Developing effective welfare measures for cattle

Ute Knierim, Christoph Winckler, Luc Mounier, Isabelle Veissier

Last authors version of : Book chapter in Marcia Endres (ed.), *Understanding the behaviour and improving the welfare of dairy cattle, burleigh dodds, 2021* ⟨10.19103/AS.2020.0084.05⟩. **Embargo until September 2 2021**

Developing effective welfare measures for cattle

Ute Knierim, University of Kassel, Germany; Christoph Winckler, University of Natural Resources and Life Sciences, Vienna, Austria; and Luc Mounier and Isabelle Veissier, Université Clermont Auvergne, INRAE, VetAgro Sup, France

Abstract

This chapter focusses on the performance characteristics a welfare measure should possess in order to be considered valid for the assessment of animal welfare. It presents a choice of validation measures that can be used to assess the welfare of cattle and discusses ways they can be collected in practice. The chapter also presents the various definitions of animal welfare and how these definitions can affect the measures that are chosen.

Key words: animal welfare; dairy cow; method validation; assessment; welfare measures

1 Introduction

Welfare is a multi-dimensional concept. It includes any aspect of the animal's life that makes it a good life to live (see for instance the definition of five freedoms by the Farm Animal Welfare Council 1992). The assessment of the welfare of an animal or a group of animals can therefore hardly rely on one single measure (Botreau et al. 2007). Rather a series of measures should be used to check the various dimensions of animal welfare.

There has been extensive discussion on whether measure x or y is valid to assess animal welfare or whether the assessment of welfare shall be performed with resource-based measures (such as space allowance, quality of the management) versus animal-based measures (such as body condition, clinical symptoms, performance or behaviour). We argue that welfare measures should undergo a precise process of validation before being considered valid or not.

The objectives of this chapter are to draw a list of performance-characteristics that a welfare measure should possess in order to be considered valid for the assessment of animal welfare. Further, we aim to present a choice of validated measures that can be used to assess the welfare of cattle and to discuss ways how they can be collected in practice. As a preamble, we will remind the reader of the various definitions of animal welfare, because they affect the measures chosen.

2 Definition of animal welfare

The Brambell committee established by the British government provided a first definition of animal welfare, by declaring that *Welfare is a wide term that embraces both the physical and mental wellbeing of the animal* (Brambell 1965). In the debate on animal welfare concepts, this very general definition has been broken down into aspects that contribute to good or poor welfare. Diverging points of view then appeared on the relative importance of these aspects. There is apparently a consensus that a prerequisite for good welfare is the absence of intense and prolonged suffering from pain, hunger, thirst, discomfort, fear or stress. There is, however, no clear agreement on other less-severe issues around biological functioning (absence of malnutrition, injury and disease) or the experience of positive experiences (Fraser 1995). Positive experiences can come from access to particularly appreciated resources (for sleeping comfort, food) or from the expression of activities for which the animals are motivated

(play activities such as calves frolicking with their tails up, exploration of the environment like pigs rooting in the ground, or cohesive social interactions like licking between animals that have affinity bonds) (Boissy et al. 2007). Fraser (1995) emphasises that increasing biological knowledge may move differing viewpoints closer to agreement, but that 'the inherently subjective element in judging the relative importance of different attributes' cannot be eliminated.

A core issue is that animals are sensitive beings, that is, they can experience affective states such as emotions, and for some authors that is why welfare makes sense (Duncan 2002, Veissier and Boissy 2007). According to these authors, it is not, for example, the disease that in itself degrades the welfare of an individual, but the fact that this individual feels sick. Here the emphasis is on mental health, and physical health is supposed to affect mental health. Discussing whether welfare is solely a matter of mental state or a matter of both mental and physical health is out of scope of the present chapter. Actually, the assessment of animal welfare is generally based on assessing the behaviour of an animal – that may reflect its mental state – and its physical state – either because it is considered to be part of welfare or because it may affect the mental state.

The different aspects of animal welfare, or principles that must be fulfilled to assure a good life, can serve as a basis to develop checklists for the assessment of animal welfare. The Farm Animal Welfare Council, which followed the Brambell Committee, defined five principles to be respected in order to achieve good welfare; their descriptions put emphasis on what animals may feel (Farm Animal Welfare Council 1992):

- Absence of prolonged hunger and thirst,
- Physical comfort,
- No injuries, illnesses or pain,
- Absence of fear and distress,
- Ability to express the normal behaviour of the species.

These principles, commonly known as the *five freedoms*, form the basis of many regulations and methods of assessing welfare. They were detailed into 12 independent criteria in the Welfare Quality® project in order to propose a common grid, making it

possible to identify a set of measures covering all aspects of welfare (Botreau et al. 2007):

- Absence of prolonged hunger: Animals should not suffer from prolonged hunger, that is, they should have a suitable and appropriate diet;
- Absence of prolonged thirst: Animals should not suffer from prolonged thirst, that is, they should have a sufficient and accessible water supply;
- Comfort around resting: Animals should have comfort when they are resting;
- Thermal comfort: Animals should be neither too hot nor too cold;
- Ease of movement: Animals should have enough space to be able to move around freely;
- Absence of injuries: Animals should be free of injuries, for example, skin damage and locomotion disorders;
- Absence of disease: Animals should be free from disease;
- Absence of pain induced by management procedures: Animals should not suffer pain induced by inappropriate management, handling, slaughter or surgical procedures (e.g. castration, dehorning);
- Expression of social behaviours: Animals should be able to express normal, non-harmful, social behaviours (e.g. grooming);
- Expression of other behaviours: Animals should be able to express other normal behaviours, that is, it should be possible to express species-specific natural behaviours such as foraging;
- Good human-animal relationship: Animals should not be afraid of humans and be handled well in all situations, that is, handlers should promote good human-animal relationships;
- Positive emotional state: Negative emotions such as fear, distress, frustration or apathy should be avoided whereas positive emotions such as security or contentment should be promoted.

With such operational definitions of animal welfare, measures can be proposed and problems quantified (Stafleu et al. 1996). Indeed, the Welfare Quality® framework is now being used as a reference to develop animal welfare assessment systems in various species, that is, cattle, pigs, hens and broilers during the Welfare Quality project but also horses, sheep, turkeys, and dolphins (Welfare Quality 2009, AWIN

2015, Clegg et al. 2015). In these assessment systems, the measures to check the various welfare aspects were chosen according to their validity, their ability to measure what is aimed to be measured with an acceptable precision, and their feasibility, that is they can be applied in field conditions.

3 Performance characteristics to define valid welfare measu

In animal welfare science, the validation process concerning measures is commonly broken down into the assessments of 'validity' and 'reliability', with different possible approaches such as determining face, concurrent or construct validity (see below) and testing intra-observer or inter-observer reliability (e.g. Scott et al. 2001, Keeling 2009). To go further and formalise the validation process, we looked towards further disciplines and their approaches of validation to see what can be learned for the animal welfare case. For instance, for analytical assays (e.g. biochemical or microbiological assays), the validation of a method (test of fitness for purpose) includes the evaluation of method performance-characteristics such as selectivity, trueness, precision, working range, sensitivity, ruggedness and matrix variation (Taverniers et al. 2004, Magnusson and Örnemark 2014). In addition, the measures need to be fit for use, that is practical aspects such as acceptability by scientific and regulatory communities (which might conform to face validity) and feasibility, including cost and time needed, must also be taken into account (EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare) 2012).

Not all performance criteria of analytical methods might be applicable to animal welfare. Here, we propose a validation process inspired from these concepts so as to strengthen the choice of welfare measures in further studies. Although terms can vary between disciplinary fields, we preferably use the definitions of performance characteristics provided by Magnusson and Örnemark (2014) in the Eurachem Guide.

3.1 Selectivity and trueness

For analytical methods, selectivity is *the extent to which the method can be used to determine particular analytes in mixtures or matrices without interferences from other components of similar behaviour* (Vessman et al. 2001). For a welfare measure, selectivity may first refer to the degree to which a measure can quantify what we want

to analyse and not something else. This property should be analysed in reference to the welfare criterion the measure is supposed to bring information about, for example, 'is a method valid to detect stress, abnormal behaviour or hunger, etc.'? It is often difficult to distinguish selectivity from trueness, which – for analytical methods – refers to the closeness of agreement between a test result and the accepted reference value of what is being measured (Thompson et al. 2002).

These two properties are often considered together and for welfare measures (lacking a reference value) they can be assessed by (Scott et al. 2001, Reenen and Engel 2004):

- Comparing the results produced by the method and those produced by another already validated method that serves as a gold standard (concurrent or criterion validity). For instance, regarding behaviour, one may compare the results obtained during a short period to those obtained by longer observations. In case scan sampling is used (i.e. one snapshot observation every fixed interval of time), the results can be compared to those obtained with continuous observations.
- Comparing the effects of conditions or treatments or demonstrating causal relationship between treatment and effect (predictive or construct validity). For instance, regarding the criterion 'Good human-animal relationship', one may ask whether an approach test – measuring at what distance one can approach an animal before it moves away – performed when the animal is feeding reflects the quality of the human-animal relationship. This question was answered in calves by comparing animals that received positive versus negative contacts with humans: the former accepted to be approached at a closer distance than the latter (Lensink et al. 2003).
- Agreeing among experts on the validity of a measure according to their experience (consensus or face validity), if concurrent or construct validity cannot be established.

It is also necessary to check for confounding factors that can influence the results. For instance, lameness can affect the results of an approach test: lame cows can be approached by a shorter distance than healthy ones, probably because the motivation

to avoid humans is counterbalanced by the pain caused by the movement (Špinka et al. 2005). In that case, lameness can be a confounding factor.

3.2 Precision

For analytical methods, precision refers to how close results are to one another (Magnusson and Örnemark 2014). It is often expressed as standard deviation or relative standard deviation. It includes:

- repeatability (sometimes called intra-assay repeatability), that is, the closeness between measures done in the same conditions, that is, by the same operator and with the same equipment,
- reproducibility (sometime called inter-assay repeatability), that is, the closeness between measures done in different conditions, for example, by different operators, or with different equipment, at different times.

For welfare measures, authors often refer to:

- *Intra-observer repeatability (or intra-observer reliability*), that is, the similarity between measures done by the same observer under similar conditions. This can be tested using video recordings for behaviours or assessing the same animals again after a short time lapse for clinical symptoms.
- *Inter-observer repeatability (or inter-observer reliability)*, that is, the similarity between measures done by several observers, under the same conditions. This may be checked by asking several observers to look at the same animals at the same time.

Test-retest repeatability (or test-retest reliability) is also sometimes used, meaning that a method should produce same results when it is applied twice within a short period of time. This property rather refers to ruggedness, which is described below. Very often, repeatability is expressed as a correlation coefficient between observations in case of numerical data, that is, frequencies of certain behaviours. Spearman, Pearson or intra-class correlations, or Kendall's coefficient of concordance for more than two observers are then calculated. In case of ordinal or nominal data, for example, lesion scores, often kappa coefficients are used (e.g. PABAK: prevalence-adjusted and bias-adjusted kappa). Although there is no natural limit when repeatability is high enough to produce trustworthy data, a correlation of above 0.70

or a concordance of more than 0.40 is generally considered as a minimum when the measure is very important and it is difficult to reach higher agreements. However, values above 0.8 for a correlation or 0.6 for a concordance should be the goal (Fleiss et al. 2003, Martin and Bateson 2007, Knierim and Winckler 2009).

3.3 Working range and sensitivity

Sensitivity is a term used with a number of differing meanings (Magnusson and Örnemark 2014). While diagnostic sensitivity refers to the ability to correctly identify true positive cases (e.g. a disease) from all assessed cases, analytical sensitivity describes how well a change in analytical results corresponds to change in the measured quantity (e.g. an analyte concentration) (Feinberg 1996). Sometimes the term is also used in the sense of limit of detection. Thus, it is always important to qualify the intended meaning of the term. Regarding analytical sensitivity, one may question which degree of differentiation of a welfare measure is necessary. Is it, for instance, sufficient to differentiate between minor versus severe lameness or mild versus severe wounds? Insufficient differentiation may level farms with truly different welfare states, while very high differentiation may negatively affect precision (i.e. repeatability, reproducibility, or reliability) of the measurement.

For analytical methods, the working range is 'the interval over which the method provides results with an acceptable uncertainty' (Magnusson and Örnemark 2014). This interval should match with the purpose of the end-user of the method. In particular, the limit of quantification, that is the smallest value that can be detected correctly, defines the lowest end of the working range. For our purpose, the limit of quantification should allow detecting a welfare problem, for example, an animal does not suffer from hunger or is not diseased. To assess whether the quantification limit of a measure is fit for the purpose, we may question whether the method can be used to detect slight welfare problems. For instance, the indicator 'sunken eyes' is used to detect dehydration in calves. This however indicates an extreme state where this animal urgently needs to be rehydrated, often by intravenous administration of an electrolyte solution. This symptom is not sensitive enough to detect thirst that may occur on a farm, if the water points are scarce and the animal needs to wait a long time before

getting access to water. Nevertheless, the symptom of sunken eyes appears appropriate to detect dehydration after very long transports during warm weather. Another example regarding the quantification limit relates to the occurrence of rare behaviours (such as certain abnormal behaviours or play in adults). This could be detected by observing animals, but not necessarily be reliably quantified within a limited time frame (Knierim and Winckler 2009).

3.4 Ruggedness

In terms of analytical method validation, the ruggedness (sometimes called robustness) is the capacity of a method to produce similar results when minor deviations are made from the experimental procedure (Magnusson and Örnemark 2014). This can be assessed by making deliberate changes to the procedure and assess the impact on the results.

For welfare measures, this may be transposed to *stability over time*, especially day-to-day variations. Indeed, the results of a welfare assessment should be representative of the long-term farm or slaughter-plant situation. Although the overall conditions on a farm may not change during a given period and the overall welfare of animals should thus be stable, there might be differences in the results obtained from some measures because the conditions for observation have slightly changed or because of random variations between days. This is particularly true for behaviours whose occurrence is determined by multiple factors. Weather conditions or occasional events like small changes in the routine management of the farm can affect the expression of behaviours. The impact of such changes on the results should be analysed. The method will be considered robust if similar recordings are achieved at different times if no major changes on the farms – that is, changes that are significant for the welfare of the animals – occurred (Knierim and Winckler 2009).

3.5 Matrix variation

For analytical methods, the matrix is the set of constituents present in the material on which the method is applied. These constituents may have an effect on the results of

the method (Feinberg 1996, Thompson et al. 2002). For instance, the determination of hormones in different tissues or species may require specific methods.

For welfare measures to be applied on farms or at slaughter, it is important to check if the measure can be applied in different farming or slaughter systems. For instance, detecting lameness in cows in a loose barn versus tie-stalls requires a different method. In Welfare Quality® lameness is assessed in loose-housed cows by making them walk in a straight line and checking if the cow bears its weight equally on the four limbs and makes regular steps. For tied cows, the observer checks if the cow stands on its four limbs when undisturbed. Then the observer makes the cow move to the left and to the right, observing how she shifts weight from foot to foot (Welfare Quality® 2009).

3.6 Fitness for use

When the assessment has to be performed on a large scale and under commercial conditions, the feasibility of a method is essential. Important issues are the time necessary to carry out a measure (e.g. long-term observations to detect changes in time budget are less feasible), the need for specific devices to perform the measure, the requirement for specific skills to perform the measure (e.g. for taking blood samples) or the cost of measures. Feasibility relates to the knowledge and devices available at present. Therefore, a measure that is considered not feasible today may become feasible in the future, for instance, thanks to automatic recordings becoming widely applicable (e.g. automatic recordings of animals' movements).

We believe that following such a validation process for welfare measures is likely to help make such measures more acceptable by a wide range of users, thus helping to improve animal welfare.

4 Choice of measures for cattle

As described above, important aspects to consider when choosing measures of welfare are validity, reliability and feasibility. In this section, we first explain the selection process using two examples from the Welfare Quality® assessment protocol for dairy cattle. Then we discuss the main drivers for the composition of assessment protocols.

The measures included in the Welfare Quality® protocol for dairy cattle are listed in Table 1 together with a semi-quantitative assessment of the respective levels of validity, reliability and feasibility. For the criterion 'Expression of social behaviours', both agonistic and affiliative behaviours were considered for the sake of content validity, that is, in order to include all relevant items of social behaviours as a welfare measure. For feasibility reasons, behaviour observations had to be limited to 2 h to allow completion of the on-farm assessment within a day. Within this time frame, on-farm observations in different housing systems and countries revealed good to very good inter-observer agreement and at least an acceptable stability over time for the incidence of behaviours such as head butts or displacements, as well as for the incidence of total agonistic interactions (additionally including fighting, chasing or chasing up from the lying area) (Laister et al. 2009a,b). Affiliative behaviours such as social licking were considered as potential measures of good welfare. However, the validity of licking as a herd measure of positive social interactions is questionable because it may just alleviate poor welfare due to social tension (Knierim and Winckler 2009). It was eventually excluded as a measure of socio-positive behaviour from the operational assessment protocol.

Table 1 shows that valid animal-based welfare measures were not available for all criteria. For example, for absence of thirst, the skin test, which uses the delay needed for the skin to resume its initial position after being pinched, has been shown lacking relation to serum osmolarity or packed cell volume (indicating the level of hydration) and to drinking behaviour (Pritchard et al. 2006). For the direct measurement of osmolarity and haematocrit counts (Knowles et al. 1995, Pritchard et al. 2006), more research is needed to establish if these indicators can help to distinguish a moderate dehydration from a severe one. Additionally, collection of a blood sample on farms is an invasive procedure and time-consuming. In view of the above concerns, recording of resource-based measures such as the number of water points, the flow of water and its cleanliness are currently considered more appropriate for assessing compliance with the criterion 'Absence of prolonged thirst'.

Although the Welfare Quality® protocols aim at allowing for a comprehensive assessment covering all dimensions of welfare, a number of main behavioural measures, for which at least face validity can be established, were not included. For example, this applies to play behaviour (Fregonesi and Leaver 2001), abnormal behaviour such as tongue rolling or intersucking (Krohn 1994, Lidfors and Isberg 2003) and comfort behaviours such as brush use (Mandel et al. 2016). All these measures require long-term direct observations for a reliable assessment and are therefore often considered not feasible in the context of on-farm assessments (Knierim and Winckler 2009).

Comparing welfare assessment protocols which are implemented in the dairy industry (Table 2) reveals marked differences in the number and type of measures included. Measures of health and physical appearance – mainly addressing the biological functioning view of welfare – are consistently included. This is less often the case for behavioural measures that provide information about affective states or the ability to perform normal behaviour. Protocols also differ substantially from each other in their comprehensiveness regarding health measures. Lameness, body condition and skin alterations appear to be 'core measures'. More detailed information on the health state is less-frequently obtained.

Feasibility in terms of time needed and associated costs is often mentioned as the main driver for the final decision on which measures to finally include in assessment protocols (Sorensen et al. 2007, Metz et al. 2015). Some limitations have to be accepted for on-farm assessments, which – as mentioned above – may lead to the omission of welfare relevant indicators such as play behaviour. The ultimate aim of the assessment also determines the content of protocols. For example, the Welfare Quality® protocols aim at providing a comprehensive picture of welfare by addressing the four principles 'Good feeding', 'Good housing', 'Good health' and 'Appropriate behaviour' (see Table 1). They have initially been developed to increase transparency of production, to integrate animal welfare in the food quality chain and to promote welfare improvement (Blokhuis et al. 2010). Farm assurance schemes may, however, only target health-centred core problems of the industry as in the case of the US-based FARM programme, or take a slightly broader approach like the AssureWel protocols

which are now used by the UK-based Soil Association (organic farming) and RSPCA Assured (welfare label). The association of health-related welfare issues with productivity and thus economic benefits may further explain the focus on health-centred measures. Respective outcomes of the assessment can also be used for decision support in herd management (Winckler 2019). However, it remains unknown whether such health-centred approaches sufficiently take societal expectations into account, which typically relate to naturalness for dairy cattle (Beaver et al. 2020).

5 Collection of animal-based welfare measures on the farm

As initially explained, the number of welfare measures is large due to the multifactorial nature of welfare, that is, because many welfare aspects need to be covered. However, in addition, almost all welfare aspects can be assessed in different ways regarding the kind of measure and the method of recording. For instance, the prevalence and the severity of lameness in a herd can be ascertained by visual scoring of the walking gait of a sample of animals that are induced to walk a certain route (e.g. Welfare Quality® 2009) or by automatic means (reviewed by Alsaaod et al. 2019). Alternatively, claw trimming or treatments of lame cows can be documented in farm records or the assessments of claw condition can be carried out at the slaughterhouse, for example, in beef bulls (e.g. Magrin et al. 2019). In this section, the pros and cons of different methods of data acquisition are discussed with reference to performance criteria explained above.

5.1 Welfare measures obtained from direct observations or assessments on the farm

Currently, the direct observation or assessment of the animals on-farm is an important component of most welfare assessments (e.g. Welfare Quality® 2009). To obtain precise (i.e. reliable) data, the assessors must be well trained and inter-observer repeatability needs to be regularly tested (Knierim and Winckler 2009). However, repeatability testing is not profane. It needs a sufficient number of samples that are as independent as possible (as a rule of thumb it should be at least 10, but preferably more). For practical reasons, some compromises are usually made. For instance, when

the outcomes are prevalences of certain health conditions on a farm, one should ideally test inter-observer repeatability regarding 10 farm prevalences or more. In practice, one will rather check repeatability at animal level rather than farm level. Usually in this case, the number of observational units can be much higher than 10 animals which is advantageous in terms of power of the reliability testing. Further examples of compromises relate to behavioural observations of groups of animals, where several places in one house or different observation times may be used as observational units, even though they are not truly independent from each other. Finally, pictures (e.g. of integument alterations of various severity), or videos, (e.g. of social interactions) can be used to supplement reliability testing. Although some differences compared to live assessments can be expected, such a procedure has the advantage that a more complete range of conditions can be assessed than on a farm. Indeed, all severity scores of health conditions and all behavioural measures provided for in the assessment protocol should be covered by the testing. For statistical reasons, different scores should be represented as evenly as possible, and zero values in metric data (for instance no agonistic behaviour) should not be over-represented. To achieve this, considerable organisational efforts (often including the use of videos or pictures) and sufficient time are commonly required and must be planned for. Also in scientific work, proper repeatability testing needs to be extended (e.g. for behavioural observations: Kaufman and Rosenthal 2009, Burghardt et al. 2012).

Direct on-farm assessments are often limited by time and labour and thus financial constraints. This can affect how representative the recorded data are of the farm situation. Sample size is one of the difficult issues. In most cases, health conditions cannot be assessed in all animals, but only in a representative sample. For instance, Welfare Quality® (2009) recommends sampling for 73 animals in a herd of 300. Many claim that such a number usually overstrains available resources and disrupts farm routine unduly. However, others regard it insufficient for the reliable estimation of a prevalence. According to the formula provided by Cochran (1977), with 73 out 300 animals, the estimation may deviate ± 10% from the true prevalence (with 95% confidence level) when supposing a prevalence around 50% (if the true prevalence is higher or lower, precision will become higher, thus 50% represents the 'worst case').

This 10% precision should be taken into account when comparing and interpreting results. For behavioural observations, the duration of observation should be long enough to fairly represent activities over a longer period. For instance, Laister et al. (2009a) compared frequencies of agonistic interactions during 4 h and 2 h of observation time after feeding and concluded that due to high correlations ($r>0.8$) the shorter observation time would be sufficient. However, within short observation times it will not be possible to reliably record rare behaviours that occur unpredictably and unevenly over time (see for instance Plesch et al. 2010 for resting behaviour).

Another difficult issue refers to the stability of snapshot assessments or observations over time, that is, whether they are representative of the general farm situation. Seasonal influences, that is, from grazing, are to be expected. Moreover, short-term events such as routine claw-trimming of the herd, herd treatments, or recent regrouping or introduction of new animals to the herd can profoundly affect the outcome of a welfare assessment. Behavioural measures are especially sensitive to such changes and they can even be biased by the presence of an unfamiliar observer. By contrast, health measures are expected to reflect medium- to long-term influences and therefore to show less short-term fluctuations. However, when Winckler et al. (2007) assessed, among others, lameness, tarsal joint lesions and soiling of hindquarters in cows every two months on eight dairy farms, they found varying correlations between prevalences from 0.48 to 0.83. Only correlations between results from single assessments and the average of all assessments showed coefficients of at least 0.70 with only few exceptions. Also Kirchner et al. (2014), applying the Welfare Quality® assessment protocol on 63 beef bull farms repeatedly after one and six months, found only few consistencies regarding animal-based measures after one month and none after six months. As a general conclusion therefore, usually repeated measurements over time are necessary for a reliable assessment of the long-term welfare level on a farm.

A final reliability aspect is the avoidance of an expectation bias (Tuyttens et al. 2014). In direct observations blinding of the assessor is not possible. Automatic or slaughterhouse recordings have an advantage in this regard. In addition, their

implementation may be less labour-intensive. However, these methods also have their specific disadvantages as outlined below.

5.2 Welfare measures obtained from farm documentation or production records

Some health conditions such as certain infections require either diagnostic procedures – which are commonly not available in the context of an on-farm assessment – or longer observation periods because they occur infrequently. Farm records of treatment incidences are often proposed as measures (e.g. Whay et al. 2003). However, such data need to be used carefully. The decision when to treat a diseased animal may greatly vary between farms. A lack of treatment does not necessarily indicate the absence of welfare problems. Instead, for example, refraining from antibiotic treatment may be due to imminent slaughter, a culling strategy, or because of label restrictions. In addition, the quality of the farm records in terms of completeness, reliable diagnosis, and distinction between prophylactic and therapeutic treatments is often insufficient. Similar considerations apply to mortality or culling rates. The decision when to cull an animal relates only to some degree to the health or welfare state of the animal. Market considerations (current prices for milk or meat, changes in subsidy programmes), breeding or eradication programmes may greatly affect this decision. Some countries, namely Scandinavian countries and Austria, run centrally organised health databases which helps to increase reliability of documentation, but possible influences of factors not related to welfare still need to be considered when interpreting these data.

A useful source of information are measures from milk performance testing, such as test-day somatic cell count in milk (as indicator of mastitis) or fat-to-protein ratios (indicating risks of metabolic disorder). Again, care is needed in the choice of specific measures and their interpretation in terms of animal welfare. For example, temporary increases in cell counts may not always be disease-related.

Farm-routine data are owned by the farmer. Depending on the framework of the welfare assessment this may pose limitations. Such data are a priority source of

information to the farmers for self-assessment of animal welfare to improve the management of the farm. However, they may not be available for third-party auditing. Their external use needs the farmer's permission, and a very careful data handling is necessary in order to protect privacy of the owner. This similarly applies to automatically recorded data on farm and slaughter records.

5.3 Welfare measures automatically recorded on the farm

Technical progress increasingly enables the use of sensors and evaluation algorithms (e.g. using artificial intelligence) to monitor a great number of health conditions or behaviours (see Chapter 6). For instance, over the last two decades, a multitude of automated methods have been developed for the detection of lameness (reviewed by Alsaaod et al. 2019). They include image-processing techniques, use of accelerometers, pressure-sensitive walkways, ground reaction force systems, four-scale weighing platforms, auditory signal analysis and indirect methods such as thermography and use of behavioural measures like different measures of feeding behaviour, frequency of visits to automated milking systems or milk production, sometimes using a combination of different measures. Alsaaod et al. (2019) list 49 peer-reviewed validation studies on this topic. It is claimed that the automated recording of lameness may be more sensitive and less prone to observational bias compared to observation by a human (Rushen et al. 2012, Alsaaod et al. 2019). Nevertheless, its application in practice lags behind. This is likely due to a number of factors: it is a long and expensive process to reach a fully automated system that produces not only raw data, but also reliable evaluations (e.g. lame/not lame). Only part of the studies listed by Alsaaod et al. (2019) has reached this final stage. The resulting high initial price of such techniques reduces the readiness of farmers to adopt them. Moreover, to date the few commercially available solutions are not completely satisfactory with regard to, for example, sensitivity (Bicalho et al. 2007). Some of the techniques developed, for example, the use of accelerometers, are only partly suitable for everyday use on farms. The equipment mounted on the animal may affect their behaviour, and even impair their welfare by being uncomfortable or causing lesions. Limited battery capacities are another issue to be solved. Moreover, the devices may

get lost, which is especially problematic when they are expensive. If they shall be used for third-party assessments, extra efforts for mounting and removing the devices may render a direct assessment more time efficient. When data from farm-owned automatic monitoring systems are to be used for third-party auditing or benchmarking between farms, then the comparability of results from different systems needs to be ascertained. This may be difficult, because sensitivity and specificity vary between systems and are affected by individual farm factors. Among them are illumination levels (for image processing), cleanliness of sensors, background noise (for acoustic processing) or individual technical configuration. Truly, repeatability testing on each farm would be necessary. Often this issue is underestimated owing to the belief that automated measurements are 'objective', that is, not biased by human perception. This is often true, but potential bias by further external factors should not be overlooked.

A further advantage of automated systems is the increased possibility to include behavioural measures into on-farm welfare assessments, both on individual and group level (Rushen et al. 2012). This may relate to, for example, lying, feeding, play and other behaviours. In addition, long-term recording of activities indicative of disease, such as coughing in relation to pneumonia in calves, will provide more reliable results than short-term direct recording of coughing. The limited time available during direct observations is one reason why currently most animal welfare protocols show a clear dominance of health measures which could be overcome at least partly by adoption of automatic recording systems.

5.4 Welfare measures obtained from slaughter records

Disease diagnoses from the official or company control in slaughterhouses can be used as welfare measures for categories of cattle slaughtered in groups, that is, the prevalence of pneumonia in veal calves. In theory they are especially valuable, because each slaughtered animal is assessed by an expert at the slaughterhouse. However, currently there is insufficient training and reliability testing of the assessors, so that outcomes ascertained by different assessors and, even more so, in different slaughterhouses are hard to compare. Another disadvantage of welfare measures from

slaughter is their retrospective nature; they cannot be used to remedy welfare problems in the animals affected, but only for the next animals that will live on the farm. We nevertheless suggest that automatic recording of welfare measures at the slaughter line has great potential which is nowadays used in poultry but should be further developed in cattle.

6 Conclusion

Welfare is by nature multi-dimensional. As a consequence, there is no unique measure for animal welfare but rather sets of measures, especially related to health and behaviour, which are to be used to address the various aspects of welfare.

As for analytical methods, it is possible to validate these measures by characterising their performances in terms of selectivity, trueness, precision, and so on. Such a process was followed to design the Welfare Quality® protocols, which include measures for which the validity, reliability (precision and ruggedness) and feasibility were deemed adequate. We acknowledge that in some rare cases; there is at present no measure sufficiently validated so that the least objectionable measures available at present have to be used (e.g. access to water instead of addressing thirst) or a specific aspect is not addressed at present (e.g. thermal comfort).

Each way of recording welfare measures in practice has pros and cons; knowing them as well as the specific challenges of different methods allows an informed choice and helps to avoid possible pitfalls. Because advantages or disadvantages of methods depend on the specific measure to be recorded, often a combination of different methods is a good solution. Lastly, the development of sensors on farms opens possibilities to monitor continuously the health and behaviour of animals and may make animal welfare measurement more viable in the future.

7 Where to look for further information

At present, not all animal welfare criteria can be checked easily in field conditions. There is for example a need to develop and validate measures to check the absence of thirst and the thermal comfort of cattle. In addition, most welfare assessment

methods are focused on animals indoors. Cattle grazing outdoors may nevertheless be exposed to welfare hazards and their welfare should be checked adequately. Sensors offer opportunities to observe animals 24h/7, however, more work is needed to be able to comprehensively assess animal welfare using sensors.

There are numerous research teams working on animal welfare over the world. The Welfare Quality Network gathers partners from the former Welfare Quality project. It aims to maintain the Welfare Quality® protocols updated according to the most recent technical and scientific knowledge. The conference Assessment of Animal Welfare at Farm and Group Level (WAFL) is held every 3 years with the aim

8 Acknowledgements

Part of this work was supported by the Swedish Centre for Animal Welfare (Project 'ENCAW Pilot Study: Impact Assessment'). The work from Isabelle Veissier, coordinator of the present chapter, is supported by the Agence Nationale de la Recherche of the French government through the programme "Investissements d'Avenir' (16-IDEX-0001 CAP 20-25 2017).

9 References

1. Alsaaod, M., Fadul, M. and Steiner, A. 2019. Automatic lameness detection in cattle. Vet. J. 246, 35–44.

2. AWIN. 2015. AWIN Welfare Assessment Protocole for Horses.

3. Beaver, A., Proudfoot, K. L. and von Keyserlingk, M. A. G. 2020. Symposium review: Considerations for the future of dairy cattle housing: An animal welfare perspective. J. Dairy Sci. 103(6), 5746–5758.

4. Bicalho, R. C., Cheong, S. H., Cramer, G. and Guard, C. L. 2007. Association between a visual and an automated locomotion score in lactating Holstein cows. J. Dairy Sci. 90(7), 3294–3300.

5. Blokhuis, H., Veissier, I., Miele, M. and Jones, B. 2010. Welfare Quality® and beyond. Acta Agric. Scand. Sect. A Anim. Sci. 60, 129–140.

6. Boissy, A., Manteuffel, G., Jensen, M. B., Moe, R. O., Spruijt, B., Keeling, L. J., 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. Physiol. Behav. 92(3), 375–397.

7. Botreau, R., Veissier, I., Butterworth, A., Bracke, M. B. M. and Keeling, L. J. 2007. Definition of criteria for overall assessment of animal welfare. Anim. Welf. 16, 225–228.

Brambell, F. W. R. 1965. Report of the Technical Committee to enquire into the welfare of animals kept under intensive livestock husbandry systems. London: Command Paper 2836, Her Majesty's Stationery Office.

8. Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrison, K. E. and Stec, C. L., Zachau, C. E. and Freeberg, T. M. 2012. Perspectives – minimizing observer bias in behavioral studies: A review and recommendations. Ethology 118, 511–517.

9. Clegg, I., Borger-Turner, J. and Eskelinen, H. 2015. C-well: The development of a welfare assessment index for captive bottlenose dolphins (Tursiops truncatus). Anim. Welf. 24(3), 267–282.

10. Cochran, W. G. 1977. Sampling Techniques (3rd edn.), New York: Wiley & Sons.

Duncan, I. J. H. 2002. Poultry welfare: science or subjectivity? Br. Poult. Sci. 43, 643-52.

11. EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). 2012. Scientific opinion on the use of animal-based measures to assess welfare of dairy cows. EFSA J. 10, 2554 [2581 pp.]. Available at: www.efsa.europa.eu/efsajournal.

12. Farm Animal Welfare Council. 1992. FAWC updates the five freedoms. Vet. Rec. 17, 357.

13. Feinberg, M. 1996. La validation des méthodes d'analyse. Une approche chimiométrique de l'assurance qualité au laboratoire, Paris: Masson, 108–124.

14. Fleiss, J. L., Levin, B. and Paik, M. C. 2003. Statistical Methods for Rates and Proportions, Hoboken, NJ: John Wiles & Sons.

15. Fraser, D. 1995. Science, values and animal welfare: Exploring the 'inextricable connection'. Anim. Welf. 4, 103–117.

16. Fregonesi, J. A. and Leaver, J. D. 2001. Behaviour, performance and health indicators of welfare for dairy cows housed in strawyard or cubicle systems. Livest. Prod. Sci. 68(2–3), 205–216.

17. Kaufman, A. B. and Rosenthal, R. 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Anim. Behav. 78(6), 1487–1491.

18. Keeling, L. 2009. Defining a framework for developing assessment systems. In: An Overview of the Development of the Welfare Quality® Project Assessment Systems. In Welfare Quality Reports 12, pp. 1–7, Cardiff, UK: Cardiff University.

19. Kirchner, M. K., Westerath, H. S., Knierim, U., Tessitore, E., Cozzi, G. and Winckler, C. 2014. On-farm animal welfare assessment in beef bulls: Consistency over time of single measures and aggregated Welfare Quality((R)) scores. Animal 8, 461–469.

20. Knierim, U. and Winckler, C. 2009. On-farm welfare assessment in cattle: Validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. Anim. Welf. 18, 451–458.

21. Knowles, T. G., Warriss, P. D., Brown, S. N., Edwards, J. E. and Mitchell, M. A. 1995. Response of broilers to deprivation of food and water for 24 hours. Br. Vet. J. 151(2), 197–202.

22. Krohn, C. C. 1994. Behavior of dairy-cows kept in extensive (loose housing pasture) or intensive (tie stall) environments. III. Grooming, exploration and abnormal-behavior. Appl. Anim. Behav. Sci. 42(2), 73–86.

23. Laister, S., Brörkens, N., Lolli, S., Zucca, D., Knierim, U., Minero, M., Canali, E. and Winckler, C. 2009a. Reliability of measures of agonistic behaviour in dairy and beef cattle. In Forkman, B. and Keeling, L. (Eds), Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves, pp. 95–112, Cardiff, UK: Cardiff University.

24. Laister, S., Brörkens, N., Minero, M., Lolli, S., Zucca, D., Knierim, U., Canali, E. and Winckler, C. 2009b. Reliability of measures of socio-positive and play behaviour in dairy and beef cattle. In Forkman, B. and Keeling, L. (Eds), Assessment of Animal

Welfare Measures for Dairy Cattle, Beef Cattle and Veal Calves, pp. 175–188, Cardiff, UK: Cardiff University.

25. Lensink, B. J., van Reenen, C. G., Engel, B., Rodenburg, T. B. and Veissier, I. 2003. Repeatability and reliability of an approach test to determine calves' responsiveness to humans – A brief report. Appl. Anim. Behav. Sci. 83(4), 325–330.

26. Lidfors, L. and Isberg, L. 2003. Intersucking in dairy cattle – review and questionnaire. Appl. Anim. Behav. Sci. 80(3), 207–231.

27. Magnusson, B. and Örnemark, U. 2014. The fitness for purpose of analytical methods – A laboratory guide to method validation and related topics. Eurachem.

28. Magrin, L., Gottardo, F., Brscic, M., Contiero, B. and Cozzi, G. 2019. Health, behaviour and growth performance of Charolais and Limousin bulls fattened on different types of flooring. Animal 13(11), 2603–2611.

29. Mandel, R., Whay, H. R., Klement, E. and Nicol, C. J. 2016. Invited review: Environmental enrichment of dairy cows and calves in indoor housing. J. Dairy Sci. 99(3), 1695–1715.

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

31. Metz, J. H. M., Dijkstra, T., Franken, P. and Frankena, K. 2015. Development and application of a protocol to evaluate herd welfare in Dutch dairy farms. Livest. Sci. 180, 183–193.

32. Plesch, G., Broerkens, N., Laister, S., Winckler, C. and Knierim, U. 2010. Reliability and feasibility of selected measures concerning resting behaviour for the on-farm welfare assessment in dairy cows. Appl. Anim. Behav. Sci. 126(1–2), 19–26.

33. Pritchard, J. C., Barr, A. R. S. and Whay, H. R. 2006. Validity of a behavioural measure of heat stress and a skin tent test for dehydration in working horses and donkeys. Equine Vet. J. 38(5), 433–438.

34. Reenen, K. and Engel, B. 2004. Validation of welfare measures. Welfare Quality Meeting, Helsinki.

35. Rushen, J., Chapina, N. and de Passillé, A.-M. 2012. Automated monitoring of behavioural-based animal welfare indicators. Anim. Welf. 21(3), 339–350.

36. Scott, E. M., Nolan, A. M. and Fitzpatrick, J. L. 2001. Conceptual and methodological issues related to welfare assessment: A framework for measurement. Acta Agric. Scand. Sect. A Anim. Sci. Suppl. 30, 5–10.

37. Sorensen, J. T., Rousing, T., Moller, S. H., Bonde, M. and Hegelund, L. 2007. On-farm welfare assessment systems: What are the recording costs? Anim. Welf. 16, 237–239.

38. Špinka, M. and Dembele, I. Panamá, J. and Stìhulová, I. 2005. Lame dairy cows have shorter avoidance distances. In 39th International Congress of the International Society for Applied Ethology, Sagamihara, Japan, p. 83.

Stafleu, F. R., Grommers, F. J. and Vorstenbosch, J. 1996. Animal welfare: evolution and erosion of a moral concept. Anim. Welf. 5, 225-234.

39. Taverniers, I., De Loose, M. and Van Bockstaele, E. 2004. Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. Trends Anal. Chem. 23(8), 535–552.

40. Thompson, M., Ellison, S. L. R. and Wood, R. 2002. Harmonized guidelines for single-laboratory validation of methods of analysis – (IUPAC technical report). Pure Appl. Chem. 74(5), 835–855.

41. Tuyttens, F. A. M., de Graaf, S., Heerkens, J. L. T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E. and Ampe, B. 2014. Observer bias in animal behaviour research: Can we believe what we score, if we score what we believe? Anim. Behav. 90, 273–280.

Veissier, I. and Boissy, A. 2007. Stress and welfare: Two complementary concepts that are intrinsically related to the animal's point of view. Physiol Behav. 92, 429-33.

42. Vessman, J., Stefan, R. I., Van Staden, J. F., Danzer, K., Lindner, W., Burns, D. T., Fajgelj, A. and Müller, H. 2001. Selectivity in analytical chemistry (IUPAC recommendations 2001). Pure Appl. Chem. 73(8), 1381–1386.

43. Welfare Quality®. 2009. Welfare Quality® Assessment Protocol for Cattle (Fattening Cattle, Dairy Cows, Veal Calves). Welfare Quality® Consortium, Lelystad, The Netherlands .

44. 45. Whay, H. R., Main, D. C., Green, L. E. and Webster, A. J. 2003. Assessment of the welfare of dairy cattle using animal-based measurements: Direct observations and investigation of farm records. Vet. Rec. 153(7), 197–202.

46. Winckler, C. 2018. Assessment of cattle welfare: Approaches, goals and next steps on farms. In Tucker, C. B. (Ed.), Advances in Cattle Welfare, pp. 55–70, Duxford, UK: Woodhead Publishing.

47. Winckler, C. 2019. Assessing animal welfare at the f arm level: Do we care sufficiently about the individual? Anim. Welf. 28(1), 77–82.

48. Winckler, C., Brinkmann, J. and Glatz, J. 2007. Long-term consistency of selected animal-related welfare parameters in dairy farms. Anim. Welf. 16, 197–199.

Table 1 Welfare principles, criteria, individual measures and type of validity available for the measures in the Welfare Quality® assessment protocol for dairy cattle

WQ principle	WQ criterion	Welfare measure	Type of validity	Inter-observers repeatability[2]	Time needed to perform a measure
Good feeding	Absence of prolonged hunger	Body condition	Concurrent validity (culling risk, infertility)	++	3 min/animal (incl. assessment of cleanliness, integument alterations, lameness, nasal/ocular/vulvar discharge, signs of diarrhoea)
	Absence of prolonged thirst	*Resource-based (provision of water)*	Face validity	Good (cleanliness of trough) to very good	15 min/unit
Good housing	Comfort around resting	Time needed to lie down	Face validity (e.g. unpleasant experience of hard surface)	+++	150 min (incl. observations of agonistic behaviour)

Last authors version of : Book chapter in Marcia Endres (ed.), *Understanding the behaviour and improving the welfare of dairy cattle, burleigh dodds, 2021* ⟨10.19103/AS.2020.0084.05⟩. **Embargo until September 2 2021**

WQ principle	WQ criterion	Welfare measure	Type of validity	Inter-observers repeatability[2]	Time needed to perform a measure
		Animals lying partly/completely outside	Face validity (unpleasant experience, injuries, soiling)	+++	
		Cleanliness	Face validity (skin inflammation, itching)	+++	See absence of prolonged hunger
	Thermal comfort	No measure available yet	–	–	–
	Ease of movement	Resource-based (housing system)	Face validity	Very good	~15 min (interview, incl. somatic cell count, management procedures, pasture access)
Good health	Absence of injury	Lameness	Construct validity (analgesics lead to lower lameness score)	++	See absence of prolonged hunger
		Integument alterations	Face validity (pain)	++	
	Absence of disease	Somatic cell count	Concurrent validity (inflammation, pain associated with mastitis)	+++	See ase of movement

Last authors version of : Book chapter in Marcia Endres (ed.), *Understanding the behaviour and improving the welfare of dairy cattle, burleigh dodds, 2021* ⟨10.19103/AS.2020.0084.05⟩. **Embargo until September 2 2021**

WQ principle	WQ criterion	Welfare measure	Type of validity	Inter-observers repeatability[2]	Time needed to perform a measure
		Coughing, nasal discharge	Face validity (respiratory disorder)	++	See absence of prolonged hunger
		Vulvar discharge	Face validity (uterine inflammation)	++	
	Absence of pain induced by management procedures	Management-based (methods used for disbudding/dehorning, castration, tail docking)	Concurrent validity (behavioural and physiological indicators of pain)	Very good	See ease of movement
Appropriate behaviour	Expression of social behaviours	Agonistic behaviours such as head butts, displacements	Concurrent validity (unpleasant, stressful situations, risk of injuries)	+++	150 min (including observations of behaviour around resting)
	Expression of other behaviours	Resource-based (access to pasture)	Concurrent validity (positive effects on e.g. lameness, skin alterations, cleanliness, mortality)	Very good	See ease of movement
	Good human-animal relationship	Avoidance distance towards human	Construct validity (quality of handling)	+++	20 min

WQ principle	WQ criterion	Welfare measure	Type of validity	Inter-observers repeatability[2]	Time needed to perform a measure
	Positive emotional state	Qualitative behaviour assessment	Face, concurrent and construct validity[1] (overall impression/ association with quantitative measures of behaviour)	+/++	25 min

1: concurrent/construct validity for experimental setups and specific test situations only, not shown for on-farm assessment of groups of animals

2: +++: Kappa, Kendall's W, r > 0.8; ++: Kappa = 0.6–0.8, Kendall's W, r = 0.7–0.8; +: Kappa = 0.4–0.6

Table 2 Measures included in protocols for cows currently being used in the dairy industry: Welfare Quality® protocol for cattle (http://www.welfarequalitynetwork.net/network/45848/7/0/40); FARM, National Dairy Farmers Assuring Responsible Management Program (https://nationaldairyfarm.com/farm-animal-care-version-4-0/); AssureWel, Advancing Animal Welfare Assurance (http://www.assurewel.org/dairycows) (modified after Winckler 2018).

*	Measure	Welfare Quality	FARM	AssureWel
Physical appearance/health	Body condition	x	x	x
	Cleanliness	x		x
	Skin alterations	x[1]	x[2]	x[1]
	Broken tails		x	x
	Lameness	x	x	x

Last authors version of : Book chapter in Marcia Endres (ed.), *Understanding the behaviour and improving the welfare of dairy cattle, burleigh dodds, 2021* ⟨10.19103/AS.2020.0084.05⟩. **Embargo until September 2 2021**

*	Measure	Welfare Quality	FARM	AssureWel
	Mastitis	x		x
	Respiratory signs	x		
	Diarrhoea/loose faeces	x		
	Vulvar discharge	x		
	Downer cows	x		
	Animals needing further care			x
	Mortality: unplanned culls/casualties	x		x
Behaviour	Agonistic behaviours	x		
	Behaviour around resting	x[3]		
	Human-animal relationship	x[4]		x[5]
	Qualitative behaviour assessment	x		

[1] hair loss, lesions, swellings at one side of the animal
[2] hock/carpal joint injury
[3] time needed to lie down, lying partly/completely outside the lying area
[4] avoidance distance towards unknown person
[5] response to stockperson