# *Inconsistency in dairy calves' responses to tests of fearfulness*

**Inconsistency in dairy calves' responses to tests of fearfulness**

Rebecca K. Meagher, Marina A.G. von Keyserlingk, Dax Atkinson, Daniel M. Weary

Animal Welfare Program, University of British Columbia, 2357 Main Mall, Vancouver, BC, Canada V6T 1Z4

Corresponding author: Rebecca K. Meagher, Email: rkmeagher@gmail.com

## Abstract

Fear is an important welfare problem for farm animals including cattle. A variety of methods of assessing fear have been proposed, but the reliability and validity of these methods, and ways of improving these characteristics, have received little study. We conducted a series of experiments to assess the consistency of dairy calves' responses of novel objects and to humans, and to investigate factors that might improve reliability. In the first experiment, latency to touch a novel object had moderate reliability ($r_s$=0.54), and latency to touch a stationary, familiar human had negligible reliability ($r_s$=0.26). Experiment 2a used the same test protocols, but with a shorter interval between repeat testing and using different stimuli in the two novel object tests; this change did not improve reliability (e.g. $r_s$=0.29 for the novel-object test). Reliability for this test was improved ($r_s$=0.58) in Experiment 2b, when the same object was used in both tests rather than a truly novel object being used the second time. Experiment 2a found ceiling effects in the response to human test associated with the short period during which approach responses were recorded. High reliability was found in Experiment 2b, where the maximum test duration was doubled, but this effect not due to the extended duration. Experiment 3 assessed reliability of a response to human approach at the farm rather than individual level, in this case assessing responses to an unfamiliar person. The proportion of calves making contact with the person was not reliable ($r_s$=0.22), but the proportion retreating from the person had moderate reliability ($r_s$=0.52). Reliability was improved by excluding data from calves that had coughs on the day of testing. Conducting multiple tests per individual using different stimuli and reporting health status of the animals are recommended for future research and animal welfare assessment schemes that include measures of fear.



## Keywords

Fearfulness; neophobia; human-animal relationship; well-being; reliability; validity

## 1. Introduction

Fear is widely recognized as a welfare concerns for cattle and other farm animals (e.g. Farm Animal Welfare Council 2009; Hemsworth et al. 2000; Jones and Boissy 2011). Fearful animals can also cause production and management challenges, including decreased productivity (e.g. Barnett et al. 1992; Hemsworth et al. 2000) and animals that are afraid of humans may be more dangerous to handle (Boivin et al. 1992; Hemsworth et al. 1989). Unfortunately, methods of assessing fear (a negative emotional state resulting from a perceived threat [Gray 1987; Ennaceur 2014]) and fearfulness (a personality trait characterized by a tendency to express fear when exposed to potentially threatening stimuli or situations) appear not to be well-validated and have uncertain reliability (Forkman et al. 2007). Of 112 papers published in this journal over a five-year period ending in August 2015 with fear* or anx* in the keywords, abstract or title, only 65 papers (or 58%) contained any form of the words reliable or repeatable anywhere in the text, and of these, only 15 actually estimated reliability. Measures also vary considerably across studies, making it difficult to extrapolate results from one approach to the next (Forkman et al. 2007).

The need for valid, reliable ways of assessing welfare in farm animals is widely recognized, to be used for example in assurance schemes for commercial farms (see Scott et al. 2001). Currently, fear is often assessed in farm animals through response to novelty (neophobia, although other factors such as exploratory motivation also influence the response), most commonly using a novel object test. Another common type of fear-related test is in response to humans (e.g. Forkman and Keeling 2009), as fear of handlers may have a major impact on the lives of intensively farmed animals. Research published to date indicates that

62 responses are not closely associated in these two contexts (e.g. Hegelund and Sorensen 2007),
63 and that separate measures may be needed. From the perspective of animal welfare,
64 fearfulness and long-lasting states of fear are of special interest, meaning that we are especially
65 interested in fear responses that are consistent over time. Unfortunately, test-retest reliability
66 (also called repeatability) is often weak making it difficult to draw strong inferences from a single
67 test.

68 In cattle, for example, the novel object test was reported to be reliable within individuals
69 between tests in at least two calf studies (using measures derived from factor analysis in Van
70 Reenen et al. 2004, and approach latency in Bokkers et al. 2009), but was unreliable in older
71 heifers and adult cows when tested using avoidance (Van Reenen et al. 2013), reactivity
72 (Gibbons et al. 2009), number of interactions and time in proximity (Kilgour et al. 2006). Results
73 have been mixed across a range of measures and ages in other studies (Graunke et al. 2013;
74 MacKay et al. 2014). Even the methods of assessing 'repeatability' vary: while most studies
75 replicate the test exactly using the same stimulus, others (e.g. Gibbons et al. 2009) instead
76 assess consistency of response across different novel stimuli because there is no way to repeat
77 a test and have it be truly novel (see e.g. Forkman et al. 2007 for a discussion of this problem).
78 Nonetheless, the novel object test has face validity, meaning that it appears sensible based on
79 our understanding of fear and comparisons with human behaviour, as judged by experts (e.g.
80 Scott et al. 2001; Whay et al. 2003). It is also one of the few tests that has undergone some
81 successful validation for cattle, suggesting it may be a true indicator of fear (based on
82 correlation with other fear- and stress-related measures and pharmacological validation using
83 anxiolytic drugs; e.g. Van Reenen et al. 2005; Van Reenen et al. 2009). Confirming or finding
84 ways to improve its reliability would thus be valuable.

85 Responses to humans (typically measured as approach or avoidance by the animal) are
86 more consistently reported to be reliable (at the individual level in calves [Rousing et al. 2005]
87 and cows [Gibbons et al. 2009; Turner et al. 2011]). However, some papers found moderate to

88 high repeatability only for some measures and time periods (Haskell et al. 2012; Mazurek et al.
89 2011; Windschnurer et al. 2008; Windschnurer et al. 2009; see also review of responses to
90 humans by de Passillé and Rushen 2005), and other studies have found no repeatability (Battini
91 et al. 2011), although all of these studies depended on some measure of avoidance or retreat
92 from a human. Fina and colleagues (2006) reported that reliability of responses to restraint
93 differed depending upon the calves' initial responses, with calm individuals remaining calm
94 across tests but fearful ones showing reduced fear over time.
95 Farm-level repeatability is also important for measures of approach or avoidance of
96 humans, because this type of measure has been proposed for use in on-farm welfare
97 assessments (e.g. Winckler et al. 2003; Winckler et al. 2007), focussing on herd-level
98 differences. Only a few papers have investigated farm-level repeatability of responses to
99 humans, all in adult cows, and studies have sometimes confounded test-retest reliability with
100 inter-observer reliability (e.g. Windschnurer et al. 2009). In these tests (based upon avoidance
101 of an approaching human) low to moderate reliability has been reported (De Rosa et al. 2003;
102 Winckler et al. 2007). Reliability can also be estimated at the level of the pen or group
103 (intermediate between individual and farm levels), and indeed some farm level estimates are
104 based upon observations of a single pen. Only one study on calves has assessed the reliability
105 of approach responses measured at the pen level, and this study reported high reliability
106 (Bokkers et al. 2009, with similar results for an avoidance measure).
107 Even among papers that claim repeatability, correlations are sometimes low. For
108 example, Turner and colleagues (2011) assessed repeatability across and within tests of fear of
109 humans in beef cattle and found the proportion of variance explained by individual consistency
110 ranged from 0.17 to 0.54. In fact, a meta-analysis of the personality literature in wild animals
111 found an average repeatability (intraclass correlation coefficient) of only 0.37 (Bell 2009), which
112 is considerably below the level generally deemed acceptable (0.6 being a traditional standard in
113 the human literature (e.g. Bruton et al. 2000, Mroczek 2007). In humans, typical correlations

over long intervals (years) are often over 0.7 in adults (Mroczek 2007). Conversely, correlation coefficients for children and college students were only 0.31 and 0.54 respectively, for major personality traits in one meta-analysis (Roberts and DelVecchio 2000). It therefore seems likely that other juvenile animals, such as calves, may also show limited correlations in their fear responses over time.

The aims of the current study were to assess the individual-level test-retest reliability of versions of novel object and response to human tests, and the farm-level test-retest reliability of a response to human test. An additional aim was to identify factors that influence reliability, enabling refinements in protocols used in future research and on-farm welfare assessments. The factors investigated included consistency of the object used in the novel object test, test duration, and calf health. We also assessed inter- and intra-observer reliability (i.e. consistency between and within people recording the data) of the measures, as these are essential to obtaining test-retest reliability.

## 2. Materials and methods

### *2.1. Experiment 1*

All of the research presented in this paper was approved by the University of British Columbia Animal Care Committee. In this experiment we used 32 Holstein bull calves, housed at the University of British Columbia Dairy Education and Research Centre. These calves also served in a concurrent study on the effects of early social housing, comparing individually housed calves (n=10), pair-housed calves (n=12), and calves kept in a complex social group with access to their dams (n=10). More detail regarding these treatments is available in Meagher et al. (2015). Pens were cleaned once per week. Calves were offered 8 L of milk per day for the first 28 d, at which time the milk ration was reduced to 6 L over 3 d, always split

between two daily feedings. This reduction was intended to stimulate solid feed intake. At approximately 58 d, calves were weaned over a 3-day period. Calves had ad libitum access to water throughout the experimental period, and access to grain (Hi-Pro Medicated Calf Starter) and a mixed ration beginning at day 5 ± 2. Health checks were performed weekly throughout the experimental period to assess symptoms of common illnesses, including respiratory and enteric disease. Calves were treated when appropriate according to standard farm protocols.

Two tests for fearfulness were used: novel object and response to human (in this case approach to a stationary, familiar person). These tests were conducted on consecutive days at approximately 41 d of age and repeated at approximately 62 d of age. The response to human test was also conducted at 25 d of age. Tests were conducted between the two daily feedings, but never within 30 min of either feeding time. Novel object tests took place in a test pen that the calves had visited twice daily (for cognitive training; see Meagher et al. 2015) for several weeks. After 2 min of habituation to the pen, the novel object (in this case, a brightly coloured ball) was lowered into the pen using a length of twine. The test lasted 10 min, and latency to make contact with the ball was recorded. The response to human tests were conducted during weekly weighing of the animals, following a similar procedure to Duve and colleagues (2012) in which calves were allowed to approach a human and then their response to weighing was assessed. In brief, the calf was released from its pen into the alley, and given up to 90 s to make contact with the stationary person. The stationary person (one person per experiment) was familiar to the calves and stood 2.4 m away. The first author (RKM, who was also familiar to the calf) stood inside the pen and recorded the latencies to touch the person. Wooden dividers blocked the view of calves on the other side of the aisle, leaving an alley approximately 1.2 m wide for the individual and pair treatments; however, calves could see into neighbouring pens on the same side of the alley as they approached the person. For the group-housed calves, the distance to the person was equivalent, but the space was wider and no other calves were in sight. The calf was then encouraged or pushed onto the scale (by the previously stationary person), and the

166 difficulty of pushing was scored by the handler on a scale of 0 to 4, with 0 indicating the calf
167 walked onto the scale with no physical guidance, and 4 that a single handler could not get them
168 on the scale alone.
169 Test-retest reliability was assessed using Spearman rank correlations due to non-
170 normality of the data. Weighted sums of Spearman correlations are presented to control for
171 effects of housing treatment (Taylor 1987). Correlation coefficients and not p-values are
172 reported, because p-values are too dependent on sample size to be very useful measures of
173 reliability (Martin and Bateson 2007). Throughout the paper, we categorize reliability as
174 negligible (correlation <0.30), low (0.30-0.49), moderate (0.50-0.69), high (0.70-0.89) or very
175 high (≥0.90) following Hinkle and colleagues (2003). For the ordinal data from scores of difficulty
176 of handling during weighing, we used two types of analysis: kappa scores for agreement on the
177 ordinal data (categorized according to Dohoo et al. 2002), and kappas combined with percent
178 agreement when converted to a binary analysis for some force needed (scores 2 to 4) versus no
179 force needed (scores 0 or 1).
180
181 *2.2. Experiment 2*
182
183 *2.2.1. General methods*
184
185 In Experiment 1, the testing schedule was partially determined by the other experiment
186 running simultaneously, and the calves had some experiences between tests that might have
187 caused changes in behaviour, including weaning from milk onto solid feed. Thus, in Experiment
188 2 we assessed the reliability of the handling and novel object responses using a shorter inter-
189 test interval and during a period of consistent management.
190 The subjects were two cohorts of Holstein calves. In Experiment 2a we used 27 calves
191 (18 male, 9 female), and in Experiment 2b we used 13 calves (all female). Calves were

192 individually housed until the end of the experiment and cared for in the same way as described
193 above, except for the following differences in feeding: no total mixed ration was provided during
194 this experiment, and calves were stepped down to 4 L of milk rather than to 6 L beginning at d
195 26. Also, for the purposes of a related experiment, 13 of the calves in the first group were given
196 a nutritional supplement with their milk, alpha S1 casein hydrosylate (Zylkène®, distributed by
197 Vétoquinol, Princeville, QC), beginning 7 d before the start of fear testing and ending on the day
198 they were moved to the group pens. This treatment did not affect any of the response measures
199 except latency in the first novel object test.
200
201 *2.2.2. Experiment 2a*
202
203 At age 36 ± 3 d, a novel object test was conducted in the home pen. This test was
204 repeated 7 d later (d 43 ± 3). Two different objects were used to maintain the novelty of the test
205 rather than conducting an exact replicate: a red and white ball, and a blue plastic basket. Half of
206 the calves received the ball in the first test and the basket in the second, and the other half
207 received the objects in the reverse order. The tests were conducted in the same way as
208 Experiment 1, but in addition to latency to make contact, total time in contact with the object was
209 recorded. All measures were assessed from video recordings by trained observers (one per
210 variable) who were blind to the study aims, and intra-observer reliability was tested by having
211 these observers score a subset of the videos a second time to ensure that they were consistent
212 in their scoring; latency to make contact was also recorded live for all calves by the first author,
213 who also assessed the other measures from a subset of videos for inter-observer reliability
214 testing. Responses to a human handler were also assessed as in Experiment 1. These tests
215 were conducted on the day following each novel object test.
216 Normality of the data was assessed using Shapiro-Wilk tests. Latency data were non-
217 normally distributed, so repeatability was assessed using Spearman rank correlations. Difficulty

of handling scores were analysed as in Experiment 1. Six calves showed symptoms of illness at some time during the testing period, primarily with enteric illness, which may have affected reaction speed and likelihood of approaching the object; these calves were excluded from the analyses.

### *2.2.3. Experiment 2b*

The second cohort of calves was used to assess whether modifying the protocols used in Experiment 2a would improve reliability. Housing, care and testing protocols were the same as in Experiment 2a, except for one change in each test. For the novel object tests, the same object was used in both tests (each calf being assigned to either the ball or the basket) rather than calves getting a different object in Test 1 and 2. For the response to human tests, the duration of the test was extended from 90 s to 180 s to reduce potential ceiling effects. The data were analysed for test-retest reliability as above, again excluding calves that were ill.

## *2.3. Experiment 3*

This experiment was conducted on 15 dairy farms in the Fraser Valley of British Columbia, Canada, with the aim of assessing farm-level repeatability in response to humans. Unweaned calves between 7 and 70 d of age were tested. Because each farm was visited twice, 6 to 8 wk apart, the individual calves tested on the second visit were a completely separate cohort, but represented the full range of ages where possible (average age in test 1: 34; test 2: 37 d). All calves were Holstein or Holstein crosses. Data were collected from a total of 677 calves, with an average of 21 calves per farm on each visit. Tests were conducted between morning and afternoon feedings and never within an hour of feeding time.

243 Fear of humans was assessed using an approaching human test, which could be
244 conducted without opening the calf pens. Unlike in the previous experiments, the human (RKM)
245 was unfamiliar to the calves. The person walked along the row of pens, parallel to them and
246 approximately 1 m from the front of each pen or hutch (space permitting). Once directly in front
247 of a pen, she then turned to face the calf and said “hello” to attract their attention (cf. Bokkers et
248 al. 2009). After pausing for 5 s to record any locomotor response, she approached the calf at a
249 pace of approximately 1 step per second (as in e.g. Windschnurer et al. 2008), and then
250 extended her arm to where the calf could reach it, with the hand flat and oriented sideways.
251 Direct eye contact was avoided (Bokkers et al. 2009). Retreats were scored on an ordinal scale
252 according to Table 1. We also recorded whether the calf touched the experimenter, and the
253 latency to do so, within 2 min. The experimenter then repeated the procedure at the next pen in
254 the row, following the same route through the pens on both visits to a farm, and never passing
255 directly in front of a calf prior to its test if at all possible. For socially-housed calves, latencies
256 and retreats for each calf in the pen were recorded.
257 Calf health was visually assessed after each test. The presence of a spontaneous
258 cough, or faecal consistency scoring greater than 2 (following McGuirk, 2013) were considered
259 indicators of illness.
260 Repeatability of the test was assessed at the farm level for the proportion of calves
261 making contact with the experimenter, since calves within a farm were non-independent, using a
262 Spearman rank correlation. Repeatability of retreats in this test was also assessed with
263 Spearman rank correlations, using three different ways of summarizing the behaviour:
264 proportion of calves retreating by the time the experimenter was at the pen with hand extended
265 (score 2 or above) or prior to extending the hand (score 3 or above), and the average score for
266 each farm.
267 One farm was excluded because a major housing change occurred between tests. On
268 the remaining 14 farms, individual calves were excluded if they showed signs of diarrhoea or

respiratory illness or both, based on the criteria above. The reliability analyses were then repeated to check for an effect of these illnesses on the results.

## 3. Results

### *3.1. Experiment 1*

Latencies to approach the novel object were moderately correlated between tests at 42 and 60 d of age, with a correlation coefficient ($r_s$) of 0.54 (n=24; Figure 1). There was little evidence of any relationship in approach latencies to the human handler between tests at 25 and 42 d of age ($r_s$ = 0.26, n=23), nor at 42 versus 60 d of age ($r_s$ = 0.21, n=26).

Difficulty of handling scores showed low reliability using the ordinal scale. Kappa values were 0.33 for day 25 vs. 42 and 0.22 for day 42 vs. 60 (indicating "fair agreement": Dohoo et al. 2003). However, 22 of 31 calves (71%) were consistent from days 25 to 42 in terms of whether any force was needed (kappa 0.44, indicating moderate reliability). For day 42 vs. 60, percent agreement was similar: 23 of 34 calves (68%; kappa 0.35).

### *3.2. Experiment 2a*

#### *3.2.1. Test-retest reliability*

The correlation between Tests 1 and 2 for latency to touch the novel object was negligible ($r_s$=0.29, n=20; Figure 2a). Excluding calves that failed to touch the object in at least one test, which often happened if calves were resting immediately before the test, perhaps reflecting drowsiness rather than increased fear or lack of interest, improved the correlation between tests ($r_s$=0.70, n=15). The reliability of time in contact with the object was low when

295 considering all calves ($r_s$=0.30, n=20), and negligible when excluding those that did not make
296 contact ($r_s$=0.02, n=16).
297 For the response to humans, a correlation between latencies in the two tests could not
298 be meaningfully assessed because only 6 of 27 calves ever made contact with the handler on
299 the first test, and of these only three also made contact during the second test. Agreement in
300 difficulty of handling scores was very low whether data were analysed as ordinal or binary
301 (kappa 0.07 and 0.03, respectively), although there was 50% agreement in the latter (10 of 20
302 calves).
303
304 *3.2.2. Intra- and inter-observer reliability*
305
306 Inter-observer reliability for latency to touch the novel object was very high ($r_s$=0.93,
307 n=27), and intra-observer reliability was also high for the subset of videos that were re-assessed
308 ($r_s$=0.81, n=15). Total time in contact also had high inter-observer reliability ($r_s$=0.70, n=10) and
309 very high intra-observer reliability ($r_s$=0.94, n=15).
310
311 *3.2.3. Experiment 2b*
312
313 Test-retest reliability for latency to approach the novel object was higher in this
314 Experiment ($r_s$=0.58, n=11; Figure 2b), but excluding non-contacts did not improve reliability
315 ($r_s$=0.32, n=10). Reliability of the response to human was high in this experiment, ($r_s$=0.76,
316 n=10; Figure 3). However, this improvement was not the result of using the extended maximum
317 test duration of 180 s; only 1 calf made contact with the handler between 90 and 180 s on both
318 tests, and artificially imposing a 90 s ceiling produced a high reliability coefficient ($r_s$=0.83,
319 n=10). The high reliability was partially due to the fact that failure to make contact within 90 s
320 was consistent among individuals: 5 of the 6 who did not make contact on the first test also

failed to make contact in the second test. Agreement in difficulty of handling scores was fair for this group (kappa 0.26), and this value was similar (0.27) for whether any force was needed to get the calf on the scale, with 7 of 11 (64%) calves in agreement.

### *3.3. Experiment 3*

Repeatability depended on the response measure and exclusions for illness, as presented in Table 2. In brief, the proportion of calves making contact with the person showed low or negligible repeatability; indeed, the slope of the relationship was negative. Retreats were moderately repeatable for the full data set. Using yes/no data for whether a calf retreated at all, before the person's arm was extended (score 3 or above) was slightly more reliable than including retreats at the time the arm was extended (score 2). The most reliable measure was the average retreat score for the farm.

Signs of illness were recorded for 68 of 599 calves on the 14 farms analysed. For three of the four response variables, excluding calves with coughs improved repeatability. Excluding calves with diarrhoea only improved repeatability for two response measures, and excluding both groups reduced repeatability for all measures relative to excluding coughs alone.

## **4. Discussion**

### *4.1. Factors influencing repeatability*

The results show varying levels of repeatability in both novel object tests and those assessing response to humans. We speculated that the low reliability in Experiment 1 was due to a long test-retest interval (approx. 20 d), combined with important management changes (including weaning from milk). Consistent with this idea, we found some improvement in

reliability estimates for the novel object test in some groups when we switched to shorter intervals (7 d) with more consistent management (pre-weaning only) in Experiment 2, and for the response to human test in Experiment 2b. Agreement in scores of difficulty of handling was typically low to fair across the experiments, although it was higher for the binary (some force vs. no force needed) scale than the ordinal scale in Experiment 1.

In Experiment 2a, the improvement in novel object reliability occurred only when including animals that were alert during testing. Unfortunately, the results of this inclusion criteria differed between Experiments 2a and b, which may reflect some instability in the correlation estimates due to the small sample sizes available (see e.g. Goodwin and Leach 2006). Based on the human literature, the sample sizes needed for stability of personality correlation estimates would be very difficult to achieve (e.g. n=250: Schönbrodt and Perugini 2013); we suggest instead the use of multi-study replication, ideally with meta-analyses, to confirm the reported effects. However, the result from 2a suggests that it would be worthwhile to investigate the benefit of a further refinement that could be used for both the novel object test and human approach tests conducted in the home pen: imposing a procedure or criteria to ensure that animals were attending to the test situation. For example, in Experiment 1 calves were moved to a testing pen and the test began shortly afterwards. This ensured that no calves were asleep or resting at the time the stimulus was presented, as well as removing possible distractions such as the presence of food. Home pen tests are desirable for practical reasons and because they avoid introducing handling effects and social isolation for group housed animals (see Forkman et al. 2007; Tecott and Nestler 2004), but in this case it seems that the costs may outweigh the benefits (the reverse may be true when measuring exploration rather than fear; see Carter et al. 2013).

In Experiment 2, we considered two additional factors thought to improve repeatability: increasing the test duration when latencies are measured, and the consistency of the novel object. Repeatability of the latency to touch humans could not be assessed in Experiment 2a

due in part to ceiling effects associated with a short test; we thus hypothesised that increasing the time allowed would improve reliability for the latency measures. The latencies in Experiment 2b did show high reliability, but this was not due to the longer tests. That said, given that ceiling effects prevented discrimination among individuals in Experiment 2a, we still contend that longer test durations improve the validity and usefulness of the test by avoiding an artificial upper limit in measures of latency. Others have similarly argued that extending test durations improves test validity (e.g. in tests of chronic anxiety in rodents; Fonio et al. 2012).

The improved repeatability of the novel object in Experiment 2b versus 2a was likely due to using a second presentation of the same object. In Experiment 2a we had used a different novel object for each test (to retain the novelty), but a disadvantage of this approach is that animals may find some objects inherently more fear-inducing than others thus making responses more variable. Although we found that using the same 'novel' object for multiple tests improved the repeatability of the test, we do not recommend this practice in future tests. Instead we argue that there is much to be gained from examining a range of objects; if individual rankings differ between arbitrarily chosen objects with no apparent biological significance, it is likely not valid to draw broad conclusions regarding 'fear of novelty' from tests with a single object.

Experiment 3 identified the role of sickness, particularly respiratory illness, in reducing reliability of responses to humans. Sickness behaviour is widely accepted to include lethargy and decreased exploratory behaviour (e.g. Millman 2007; Swiergiel and Dunn 2007). A recent study in calves found that respiratory illness and fever decreased probability of calves approaching novel objects and stationary humans; diarrhoea did not immediately have this effect, although during recovery from this ailment calves were less likely to approach people (Cramer and Stanton 2015). Changes in health status could thus reduce repeatability of the results for both types of test. Cramer and Stanton (2015)'s findings mirror the current results, in that excluding calves with signs of respiratory illness most consistently improved the correlation

399 between tests across variables. Excluding calves with signs of diarrhoea or both illnesses was
400 less helpful, although this may have been due to the reduced sample size (Goodwin and Leech
401 2006), and this should therefore be retested in a larger sample of calves. While the differences
402 in reliability estimates in this experiment were relatively small, collectively, these findings
403 support our choice to exclude animals that were sick around the day of testing in Experiment 2.
404 Unfortunately, health checks were not conducted on test days in Experiment 1. In future, health
405 status should be addressed when reporting responses to these tests.
406 Another lesson from Experiment 3 was that the proportion of calves making contact with
407 an unfamiliar human has low repeatability relative to other response measures. This is
408 surprising since this measure, and the related measure of latency to contact, are commonly
409 used (e.g. Bokkers et al. 2009; Forkman and Keeling 2009). We found that the most reliable
410 response measure was the retreat score. For retreat as a yes/no variable, which is simpler to
411 record, particularly when calves are group-housed, the correlation between visits was slightly
412 higher when counting retreats before the researcher's arm was extended versus retreats at the
413 time the arm was extended. Although the difference was small, it may reflect inconsistency in
414 the behaviour of the test person, such as speed of arm extension or positioning of the hand
415 relative to the calf. Repeatability of the retreat measures at farm level was comparable to the
416 individual-level results using latencies in Experiment 1.
417 Several factors that could influence repeatability of tests of fear were not investigated
418 here. As described by Waiblinger and colleagues (2006), human-animal relationship tests in
419 farm animals can be influenced by many factors, including interference by neighbouring
420 animals, exploratory, social, feeding and lying motivations, and social isolation. Feeding
421 motivation was relatively constant within each of the experiments in the current study (as tests
422 were held outside of regular feeding times, although this was not a perfect control since the
423 testing window was relatively large for practical reasons, likely increasing variation between
424 days), and social motivation and responses to isolation were not relevant in most cases.

425 Interference by neighbouring animals was not an issue during the novel object tests, since the
426 calves were alone during testing, and were minimal throughout Experiment 3 since most calves
427 were housed alone and vocalizations were not common. However, it may have been an issue in
428 the response to human tests of Experiments 1 and 2, as calves could walk past the pens of
429 neighbours. In Experiment 3, there may also have been fluctuations in farm practice such as
430 staff members providing most care to the calves, or feeding times. However, this will be the
431 reality for any on-farm work and such variation must be accepted except where changes are
432 predictable (e.g. due to season) and can thus be accounted for in the study design.
433 One effect that has not been directly investigated in this context, but which is known to
434 play a role in animals' responses to potentially threatening stimuli, is laterality. Vertebrates,
435 including cows, typically prefer to view threatening stimuli from the left eye (Robins and Phillips
436 2010), and the eye that first sees a stimulus can influence escape responses (e.g. Austin and
437 Rogers 2007). It would be of interest to test whether inconsistency in the orientation of cattle
438 relative to fear-inducing stimuli can explain differences in responses on repeated tests. Testing
439 this idea will require a test environment that allows control of presentation side.
440
441 *4.2. Strategies for using tests with limited repeatability*
442
443 Even if protocols are refined to reduce noise, there are likely limitations in the level of
444 repeatability that can be achieved. As discussed in the Introduction, the average repeatability
445 reported for personality traits of wild animals is only 0.37 (Bell et al. 2009). How consistency of
446 behaviour in farm animals will compare is difficult to predict. As de Passillé and Rushen (2005)
447 point out, even where there are moderate, statistically significant correlations, a large number of
448 animals will be "misclassified" by a single test. These limitations do not necessarily prevent the
449 tests from being useful; despite their typically low reliability, personality tests in wild animals can
450 still predict ecologically or practically important outcomes (e.g. Smith and Blumstein 2008). In

451 the experiments described here, despite low to moderate reliability, the tests conducted at 41 d
452 of age in Experiment 1 were able to detect some effects of treatment that correspond with
453 theory: fear of novelty was higher in calves reared in simpler, more socially restricted housing
454 (Meagher et al. 2015). Human personality studies typically report repeatability estimates
455 averaging 0.7 to 0.8 for the Big Five factors of personality (e.g. Gnambs 2014, Mroczek 2007),
456 but these factors are typically derived from multi-item scales. Having only one or two measures,
457 which is the norm in animal studies, is expected to increase measurement error (Credé et al.
458 2012).

459 A common recommendation when assessing traits is to conduct repeated tests and sum
460 or average responses. However, in the case of novelty, repeated testing is logically problematic
461 (see Forkman et al. 2007), as the object is no longer novel when presented a second time; even
462 if the object is changed, the test procedure becomes less novel. One approach to circumvent
463 this difficulty is to consider decreases in fear as an acceptable result when assessing reliability
464 (e.g. Meagher et al. 2011); repeated testing can then still be used to draw inferences, because
465 differences in habituation or sensitization rates may also be consistent, welfare-relevant
466 individual traits (Jones and Boissy 2011).

467 The results from the current study suggest that multiple tests might be needed, but using
468 a range of objects or other stimuli, given the differences in individual rankings depending on the
469 objects used. Similarly, Ramos (2008) argues that for measuring trait anxiety (and/or modelling
470 human anxiety disorders), conducting multiple types of tests is necessary. He argues that these
471 should be conducted simultaneously if the alternative is placing the animal in the same test
472 chamber or apparatus multiple times, but this would not allow assessment of how much of the
473 response is due to temporary states present at the time of testing. One difficulty with
474 recommending multiple tests is the time and expense, for example, of conducting multiple visits
475 to farms for on-farm welfare audits. Current protocols sometimes focus on ensuring inter-

observer reliability (e.g. Wemelsfelder and Lawrence 2001), but our results indicate that this is not sufficient for producing repeatability.

Farm-level repeatability could conceivably be attained without individual-level repeatability, if problems with the latter are due only to the inherent problems in repeating a test involving novelty. As long as the results are repeatable within farm using new groups of animals, this would not be a major concern for farm-level investigations. Understanding why results change within individuals is nonetheless important, since differences due to age or season should be taken into account when selecting samples and testing times (see Haskell et al. 2012).

*4.3. Outstanding concerns regarding test validity*

The discussion above was focused primarily on practical issues regarding reliability of fear tests, but even if these issues can be resolved questions remain about test validity. Very little validation testing has been done for response to human tests in calves, including the voluntary approach-type tests used here, although some studies indicate that both voluntary approach to humans and avoidance distance are influenced by rough or gentle handling (Lensink et al. 2000; Schuetz et al. 2012; Windschnurer et al. 2009**).** The tests of neophobia used in farm animals, such as novel object (e.g. Misslin and Ropartz 1981) and open field tests (Hall 1936; Archer 1973), are largely adapted from tests originally developed and validated for laboratory species. In some cases, the rationale for the test was based on the behavioural ecology of the rodent species, and applicability to other species is questionable. For example, the open field test makes sense for rats and mice that fear open areas (presumably because these are associated with increased predation risk; e.g. Lister 1987; Ohl 2003; Rodgers 1997), but cattle are too large to be at risk of overhead predators and are adapted to life in open habitats. The novel object test is expected to apply more broadly (e.g. Russell 1973), but in

some cases species-specific responses, such as burying, need to be taken into account (e.g. Misslin and Ropartz 1981). As noted above, the object-specificity of the test results in Experiment 2 also raises some concerns about its validity as a general measure of neophobia. A second potential problem is that, even in laboratory animals where these tests have been better validated and sometimes proved useful in drug screening, the validity of some tests (e.g. the open field) has also been called into question (e.g. Ennaceur 2014). Known issues from the laboratory animal literature include sensitivity to environmental variables unrelated to the intended treatment, reducing external validity (Garner 2005) and preventing accurate measures of trait anxiety because they are overshadowed by the effects of temporary states (e.g. Ohl 2003; Sylvers et al. 2011). This is likely one reason for failures to replicate results in different laboratories (e.g. Dawson and Tricklebank 1995; Sousa et al. 2006; Wurbel, 2002).

Although careful attention to the methodological factors described above will likely reduce problems of poor reliability and aid in the interpretation of data, the use of short-term tests may be inherently problematic if the aim is to assess consistent traits in animals. Temperament ratings by people who can integrate behaviour over time are one suggested alternative (see Carlstead et al. 1999; Meagher 2009), but the relationship between these measures and standard tests is not well understood (e.g. de Passillé and Rushen 2005). Finally, the same underlying motivation can be expressed very differently depending on the testing situation (e.g. approaching to bury an object when possible versus retreating from it if not), potentially leading to misinterpretations regarding fearfulness (Franks et al. 2012). More species-specific validations of the different types of fear test, taking into account natural behaviour, are thus needed.

## 5. Conclusions

Moderate test-retest reliability seems achievable for both novel object and response to human tests in dairy calves. It is, however, contingent on allowing sufficient time for the behavioural response, and excluding calves with respiratory illness and perhaps other forms of illness if replications of this work can confirm that they decrease reliability. In the case of novel object tests, moving subjects to a testing pen or otherwise assuring that calves are alert at the beginning of the test and not distracted by competing motivations will also help. For tests using an unfamiliar human as the stimulus, moderate repeatability was only achieved for retreat scores and not for likelihood of making contact with the person. None of the protocols assessed provided consistently high repeatability, and results of neophobia tests seem to be dependent on the specific stimuli chosen. For these reasons, we suggest that future research use multiple tests to assess fearfulness or anxiety, using different stimuli.

**Acknowledgements**

We are grateful to the students and staff of the UBC Dairy Education and Research Centre, especially to Alan Makarewicz, Tatiane Vito Camiloti, Nancy Chen, Justine Gallo, Annett Gefrom, Clémence Messant, Pauline Gautier, Sara McNamara, Yasmine Yavari & Ty Chapman for help with calf care, and to João Cardoso Costa for helpful discussions. Funding was provided by a Discovery Grant (RGPIN 262278-10) from the Natural Sciences and Engineering Research Council of Canada (NSERC) to DMW.

**References**

Archer, J., 1973. Tests for emotionality in rats and mice: A review. Anim. Behav. 21, 205-235.

Austin, N.P., Rogers, L.J., 2007. Asymmetry of flight and escape turning responses in horses. Laterality 12, 464-474.

553 Barnett, J., Hemsworth, P., Newman, E., 1992. Fear of Humans and its Relationships with
554 Productivity in Laying Hens at Commercial Farms. Br. Poult. Sci. 33, 699-710.
555 Battini, M., Andreoli, E., Barbieri, S., Mattiello, S., 2011. Long-term stability of Avoidance
556 Distance tests for on-farm assessment of dairy cow relationship to humans in alpine traditional
557 husbandry systems. Appl. Anim. Behav. Sci. 135, 267-270.
558 Bell, A.M., Hankison, S.J., Laskowski, K.L., 2009. The repeatability of behaviour: a meta-
559 analysis. Anim. Behav. 77, 771-783.
560 Boivin, X., Le Neindre, P., Chupin, J.M., Garel, J.P., Trillat, G., 1992. Influence of breed and
561 early management on ease of handling and open-field behaviour of cattle. Appl. Anim. Behav.
562 Sci. 32, 313-323.
563 Bokkers, E.A.M., Leruste, H., Heutinck, L.F.M., Wolthuis-Fillerup, M., van der Werf, J.T.N.,
564 Lensink, B.J., van Reenen, C.G., 2009. Inter-observer and test-retest reliability of on-farm
565 behavioural observations in veal calves. Anim. Welf. 18, 381-390.
566 Bruton, A., Conway, J.H., Holgate, S.T., 2000. Reliability: What is it, and how is it measured?
567 Physiotherapy 86, 94-99.
568 Carlstead, K., Mellen, J., Kleiman, D.G., 1999. Black rhinoceros (Diceros bicornis) in US zoos: I.
569 Individual behavior profiles and their relationship to breeding success. Zoo Biol. 18, 17-34.
570 Carobrez, A.P., Bertoglio, L.J., 2005. Ethological and temporal analyses of anxiety-like
571 behavior: The elevated plus-maze model 20 years on. Neuroscience & Biobehavioral Reviews;
572 Defensive Behavior 29, 1193-1205.
573 Carter, A.J., Feeney, W.E., Marshall, H.H., Cowlishaw, G., Heinsohn, R., 2013. Animal
574 personality: what are behavioural ecologists measuring? Biological Reviews 88, 465-475.
575 Cramer, M.C., Stanton, A.L., 2015. Associations between health status and the probability of
576 approaching a novel object or stationary human in preweaned group-housed dairy calves. J.
577 Dairy Sci. 98, 7298-7308.

578 Credé, M., Harms, P., Niehorster, S., Gaye-Valentine, A., 2012. An evaluation of the
579 consequences of using short measures of the Big Five personality traits. J. Pers. Soc. Psychol.
580 102, 874-888.
581 Dawson, G.R., Tricklebank, M.D., 1995. Use of the elevated plus maze in the search for novel
582 anxiolytic agents. Trends Pharmacol. Sci. 16, 33-36.
583 de Passillé, A.M., Rushen, J., 2005. Can we measure human-animal interactions in on-farm
584 animal welfare assessment? Appl. Anim. Behav. Sci. 92, 193-209.
585 De Rosa, G., Tripaldi, C., Napolitano, F., Saltalamacchia, F., Grasso, F., Bisegna, V., Bordi, A.,
586 2003. Repeatability of some animal-related variables in dairy cows and buffaloes. Anim. Welfare
587 12, 625-629.
588 Duve, L.R., Weary, D.M., Halekoh, U., Jensen, M.B., 2012. The effects of social contact and
589 milk allowance on responses to handling, play, and social behavior in young dairy calves. J.
590 Dairy Sci. 95.
591 Ennaceur, A., 2014. Tests of unconditioned anxiety — Pitfalls and disappointments. Physiol.
592 Behav. 135, 55-71.
593 Farm Animal Welfare Council, 2009. Farm animal welfare in Great Britain: Past, present and
594 future.
595 Fina, M., Casellas, J., Manteca, X., Piedrafita, J., 2006. Analysis of temperament development
596 during the fattening period in the semi-feral bovine calves of the Alberes Massif. Animal
597 Research 55, 389-395.
598 Fonio, E., Benjamini, Y., Golani, I., 2012. Short and Long Term Measures of Anxiety Exhibit
599 Opposite Results. Plos One 7, e48414.
600 Forkman, B., Boissy, A., Meunier-Salauen, M.C., Canali, E., Jones, R.B., 2007. A critical review
601 of fear tests used on cattle, pigs, sheep, poultry and horses. Physiol. Behav. 92, 340-374.
602 Forkman, B., Keeling, L.J., 2009. Assessment of animal welfare measures for dairy cattle, beef
603 bulls and veal calves. Welfare Quality Reports 11, 1-314.

604 Franks, B., Higgins, E.T., Champagne, F.A., 2012. Evidence for Individual Differences in
605 Regulatory Focus in Rats, Rattus norvegicus. J. Comp. Psychol. 126, 347-354.
606 Garner, J.P., 2005. Stereotypies and Other Abnormal Repetitive Behaviors: Potential Impact on
607 Validity, Reliability, and Replicability of Scientific Outcomes. ILAR 46, 106-117.
608 Gibbons, J., Lawrence, A., Haskell, M., 2009. Responsiveness of dairy cows to human
609 approach and novel stimuli. Appl. Anim. Behav. Sci. 116, 163-173.
610 Gnambs, T., 2014. A meta-analysis of dependability coefficients (test–retest reliabilities) for
611 measures of the Big Five. Journal of Research in Personality 52, 20-28.
612 Goodwin, L.D., Leech, N.L., 2006. Understanding correlation: Factors that affect the size of *r*.
613 The Journal of Experimental Education 74, 251-266.
614 Graunke, K.L., Langbein, J., Repsilber, D., Schoen, P., 2013. Objectively measuring behaviour
615 traits in an automated restraint-test for ungulates: towards making temperament measurable. J.
616 Agric. Sci. 151, 141-149.
617 Gray, J.A., 1987. The Psychology of Fear and Stress. Cambridge University Press, Cambridge.
618 Hall, C.S., 1936. Emotional behaviour in the rat. III. The relationship between emotionality and
619 ambulatory activity. J. Comp. Psych. 22, 345-352.
620 Haskell, M.J., Bell, D.J., Gibbons, J.M., 2012. Is the response to humans consistent over
621 productive life in dairy cows? Anim. Welfare 21, 319-324.
622 Hegelund, L., Sorensen, J.T., 2007. Measuring fearfulness of hens in commercial organic egg
623 production. Anim. Welfare 16, 169-171.
624 Hemsworth, P.H., Barnett, J.L., Coleman, G.J., Hansen, C., 1989. A Study of the Relationships
625 between the Attitudinal and Behavioral Profiles of Stockpersons and the Level of Fear of
626 Humans and Reproductive-Performance of Commercial Pigs. Appl. Anim. Behav. Sci. 23, 301-
627 314.
628 Hemsworth, P., Coleman, G., Barnett, J., Borg, S., 2000. Relationships between human-animal
629 interactions and productivity of commercial dairy cows. J. Anim. Sci. 78, 2821-2831.

630 Jones, B., Boissy, A., 2011. Fear and other negative emotions, in: Appleby, M.C., Mench, J.A.,
631 Olsson, I.A.S., Hughes, B.O. (Eds.). CABI, Cambridge, UK, pp. 78-97.
632 Kilgour, R.J., Melville, G.J., Greenwood, P.L., 2006. Individual differences in the reaction of beef
633 cattle to situations involving social isolation, close proximity of humans, restraint and novelty.
634 Appl. Anim. Behav. Sci. 99, 21-40.
635 Lensink, B., Boivin, X., Pradel, P., Le Neindre, P., Veissier, I., 2000. Reducing veal calves'
636 reactivity to people by providing additional human contact. J. Anim. Sci. 78, 1213-1218.
637 Lister, R.G., 1987. The use of a Plus-Maze to Measure Anxiety in the Mouse.
638 Psychopharmacology (Berl. ) 92, 180-185.
639 MacKay, J.R.D., Haskell, M.J., Deag, J.M., van Reenen, K., 2014. Fear responses to novelty in
640 testing environments are related to day-to-day activity in the home environment in dairy cattle.
641 Appl. Anim. Behav. Sci. 152, 7-16.
642 Martin, P., Bateson, P., 2007. Measuring Behaviour: An Introductory Guide. Cambridge
643 University Press, Cambridge, UK.
644 Mazurek, M., McGee, M., Crowe, M.A., Prendiville, D.J., Boivin, X., Earley, B., 2011.
645 Consistency and stability of behavioural fear responses of heifers to different fear-eliciting
646 situations involving humans. Appl. Anim. Behav. Sci. 131, 21-28.
647 McGuirk, S., 2013. Calf health scoring chart. 2015.
648 Meagher, R.K., 2009. Observer ratings: Validity and value as a tool for animal welfare research.
649 Appl. Anim. Behav. Sci. 119, 1-14.
650 Meagher, R.K., Daros, R.R., Costa, J.H.C., von Keyserlingk, Marina A. G., Hötzel, M.J., Weary,
651 D.M., 2015. Effects of Degree and Timing of Social Housing on Reversal Learning and
652 Response to Novel Objects in Dairy Calves. PLoS ONE 10, e0132828.
653 Meagher, R.K., Duncan, I., Bechard, A., Mason, G.J., 2011. Who's afraid of the big bad glove?
654 Testing for fear and its correlates in mink. Appl. Anim. Behav. Sci. 133, 254-264.

655 Millman, S.T., 2007. Sickness behaviour and its relevance to animal welfare assessment at the
656 group level. Anim. Welfare 16, 123-125.
657 Misslin, R., Ropartz, P., 1981. Responses in Mice to a Novel Object. Behaviour 78, 169-177.
658 Mroczek, D.K., 2007. The analysis of longitudinal data in personality research, in: Robins, R.W.,
659 Fraley, R.C., Krueger, R.F. (Eds.). Guilford Press, New York, USA, pp. 543-556.
660 Ohl, F., 2003. Testing for anxiety. Clinical Neuroscience Research 3, 233-238.
661 Ramos, A., 2008. Animal models of anxiety: do I need multiple tests? Trends Pharmacol. Sci.
662 29, 493-498.
663 Roberts, B., DelVecchio, W., 2000. The rank-order consistency of personality traits from
664 childhood to old age: A quantitative review of longitudinal studies. Psychol. Bull. 126, 3-25.
665 Robins, A., Phillips, C., 2010. Lateralised visual processing in domestic cattle herds responding
666 to novel and familiar stimuli. Laterality 15, 514-534.
667 Rodgers, R.J., 1997. Animal models of 'anxiety': where next? Behav. Pharmacol. 8, 477-496.
668 Rousing, T., Ibsen, B., Sorensen, J.T., 2005. A note on: On-farm testing of the behavioural
669 response of group-housed calves towards humans; test-retest and inter-observer reliability and
670 effect of familiarity of test person. Appl. Anim. Behav. Sci. 94, 237-243.
671 Russell, P.A., 1973. Relationships between exploratory behaviour and fear: a review. British
672 Journal of Psychology 63, 417-433.
673 Schönbrodt, F.D., Perugini, M., 2013. At what sample size do correlations stabilize? J. Res.
674 Personality 47, 609-612.
675 Schuetz, K.E., Hawke, M., Waas, J.R., McLeay, L.M., Bokkers, E.A.M., van Reenen, C.G.,
676 Webster, J.R., Stewart, M., 2012. Effects of human handling during early rearing on the
677 behaviour of dairy calves. Anim. Welfare 21, 19-26.
678 Scott, E.M., Nolan, A.M., Fitzpatrick, J.L., 2001. Conceptual and Methodological Issues Related
679 to Welfare Assessment: A Framework for Measurement. Acta Agriculturae Scandinavica,
680 Section A — Animal Science 51, 5-10.

681 Smith, B.R., Blumstein, D.T., 2008. Fitness consequences of personality: a meta-analysis.
682 Behav. Ecol. 19, 448-455.
683 Sousa, N., Almeida, O., Wotjak, C., 2006. A hitchhiker's guide to behavioral analysis in
684 laboratory rodents. Genes Brain and Behavior 5, 5-24.
685 Swiergiel, A.H., Dunn, A.J., 2007. Effects of interleukin-1β and lipopolysaccharide on behavior
686 of mice in the elevated plus-maze and open field tests. Pharmacology Biochemistry and
687 Behavior 86, 651-659.
688 Sylvers, P., Lilienfeld, S.O., LaPrairie, J.L., 2011. Differences between trait fear and trait
689 anxiety: Implications for psychopathology. Clin. Psychol. Rev. 31, 122-137.
690 Taylor, J., 1987. Kendall and Spearman Correlation-Coefficients in the Presence of a Blocking
691 Variable. Biometrics 43, 409-416.
692 Tecott, L.H., Nestler, E.J., 2004. Neurobehavioral assessment in the information age. Nat.
693 Neurosci. 7, 462-466.
694 Turner, S.P., Navajas, E.A., Hyslop, J.J., Ross, D.W., Richardson, R.I., Prieto, N., Bell, M.,
695 Jack, M.C., Roehe, R., 2011. Associations between response to handling and growth and meat
696 quality in frequently handled *Bos taurus* beef cattle. J. Anim. Sci. 89, 4239-4248.
697 Van Reenen, C.G., Van der Werf, J.T.N., O'Connell, N.E., Heutinck, L.F.M., Spoolder, H.A.M.,
698 Jones, R.B., Koolhaas, J.M., Blokhuis, H.J., 2013. Behavioural and physiological responses of
699 heifer calves to acute stressors: Long-term consistency and relationship with adult reactivity to
700 milking. Appl. Anim. Behav. Sci. 147, 55-68.
701 Van Reenen, C.G., Hopster, H., Van der Werf, J.T.N., Engel, B., Buist, W.G., Jones, R.B.,
702 Blokhuis, H.J., Korte, S.M., 2009. The benzodiazepine brotizolam reduces fear in calves
703 exposed to a novel object test. Physiol. Behav. 96, 307-314.
704 Van Reenen, C.G., O'Connell, N.E., Van der Werf, J.T.N., Korte, S.M., Hopster, H., Jones, R.B.,
705 Blokhuis, H.J., 2005. Responses of calves to acute stress: Individual consistency and relations
706 between behavioral and physiological measures. Physiol. Behav. 85, 557-570.

707 Van Reenen, C., Engel, B., Ruis-Heutinck, L., Van der Werf, J., Buist, W., Jones, R., Blokhuis,
708 H., 2004. Behavioural reactivity of heifer calves in potentially alarming test situations: a
709 multivariate and correlational analysis. Appl. Anim. Behav. Sci. 85, 11-30.
710 Waiblinger, S., Boivin, X., Pedersen, V., Tosi, M., Janczak, A.M., Visser, E.K., Jones, R.B.,
711 2006. Assessing the human-animal relationship in farmed species: A critical review. Appl. Anim.
712 Behav. Sci. 101, 185-242.
713 Wemelsfelder, F., Lawrence, A.B., 2001. Qualitative Assessment of Animal Behaviour as an
714 On-Farm Welfare-monitoring Tool. Acta Agriculturae Scandinavica, Section A — Animal
715 Science 51, 21-25.
716 Whay, H.R., Main, D.C.J., Green, L.E., Webster, A.J.F., 2003. Animal-based measures for the
717 assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion.
718 Anim. Welfare 12, 205-217.
719 Winckler, C., Brinkmann, J., Glatz, J., 2007. Long-term consistency of selected animal-related
720 welfare parameters in dairy farms. Anim. Welfare 16, 197-199.
721 Winckler, C., Capdeville, J., Gebresenbet, G., Horning, B., Roiha, U., Tosi, M., Waiblinger, S.,
722 2003. Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo.
723 Anim. Welfare 12, 619-624.
724 Windschnurer, I., Barth, K., Waiblinger, S., 2009. Can stroking during milking decrease
725 avoidance distances of cows towards humans? Anim. Welfare 18, 507-513.
726 Windschnurer, I., Boivin, X., Waiblinger, S., 2009. Reliability of an avoidance distance test for
727 the assessment of animals' responsiveness to humans and a preliminary investigation of its
728 association with farmers' attitudes on bull fattening farms. Appl. Anim. Behav. Sci. 117, 117-
729 127.
730 Windschnurer, I., Schmied, C., Boivin, X., Waiblinger, S., 2008. Reliability and inter-test
731 relationship of tests for on-farm assessment of dairy cows' relationship to humans. Appl. Anim.
732 Behav. Sci. 114, 37-53.

Wurbel, 2002. Behavioral phenotyping enhanced - beyond (environmental) standardization. Genes, brain, and behavior 1, 3.

## Tables

### Table 1

Fear scoring system in Experiment 3 based on stage at which the calf retreated from the approaching experimenter. The experimenter approached the calf or calves in the home pen, in a standardized way each time, and the calf was given a total of 2 min to approach or retreat.

| Score | Description |
|---|---|
| 9 | Retreat before arrive at pen |
| 8 | Retreat when face pen |
| 7 | Retreat when speak |
| 4-6 | Retreat during approach (each step the experimenter took towards the pen before a retreat reducing the score by 1) |
| 3 | Retreat when reached front of pen |
| 2 | Retreat when extend arm |
| 1 | Retreat during remainder of test |
| 0 | No retreat |

**Table 2**

Spearman correlation coefficients from Experiment 3. Coefficients describe the repeatability of responses to the approaching human test on commercial farms, depending on the response variable and exclusion criteria**.** Each of 14 farms was tested on two occasions. The highest coefficient for each response variable is indicated in bold.

| | **Exclusion criteria** | | | |
|---|---|---|---|---|
| **Response variable** | Sick calves included | Calves with diarrhoea excluded | Calves with coughs excluded | Both excluded |
| Proportion of calves that made contact | 0.222 | 0.279 | **0.332** | 0.253 |
| Proportion retreating when arm extended or before | **0.508** | 0.450 | 0.494 | 0.486 |
| Proportion retreating before arm extended | 0.516 | 0.477 | **0.521** | 0.486 |
| Average retreat score | 0.538 | **0.587** | 0.582 | 0.560 |

## Figures

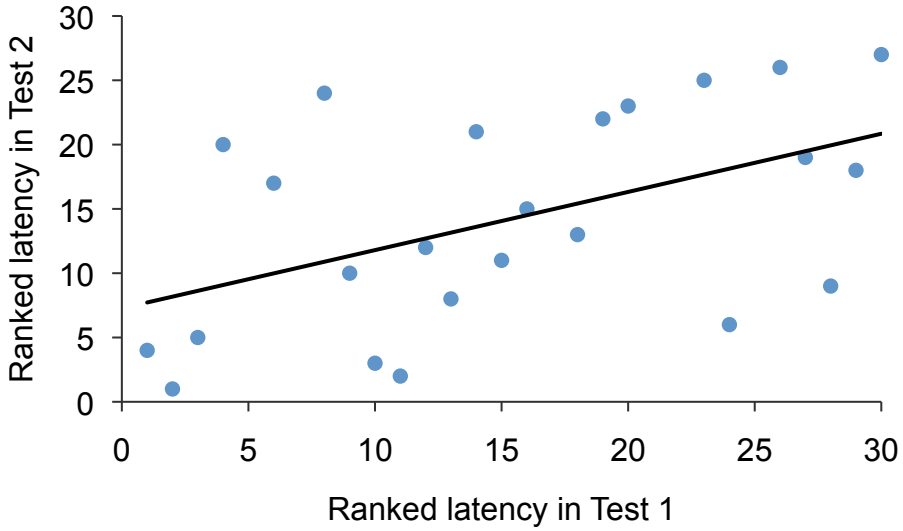
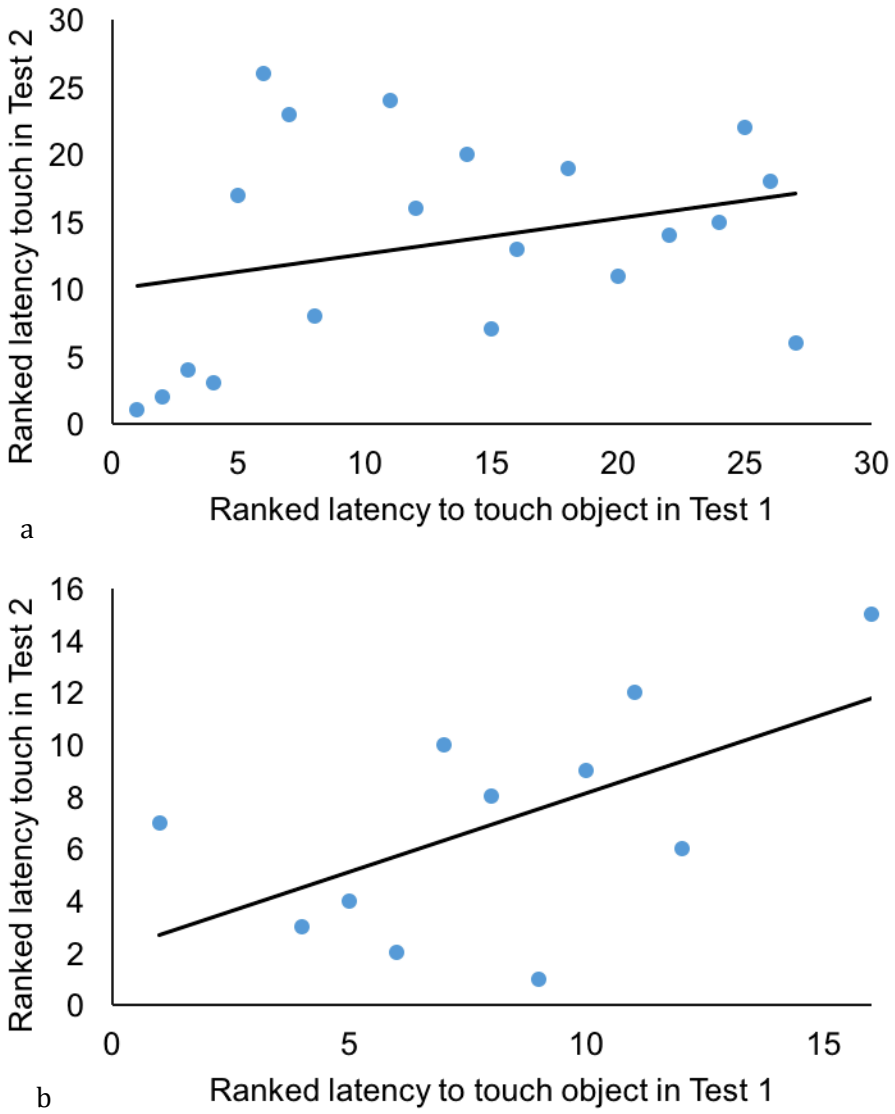



**Fig. 1.** Rank correlation between dairy calves' latencies to touch a novel object across two tests in Experiment 1. Calves were tested at approximately 41 and 62 d of age, with the object being a colourful ball.



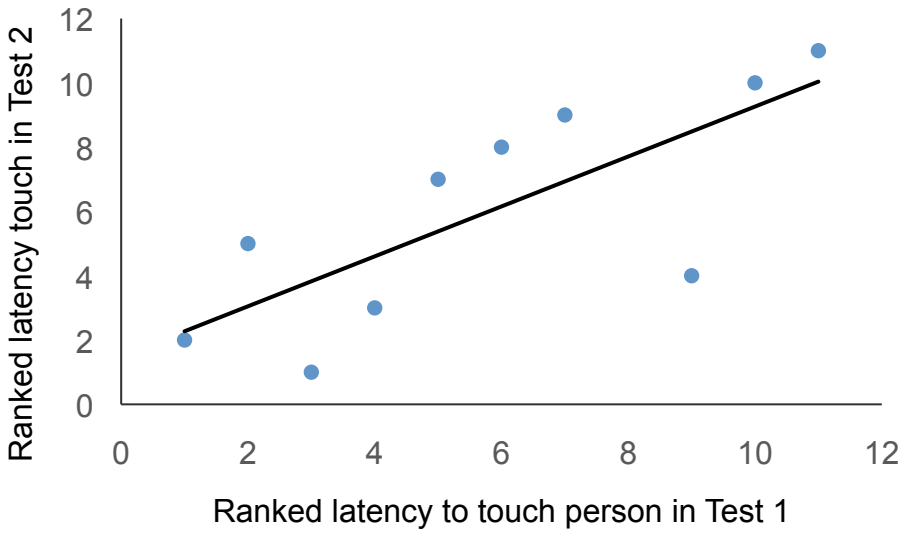

761

762 **Fig. 2.** Rank correlation between calves' latencies to touch a novel object across two tests in

763 Experiments 2a and b. Calves were tested at approximately 36 and 43 d of age, using a

764 different object each time in a) versus the same object in b).

765




766

767 **Fig. 3.** Rank correlation between calves' latencies to touch a familiar handler across two tests in

768 Experiment 2b. Calves were let out of their pens and given up to 3 min to touch a stationary

769 person.

770