ANIMAL WELL-BEING AND BEHAVIOR

On-farm welfare assessment of commercial Pekin ducks: a comparison of methods

Essam Abdelfattah,[*,†] Giuseppe Vezzoli,[*,‡] and Maja M. Makagon[*,1]

*Center for Animal Welfare, Department of Animal Science, University of California Davis, Davis, CA, USA; †Department of Animal Hygiene, Behavior and Management, Faculty of Veterinary Medicine, Benha University, Benha, Egypt; and ‡School of Mathematics and Science, College of the Desert, Palm Desert, CA, USA

ABSTRACT Although a number of welfare assessment methods have been developed for poultry, none have been evaluated for use in commercial duck farms. The primary objective of the study was to evaluate the inter-rater reliability and relative accuracy of 4 duck welfare assessment strategies. Over 2 experiments, 12 flocks of commercial meat ducks (5,850 to 6,300 ducks/flock) aged 30 to 34 D were evaluated. During experiment 1, six flocks were evaluated using 2 welfare assessment methods: transect walks (**TW**) and catch-and-inspect (**CAI**). During TW, 2 observers walked predetermined transects along the length of the house and recorded the number of ducks per transect that were featherless, were dirty, were lethargic, had bloody feathers, had infected eyes, and/or had plugged nostrils or were found dead. During CAI, a total of 150 ducks per flock were corralled and individually evaluated. The same welfare indicators were assessed using both methods. During experiment 2, six flocks were initially evaluated using CAI, TW, and a distance evaluation (**DE**; a total of 50 ducks per flock evaluated from a walking distance) and then reassessed within 24 h during the loadout (**LO**) process. Data were analyzed in SAS (version 9.4) to determine the observer and method effects on the incidence of welfare indicators. Interobserver reliability was high ($P > 0.05$) across methods for most welfare indicators. The assessment method affected the measured outcome variables in both experiments ($P < 0.05$). CAI resulted in higher estimated incidences of most welfare indicators than TW (experiment 1 and 2) and LO (experiment 2). DE yielded intermediate results compared with other methods (experiment 2). Results obtained using TW and LO were most similar, the only difference being the number of dead birds observed using each method ($P < 0.0001$). The average time required for CAI, TW, DE, and LO was 2.40 ± 0.004, 1.12 ± 0.02, 1.54 ± 0.001, 3.56 ± 0.006 h, respectively. Bootstrapping analyses showed that the observed welfare indicator prevalence estimates were affected by the number of transects (TW) and number of birds (CAI) sampled.

Key words: welfare assessment, individual sampling, Pekin duck, transect walks, loadout

2020 Poultry Science 99:689–697
https://doi.org/10.1016/j.psj.2019.10.006

INTRODUCTION

Animal welfare assessments are routinely conducted as a part of duck farm audits and in research. Such assessments typically include animal-based measures, which are thought to provide a more accurate representation of the animal's actual state of welfare (Blokhuis et al., 2003). Duck welfare assessments have been conducted by walking ducks into catch pens and inspecting individuals from among the penned sample group (Karcher et al., 2013; Fraley et al., 2013; Colton and Fraley, 2014), and by visual inspection of ducks from a distance (Jones and Dawkins, 2010). Although the methods resemble those traditionally used to assess the welfare of other poultry species (Bright et al., 2006; Welfare Quality, 2009; Marchewka et al., 2013, 2015), the type of information they provide and the reliability with which they can be used by multiple scorers have not been evaluated in duck production settings.

Practical animal welfare assessment strategies should be easy to apply by diverse assessors and reflect the actual state of welfare of individuals within the sampled population (Butterworth et al., 2011). Inspection of individuals selected from among a sample captured throughout the barn (catch-and-inspect [**CAI**]) is among the most popular methods for evaluating the welfare of poultry. The CAI assessment strategy has, for example,

Received June 12, 2019.
Accepted October 4, 2019.

[1]Corresponding author: mmakagon@ucdavis.edu

been incorporated into the Welfare Quality protocols for poultry (Welfare Quality, 2009). This method of assessment allows for a close inspection of each individual bird. However, because it involves catching and handling of the birds, it is likely stressful for the birds. Furthermore, it is time-consuming, which limits the evaluation to a subset of the flock. Bright and colleagues (2006) reported that the sample size of 100 birds per flock, which is typically used in poultry assessments, was insufficient when the welfare indicator assessed varied in prevalence across flocks. It is, therefore, possible to acquire biased results when applying this method to a poultry welfare assessment protocol.

The transect walk (**TW**) assessment strategy allows for the inspection of an entire flock. The TW method expands upon and formalizes the type of information that can be obtained by poultry farmers as they walk through the barn and visually inspect their flocks during daily checks. Rather than inspecting a subsample of birds, assessors walk the length of the barn back and forth along predetermined paths and record all incidences of birds that show welfare impairments. The TW method has been identified as a viable strategy for evaluating from a distance the welfare of broilers (Marchewka et al., 2013; BenSassi et al., 2019a,b) and turkeys (Marchewka et al., 2015, 2019; Ferrante et al., 2018) and has been incorporated into the AWIN Welfare Assessment Protocol for Turkeys (AWIN, 2015).

Distance evaluation (**DE**) has been proposed as another alternative to CAI. The method does not require birds to be handled, reducing the amount of stress they may experience and increasing the number of birds that can be sampled as compared with CAI. DE has been formally evaluated as a method for assessing feather coverage in chickens yielding promising results (Bright et al., 2006). A disadvantage for DE is that the method still requires subsampling of birds. Marchewka et al. (2015) used a modified version of DE to evaluate turkeys as they were loaded out from the barn onto trucks ahead of slaughter (loadout [**LO**]). The scorers stood stationary and evaluated small groups of turkeys as they were herded toward the loading ramp. Although impractical as an audit tool do to its time requirement, LO allows for the evaluation of each bird in the flock as it walks onto the loading dock, while ensuring that birds are not double counted.

Given the lack of information about how the impact of welfare assessment strategy may impact commercial duck welfare assessment outcomes, the objectives of this study were to identify information trade-offs associated with 4 duck welfare assessment strategies: TW, CAI, DE, and LO. Specifically, over the course of 2 experiments, we compared the relative accuracy of the 4 welfare assessment methods when applied to commercial duck flocks and the inter-rater reliability associated with each method. Experiment 1 evaluated CAI and TW, as these methods have been incorporated into welfare assessment protocols (Welfare Quality, 2009; AWIN, 2015). Results from experiment 1 were used to refine data collection during experiment 2, which additionally included an evaluation of DE and LO sampling methods. We focused the study on ducks that were near or at processing age as we assumed that welfare issues would be most severe and most visible at this age. In addition, testing birds just before they were moved to the processing plant allowed us to compare the outcomes of the 3 on-farm assessment strategies to welfare assessments conducted when birds were loaded out of the barn and onto trucks (LO).

MATERIALS AND METHODS

Housing and Ducks

Data were collected on commercial duck farms located in Indiana, USA. All farms were contracted by a single duck company (Maple Leaf Farms Inc., Milford, IN) and were therefore managed using similar practices and husbandry protocols. All duck barns had plastic slatted floors and were equipped with nipple watering systems. *Ad libitum* feed was provided in round poultry feeders. Flock sizes ranged from 5,850 to 6,300 ducks of a single commercial strain, with a target space allocation of 0.16 m^2 per duck. Experimental procedures were approved by the Institutional Animal Care and Use Committee at the University of California, Davis (Protocol No. 20198).

Data Collection

Experiment 1 A total of 6 flocks of commercial Pekin ducks, 2 from each of 3 farms, were evaluated over the course of 1 wk in September 2016. The ducks were assessed at 30 D of age, approximately 5 D before they were transported off the farm for processing. During data collection, ducks were scored simultaneously but independently (without discussion) by 2 trained observers using the welfare indicator category definitions provided in Table 1. Briefly, feather quality, feather cleanliness, eye condition, nostril condition, gait score (TW only), and footpad quality (CAI only) were assessed on a 3-point scale: 0 (best condition), 1 (moderate condition), and 2 (worst condition). The presence of blood on feathers and lethargic appearance were marked as present or absent. We collected data on each flock using CAI and TW in random order. During CAI, 150 ducks per flock from across 5 locations distributed through the width and length of the barn were evaluated. At each location, a group of approximately 30 ducks was corralled into a mobile plastic pen. Individual birds were then picked up by farm staff and evaluated by each of the 2 observers. TW was based on the methodology previously described for broilers (Marchewka et al., 2013) and turkeys (Marchewka et al., 2015). Seven longitudinal transects were identified per barn, delineated by the location of feeders and drinker lines (Figure 1). Their widths varied from 2.74 m to 1.28 m. Two observers slowly walked the length of each transect and recorded the number of birds within the width of the

Table 1. Description of welfare indicators used for welfare assessment of commercial Pekin ducks. Gait was measured during TW only during experiment 1 and TW, DE, and LO during experiment 2. Footpads were assessed only during CAI.

Indicator	Score	Description
Feather quality[1]	1	Damaged (worn or deformed) feathers, or one or more featherless areas < 5 cm in diameter at the largest extent
	2	At least one featherless area ≥5 cm in diameter at the largest extent
Feather cleanliness[1]	1	Staining on down or feathers < 5 cm in diameter, includes discoloration due to adhering dirt or manure
	2	Staining on down or feathers ≥ 5 cm in diameter, includes discoloration due to adhering dirt or manure
Blood on feather	-	Fresh or old blood visible on the back and/or wings
Eye	1	Dirt or staining around the eye area, or presence of wet eye ring
	2	Inflamed eyelids, conjunctivitis, eyes sealed shut, or evidence if blindness
Nostril	1	Air passageways blocked with dust or mucus from inside the nostril cavity
	2	Air passageways blocked from outside (or inside and outside), the nostril opening is plugged
Lethargic	-	Bird showing general signs of impaired health: head is pulled into the body, bird is huddling with disarranged feathers. These birds are usually found in a resting position
Dead	-	Dead
Gait (TW)	1	Duck walks with slight limp, or walking is labored due to crossed feet or awkward strut (ex. visible limping and stiffing of legs)
	2	Duck is reluctant to walk, will only walk short distances when encouraged, typically due to obvious leg problems (e.g., swelling of joints, obvious injury)
Footpad (CAI)	1	Calluses or lesions cover < 50% of the pad area and are free of blood
	2	Callouses or lesions cover ≥50% or more of pads, and/or presence of bloody lesions

Abbreviations: CAI, catch-and-inspect; DE, distance evaluation; LO, loadout; TW, transect walks.

[1]During Experiment 2 feather quality and cleanliness were score separately for the wings (at least 1 affected), back (as distinct from wings, neck and rump), and flanks.

transect and within 1 m of the observer that were showing any of the predefined welfare issues (Table 1). Only the incidence of birds observed having blood on feathers or a lethargic appearance or scoring 1 or 2 in any of the other indicator categories was recorded on farm. The proportion of birds with scores of 0 in any of the indicator categories was calculated using this information. TW were conducted in semirandom order in that we avoided walking sequentially through adjoining transects. A timer was used to determine the amount of time it took for the completion of each assessment.

Data collection was preceded by a training day during which the 2 observers involved in the study practiced assessing ducks until reaching >95% inter-rater agreement for each assessment method. Neither of the observers involved in experiment 1 (E.A. and G.V.) had previous experience assessing duck welfare, although one (G.V.) had previously used CAI to evaluate chicken flocks and TW to evaluate turkey flocks (Marchewka et al., 2015).

Figure 1. Each house was divided longitudinally into 7 transects from wall to wall. Two observers walked the length of each transect in a semirandom (evaluation of adjacent transects was avoided) order that was determined independently for each flock and recorded the number of birds showing any of the predefined welfare problems. During the walk, the observers score the indicators in a zone covered by a semicircular area, 1 m in front of the observer. The location of drinkers and feeders marked the edges of the transects, whereas the red line delineates the path traveled by the observers. A sample transect sampling order is provided in red numbers.

The flocks evaluated during the training session were not included in the analysis.

Experiment 2 A second experiment was conducted over the course of 2 wk in October 2017. Once again, 2 flocks were evaluated per each of 3 farms, this time at 34 D of age (±1 D). Farms that participated in experiment 2 were different from those visited during experiment 1. Two observers (E.A. and one new observer) simultaneously but independently evaluated each flock using 4 welfare assessment strategies. As in experiment 1, data collection was preceded by a training day during which the 2 observers practiced assessing ducks until reaching >95% inter-rater agreement for each assessment method. The first 3, CAI, TW, and a visual evaluation of small groups of ducks (distance evaluation [**DE**]), were conducted in random order on the same day. Each flock was evaluated a fourth time when the flock was moved out of the barn and walked up a ramp onto transportation trucks (LO), which occurred within 24 h of the initial on-farm assessment. Ducks were scored using welfare indicator definitions provided in Table 1. The welfare categories used were slightly modified from those utilized during experiment 1. Specifically, as minor eye staining and obstruction of the nostrils within the nasal cavity (score 1) required closer inspection, only occurrences of nostril and eye condition score 2 were recorded during TW and LO. Gait was scored during TW, DE, and LO. In addition, for all methods, the scoring of feather quality and condition was subdivided in accordance with body area (wing, back, and underwing region). We postulated that injury and staining to these distinct body parts may have different causes and were therefore interested in determining whether assessment method affected the inter-rater reliabilities. However, whether a spot of the feathers was caused by a stain (ex. manure) or actual damage proved difficult to determine from a distance (TW and LO). Similarly, the

5-cm stain diameter cutoff was difficult to adhere to using at TW and LO. Feather quality and feather cleanliness scores were therefore combined ahead of analysis into a single score, indicating the presence or absence of stained or damaged feathers of a visible size.

CAI and TW were conducted as described for experiment 1. During DE, the 2 observers slowly walked along the house stopping at 3 randomly selected locations to visually evaluate groups of 13–20 birds. A total of 50 ducks per flock were assessed using this method. During LO, groups of approximately 80 to 100 ducks were herded into a corridor that led to the loading ramp. Observers stood behind a partition on the loading dock, positioned so that they could evaluate the ducks from multiple angles as walked passed through a door into the loading area and turned away from the observers and onto the loading ramp. The ducks moved in a single direction, which ensured that individual birds were not counted twice. Once the LO process was completed, the observers walked through the barn to evaluate any remaining birds (i.e., those deemed unfit for transport). The loadouts occurred between 10:00 pm and 8:00 am. The observers wore head lamps and were further assisted by the truck's back lights. As in experiment 1, a timer was used to determine the amount of time it took for the completion of each assessment.

Statistical Analysis

Before analysis, the number of ducks exhibiting each welfare parameter was converted into a flock percentage by dividing the total number of observed incidences within each welfare category (Table 1) by the number of ducks assessed (150 for CAI, 50 for DE, and the total number of ducks in the flock at the time of assessment for TW and LO).

Data collected during experiment 1 were analyzed using SAS, version 9.4, for Windows (SAS Institute Inc., Cary, NC). Data were checked for normality using the PROC UNIVARIATE procedure. If a normal distribution could not be achieved, the nonparametric Kruskall–Wallis test was used, and pairwise comparisons between methods were performed using the nonparametric Wilcoxon test. Otherwise, data were analyzed using the PROC MIXED model with the flock as the experimental unit. The model included observer, flock, and their interaction as fixed factors and farm as a random statement. For indicators characterized by high inter-rater reliability, an average of the 2 observers' scores was used for subsequent analyses. Otherwise, scores obtained by E.A. were used. To test the method effect, the model included assessment method as a fixed factor, farm as a random factor, and the flock as a repeated measure. The pairwise comparisons between methods were tested at a total significance level of 0.05 using the Tukey–Kramer adjustment for multiple comparisons. During CAI and TW assessments, eye condition S2 and the number of lethargic birds were not frequently seen, and the data were subjected to Chi-square analysis.

For method comparisons in experiment 2, an independent mixed-model repeated-measures ANOVA was performed for each welfare indicator using the PROC MIXED model in SAS, version 9.4. The model included the method of assessment (CAI, TW, DE, and LO), observer, and their interactions fixed factors. The farm was included as a random effect, and the flock was specified a repeated effect. Least squares mean differences were adjusted for multiple comparisons using the post hoc Tukey test. Least squares means and SEM were reported for each welfare indicator, and the level of statistical significance was reported at a P-value of ≤ 0.05. Data that were not normally distributed after transformation were analyzed using nonparametric tests (Mann–Whitney–Wilcoxon test and Kruskal–Wallis test).

The bootstrapping analysis was applied to data from experiment 1 and 2 (12 flocks total), separately for TW and CAI. The bootstrapping analysis involved taking simulated random samplings combinations from the original data set using Monte Carlo methods. Only indicators that were evaluated the same way in both experiments were subjected to the analysis. For TW, this included eye condition score 2, nostril condition score 2, bloody feather, gait score 2, and the number of lethargic birds. For CAI, the analyzed indicators were bloody feathers, eye condition score 1 and 2, nostril condition score 1 and score 2, the number of lethargic birds, and footpad scores 1 and 2. For TW data, expected mean and SE of the data set for each welfare indicator was calculated by taking random samples of 2 transect (20% of the information) or combinations of 3, 4, 5, and 7 transects (40, 60, 80, and 100% of information, respectively), while for CAI data, expected mean and SE of the data set for each welfare indicator was calculated by taking random samples of individual evaluation of 25, 50, 100, or 150 birds. Simulations were run 10,000 times per flock per welfare indicator. The obtained mean values for 20, 40, 60, 80, and 100% of information for all variables during TW and the obtained mean values for individual evaluation of 25, 50, 100, and 150 birds for all variables after 10,000 simulations were box plotted. All bootstrapping data analysis was conducted using R software (version 3.5.1; R Foundation, Vienna, Austria). We compared the difference between sample sizes and transect numbers by computing the 95% CI of their pairwise differences using the mcmc package in R. Two sample sizes were considered to be different if the corresponding 95% CI on their pairwise differences did not include the null value zero.

RESULTS

Experiment 1

Observers differed only in their prevalence estimates for feather cleanliness score 2 ($P = 0.01$) when assessed using CAI ($P = 0.01$) and TW ($P = 0.01$). Flock affected the prevalence of most welfare indicators (all $P < 0.05$), except for nostril condition score 2 ($P = 0.66$) and gait

score 2 ($P = 0.20$) when assessed using TW and feather quality score 2 ($P = 0.46$) when assessed by CAI. An interaction between flock and observer was reported for the incidence of feather cleanliness scores 1 ($P = 0.02$) and 2 ($P = 0.02$) during CAI. The house and observer interaction had no effect on any of the measured variables during TW method (all $P > 0.05$).

The assessment method affected the outcomes, as presented in Table 2. Briefly, CAI resulted in higher estimates for the incidence of featherless (score 1 and 2), feather cleanliness (score 1 and 2), bloody feathers, eye condition score 1, and nostril condition score 1 than TW (all $P < 0.0001$). The assessment method did not affect the estimated flock prevalence of eye condition score 2, nostril condition score 2, and the number of lethargic birds (all $P > 0.05$).

Experiment 2

Interobserver agreement was high, regardless of the assessment method used. The sole discrepancy was the estimate of the incidence of bloody feathers ($P = 0.01$), which differed between observers when assessed using CAI. As in experiment 1, flock impacted most of the welfare indicator estimates (all $P < 0.05$), except for back quality scores 1 and 2, back cleanliness score 1, underwing quality score 1, and underwing cleanliness scores 1 and 2 when measured using CAI (all $P > 0.05$), and incidence of lethargic birds when measured by TW ($P = 0.27$). An interaction between observer and flock was noted for underwing cleanliness score 2 ($P = 0.02$), bloody feathers ($P = 0.005$), and footpad score 1 ($P = 0.01$) and score 2 ($P < 0.001$) using CAI and for back ($P = 0.02$) and underwing ($P = 0.005$) feather condition using TW.

The assessment method affected most measured variables (all $P \leq 0.05$, Table 3), except eye score 2 ($P = 0.405$), gait score 2 ($P = 0.141$), and the number of lethargic birds ($P = 0.422$). Where differences were noted, CAI yielded higher incidence estimates than TW and LO. Estimates provided by DE were intermediate. We did not find any differences in data collected using TW and LO ($P \geq 0.05$), except for the number of dead birds.

Time Requirement

Across the flocks sampled during experiment 1 and 2 (12 flocks total), the time required for penning and individual evaluation (CAI) of 150 ducks per flock sampled in 5 areas of the house averaged 2.42 ± 0.004 h, while the time required for TW assessment of a flock (all 7 transects) averaged 1.15 ± 0.002 h. The average time required for DE and LO, which were used to evaluate the 6 flocks enrolled in experiment 2, was 1.54 ± 0.001 and 3.56 ± 0.006 h, respectively.

Bootstrapping Analysis

For CAI, the resulting expected mean for all welfare indicators subjected to bootstrapping analysis was similar to the observed mean value by using a sample size of 25 ducks. However, SE was different among sample sizes used. Increasing sample size from 25 birds to 100 and 150 birds decreased the SE (representative example in Table 4 and Figures 2A and 2B). To determine the minimum necessary number of transects required to obtain a reliable estimation on the welfare status of the duck flock, bootstrapping technique was applied to TW data. It was observed that the obtained mean for each welfare indicator after bootstrapping was similar to the observed mean value by using as little as 2 transects (20% of information). But the value of the SEM stabilized when data from a minimum 4 transects per flock were included (representative example in Table 5 and in Figures 3A and 3B).

DISCUSSION

Animal-based measures are incorporated into many modern animal welfare assessment schemes. By way of 2 experiments, we evaluated the inter-rater reliabilities and relative accuracies of animal-based measures

Table 2. Results of experiment 1; the percentages (LSM ± SEM) of ducks per flock for each welfare indicator as expressed for each assessment method.

Welfare indicators (%)	Catch-and-inspect method	Transect walks method	*P*-values
Feather quality score 1[1]	6.05 ± 0.97[a]	0.10 ± 0.97[b]	<0.0001
Feather quality score 2[1]	9.05 ± 0.47[a]	1.03 ± 0.47[b]	<0.0001
Feather cleanliness score 1[1]	8.500 ± 1.71[a]	0.30 ± 1.71[b]	<0.0001
Feather cleanliness score 2[1]	8.99 ± 1.46[a]	0.70 ± 1.46[b]	0.0250
Blood on feather[1]	38.05 ± 3.30[a]	2.01 ± 3.30[b]	<0.0001
Eye score 1[1]	17.38 ± 3.15[a]	0.44 ± 3.15[b]	<0.0001
Eye score 2[1]	0.11 ± 0.04	0.01 ± 0.04	0.337
Nostril score 1[1]	45.72 ± 3.73[a]	0.51 ± 3.73[b]	0.0039
Nostril score 2[1]	5.11 ± 2.39	0.11 ± 2.39	0.07
Lethargic[1]	0.61 ± 0.38	0.01 ± 0.38	0.247
Gait score 1	N/A	0.45 ± 0.06	-
Gait score 2	N/A	0.18 ± 0.04	-
Footpad score 1	38.11 ± 6.47	N/A	-
Footpad score 2	5.11 ± 1.94	N/A	-

[a,b]Within the same row signifies statistical difference ($P > 0.05$).
[1]Transformed using logarithmic transformation ahead of analysis.

Table 3. Results of experiment 2; the percentages (LSM ± SEM[1]) of ducks per flock for each welfare indicator as expressed for each assessment method.

Welfare indicators (%)	Catch-and-inspect	Transect walks	Distance evaluation	Loadout	*P*-value
Wing condition[2]	22.02 ± 1.50[a]	4.38 ± 1.50[c]	7.53 ± 1.48[b]	2.90 ± 1.31[c]	<0.0001
Back condition	7.76 ± 1.14[a]	1.85 ± 1.14[b]	1.97 ± 1.13[b]	1.05 ± 1.00[b]	0.0002
Underwing condition	17.06 ± 2.81[a]	2.37 ± 2.81[b]	14.39 ± 2.78[a]	4.07 ± 2.46[b]	0.0042
Blood on feather	24.30 ± 2.82[a]	4.35 ± 2.82[c]	14.19 ± 2.79[b]	4.79 ± 2.47[c]	<0.0001
Eye score 2	0.00 ± 0.03	0.07 ± 0.03	0.02 ± 0.03	0.01 ± 0.03	0.4047
Nostril score 2[2]	4.90 ± 1.19[a]	0.84 ± 1.19[b,c]	2.68 ± 1.18[a,b]	0.53 ± 1.05[c]	0.0009
Gait score 2	N/A	0.86 ± 0.21	0.44 ± 0.21	0.42 ± 0.20	0.1412
Lethargic	3.20 ± 1.63	0.04 ± 1.63	0.00 ± 1.61	0.00 ± 1.61	0.4224
Dead	N/A	0.05 ± 0.09[b]	0.00 ± 0.07[b]	0.66 ± 0.08[a]	<0.0001

[a,b,c]Within the same row signifies statistical difference ($P > 0.05$).

[1]Least squares means are expressed with ± standard errors obtained from original data; reported differences are based on normalized data after log transformation.

[2]log transformed using log10 function.

obtained using a total of 4 welfare assessment strategies carried out on commercial duck farms. For 2 of the strategies (CAI and TW), we furthermore evaluated how samples size affects obtained results.

All 4 of the examined methods were associated with high inter-rater reliabilities. This is a promising finding, particularly because the observers involved had no experience with ducks at the start of the study and received minimal training. High overall inter-rater reliability among observers with limited training and experience has previously been reported for CAI and TW methods when applied to boilers (Marchewka et al., 2013) and turkeys (Marchewka et al., 2015), whereas good inter-rater reliability for a version of DE (assessment from 2 m away) was reported by Bright et al. (2006) who used the method to evaluate prevalence of feather loss in laying hens. In the present study, observers differed mainly in their estimation of feather-related metrics (feather cleanliness score 2 during experiment 1 and blood on feathers during experiment 2) when evaluated using CAI. Estimates of the prevalence of feather cleanliness score 2 also differed for estimates obtained using TW during experiment 1. Pooling feather metrics into a single score, which was carried out during experiment 2, alleviated the observer effect. It is likely that that the observers had a difficult time estimating the 5-cm diameter cutoff, which was used to distinguish feather scores of 1 vs. 2. A similar observation was made by Marchewka et al. (2019) and BenSassi et al. (2019a) who used TW to evaluate turkey and broiler flocks, respectively. Marchewka et al. (2019) reported that observers differed in their estimations of the prevalence of back and tail wounds within turkey flocks. However, the observer effect disappeared when the 2 indicators were pooled. Similarly, observers found it difficult to differentiate among immobile and lame broilers as reported by BenSassi et al. (2019a), although showed a high degree of

Table 4. Mean and SE for bloody feathers and footpad lesions of simulated CAI data of 25, 50, 100, and 150 ducks using the bootstrapping technique for 12 flocks (experiment 1 and 2).

		Information used (number of ducks evaluated)				
Welfare indicators	Flock	25	50	100	150	*P*-value[1]
Bloody feathers	1	0.3700 ± 0.001	0.3722 ± 0.001	0.3704 ± 0.000	0.3708 ± 0.000	0.967
	2	0.2400 ± 0.001	0.2387 ± 0.000	0.2379 ± 0.000	0.2384 ± 0.000	0.599
	3	0.4127 ± 0.001	0.4134 ± 0.001	0.4130 ± 0.000	0.4133 ± 0.000	0.655
	4	0.4142 ± 0.001	0.4130 ± 0.001	0.4130 ± 0.000	0.4133 ± 0.000	0.408
	5	0.3600 ± 0.001	0.3601 ± 0.001	0.3598 ± 0.000	0.3600 ± 0.000	0.893
	6	0.3007 ± 0.001	0.3008 ± 0.001	0.3001 ± 0.000	0.3000 ± 0.000	0.184
	7	0.2337 ± 0.001	0.2323 ± 0.000	0.2335 ± 0.000	0.2333 ± 0.000	0.767
	8	0.1534 ± 0.001	0.1540 ± 0.000	0.1534 ± 0.000	0.1533 ± 0.000	0.506
	9	0.3667 ± 0.001	0.3665 ± 0.001	0.3666 ± 0.000	0.3667 ± 0.000	0.915
	10	0.2263 ± 0.001	0.2267 ± 0.001	0.2266 ± 0.000	0.2267 ± 0.000	0.718
	11	0.5735 ± 0.001	0.5736 ± 0.001	0.5729 ± 0.000	0.5733 ± 0.000	0.614
	12	0.2199 ± 0.001	0.2193 ± 0.001	0.2201 ± 0.000	0.2200 ± 0.000	0.508
Footpad score 1	1	0.4380 ± 0.001	0.4363 ± 0.001	0.4369 ± 0.000	0.4371 ± 0.000	0.578
	2	0.4764 ± 0.001	0.4765 ± 0.001	0.4768 ± 0.000	0.4768 ± 0.000	0.559
	3	0.3873 ± 0.001	0.3873 ± 0.001	0.3867 ± 0.000	0.3867 ± 0.000	0.295
	4	0.3799 ± 0.001	0.3799 ± 0.001	0.3801 ± 0.000	0.3800 ± 0.000	0.833
	5	0.2066 ± 0.001	0.2075 ± 0.000	0.2070 ± 0.000	0.2067 ± 0.000	0.642
	6	0.2801 ± 0.001	0.2810 ± 0.000	0.2802 ± 0.000	0.2800 ± 0.000	0.505
	7	0.3127 ± 0.001	0.3137 ± 0.001	0.3132 ± 0.000	0.3133 ± 0.000	0.604
	8	0.3468 ± 0.001	0.3465 ± 0.001	0.3461 ± 0.000	0.3467 ± 0.000	0.783
	9	0.3269 ± 0.001	0.3262 ± 0.001	0.3261 ± 0.000	0.3267 ± 0.000	0.840
	10	0.2597 ± 0.001	0.2604 ± 0.001	0.2598 ± 0.000	0.2600 ± 0.000	0.993
	11	0.1800 ± 0.001	0.1799 ± 0.001	0.1801 ± 0.000	0.1800 ± 0.000	0.909
	12	0.2997 ± 0.001	0.3001 ± 0.001	0.3000 ± 0.000	0.3000 ± 0.000	0.816

[1]One-way ANOVA and post hoc tests using pairwise *t*-test.

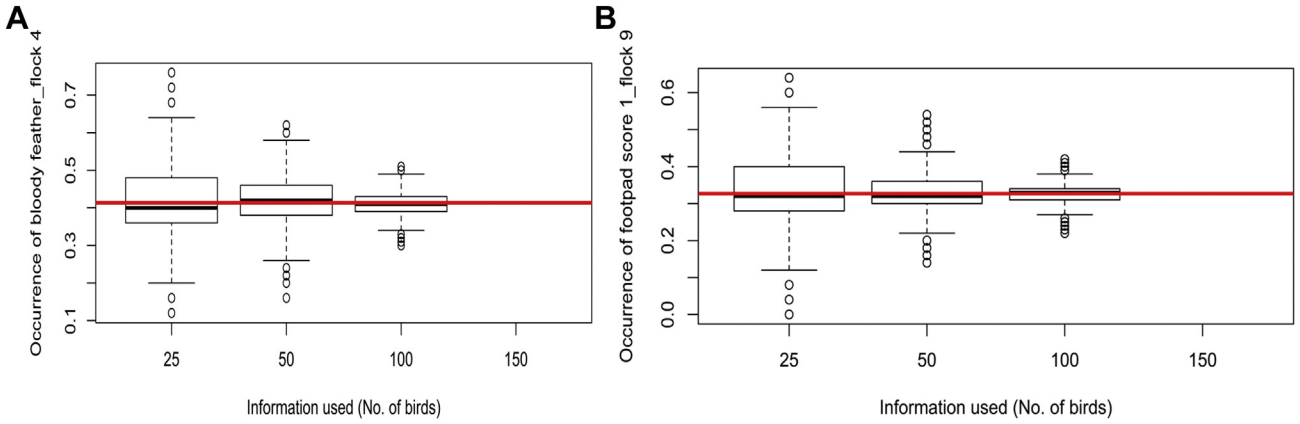

Figure 2. Mean values and SEM of the data set for bloody feather (A) and footpad score 1 (B) were calculated by taking random samples of individual evaluation of 25, 50, 100, and 150 birds. Simulations were run 10,000 times using bootstrapping technique in R.

inter-rater reliability when leg issues were pooled into a single score.

Among the evaluated assessment methods, TW and LO had the highest degree of agreement. The sole disagreement was in the number of dead ducks found, which was higher when assessed using LO than any other method. This disparity may reflect the relative ease with which dead birds can be detected using this method because these birds would have been left behind in an otherwise empty barn. In addition, some of the ducks deemed unfit for transport would have been culled during the LO process, particularly if they required immediate euthanasia. This would have further inflated the number of birds within this category. Marchewka et al. (2015) similarly speculated that increased visibility of birds that would otherwise be cowering or laying on the ground contributed to the relatively higher prevalence of sick, immobile, and dead turkeys, which they observed using LO vs. other assessment methods. Like in the present study, they reported an overall high degree of agreement between outcomes of TW and LO assessments.

The CAI method resulted in higher estimated incidences of most welfare indicators than TW (experiment 1 and 2) and LO (experiment 2). DE (experiment 2) resulted in prevalence estimates that were intermediate: in between those reported by CAI and TW or LO, in agreement with CAI, or with TW and LO. Bright et al. (2006) previously reported a good level of agreement between the CAI and a version of the DE method applied to laying hens. In the present study, differences among the results obtained using the 4 assessment

Table 5. Mean and SE for gait score 2 and bloody feather of simulated transect walks data of 2, 3, 4, 5, and 7 transects using the bootstrapping technique for 12 flocks (experiment 1 and 2).

Welfare indicators		Information used (number of transects)					
Gait score 2	Flock	2	3	4	5	7	*P*-value[1]
	1	0.2702 ± 0.001	0.2714 ± 0.001	0.2692 ± 0.000	0.2699 ± 0.000	0.2703 ± 0.000	0.961
	2	0.2488 ± 0.001	0.2483 ± 0.001	0.2481 ± 0.000	0.2491 ± 0.000	0.2491 ± 0.000	0.655
	3	0.1861 ± 0.001	0.1860 ± 0.001	0.1872 ± 0.000	0.1861 ± 0.000	0.1861 ± 0.000	0.993
	4	0.3645 ± 0.002	0.3572 ± 0.002	0.3577 ± 0.001	0.3570 ± 0.001	0.3579 ± 0.001	0.899
	5	0.0701 ± 0.002	0.0692 ± 0.002	0.0692 ± 0.000	0.0694 ± 0.000	0.0692 ± 0.000	0.229
	6	0.1530 ± 0.001	0.1528 ± 0.001	0.1519 ± 0.000	0.1522 ± 0.000	0.1524 ± 0.000	0.373
	7	0.4588 ± 0.002	0.4580 ± 0.001	0.4592 ± 0.000	0.4586 ± 0.000	0.4575 ± 0.000	0.648
	8	0.3785 ± 0.001	0.3777 ± 0.001	0.3801 ± 0.000	0.3791 ± 0.000	0.3793 ± 0.000	0.414
	9	0.5177 ± 0.003	0.5182 ± 0.002	0.5201 ± 0.001	0.5172 ± 0.001	0.5182 ± 0.001	0.993
	10	1.433 ± 0.003	1.437 ± 0.002	1.434 ± 0.001	1.434 ± 0.001	1.436 ± 0.001	0.670
	11	1.056 ± 0.003	1.052 ± 0.002	1.050 ± 0.001	1.053 ± 0.001	1.051 ± 0.000	0.242
	12	1.141 ± 0.003	1.142 ± 0.002	1.145 ± 0.002	1.143 ± 0.001	1.144 ± 0.001	0.330
Bloody feather	1	1.697 ± 0.004	1.695 ± 0.002	1.695 ± 0.002	1.700 ± 0.001	1.699 ± 0.000	0.314
	2	1.460 ± 0.002	1.461 ± 0.001	1.461 ± 0.001	1.460 ± 0.000	1.460 ± 0.000	0.780
	3	2.397 ± 0.003	2.403 ± 0.002	2.399 ± 0.002	2.399 ± 0.001	2.401 ± 0.000	0.483
	4	3.075 ± 0.004	3.065 ± 0.003	3.066 ± 0.002	3.068 ± 0.001	3.067 ± 0.000	0.186
	5	2.074 ± 0.004	2.068 ± 0.003	2.078 ± 0.002	2.073 ± 0.001	2.073 ± 0.000	0.785
	6	1.725 ± 0.002	1.724 ± 0.002	1.726 ± 0.001	1.725 ± 0.001	1.727 ± 0.000	0.391
	7	1.341 ± 0.003	1.342 ± 0.002	1.344 ± 0.001	1.344 ± 0.001	1.344 ± 0.000	0.350
	8	1.503 ± 0.004	1.500 ± 0.002	1.504 ± 0.001	1.501 ± 0.001	1.501 ± 0.000	0.770
	9	6.016 ± 0.038	5.986 ± 0.028	5.988 ± 0.021	5.986 ± 0.015	5.994 ± 0.000	0.646
	10	10.14 ± 0.030	10.11 ± 0.022	10.11 ± 0.016	10.10 ± 0.011	10.11 ± 0.000	0.401
	11	5.806 ± 0.024	5.780 ± 0.018	5.777 ± 0.013	5.803 ± 0.009	5.798 ± 0.000	0.856
	12	4.501 ± 0.015	4.528 ± 0.010	4.544 ± 0.007	4.524 ± 0.004	4.525 ± 0.000	0.129

[1]One-way ANOVA and post hoc tests using pairwise *t*-test.

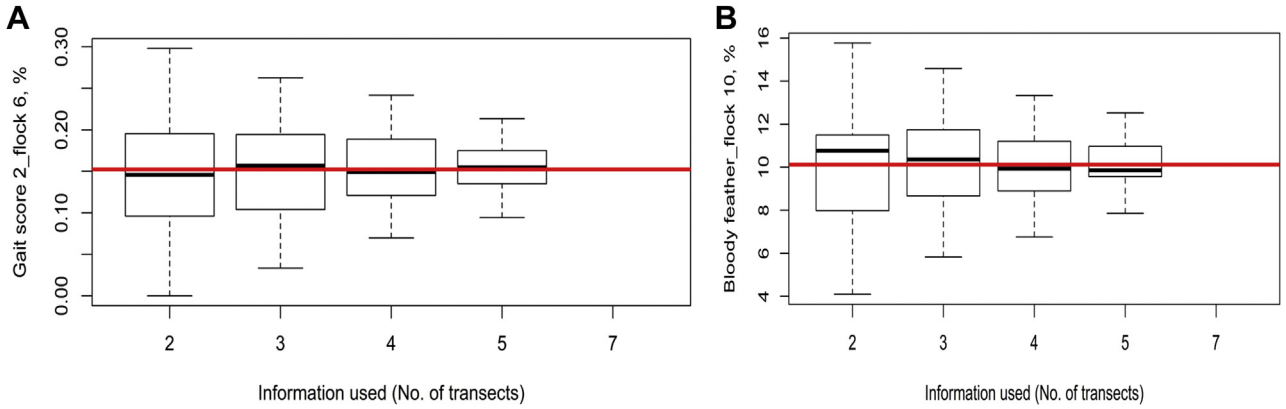

Figure 3. Mean values and SEM for gait score 2 and (A) blood on feather (B) expressed as percentage for 2, 3, 4, 5, and 7 transects used in 10,000 simulation using bootstrapping technique in R.

methods could be attributed to differences in sample size, the need to corral birds, the distance from which birds were evaluated, or a combination of these factors. One additional bird within a specific welfare indicator category translates to a prevalence increase of 0.67% when based on a sample of 150 birds (CAI) or 2.0% for a sample size of 50 (DE). Meanwhile, the impact of a single bird on total prevalence is relatively small for TW and LO, as these methods allow for the inspection of most or all individuals within a flock. Sample size does not, however, explain why the highest prevalence of many of the indicators was observed during CAI, but not DE. A possible contributing factor could be that the process of corralling birds, which is only conducted during CAI, may result in sampling bias. Marchewka et al. (2013) suggested that birds with mobility problems and other characteristics that prevent them from escaping are likely to be over-represented when birds are corralled.

Yet another reason behind the variation in prevalence estimates across the evaluated methods could be the distance and angle from which the birds were inspected. CAI allowed the observers to inspect the birds from all sides and up close, whereas birds were evaluated from above and from a distance during DE, TW, and LO. This difference in the resolution or detail with which animals could be assessed could explain why prevalence estimates of feather cleanliness and feather quality scores were all higher when CAI vs. TW was used in experiment 1. It was likely more difficult to differentiate from a distance, with a high degree of confidence, the size of the affected area (greater or equal to or smaller than 5 cm) and whether feather damage was due to staining, feather loss, or both. Although all the feather scores were converted into a single binary score during experiment 2, estimates related to feather condition were again higher when obtained using CAI. It is possible that using CAI observers were able to identify feather deficiencies that were too small to be seen see from a distance or obstructed from view by the duck's own feathers or by the presence other ducks in the pen. The numerically large difference in the estimate of feather damage on the wing lends support to this interpretation. Other indicators, such as nostril condition, may also have been impacted by the distance from which the assessment was made.

The desire to obtain the largest possible sample to ensure the most representative welfare assessment results has to be balanced against logistical constraints, which include labor costs and flock assessment time requirements. We found that assessment of a flock via TW required a smaller time investment than the evaluation of 150 ducks across 5 sampling locations using CAI. These results replicate findings reported by Marchewka et al. (2013) who compared the time requirement for broiler welfare assessment using CAI and TW. Following Marchewka et al. (2015), we used bootstrapping analyses to determine whether the number of sampled ducks or transects using CAI and TW could be reduced without affecting data accuracy. The results suggest that sampling of 4 transects could prevalence estimates comparable with those obtained using 7 transects. When using fewer transects is advisable that the transects used be spread across the farm to decrease the likelihood of double counting birds as they move across transects, and that transects of the same dimensions be used to the extent possible. It should be noted that as in previous work (Marchewka et al., 2013, 2015, 2019; BenSassi et al., 2019a,b), drinkers and feeders were used as transect delineators in the present study. However, unlike in previous studies, the lines and feeders were not equally distributed resulting in transects of varying widths. For our analysis, we assumed that birds would be distributed evenly across all of the transects but had no way to evaluate this assumption. Therefore, while promising, the result of the bootstrapping analysis on data collected using TW would benefit from independent replication.

The bootstrapping analysis for CAI showed that the prevalence estimates obtained using this method were affected by sample size. Given that the evaluation of 150 ducks took over 3 h, it is not likely feasible to increase the sample size, particularly if using the assessment strategy as part of an audit program. The balance between

information obtained using assessment strategies that involve corralling and handling of birds and the time required to conduct such assessment has been identified as one of the reasons why animal-based assessment strategies for poultry, such as the Welfare Quality Assessment for Broilers, have not been widely implemented (De Jong et al., 2016). Further research involving sampling of more than 150 birds per flock is necessary to identify the specific duck welfare indicators that are most affected by the sampling size and whether simplification of the scoring system or the number of categories used could improve the practicality of this welfare assessment strategy.

The 2 experiments described information trade-offs associated with 4 welfare assessment methods applied to commercial duck flocks. Each method was revealed to have its own set of pros and cons with respect to the amount of time it required, the percentage of the flock that could be sampled, and the about of detail that could be observed about the welfare status of individual birds. It is yet to be determined which of the tested method produces estimates closest to actual prevalence. Marchewka et al. (2015) used the LO as the reference method arguing that it allowed for an assessment of all birds in the flock. Numerous studies have shown that larger numbers of individuals need to be sampled to ensure data accuracy, especially when the actual prevalence of the condition in question is low within a flock (Farver, 1984; Van Os et al., 2018; Mullan et al., 2009), or when the prevalence varies widely across flocks (Bright et al., 2006). These reports lend support to the use of methods such as TW and LO that allow for a greater proportion of the flock to be evaluated. However, the benefits of any sampling method must be balanced against its limitations. The reported results raise questions about the amount of detail that can be recorded using the distance evaluation methodologies that allow for evaluation of a large proportion of the flock (TW and LO) vs. those that allow for close up inspection of a relatively small sample of birds (CAI). A question that remains to be answered is the degree to which the detail obtained via CAI and lost when using TW and LO (ex. ability to note injuries and feather damage <5 cm in diameter) are important to the overall metric of flock welfare. Further research should also investigate the trade-offs of animal welfare sampling strategies on assessment of duckling at younger ages.

ACKNOWLEDGMENTS

The authors would like to thank Maple Leaf Farms Inc. for providing facilities, staff, and help with data collection, and especially Steven Corbitt, Andreas Schlatler, Ben Fetrow, Zach Tucker, and Dan Shafer for their valuable input. The authors thank Margaret Ann De Luz for helping with data collection and the undergraduate and graduate student members of the Makagon lab for their feedback and for assistance with data entry and management. The authors extend their gratitude to Neil Willits (UC Davis Department of Statistics) for proving statistical consultation and to two anonymous reviewers for providing feedback on an earlier draft of this manuscript.

REFERENCES

AWIN. 2015. AWIN welfare assessment protocol for turkeys. Accessed Dec. 2019. https://air.unimi.it/retrieve/handle/2434/269107/384771/AWINProtocolTurkeys.pdf.

BenSassi, N., X. Averós, and I. Estevez. 2019a. The potential of the transect method for early detection of welfare problems in broiler chickens. Poult. Sci. 98:522–532.

BenSassi, N., X. Averós, and I. Estevez. 2019b. Broiler chickens on-farm welfare assessment: estimating the robustness of the transect sampling method. Front. Vet. Sci. 6:236.

Blokhuis, H. J., R. B. Jones, R. Geers, M. Miele, and I. Veissier. 2003. Measuring and monitoring animal welfare: transparency in the food product quality chain. Anim. Welf. 12:445–455.

Bright, A., T. A. Jones, and M. S. Dawkins. 2006. A non-intrusive method of assessing plumage condition in commercial flocks of laying hens. Anim. Welfare 2006 15:113–118.

Butterworth, A., J. A. Mench, and N. Wielebnowski. 2011. Practical strategies to assess (and improve) welfare'. In: M. C. Appleby, J. A. Mench, I. A. S. Olsson and B. O. Hughes (Eds.), Animal Welfare. CABI Publishing Oxford, pp. 200–214, 2011. ISBN: 9781845936594.

Colton, S., and G. S. Fraley. 2014. The effects of environmental enrichment devices on feather picking in commercially housed Pekin ducks. Poult. Sci. 93:2143–2150.

de Jong, I. C., V. A. Hindle, A. Butterworth, B. Engel, P. Ferrari, H. Gunnink, T. Perez Moya1, F. A. M. Tuyttens, and C. G. van Reenen. 2016. Simplifying the Welfare Quality® assessment protocol for broiler chicken welfare. Animal 10:117–127.

Farver, T. B. 1984. Some practical considerations in sampling livestock populations to estimate disease prevalence and other parameters. Prev. Vet. Med. 2:453–462.

Ferrante, V., S. Lolli, L. Ferrari, T. T. N. Watanabe, C. Tremolada, J. Marchewka, and I. Estevez. 2018. Differences in prevalence of welfare indicators in male and female turkey flocks (Meleagris gallopavo). Poult. Sci. 98:1568–1574.

Fraley, S. M., G. S. Fraley, D. M. Karcher, M. M. Makagon, and M. S. Lilburn. 2013. Influence of plastic slatted floors compared with pine shaving litter on Pekin Duck condition during the summer months. Poult. Sci. 92:1706–1711.

Jones, T. A., and M. S. Dawkins. 2010. Environment and management factors affecting Pekin duck production and welfare on commercial farms in the UK. Br. Poult. Sci. 51:12–21.

Karcher, D. M., M. M. Makagon, G. S. Fraley, S. M. Fraley, and M. S. Lilburn. 2013. Influence of raised plastic floors compared with pine shaving litter on environment and Pekin duck condition. Poult. Sci. 92:583–590.

Marchewka, J., I. Estevez, G. Vezzoli, V. Ferrante, and M. M. Makagon. 2015. The transect method: a novel approach to on-farm welfare assessment of commercial turkeys. Poult. Sci. 94:7–16.

Marchewka, J., T. T. N. Watanabe, V. Ferrante, and I. Estevez. 2013. Welfare assessment in broiler farms: transect walks versus individual scoring. Poult. Sci. 92:2588–2599.

Marchewka, J., G. Vasdal, and R. O. Moe. 2019. Identifying welfare issues in turkey hen and tom flocks applying the transect walk method. Poult. Sci. 98:3391–3399.

Mullan, S., W. J. Browne, S. A. Edwards, A. Butterworth, H. R. Whay, and D. C. J. Main. 2009. The effect of sampling strategy on the estimated prevalence of welfare outcome measures on finishing pig farms. Appl. Anim. Behav. Sci. 119:39–48.

Van Os, J. M. C., C. Winckler, J. Trieb, S. V. Matarazzo, T. W. Lehenbauer, J. D. Champagne, and C. B. Tucker. 2018. Reliability of sampling strategies for measuring dairy cattle welfare on commercial farms. J. Dairy Sci. 101:1495–1504.

Welfare Quality® 2009. Welfare Quality® Assessment Protocol for Poultry (Broilers, Laying Hens). Welfare Quality® Consortium, Lelystad, The Netherlands.