Applied Animal Behaviour Science 133 (2011) 254–264

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim

Who's afraid of the big bad glove? Testing for fear and its correlates in mink

Rebecca K. Meagher*, Ian Duncan, Allison Bechard, Georgia J. Mason

Animal and Poultry Science Department, University of Guelph, 50 Stone Road East, Building #70, Guelph, Ontario, Canada N1G 2W1

ARTICLE INFO

Article history:
Accepted 16 May 2011
Available online 12 June 2011

Keywords:
Fear
Temperament
Mink
Welfare
Reproductive success
Reliability
Validity

ABSTRACT

Fear in farm animals is a welfare and economic concern. For Scandinavian mink, the "stick test" is common for assessing fearfulness: a spatula is inserted into the cage and minks' immediate responses are noted. However, on Ontario farms, fearfulness in the stick test was very rare and aggressive responses were prevalent, rendering this test poor for testing hypotheses related to fear and welfare. We therefore developed a modified version, the "glove test", where the finger of a handling glove is inserted into the cage. This proved more sensitive than the stick test for detecting fearfulness in Ontario mink (20% vs. 2.5%, $P < 0.0001$); and successfully reduced aggressive responding (22% vs. 41%, $P < 0.0001$). When test-retest reliability was assessed conventionally, it was only moderate (e.g., 37.5% mink behaved the same in three tests). However, it is biologically realistic to expect habituation over repeated trials (e.g., reduced fearfulness), and treating such changes as acceptable, results were reliable for 76% of mink over three tests. Reliability could be further improved by taking location into account, since some mink were unresponsive only if in the nest box, half-asleep (Experiment 3: kappa of 0.38 if never in nest box vs. kappa of 0.002 if were). Glove tests had construct validity: mink immediately classified as "fearful" spent more time exhibiting other fear-related behaviours ("ambivalence": mean 30 s vs. 4 s, $P = 0.009$), while mink immediately classified as "curious" then spent more time investigating the glove (mean 103 s vs. 57 s, $P < 0.0001$). Glove tests also revealed expected temperament differences between Black and Pastel colour-types, with Pastels being less fearful ($P = 0.001$). Finally, we tested whether fearfulness in the glove test is associated with decreased reproductive success. Pastel mink that were fearful during the presumed gestation period were less likely to reproduce ($P = 0.006$). Like stick tests, glove tests are thus practical, valid and reliable for assessing fearfulness in farmed mink, but they are better for detecting fearfulness in low fear populations.

1. Introduction

Fear in farm animals has serious economic, practical and welfare implications (Hemsworth, 2003; Jones, 1997). Attention has therefore been paid to decreasing fear through altered management (e.g., Hemsworth et al., 1989) or selective breeding (e.g., Hansen, 1996). Methods of fear assessment vary, however, and many have been insufficiently validated (Forkman et al., 2007). Farmed mink (*Neovison vison*) exhibit some fear of humans, since they are less domesticated than other livestock (Nimon and Broom, 1999), and are handled only sporadically (e.g., for relocating or vaccinating) by stockpersons wearing thick gloves to protect against bites (European Commission, 2001). This handling is acutely stressful (Hansen and Damgaard, 1991; Korhonen et al., 2000). The fear mink experience under typical farm practices is thus a potential welfare issue.

* Corresponding author. Tel.: +1 519 824 4120x53557; fax: +1 519 836 9873.
E-mail address: rmeagher@uoguelph.ca (R.K. Meagher).

0168-1591/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.applanim.2011.05.009

The most common test for fearfulness is the "stick test" (e.g., Hansen, 1996; Kirkden et al., 2010), which categorizes mink as fearful, curious (sometimes called "confident"), or aggressive based on their immediate response to a wooden spatula held through the mesh of the cage. This test, developed and validated in Scandinavia, has proved very successful (see e.g., Kirkden et al., 2010): it is rapid, averaging 9 s per subject; can have good inter-observer reliability (Hansen and Møller, 2001); and has been used to select lines of fearful or confident mink which then differ consistently in physiological stress responses (e.g., Hansen, 1997) and their behaviour in diverse other fear tests (Malmkvist and Hansen, 2002). We therefore wished to use the stick test on Ontario farms, to test welfare-relevant hypotheses about relationships between fearfulness and stereotypic behaviour, reproductive success, and environmental enrichment. However, informal pilots soon revealed that very few Ontario mink showed fearful responses, giving the stick test minimal usefulness for distinguishing between individuals. A more sensitive test (i.e., able to detect fear in more animals) was therefore needed. Aggressive responses were also very common: a problem for our research because it is unclear how to interpret these in term of welfare.

On Scandinavian farms, the "Trapezov hand test" is the recommended alternative test with higher sensitivity to fear, being more threatening: the cage lid is opened, a gloved hand is reached in, and if possible, the mink is touched; the mink's response is scored (Malmkvist and Hansen, 2002; Trapezov, 2000). However, this test was impossible on our farms because cage design differed from that typical of European farms: the nest box is placed at the back of the cage rather than the front; it sits within a hole in the cage roof; and the only way to open the cage is to lift this nest box out. Reaching in to conduct a hand test would thus be very slow; would involve flushing out any mink in nest boxes; and most importantly, would leave the animal with no clear place for retreat (the nest box is gone, plus the hand is entering from the back of the cage, but the human tester is standing at the front), making fearful responses challenging to score. We thus needed to develop and validate a variant on the stick test that would be more appropriate for North American farms.We therefore modified the stick test by using a well-used handling glove – a leather gauntlet used on-farm for catching animals – in place of the stick. We hypothesised that this would be more aversive than a stick, thence better at eliciting fear, because the mink may associate the glove with past handling events; it also smelled of the scent that mink release when frightened (Dunstone, 1993). This was expected to increase the probability of an active response, including fearful ones, by increasing arousal and vigilance (cf. Zalaquett and Thiessen, 1991). The second aim of using a glove was to reduce aggressive responses that might reflect predation. The spatula used in the stick test is similar in length to minks' average prey (15–20 cm: Dunstone and Sinclair, 1978); biting it often involves attempts to pull it into the cage (Kirkden et al., 2010); and mink are known to chew and hoard small objects (Axelsson et al., 2009, personal observation). This suggests that at least some 'aggressive' stick-biting may reflect predatory or even playful motivations to obtain it for carrying, chewing and hoarding. We hypothesised that the finger of a glove might be less likely to elicit such responses because more aversive than the stick, and part of an object larger than typical prey.

Here we present three experiments exploring this new test's utility. In Experiment 1, we assessed its test-retest reliability (repeatability), and its construct validity (its ability to measure what it is meant to be measuring: Cronbach and Meehl, 1955; Meagher, 2009). In Experiment 2, we replicated our previously-unrecorded pilot study, directly comparing glove and stick tests to determine whether the former was indeed more sensitive to fear and less prone to eliciting aggression. In Experiment 3, we used the glove test to assess whether fearfulness predicts reduced reproductive success, as it does in some livestock (e.g., Hemsworth et al., 1989; Hemsworth, 2003).

2. Experiment 1: Validation of the glove test

2.1. Methods

2.1.1. Subjects and timing

All mink were adult females of two colour-types, Black and Pastel, on a commercial farm, individually housed in standard cages, in several different sheds. These cages are arranged in long rows with opaque partitions between them, and furnished with a wooden nest box. Tests were conducted in September, and immediately post-feeding to ensure that as many animals as possible were awake. A total of 187 mink (113 Blacks, 80 Pastels) were tested multiple times to assess test-retest reliability, while an additional 46 (18 Blacks, 28 Pastels) were tested once in a longer test for assessing construct validity.

2.1.2. General test protocol

The experimenter held the glove against the cage-front with one empty finger extended through the wire. So that all subjects had the same pre-test exposure to the experimenter, the test was always first performed on the mink in the neighbouring cage, even if that animal was not included in the study, and a well-used glove was used so that it would not shift from odor-free to odorous during the course of testing. The mink was exposed to the stimulus for up to 10 s, or 30 s if it apparently failed to notice the glove (see below).

As is standard in stick tests, mink were categorized according to their immediate reaction (e.g., Hansen, 1996; Kirkden et al., 2010; Korhonen et al., 2002). The categories used, modelled on those of Hansen (1996) were: (1) fearful, if the mink retreated (note that withdrawing into the nest box from a lying position with the head out of the box was considered a retreat); but in contrast to Hansen (1996), we also included in this category mink that remained standing still at the far side of the cage and oriented to the stimulus for at least 10 s, because no retreat was possible for these subjects and their 'freezing' suggested fear (cf. Blanchard and Blanchard, 1969; Malmkvist, 2001); (2) curious, if it approached and made contact with the glove without a hard bite, i.e., without closing its teeth; (3) aggressive, if it gave a hard bite to the glove, clamping its teeth together fully; and (4) other, if the mink was alert and initially oriented to the front of the cage but did not respond in any of the above ways, i.e., it either exhibited no obvious behavioural response (thus called "unresponsive"), or performed some other activity such as stereotypic behaviour. Mink that remained stationary and oriented to the stimulus fell into the "other" category if they were not standing at the far back of the cage, as described for "fearful", because it was less clear whether this was a 'freezing' response.

Any mink that could not be tested and thus assigned to a category was excluded from analyses. This occurred if an individual was out of sight in the nest box (i.e., lying with the head down, in which case they were not visible through the opening) or resting/sleeping either in the nest box or the open cage such that it did not appear to notice the glove within 30 s on two separate attempts. In some cases, the decision that mink were untestable was also made at the group level: if more than half of the remaining animals in a section of a shed were resting, it was assumed that they had entered their post-feeding rest period (Hansen and Møller, 2008) and testing ceased in that section.

2.1.3. Reliability

The test was repeated on five consecutive days to investigate patterns of change. The order in which the sheds were tested was reversed on alternate days (thus no group was always tested first or last). Traditional test-retest reliability is based on the proportion of subjects showing identical responses from one trial to the next. However, some decrease in fear due to habituation was expected over repeated tests (e.g., Kirkden et al., 2010). This has been observed in the stick test (Malmkvist and Hansen, 2001), and more generally, decreased responsiveness is expected with repeated presentations of a stimulus that does not either provide a reward or cause harm (see e.g., Domjan, 2003). Thus, we predicted that with re-retesting, some initially fearful mink would become curious, while fearful, curious or aggressive mink might become unresponsive. We judged that only changes that did not fit these expected patterns would indicate a problem with the test that would justify calling it "unreliable". Therefore, in addition to testing for test-retest reliability in the orthodox manner, we also assessed reliability in a more biologically-relevant way taking these predicted changes in response into account.

Such habituation effects might complicate the interpretation of test-retest reliability measures, but it is important to determine whether changes over time are due to habituation or are apparently random. In the latter case, the use of even a single test would have questionable biological relevance. In the former, a single test should instead reflect temperament; furthermore, repeated testing could be used to increase the data per individual and thus the statistical power (by summing results; cf. Kenttämies et al., 2002), and/or to reduce noise by eliminating those individuals whose responses seem to vary due to chance effects (see Experiment 3).

2.1.4. Construct validity

A separate, naïve group of 46 mink was screened in one glove test. This test was carried out as above except that the test duration was 3 min instead of 10 s or less. The aim was to check that the immediate response used to classify the mink predicted other aspects of behaviour consistent with that classification. Quantitative data were recorded, using a stopwatch, for durations of "ambivalent behaviour" and total interaction with the stimulus. Ambivalent behaviour was defined as alternating approach and withdrawal while oriented to the glove, an "oscillation" associated with conflicts between motivations to avoid and explore (Gray, 1987; Miller, 1944); it was thus expected to be most common in mink categorized as fearful based on their immediate response. Those mink instead categorized as curious based on their immediate response were expected to show more prolonged interaction with the glove over the 3 min, compared with fearful mink or those initially classified as "unresponsive".

2.1.5. Statistical analysis

For the reliability tests, kappa statistics were calculated (Cohen's kappa for comparisons between two tests; Fleiss' kappa for more than two tests). Because many mink were not alert every day and could therefore not be tested five times, these statistics were calculated across the first three tests, as well as being calculated across all five tests in the subset of animals for which that was possible. Mink that could not be tested at least three times were excluded from the reliability analysis. In addition, to account for expected changes due to habituation, we calculated the percent of mink that followed this predicted habituation pattern. Before this, McNemar's tests were used to check for population-level changes confirming this type of habituation. Differences between colour-types were assessed using Pearson's chi-square tests on data from the first test. The above analyses were conducted using Minitab 14 (Minitab Inc., PA, USA).

Behavioural differences between temperament categories, as assigned according to immediate response, were analysed first using General Linear Models to include colour-type. Colour-type was not significant in these validation tests and was therefore removed so that simple Analysis of Variance (ANOVA) could be employed. Normality was assessed using Anderson–Darling tests, and Levene's tests for equal variances were conducted. Where variances were not equal ($P < 0.05$), Welch's ANOVAs were used to accommodate differences in variance (Welch, 1951); planned contrasts were assessed in the same way. These analyses were carried out in JMP 8 (SAS Institute Inc., NC, USA).

2.2. Results

2.2.1. Behaviour in the five reliability tests

Twenty-seven mink could not be tested at least three times, and so were excluded. Of the remaining 160, 75 yielded data from all five tests. In the full sample of 160, 20% were fearful on the first test; 62%, curious; 10%, aggressive; and 8% were categorized as "other". Mink categorized as "other" were all behaviourally unresponsive (i.e., exhibited little to no locomotion during the test), and so are called "unresponsive" henceforth.

Temperament distributions on the first test varied between colour-types ($\chi^2 = 14.9$, d.f. = 3, $P = 0.002$). Blacks were more likely than Pastels to be fearful (29% vs. 7%;

Table 1
Pairwise reliability comparisons for glove tests repeated once a day for five days.

Tests compared	Kappa (N = 160)
1 vs. 2	0.3184 ± 0.0567
2 vs. 3	0.3623 ± 0.0535
3 vs. 4	0.4904 ± 0.0553
4 vs. 5	0.4574 ± 0.0805

Values are kappa ± standard error. Kappa values greater than 0.40 indicate “moderate reliability” (Dohoo et al., 2003).

$\chi^2 = 11.8$, d.f. = 1, $P = 0.001$), while Pastels were more likely to be unresponsive (13% vs. 4%; $\chi^2 = 4.1$, d.f. = 1, $P = 0.04$).

2.2.2. Test-retest reliability

Test-retest reliability was fairly poor, assessed in the traditional way without taking predicted habituation patterns into account (such that all changes in temperament category, even expected ones, were counted as unreliable). Thus for the 160 mink for which reliability could be calculated over three tests, only 60 (37.5%) behaved consistently. This produced a Fleiss’ kappa of 0.33, indicating only “fair agreement” (Dohoo et al., 2003). For the subset tested the full five times, 20 individuals (26.7%) were consistent across all tests, and Fleiss’ kappa was 0.40. Reliability seemed to increase somewhat over repeated tests, as shown in Table 1, with pairwise kappas between the later tests being over 0.40, thus indicating “moderate reliability” (Dohoo et al., 2003).

However, population-level changes over repeated testing showed that mink were following the pattern predicted by habituation (Fig. 1). Thus, fewer of these 75 mink tended to be fearful by the last test (McNemar’s test, $\chi^2 = 3.0$, $P = 0.08$), while significantly more were unresponsive by the last test (McNemar’s test, $\chi^2 = 10.7$, $P = 0.001$). (More mink also became aggressive by the last test: McNemar’s test, $\chi^2 = 7.4$, $P = 0.007$). When the two types of change expected due to habituation (becoming less fearful/more unresponsive) were taken into account, a more optimistic view of reliability emerged: 122 mink (76.2%) were consistent or changed in predicted directions over three tests. Over five tests, however, this value was lower, with only 35 of 75 (46.7%) fitting these criteria.

2.2.3. Construct validity

Sample sizes, means and standard errors for both validation measures are presented in Table 2. For ambivalence duration, the assumption of equal variances was not met; this variable was also non-normal (ambivalence AD = 7.6, $P < 0.005$), as were the residuals obtained from ANOVAs, nor could this be corrected by transformation. However, the residuals appeared approximately normal by inspection even for untransformed data. Because the equivalent non-parametric test (Kruskal–Wallis) is not advised if variances are non-homogenous (Maxwell and Delaney, 2004), and ANOVAs are robust to deviations from normality (see e.g., Cochran and Cox, 1957) we therefore used Welch’s ANOVAs to test the differences in the means of the untransformed data for those two variables.

Mink classified into different temperament categories based on their immediate responses differed in both duration of contact and ambivalence (Table 2). Curious mink spent more time in contact with the glove than did fearful or unresponsive mink (contrast of curious vs. other categories: $F_{1,43} = 23.34$, $P < 0.0001$; Fig. 2). Fearful mink, meanwhile, displayed more ambivalence than did mink in the other two categories (Fig. 3; contrast $F_{1,9} = 11.06$, $P = 0.009$).

2.3. Discussion

Categorizing mink via their immediate response to the glove in a single test had construct validity: mink categorized as fearful on this basis then exhibited conflict between approach and avoidance motivations during more in-depth testing, while mink categorized as curious then exhibited the most sustained interest in the glove. The test also revealed expected differences between colour-types: Black mink, generally considered more nervous and restless than Pastels (European Commission, 2001), were more likely to be fearful. The test’s repeatability over consecutive days appeared low; by comparison, Hansen and Møller

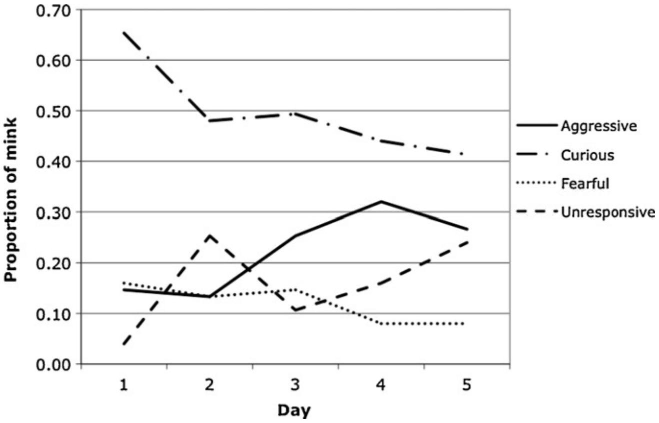

Fig. 1. Population-level changes in temperament distribution over repeated tests (Experiment 1). Proportions of mink falling into each temperament category, including only the 75 mink for which five tests could be completed. Note that no directional prediction was made for the observed change in aggression.

Table 2
Behaviour in validation tests: tests for unequal variances between temperament categories, as determined by minks' immediate responses to the stimulus.

Measure	Curious ($N=22$)	Fearful ($N=9$)	Unresponsive ($N=15$)	Levene's test P-value[a]	Difference in means
Duration of contact (s)	103.0 ± 6.0	58.0 ± 6.2	57.1 ± 10.5	0.08	$F_{2,43}=12.10$, $P<0.0001$
Duration of ambivalence (s)	1.0 ± 3.0	30.6 ± 4.7	9.4 ± 3.6	0.0002	$F_{2,14.5}=14.16$, $P=0.004$

Data presented are means ± standard error; test duration was 180 s.

[a] Test for equal variances among the three categories.

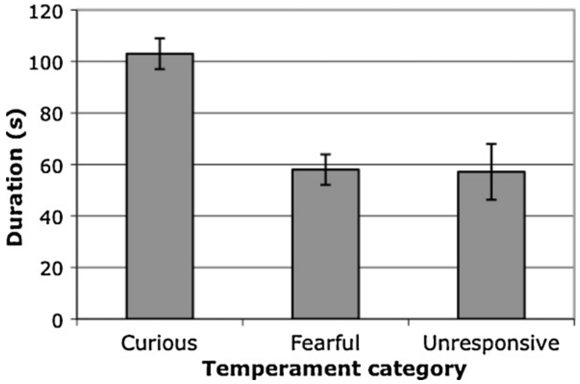

Fig. 2. Average total duration of contact with glove over 3 min test, grouped by immediate response. Data are means ± standard error.

(2001) present data from two pairs of stick tests from which we can calculate kappas of 0.58 and 0.77 for test–retest reliability. However, much of the apparent 'unreliability' in the current experiment was successfully explained by habituation and therefore did not invalidate the test. This issue of habituation has not yet been raised with regard to the stick test, likely because when multiple tests have been conducted, they were spread over a long period of time, which reduces habituation effects (Domjan, 2003). However, in some experiments, temperament must be determined over a relatively short time period to test causal relationships. Repeated testing is useful in those situations (see Section 2.1.3), and so habituation effects are important to consider. The one change observed that was not predicted *a priori* from habituation was the increase of aggression over time, addressed in Section 5.

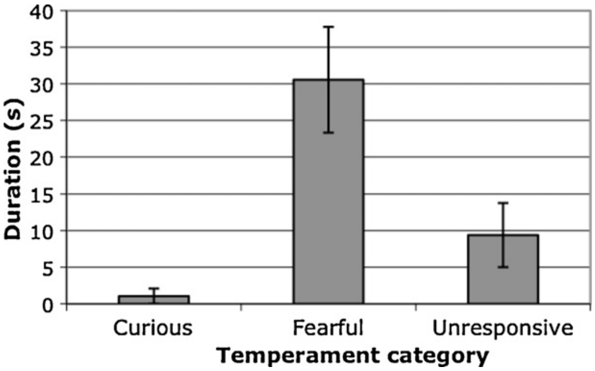

Fig. 3. Average total duration of ambivalence over 3 min test, grouped by immediate response. Data are means ± standard error of time spent alternating between approach and avoidance of the glove.

Several times, mink were asleep immediately prior to the test: a phenomenon that we suspect adds noise to the test results. When sleeping mink awoke and saw the glove, they were tested, but were often awake only briefly, and seemed more likely than others to be classified as "unresponsive". Mink that were in this 'drowsy' state on one test day but not on the following one might therefore be expected to change apparent temperament categories, not because of progressive habituation but because of the chance events of being awake or asleep as the tester approached–chance events that might then reduce the test's apparent reliability. We suspect that not only is the "unresponsive" category subject to chance events like this, but also that it is motivationally heterogeneous: the lack of an active response (whether contacting the glove or withdrawing from it) may thus sometimes reflect temporary 'drowsiness', as above, but in other cases it could be due either to a genuine lack of interest, or in other cases still, to fear causing freezing (mink labelled "unresponsive" did display some ambivalent behaviour during the longer tests of the validation phase). The idea that "unresponsive" may be a heterogeneous category is returned to in Section 5.

3. Experiment 2: Comparison of glove and stick tests

3.1. Methods

3.1.1. Subjects and timing

This experiment aimed to formally replicate our informal pilot observations on the low sensitivity of the stick test for detecting fear in Ontario mink. Subjects were 121 adult female Black mink at another commercial farm. Tests were conducted in December, one to two weeks after pelting had been completed on the farm.

3.1.2. Test protocol

Based on Experiment 1 findings, we modified the test protocol in two ways. First, we added a statistical control for whether mink were in the nest box, where most sleeping occurs, and thus likely to be 'drowsy'. Second, we increased the time allowed for mink to respond from 10 s to 30 s. This was to increase the likelihood that when individuals were categorized as "unresponsive", it reflected a genuine temperament trait rather than a temporary 'drowsiness'.

All tests were conducted in the morning, beginning approximately 3 h before feeding time. On the first day, half of the mink were exposed to the stick test, and half to the glove test. The following day, this was reversed to control for effects of test order. The protocol from Experiment 1 was followed for both stimuli, except that, as discussed above, mink were allowed 20 s longer to respond. Location

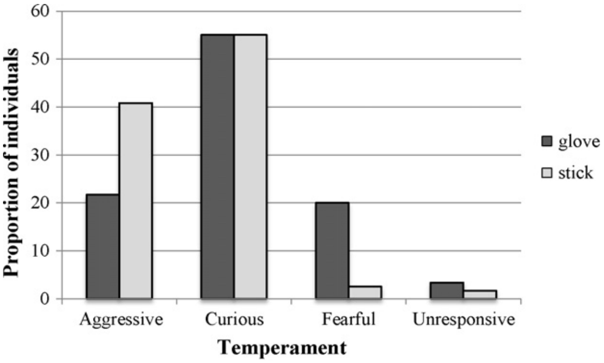

Fig. 4. Comparison of temperament distributions for stick and glove tests. $N = 120$, all exposed to both tests.

(whether in or out of the nest box) was also now noted at the start of each test to control for drowsiness. As before, if a sleeping individual did not wake up when the stimulus was placed in the cage on two 30-s attempts, no category was assigned.

The proportions of mink categorized as fearful and aggressive in each test were compared using Fisher's exact tests, or McNemar's tests where repeated measures from the same individuals were involved. Kappa statistics for agreement between the two tests were calculated, both overall and split by day.

3.2. Results

3.2.1. Temperament distributions

The proportions of mink assigned to each category are shown in Fig. 4, and comparisons on an individual level in Table 3. One individual was excluded for remaining asleep during attempted tests, leaving 120 mink in the study. Significantly more mink were fearful in the glove than the stick test (McNemar's test, $\chi^2 = 19.2$, $P < 0.0001$). Test order affected fearfulness in the glove test: mink were more likely to exhibit fear if they were glove-tested on the first than on the second day (Fisher's exact test, $P = 0.0004$). Results of the two tests were therefore also compared separately for each day. The proportion of mink that were fearful was significantly higher using the glove test than the stick test on Day 1 (Fisher's exact test, $P < 0.0001$); the effect was in the same direction on Day 2, albeit not significantly (Fisher's exact, $P = 0.119$). No significant effect of test day on fearfulness was detected for the stick test (Fisher's exact test, $P > 0.10$), but statistical power was very low (see Table 3) and the only three individuals that exhibited a fearful response to the stick were tested on the first rather than the second day.

Meanwhile, significantly more mink were aggressive in the stick test than the glove test (McNemar's test, $\chi^2 = 18.2$, $P < 0.0001$). As above, test day affected the glove test: more mink tended to exhibit aggression in this test if this was the second test to which they were exposed ($\chi^2 = 3.1$, $P = 0.076$). Again, test day did not significantly affect the stick test (Fisher's exact, $P > 0.10$).

3.2.2. Correlation between tests

Table 3 compares the two tests' results on an individual basis. Of 120 mink, 68 (57%) were assigned the same category using both tests. This produced a kappa of 0.2729. For agreement in simply categorizing mink as fearful or not, the kappa value was very low at 0.0969, considered poor reliability. Relatively few mink were in the nest box at the beginning of the test, and so reliability was not improved by excluding them (kappa = 0.2669, $N = 111$).

3.3. Discussion

Our previous impressions from pilot trials were confirmed: the stick test categorized hardly any Ontario mink as fearful, but many as "aggressive". In contrast, the lowest reported prevalence of fear in adult females in Scandinavian studies using the stick test was 22% (Møller and Hansen, 2001) compared to our 2.5%. The prevalence of aggression was also low e.g., 12% (Møller and Hansen, 2001) vs. our 41%. Experiment 2 also supported our hypothesis that the glove test would be more sensitive than the stick test to fear, and less prone to eliciting aggression. Contrary to our expectations, the mink's initial location did not seem relevant in this experiment, but this was likely because only nine individuals were in the nest box at the start of either test, and most were awake so unlikely to be 'drowsy'.

Because of the shifts in the number of mink exhibiting fear or aggression, agreements between stick and glove tests were poor, especially for categorizing fear. This is somewhat surprising given that past research suggests that fear in the stick test generalizes to other tests (Malmkvist and Hansen, 2002). However, the floor effects on the stick test data here (which classed only three of 120 mink as fearful) made it statistically impossible to determine whether individuals that were fearful in that test were also fearful in the glove test. A second possible explanation for the lack of agreement between tests is that both the test person and general testing procedure were completely novel on the first test day, but not the second. The glove test

Table 3
Comparison of results in the glove test vs. stick test applied to the same sample of mink.

	Stick test				
	Aggressive	Curious	Fearful	Unresponsive	Totals for glove test
Glove test					
Aggressive	23	3	0	0	26
Curious	23	42	1	0	66
Fearful	3	18	2	1	24
Unresponsive	0	3	0	1	4
Totals for stick test	49	66	3	2	120

seemed especially sensitive to such 'order effects', as discussed below.

The glove test's sensitivity to fear appeared affected by the novelty of the situation (e.g., the unfamiliarity of the test person): more mink were fearful if given this test on the first rather than the second day. The reverse was true for aggression. These findings are rather consistent with the habituation data from Experiment 1, and suggest that being naïve rather than having been previously tested (in *any* test) makes mink more fearful and less aggressive in glove tests. In Experiment 2 (unlike Experiment 1), the subjects were also naïve to the experimenter on the first day: unfamiliarity of the tester may thus interact with the stimulus presented to affect the sensitivity of a fear test. A longer interval between tests might reduce this 'day effect', but if maximum sensitivity to fear is needed for the purposes of research or animal selection, then these data suggest results of the first glove test should be used.

4. Experiment 3: Reproductive correlates of fear in the glove test

This final experiment aimed to determine whether fearfulness predicted poor reproductive success, since reproductive suppression is a common consequence of chronic stress (Wingfield and Sapolsky, 2003), and fear of humans specifically can reduce reproductive output (see Section 1).

4.1. Methods

4.1.1. Subjects and locations

Subjects were 546 female mink in their second or third breeding seasons, spread across three farms. All had been mated in late February. Of these, on Farm 1 there were 147 Blacks and 148 Pastels, on Farm 2, 140 Blacks, and on Farm 3, 111 Pastels. Farms 1 and 2 were those used in Experiments 1 and 2, respectively. One researcher worked at Farm 1, another at Farms 2 and 3.

4.1.2. Test protocol and measures of reproductive success

Due to concerns about reliability, we continued testing each mink three times. This allowed us to eliminate individuals that changed responses seemingly randomly, since we were not confident that these responses reflected underlying temperament. The results of the sensitive first test were then used to test our hypotheses for all remaining mink. The protocol was thus the same as in Experiment 2, but repeated three times over a four-day period; and mink categorized according to their response on the first test, but only if they were then consistent across all three tests or if any change was in a direction predicted by habituation (see Experiment 1). Mink not meeting this criterion were excluded from analyses. To control for the possible confound of 'drowsiness', reliability was then compared between individuals that were sometimes in the nest box during tests and those that were not. All testing occurred in early April, during the period between mating and being moved to the cages where they would give birth.

Reproductive success was assessed via failure to have a litter (hereafter termed 'barrenness'), litter size counted on the day after birth, and kit mortality in the first three weeks, assessed as a proportion of original litter size. Body fat was scored by the farmers on a three-point scale just after mating, because this influences reproduction (Baekgaard et al., 2007; Malmkvist and Palme, 2008).

4.1.3. Statistical analysis

Glove test reliability was assessed again using kappa statistics, overall and split by farm. Population level changes in habituation were assessed using McNemar's tests comparing Tests 1 and 3 for each category, split by colour; for simplicity of presentation, farms were pooled unless the relationship differed between them. Likelihood of barrenness was compared between fearful mink and all others within each colour-type using Pearson's Chi-square tests, or Fisher's exact tests if any cells had counts of less than five. Because litter size and infant mortality data were not normally distributed according to Shapiro–Wilk tests, generalized linear models were used to test whether fearfulness predicted either variable. For testing whether fearfulness predicted litter size, we used a model with a Poisson distribution and a log link function. The relationship of infant mortality with fearfulness was assessed in a binomial logistic regression. Both models controlled for farm, colour-type and, in the case of litter size, body fat score. All analyses were conducted in JMP 8 (SAS Institute Inc., NC, USA).

4.2. Results

4.2.1. Reliability of the glove test and habituation effects

Sixteen mink were excluded because they were not tested three times due to remaining asleep. For the 530 others, the overall kappa was 0.33. However this value was affected by whether mink were in the nest box when tested: thus for the 405 individuals that were outside the nest box during all three glove tests, the kappa statistic for agreement among all three was 0.38 ("moderate"), compared to just 0.002 for the 125 that were in the nest box at least once. Reliability also differed between farms: overall kappa values were 0.40, 0.21 and 0.28 at Farms 1, 2 and 3, respectively.

As in Experiment 1, population-level changes in temperament from the first to the last test were consistent with habituation (Fig. 5). Fear decreased from 13% to 4% in Blacks (McNemar's $\chi^2 = 13.4$, $P < 0.0001$), and no Pastels were fearful on Test 3 while 5% had been on Test 1. Unresponsiveness increased from 5% to 31% in Blacks and from 10% to 31% in Pastels (McNemar's $\chi^2 = 66.4$, 43.8 respectively; both $P < 0.0001$). Aggression increased only at Farm 1, in both colour-types (Blacks: 24% to 38%, McNemar's $\chi^2 = 13.4$, $P = 0.0002$; Pastels: 36% vs. 49%, McNemar's $\chi^2 = 8.5$, $P = 0.004$).

4.2.2. Colour and farm effects on fear

Overall, 12% of categorizable mink were fearful in the first test; 45%, curious; 37%, aggressive, and 7% unresponsive. As in Experiment 1, Blacks were more likely than Pastels to be fearful on Farm 1, the only farm with both colour-types: $\chi^2 = 13.4$, d.f. = 1, $P = 0.0002$). Fearfulness also differed between farms: a chi-square test between farm by

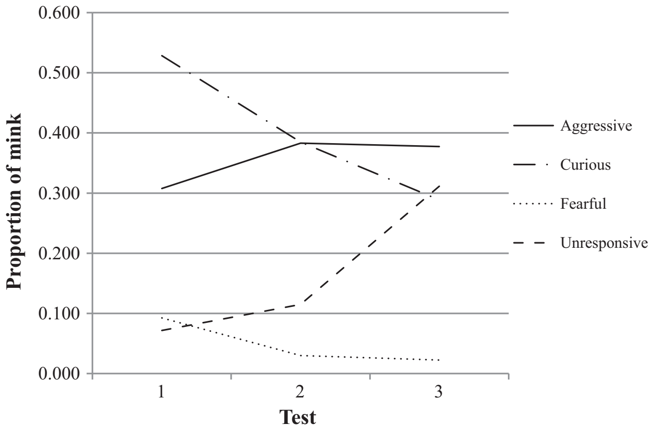

Fig. 5. Population-level changes in temperament distribution over three tests (Experiment 3). Proportion of all mink tested, with colour-types pooled ($N = 530$).

colour groups was significant ($\chi^2 = 12.8$, d.f. = 3, $P = 0.005$), due to a difference between Pastels at Farms 1 and 3 ($\chi^2 = 8.9$, d.f. = 1, $P = 0.003$).

4.2.3. *Fear vs. reproductive success*

Among Pastels, fearful mink were more likely to be barren (Fisher's exact, $P = 0.006$; Fig. 6). When split by farm, this difference appeared only at Farm 3 (Fisher's exact $P = 0.04$; 44% of 9 fearful individuals vs. 14% of 86 non-fearful individuals); but this could reflect low statistical power since only four Pastels were barren at Farm 1 (and Farm 2 had no Pastels). There was no significant difference between fearful and non-fearful mink among Blacks. Among mink that successfully produced kits, no significant relationship between fearfulness on litter size or kit mortality were detected ($P > 0.05$).

4.3. *Discussion*

Previous studies provided evidence of relationships between fearfulness and reproduction in mink, but its manifestations have varied. Malmkvist et al. (1997) found that mink from a line selected for fearfulness mated later in the season than did those from a confident line, suggesting reduced motivation to mate; however, once mated, their reproductive success was unaffected. Korhonen et al. (2002) by contrast, found no difference in mating dates but that females who were fearful in stick tests then gave birth to smaller litters than did confident females. Among primiparous females, rates of infertility were also higher in fearful individuals, as in Pastels in the current study, but this difference was not statistically significant. Fearfulness has likewise been linked to infertility or decreased rate of breeding in a variety of other captive species, from chickens (Shabalina, 1984) to cheetahs (Wielebnowski, 1999). Section 5 discusses potential mechanisms mediating the link between barrenness and fearfulness in our mink.

5. General discussion

In stick tests, our Ontario mink did not react as expected from published Scandinavian data, being far less fearful and more aggressive. It is unclear why – differences in age, genetics, husbandry, and subtle aspects of the test such as season and the height and sex of testers could all play roles – but our need for a test with greater resolution and sensitivity, applicable to North American-style cages that preclude using the Trapezov hand test, led to us successfully modifying the stick test into a glove test. This combined the basic stimulus of the Trapezov hand test

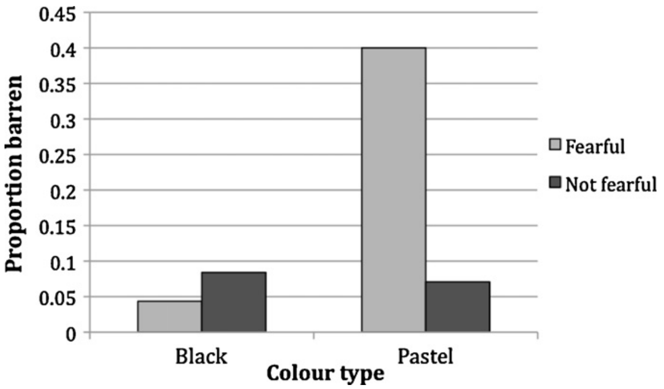

Fig. 6. Proportion of fearful vs. non-fearful mink that failed to reproduce. $N = 237$ Black (23 fearful), 236 Pastel (10 fearful).

with the advantages of the stick test, being quick and simple because conducted from outside the cage, and therefore able to be used on large numbers of animals and by testers inexperienced at handling mink. Furthermore, the glove test did indeed prove more sensitive than the stick test for investigating fear, increasing the number of mink classified as fearful in our tests by about eightfold. This is important for populations where fearfulness is relatively rare, and so where tests using more neutral stimuli might not detect fear in enough animals to assess its correlates reliably.

The glove test thus adds to a range of existing temperament tests that differ in how challenging they are for mink, ranging from the Trapezov hand test at the most stressful end of the spectrum, best for identifying individual differences within very confident populations (Kirkden et al., 2010), through to the least threatening "food" or "titbit" tests (where the human offers food; see Hansen and Møller, 2001) and "walk by" tests (where the tester simply passes the cage; e.g., Harri et al., 1998) most suitable for very fearful populations where other tests would yield ceiling effects. As an additional nuance, our data showed the glove test to be more fear-inducing on the first test, and likely even more so when conducted by an unfamiliar tester; the sensitivity of the glove test can thus be modified according to how exactly it is applied. In Experiment 3, for instance, classifying mink according to sustained reactions instead of just the first test would have left us with only three fearful individuals (although these were undoubtedly very reliably fearful). Both which test and what method of application is ideal will therefore depend on the population characteristics, and also on the test's aims. For example, the glove test might be useful in genetically selecting for very calm animals at the population level, and where fear levels are already fairly low; in this case, maximum sensitivity and speed are desirable, and farmers might only conduct the test once. For research, however, it is more crucial that the classification accurately represents underlying temperament, and so multiple tests may be preferable in order to reduce noise by eliminating individuals whose responses change unpredictably (see Section 2.1.3).

Glove tests also proved to have good construct (convergent) validity, and practical usefulness. Construct validity was good when assessed formally (Experiment 1), and the observed differences between colour-types provided further, albeit less formal, support for the test's validity. The glove test's validity was also supported by its usefulness in predicting (or detecting in advance) a biologically relevant variable, barrenness. This observed relationship between barrenness and greater fearfulness could be mediated in several different ways, which can now be empirically investigated. Fearfulness could reduce reproductive success through the activation of the HPA axis, which adversely affects reproductive hormones (Sapolsky et al., 2000; von Borell et al., 2007). Fearful individuals may also copulate less frequently, due to decreased receptivity to males or because the difficulty of handling these individuals results in farmers making fewer attempts to breed them. In the current study, mink were tested for fearfulness after they should have been pregnant; it is thus also possible that any that had not successfully conceived were in a different hormonal state from pregnant females, with this then affecting responses to the test. Until this possibility has been investigated, it is uncertain whether fearfulness is truly predictive of reproductive success. As a final testable hypothesis, those that had many unsuccessful mating attempts might have been handled more often, and so been more likely to develop fear of humans with handling gloves. The test's decision validity (ability to predict performance, i.e., its usefulness for making management decisions) remains to be evaluated, for example by determining whether it can predict barrenness when assessed before the mating season, or be used in genetic selection for temperament.

Most mink did not show identical responses each day when tested daily. Nevertheless, this is not grounds for discarding the glove test in favour of other tests. First, if conventional reliability tests for the other widely-used tests for mink temperament have ever been conducted, they have not been published in accessible literature, so their reliabilities are essentially unknown. Second, most changes in reaction seemed to be due to normal processes of habituation. Third, reliability appeared to increase over repeated tests, as if responses after the first or second test are more stable. Fourth, there seems scope for improving reliability through training the tester, in that a possible explanation for lower reliabilities at Farms 2 and 3 in Experiment 3 was that the researcher here was less experienced than that working at Farm 1. Fifth, reliability was greatly improved by excluding mink that were awake but in the nest box at testing, and might be improved yet further still by additional refinements in categorization as suggested below.

Our data highlighted "unresponsiveness" (sometimes called "undecided") as a category needing further research. We suspect that this category encompasses mink that are genuinely uninterested in the stimulus; mink that are scared and so 'freezing'; and mink that are only half-awake. Refining future methods might therefore involve better controls for each mink's 'drowsiness' at the time of the test, for example by excluding only mink with eyes closed or head tucked prior to the test if this is visible in the nest box. Our method of identifying 'freezing' in mink that are standing as far as possible from the tester by their immobility and sustained attention to the stimulus (see 2.1.2) could also be extended to mink in any location. Future research could assess sympathetic responses (e.g., via telemetry devices) to determine whether mink suspected to be 'freezing' are indeed frightened by the stimulus. "Aggression" is another category whose motivational basis and welfare significance remain unclear. From our data, it is evident that aggression levels depend on the test stimulus. The lower aggression in the glove test than the stick test does not seem to be due to greater fear of the glove inhibiting aggression, since most mink that were aggressive only towards the stick were curious, not fearful, in the glove test. This is consistent with our suggestion in the Introduction that "aggression" towards the stick may reflect something other than a response to the human tester, such as predation or exploration directed at the small object. This hypothesis could be tested by comparing test results with their responses to being handled and with their feeding motivation, as well as testing whether aggression increases if they are always allowed to carry the stick away at the end of the test. Similarly, the increased aggression over repeated tests in Experiments 1 and 3 cannot be explained by initial fear: mink that became aggressive had typically begun as curious. Some sensitization may have occurred with repeated disturbances by a human, a hypothesis that could now be tested by comparing results of a single test between animals that had been frequently disturbed by humans in other ways over the preceding days.

Thus in the glove test, as in other temperament tests, further research could refine how mink are classified, and also yield a better fundamental understanding of what some responses mean. However, even without these additional data, when used carefully by trained researchers, the glove test is a useful, valid tool for assessing mink fearfulness.

6. Conclusions

The glove test is a promising and valid test of temperament, including fearfulness, for farmed mink, especially useful for populations where very few mink exhibit fear in the stick test. Some further refinement is needed to improve its repeatability and thereby reduce data loss from individuals that are inconsistent across tests. Using the current method, however, we were able to detect expected temperament differences between colourtypes, and a negative relationship between fearfulness and reproductive success. This suggests that reducing fearfulness through selective breeding or changes in management may improve mink productivity as well as welfare.

Acknowledgements

Many thanks are due to the owners of these farms, Kirk Rankin, Ted Parkinson and Lyn Parkinson, for their cooperation. We would also like to thank Mike Walker and anonymous reviewers for their comments on earlier drafts of the manuscript, and NSERC and the Canada Mink Breeders Association for funding.

References

Axelsson, H.M.K., Aldén, E., Lidfors, L., 2009. Behaviour in female mink housed in enriched standard cages during winter. Appl. Anim. Behav. Sci. 121, 222–229.

Baekgaard, H., Hansen, M.U., Sønderup, M., 2007. The influence of body condition on breeding results and early kit mortality. In: NJF Seminar No. 403 , Kolding, Denmark, 13–15 August.

Blanchard, R.J., Blanchard, D.C., 1969. Crouching as an index of fear. J. Comp. Physiol. Psychol. 67, 370–375.

Cochran, W.G., Cox, G.M., 1957. Experimental Designs. John Wiley and Sons, Inc, New York, p. 91.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52, 281–302.

Dohoo, I., Martin, W., Stryhn, H., 2003. Veterinary Epidemiologic Research. AVC Inc., Charlottetown.

Domjan, M., 2003. The Principles of Learning and Behavior. Wadsworth, Belmont, USA, pp. 39–41.

Dunstone, N., 1993. The Mink. T. and A.D. Poyser Ltd., London.

Dunstone, N., Sinclair, W., 1978. Comparative aerial and underwater visual acuity of the mink, *Mustela vison* schreber, as a function of discrimination distance and stimulus luminance. Anim. Behav. 26, 6–13.

European Commission, Scientific Committee on Animal Welfare. 2001. The welfare of animals kept for fur production. http://ec.europa.eu/food/fs/sc/scah/out67_en.pdf.

Forkman, B., Boissy, A., Meunier-Salauen, M.C., Canali, E., Jones, R.B., 2007. A critical review of fear tests used on cattle, pigs, sheep, poultry and horses. Physiol. Behav. 92, 340–374.

Gray, J.A., 1987. The Psychology of Fear and Stress. Cambridge University Press, Cambridge.

Hansen, S.W., 1996. Selection for behavioural traits in farm mink. Appl. Anim. Behav. Sci. 49, 137–148.

Hansen, S.W., 1997. Selection for tame and fearful behaviour in mink and the effect on the HPA axis. In: Proceedings of the 31st International Congress of the ISAE , Prague, Czech Republic, p. 72.

Hansen, S.W., Damgaard, B.M., 1991. Effect of environmental stress and immobilization on stress physiological variables in farmed mink. Behav. Process. 25, 191–204.

Hansen, S.W., Møller, S.H., 2001. The application of a temperament test to on-farm selection of mink. Acta Agric. Scand., Sect. A 30, 93–98.

Hansen, S.W., Møller, S.H., 2008. Diurnal activity patterns of farm mink (*Mustela vison*) subjected to different feeding routines. Appl. Anim. Behav. Sci. 111, 146–157.

Harri, M., Mononen, J., Rekilä, T., Korhonen, H., Niemelä, P., 1998. Effects of top nest box on growth, fur quality and behaviour of blue foxes (*Alopex lagopus*) during their growing season. Acta Agric. Scand., Sect. A 48, 184–191.

Hemsworth, P.H., 2003. Human–animal interactions in livestock production. Appl. Anim. Behav. Sci. 81, 185–198.

Hemsworth, P.H., Barnett, J.L., Coleman, G.J., Hansen, C., 1989. A study of the relationships between the attitudinal and behavioral profiles of stockpersons and the level of fear of humans and reproductive-performance of commercial pigs. Appl. Anim. Behav. Sci. 23, 301–314.

Jones, R.B., 1997. Fear and distress. In: Appleby, M.C., Hughes, B.O. (Eds.), Animal Welfare. CAB International, New York, pp. 75–88.

Kenttämies, H., Nordrum, N.V., Brenøe, U.T., Smeds, K., Johannessen, K.R., Bakken, M., 2002. Selection for more confident foxes in Finland and Norway: heritability and selection response for confident behaviour in blue foxes (*Alopex lagopus*). Appl. Anim. Behav. Sci. 78, 67–82.

Kirkden, R.D., Rochlitz, I., Broom, D.M., Pearce, G.P. (2010). Assessment of on-farm methods to measure confidence in mink and foxes on Norwegian farms. Report prepared for Dyrevernalliansen (Norwegian Animal Protection Alliance), Oslo, Norway. Cambridge University Animal Welfare Information Centre, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK, 43 pp.

Korhonen, H., Hansen, S.W., Malmkvist, J., Houbak, B., 2000. Effect of capture, immobilization and handling on rectal temperatures of confident and fearful male mink. J. Anim. Breed. Genet. 117, 337–345.

Korhonen, H.T., Jauhiainen, L., Rekila, T., 2002. Effect of temperament and behavioural reactions to the presence of a human during the pre-mating period on reproductive performance in farmed mink (*Mustela vison*). Can. J. Anim. Sci. 82, 275–282.

Malmkvist, J., 2001. Fear in farm mink (*Mustela vison*) – consequences of behavioural selection. Ph.D. thesis, University of Copenhagen.

Malmkvist, J., Hansen, S.W., 2001. The welfare of farmed mink (*Mustela vison*) in relation to behavioural selection: a review. Anim. Welf. 10, 41–52.

Malmkvist, J., Hansen, S.W., 2002. Generalization of fear in farm mink, *Mustela vison*, genetically selected for behaviour towards humans. Anim. Behav. 64, 487–501.

Malmkvist, J., Palme, R., 2008. Periparturient nest building: Implications for parturition, kit survival, maternal stress and behaviour in farmed mink (*Mustela vison*). Appl. Anim. Behav. Sci. 114, 270–283.

Malmkvist, J., Houbak, B., Hansen, S.W., 1997. Mating time and litter size in farm mink selected for confident or timid behaviour. Anim. Sci. 65, 521–525.

Maxwell, S.E., Delaney, H.D., 2004. Designing Experiments and Analyzing Data: A Model Comparison Perspective, second Ed. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 141–142.

Meagher, R.K., 2009. Observer ratings: validity and value as a tool for animal welfare research. Appl. Anim. Behav. Sci. 119, 1–14.

Miller, N.E., 1944. Experimental studies of conflict. In: Hunt, J.M. (Ed.), Personality and the Behavior Disorders, vol. 1. The Ronald Press Company, New York, pp. 431–465.

Møller, S.H., Hansen, S.W., 2001. Assessment of mink welfare at farm level. In: NJF Seminar No. 331 , Snekkersten, Denmark, October 1–3.

Nimon, A.J., Broom, D.M., 1999. The welfare of farmed mink (*Mustela vison*) in relation to housing and management: a review. Anim. Welf. 8, 205–228.

Sapolsky, R.M., Romero, L.M., Munck, A.U., 2000. How do glucocorticoids influence stress responses? Integrating permissive, suppressive, stimulatory, and preparative actions. Endocr. Rev. 21, 55–89.

Shabalina, A.T., 1984. Dominance rank, fear scores and reproduction in cockerels. Brit. Poultry Sci. 25, 297–301.

Trapezov, O.V., 2000. Behavioural polymorphism in defensive behaviour towards man in farm raised mink (*Mustela vison* Schreber, 1777). Scientifur 24, 103–109.

von Borell, E., Dobson, H., Prunier, A., 2007. Stress, behaviour and reproductive performance in female cattle and pigs. Horm. Behav. 52, 130–138.

Welch, B.L., 1951. On the comparison of several mean values: an alternative approach. Biometrika 38, 330–336.

Wielebnowski, N.C., 1999. Behavioral differences as predictors of breeding status in captive cheetahs. Zoo Biol. 18, 335–349.

Wingfield, J.C., Sapolsky, R.M., 2003. Reproduction and resistance to stress: when and how. J. Neuroendocrinol. 15, 711–724.

Zalaquett, C., Thiessen, D., 1991. The effects of odors from stressed mice on conspecific behavior. Physiol. Behav. 50, 221–227.