The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK

Animal Welfare 2009, 18: 451-458
ISSN 0962-7286

On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach

U Knierim*† and C Winckler‡

† Department of Farm Animal Behaviour and Husbandry, Faculty of Organic Agricultural Science, University of Kassel, Nordbahnhofstr 1a, 37213 Witzenhausen, Germany
‡ Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Applied Life Sciences, Gregor-Mendel-Str 33, 1180 Vienna, Austria
* Contact for correspondence and requests for reprints: knierim@wiz.uni-kassel.de

Abstract

This paper discusses the current state of development of on-farm cattle welfare assessment systems with special regard to the approach of Welfare Quality® that focuses on animal-related measures. The central criteria validity, reliability and feasibility are considered with regard to selected welfare measures. All welfare measures incorporated into the Welfare Quality® protocol possess face validity, but for most of them construct or criterion validity as, eg shown for lameness, have not been demonstrated. Exemplarily the cases of qualitative behaviour assessment and measurement of avoidance distance towards humans or social licking are discussed. Reliability issues have often been neglected in the past and require more thorough investigation and discussion in the future, especially with respect to appropriate test statistics and limits of acceptability. Means of improving reliability are the refinement of definitions or recording methods and training. Consistency of results over time requires further attention, especially if farms are to be certified, based on infrequent recordings. Considering feasibility, time constraints are the main concern for assessment systems that focus seriously on animal-based measures; currently they require several hours of on-farm recordings, eg about 6 h for a herd of 60 dairy cows. The Welfare Quality® project has promoted knowledge and discussion about validity, reliability and feasibility issues. Many welfare measures applied in the Welfare Quality® on-farm assessment approach can be regarded sufficiently valid, reliable and feasible. However, there are still a considerable number of challenges. They should be tackled while using the present assessment system in order to constantly improve it.

Keywords: *animal welfare, cattle, certification, reliability, validity, welfare assessment*

Introduction

Assessment of animal welfare levels on farms is an ongoing challenge for animal welfare scientists, and already a large body of literature is available on different approaches to such assessments (eg Sørensen *et al* 2001; Bartussek 2001; Bracke *et al* 2001; Botreau *et al* 2007; Main *et al* 2007). Additionally, there are several critical reviews of the different approaches available (eg Johnsen *et al* 2001; Spoolder *et al* 2003; Knierim *et al* 2004). Publications were especially stimulated by the International Workshop on the Assessment of Animal Welfare at Farm or Group Level which has now taken place four times (including this edition) (Sørensen & Sandøe 2001; Webster & Main 2003; Winckler *et al* 2007a). However, work on welfare assessment, on-farm, is not published exclusively in these special issues and began earlier than 2001. But this earlier work (as with much of the current literature) on welfare assessment often concentrated on experimental approaches rather than on-farm applications (eg Broom 1991; Rushen & de Passillé 1992; Dawkins 2004).

In brief, a number of the major contentious issues with regard to on-farm assessment relate to the adequate selection of welfare measures and of appropriate methods of measurement, how to aggregate the different measures into an overall assessment and how to cope with the practical constraints of implementation. Regarding the selection of measures, one important distinction is whether they are resource- or management-based on the one hand (also called design criteria [Rushen & de Passillé 1992] or influencing factors [Waiblinger *et al* 2001]) or animal-based on the other hand (also called performance criteria [Rushen & de Passillé 1992] or welfare indicators [Waiblinger *et al* 2001]). Different welfare assessment systems use these categories of measures to largely varying degrees with profound effects on validity, reliability and feasibility of these systems.

Welfare Quality®, a European research project on integration of animal welfare in the food quality chain, began in 2004 with a total of 44 participating institutions in 17 countries. One of its central aims is the development of

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

on-farm welfare assessment systems that focus on animal-based measures and are scientifically sound as well as feasible (Blokhuis *et al* 2003). In terms of feasibility, it should be possible for a single observer to carry out a farm assessment during a one-day visit. The Welfare Quality® assessment will allow welfare-specific product information, thus becoming a means for farm welfare certification. At the same time, it may also facilitate advice on potential improvements. However, real-time monitoring of the state of animal welfare is outwith the scope of this assessment approach.

This paper has the objective of utilising four years of participation in this field, with regard to cattle, in order to discuss the current state of development of on-farm cattle welfare assessment systems with special reference to the Welfare Quality® approach. It is outwith the scope of our review to tackle questions pertaining to the overall design of assessment systems (see eg Botreau *et al* 2007) or aggregation of results (see, eg Botreau *et al* 2009). Instead, we will concentrate on a discussion of the three central criteria: validity, reliability and feasibility with regard to exemplary cattle welfare measures. While these criteria are inextricably linked with the methodology of measurement, space constraints make it impossible to become overly specific in this regard. The choice of examples was guided by our experience with certain measures and their suitability in illustrating critical points that we regard of importance to the ongoing discussion regarding on-farm welfare assessment.

Validity

The main concern regarding welfare assessment is the extent to which we are actually measuring what we are supposed to be measuring. At the single welfare measure level, the major criticism of resource- and management-based measures is that their validity is potentially low due to their indirect nature and complex interaction with other resource and management conditions as well as the animal itself, leading to largely unpredictable outcomes (Waiblinger *et al* 2001). Animal-based measures have a theoretical advantage since they reflect directly how the animal is faring, although reliability problems (see below) may limit this advantage to a certain extent.

All the welfare measures incorporated into the Welfare Quality® protocol possess face validity (Winckler *et al* 2003), ie they are thought to be valid as judged by experts (Scott *et al* 2001; Whay *et al* 2003). However, the quantification of the welfare level is less obvious. For instance, it is an open question, with regard to lesions, the extent to which callosities, slight swellings or open wounds impair welfare. The solution to this general problem adopted in the Welfare Quality® project, is to use expert judgements (Botreau *et al* 2009). However, it would be extremely helpful to gain a greater insight into this issue from future studies.

In comparison with attested face validity, construct and criterion validity provide greater credibility. Construct validity describes the scenario whereby an expected relationship between welfare and a measure has been confirmed experimentally while criterion validity is based on the relationship of the measure in question to another welfare-relevant measure (Scott *et al* 2001). For instance, for lameness scoring, Rushen *et al* (2007) provided additional construct validation. They demonstrated, through the use of local anaesthesia that, in most cases, pain tends to be the cause of gait aberrations. Moreover, even in instances in which pain may not play a role, impaired mobility, which restricts the ability to reach resources or to cope with agonistic encounters will impact on welfare. For example, Borderas *et al* (2008) found the frequency of visits of dairy cows to an automatic milking system was related to their locomotory ability (indicating criterion validity). However, for most other measures, construct or criterion validity has yet to be demonstrated. In the following, we will give three examples in which doubts regarding validity as on-farm welfare measures have been or may be expressed.

Qualitative behaviour assessment (Wemelsfelder & Lawrence 2001) is a relatively new method, challenging conventional theory of science and which, in turn, may be criticised for being anthropomorphic judgements of uncertain validity (Wemelsfelder *et al* 2001). Until now, validation work has mainly been done in pigs for the original free-choice profiling approach, ie where descriptive terms of the mood of an animal are freely generated by a number of observers (Wemelsfelder *et al* 2001, 2009). However, in the Welfare Quality® assessment protocol, a list with fixed terms is used in order to allow for a more standardised procedure. Currently, only inter-observer reliability has been tested for this approach (Wemelsfelder *et al* 2008). Data generated in the Welfare Quality® project should be used to examine associations of the ascribed mood of the animals in the assessed groups with quantitative welfare measures. Although it is unrealistic to expect every single other welfare measure to correspond to the animals' mood, at least certain health or behavioural impairments can be expected to be associated with negative emotional states.

Avoidance distance towards an experimenter, indicating the quality of the human-animal relationship, is another measure that is very critically discussed with respect to validity for the on-farm welfare assessment (de Passillé & Rushen 2005). A considerable number of different critical points are raised that basically relate to the method of measurement and the interpretation of the results. To name only a few, in terms of methodology, de Passillé and Rushen (2005) point at the apparent sensitivity of the measure to changes in supposedly minor parameters, such as clothing of the assessor or exact location of the test that reduce validity. Concerning the interpretation of results, they argue, for example, that the predictive value of the measure for responses of cattle to humans in other contexts (eg during milking) found in different investigations was moderate at best. At the same time, the size of demonstrated significant treatment effects was often relatively low, leading de Passillé and Rushen (2005) to ask, whether the measure really allows differentiation between farms with and without welfare problems in the area of human-animal relationship. While the aspect of differentiation between farms can be further explored using data from the Welfare

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

Quality® project, there are other questions which deserve further investigation experimentally.

Social licking is often mentioned as a potential indicator of positive feelings (Knierim *et al* 2001; Winckler *et al* 2003; Boissy *et al* 2007), but in our view validity questions are yet to be adequately addressed for this measure. Therefore, they shall be discussed in greater detail. Cows receiving licks frequently display behavioural signs of enjoyment, such as partly closing their eyes. Based on this observation, Sato *et al* (1991) postulate a calming effect of social licking. Physiological measurements, namely of heart rate of animals receiving grooming, confirmed this hypothesis in primates (Boccia *et al* 1989; Aureli *et al* 1999) and cows (Sato & Tarumizu 1993). However, in all cases, sample size was very small and statistical analysis of the data questionable, eg treating dependent data as if they were independent. Moreover, in order to use social licking as an on-farm welfare measure, the underlying hypothesis would be that cattle in herds with higher social licking levels are feeling better than those in herds with lower social licking levels. However, there are indications that this is not necessarily the case. For instance, a number of authors (Reinhardt 1980; Sato *et al* 1991; Waiblinger *et al* 2002) propose that social licking might serve to reduce tensions. Currently, empirical evidence for this hypothesis is scarce. Waiblinger *et al* (2002), in a group of 19 beef suckler cows subjected to a short-term reduction of feeding places, did not observe more frequent social licking during times of increased competition, but did so the following day, which they interpreted as appeasement activities after a situation of increased tension. Emmerig (2004), in a cross-over experiment with two groups of approximately 20 horned dairy cows, found that different conditions of longer-term space restriction resulted in more licking bouts, more total time spent licking and more different cows involved in social licking. However, this was a small pilot study with only one replication and no confirmatory statistical analysis. Nevertheless, this potentially means that higher licking levels in a herd might be an indirect reflection of greater levels of conflict within this herd. This may also be one possible explanation for cows kept in tie stalls showing higher frequencies of social licking than cows in loose housing as found by Krohn (1994) and by ourselves (Laister *et al* unpublished data) in 31 loose and 12 tied housing dairy farms in Austria, Germany and Italy (0.31 [± 0.28] vs 1.10 [± 0.66] events per animal and per hour.

Forced spatial vicinity between the cows and impossibility of avoidance might lead to increased attempts to reduce social tension by licking. However, in contrast, increased familiarity between the pairs of tethered cows may also be a contributory factor for higher licking levels (Sato *et al* 1993).

Other authors propose that social licking increases under more restrictive housing conditions (Reinhardt 1980), where licking could be a way of coping with restrictive conditions by self-narcotisation (Fraser & Broom 1990). From studies with rats and primates it is known that opioids are involved in allogrooming both in the receiver and the actor (Keverne *et al* 1989; Niesink & van Ree 1989). Both appeasement and self-narcotisation might be reflected by a decrease in heart rate. Interestingly, in tethered animals, we found more pronounced and consistent heart rate reductions in actors than in receivers, especially in those that licked spontaneously (Laister *et al* unpublished data). Finally, licking could be a way of reducing boredom or oral understimulation (Fraser & Broom 1990). From the current state of knowledge, or rather, of informed speculations, we conclude that although an association of social licking with positive feelings on the individual level can be expected, it may, in certain cases, merely alleviate poor welfare. Future research should try to identify better measures concerned with social licking that reflect more closely the actual affective state of the animals at the herd level.

Reliability

A fundamental requirement of an assessment system is sufficient reliability of the measures applied. This includes the notion that different assessors with a certain degree of training should achieve the same results as far as possible (inter-observer reliability) or that results are largely the same in repeated tests with the same subjects (test-retest reliability). In the following section we will address the difficult question of limits of acceptability of agreements and how to improve agreements. Further, we will discuss the special case of longer-term consistency of results for assessments serving certification purposes and will draw attention to the problems that arise when recording sporadic behaviours.

The paucity of information on achieved reliability measures for welfare indicators is remarkable. This applies especially to observations of spontaneous behaviour, scoring of the integument or signs of clinical disease, while considerably more published reliability data are available on lameness scoring (eg Winckler & Willen 2001; de Rosa *et al* 2003; Engel *et al* 2003; O'Callaghan *et al* 2003; March *et al* 2007; Rushen *et al* 2007; Borderas *et al* 2008; Thomsen *et al* 2008) and behavioural tests regarding the human-animal relationship (eg de Rosa *et al* 2003; Lensink *et al* 2003; Waiblinger & Menke 2003; Rousing & Waiblinger 2004; Waiblinger *et al* 2007).

The somewhat alarming result of a majority of the publications cited above and of reliability testing within the Welfare Quality® project (eg Brenninkmeyer *et al* 2007) is that robust agreement between different assessors and even within assessors (intra-observer reliability) is often not easy to achieve. While no unequivocal scientific criteria for the setting of limits for 'good' or 'acceptable' agreement are available, some opinions are given in the literature. For instance, with respect to the use of correlation coefficients, Martin and Bateson (2007) state that, regarding "an important category that is difficult to measure, a rough guideline for acceptability might be a correlation coefficient of at least 0.7." Regarding Kappa values, different limits are given. Mostly, the maximum lower limit is set at 0.4 (Fleiss *et al* 2003). However, a Kappa coefficient of 0.4 may, in an admittedly extreme case, mean that two assessors gait scoring the same 100 cows, may have found lameness prevalences of 50 or 80%, respectively, with 70%

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

agreement between their dichotomous scorings. Similarly, a correlation coefficient of 0.7 means that less than 50% of the variance in the assessments is common between two assessors. We agree with de Passillé and Rushen (2005) that for welfare assessments that aim to classify farms according to their animal welfare level and which may have economic consequences for the farmers, higher reliability needs to be achieved than at the limits discussed before. At the same time, we similarly agree that it would be necessary to also test reliability at farm instead of individual animal (or group) level, and additionally at the level of an aggregated welfare assessment. This is yet to be done and is another important task for the future.

Means to improve reliability are straightforward: refining definitions or data recording design and training. Refining definitions can be an improvement of the description of categories as has been attempted, for instance, by Thomsen *et al* (2008), building on a number of existing lameness scoring systems. A further efficient way of increasing the proportion of observers allocating the same score may also be to merge several more detailed classes into fewer classes, as has been done by Brenninkmeyer *et al* (2007) and March *et al* (2007). They merged a 5-point lameness scale into a 2-point scale (lame/not lame) which led to improvements in PABAKs (prevalence adjusted bias adjusted kappa coefficient; Byrt *et al* 1993) of up to 0.3. It should be mentioned, however, that the PABAK does not differentiate between small and large disagreements. The improvement in agreements by merging scores is associated with the cost of lower discrimination ability which may or may not be a problem depending on the goal of the assessment. Similar effects, in terms of both improved agreement and decreased discrimination ability, are obtained by allowing greater differences between observers (Engel *et al* 2003). This means, for example, that deviations by one score count as agreement.

Regarding the efficiency of training, varying results can be found depending on the assessors' previous experience, difficulty of the assessment (both in terms of the assessment system and the animals to be assessed) and intensity of training. Thomsen *et al* (2008), after very brief training of experienced assessors, found even a slight decrease in intra-observer agreement regarding gait scoring. However, in general, training will lead to improvements, but a considerable number of training assessments might be necessary (Brenninkmeyer *et al* 2007; March *et al* 2007). This requires time and resources that are often limited. Also, some individual variation in assessment ability will remain that also affects training needs (Engel *et al* 2003).

The repeatability of assessments over time is especially important and constitutes a special case where welfare assessment protocols are going to be used for certification purposes. In order to be cost effective, such assessments will take place in longer intervals of supposedly more than six months. This means that assessment results need to be representative of the longer-term farm situation instead of being sensitive to changes in environmental or internal conditions that are largely insignificant for the welfare state of the animals. Thus, similar recordings should be achieved at different times if no major changes on the farms occurred. Published data on test-retest reliability with a number of days in between, here on responses towards an unknown approaching person, on-farm (Lensink *et al* 2003; Rousing & Waiblinger 2004), again show the difficulty of achieving the limits discussed above, with resulting misclassifications of animals in one or the other test of up to 48% (de Passillé & Rushen 2005). In our investigations (eg Laister *et al* 2007; Plesch *et al* 2008), we looked at Kendall rank correlations between results of behavioural observations at farm level at three different observation days that were 60 and 120 days apart as a reliability measure. Many of the observed behavioural categories did not show satisfactory correlations over time. However, even for behavioural categories exceeding the threshold for Kendall's W of 0.7, such as social licking in beef bulls, it can be seen in Figure 1(a) that for many farms ranks varied considerably between observation days. We concede that a ranking of farms (which is the basis for calculating Kendall's W) is predominantly of interest for benchmarking in the context of advisory activity where longer-term reliability problems are less critical. However, Figure 1(b) shows that the absolute values similarly show great variability from visit-to-visit. For certification purposes, one or more limits have to be defined which will allow classification of the farms. A limit of one licking event per animal and hour, just to create an arbitrary example, would lead to nine farms (50%) being misclassified at one or another visit. Possible reasons for variations in the measures might have been slight changes in climate, herd composition, internal states of the animals etc, all of which are within a usual variation in farm conditions that should not affect the principal welfare assessment of the farms.

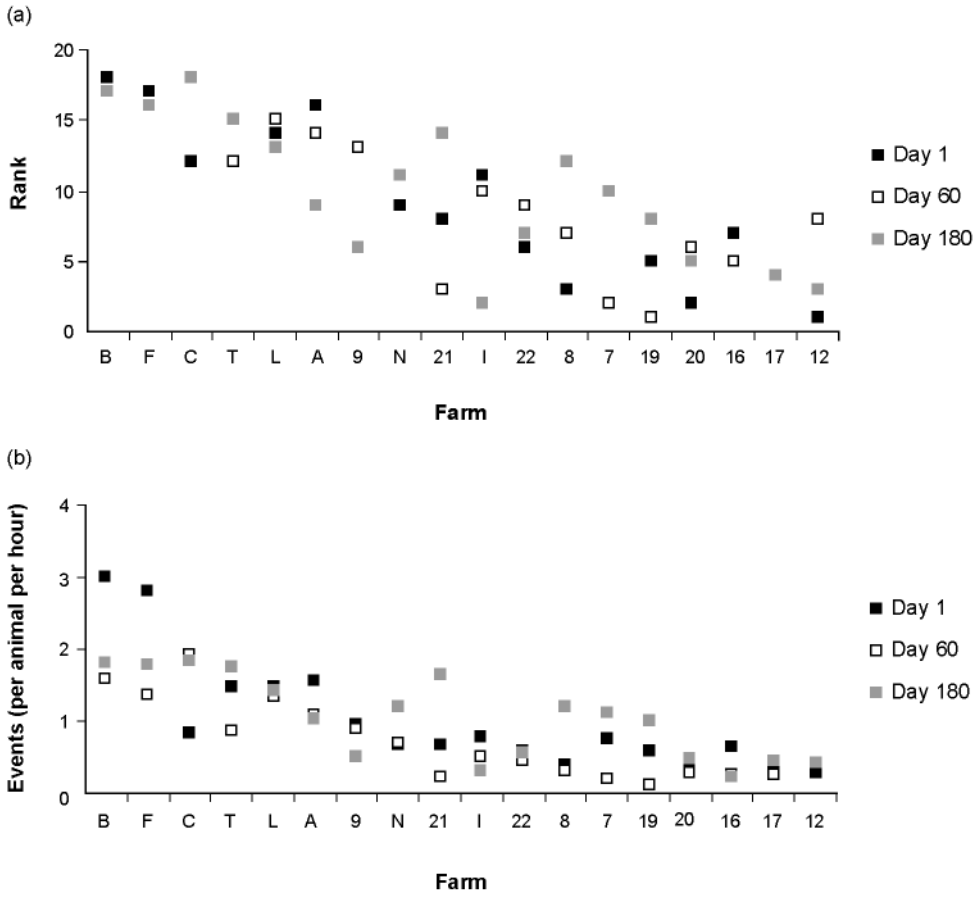
Winckler *et al* (2007b), during five visits at two-month intervals on eight dairy farms in which they took a number of behavioural and clinical measures, similarly found relatively low repeatability over time. Only for avoidance distances towards an unknown person were correlation coefficients consistently over 0.7 found (r_s = 0.76 to 0.81; $P < 0.05$). However, for prevalences of lameness and lesions of the tarsal joint, cleanliness of the hind leg and avoidance distance correlations between single visits and the average, values were almost exclusively above 0.7. Taking these initial investigations into account, the discussion about solutions for the basic problem of variability of measures requires to be intensified.

In general, we believe that three issues need much more attention in future welfare studies: (i) reliability needs to be tested and reported as a precondition for any use of welfare measures in general; (ii) more discussion and perhaps standardisation is needed on appropriate statistical tests of reliability and (iii) limits for acceptability need to be reconsidered, in, of course, the context of the objectives of the use of the welfare measures.

Finally, the problem of sporadic behaviours that cannot reliably be recorded in a short time, even if they are important will be highlighted. Examples of such behaviours

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

Figure 1



Ranking according to (a) social licking events as well as (b) incidence of social licking of 18 beef cattle farms sorted by mean rank on three different days (weight class > 350 kg; day 60 and day 180 = days apart from first visit, Kendall's *W* = 0.73; Laister *et al* unpublished data).

are play, abnormal rising or lying down, tongue rolling or slipping. Play, for instance, is one of the uncontested indicators of good welfare (Knierim *et al* 2001; Winckler *et al* 2003; Boissy *et al* 2007) and according to our own observations does not only occur in young, but also in adult cattle. However, its average occurrence is so low that no representative data can be recorded during short-term observations. In our investigations, we set arbitrary limits of, for instance, 1.0 events of abnormal rising or lying down occurrence, such as horse-like rising per hour (Plesch *et al* 2008) or of 0.1 events of other behaviours per animal and hour (Laister *et al* 2007). However, such limits also require further discussion. A future task may additionally be to develop methods that allow consideration of important but rare behaviours, for instance by automatic recording. Approaches that show promise include accelerometers or image analysis techniques as have currently been developed, eg for lameness detection (Bahr *et al* 2008).

Feasibility

In general, the Welfare Quality® assessment protocols proved to work well in the on-farm situation. Some measures need to be carried out at certain times of the day (eg social behaviour observations after feed delivery or after feed on the feed bunk is moved towards the feed fence in order to bring it within the reach of animals), but there is also some flexibility in the system, eg the timing of the farmer interview can vary. However, in very large herds, the identification of representative samples may be problematic.

The main constraint is the total time required for the current assessment protocols. Given the fact that some measures, such as clinical scoring or avoidance distance recording at the feed place are assessed in a representative sample, the time needed depends, to a certain extent, on herd size. Based on first experiences, this amounts to approximately 5.5 h net time for a herd of 60 dairy cows (Table 1), ie this is the time needed for the recordings and farmer's interview conducted by a single assessor.

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

Table 1 Average time needed for the application of the Welfare Quality® assessment protocol in dairy cattle and estimates for the duration of farm visits depending on herd size (for full protocol see Welfare Quality® Consortium 2009).

	Mean (± SD) duration (min) (n = 40–49)	Estimated time required			
Herd size	–	25	60	100	200
Sample size for measurements in individual animals	–	20	37	49	65
Avoidance distance at feed place	1.0 (± 0.4)[1]	20	37	49	65
Qualitative behaviour assessment	25 (± 4)[2]	25	25	30	30
Behaviour recordings (social, resting, coughing, sneezing)	147 (± 16)[2]	145	143	155	155
Clinical scoring (cleanliness, body condition, injuries etc)	3.0 (± 0.8)[1]	60	111	147	195
Resources/management (checklist and questionnaire)	*	15	15	15	15
Total time (h)	–	**4.4**	**5.6**	**6.6**	**7.7**

[1] per animal; [2] per farm; * only estimates available.

In terms of the costs of the assessment, it is conceivable to aim for a reduction in the time needed to implement the protocol. Working towards reducing the protocol is currently being undertaken. In line with the aim to focus on animal-based measures, currently the animal-based measures in the protocol amount to almost two-thirds of the measures. It can be seen from Table 1 that the one-third of resource- and management-based measures require minimal time compared with the animal-based measures. It remains a challenge to reduce the time needed for the assessment whilst still fulfilling the claim of an on-farm welfare assessment system that focuses on animal-based measures and is scientifically sound.

As far as farmers were concerned, the duration of farm visits was, however, not regarded problematic. In a beef cattle implementation study with 90 beef farmers interviewed in Austria, Italy and Germany, 7 h were mentioned as being an acceptable duration of farm visits (Kirchner *et al* 2008). The vast majority of farmers responded positively about the application of the protocols. This may be partly due to the fact that they are not heavily involved in data collection. Perhaps even more important is the high level of interest they show in the animal-based parameters, which is information they are usually not provided with.

Conclusion and animal welfare implications

Animal-related measures for the on-farm cattle welfare assessment bear a high chance of validity. Therefore, the approach taken by, eg the Welfare Quality® project to combine current knowledge and to further develop and test animal-based measures can be regarded a big essential step. However, from our experience, there are still a considerable number of challenges which should be tackled while using the present assessment system in order to constantly improve it. Among others, these include deeper insight into the biological significance of differing degrees of welfare impairment of apparently valid parameters, observer training and inter-observer reliability issues, and consistency of measures over time. While welfare assessment systems focusing seriously on animal-based measures currently require several hours of on-farm recordings, it is hoped that the development of new methods, technical solutions (such as automatic behaviour recording or use of already existing databases) may allow time to be saved or the coverage of aspects that currently cannot be assessed. Furthermore, areas thus far excluded, such as pasture-based production systems, should also be covered in future.

Acknowledgements

The studies referred to are part of the Welfare Quality® research project which has been co-financed by the European Commission, within the 6th Framework Programme, contract Number. FOOD-CT-2004-506508. The text represents the authors' views and does not necessarily represent a position of the Commission who will not be liable for the use made of such information.

References

Aureli F, Preston SD and de Waal FB 1999 Heart rate responses to social interactions in free-moving rhesus macaques (*Macaca mulatta*): A pilot study. *Journal of Comparative and Physiological Psychology 113*: 59-65

Bahr C, Leroy T, Song XiangYu, Maertens W, Vranken E, Nuffel A van, Vangeyte J, Sonck B and Berckmans D 2008 Automatic detection of lameness in dairy cattle by vision analysis of cows' gait. *Agricultural and Biosystems Engineering for a Sustainable World. International Conference on Agricultural Engineering*. 23-25 June 2008, Hersonissos, Greece

Bartussek H 2001 An historical account of the development of the animal needs index ANI-35L as part of the attempt to promote and regulate farm animal welfare in Austria: An example of the interaction between animal welfare science and society. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 34-41

Blokhuis HJ, Jones RB, Geers R, Miele M and Veissier I 2003 Measuring and monitoring animal welfare: Transparency in the food product quality chain. *Animal Welfare 12*: 445-455

Boccia M, Reite M and Laudenslager M 1989 On the physiology of grooming in a pigtail macaque. *Physiology & Behavior 45*: 667-670

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press

Boissy A, Manteuffel G, Jensen MB, Oppermann Moe R, Spruijt BM, Keeling L, Winckler C, Forkman B, Dimitrov I, Langbein J, Bakken M, Veissier I and Aubert A 2007 Assessment of positive emotions in animals to improve their welfare. *Physiology & Behavior 92*: 375-397

Borderas TF, Fournier A, Rushen J and de Passillé AM 2008 Effects of lameness on dairy cows' visits to automatic milking systems. *Canadian Journal of Animal Science 88*: 1-8

Botreau R, Veissier I and Perny P 2009 Overall assessment of cow welfare; strategy adopted in Welfare Quality® *Animal Welfare 18*: 363-370

Botreau R, Veissier I, Butterworth A, Bracke MBM and Keeling LJ 2007 Definition of criteria for overall assessment of animal welfare. *Animal Welfare 16*: 225-228

Bracke MBM, Metz JHM and Spruijt BM 2001 Development of a decision support system to assess farm animal welfare. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 17-20

Brenninkmeyer C, Dippel S, March S, Brinkmann J, Winckler C and Knierim U 2007 Reliability of a subjective lameness scoring system for dairy cows. *Animal Welfare 16*: 127-129

Broom DM 1991 Animal welfare: Concepts and measurement. *Journal of Animal Science 69*: 4167-4175

Byrt T, Bishop J and Carlin JB 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46*: 423-429

Dawkins MS 2004 Using behaviour to assess animal welfare. *Animal Welfare 13*: 3-7

de Passillé AM and Rushen J 2005 Can we measure human-animal interactions in on-farm animal welfare assessment? Some unresolved issues. *Applied Animal Behaviour Science 92*: 193-209

de Rosa G, Tripaldi C, Napolitano F, Grasso F, Bisegna V and Bordi V 2003 Repeatability of some animal related variables in dairy cows and buffaloes. *Animal Welfare 12*: 625-629

Emmerig H 2004 *Behavioural indicators of good welfare in dairy cows, an exploratory approach*. Bachelor Thesis, Faculty of Organic Agricultural Sciences, University of Kassel, Germany

Engel B, Bruin G, Andre G and Buist W 2003 Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows. *Journal of Agricultural Science 140*: 317-333

Fleiss JL, Levin B and Paik MC 2003 *Statistical Methods for Rates and Proportions* pp 598-626. John Wiles & Sons: Hoboken, NJ, USA

Fraser AF and Broom DM 1990 *Farm Animal Behaviour and Welfare*, *Third Edition*. Baillière Tindall: London, UK

Johnsen PF, Johannesson T and Sandøe P 2001 Assessment of farm animal welfare at herd level: Many goals, many methods. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 26-33

Keverne EB, Martensz ND and Tuite B 1989 Beta-endorphin concentrations in cerebrospinal fluid of monkeys are influenced by grooming relationships. *Psychoneuroendocrinology 14*: 155-161

Kirchner M, Schulze Westerath H, Tessitore E, Cozzi G, Knierim U and Winckler C 2008 Perception and attitudes of beef farmers towards the Welfare Quality® assessment system. In: Koene P (ed) *Book of Abstracts 4th International Workshop on the Assessment of Animal Welfare at Farm or Group Level* p 159. 10-13 September 2008, Ghent, Belgium

Knierim U, Carter CS, Fraser D, Gärtner K, Lutgendorf SK, Mineka S, Panksepp J and Sachser N 2001 Good Welfare: Improving Quality of Life. In: Broom DM (ed) *Coping with Challenge: Welfare in Animals including Humans. Dahlem Workshop Report* pp 79-100. Dahlem University Press: Berlin, Germany

Knierim U, Sundrum A, Bennedsgaard T, Roiha U and Johnson PF 2004 Assessing animal welfare in organic herds. In: Vaarst M, Roderick S, Lund V and Lockeretz W (eds) *Animal Health and Welfare in Organic Agriculture* pp 189-203. CAB International: Wallingford, UK

Krohn CC 1994 Behaviour of dairy cows kept in extensive (loose housing/pasture) or intensive (tie stall) environments. III. Grooming, exploration and abnormal behaviour. *Applied Animal Behaviour Science 42*: 73-86

Laister S, Hesse N, Zucca D, Knierim U, Minero M, Canali E and Winckler C 2007 Suitability of selected behavioural indicators for on-farm welfare assessment in loose housed dairy cattle. In: Galindo F and Alvarez L (eds) *Proceedings of the 41st International Congress of the ISAE* p 94. 30 July-3 August 2007, Merida, Mexico

Laister S, Regner A-M, Zenger K, Winckler C, Hesse N, Quast R and Knierim U Validation of social licking as an indicator for positive emotions. *Unpublished Report EU Project, Welfare Quality®* Lelystad, The Netherlands

Lensink BJ, Van Reenen CG, Engel B, Rodenburg TB and Veissier I 2003 Repeatability and reliability of an approach test to determine calves' responsiveness to humans: 'a brief report'. *Applied Animal Behaviour Science 83*: 325-330

Main DCJ, Whay HR, Leeb C and Webster AJF 2007 Formal animal-based welfare assessment in UK certification schemes. *Animal Welfare 16*: 233-236

March S, Brinkmann J and Winckler C 2007 Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. *Animal Welfare 16*: 131-133

Martin P and Bateson P 2007 *Measuring Behaviour*. Cambridge University Press: Cambridge, UK

Niesink RJM and van Ree JM 1989 Involvement of opioid and dopaminergic systems in isolation-induced pinning and social grooming of young rats. *Neuropharmacology 28*: 411-418

O'Callaghan KA, Cripps PJ, Downham DY and Murray RD 2003 Subjective and objective assessment of pain and discomfort due to lameness in dairy cattle. *Animal Welfare 12*: 605-610

Plesch G, Brörkens N, Laister S, Winckler C and Knierim U 2008 Reliability testing concerning behaviour around resting in dairy cows. In: Koene P (ed) *Book of Abstracts 4th International Workshop on the Assessment of Animal Welfare at Farm or Group Level* p 87. 10-13 September 2008, Ghent, Belgium

Reinhardt V 1980 *Untersuchungen zum Sozialverhalten des Rindes*. Birkhäuser Verlag: Stuttgart, Germany. [Title translation: Investigations on the social behaviour of cattle]

Rousing T and Waiblinger S 2004 Evaluation of on-farm methods for testing the human-animal relationship in dairy herds with cubicle loose housing systems: test-retest and inter-observer reliability and consistency to familiarity of test person. *Applied Animal Behaviour Science 85*: 215-231

Rushen J and de Passillé AMB 1992 The scientific assessment of the impact of housing on animal welfare: a critical review. *Canadian Journal of Animal Science 72*: 721-743

Rushen J, Pombourcq E and de Passillé AM 2007 Validation of two measures of lameness in dairy cows. *Applied Animal Behaviour Science 106*: 173-177

Sato S and Tarumizu K 1993 Heart rates before, during and after allo-grooming in cattle (*Bos taurus*). *Journal of Ethology 11*: 149-150

Sato S, Sako S and Maeda A 1991 Social licking patterns in cattle (*Bos taurus*): influence of environmental and social factors. *Applied Animal Behaviour Science 32*: 3-12

Sato S, Tarumizu K and Hatae K 1993 The influence of social factors on allogrooming in cows. *Applied Animal Behaviour Science 38*: 235-244

Scott EM, Nolan AM and Fitzpatrick JL 2001 Conceptual and methodological issues related to welfare assessment: A framework for measurement. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 5-10

Sørensen JT and Sandøe P 2001 Assessment of animal welfare at farm or group level. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*

Sørensen JT, Sandøe P and Halberg N 2001 Animal welfare as one among several values to be considered at farm level: The idea of an ethical account for livestock farming. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 11-16

Spoolder HAM, de Rosa G, Hörning B, Waiblinger S and Wemelsfelder F 2003 Integrating parameters to assess on-farm welfare. *Animal Welfare 12*: 529-534

Thomsen PT, Munksgaard L and Tøgersen FA 2008 Evaluation of a lameness scoring system for dairy cows. *Journal of Dairy Science 91*: 119-126

Waiblinger S and Menke C 2003 Influence of sample size and experimenter on reliability of measures of avoidance distance in dairy cows. *Animal Welfare 12*: 585-589

Waiblinger S, Fresdorf A and Spitzer G 2002 The role of social licking in cattle for conflict resolution. *Proceedings of the 1st European Conference of Behavioural Biology* p 122. 1-4 August 2002, Münster, Germany

Waiblinger S, Knierim U and Winckler C 2001 The development of an epidemiologically based on-farm welfare assessment system for use with dairy cows. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30:* 73-77

Waiblinger S, Mülleder C, Schmied C and Dembele I 2007 Assessing the animals' relationship to humans in tied dairy cows: between-experimenter repeatability of measuring avoidance reactions. *Animal Welfare 16*: 143-146

Webster AJF and Main DCJ 2003 Special Issue: Proceedings of the 2nd International Workshop on the Assessment of Animal Welfare at Farm or Group Level. *Animal Welfare 12(4)*

Welfare Quality® Consortium 2009 *Welfare Quality® Assessment Protocol for Cattle.* Lelystad: The Netherlands

Wemelsfelder F and Lawrence AB 2001 Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. *Acta Agriculturae Scandinavica Section A Animal Science* Supplementum *30*: 21-25

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2001 Assessing the 'whole animal': a free choice profiling approach. *Animal Behaviour 62*: 209-220

Wemelsfelder F, Knierim U, de Rosa G, Napolitano F and Haslam S 2008 The development of qualitative behaviour assessment as an on-farm welfare inspection tool. In: Koene P (ed) *Book of Abstracts 4th International Workshop on the Assessment of Animal Welfare at Farm or Group Level* p 52. 10-13 September 2008, Ghent, Belgium

Wemelsfelder F, Nevison I and Lawrence AB 2009 The effect of perceived environmental background on qualitative assessments of pig behaviour. *Animal Behaviour 78*: 477-484

Whay HR, Main DCJ, Green LE and Webster AJF 2003 Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: Consensus of expert opinion. *Animal Welfare 12*: 205-217

Winckler C and Willen S 2001 The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agriculturae Scandinavica Section A Animal Science Supplementum 30*: 103-107

Winckler C, Baumgartner J and Waiblinger S 2007a Proceedings of the 3rd International Workshop on the Assessment of Animal Welfare at Farm or Group Level. *Animal Welfare 16(2)*

Winckler C, Brinkmann J and Glatz J 2007b Long-term consistency of selected animal-related welfare parameters in dairy farms. *Animal Welfare 16*: 197-199

Winckler C, Capdeville J, Gebresenbet G, Hörning B, Roiha U, Tosi M and Waiblinger S 2003 Selection of parameters for on-farm welfare assessment protocols in cattle and buffalo. *Animal Welfare 12*: 619-624

https://doi.org/10.1017/S0962728600000865 Published online by Cambridge University Press