

Scotland's Rural College

## Trained-user opinion about Welfare Quality® measures and integrated scoring of dairy cattle welfare

de Graaf, S; Ampe, B; Winckler, C; Radeski, M; Mounier, L; Kirchner, MK; Haskell, MJ; van Eerdenburg, FJCM; Boyer des Roches, A; Andreasen, SN; Bijttebier, J; Lauwers, L; Verbeke, W; Tuytens, FAM

*Published in:*  
Journal of Dairy Science

*DOI:*  
[10.3168/jds.2016-12255](https://doi.org/10.3168/jds.2016-12255)

First published: 29/05/2017

*Document Version*  
Peer reviewed version

[Link to publication](#)

### *Citation for published version (APA):*

de Graaf, S., Ampe, B., Winckler, C., Radeski, M., Mounier, L., Kirchner, MK., Haskell, MJ., van Eerdenburg, FJCM., Boyer des Roches, A., Andreasen, SN., Bijttebier, J., Lauwers, L., Verbeke, W., & Tuytens, FAM. (2017). Trained-user opinion about Welfare Quality® measures and integrated scoring of dairy cattle welfare. *Journal of Dairy Science*, 100(8), 6376 - 6388. Advance online publication. <https://doi.org/10.3168/jds.2016-12255>

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **Interpretive summary**

2 **Trained-user opinion about Welfare Quality® measures and integrated scoring of dairy**

3 **cattle welfare. By de Graaf et al.** Overall welfare assessments can be used for

4 communication with consumers (labelling), as incentive for welfare improvements and as

5 regulative target. Such assessments should be valid, reliable and balance the relative

6 importance of various welfare measures. The Welfare Quality® (WQ) approach is one of the

7 most well-known methods for overall welfare assessment. However, the present study shows

8 that the WQ integration method should be revised if it is to correspond with the opinions of

9 trained users of the WQ protocol for dairy cattle.

10 WELFARE QUALITY® VS. TRAINED-USER OPINION

11 **Trained-user opinion about Welfare Quality® measures and integrated scoring of dairy**

12 **cattle welfare**

13 S. de Graaf<sup>\*, †</sup>, B. Ampe<sup>\*</sup>, C. Winckler<sup>‡</sup>, M. Radeski<sup>§</sup>, L. Mounier<sup>#</sup>, M.K. Kirchner<sup>||</sup>, M.J.

14 Haskell<sup>||</sup>, F.J.C.M. van Eerdenburg<sup>\*\*</sup>, A. de Boyer des Roches<sup>#</sup>, S.N. Andreasen<sup>||</sup>, J.

15 Bijttebier<sup>\*</sup>, L. Lauwers<sup>\*, †</sup>, W. Verbeke<sup>†</sup>, F.A.M. Tuytens<sup>\*</sup>

16

17 <sup>\*</sup>Institute for Agricultural and Fisheries Research (ILVO), Burg. van Gansberghelaan 92,

18 9820 Merelbeke, Belgium

19 <sup>†</sup>Department of Agricultural Economics, Ghent University, Coupure links 653, 9000 Ghent,

20 Belgium

21 <sup>‡</sup>Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University

22 of Natural Resources and Life Sciences, Gregor-Mendel Straße 33, 1180 Vienna, Austria

23 <sup>§</sup>Animal Welfare Center, Faculty of Veterinary Medicine, Ss. Cyril and Methodius University,

24 Lazar Pop-Trajkov 5-7, 1000 Skopje, Republic of Macedonia

25 #UMR1213 Herbivores, INRA, VetAgro Sup, Clermont université, université de Lyon, F-  
26 63122 Saint-Genès-Champanelle, France

27 †University of Copenhagen, Dept. of Veterinary and Animal Sciences, Section of Animal  
28 Welfare and Disease Control, Grønnegårdsvej 8, 1870 Frederiksberg, Copenhagen, Denmark

29 ‡SRUC, West Mains Road, Edinburgh EH9 3JG, Scotland, United Kingdom

30 \*\*Department of Herd Animal Health, Utrecht University, 3508 TD Utrecht, The Netherlands

31 Corresponding Author: Frank Tuytens. Email: Frank.Tuytens@ilvo.vlaanderen.be

## 32 **ABSTRACT**

33 The Welfare Quality® (WQ) protocol for on-farm dairy cattle welfare assessment describes  
34 27 measures and a step-wise method to integrate values for these measures into 12 criteria  
35 scores, grouped further into four principle scores and finally into an overall welfare  
36 categorization with four levels. We conducted an online survey to examine whether trained  
37 users' opinions of the WQ protocol for dairy cattle correspond with the integrated scores  
38 (criteria, principles and overall categorization) calculated according to the WQ protocol. First,  
39 the trained users' scores (n = 8 - 15) for reliability, validity and their ranking of the  
40 importance of all measures for herd welfare were compared to the degree of actual impact of  
41 these measures on the WQ integrated scores. Logistic regression was applied to identify the  
42 measures that affected the WQ overall welfare categorization into the 'not classified' or  
43 'enhanced' categories for a database of 491 European herds. The smallest multivariate model  
44 whilst maintaining the highest % of both sensitivity and specificity for the 'enhanced'  
45 category contained six measures, the model for not-classified contained four measures. Some  
46 of the measures that were ranked as least important by trained users (e.g. measures relating to  
47 drinkers) had the highest influence on the WQ overall welfare categorization. Conversely,  
48 measures rated as most important by the trained users (e.g. lameness and mortality) had a  
49 lower impact on the WQ overall category. In addition, trained users were asked to allocate

50 'criterion' and 'overall' welfare scores to seven focal herds selected from the database (n =  
51 491 herds). Data on all WQ measures for these focal herds relative to all other herds in the  
52 database were provided. The degree to which expert scores corresponded to each other, the  
53 systematic difference and the correspondence between median trained-user opinion and the  
54 WQ criterion scores were then tested. The level of correspondence between expert scoring vs.  
55 WQ scoring for 6 of the 12 criteria and for the overall welfare score was low. The WQ scores  
56 of the protocol for dairy cattle thus lacked correspondence with trained users on the  
57 importance of several welfare measures.

58 **Keywords:** animal welfare, welfare assessment, trained-user opinion, Welfare Quality®

## 59 INTRODUCTION

60 Assessing animal welfare is a highly complex task. Animal welfare is a multidimensional  
61 concept, which calls for a multi-criteria assessment using a multitude of welfare-indicators  
62 (Mason and Mendl 1993; Fraser et al., 1997). To express the overall welfare status of a group  
63 of (farm) animals in one score or index, indicator data should be integrated which requires  
64 interpretation and balancing. The lack of a 'gold standard' for animal welfare assessment (i.e.  
65 there is no standardized and commonly agreed-on method for assessing the overall welfare  
66 status of a group of farm animals) implies that some degree of subjectivity is inevitable when  
67 weighting different measures (Spoolder et al., 2003). To be widely accepted, an overall  
68 welfare index ought to correspond with society's concept of animal welfare and with the  
69 opinion of experts, i.e. people who are seen by society to have adequate knowledge and  
70 expertise about animal welfare. However, opinions on the concept of animal welfare may  
71 differ between and even within experts and society. For example, producers tend to highlight  
72 basic health and functioning of farm animals while non-producers tend to emphasize the need  
73 for a natural living environment of farm animals (reviewed by Sørensen and Fraser, 2010). It

74 can be argued that for people without expertise in dairy cattle welfare and the specific welfare  
75 measures involved, it is too difficult to adequately balance the importance of different welfare  
76 measures. It has been shown that providing detailed information about on-farm collection  
77 methods of welfare measures, significantly influences the relative weights they are given by  
78 experts (Rodenburg et al., 2008). Therefore, the current study elicited experienced animal  
79 scientists on the specific welfare measures involved only.

80 To date, the Welfare Quality<sup>®</sup> (WQ) protocols are most likely the most renowned and  
81 comprehensive method for overall welfare assessment of different farm animal species  
82 (chickens, pigs and cattle) (Welfare Quality, 2009). Unlike some other welfare assessment  
83 protocols, WQ relies predominantly on animal-based measures. Resource-based and  
84 management-based measures, in contrast, mostly reflect risk factors for welfare impairments  
85 instead of directly measuring welfare (Blokhuis et al., 2003; 2010). The WQ protocols are  
86 based on four main welfare principles ('good feeding', 'good housing', 'good health' and  
87 'appropriate behavior') which are split into 12 independent welfare criteria (Table 1). Various  
88 welfare measures (n = 27 for dairy cows) were selected by animal scientist to assess these  
89 welfare criteria, based on validity, reliability and feasibility to perform on-farm. The WQ  
90 protocol describes three steps to integrate these welfare measures into an overall final welfare  
91 category. Methods of integration aim to be widely acceptable by society and are therefore  
92 based upon expert opinion of social and animal scientists and stakeholders (Botreau et al.,  
93 2007), depending on the integration step. For interpretation of measures into criteria scores,  
94 animal scientists were consulted (n = 6) who were involved in the choice and development of  
95 the WQ measures (Botreau et al., 2008). They were asked to score several situations which  
96 could occur on-farm per criterion (e.g. for integument alterations within the criterion 'absence  
97 of injuries', experts were asked to score 11 hypothetical farms with varying prevalence of  
98 hairless patches, wounds and swellings). Calculation of criterion scores is based on expert

99 scoring. For aggregation from criteria to principle scores, social scientists were involved as  
100 well, using a similar approach. For the final step, several scenarios for reference profiles were  
101 developed to aggregate principle scores into an overall category. These scenarios were tested  
102 for 69 European dairy farms (Austrian, German and Italian) to firstly compare their ability to  
103 discriminate between farms. Secondly, stakeholders were consulted to assess which scenario  
104 was most appropriate and thirdly, the degree to which each scenario matched with the general  
105 impression of observers for 44/69 dairy farms was assessed. The four overall categories  
106 ('excellent', 'enhanced', 'acceptable' or 'not classified' (Welfare Quality<sup>®</sup>, 2009)) were  
107 constructed to reflect both the multi-dimensional nature of welfare and the relative importance  
108 of the various welfare measures using mathematical operators which limit the amount of  
109 compensation which may occur between welfare measures, i.e. when a combination of  
110 positive scores compensate for one negative score (Botreau et al., 2009).

111 Recent critical evaluations of the WQ integration methods indicate that in the dairy cattle  
112 protocol a few resource-based measures appear to have a disproportionately large influence  
113 on integrated scores (Heath et al., 2014; de Vries et al., 2014). For example, the measures for  
114 the criterion 'absence of prolonged thirst' (i.e. number, adequate functioning and cleanliness  
115 of drinkers) have a relatively large influence on integrated scores, although they are criticized  
116 for their low or undocumented validity (Knierim and Winckler, 2009; de Vries et al., 2013; de  
117 Jong et al., 2016; Tuytens et al., 2014). In contrast, some of the most pressing welfare  
118 problems for dairy cattle as highlighted by epidemiological studies (de Boyer des Roches et  
119 al, 2014; Main et al, 2003; Whay et al, 2003) and assessed by experts (i.e. mortality, lameness  
120 and mastitis, Lievaart and Noordhuizen, 2011; Nielsen et al., 2014; Whay et al., 2003), had a  
121 smaller influence on overall welfare categorization (de Vries et al., 2013; Heath et al., 2014;  
122 Buijs et al., 2016) These findings point towards potential discrepancies between the welfare  
123 assessment in dairy cattle of certain welfare experts and the WQ scores.

124 The WQ protocols were designed with the intention of modifying and updating assessment  
125 methods according to advances in animal welfare science. Currently, a large group of  
126 researchers has become familiar with the protocol and many farm visits have been performed  
127 by these researchers (further referred to as ‘trained users’), allowing for a thorough evaluation  
128 of the impact which measures have on overall welfare categorization. Therefore, analyzing  
129 the correspondence between WQ integrated scores and the opinion of such trained users has  
130 now become feasible. Hence, the objective of the current study was to analyze  
131 correspondence between welfare assessment by trained users and the WQ scores (criterion  
132 and overall welfare category). We performed this by examining whether measures which  
133 impact WQ categorization most are also those which are deemed most important by trained  
134 users.

## 135 MATERIALS AND METHODS

### 136 *WQ Protocol*

137 A brief description of the WQ protocol for on-farm dairy cattle welfare assessment is  
138 presented below; the full protocol can be found at <http://www.welfarequalitynetwork.net/>. In  
139 short, the protocol describes 27 on-farm welfare measures (Table 1) that are subsequently  
140 integrated in a 3-step process to arrive at an overall welfare category. First, 27 welfare  
141 measures of various scales are combined into scores for 12 welfare criteria on a scale of 0  
142 (worst) – 100 (best) (Table 1), using various aggregation methods (for details see Welfare  
143 Quality<sup>®</sup>, 2009). Second, criteria are integrated into scores for four welfare principles using  
144 Choquet integrals, algorithmic operators which ensure that a poor score cannot be fully  
145 compensated by a better score in another criterion (Botreau et al., 2008). Principle scores can  
146 range from 0 (worst) to 100 (best). The third and final integration step is an outranking  
147 procedure from principle scores, arriving at an overall welfare category. Dairy welfare in a

148 herd is considered ‘excellent’ when that herd scores >50 for each principle and >75 on two of  
149 them. When a herd scores >15 on each principle and >50 on at least two of them, it is  
150 classified as ‘enhanced’. ‘Acceptable’ herds score >5 for all principles and >15 for at least  
151 three principles. Herds that do not reach the thresholds for the category ‘acceptable’ are  
152 considered ‘not classified’. These reference profiles for overall welfare categorization were  
153 based on data from 69 herd assessments in the European Union (Botreau et al., 2009).

154 <Table 1>

### 155 *Collating WQ Data*

156 Datasets of assessments using the WQ protocol for on-farm dairy cattle welfare were collated  
157 from seven European research institutes. Data from 10 countries (Macedonia, The  
158 Netherlands, France, Belgium, Scotland, Denmark, Romania, Northern Ireland, Spain and  
159 Austria) and 491 herds were used. The collected samples were selected to be representative  
160 for 1) small scale dairy herds in Macedonia (n = 12); 2) non-organic and non-tie stall dairy  
161 herds in The Netherlands (n = 60) and France (n = 128); 3) random herds with individual  
162 Somatic Cell Count data available (SCC, to be able to calculate WQ scores) in Belgium (n =  
163 140), Scotland (n = 16) and Denmark (n = 42); 4) typical herds for the regional low-input  
164 herding systems in Romania, Northern Ireland and Spain (n = 30); and 5) loose housed dairy  
165 herds with at least 20 cows in Austria (n = 65). Integrated WQ scores were calculated from  
166 raw data using a custom-made integration procedure programmed in R 3.2.2 (R Foundation  
167 for Statistical Computing, Vienna, Austria). The R integration program is available on  
168 request. The resulting welfare scores were in agreement with the INRA WQFA webtool  
169 (<http://www1.clermont.inra.fr/wq/>), in which WQ measure scores can be entered (for dairy  
170 cows, fattening cattle, growing pigs and broilers), and WQ criteria, principle and  
171 categorization scores are provided.



172 **Survey**

173 The survey was sent to 31 trained users, partially completed by 14 - 15 (depending on the  
174 question) and totally completed by 8 trained users. The survey was sent to animal welfare  
175 scientists which the co-authors knew to be experienced in the WQ assessment protocol for  
176 dairy cow welfare. These trained users were in turn asked to provide contact details of any  
177 additional animal welfare scientists which would be suitable (trained to use the WQ protocol).  
178 All trained users who filled out the survey, were not involved in creating the survey. All  
179 trained users had experience with the WQ protocol for dairy cattle (i.e. were trained to  
180 perform the WQ protocol for dairy cattle and had performed on-farm WQ assessment of dairy  
181 herds), were animal scientists and had authored at least 1 peer-reviewed scientific paper about  
182 dairy cattle welfare involving the WQ protocol. Trained users were all European and a total  
183 of 8 different nationalities was represented (British, Spanish, Macedonian, Dutch, Finnish,  
184 Austrian, German and French). Trained users were surveyed on their judgement of the  
185 reliability, validity and importance of all WQ measures. In questions based on data from the  
186 WQ EU database, they were asked to score the farms for each WQ criteria and to assign an  
187 overall welfare score.

188 ***Reliability, Validity and Ranking of all WQ Measures for Dairy Cattle.*** The trained  
189 users were asked to indicate how acceptable they judged the reliability and validity of all  
190 measures using a tagged visual analogue scale from 0 to 100. Tags were ‘not acceptable  
191 (<25)’, ‘just acceptable (25 – 50)’, ‘acceptable (50 – 75)’, and ‘very acceptable (75 – 100)’.  
192 ‘Reliability’ was defined in the survey as ‘a combination of inter-observer, intra-observer and  
193 test-retest reliability’. ‘Validity’ was defined as ‘the measure measures what it is supposed to’.  
194 Trained users were then asked to rank all WQ measures according to importance for the  
195 overall welfare status of a herd of dairy cattle from 1 (most important) – 27 (least important).

196 It was mentioned that for ranking, (inter alia) reliability, validity, perceived relevance and  
197 prevalence may be considered.

198 ***Expert scoring based on all WQ measurements.*** The trained users were then asked to  
199 score overall welfare based on all measures from the WQ protocol. They were shown one  
200 figure with box plots for all measures (part of the figure for one criterion: Figure 1). These  
201 showed the same herds as in the first figure using the same colored triangles. Trained users  
202 were asked to score overall welfare of 7 focal herds using a 0-100 tagged visual analogue  
203 scale, with the tags ‘not classified’ (< 20), ‘acceptable’ (20 – 55), ‘enhanced’ (55- 80) and  
204 ‘excellent’ (>80). For this purpose, we randomly selected five herds from the ‘acceptable’  
205 welfare category and two herds from the ‘enhanced’ category out of the entire dataset. This  
206 reflects the distribution of the dataset in which 1.8% of the herds were categorized as ‘not  
207 classified’ (9 herds), 62.7% as ‘acceptable’ (308 herds), 35.4% as ‘enhanced’ (174 herds) and  
208 none as ‘excellent’.

209 <Figure 1>

210 ***Comparing WQ Criteria Scores Using Trained-user Opinion.*** To assess the degree to  
211 which integrated WQ criteria scores correspond to trained-user opinion, the trained users were  
212 shown graphs of all measures per criterion separately, showing the distribution of all herds in  
213 the database (example of one criterion: Figure 2, data shown in Table 2). The ‘focus herds’  
214 were highlighted using triangles in different colors, and tables stated the data for each.  
215 Trained users were asked to score the herds for all 11 criteria (excluding the criterion ‘thermal  
216 comfort’, as this is not measured on-farm for dairy cattle) on a 0-100 tagged visual analogue  
217 scale using the tags ‘not classified (< 20)’, ‘acceptable (20 – 55)’, ‘enhanced (55- 80)’ and  
218 ‘excellent (>80)’.

219 <Figure 2>

220 ***Statistical analysis***

221 The statistical analysis was performed in R 3.2.2 (R Foundation for Statistical Computing,  
222 Vienna, Austria). The analyzed data (except overall welfare categorization) were considered  
223 to be sufficiently normally distributed, based on the graphical evaluation (histogram and QQ-  
224 plot) of the residuals.

225 ***Reliability, Validity and Ranking of all WQ Measures for Dairy Cattle.*** To examine the  
226 influence of median reliability and validity scores and their interaction on median ranking of  
227 all measures, we used a linear mixed regression model with reliability and validity scores as  
228 independent variables, and importance rank as dependent variable. A random effect for expert  
229 was included in the model to account for the repeated measures.

230 ***Predicting Overall Welfare Categorization Using WQ Measures.*** To analyze which  
231 measures affected the WQ overall categorization both into the lowest (not classified) and the  
232 highest (enhanced, as no farms were categorized as excellent) categories, welfare categories  
233 of the entire European dataset (n = 491) were divided into two binary variables (1=enhanced,  
234 0=other for variable 1; and 1=not classified, 0=other for variable 2). Logistic regression was  
235 used to identify measures that affected overall categorization both univariate and multivariate.  
236 For the latter, a model was built using stepwise forward selection, retaining measures with a P  
237 < 0.05 while maintaining the highest R2. Collinearity was checked for measures used within  
238 the models. Model outcome was assessed by calculating specificity and sensitivity using the  
239 following formulae:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

240 Where TN = true negatives, FP = false positives, TP = true positives and FN = false  
241 negatives. Negatives were those farms categorized as ‘other’ and positives were those farms  
242 categorized as either ‘enhanced’ for the first binary variable or ‘not classified’ for the second.

243 ***Comparing WQ Criteria Scores with Trained-user Opinion.*** To assess the systematic  
244 difference between the median trained-user opinion score and the WQ criteria scores for each  
245 focal herd (n = 7), a paired t-test was performed. To model the correspondence of median  
246 scores allocated by the trained users and the WQ criteria scores, a linear model was fitted and  
247 the coefficient of determination was calculated. Additionally, the intra-class correlation  
248 coefficient (ICC) was calculated to assess the degree of coherence between individual trained-  
249 user opinions.

## 250 **RESULTS**

### 251 ***Perceived Reliability, Validity and Ranking of WQ Measures***

252 Median validity and reliability scores for all measures were ‘acceptable’ to ‘very acceptable’  
253 (i.e. median scores > 50, Table 3). Nevertheless, there was variation in median scores for the  
254 various measures, ranging from 60 to 100 and from 50 to 90 for reliability and validity  
255 respectively. Highest median ranking was attached to ‘lameness score’ (rank 2), ‘body  
256 condition score’ (4), ‘mortality rate’ (7) and ‘integument alterations’ (7). ‘Lameness score’  
257 and ‘integument alternations’ received the highest median validity scores (89 and 90,  
258 respectively), along with ‘lying outside the lying area’ (89) and ‘tail docking method’ (88).  
259 ‘Tied versus loose housing’ (100), measures of drinker space (‘Centimeters of trough per cow  
260 (minimum 6 cm), number of water bowls per cow (minimum 0.10) and at least two drinkers  
261 available for each cow’ (93) and ‘water flow’ (90) received the highest median reliability  
262 scores. The measure ‘Qualitative Behavior Assessment’ (QBA) was given the worst median  
263 importance rank (22), the lowest median reliability score (60), and was among the lowest  
264 median validity scores (57). Measures of drinker space was given the lowest median validity  
265 score (50). ‘Water flow’ was among the lowest ranking measures in terms of importance (20)  
266 and among the lowest median validity scores (60) as well. Highest variation in reliability

267 scores between trained users (SD) was found for QBA (32), and lowest variation for ‘Body  
268 condition score’(10). For validity scores, highest variation between trained users was found  
269 for validity scores of ‘water flow’(28) and lowest for integument alterations (8). For ranking,  
270 scores for ‘Tail docking method’, ‘Head butts and displacements’ and ‘Avoidance distance  
271 test’ (9) were most variable and scores for ‘Mortality’ and ‘Integument alterations’ were least  
272 variable (4).

273

274 <Table 3>

275

276 The importance rank of the measure was negatively associated with both the reliability and  
277 validity scores, although validity had a somewhat higher estimate (i.e. higher importance as  
278 indicated by a lower ranking was associated with higher reliability and validity scores) ( $P =$   
279  $0.03$  for both, estimates  $-0.66$  and  $-0.74$ , respectively, adjusted  $R^2 = 0.20$ ). A very small but  
280 significant interaction was found between reliability and validity scores, where they did not  
281 strengthen each other’s negative effect on ranking ( $P = 0.048$ , estimate =  $-0.009$ ).

### 282 ***Predicting Overall Welfare Categorization Using WQ Measures***

283 When analyzed univariately, 20 out of 41 measures significantly ( $P < 0.05$ ) affected overall  
284 welfare categorization into the ‘enhanced’ category (Table 4), and 11 measures significantly  
285 affected categorization into the ‘not classified’ category for the entire European dataset ( $n =$   
286 491).

287 <Table 4>

288 The multivariable model that had the fewest variables whilst maintaining the highest % of  
289 both sensitivity and specificity (67% and 85%, respectively) for the ‘enhanced’ category  
290 contained the following measures (from most to least influence): ‘at least two drinkers/cow’,  
291 ‘water flow’, ‘% of animals lying outside the lying area’, ‘mean time needed to lie down’

292 'drinker cleanliness', and '% of animals with at least one lesion/swelling' (Table 5). For not-  
293 classified, the measures (from most to least influence) 'at least two drinkers/cow', 'number of  
294 lean cows', 'QBA-index' and 'number of displacements/cow/h.' contributed to the model  
295 with fewest variables but the highest sensitivity (44%) and specificity (100%).  
296 <Table 5>

### 297 *Comparing WQ Overall Welfare Category and Criteria Scores with Trained-user Opinion*

298 For 2 of 5 'acceptable' herds and for 1 of 2 'enhanced' herds, the majority of trained users (n  
299 = 8) scored in accordance with WQ (Figure 3). Regarding scores that were not in accordance  
300 with WQ, the vast majority were a lower category than the WQ calculation (25 of 29 expert  
301 scores). Overall, ICC for overall welfare scores by trained users was 0.5.

302 <Figure 3>

303 The criteria 'absence of injuries', 'absence of pain induced by management procedures',  
304 'expression of social behavior' and 'good human-animal relationship' were systematically  
305 scored lower by trained users than the WQ score (Table 6). The expert and WQ scores were  
306 not significantly related for two criteria: 'absence of prolonged thirst' and 'absence of  
307 prolonged hunger' (Table 6). The correspondence between trained users was insufficient (ICC  
308 < 0.6) for two criteria, namely 'absence of injuries' and 'absence of disease'. The number of  
309 measures within a criterion tended to be negatively related to ICC (P = 0.06, estimate = -  
310 0.04).

311 <Table 6>

312

## DISCUSSION

313 This study gives insight into the relation of integrated scores of the WQ dairy cattle protocol  
314 with trained-user opinion. The specific research design imposes some limitations, but also  
315 provides challenges for future research. For example, we chose to only select dairy cattle

316 welfare trained users who were trained users of the WQ dairy cattle protocol. This ensured  
317 that trained users had a proper knowledge of the protocol and all measures, but limited the  
318 number of possible respondents. The results show discrepancies between trained-user opinion  
319 and WQ scores.

### 320 *Trained-user Opinion on Ranking, Reliability and Validity of Measures*

321 The measures that the trained users ranked highest in terms of perceived importance for the  
322 overall welfare status of a herd (viz. ‘lameness score’, ‘body condition score’, ‘mortality rate’  
323 and ‘integument alterations’) are in agreement with earlier studies in which dairy cattle  
324 welfare trained users were asked to score the importance of welfare measures (Nielsen et al.,  
325 2014; Lievaart and Noordhuizen, 2011; Whay et al., 2003). Reliability and validity scores  
326 both influenced ranking positively (based on the negative relationship between reliability and  
327 validity scores and ranking), but did not positively interact. This means that highest ranked  
328 measures in the current study did not necessarily receive the highest validity *and* reliability  
329 scores. In addition, although the set-up of this study was such that trained users had to  
330 consider validity and reliability before ranking, other (unknown) factors appeared to influence  
331 the trained users’ opinion on the importance of the various measures for overall herd welfare  
332 as well (further supported by the models’ low  $R^2$  of 0.20). This was the case for lameness, for  
333 example, which was ranked highest for importance although its reliability was among the  
334 lowest.

335

336 Overall, QBA was scored among the lowest by the trained users with regard to reliability and  
337 validity (although still within the ‘acceptable’ range) and was ranked lowest on importance  
338 for dairy cattle welfare status. The QBA is a method that uses descriptors such as ‘frustrated’  
339 or ‘content’, to interpret the behavior and body language of an animal by integrating these  
340 details of animal behavior into a qualitative judgment of overall welfare state (Wemelsfelder,

341 2001; Rousing and Wemelsfelder, 2006; Wemelsfelder, 2007). Inter-observer reliability was  
342 tested and deemed acceptable for a QBA method using ‘free’ descriptors (i.e. not set but  
343 determined by observers themselves) and was validated by correlating results to behavioral  
344 observations (Rousing and Wemelsfelder, 2006; Napolitano et al., 2012). The fixed-term-  
345 method and specific set of descriptors used in the WQ protocol were tested for inter-observer  
346 reliability in a study by Bokkers et al., (2012) and judged as not satisfactory by the authors  
347 involved (i.e. Kendall’s coefficient of concordance  $< 0.7$ ), whereas Wemelsfelder et al. (2009)  
348 reported satisfactory observer agreement in beef, dairy cattle and veal calves of those  
349 descriptors. In addition, recently published papers demonstrated internal validity by testing  
350 correlation between QBA and other behavioral and physiological measures (Coignard et al.,  
351 2014; Phythian et al., 2016; Serrapica et al., 2017).

352

353 While some measures scored highest for reliability, they scored lowest for validity, e.g.  
354 measures related to the criterion ‘absence of prolonged thirst’ (‘centimeters of trough per  
355 cow’), or were ranked lowest on importance for dairy cattle welfare (‘water flow’). Criticism  
356 expressed in earlier studies for these measures is related to their resource-based nature and the  
357 impact these specific measures have on the WQ integrated scores, while preference generally  
358 shall be given to animal-based measures (de Vries et al., 2013; Heath et al., 2014; Buijs et al.,  
359 2016). Measuring functioning of water points, water provision and water cleanliness refers to  
360 assessing a risk for cows being in a certain welfare state and may therefore in some cases not  
361 be the most valid measure of an actual welfare state in dairy cattle, in this case due to  
362 prolonged thirst. Additionally, to our knowledge, no actual validity testing of the WQ drinker  
363 measures has occurred. This could explain the relatively low perceived validity score attached  
364 by the trained users to these measures. Further testing of reliability and validity on certain  
365 measures is needed, based on the results of the current study and previous research (Knierim



366 and Winckler, 2009). If from such studies it appears that measures are not sufficiently reliable  
367 or valid, then research should be performed to propose improved measures.

368

369 The trained users did not always agree on the relative importance for the overall welfare status  
370 of dairy herds of different welfare measures (given the high variations in ranking and  
371 reliability and validity scores between trained users). This possibly reflects diverging views in  
372 what trained users find most important for dairy cattle welfare, as Fraser et al., (1997) showed  
373 in his study on animal welfare conceptualization among animal welfare scientists. This  
374 indicates that when using trained-user opinion to determine weights for various measures,  
375 such variation should be accounted for when selecting the expert panel. Therefore, it is not  
376 likely that an overall welfare score will always perfectly reflect an individual trained users'  
377 opinion. Methods to achieve more consensus among trained users exist. Examples are  
378 deliberative processes using a workshop like performed by Rodenburg et al. (2008), or more  
379 complex processes like a 'Delphi' method with multiple rounds of expert elicitation and  
380 feedback (Linstone and Turoff, 1975).

381

### 382 *Comparison of the measures' impact on overall welfare categorization and trained-user* 383 *opinion*

384 Compared to previous studies (Heath et al., 2014; Buijs et al., 2016), more measures affected  
385 both the 'enhanced' and the 'not classified' categorization in the current study. This is likely  
386 due to a larger variation in data in the current study which used a much larger (and diverse, as  
387 data was collected in more than one country) database compared to both other studies. To  
388 specify, the current sample comprised of 491 herds, as opposed to 92 herds and 22 flocks for  
389 Heath et al., (2014) and Buijs et al., (2016) respectively. In accordance with Heath et al.  
390 (2014) drinker measures had the biggest influence for both the enhanced and not classified

391 models, while in the current study these received some of the lowest ranks and/or validity  
392 scores by the trained users. Additionally, the QBA score which scored lowest overall was  
393 among the best predictors for the ‘not classified’ categorization. By contrast, although there is  
394 often little agreement among trained users on importance of various welfare measures, some  
395 measures which are regarded as highly important to cattle welfare by certain welfare trained  
396 users, did not have a great influence on the overall welfare status categorization. For example,  
397 although ‘lameness score’ and ‘mortality rate’ contributed to the ‘enhanced’ categorization in  
398 univariate models, they did not when combined into a multivariable model. These results  
399 show that the relative influence of measures on WQ integrated scores may not be in  
400 accordance with the trained users’ opinion of this study. We tested this by comparing expert  
401 scoring of WQ criteria and overall welfare with calculated WQ scores.

#### 402 ***Comparing WQ Integrated Scores with Trained-user Opinion***

403 ***Overall welfare category.*** For only three out of the seven herds, the majority of trained  
404 users scored in accordance with the WQ overall welfare categorization. The two herds that  
405 were scored as ‘not classified’ by at least half of the trained users (Herds 3 and 7) both scored  
406 badly (i.e. relatively high prevalence) on measures that were ranked as highly important by  
407 the trained users, namely lesions/swellings and moderately lame cows.

408 Variation between trained users was shown for the overall welfare scoring, given the  
409 relatively low ICCs. This was also shown for criteria scores, where ICCs tended to be lower  
410 for criteria which contain the most measures. This can indicate that 1) trained users did not  
411 agree on their assessment of overall welfare caused by a different view of animal welfare (as  
412 mentioned above) and/or 2) some trained users may have had difficulties in aggregating many  
413 welfare measures into one overall score. The latter explanation is supported by that fact that  
414 six of the 14 trained users who completed the questions on criterion scores, did not complete  
415 the question on overall welfare scores.

416            *Criteria scores.* The criteria ‘absence of injuries’, ‘absence of pain induced by  
417 management procedures’, ‘expression of social behavior’ and ‘good human-animal  
418 relationship’ were systematically scored lower by trained users than the WQ integrated scores.  
419 In the WQ protocol, poor scores have more influence on integrated scores than good scores  
420 (Buijs et al., 2016). Therefore, lower scores on each of these criteria would have a major  
421 effect on principle scores and overall welfare category.

422            The correspondence between the expert and WQ score for the criterion ‘absence of  
423 prolonged thirst’ was extremely low. The finding that the trained users considered some of  
424 these measures of relatively poor validity may partly explain this lack of correspondence. It is  
425 a strong indication that trained users of the present study did not agree with the way that the  
426 criterion score for ‘absence of prolonged thirst’ is calculated in the WQ protocol.

427            Four complementary explanations can be put forward for the poor correspondence  
428 between trained users’ scores and WQ integrated scores. First, except for the first step of the  
429 integration procedure, WQ consulted a much wider group of stakeholders (including animal  
430 scientists, social scientists, producers and retailers.) than we did in the current study. These  
431 stakeholders’ views on the relative impact of the various measures on dairy cattle welfare may  
432 differ substantially from those of the trained users in the current study. We opted to limit the  
433 current study to trained users only, because it could be argued that they are best qualified to  
434 assess overall dairy cattle welfare state and the relative importance of the various WQ  
435 measures.

436            Second, as the protocol was not yet published when stakeholder opinion was elicited  
437 during the WQ project, they could not have gained as much experience in performing the  
438 various WQ measures as the trained users in this study. It has previously been shown, that  
439 detailed information on welfare measures (e.g. practical implications) can significantly

440 influence relative weight attributed by trained users to these welfare measures (Rodenburg et  
441 al., 2008).

442 Third, there was considerable variation between trained users in the present study  
443 regarding importance ranking, although there is no information readily available on the degree  
444 of variation between the original WQ trained users. The variation in prioritizing certain  
445 aspects of welfare in the current group of trained users could arise from different concepts of  
446 animal welfare, like Fraser et al., (2008) described as ‘basic health and functioning’, ‘natural  
447 living’, and ‘affective states’.

448 Fourth, WQ integration methods likely contribute to differences between trained-user  
449 opinion and WQ integrated scores. De Graaf et al., (2016) identified two factors which  
450 influence the impact a measure has on the integrated WQ scores, but which seem unintended  
451 by the WQ consortium. Namely, 1) the number of integrated measures per criterion or  
452 principle, and 2) the various aggregation methods of measures into criteria scores which  
453 influence the impact individual measures have on integrated scores. In the present study a low  
454 level of correspondence between welfare measures which impact WQ categorization most and  
455 which were scored as most important by trained users was found. Also, poor correspondence  
456 between trained-user opinion and some criterion scores indicated that this lack of  
457 correspondence already starts in the first step of integration.

458 These findings indicate a lack of correspondence between WQ-welfare scores and  
459 trained users’ assessment of herd welfare. The opinion of these trained users is the only ‘silver  
460 standard’ we have to validate animal welfare integrated scores, since they are arguably best  
461 equipped to assess and quantify the welfare of a given herd. Moreover, these trained users  
462 may be considered authorities for animal welfare assessment in society, and it is important  
463 that scientists who use this method support it. Future research could focus on determining  
464 whether the way trained users assess welfare is in correspondence with other stakeholders’

465 assessment. Improvements for WQ may be derived from the observed discrepancies between  
466 WQ overall welfare assessment and that of the trained users. In some cases, the trained users  
467 scored lower than WQ and in other cases (water provision) they were less stringent. Because  
468 WQ allocates more weight to low scores this is likely to have a significant impact on the  
469 overall assessment. For example, higher criterion-scores for absence of thirst (following our  
470 trained users' opinion) would reduce the impact of this criterion on the overall assessment. On  
471 the contrary, lameness should be given more impact since our trained users ranked this as  
472 highly important.

### 473 **CONCLUSION**

474 Trained-user opinion on the most and least important measures for the overall welfare status  
475 of a herd did not correspond well with the influence of these measures on the WQ overall  
476 welfare categorization. Some of the measures that were ranked as least important for herd  
477 welfare by trained users (e.g. measures relating to drinkers) had the highest influence on the  
478 WQ overall welfare categorization. On the contrary, measures ranked as most important by  
479 the trained users (e.g. lameness and mortality) had a lower impact on the WQ overall  
480 category. In addition, results indicate poor correspondence between trained users' scoring and  
481 6 of 11 WQ-criteria and the overall welfare category. In both cases, trained users mostly  
482 allocated more negative scores, indicating a lower level of welfare. The WQ scores of the  
483 protocol for dairy cattle thus lacked correspondence with selected trained users on the  
484 importance of several welfare measures.

### 485 **ACKNOWLEDGEMENTS**

486 We thank all trained users who filled out the survey and Miriam Levenson for language  
487 editing.

488

## REFERENCES

489 Bokkers, E. A. M., M. de Vries, I. C. M. A. Antonissen and I. J. M. de Boer. 2012. Inter-and  
490 intra-observer reliability of experienced and inexperienced observers for the Qualitative  
491 Behavior Assessment in dairy cattle. *Anim. Welf.* 21:307-318.

492

493 Blokhuis, H. J., R. B. Jones, R. Geers, M. Miele and I. Veissier. 2003. Measuring and  
494 monitoring animal welfare: Transparency in the food product quality chain. *Anim. Welf.*  
495 12:445-455.

496

497 Blokhuis, H. J., I. Veissier, M. Miele and B. Jones. 2010. The Welfare Quality® project and  
498 beyond: Safeguarding herd animal well-being. *Acta Agric. Scand. A Anim. Sci.* 60:129-140.

499

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

502

503 Botreau, R., J. Capdeville, P Perny, & I. Veissier. 2008. Multicriteria evaluation of animal  
504 welfare at farm level: an application of MCDA methodologies. *Found. Comp. Decis. Sci.* 33:  
505 287-317.

506

507 Botreau, R., I. Veissier and P. Perny. 2009. Overall assessment of animal welfare: strategy  
508 adopted in Welfare Quality®. *Anim. Welf.* 18:363-370.

509

510 Buijs, S., B. Ampe and F. A. M. Tuytens. 2016. Sensitivity of the Welfare Quality® Broiler  
511 chicken protocol to differences between intensively reared indoor flocks: which factors  
512 explain overall categorization? *Animal* 15:1-10.

513

514 de Boyer des Roches, A., I. Veissie, M. Coignard, N. Bareille, R. Guatteo, J. Capdeville, E.

515 Gilot-Fromont and L. Mounier. 2014. The major welfare problems of dairy cows in French

516 commercial farms: an epidemiological approach. *Anim. Welf.* 23: 467-478.

517 Coignard M, R. Guatteo, I. Veissier, A. Lehebel, C. Hoogveld, L. Mounier and N. Bareille.

518 2014. Does milk yield reflect the level of welfare in dairy herds? *Vet. J.* 199:184-187.

519

520 de Graaf, S., B. Ampe, S. Buijs, S. N. Andreasen, A. De Boyer Des Roches, F. J. C. M. van

521 Eerdenburg, M. J. Haskell, M. K. Kircher, L. Mounier, M. Radeski, C. Winckler, J. Bijttebier,

522 L. Lauwers, W. Verbeke and F. A. M. Tuyttens. 2016. Sensitivity of the integrated Welfare

523 Quality® scores of the dairy cattle protocol to changes in individual measures. Page 12 in

524 Proc. Benelux ISAE Conf 2016, Berlicum, The Netherlands.

525

526 de Jong, I. C., V. A. Hindle, A. Butterworth, B. Engel, P. Ferrari, H. Gunnink, T. P. Moya, F.

527 A. M. Tuyttens and C. G. Van Reenen. 2016. Simplifying the Welfare Quality® assessment

528 protocol for broiler chicken welfare. *Animal* 10, 117-127.

529

530 de Vries, M., E. A. M. Bokkers, G. van Schaik, R. I. Botreau, B. Engel, T. Dijkstra and I. J.

531 M. de Boer. 2013. Evaluating results of the Welfare Quality multi-criteria evaluation model

532 for categorization of dairy cattle welfare at the herd level. *J. Dairy Sci.* 96:6264-6273.

533

534 Fraser, D., D. M. Weary, E. A. Pajor and B. N. Milligan. 1997. A scientific conception of

535 animal welfare that reflects ethical concerns. *Anim. Welf.* 6:187-205.

536

537 Fraser, D. 2008. Understanding animal welfare. *Acta Vet. Scand.* 50, 1.

538

539 Heath, C. A. E., W. J. Browne, S. Mullan and D. C. J. Main. 2014. Navigating the iceberg:  
540 reducing the number of parameters within the Welfare Quality<sup>®</sup> assessment protocol for dairy  
541 cows. *Animal* 8:1978-1986.

542

543 Knierim, U., and C. Winckler. 2009. On-farm welfare assessment in cattle: validity, reliability  
544 and feasibility issues and future perspectives with special regard to the Welfare Quality<sup>®</sup>  
545 approach. *Anim. Welf.* 18:451-458.

546

547 Lievaart, J. J. and J. P. T. M. Noordhuizen. 2011. Ranking experts' preferences regarding  
548 measures and methods of assessment of welfare in dairy herds using Adaptive Conjoint  
549 Analysis. *J. Dairy Sci.* 94:3420-3427.

550

551 Linstone, H. A. and M. Turoff. 1975. *The Delphi method: Techniques and applications.*  
552 Addison-Wesley, London, UK.

553

554 Mason, G. and M. Mendl. 1993. Why is there no simple way of measuring animal  
555 welfare?. *Anim. Welf.* 2:301-319.

556

557 Napolitano, F., G. De Rosa, F. Grasso and F. Wemelsfelder 2012. Qualitative behavior  
558 assessment of dairy buffaloes (*Bubalus bubalis*). *Appl. Anim. Behav. Sci.* 141:91-100.

559

560 Nielsen, B. H., A. Angelucci, A. Scalvenzi, B. Forkman, F. Fusi, F. A. M. Tuyttens, H. Houe ,  
561 H. Blokhuis, J. T. Sørensen, J. Rothmann, L. Matthews, L. Mounier, L. Bertocchi, M.

562 Richard, M. Donati, P. P. Nielsen, R. Salini, S. de Graaf, S. Hild, S. Messori, S. S. Nielsen, V.



563 Lorenzi, X. Boivin P. T. and Thomsen. 2014. Use of animal based measures for the  
564 assessment of dairy cow welfare-ANIBAM. EFSA External scientific report.  
565

566 Main, D. C. J., H. R. Whay, L. E. Green and A. J. F. Webster. 2003. Effect of the RSPCA  
567 Freedom food scheme on the welfare of dairy cattle. *Vet. Rec.* 153: 227-231.

568 Phythian, C. J., E. Michalopoulou, P. J. Cripps, J. S. Duncan and F. Wemelsfelder. 2016. On-  
569 farm qualitative behaviour assessment in sheep: Repeated measurements across time, and  
570 association with physical indicators of flock health and welfare. *Appl. Anim. Behav. Sci.*  
571 175:23-31.

572

573 Rodenburg, T. B., F. A. M. Tuytens, K. De Reu, L. Herman, J. Zoons and B. Sonck. 2008.  
574 Welfare assessment of laying hens in furnished cages and non-cage systems: assimilating  
575 trained user opinion. *Anim. Welf.* 17:355-361.  
576

577 Rousing, T. and F. Wemelsfelder. 2006. Qualitative assessment of social behavior of dairy  
578 cows housed in loose housing systems. *Appl. Anim. Behav. Sci.* 101:40-53.  
579

580 Serrapica, M., X. Boivin, M. Coulon, A. Braghieri and F. Napolitano. 2017. Positive  
581 perception of human stroking by lambs: Qualitative behaviour assessment confirms previous  
582 interpretation of quantitative data. *Appl. Anim. Behav. Sci.* 187:31-37.  
583

584 Sørensen, J. T. and D. Fraser. 2010. On-farm welfare assessment for regulatory purposes:  
585 Issues and possible solutions. *Livest. Sci.* 131:1-7  
586

587 Spoolder, H., G. De Rosa, B. Horning, S. Waiblinger and F. Wemelsfelder. 2003. Integrating  
588 parameters to assess on-farm welfare. *Anim. Welf.* 12:529-534.

589

590 Tuytens, F. A. M., F. Vanhonacker, E. Van Poucke and W. Verbeke. 2010. Quantitative  
591 verification of the correspondence between the Welfare Quality® operational definition of  
592 farm animal welfare and the opinion of Flemish farmers, citizens and vegetarians. *Livest. Sci.*  
593 131: 108-114.

594

595 Tuytens, F. A. M., S. de Graaf, J. L. Heerkens, L. Jacobs, E. Nalon, S. Ott, L. Stadig, E. Van  
596 Laer and B. Ampe. 2014. Observer bias in animal behavior research: can we believe what we  
597 score, if we score what we believe?. *Anim. Behav.* 90:273-280.

598

599 Whay, H. R., D. C. J. Main, L. E. Green and A. J. F. Webster. 2003. Assessment of the  
600 welfare of dairy cattle using animal-based measurements: direct observations and  
601 investigation of farm records. *Vet. Rec.* 153: 197-202.

602

603 Wemelsfelder, F. and A. B. Lawrence. 2001. Qualitative assessment of animal behaviour as  
604 an on-farm welfare-monitoring tool. *Acta Agric. Scand. A Anim. Sci.* 51:21-25.

605

606 Wemelsfelder, F., F. Millard, G. De Rosa, and F. Napolitano. 2009. Qualitative Behaviour  
607 Assessment, in: Forkman, B., Keeling, L.J. (Eds.), *Welfare Quality Reports No. 11.*, Welfare  
608 Quality Consortium, Lelystad, the Netherlands, 215-224

609

610 Wemelsfelder, F. 2007. How animals communicate quality of life: the qualitative assessment  
611 of behaviour. *Anim. Welf.* 16:25-31.

612

613 Welfare Quality® Consortium. 2009. *Welfare Quality® Assessment Protocol for Cattle.*

614 Lelystad, The Netherlands. <http://www.welfarequalitynetwork.net/> accessed on: 4-11-2016.

615

616 Whay, H. R., D. C. J. Main, L. E. Greent and A. J. F. Webster. 2003. Animal-based measures  
617 for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of trained  
618 user opinion. *Anim. Welf.* 12:205-217.

619 *Table 1: Principles, the corresponding criteria and measures used in the Welfare Quality®*

620 *assessment protocol for dairy cows*

Principles	Criteria	Measures
Good feeding	Absence of prolonged hunger	Body Condition Score (% very lean animals)
	Absence of prolonged thirst	Availability & cleanliness water
Good housing	Comfort around resting	Lying down duration; collisions during lying down; on edge/outside of lying area; cleanliness
	Thermal comfort	No measure for dairy cattle
	Ease of movement	Free stalls or presence of tethering and exercise
Good health	Absence of injuries	Lameness; integument alterations
	Absence of disease	Respiration/digestive diseases; mastitis; mortality; dystocia, downer cows
	Absence of pain induced by management procedures	Mutilations (dehorning; tail docking; use of anesthetics/analgesics)
Appropriate behavior	Expression of social behavior	Incidence agonistic interactions
	Expression of other behaviors	Access to pasture
	Good human-animal relationship	Avoidance distance at feeding place
	Positive emotional state	Qualitative Behavioral Assessment

621

622 *Table 2: Measure values of each of the seven herds presented to trained users in the survey*

Criteria, measures	Herd #	1	2	3	4	5	6	7
<b>Absence of prolonged hunger</b>								
% of lean cows		0	3	17	5	11	3	24
<b>Absence of prolonged thirst</b>								
Number of water bowls/cow		0.5	0.0	0.0	0.0	0.0	0.6	0.05
Trough length/cow (cm)		0.0	7.9	4.7	28.6	9.0	0.0	0.0
Drinker cleanliness		Yes	Yes	No	Yes	Yes	Yes	Yes
At least 2 drinkers/cow		No	Yes	No	No	Yes	No	Yes
<b>Resting comfort</b>								
Mean time needed to lie down (s)		4.6	4.6	7.5	4.1	6.6	5.4	6.8
% of cows colliding with housing equipment		16	15	72	0	37	8	33
% of cows lying outside of lying area		50	11	0	0	0	35	0
% of cows with dirty flanks		34	55	81	14	67	79	70
% of cows with dirty lower legs		57	37	85	38	20	79	100
% cows with a dirty udder		18	21	77	10	42	48	95
<b>Ease of movement</b>								
Loose (L) or tied (T) housing		T	L	L	L	L	T	L
<b>Absence of injuries</b>								
% of moderately lame cows		0	13	88	0	23	0	84
% of severely lame cows		32	0	12	10	17	27	5
% of cows with at least one lesion		7	12	72	28	13	20	68
% of cows with no lesions but at least one hairless patch		98	18	28	38	21	100	32
<b>Absence of disease</b>								
Number of coughs/cow/minute		0.05	0.00	0.13	0.10	0.06	0.17	0.00
% cows with nasal discharge		59	0	0	0	5	18	0
% cows with ocular discharge		0	0	0	0	0	0	0
% cows with hampered respiration		0	0	0	5	0	0	0
% cows with diarrhea		5	0	0	0	0	0	16
% cows with vulvar discharge		0	0	0	0	0	0	3
% cows with SCC <sup>1</sup> > 400.000		8	21	25	0	14	8	12
% cows mortality		5	3	4	0	4	3	4
% calvings with dystocia		0	21	0	0	1	6	3
% downer cows		0	6	0	0	0	6	5
<b>Absence of pain induced by management procedures</b>								
Dehorning method, Thermal (T), Caustic paste (P) or None (N)		T	P	P	N	P	P	T
Use of Analgesics		No	Yes	Yes	No	Yes	No	No
Use of Anesthetics		No	Yes	Yes	No	Yes	No	Yes
<b>Expression of social behavior</b>								
Number of Head butts/cow/15 min.		0.8	4.0	0.7	0.0	0.4	0.4	1.0
Number of Displacements/cow/15 min.		0	1.2	0.1	0.0	0.2	0.0	0.8

---

min.							
Expression of other normal behavior							
Number of hours on pasture	214	180	0	0	0	214	195
Number of days on pasture	19	9	0	0	0	8	9
Human-animal relationship							
% of cows that could be touched	36	55	59	100	55	44	30
% closer than 50 cm but not touched	11	36	37	0	26	2	35
% between 50 and 1 m	23	9	9	0	11	14	24
% > 1 m	30	0	0	0	9	41	11
Positive emotional state							
QBA <sup>2</sup> score	43	40	8	91	77	66	54

---

623 <sup>1</sup>Somatic Cell Count

624 <sup>2</sup>Qualitative Behaviour Assessment

625 *Table 3: Median (interquartile range) reliability and validity scores and rankings for each WQ*  
 626 *measure by trained users*

	Reliability score (n = 15)	Validity score (n = 15)	Ranking (n = 13)
Body condition score	89 (11)	79 (35)	4 (8)
Centimeters of trough per cow (minimum 6 cm), number of water bowls per cow (minimum 0.10) and at least two drinkers available for each cow	93 (15)	50 (34)	13 (6)
Water cleanliness, judged visually	80 (28)	70 (36)	19 (9)
Water flow	90 (33)	60 (40)	20 (15)
Time needed to lie down	75 (38)	78 (21)	9 (7)
Cows colliding with housing	70 (39)	82 (28)	16 (10)
Cows lying outside of lying area	85 (33)	89 (28)	16 (10)
Cleanliness of udders, flanks and lower legs	75 (12)	81 (24)	15 (5)
Tied versus loose housing	100 (6)	84 (28)	11 (13)
Lameness score	69 (36)	89 (11)	2 (2)
Integument alterations	75 (15)	90 (14)	7 (4)
Coughing	69 (44)	75 (35)	19 (13)
Nasal discharge	84 (35)	80 (11)	18 (8)
Ocular discharge	85 (31)	80 (12)	18 (11)
Hampered respiration	88 (36)	86 (12)	21 (12)
Diarrhea	75 (21)	70 (22)	15 (8)
Vulvar discharge	77 (39)	86 (14)	18 (8)
Somatic cell count >400.000	83 (19)	81 (11)	13 (14)
Mortality	79 (47)	81 (16)	7 (6)
Dystocia	79 (37)	80 (17)	13 (10)
Downer cows	79 (47)	81 (16)	15 (14)
Dehorning method	90 (26)	86 (16)	11 (10)
Tail docking method	95 (16)	88 (17)	17 (18)
Head butts and displacements	70 (26)	75 (17)	14 (16)
Access to pasture (number of hours and number of days on pasture)	90 (18)	75 (33)	19 (8)
Avoidance distance test	66 (24)	76 (28)	17 (15)
Qualitative Behavior Assessment	60 (37)	57 (20)	22 (11)

627

628

629 *Table 4: P-values of the univariate logistic regression models examining predictability of*  
630 *single measures for a herd to be categorized as ‘Enhanced’ or ‘Not classified’ based on the*  
631 *collated European dataset (n = 491)*

Criteria, Measures	Enhanced	Not classified
<b>Absence of prolonged hunger</b>		
% of lean cows	<0.001	<0.001
<b>Absence of prolonged thirst</b>		
Number of water bowls	0.070	0.863
Water flow	<0.001	0.505
Trough length/cow (cm)	0.001	0.008
At least 2 drinkers/cow	<0.001	0.006
Drinker cleanliness	<0.001	0.068
<b>Resting comfort</b>		
Mean time needed to lie down	<0.001	0.577
% of cows colliding with housing	<0.001	0.365
% of cows lying outside of lying area	<0.001	0.014
% of cows with dirty flanks	0.101	0.172
% of cows with dirty lower legs	0.023	0.110
% cows with a dirty udder	0.374	0.258
<b>Ease of movement</b>		
Loose or tied housing	<0.001	0.016
<b>Absence of injuries</b>		
% of moderately lame cows	0.002	0.392
% of severely lame cows	<0.001	0.096
% of cows with at least one lesion/swelling	<0.001	0.014
% of cows with at least one hairless patch	0.141	0.075
<b>Absence of disease</b>		
Number of coughs/cow/minute	0.168	0.350
% cows with nasal discharge	0.092	0.165
% cows with ocular discharge	0.044	0.426
% cows with hampered respiration	0.293	0.385
% cows with diarrhea	0.386	0.546



% cows with vulvar discharge	0.588	0.936
% cows with SCC >400.000	0.130	0.014
% cows mortality	<0.001	0.189
% calvings with dystocia	0.619	0.841
% downer cows	0.742	0.423
Absence of pain induced by management procedures		
Method dehorning	0.130	0.021
Use of analgesics during/after dehorning	0.618	0.540
Use of anesthesia during dehorning	0.759	0.110
Method tail docking	0.150	0.974
Use of analgesics during/after tail docking	0.011	0.008
Use of anesthesia during tail docking	0.025	0.010
Expression of social behavior		
Head butts/cow/15 min.	0.033	0.759
Displacements/cow/15 min.	0.615	0.159
Expression of other normal behavior		
Number of hours on pasture	0.467	0.153
Number of days on pasture	0.810	0.454
Human-animal relationship		
% of cows that could be touched	0.711	0.188
% of cows that can be approached < 50 cm but not touched	0.012	0.379
% of cows that can be approached by 50 – 1 m	0.253	0.924
% of cows that can't be approached (> 1 m)	0.011	0.547
Positive emotional state		
QBA index score	0.079	<0.001

633 *Table 5: P-values and model estimates of measures in the multivariate logistic regression*  
634 *models predicting a herd to be categorized as 'Enhanced' or 'Not classified' based on the*  
635 *collated European dataset (n = 491)*

Outcome variables	Enhanced model		Not classified model	
	Estimate	P-value	Estimate	P-value
Number of lean cows	-	-	1.8	<0.001
Water flow	1.1	<0.001	-	-
At least 2 drinkers/cow	2.4	<0.001	-3.7	0.007
Drinker cleanliness	0.6	<0.001	-	-
Mean time needed to lie down	-0.7	<0.001	-	-
% of cows lying outside of lying area	-0.9	<0.001	-	-
% of cows with at least one lesion/swelling	-0.5	<0.001	-	-
Number of displacements/cow/h.	-	-	0.7	0.043
QBA index score	-	-	-1.6	0.002

636

637 *Table 6: Systematic t-test P-value, Linear Regression R<sup>2</sup> and ICC of WQ integrated scores and trained*  
 638 *user median scores (n =14) for the focus herds (n = 7) for each WQ criterion*

Criteria	Median (IR) <sup>1</sup> WQ score	Median (IR) <sup>1</sup> expert score	Systematic t- test P-value	Regression R <sup>2</sup>	ICC
Absence of prolonged hunger	67 (39)	50 (75)	0.475	0.237	0.6
Absence of prolonged thirst	20 (97)	50 (71)	0.737	0.007	0.7
Comfort around resting	27 (20)	25 (33)	0.181	0.880**	0.8
Freedom of movement	100 (33)	90 (90)	0.125	1.000***	1.0
Absence of injuries	28 (19)	18 (29)	0.006	0.926***	0.5
Absence of disease	40 (32)	42 (34)	0.296	0.903**	0.4
Absence of pain induced by management procedures	58 (18)	10 (50)	0.023	0.521*	0.8
Expression of social behavior	84 (24)	58 (50)	0.020	0.869**	0.6
Expression of other normal behavior	73 (78)	60 (78)	0.828	0.978***	0.9
Good human-animal relationship	54 (37)	52 (50)	0.023	0.984***	0.7
Positive emotional state	54 (32)	50 (37)	0.901	0.997***	0.8

639 <sup>1</sup>IR = Interquartile Range

640 \*P < 0.05

641 \*\*P < 0.01

642 \*\*\*P < 0.001

643 **Figures**

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662 de Graaf, Figure 1

663

664

665

666

667

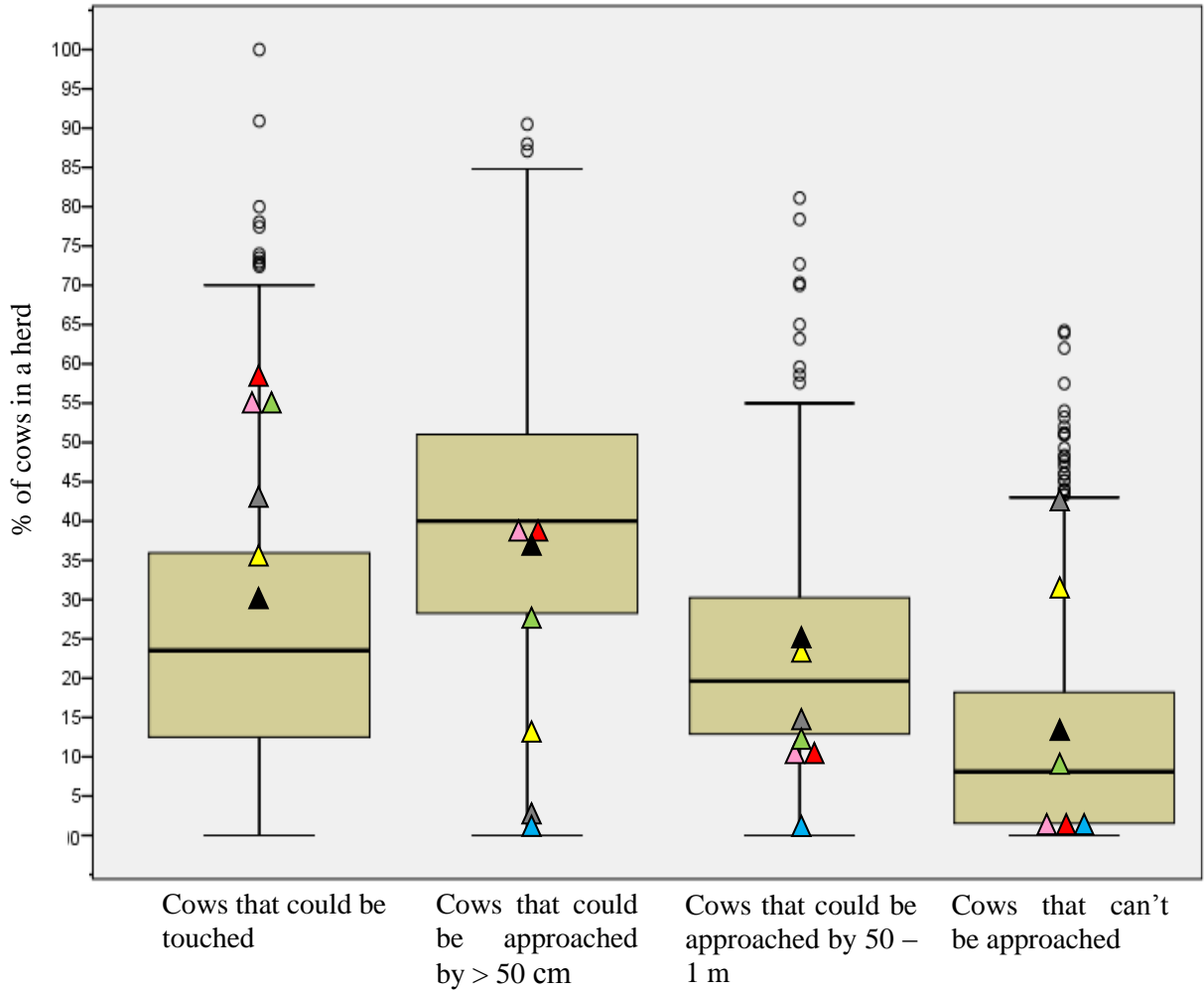
668

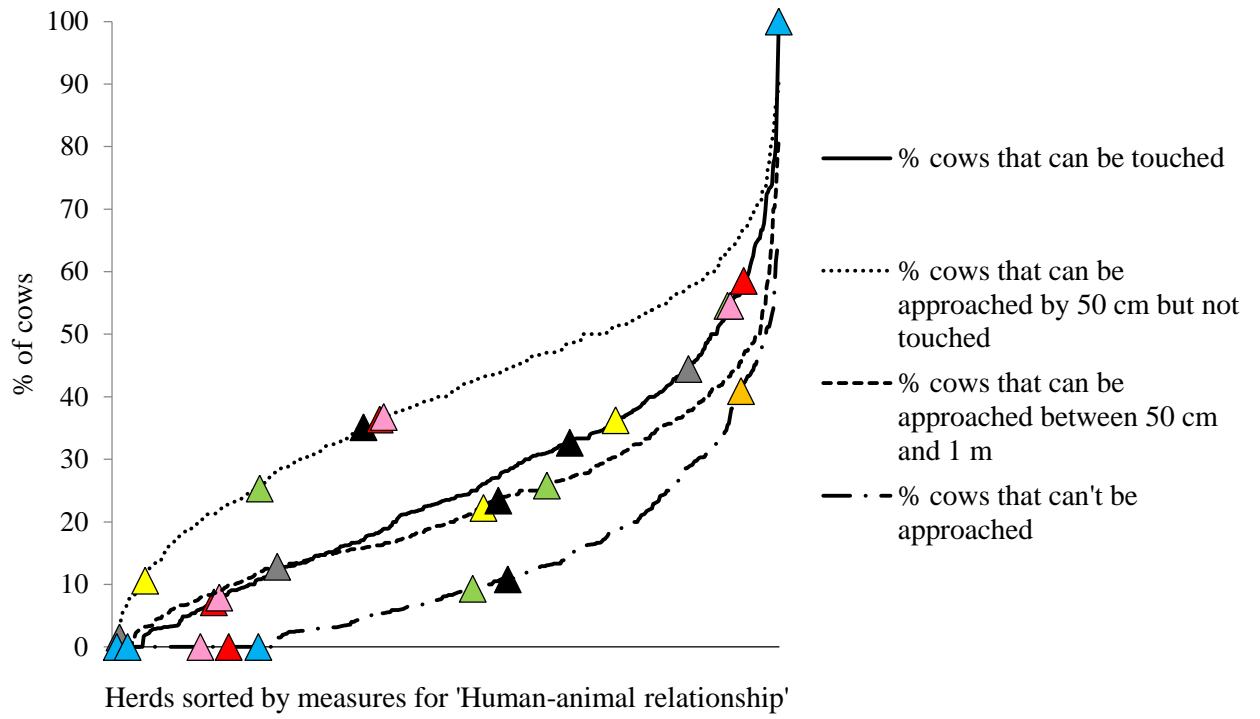
669

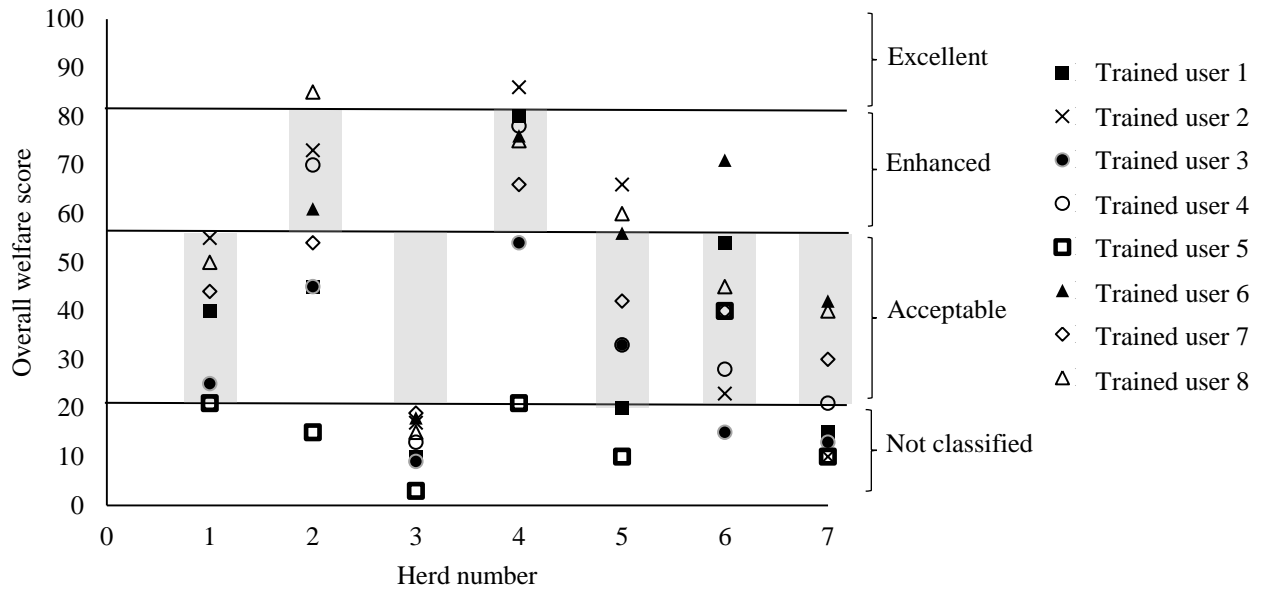
670

671

672







675

676 de Graaf, Figure 3

677 **Figure captions**

678 **Figure 1** Example boxplot figure from the survey among trained users, portraying the  
679 distribution of all herds in the database (n = 491) for the measures of the ADF, within the  
680 criterion 'Human-animal relationship'. Colored triangles mark the seven focus herds.

681

682 **Figure 2** Example figure from the survey among trained users, portraying the distribution of  
683 all herds in the database (n = 491) for the measures of the Avoidance Distance at the the Feed  
684 rack test (ADF), within the criterion 'Human-animal relationship'. Colored triangles mark the  
685 seven focus herds.

686

687 **Figure 3** Overall welfare score for all seven focus herds by eight trained users, grey boxes  
688 indicate WQ overall welfare category

689

690