15	-0.04	0.05	0.21
Diff. ³ 1	51	0.00	0.05	-0.10	-0.04	0.04	0.11
Diff. 2	51	0.02	0.05	-0.08	-0.02	0.05	0.12

¹Twenty-fifth and 75th percentile, respectively.

²Averages of changes in BCS from the first to second scoring among instructor (I) and student (S) classifiers, respectively.

³Respective differences in BCS between student and instructor classifiers on the first and second scoring, measured 2.5 h apart.

Table 3. Distribution of weighted kappa values within classifiers, in student-student pairs, and in student-instructor pairs

Classifier	Observed units	n	Mean	SD	Minimum	Percentile ¹		Maximum
						25	75	
Within students	BCS 1 vs. BCS 2	51	0.55	0.11	0.22	0.50	0.63	0.75
Within instructors	BCS 1 vs. BCS 2	6	0.91	0.04	0.86	0.88	0.92	0.98
Student-student pairs	BCS 1	197	0.50	0.12	0.17	0.41	0.57	0.78
	BCS 2	197	0.64	0.09	0.41	0.58	0.70	0.82
Student-instructor pairs	BCS 1	51	0.62	0.10	0.33	0.57	0.68	0.84
	BCS 2	51	0.74	0.07	0.55	0.69	0.79	0.85

¹Twenty-fifth and 75th percentile, respectively.

among instructors. This might be explained as a training effect. That is, the students who differed systematically from the instructors would tend to have changed toward the instructors' level. Nonequality of strata in the group of instructor classifiers was statistically nonsignificant ($P = 0.16$). Consequently, kappa values of the individual instructors were collapsed into an overall kappa value that emerged as 0.94 (95% confidence interval: 0.91 to 0.97). None of the instructors (within-instructor kappa) had values below 0.86. The strata in the group of student classifiers were not equal ($P = 0.01$).

Table 3 also shows distributions of weighted kappa values calculated for all the pairs of students who scored the same cows on the same day ($n = 197$). The values were, on average, 0.14 units higher on the second BCS compared with the first BCS. The minimum value increased from 0.17 to 0.41, whereas the maximum values were virtually the same on the first and second BCS. The strata in the group of student pairs were not equal on the first BCS ($P < 0.01$), but nonequality of strata was statistically nonsignificant on the second BCS ($P = 0.29$). That made it valid to estimate the pooled kappa value from the second BCS at 0.67 (95% confidence interval: 0.65 to 0.68). Consequently, the students' scores were more homogeneous on the second than on the first BCS. Mean kappa values from the 51 student-instructor comparisons increased from 0.62 on the first BCS to 0.74 on the second BCS. The corresponding minimum values increased from 0.33 to 0.55, whereas the maximum values largely remained the same. On the first BCS, the strata were not equal ($P = 0.01$). On the second BCS, nonequality of strata was statistically nonsignificant ($P = 0.78$). It was therefore possible to estimate the pooled kappa value from the second BCS at 0.76 (95% confidence interval: 0.74 to 0.78). In one herd, the students obtained consistently poorer kappa values on the first occasion (approximately 0.10 to 0.20 units lower than the other student groups), probably reflecting a less experienced group of students. In summary, the current findings showed that BCS allocated

by students became more homogeneous and more similar to those recorded by the instructors from the first to second scoring.

It may be relevant or practical simply to identify fat or thin cows. To judge the effectiveness of BCS as a tool for identifying fat and thin cow groups, instructor scores <3.00 and >3.75 were dichotomized. In a comparison of the first and second BCS, only 1 score out of 117 scores was misclassified. That is, extreme values in the BCS do not appear to be more difficult to distinguish.

Kappa values greater than 0.80 usually are regarded as indicative of excellent agreement in clinical observations. Kappa has widely recognized limitations as an indicator of agreement (Ersbøll et al., 2004). It is sensitive to the choice of scale; therefore, comparison of agreement studies using different scales with different numbers of categories can be problematic. The prevalence of conditions in each category may also affect kappa values. Classifiers used the same number of categories in all comparisons. The distribution of BCS was very similar in all 6 largely independent trials in the current study (specific strata not shown), and estimates within classifier were based on exactly the same cows, with only a 2.5-h interval between scorings. For these reasons, kappa values within the current study are comparable. However, comparison with other studies can be problematic because the scales may be different. The substantial effect of weight type just described also shows that such comparison is problematic unless all details about the estimation methods are presented.

Apart from reports of studies of conformation scores for breeding programs, which have used correlation coefficients from mixed models as repeatability estimates, no studies have apparently been published on the variability of scores obtained within and among classifiers, for example, by a large number of veterinarians or other consultants doing multiple BCS on the same cows in the field. Calavas et al. (1998) conducted a study with 13 classifiers who evaluated BCS twice on 48 ewes. They used a 0- to 5-point scale with 0.25-unit increments. Within-classifier kappa values were 0.44 to 1.00

and between-classifier kappa values were 0.03 to 0.58. Audigé et al. (1998) and Van Steenberghe (1989) conducted within-classifier studies with deer and pigs, respectively (using correlation coefficients). They also found that within-classifier correlations were higher than those between classifiers. Those findings also make sense intuitively.

Numerous small-scale studies of agreement are no doubt performed during the development of measuring protocols for research projects in various species (e.g., Manske, 2002; Petersen et al., 2004), but they are rarely reported separately. In addition, numerous studies have been done on the agreement of clinical assessments from other disciplines, especially human medicine. The research of Al-Shahi et al. (2002), Dixon and Johnston (1972), Klinkhoff et al. (1988), Molyneux et al. (1999), Thomsen et al. (2002), and Wallace et al. (2001) are examples of work in which study designs were similar to those in the current study. Most studies seem to apply binary scores, which make comparisons with the current findings very difficult. However, it is rare to find kappa values as high as those reported here. In the original description of the development of their visual BCS, Ferguson et al. (1994) did not apply the kappa coefficient to assess agreement. Their description of the proportion of observations with disagreement may be misleading because it does not take into account agreement by chance. However, because the scale and distribution of values from their study are very similar to those in the current study, a comparison of the total proportion of agreement is appropriate, as shown earlier in this section.

The weighted kappa values of the student classifiers ranged from very poor (0.17) to good (0.78). However, it is important to note that kappa values may be very high even if the scale is used incorrectly (i.e., there may be poor validity and high precision). In such a situation, it would obviously be easy to adjust the recordings if the difference between the true condition and the actual measurement were known. The tools described by Dechow et al. (2003) and Veerkamp et al. (2002) could also be used to evaluate and adjust such data for breeding purposes. However, for management purposes, efforts to secure validity and precision in line with those of the instructors in this study do appear to be more worthwhile, because that would allow the scores to be used for decision making at the cow level (e.g., in decisions about the length of dry periods or the start of insemination). If the quality of data improved to this level, the value of such data for breeding purposes would also increase. Providing training, as described for the current study, gave an indication of the effort needed to achieve a substantial reduction of intra- and interclassifier variability (with increases in mean kappa values

up to approximately 0.20 units). The studies by Dixon and Johnston (1972) and Klinkhoff et al. (1988) indicate that the majority of the learning effect comes from the first training session.

Potential Sources of Error

Some potential sources of error should be mentioned. There appears to be a general reluctance for classifiers to use the end points of a scale, especially when they are not experienced. In contrast with what was expected from measurements with poor precision, such behavior may explain why the variance in BCS recorded by the students was similar to the variance of BCS by the instructors. This unanticipated finding is consistent with earlier findings (see Veerkamp et al., 2002) showing that a comparison of the means and standard deviations of classifiers' scores does not always identify poor precision. Kappa values within students probably were underestimated as a result of the fact that the hands-on training between the first and second BCS changed the use of the BCS scale for some students (whereas the BCS scale was virtually constant among the instructors). Overestimation also may have occurred because the possibility that some classifiers were able to remember some cows from the first to second BCS cannot be excluded. However, such carryover effects are most likely to have occurred in connection with cows with extreme scores, and such cows were rare in the population studied. In addition, if students did remember the scores of some cows, it is highly unlikely that different students would remember the same cows. In the event that they remembered different cows, the kappa values of student pairs on the second occasion should not be higher than on the first occasion. In the current study, the kappa values of students were, on average, 0.14 units higher on the second than on the first BCS despite the risk of underestimation mentioned. These results support the interpretation that the limited training effort applied in this study did cause a substantial improvement in both the validity and precision of the BCS among students.

The Study Context

It would not be justified to claim that the group of workshop students represented a random sample of dairy veterinarians in Denmark. However, because of their willingness to pay a considerable course fee, spend several days in the course, and prepare course work, all students were veterinarians with a strong interest in advanced dairy herd management. Informal information about the students also showed that they had variable BCS experience, ranging from none to several

years of routine BCS recording. The population of student classifiers therefore represented a broad spectrum of practicing dairy veterinarians serving a broad spectrum of dairy herds—that is, as far as developing dairy management for the future was concerned, they were the target group. In Denmark, those veterinarians typically work with 100- to 500-cow herds. Consequently, they would be able to collect BCS of high quality from a considerable number of cows. As demonstrated by Dechow et al. (2003), such data might be of value for breeding programs. The group of instructors, and the efforts they made in connection with the development of a measurement protocol, also represented a practical and feasible organizational achievement among practicing consultants.

The Analytical Methods

A mixed-model approach to the analysis of BCS would undoubtedly have been much more efficient than the current approach if model assumptions held and the derived repeatability estimates were valid and were precise estimates of agreement. Direct comparisons of kappa values and repeatability estimates from mixed models have not been published regarding BCS, but such a comparison would be of interest.

Different agreement might exist if the scale values for body condition were more extreme than those observed in the current study. Estimates of agreement would likely be better if extreme values were included. Bayesian threshold modeling (Baadsgaard and Jørgensen, 2003) is a method that estimates the “true state,” given 2 independent and imperfect observations. This approach probably would be an efficient tool to examine the effects of selecting a series of threshold values. That is, some BCS values might be more difficult to distinguish than others, or some classifiers might interpret and use the scale differently. If that were the case, training sessions might be further refined to address those issues. Comparison of model fit among competing models should be incorporated, which was not possible with the kappa approach. However, preliminary evaluations indicate that the number of observations per strata was too low to apply this model type to the current data.

Perspectives

The high degree of precision and validity in the BCS assigned by instructors justifies the use of their BCS as a gold standard for evaluating the BCS assigned by students during training sessions. It also showed that, as long as the classifiers were trained sufficiently, individual cows and herds could be compared (or ranked)

across classifiers by means of this type of score (benchmarking). That is, a single classifier would be able to communicate results pertaining to a particular herd to another classifier working with another herd. This is an important finding, because this opportunity ought to reduce, or even eliminate, the need for statistical adjustment of estimates of the kind that would require veterinarians to conduct body condition scoring in the same herds.

The study by Houe et al. (2002) indicated that it was easier to obtain good agreement for pathological conditions than for scores such as the BCS. Consequently, given the finding that excellent validity and precision of BCS can be obtained, a similar quality of records related to other clinical recordings might be expected.

Currently, new reasons for insisting on systematic clinical observations and data collection are emerging. Public authorities, milk processing plants, and farmers' organizations are increasingly interested in some of the quality assurance programs (e.g., certification) and the effectiveness of the tools used to detect illegality. For example, in Denmark there is a law against shipping a cow for slaughter if she is in the last tenth of her pregnancy, and there have been suggestions that a new law regulating the use of antibiotics should introduce a requirement for systematic clinical assessments. Legal restraints are not meaningful if such clinical assessments are not highly standardized. This study demonstrates that a high degree of standardization can be obtained with training.

Skepticism about the use of clinical observations in management and quality assurance seems to be widespread. This skepticism is reflected in the frequent use of the term “subjective” to characterize clinical observations as a tool to estimate the given state of an individual (e.g., Veerkamp et al., 2002). The term subjective seems to be perceived as an assessment that is unique to the individual classifier. To generally categorize clinical assessments as subjective is objectionable because clinical sciences indicate that it is possible to make such assessments comparable across persons. Logically, if that is possible, an objective measurement is obtainable. However, the aforementioned examples of potential applications justify efforts to develop well-defined scores of important traits and to document their quality.

CONCLUSIONS

There was excellent agreement ($\text{kappa} \geq 0.86$) between repeated BCS recorded on separate occasions for the same cows by a team of experienced classifiers who had worked systematically to develop and harmonize a measurement protocol for more than 3 yr. In addition,

the BCS provided by multiple classifiers from this team were comparable across herds. This legitimizes the use of BCS for benchmarking both at the cow and the herd level if efficient training has been received.

The within-classifier and between-classifier kappa values were in the ranges of 0.22 to 0.75 and 0.17 to 0.78, respectively, in a group of practicing dairy veterinarians. Many of the veterinarians obtained estimates of average BCS that differed considerably from the BCS assigned by experienced classifiers. Between-classifier comparisons of herd BCS are not warranted unless validation has been performed. If scores are collected by multiple classifiers, a valid but imprecise estimate of the true population mean of BCS may be obtained, even if the classifiers are inexperienced. Substantial improvements in validity and precision were brought about by limited training.

ACKNOWLEDGMENTS

The very constructive and encouraging comments and suggestions from 3 anonymous reviewers and the editor are greatly appreciated.

REFERENCES

- Al-Shahi, R., N. Pal, S. C. Lewis, J. J. Bhattacharya, R. J. Sellar, and C. P. Warlow. 2002. Observer agreement in the angiographic assessment of arteriovenous malformations of the brain. *Stroke* 33:1501–1508.
- Audigé, L., P. R. Wilson, and R. S. Morris. 1998. A body condition score system and its use for farmed red deer hinds. *N. Z. J. Agric. Res.* 41:545–553.
- Baadsgaard, N. P., and E. Jørgensen. 2003. A Bayesian approach to the accuracy of clinical observations. *Prev. Vet. Med.* 59:189–206.
- Calavas, D., P. Sulpice, E. Lepetitcolin, and F. Bugnard. 1998. Appréciation de la fidélité de la pratique d'une méthode de notation de l'état corporel des brebis dans un cadre professionnel [Assessing the accuracy of the practice of a method of scoring the body condition of ewes within a professional framework]. *Vet. Res.* 29:129–138. In French.
- Dechow, C. D., G. W. Rogers, L. Klei, and T. J. Lawlor. 2003. Heritabilities and correlations among body condition score: Dairy form and selected linear type traits. *J. Dairy Sci.* 86:2236–2242.
- Dixon, R. A., and S. M. Johnston. 1972. Sources of variation in clinical observations: Problems of teaching and some results. *Meth. Inform. Med.* 11:177–182.
- Ersbøll, A. K., J. Bruun, and N. Toft. 2004. Data analysis. Chap. 13 in *Introduction to Veterinary Epidemiology*. H. Houe, A. K. Ersbøll, and N. Toft, ed. Biofolia, Frederiksberg, Denmark.
- Evans, D. G. 1978. The interpretation and analysis of subjective body condition scores. *Anim. Prod.* 26:119–125.
- 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.
- Houe, H., M. Vaarst, and C. Enevoldsen. 2002. Clinical parameters for assessment of udder health in Danish dairy herds. *Acta Vet. Scand.* 43:173–184.
- Klinkhoff, A. V., N. Bellamy, C. Bombardier, S. Carette, A. Chalmers, J. M. Esdaile, C. Goldschmidt, P. Tugwell, H. A. Smythe, and W. W. Buchanan. 1988. An experiment in reducing interobserver variability of the examination of joint tenderness. *J. Rheumatol.* 15:492–494.
- Kristensen, T. 1986. Method for estimation of body condition in dairy cows. Page 59 in *Research in cattle production systems*. Rep. No. 615. National Institute of Animal Science, Copenhagen, Denmark. In Danish with English subtitles and summary.
- Manske, T. 2002. Hoof Lesions and Lameness in Swedish Dairy Cattle. Doctoral thesis. Department of Animal Environment and Health, Swedish University of Agricultural Sciences, Skara, Sweden. Available: http://diss-epsilon.slu.se/archive/00000081/01/Ram_Manske.pdf Accessed Oct. 17, 2002.
- Markusfeld, O., N. Galon, and E. Ezra. 1997. Body condition score, health, yield, and fertility in dairy cows. *Vet. Rec.* 141:67–72.
- Molyneux, P. D., D. H. Miller, M. Filippi, T. A. Yousry, E. W. Radu, H. J. Ader, and F. Barkhof. 1999. Visual analysis of serial T2-weighted MRI in multiple sclerosis: Intra- and interobserver reproducibility. *Neuroradiology* 41:882–888.
- Nicholson, M. J., and A. R. Sayers. 1987. Repeatability, reproducibility and sequential use of condition scoring of *Bos indicus* cattle. *Trop. Anim. Health Prod.* 19:127–135.
- Petersen, H. H., C. Enøe, and E. O. Nielsen. 2004. Observer agreement on pen level prevalence of clinical signs in finishing pigs. *Prev. Vet. Med.* 64:147–156.
- Roche, H. R., P. G. Dillon, C. R. Stockdale, L. H. Baumgard, and M. J. VanBaale. 2004. Relationships among international body condition scoring systems. *J. Dairy Sci.* 87:3076–3079.
- SAS Institute. 1996. SAS/STAT Software: Changes and Enhancements Through Release 6.11. SAS Institute Inc., Cary, NC.
- Thomsen, N. O., L. O. Olsen, and S. T. Nielsen. 2002. Kappa statistics in the assessment of observer variation: The significance of multiple observers classifying ankle fractures. *J. Orthop. Sci.* 7:163–166.
- Veerkamp, R. F., C. L. M. Gerritsen, E. P. C. Koenen, A. Hamoen, and G. De Jong. 2002. Evaluation of classifiers that score linear type traits and body condition score using common sires. *J. Dairy Sci.* 85:976–983.
- Van Steenberghe, E. J. 1989. Description and evaluation of a linear scoring system for exterior traits in pigs. *Livest. Prod. Sci.* 23:163–181.
- Wallace, M. B., R. H. Hawes, V. Durkalski, A. Chak, S. Mallery, M. F. Catalano, M. J. Wiersema, M. S. Bhutani, D. Ciaccia, M. L. Kochman, F. G. Gress, A. Van Helse, and B. J. Hoffman. 2001. The reliability of EUS for the diagnosis of chronic pancreatitis: Interobserver agreement among experienced endosonographers. *Gastrointest. Endosc.* 53:294–299.