

Weighting of different measurers in the assessment of farm animal welfare: a challenge

T. Johannsson¹, L. Alban² & P.F. Johnsen³

¹ The Agricultural University of Iceland, Hvanneyri, IS-³¹¹ Borgarnes, Iceland

² Danish Bacon & Meat Council, Axeltorv ³, DK-¹ 609 Copenhagen V, Denmark

³ The Royal Veterinary and Agricultural University, Groennegaardsvej 8, DK-¹ 870 Frederiksberg, Denmark

Summary

Weighting various welfare parameters represents a great challenge in the assessment of farm animal welfare at herd level. Weighting consist of two separate processes: first, all parameter measurements must be transformed into a common scale, where one end of the scale represents the best state of welfare and the other end the poorest state. This transformation involves determination of which levels of different parameters are equally important for the welfare of the considered animals. Parameter levels which are considered equally important belong to the same category on the scale. Another aspect of scales is whether they are numerical or ordinal and how many categories they have. The fewer categories the scale has, the fewer nuances is it capable of expressing. When numerical scales are applied it must be ensured that the distance between categories 1 and 2 is equal to the distance between 2 and 3 and so forth. This problem does not exist in ordinal scales.

When all Parameters have been transformed to a common scale, the results of each parameter are combined to an aggregated welfare assessment. The most important element of this combination is whether, and to what degree positive aspects can compensate for negative aspects of animal welfare. Existing welfare assessment systems propose three different approaches: no compensation, partial compensation, and full compensation.

In the paper, different techniques to construct a common scale are critically discussed as well as the question of how the method of compensation affects the results of the final welfare assessments.

Keywords: animal welfare, welfare assessments, scales, weighting, compensation.

Introduction

It is often claimed that it is impossible to compare different kinds of welfare parameters, such as teat tramping in dairy cows and high animal density. The widespread acceptance of this claim can be attributed partly to confusion about how to define animal welfare. Different definitions stress different aspects of welfare – e.g. a hedonistic definition of animal welfare favors the absence of pain, while a naturalistic view emphasis “natural” living conditions. However, the problem also persists within a given definition of welfare. Teat tramping causes brief, intense pain, while crowding involves a long-lasting, complex experience of social encounters. The former usually only affects a small fraction of a herd, while the latter affects the majority of cows in the herd. At first glance the two parameters seems so different that any comparison appears to be impossible. However, in this paper, we will argue that comparison is indeed possible and that it is necessary to do so in order to assess welfare at herd level. We will also discuss some important aspects of how this can be done and point out the biggest challenges in the weighting process.

Comparing the incomparable

Our first line of argument is that most people all the time make complex decisions about their own lives: whether to watch television or take a refreshing walk; how to spend their salary; and so on. Again, in many cases people make similar decisions on behalf of their children or friends, and the basic idea here, of an “enlightened emperor”, assumes that it is indeed possible for a person to make sophisticated overall evaluations of another person’s welfare.

If we accept that people are able to make valid comparison of different options for themselves as well as for their relatives and friend, then it seems plausible that this ability may, at least to a certain degree, be extended to farm animals.

Normally, when people make decisions to optimize their quality of life, they express what we can call “global preferences”. When, for example, evaluating the preferable aspects of seeing a good movie, we do not evaluate separately the value of the smell of the popcorn, the softness of the chairs, the sound effects and so on. We simply have a notion of going to the movies’ that we compare with other alternatives. However, in the case of animal welfare assessment, it can be advantageous to separate different aspects of the various situations we want to compare.

First, it must be emphasized that different parameters cannot be weighed against each other. Instead the focus must be on the comparison of individual measurements of the parameters. To take an example, it does not make sense to claim that animal density is more, or less, important for animal welfare than teat tramping. Obviously, there will be levels of teat tramping that are less harmful than certain levels of animal density and vice versa. The task, therefore, is to decide whether a certain levels of animal density is better, or worse, for the welfare status of herd than certain level of teat tramping. And this ought to be possible, given all available scientific data. On this approach, then, certain levels of teat tramping and animal density will be evaluated as equally important welfare determinants in a given herd of animals. Indeed, the aim will be to score every measurement of a parameter will receive a score, and this unified welfare score can be compared across parameters.

The scale does not need to be very detailed. The minimum is, of course, two, and indeed this is common in certification systems: a measurement of a parameter is either above or below a certain limit, e.g. “do the cows have access to exercise or not”. Information on animal welfare is considerably reduced when using only two levels of welfare and this reduction is rarely defensible. Terms like “outdoor access” may cover everything from 24 hrs access to pasture the whole year to a few hours access to a muddy exercise area. The upper limit of the number of levels is constrained by the ability of the designer to assist scores to measurements of parameters. It is our experience that it is difficult to distinguish between more than five levels and that further differentiation seldom is needed.

Defining these unified scores represents one of the biggest challenges in the whole process of welfare assessment. The immediate approach is to start by defining the endpoints of the scale and then score the intermediate levels. If we consider “body conditioning score” (BCS), a trained person would almost certainly have a fairly clear notion of “a very skinny cow” and “a very fat cow”. Somewhere in between, there would be “a normal cow”. Thus we have score 1,3 and 5. Scores 2 and 4 cover the areas not already covered by scores 1, 3 and 5.

However, there are two serious problems with this approach. First, the welfare of the cow can be threatened, both by very low and very high BCS, and the scale does not indicate whether score 1 is better or worse than score 5. This, the score must be transformed to a scale where such a distinction is possible. The second problem is more challenging. Even though we can rate a single parameter, like BCS, on a 1-5 scale, it does not follow that it can be compared to another parameter; say “cleanliness score”, also graded on a 1-5 scale. That would only be possible, if the two parameters are equally important for the welfare of the

animals, that is, if any of the five scores for “body condition” is comparable to the corresponding five scores of “cleanliness”. However, it is rather obvious that the lowest possible level of cleanliness will never be comparable to the lowest possible level of body condition, in terms of animal welfare. Therefore, the scale must be weighed. At first glance, it might seem advantageous to multiply the score for each parameter with an appropriate weight. If we find “cleanliness” only half as important as “body condition” we multiply the scores for “body condition” we multiply the scores for “body condition” with 2, before we continue with the calculations. However, this approach implies that any measurement of “body condition” is twice as important as any corresponding measurement of “cleanliness”. Let’s say an animal has the second best score in both parameters, this approach would claim, that the relatively good “body condition” contributes twice as much to the animal welfare, as it being relatively clean. We have doubts whether this is defensible.

It appears to us, that the most appropriate method for weighting is to use different parts of the scale for different parameters. Given a 1-5 scale where 1 is worst and 5 best, “body condition” may be graded from 1 to 5 where as “cleanliness” only goes from 3 to 5. Thus, on some levels (3,4 & 5) the parameters may be compared, but “cleanliness” can never influence welfare as seriously as “body condition”. This approach has been chosen for example in the Austrian TGI 35 L (Bartussek, 1999) and the German TGI 200 (Sundrum et al., 1994).

Given the general framework of techniques for weighting, the problem still remains how to decide which parameter levels are in fact equally important for animal welfare? The common answer is that we must rely on the principles of analogous (Stafley, et al., 1992) and homologous (Sandoe & Simonsen, 1992) comparison, when possible and of course use all available data about the state of the animals as well as the environment (Bracke, et al., 1999). In our experience, however, these general guidelines are of limited help. The English philosopher Jeremy Bentham (1748-1832) proposed a more analytical approach. He suggested, that the importance of a parameter measurement (which he called “circumstances”) could be evaluated according to the following seven “attributes” (Bentham, 1970):

1. Intensity
2. Duration
3. Certainty or uncertainty
4. Propinquity or remoteness
5. Fecundity
6. Purity
7. Extent

The first four elements relate exclusively to the pleasure or pain caused by the “circumstance” itself. The Fifth and sixth elements refer to the probability of the pleasure or pain being followed by other similar or dissimilar sensations, and the seventh element brings into the equation the number of individuals that are affected. Elements 3 (at least in part), 4, 5 and 6 are of no real importance if the welfare of the herd is only to be assessed for a defined period of time. Instead the concept of “frequency” should be used to evaluate how often certain “circumstances” take place within a defined period.

This approach has been applied for evaluating the welfare impact of various diseases in dairy cows and pigs (Alban, 1995). Here, the welfare impact of a given disorder was quantified in terms of the combination of “the intensity of the pain caused by the disorder”, “it’s duration” and “it’s prevalence in a herd”. The assessment of the “intensity of the pain” was based on clinical, veterinary knowledge of how the animals were affected by the various disorders.

Various scientific disciplines may find it hard to agree on how claim that it is only possible to assess and monitor animal welfare at herd level if the different indicators of welfare are compared on a common scale.

Ordinal or numerical scales

Johnsen et al. (2000) reviewed nine European assessment systems and found that there exist many types of scales in welfare assessment. As discussed before, it is preferable that one end of the scale represents the best state of welfare, whereas the other end represents the poorest state of welfare, which, in statistical terms, means that the scales must be ordinal-, interval- or ratio scales. In ordinal scales (e.g. A, B, C, D where A is best and D is worst), any measurement in level B, is worse than any measurement in level A, and better than measurements in both C and D, and so on. However, it is not possible to determine how much B differs from A or C. If the measurements of parameters cannot be fitted onto an ordinal scale, they cannot, per definition, be used to assess animal welfare. An example of a parameter is the mean BCS for a herd of animals. All we can do is to calculate the proportion of animals that receive each score.

In interval scales (e.g. -2, -1, 0, 1, 2 where -2 is worst and 2 is best) it is additionally possible to give a meaningful measure of the distance between levels. This, in any measurement in level -1 is equally far from any measurement in level -2 as to any measurement in level 0. Consequently, the scores for several parameters can be added and subtracted.

Ratio scales differ from interval scales in that they have a true zero point, below which no measurements can be made (e.g. 0, 1, 2, 3, 4, where 0 is worst and 4 is best). This means that in addition to the distance between different levels, it is possible to state that every measurement in level 2, is twice as good as any measurement in level 1, and half as good as any measurement in level 4. Consequently, when interval- or ratio scales are chosen it is an extremely complex process to determine which score are to be given to the various parameter measurements.

It is obvious, that as long as no accepted “welfare unit” exist, it is not statistically valid to assess welfare by other scales than ordinal. There is, however, some tradition for treating ordinal data as if they were numerical. This is commonly done in BCS and has also been done for evaluation of behavioral responses of cows towards humans (Munksgaard et al., 1997). Moreover both TGI 35 L and TGI 200 use these apparently invalid statistics. However, it may be argued that in some cases, the statistical errors caused by treating truly ordinal variables as numerical, are outweighed by the advantages this gives. We suggest that interval scales should be used to compare various parameter measurements in a welfare assessment. Thus, the scale can be extended in both ends, but still provides a measure of the relative weight of each measurement.

Compensation

Once all measurements in a welfare assessment protocol has been transposed to a common scale, the next task is to combine the different measurements into an aggregated welfare assessment for the entire herd. This includes the question of compensation, i.e. whether, and to what extent positive factors can compensate for negative welfare reducing factors. In methods for assessment of animal welfare at herd level there appear to be three levels of compensation: *no* compensation, *partial* compensation and *full* compensation (Johnsen et al., 2000).

No compensation

Most certification standards for animal welfare do not allow any compensation between parameters. Consequently, a predefined minimum level must be attained for every parameter included in the assessment in order to reach a certain standard of welfare, (see e.g. R.S.P.C.A. Freedom Food Scheme (RSPCA, 1998)). The results of assessment are usually

presented on a binomial scale – either the herd meets the requirements or it does not – however, technically a scale with several steps could be applied.

The advantage of not allowing any compensation between parameters is that it results in a clear cut conclusion, which is easy to communicate to non-specialists (farmers, consumers, decision-makers etc.). The welfare status is either above or below a predefined level of welfare. If above, the level of welfare is acceptable, if below it is unacceptable. The disadvantage is that the method is crude. A single parameter may alter the final result from being “acceptable” to “unacceptable”. Moreover, and more important, the minimum level of welfare may easily become the maximum level, as there is no incentive for the producers to improve the welfare of the herd beyond the minimum level.

Partial compensation

The term “partial compensation” refers to the situation where a high value of one parameter can compensate for a low value of parameters, which belong to the same category of influence. Typically, the parameters are grouped with respect to the areas they influence, e.g. social behavior, resting or feeding, and compensation can only occur between parameters within the same group.

An example of this method is found in welfare assessment system developed in France by Capdeville & Veissier (1999). According to this method, compensation can take place within five predefined groups of parameters. These groups are based on the well-known concept of “five freedoms” (Brambell, 1965) and are selected to reflect certain “criteria” of the welfare in the herd.

The advantage with partial compensation is that related parameters are allowed to compensate for each other. The disadvantage, however, is that it is very difficult to define the categories, within which compensation can occur. If compensation only takes place between parameters within certain groups, the content of the groups becomes very important.

To take an example, three parameters, “softness of the floor”, “amount of straw used”, and “access to pasture”, are used to describe the resting situation for a cow in a tie stall. The question is, whether the effect of “hard floor” is equally compensated for by applying suitable amount of straw and by access to pasture. Most observers would probably claim that adequate straw is a better compensation for a hard floor than access to pasture, since the presence of straw directly alters the hardness of the floor (the floor and straw parameters overlap). However, it can also be argued that access to pasture improves leg health and thus the cow may rest more comfortably. In this case, the compensatory capacity of the parameters appears to vary. Additionally, the same parameters might belong to more than one category, and as a consequence gain more weight compared to other parameters.

Full compensation

The Austrian TGI 35L (Bartussek, 1999) and German TGI 200 (Sundrum et al., 1994) are example of so-called index systems for welfare assessment at herd level. They allow compensation between all types of parameters. These systems divide the parameters into categories, five and seven, respectively, and separate sub-totals are calculated for each category. The sun-totals, however, are added into a final welfare score for the herd, and consequently full compensation takes place. The strength of this method is that it is possible to calculate an overall welfare score, which can be compared across farms and over time. However, the method relies on the convenient, but highly doubtful assumption that high level of any parameter can compensate for low level of any other parameter.

Whether and how compensation is restricted within an assessment protocol is a matter of great significance. Overlapping parameters need to be identified and prevented from giving additive effects.

The effect of compensation method

To illustrate how the above discussion compensation methods may alter the results of welfare assessment, the following example was constructed, based on welfare assessment of dairy cows in two cubicle barns. For the sake of the illustration, only five parameters, belonging to two categories are shown. It is assumed that all parameters can be evaluated on a scale from 1 to 5 where 5 represent the most favorable level of animal welfare. It is also assumed, that score 1 and 2 reflect circumstances that are not acceptable as “welfare friendly”.

	<i>Resting</i>		<i>Social conditions</i>		
	Dimensions of lying area	Softness of lying area	Animals per feeding site	Animal density in walking area	Regrouping
	1=restrictive	1=hard	1=many	1=high	1=frequent
	5=non-restrictive	5=soft	5=few	5=low	5=rare
Herd A	2	2	3	3	5
Herd B	3	3	5	1	3
<i>Results</i>	No compensation	Partial Compensation		Full compensation	
Herd A	Not acceptable standard of animal welfare	Not acceptable standard of animal welfare		Acceptable standard of animal welfare	
Herd B	Not acceptable standard of animal welfare	Acceptable standard of animal welfare		Acceptable standard of animal welfare	

If no compensation was allowed, neither herd would reach acceptable standards of animal welfare. In Herd A, two parameters fails in the category “resting” but Herd B one parameter is below 3. Herd A fails in the category “resting” but Herd B in “Social conditions”.

If, on the other hand, partial compensation was used, we could calculate means for the two categories: In Herd A, “resting” has mean of 2 and “social” receives 3, 7. The herd does therefore not reach acceptable level. In Herd B, the mean for both “resting” and “social” is 3, which is acceptable, even though one of the parameters scores only 1.

Finally, given full compensation, the overall means for the two herds are equal; both have 3 points. Consequently, both have acceptable standards of welfare.

The example illustrates, that depending on the method of compensation, we might conclude that neither of the herds, one of them, or both, have an acceptable standards of animal welfare.

Concluding remarks

Welfare assessment on herd level is practiced in several European countries and several research projects on the subject are current in progress (Johnsen et al., 2000). One of the main issues in welfare assessment is whether and how different measurements of welfare should be weighed against each other. We have argued, that a true assessment of animal welfare, for not to speak of monitoring and comparison, is impossible if no weighting takes place. We have also suggested a rudimentary set of principles for weighting different measurements and discussed advantages and disadvantages of different types of scale.

Finally, we have touched upon the term “compensation between measurements” and demonstrated that different methods of compensation may have considerable effect on the interpretation of herd recordings. Future work includes development of methods to identify

compensation measurements and quantify the level of compensation in each case. To our knowledge, this has not yet been done in any methods of welfare assessment systems and is therefore a major task in the future development of welfare assessment at herd level.