The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2019, 28: 67-76
ISSN 0962-7286
doi: 10.7120/09627286.28.1.067

Aggregating animal welfare indicators: can it be done in a transparent and ethically robust way?

P Sandøe*†‡, SA Corr§, TB Lund‡ and B Forkman†

† Department of Veterinary and Animal Sciences, University of Copenhagen, Grønnegårdsvej 8, 1870 Frederiksberg C, Denmark
‡ Department of Food and Resource Economics, University of Copenhagen, Rolighedsvej 25, 1958 Frederiksberg C, Denmark
§ School of Veterinary Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow G61 1QH, UK
* Contact for correspondence and requests for reprints: pes@sund.ku.dk

Abstract

A central aim of animal welfare science is to be able to compare the effects of different ways of keeping, managing or treating animals based on welfare indicators. A system to aggregate the different indicators is therefore needed. However, developing such a system gives rise to serious challenges. Here, we focus specifically on the ethical aspects of this problem, taking as our starting point the ambitious efforts to set up an aggregation system within the project Welfare Quality® (WQ). We first consider the distinction between intra- and inter-individual aggregation. These are of a very different nature, with inter-individual aggregation potentially giving rise to much more serious ethical disagreement than intra-individual aggregation. Secondly, we look at the idea of aggregation with a focus on how to compare different levels and sorts of welfare problems. Here, we conclude that animal welfare should not be understood as a simple additive function of negative or positive states. We also conclude that there are significant differences in the perceived validity and importance of different kinds of welfare indicators. Based on this, we evaluate how aggregation is undertaken in WQ. The main conclusion of this discussion is that the WQ system lacks transparency, allows important problems to be covered up, and has severe shortcomings when it comes to the role assigned to experts. These shortcomings may have serious consequences for animal welfare when the WQ scheme at farm or group level is applied. We conclude by suggesting ways to overcome some of these shortcomings.

Keywords: *aggregation, animal welfare, ethics, expert opinion, farm animals, Welfare Quality®*

Introduction

A key aim of animal welfare science is to be able to compare the effects of different ways of keeping, managing or treating animals. Typically, groups of animals, or the same flock or herd of animals at different times, are exposed to different forms of housing, management procedures or other treatments. A number of indicators may be used to measure welfare between groups of animals or in the same group or individual over time. Sometimes it is straightforward to add up things to be able to rank different forms of housing, management procedures or other treatments in terms of animal welfare outcome. Other times it may be more difficult, due to different welfare indicators pointing towards different welfare outcomes.

Traditionally, farm animal welfare research has focused on applying single welfare indicators, often in an experimental setting; and therefore the issue of aggregation has largely been avoided. This has changed gradually since the 1990s, beginning with the development of systems for assessing welfare impact on laboratory animals (eg Porter 1992; Stafleu *et al* 1999). Since around 2000, initiatives have been developed to assess farm animal welfare at group level, which have given rise to more systematic discussions about how to aggregate different welfare indicators (cf, for example, Capdeville & Veissier 2001; Spoolder *et al* 2003). These efforts have so far culminated in Welfare Quality® (WQ), a large project funded by the EU Commission that developed protocols to measure the welfare of cattle, pigs and hens at farm level (for further information, see Keeling 2009).

The WQ protocols take as their starting point a comprehensive definition of farm animal welfare in terms of four principles: Good feeding; Good housing; Good health; and Appropriate behaviour. These are subdivided into 12 welfare criteria. Each criterion is measured by a number of indicators that are dependent on the type of animal being studied. In the case of dairy cows, for example, there are 31 indicators, primarily focusing on the states of the animals themselves (so-called animal-based indicators), rather than the resources provided to the animals. So-called environmental-based indicators (relating to availability of resources) are only used when animal-based indicators are not available or are deemed less feasible or reliable, eg in the case of thirst, where availability of drinkers is used as a proxy indicator.

WQ enables all farms or groups of animals within the farm covered by the protocols to be grouped into four categories: 'Excellent', 'Enhanced', 'Acceptable', and 'Not-classified'. Assignment to one of the four categories is determined using a comprehensive dataset relating to the many indicators that serve as the basis of the system. Specifically, the results for up to 31 different indicators are transformed through four stages of aggregation into one of the four aforementioned categories.

The aggregation procedures used in WQ have caused some controversy in the academic literature about whether the outcome of WQ aggregations is in line with expert opinion (de Graaf *et al* 2017). However, little focus has so far been put on the ethical aspects of aggregation, eg on whether it is acceptable to simply add up welfare states, not only within the life of the individual animal, but also across animals. Researchers in the WQ team have had some focus on what they call the problem of 'compensation' (Veissier *et al* 2009) which, roughly speaking, is about whether a bad thing happening to one or more animals may be offset by avoiding other bad things happening, but little has been done to unpack and analyse this and other ethical issues potentially raised by aggregation. To do this is a main aim of this paper.

In the paper, we first consider the idea of aggregation in light of the distinction between intra- and inter-individual aggregation. Secondly, we look at the idea of aggregation in terms of comparing different levels and sorts of welfare problems. Based on this, finally, we critically review how aggregation is undertaken in WQ.

Intra- and inter-individual aggregation: ethical implications

Life involves many choices that result not only in good things, but different mixes of good and bad things. This is also the case when society makes choices on behalf of animals. These choices, in our view, fall in two distinct groups. The first group consists of choices where only one animal is involved, and the task is to balance the potential good and bad things to achieve the best possible outcome for that particular individual. This kind of choice requires what we will call *intra-individual* aggregation. The other group consists of choices where a group of animals is involved and the task is to find out what is best for the group. This gives rise to what we call *inter-individual* aggregation.

Two examples may serve to illustrate what each involves:

Suppose a dog is suffering from cruciate ligament disease, a common canine disease in which the ligaments stabilising the knee deteriorate and rupture. This condition can be managed either conservatively (ie non-surgically) or with surgery. With conservative management, most dogs will improve to some degree if their exercise is restricted for several months (Wurcherer *et al* 2013) but will typically end up with variable lameness, a thickened knee, and moderate to severe osteoarthritis in the joint. There are many surgical treatment options, of which a TPLO procedure (involving cutting and rotating the bone and then stabilising it with a bone plate) is considered to be one of the best, based on current scientific evidence (Krotscheck *et al* 2016). The dog will usually be discharged the day after surgery on painkillers, and it will be exercise-restricted for about six weeks, at which stage it will be sedated and have x-rays taken to assess bone healing. If everything is healing well, the dog will gradually be returned to full exercise over the following six weeks, and can be expected to have minimal lameness, and minimal osteoarthritis, for the rest of its life.

When deciding on this case, assuming that cost is not an issue, the dog owner, assisted by the veterinarian, will have weighed up the different good and bad things for the dog in the two possible ways of dealing with the disease. This may not be easy for several reasons. Firstly, the outcome of each decision is uncertain. With conservative therapy the dog may end up more or less lame and it may or may not develop severe osteoarthritis in its knee. With surgery, severe complications may occur, even though they are rare. Secondly, there may be uncertainty about how much the different issues facing the dog in each scenario will matter from the point of view of the dog. Will the stay at the veterinary surgery away from the owner be traumatic? How much will a certain degree of lameness matter to the dog?

So, when considering this case there is, of course, some uncertainty about the outcome of two treatment choices. This uncertainty mainly concerns the actual outcome as just described, but there may also be an ethical twist to this. Thus, it may be argued that the surgical option simultaneously presents the prospect of the best possible long-term outcome for the dog, but also the greatest risk of severe complications. So, depending on whether one focuses on the 'most likely', or on the 'worst possible' outcome, ie on whether one is a risk-taker or is risk-averse, different conclusions may be reached (and different treatments chosen).

Another factor that may affect one's decision is how one weighs the different aspects of dog welfare involved. In both cases, there will be some element of pain and discomfort, but in the case of surgery the dog may also experience fear and loneliness when taken to the veterinarian and separated from the owner. Depending on how much weight is put on preventing physical discomfort and pain compared to other aspects of psychological well-being, the two decisions may be viewed very differently. (More on this in the following section).

The second example concerns the decision on whether to tail-dock a group of piglets that are going to live their lives as slaughter pigs in an intensive pig production unit. Tail-docking typically involves cutting off 30–50% of the tail of newborn piglets to prevent tail-biting. It is widely agreed among researchers and other experts in the field that although tail-docking does not completely prevent tail-biting, it significantly reduces the number of pigs that have their tails bitten (eg Valros & Heinonen 2015). While tail-docking involves an unpleasant procedure causing some pain and discomfort, the welfare problems experienced by animals that are severely tail-bitten are clearly much greater, not least since bitten tails often become infected. When farmers consider whether to tail dock their piglets

https://doi.org/10.7120/09627286.28.1.067 Published online by Cambridge University Press

Figure 1

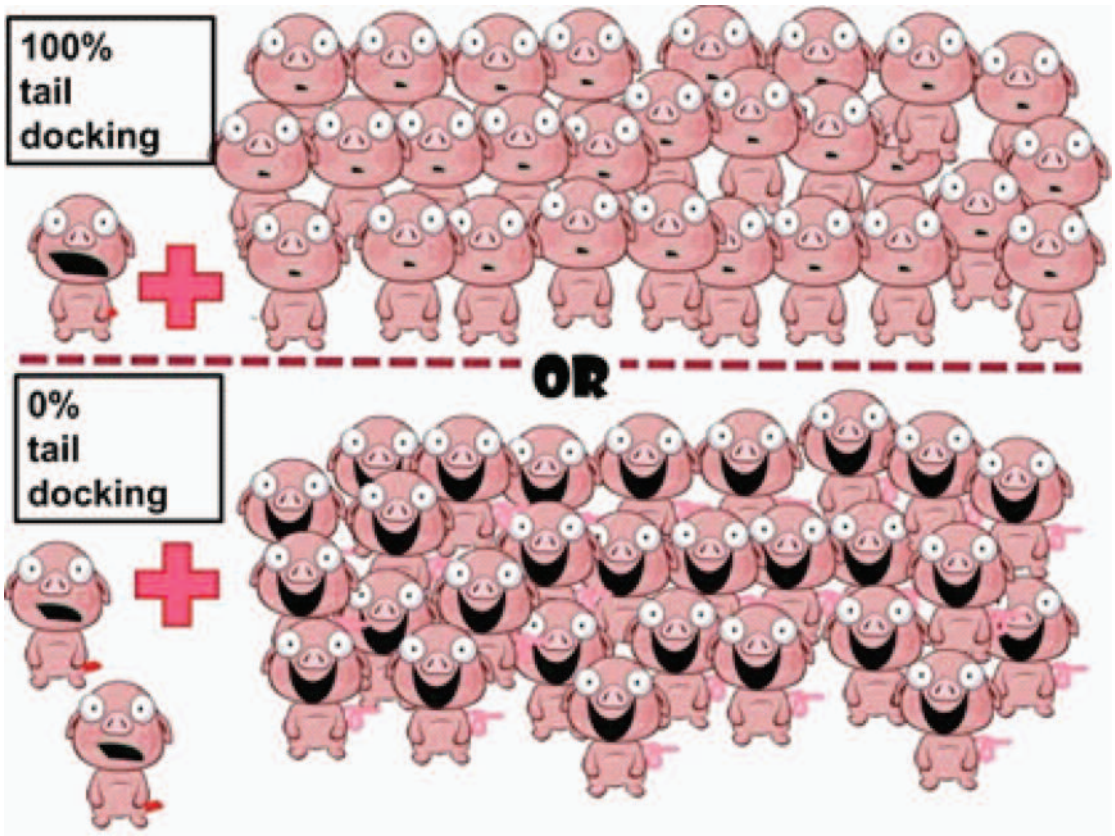

Illustration of the challenge of inter-individual aggregation. Theoretical illustration of the result of a 100 versus a 0% tail-docking policy. Tail-docking causes some pain to 100% of the pigs and reduces the risk of tail-biting (approximately two-fold, based on available abattoir data). Even though the risk for tail-biting might be higher if tails are not docked, and the pain caused by biting more intense than the pain caused by docking, the non-bitten pigs are fully spared the pain due to tail-docking in a 0% docking scenario. Pigs that are both docked and bitten suffer the most pain. (Figure and part of original caption re-printed from Valros and Heinonen (2015), published by BioMed Central, with kind permission from A Valros).

they are not, or not solely, driven by a concern for the welfare of their animals — economic considerations also play a role, since tail-biting may give rise to serious infections which lead to carcase condemnation and in other ways have a negative effect on productivity. Therefore, farmers may have a strong economic incentive to tail dock. However, here, we will try to consider the issue only from the point of view of the pig.

In one way, this case is similar to the previously discussed dog case, since both compare physical pain and discomfort. However, in another way it is much more complicated, since it is about distributing welfare outcomes among different animals in a group, rather than just securing the best welfare for one individual. It is about inter-individual aggregation where the complicating factor is that different consequences befall different individuals, subsequently raising questions about fairness. The situation may be illustrated by means of Figure 1.

Figure 1 displays a herd of around 30 pigs. In the situation depicted at the bottom, there is no tail-docking, with the result that we have 28 pigs with no negative welfare consequences related to tail-docking and two pigs with a severe welfare problem due to tail bite. In the situation depicted at the top, all pigs are tail-docked with the result that one pig is spared the consequences of tail-biting, but all the pigs have to endure a lesser amount of pain and discomfort. The question then is how to aggregate across the different individuals in each situation. Which situation is preferable from the point of view of the involved pigs? (The choice described here is deliberately simplified). A further complicating factor that should be included in real life is how being allowed to tail dock will affect the way the farmer cares for his animals. Valros and Heinonen (2015) claim that the farmer, if she or he is not able to tail dock, will take better care of the pigs to try to limit tail-biting. However, D'Eath and colleagues (2014) express a more pessimistic view on this.

Some moral philosophers think that the intra- and inter-individual aggregation cases are in principle not that different. A case in point is RM Hare, the leading British moral philosopher in the post Second World War period. In his book, *Moral Thinking* (1981), he claims that aggregation across individuals can be reduced to aggregation within the life of one individual. To find out which situation is morally preferable, therefore, one would have to go through the

Table 1 Distribution of gait scores of Danish broilers at three points in time.

Year/Gait score	Gait score 0 (%)	Gait score 1 (%)	Gait score 2 (%)	Gait score 3 (%)	Gait score 4 (%)	Gait score 5 (%)
1998/99 (A)	25.0	20.6	24.3	24.3	5.5	0.3
2004/05 (B)	19.5	33.2	34.1	12.6	0.5	0.0
2011 (C)	0.6	12.2	83.3	3.7	0.2	0.0

Distribution of gait scores (percentages of flock level prevalence) measured on Danish broiler farms in 1998/99 (Sanotra *et al* 2001), in 2004/05 (Petersen 2006), and in 2011 (Rasmussen & Spangbjerg 2012).

following thought experiment: imagine that you will live the life of all the 30 involved pigs, one after the other. Then ask yourself the following question: which of the two alternatives would I prefer, given that I am going to live the lives of all the affected pigs? Here, it is likely that 0% tail-docking would be preferable: even though there will be extra miserable moments, these could plausibly be outweighed by a better quality of life in the vast majority of imagined consecutive lives. The overall idea of Hare's position is that when aggregating across individuals one should aim to get the greatest total sum of welfare, which is in accordance with the ethical position called utilitarianism.

Others, like the American political philosopher, John Rawls, would object to this line of thought. He would claim that a difference of crucial importance is being overlooked: the difference between individuals. When aggregation is done within one life, all the costs and the overall net benefit fall on the same individual. This seems fair. But when aggregation is carried out across individuals, the benefits and costs fall on different individuals. Here, there will be winners and losers, and if the winner takes it all, the loser will get an unfair deal. Thus, Rawls famously complains that utilitarianism "adopt[s] for society as a whole the principle of choice for one man" (Rawls 1971; p 24). (Rawls only considered the issue of aggregation in relation to humans and he would not, for reasons that we will not discuss here, have sympathised with the vastly simplified extension of his views presented here to cover animals).

Rawls' alternative to the utilitarian view of aggregation across individuals, which favours the "greatest sum of welfare across individuals", is the famous "difference principle" according to which the aim is to secure the "greatest benefit of the least advantaged" (Rawls 1993; p 6). Given the choice between 0 or 100% tail-docking of piglets, it can be argued that tail-docking is the preferable option since it will improve the situation of half of the least advantaged, ie the pigs that would otherwise have had their tails bitten (based on the assumption that tail-biting gives rise to problems of an order of magnitude much greater than tail-docking would have done for the bitten pigs).

The utilitarian principle that aggregation should be done to secure the greatest sum of welfare across individuals, and the difference principle, according to which aggregation should be done to achieve the greatest advantage for the most disadvantaged individuals, can be viewed as two extremes on a scale where numerous middle positions are conceivable. One such middle position suggested by the British moral philosopher Derek Parfit is the so-called priority view, according to which aggregation should proceed with extra weight being given to worse-off individuals (Parfit 1997; Arneson 2000). Thus, the priority view reflects a middle ground, giving special weight to the plight of the worst off (like the difference principle), yet still considering the plight of the better off, so that all individuals count (as in utilitarianism) in the aggregation process.

So, inter- compared to intra-individual aggregation does give rise to a new level of moral complexity concerning how to balance welfare costs and benefits between affected individuals. If consequences are simply added up this will reflect a utilitarian view on aggregation, which in many ways is morally controversial. Other ways of adding up welfare consequences across individuals, including those found in WQ, will also potentially give rise to controversies. Thus, there is no morally neutral way of aggregating welfare across individuals, and the way forward must therefore involve complete openness about how aggregation is done, on the moral implications, and about alternative ways of aggregating.

Adding up different kinds of problems: ethical implications

Welfare issues are diverse — they may be of different kinds and come in different degrees of intensity. How to aggregate across different welfare states and degrees of intensity may give rise to ethical challenges, irrespective of whether the issue is intra- or inter-individual aggregation. For example, it will require ethical thinking to consider how suffering pain compares to the lack of ability to express natural behaviour. Likewise, it may, for example, be an ethical issue to compare mild pain to severe pain, since it may be argued that low levels of pain does not lead to suffering. (For a more full discussion of different conceptions of animal welfare, see Weary & Robbins 2019; this issue).

Beginning with the latter issue: scales on which things like pain are measured are ordered, ranging from no pain to very severe pain. However, the way pain matters is not necessarily similar to the underlying scale. Thus, if pain is scored on a scale from 0 to 10 it does not follow that ten days with pain at level 1 are equivalent to one day with pain at level 10, or that 10 animals with pain score 1 are equivalent to one animal with pain score 10. A real-life example may serve to illustrate this: the 'gait score' of broilers, used as an indicator of pain and discomfort due to leg problems. Leg

https://doi.org/10.7120/09627286.28.1.067 Published online by Cambridge University Press

problems are affected by a number of factors, the most important of which is the very rapid growth rate for which the birds have been genetically selected over many generations. Gait score is measured on a scale from 0 to 5, where 0 is given to birds that walk perfectly and 5 is given to birds unable to walk. Gait scores were measured in Danish broiler flocks in 1998/99, 2004/05 and in 2011 (see Table 1).

In 1998/99, the distribution of gait scores of Danish broilers (situation A, the specific combination of genotype and management at that time) showed a relatively high proportion of birds with gait scores of 3, 4 and 5. In 2004/05, mainly due to changes in genotype (situation B), a much lower proportion of gait scores 3, 4 and 5 was reported together with a higher proportion of gait scores 1 and 2. In 2011, the distribution changed again due largely to a changed genotype (situation C), resulting in very few birds with gait scores 3, 4 and 5 and also with gait scores 0 and 1. So, how should this be assessed? Should C be classified as worst, since here the fewest birds have very good gait scores (0 and 1)? Or should A be classified as worst, since the highest number of poor gait scores (3, 4, and 5) are seen here?

While there has been a consistent reduction in the % of high gait scores (3, 4, and 5) from A to B to C, the number of birds with gait score 2 has increased by nearly 60% (from 24.3% at A to 83.3% at C). Gait score 4, for example, represents a bird that has a severe gait defect and therefore may be expected to be in severe pain. The bird is capable of walking, but only with difficulty and when driven or strongly motivated. Otherwise it squats down at the first available opportunity (Kestin *et al* 1992). A decrease from 5.5 to 0.2% in gait score 4 therefore means far fewer birds are likely to be suffering severe pain. In contrast, gait score 2 represents a bird with a gait abnormality that does not affect its ability to move, and as such might be associated with relatively mild pain, but in many more birds.

Although this example is similar to the tail-bite case discussed in the previous section, considering a larger number of individuals in mild pain versus a few individuals in severe pain, the point we want to make is a different one. It concerns what counts as suffering. The birds with gait scores 3, 4 and 5 clearly seem to be suffering (Caplen *et al* 2014). If they are given a painkiller, many of them will walk normally (McGeown *et al* 1999), suggesting that, untreated, those birds feel pain to a degree that it either badly limits their ability to walk or completely prevents them from doing so. So, if their ability to walk is severely impaired, it is a sign of suffering. On the other hand, mild lameness, as found in gait scores 1 and 2, may be something that the bird is able to cope with, which is therefore a less serious welfare problem (although few studies have been done on the welfare consequences of gait scores 1 and 2). The point we are making here may be described in more general terms as follows: suffering, which is a serious welfare problem, consists of negative mental states. However, due to the ability of most animals to adapt and cope, suffering is not a simple additive function of these negative mental states. Rather, suffering only occurs when a certain threshold is reached due to the intensity, combination or duration of negative states. An example of such a combination may be a low number of pecks that may not add up to a welfare problem, but if combined with a slight feed restriction and insufficient access to a dry resting area, the combined effect may be greater than the simple additive value of the isolated effect of each. It could be argued that the overall reduction in welfare caused by these combined stressors therefore cannot be assessed on the basis of the individual measures, but needs to be assessed using an overall measure, such as cortisol, telomere length (see Bateson & Poirier 2019; this issue) or an emotional bias test. (For a more thorough discussion of the notion of suffering see Weary 2014).

It should be added that scientists know very little about the way relatively small independent stressors interact in a single animal. Some of the few studies indicate that they in combination may have a very strong effect (eg Ritter *et al* 2009). Furthermore, there may be individual differences in the ability to adapt and cope (cf the notion of a pain threshold). Therefore, simply 'averaging' negative states over time and/or across individuals may give a very poor measure of the amount of suffering. Instead, it seems necessary to take account of the intensity, duration and combination of negative states within individuals, as well as the individual's perceived control over the situation.

So far, we have been talking about welfare as defined by avoiding pain and other negative mental states. However, narrowing down the notion of welfare in this way is, of course, highly controversial, not only because of the omission of positive mental states but also because there are definitions of 'welfare' to be found within the field of animal welfare science that do not share the assumption that welfare is all about mental states. As aptly phrased by Fraser and colleagues (1997), besides feelings (mental states), welfare can be defined in terms of function and natural living. Or, in more traditional philosophical vocabulary, there are different theories of well-being, one of which, hedonism, defines welfare in terms of mental states, whereas others focus in addition, or exclusively, on preference satisfaction or perfection (understood in terms of flourishing or performing natural behaviour) (Appleby & Sandøe 2002).

The conclusion of this section is that aggregation raises challenging issues about how to add up different mental states such as pain, where the resulting welfare is not a simple additive function of these states, and about how to deal with indicators where there are varying levels of agreement concerning their validity.

Considering aggregation within WQ from an ethical perspective

The aim of WQ was to develop a methodology for the assessment of welfare at farm level to enable the labelling of products of animal origin with regard to the welfare state. To create easily understandable labels, the aggregation of animal welfare measures is of prime importance. Aggregation within WQ therefore concerns how to add up the results of a wide variety of indicators across individuals within one farm.

In light of the considerations presented above it is appropriate to raise the following five questions concerning the aggregation undertaken in WQ, which will serve to structure the discussion of this section:

- 1) How are different indicators relating to an individual animal aggregated within WQ to reach a verdict about the welfare of that animal?
- 2) How are different welfare indicators aggregated across animals within WQ?
- 3) How are mild levels of welfare problems aggregated with more severe ones within WQ?
- 4) How are indicators with different levels of perceived importance aggregated within WQ?
- 5) What is the role of experts in aggregation within WQ and how is it justified?

1) How are different indicators relating to an individual animal aggregated within WQ to reach a verdict about the welfare of that animal?

According to some influential ethical theories (see the first section of this paper), the relative negative load on the welfare of individual animals, and not just the sum of welfare scores across individuals, matters ethically. Therefore, it is important to aggregate welfare at the level of the individual animal. However, the short answer to this question is that it does not happen in the WQ protocols. This is because WQ aggregates across indicator values without considering the welfare status of individual animals (Veissier *et al* 2011). So, for example, in the case of dairy cows, the basic input into the aggregation process concerns the number or proportion of cows that are lame, that are emaciated, that collide with the housing equipment when lying down, and that show fear of humans measured by a long avoidance distance — to mention four indicators covering all four principles in the WQ definition of animal welfare. However, there is no indication of whether the same cows that are lame are also those that are emaciated, spend a long time lying down and show fear of humans.

To understand why this may pose a problem, consider, for example, two farms with exactly the same load of measured problems but with an important difference: in one farm, all the problems are borne by 10% of the animals, whereas on the other farm the problems are evenly divided between 30% of the animals. Clearly there will be more intense suffering (by a smaller number of animals) on the first farm but this difference will not be noted in WQ because the system does not look at how welfare problems are distributed between individuals.

WQ measures the welfare of individuals at one specific time within one farm, yet emphasis in the WQ protocols is given to indicators that are repeatable over time. As the indicators are not recorded over time, this is a limitation when it comes to following individual animals. The repeatability is therefore on farm level, not on an individual animal level, but it seems likely that the degree of suffering is linked to the duration of problems *within* individuals across time (see Houe *et al* 2011 for an account of how this may work out in real life situations).

In conclusion, whereas WQ keeps record of the total load of problems in a group, the way this load is distributed across individual animals remains unacknowledged and unmeasured in WQ, both from a cross-sectional and longitudinal perspective, as we have argued above. This is especially a concern for animals spending several years on-farm, eg dairy cattle, but may be less of a problem for the assessment of broiler welfare, for example.

2) How are different welfare indicators aggregated across animals within WQ?

This question is, like the previous one, justified by the apparent ethical significance of intra-individual aggregation, and by the importance attached by WQ to avoiding 'compensation.' However, as follows from our discussion of question 1, question 2 contains the same false assumption, ie that aggregation in WQ takes place across animals. Aggregation within WQ is done across indicators, not animals — which is an important ethical limitation to bear in mind in what follows.

When it comes to aggregation across indicators in WQ, the first step involves different scores relating to a certain indicator being transformed into a welfare score ranging from 0 (worst) to 100 (best). To illustrate this let's take the indicator 'lameness' for dairy cows. Initially, an index for lameness ranging from 0 to 100 is defined within WQ. This index score is a linear and additive function of the relative number of lame cows (weighted for severity), so that 100% severely lame cows gives a score of 0, and 100% non-lame cows gives a score of 100 (Veissier *et al* 2009; p 22). Subsequently, this score is transformed by means of a non-linear function derived from expert consultation. The specific non-linear function employed by WQ means that the addition of a lame cow will have a much greater effect on the welfare score of a farm with low levels of lameness compared to one with high levels of lameness (for a discussion of this claim and its implications, see Sandøe *et al* 2017).

Three further steps of aggregation follow: from indicators to criteria, from criteria to principles and from principles to the four final values (Excellent, Enhanced, Acceptable, and Not classified). In these three steps, weightings are made based on consultations with experts and/or other stakeholders. In the final step, where scores ranging from 0 to 100 for each of the four principles are aggregated into the four final categories, some pragmatism seems to enter into how the aggregation procedure is constructed.

As a first step, aspiration values for each principle were defined as 80, 55, and 20, respectively (where over 80 is Excellent, between 55–80 is Enhanced, between 20–55 is Acceptable, and below 20 is Not classified). These values were defined based on actual distributions (Veissier *et al* 2009; p 28) rather than on expert views, and it was taken for granted that farms were fairly evenly distributed between the four values.

The researchers, together with the Welfare Quality® Advisory Committee, next defined how to go from the values for the four principles to the final value for the farm.

https://doi.org/10.7120/09627286.28.1.067 Published online by Cambridge University Press

Here, different rules were considered. The first and simplest one was 'unanimity' which means that a farm needs to reach the aspiration value of a given category for all welfare principles to be assigned to that category. For instance, to be Enhanced, a farm would need to score at least 55 on all principles (Veissier *et al* 2009; p 28). However, this rule was rejected because it did not seem 'realistic', based on the observation that half of the farms visited would be scored as Not classified and the other half as only Acceptable (Veissier *et al* 2009; p 28).

There are several perspectives from which the WQ approach to aggregating can be discussed. The first and most obvious will relate to the official ambition of the WQ aggregation system to avoid or at least limit 'compensation', which we previously tentatively defined as whether a bad thing happening to one or more animals may be offset by avoiding other bad things happening. Unfortunately, the researchers behind WQ do not define what exactly is meant by compensation, nor the intended limit to compensation. Thus, compensation is not considered by the WQ researchers in the context of the principles of aggregation presented above, even though it has been suggested that some sort of middle position between utilitarianism and what we termed the difference principle is intended (Botreau *et al* 2007). So, it is difficult to tell whether the degree of compensation found is acceptable or not as measured by the standards set by the WQ researchers.

However, some studies have found that, as measured by expert opinion (which in many ways seems to be the gold standard for WQ), the level of compensation is too high. Thus, Sandøe *et al* (2017), in a case study involving 44 Danish dairy farms, demonstrated that welfare problems perceived as unacceptable by the experts were effectively hidden by good scores for other aspects of welfare — indicating that the WQ system allows compensation at a level that the surveyed experts would find unacceptable. Similarly, in a Dutch study of 196 dairy herds (de Vries *et al* 2013), it was found that "except for one herd, a high prevalence of (severely) lame cows did not result in herds being classified as unacceptable" (p 6271). The authors of the study combined this finding with a reference to two other studies (Whay *et al* 2003; Lievaart & Noordhuizen 2011) showing that animal welfare experts rank "lameness as the most important measure of dairy cattle welfare" (p 6271). The authors also highlighted the mechanisms in the WQ system that give rise to this problematic sort of compensation. In addition, a Belgian study (de Graaf *et al* 2017) found a low level of agreement between the overall WQ welfare score and scoring by experts from multiple European countries. So it is fair to say that, as measured by expert opinion, WQ does not succeed in preventing or limiting compensation to an acceptable degree.

However, it is also important to consider that the problems encountered are to some extent endemic to scoring systems of the kind considered here. Animal welfare is widely recognised as being a multidimensional phenomenon, and it therefore makes sense to use a number of different measures to capture the overall welfare of an animal, or animals, on a farm. However, as the number of measures increases, any assessment system that does not use cut-offs for the individual measures, where going beyond the cut-off has the consequence that the welfare of the group will be deemed unacceptable, will see the importance of each individual measure decrease. Any animal welfare protocol will therefore need to balance the multidimensional nature of animal welfare with the dilution of the individual measures.

A further concern relates to the four steps of aggregation each with ethically motivated weightings found in the WQ system. Because weighting is done in several iterations, the weighting and the ethical judgements are likely to become progressively more opaque. Experts may know what they are doing when they indicate their moral acceptance of the levels of welfare problems at the most basic level of transforming different indicator scores into welfare scores. However, it is very unlikely that they will maintain the overview and not lose track of the underlying weightings of the previous step(s) when aggregating in three further iterations.

Another concern here is the pragmatism with which WQ defines acceptable welfare, not based on what experts or other stakeholders consider to be acceptable, but seemingly based on the tacit assumption that only a minority of commercial farms will have an unacceptable level of animal welfare. Clearly, this could be seen as a politically convenient assumption, but it risks undermining the credibility of WQ among those who are worried about modern, intensive animal production. Even though the attempt to deal with the worst problems may be appreciated, there may also be the concern that the system, overall, serves to 'cover-up' the problematic practices found in modern animal production.

So, the conclusion of this section is that the way aggregation of welfare indicators is undertaken in WQ creates a number of critical problems and limitations: the focus is on indicators, not individuals; unacceptable compensation (as measured by expert opinion) is not prevented; due to four successive iterations of weightings the aggregation procedure lacks transparency; and assumptions about what is a 'realistic' requirement for an acceptable level of welfare may undermine the credibility of the system.

3) How are mild levels of welfare problems aggregated with more severe ones within WQ?

The WQ system does reflect a justified concern for putting more emphasis on severe compared to mild problems. The case of lameness in dairy cows may again serve as an illustration. Three possible values of lameness are recorded: 'non-lame', 'mildly lame' and 'severely lame'. Mild lameness is, according to WQ, given a weighting of 2 and severe lameness a weighting of 7, so that in terms of welfare impact, one severely lame cow would be equivalent to three and a half mildly lame cows. So, clearly an effort is made here to make more severe problems stand out compared to the mild ones.

4) How are indicators with different levels of perceived importance aggregated within WQ?

As already mentioned, WQ takes as its starting point a comprehensive definition of farm animal welfare in terms of four principles, Good feeding, Good housing, Good health, and Appropriate behaviour. These four principles and the underlying 12 criteria were constructed by a combination of deliberation among the researchers involved in WQ and input from the social science wing of WQ about public perception of animal welfare. In line with previous studies (eg Lassen *et al* 2006) it was found that the general public emphasised the importance of positive welfare and natural living.

While including the views of all relevant parties makes room for a comprehensive account of animal welfare, it risks creating another layer of problems, where the different elements have to be balanced against each other. For example, how much weight should be given to natural living and positive welfare compared to the prevention of suffering? Here, there is a real danger that bad scores for welfare problems considered highly important by experts (eg severe lameness in dairy cattle), may end up being counter-balanced by good scores for problems experts consider less important (eg availability of drinkers) (de Vries *et al* 2013).

5) What is the role of experts in aggregation within WQ and how is it justified?

Whereas input was provided from studies of public perception regarding the overall architecture of the WQ definition of animal welfare in terms of the four principles and the underlying 12 criteria, the input into the aggregation procedure mainly came from expert consultation. One motivation for this difference could be that the definition of animal welfare is taken to be of an ethical nature while the aggregation procedure is more technical. However, although there are obviously technical issues involved in setting up an aggregation procedure, at the core, many decisions, as we have seen, are clearly of an ethical nature.

Take, for example, the transformation of a lameness index for dairy cows into a welfare score described under question 2. To achieve this, experts were asked to score the welfare impact of the different values of the lameness index on a scale concerned partly with whether action *could*, *should* or *must* be taken. Thereby, a welfare index ('how bad for the cows?') is transformed into an ethics score ('how acceptable?'). However, the respondents were experts in the field of welfare, not ethics, and furthermore, the question of whether any individual can be an expert in ethics in the relevant sense is certainly open to debate.

It is important here to highlight the results of a study (Tuyttens *et al* 2010) that found that three different groups of stakeholders, 'farmers, citizens and vegetarians' disagreed with both the WQ aggregation procedure, and with each other. So, there is reason to believe that the ethical views of the experts may differ from those of other stakeholders.

Indeed, the same may be true for scientists. When it comes to different welfare indicators, recent evidence (Sandøe *et al* 2017) suggests that there are big differences in their perceived validity and importance between experts. Some indicators, notably those linked to pain, such as lameness, appear to be widely perceived as valid and important, whereas others are much more controversial. This gives rise to a dilemma when it comes to setting up ways to aggregate animal welfare. Either one focuses on indicators that command wide agreement, resulting in a very narrow view of welfare, or one goes for a more comprehensive notion of welfare, with the potential for a high level of disagreement concerning the validity of the results.

Animal welfare implications

Methods of aggregation may be of huge consequence for animal welfare when schemes for assessing animal welfare on farm or group level are applied. They will, among other things, define the relative weight assigned to different aspects of animal welfare, and they will define how to add up across individuals, and across different indicators of animal welfare.

Since defining an aggregation system will inevitably involve taking controversial ethical decisions, notably on how to balance the total or average welfare of the affected animals against concern for the worst affected animals, there is a crucial need for transparency as to how aggregation is carried out. Without such transparency, there is a danger that welfare issues that matter most to some stakeholders will inadvertently be ignored.

Also, it is important to recognise the political nature of some decisions relating to aggregation, not least when it comes to defining the line between acceptable and unacceptable levels of welfare. What, to one stakeholder, may seem to be a sensible and pragmatic tool to improve animal welfare over time may be seen by another as an attempt to cover-up unacceptable animal welfare problems.

WQ is, to date, the most sophisticated attempt within animal welfare science to set up an aggregation system in combination with a comprehensive set of animal welfare indicators. This is a highly positive initiative and should be commended. However, we argue here that the aggregation system is in several ways problematic and would benefit from a fundamental overhaul.

Some points that could be considered in such an overhaul include:

- 1) Making sure that severe welfare problems, such as lameness in dairy cows, are not hidden;
- 2) Making ethical decisions more transparent at all levels, depending on the aim of the aggregation procedure; and
- 3) Involving other stakeholders so that their views are represented when such ethical decisions are taken.

We are aware that these points (particularly the third one) so far remain vague and that it is much easier to criticise an existing aggregation system, such as that used in WQ, than to develop a feasible alternative.

https://doi.org/10.7120/09627286.28.1.067 Published online by Cambridge University Press

Acknowledgements

The authors are grateful for very thorough and perceptive comments made by three anonymous referees to an earlier version of this paper. Peter Sandøe would like to acknowledge his indebtedness to earlier collaboration on the issue of aggregation with Karsten Klint Jensen.

References

Appleby MC and Sandøe P 2002 Philosophical debate on the nature of well-being: Implications for animal welfare. *Animal Welfare 11*: 283-294

Arneson R 2000 Luck egalitarianism and prioritarianism. *Ethics 110*: 339-349. https://doi.org/10.1086/233272

Bateson M and Poirier C 2019 Can biomarkers of biological age be used to assess cumulative lifetime experience? *Animal Welfare 28*: 41-56. https://doi.org/ 10.7120/09627286.28.1.041

Botreau R, Bracke MBM, Perny P, Butterworth A, Capdeville J, Van Reenen CG and Veissier I 2007 Aggregation of measures to produce an overall assessment of animal welfare. Part 2: Analysis of constraints *Animal 1*: 1188-1197. https://doi.org/10.1017/S1751731107000547

Capdeville J and Veissier I 2001 A method of assessing welfare in loose housed dairy cows at farm level, focusing on animal observations. *Acta Agriculturae Scandinavica, Section A - Animal Science 51(S30)*: 62-68

Caplen G, Hothersall B, Nicol CJ, Parker RMA, Waterman-Pearson AE, Weeks CA and Murrell JC 2014 Lameness is consistently better at predicting broiler chicken performance in mobility tests than other broiler characteristics. *Animal Welfare 23*: 179-187. https://doi.org/10.7120/09627286.23.2.179

D'Eath RB, Arnott G, Turner SP, Jensen T, Lahrmann HP, Busch ME, Niemi JK, Lawrence AB and Sandøe P 2014 Injurious tail biting in pigs: how can it be controlled in existing systems without tail docking? *Animal 8*: 1479-1497. https://doi.org/10.1017/S1751731114001359

de Graaf S, Ampe B, Winckler C, Radeski M, Mounier L, Kirchner MK, Haskell MJ, van Eerdenburg FJCM, de Boyer des Roches A, Andreasen SN, Bijttebier J, Lauwers L, Verbeke W and Tuyttens FAM 2017 Trained-user opinion about Welfare Quality® measures and integrated scoring of dairy cattle welfare. *Journal of Dairy Science 100*: 6376-6388. https://doi.org/10.3168/jds.2016-12255

de Vries M, Bokkers EAM, van Schaik G, Botreau R, Engel B, Dijkstra T and de Boer IJM 2013 Evaluating results of the Welfare Quality® multi-criteria evaluation model for classification of dairy cattle welfare at the herd level. *Journal of Dairy Science 96*: 6264-6273. https://doi.org/10.3168/jds.2012-6129

Fraser D, Weary DM, Pajor EA and Milligan BN 1997 A scientific conception of animal welfare that reflects ethical concerns. *Animal Welfare 6*: 187-205

Hare RM 1981 *Moral Thinking – Its Level, Method, and Point.* Clarendon Press: Oxford, UK. https://doi.org/10.1093/0198246609.001.0001

Houe H, Sandøe P and Thomsen PT 2011 Welfare assessments based on lifetime health and production data in Danish dairy cows. *Journal of Applied Animal Welfare Science 14*: 255-264. https://doi.org/10.1080/10888705.2011.576984

Keeling L 2009 *An Overview of the Development of the Welfare Quality® Assessment Systems.* Welfare Quality® Reports no 12: Cardiff University, UK

Kestin SC, Knowles TG, Tinch AE and Gregory NG 1992 Prevalence of leg weakness in broiler chickens and its relationship with genotype. *Veterinary Record 131*: 190-194. https://doi.org /10.1136/vr.131.9.190

Krotscheck U, Nelson SA, Todhunter RJ, Stone M and Zhang Z 2016 Long term functional outcome of TTA vs TPLO and extracapsular repair in a heterogeneous population of dogs. *Veterinary Surgery 45*: 261-268. https://doi.org/10.1111/vsu.12445

Lassen J, Sandøe P and Forkman B 2006 Happy pigs are dirty! – conflicting perspectives on animal welfare. *Livestock Science 103*: 221-230. https://doi.org/10.1016/j.livsci.2006.05.008

Lievaart JJ and Noordhuizen JPTM 2011 Ranking experts' preferences regarding measures and methods of assessment of welfare in dairy herds using Adaptive Conjoint Analysis. *Journal of Dairy Science 94*: 3420-3427. https://doi.org/10.3168/jds.2010-3954

McGeown D, Danbury TC, Waterman-Pearson AE and Kestin SC 1999 Effect of carprofen on lameness in broiler chickens. *Veterinary Record 144*: 668-671https://doi.org/10.1136/vr.144.24.668

Parfit D 1997 Equality and priority. *Ratio 10*: 202-221. https://doi.org/10.1111/1467-9329.00041

Petersen JS 2006 Benmonitoreringsprojektet 2005. In: *Årsberetningen fra Det Danske Fjerkræraad* pp 16-19. Det Danske Fjerkræraad: Copenhagen, Denmark

Porter DG 1992 Ethical scores for animal experiments. *Nature 356*: 101-102. https://doi.org/10.1038/356101a0

Rasmussen IK and Spangberg A 2012 *Screening af slagtekyllingers gangegenskaber anno 2011*. Videncentret for Landbrug, Fjerkræ: Århus, Denmark

Rawls J 1971 *A Theory of Justice. Revised Edition 1999*. Harvard University Press: Cambridge, USA

Rawls J 1993 *Political Liberalism*. Columbia University Press: New York, USA

Ritter MJ, Ellis M, Anderson DB, Curtis SE, Keffaber KK, Killefer J, McKeith FK, Murphy CM and Peterson BA 2009 Effects of multiple concurrent stressors on rectal temperature, blood acid-base status, and longissimus muscle glycolytic potential in market-weight pigs. *Journal of Animal Science 87*: 351-362. https://doi.org/10.2527/jas.2008-0874

Sandøe P, Forkman B, Hakansson F, Andreasen SN, Nøhr R, Denwood M and Lund TB 2017 Should the contribution of one additional lame cow depend on how many other cows on the farm are lame? *Animals 7(12)*: 1-13. https://doi.org/10.3390/ani7120096

Sanotra GS, Lund JD, Ersbøll AK, Petersen JS and Vestergaard KS 2001 Monitoring leg problems in broilers: a survey of commercial broiler production in Denmark. *World's Poultry Science Journal 57*: 55-69. https://doi.org/10.1079 /WPS20010006

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

Stafleu FR, Tramper R, Vorstenbosch J and Joles JA 1999 The ethical acceptability of animal experiments: a proposal for a system to support decision-making. *Laboratory Animals 33*: 295-303. https://doi.org/10.1258/002367799780578255

Tuyttens FAM, Vanhonacker F, Van Poucke E and Verbeke W 2010 Quantitative verification of the correspondence between the Welfare Quality® operational definition of farm animal welfare and the opinion of Flemish farmers, citizens and vegetarians. *Livestock Science 131*: 108-114. https://doi.org/10.1016/j.livsci.2010.03.008

Valros A and Heinonen M 2015 Save the pig tail. *Porcine Health Management 1*: 1-7. https://doi.org/10.1186/2055-5660-1-2

Veissier I, Botreau R and Perny P 2009 Scoring animal welfare: difficulties and Welfare Quality® solutions. In: Keeling L (ed) *An Overview of the Development of the Welfare Quality® Assessment Systems* pp 15-32. Welfare Quality® Reports no 12: Cardiff University, UK

Veissier I, Jensen KK, Botreau R and Sandøe P 2011 Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. *Animal Welfare 20*: 89-101

Weary DM 2014 What is suffering in animals? In: Appleby MC, Weary DM and Sandøe P (eds) *Dilemmas in Animal Welfare*. CABI: Wallingford, UK. https://doi.org/10.1079/9781 780642161.0188

Weary DM and Robbins J 2019 Understanding the multiple conceptions of animal welfare. *Animal Welfare 28*: 33-40. https://doi.org/10.7120/09627286.28.1.033

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

Wucherer KL, Conzemius MG, Evans R and Wilke VL 2013 Short-term and long-term outcomes for overweight dogs with CCL rupture treated surgically or non-surgically. *Journal of the American Veterinary Medical Association 242*: 134-172. https://doi.org/10.2460/javma.242.10.1364

https://doi.org/10.7120/09627286.28.1.067 Published online by Cambridge University Press