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

Animal Welfare 2022, 31: 163-173
ISSN 0962-7286
doi: 10.7120/09627286.31.2.001

Validating owner-reporting of feather condition of pet Psittaciformes using photographs

EL Mellor*[†‡], M Mendl[†], G Mason[§], C Davison[†], Y van Zeeland[#] and IC Cuthill[‡]

[†] Bristol Veterinary School, University of Bristol, Langford House, Langford, Bristol BS40 5DU, UK
[‡] School of Biological Sciences, University of Bristol, Bristol Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[§] College of Biological Science, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada
[#] Department of Clinical Sciences of Companion Animals, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 108, 3584 CM Utrecht, The Netherlands
* Contact for correspondence: emma.mellor@bristol.ac.uk

Abstract

Reporting of outcome variables by caregivers in welfare studies is commonplace but is open to subjective bias and so requires validation. Biases can occur in either direction: familiarity with an animal allows a deeper insight into welfare problems, but also can lead to reticence in admitting that an animal in one's care is experiencing problems. Here, we aim to validate owner-reporting of plumage condition of pet parrots, including those with self-inflicted feather-damaging behaviour (FDB), by comparing owners' scores of feather condition with those of two independent raters, blind to the owners' and each other's assessments. We surveyed pet parrot owners to collect data on basic demographics and feather condition, and requested four standardised photographs of birds. We received 259 responses (17% of the 1,521 people contacted); 78 sets of images of appropriate quality for assessment by raters were provided. Mean percentage agreement between owners' and raters' scores was mostly fair to substantial using Cohen's kappa; however, raters scored a greater proportion of feather damage than did owners. Overall, our results indicate owner-reporting of feather condition, including FDB, to be generally reliable and consistent with independent assessment of photographs. As the use of photographs can be limited by image quality, a failure to represent the long-term state of a parrot, and the potential for incorrect recording if assessed without relevant information (eg on moulting), this evidence that owner-reports can be reliable opens the door for larger-scale surveys of the extent of welfare-relevant problems.

Keywords: *animal welfare, feather-damaging behaviour, feather picking, feather plucking, plumage, psittacine*

Introduction

Caregivers' reports of animals' behaviour, longevity, and/or disease are often used as outcome variables in studies of welfare and life-history (Müller *et al* 2011; McDonald Kinkaid 2015; Pollard *et al* 2019; Finnegan *et al* 2020; Mellor 2020). Such reporting comes with benefits over professionally reported outcomes including deeper personal knowledge of the individual animal, speed of data collection, and larger sample sizes (Mellor *et al* 2018). For instance, Potter *et al* (2017) and Pollard *et al* (2019) both reported that only ~50% of horse owners had a veterinarian confirm diagnosis of laminitis, meaning that relying solely on veterinary-diagnosed incidences may underestimate the extent of behavioural problems and/or disease. Validation of caregiver reporting, though, is less common (but see, for example, Pollard *et al* 2017 and Malalana *et al* 2019). Validation is an important consideration, as mild cases of behavioural problems and diseases might be undiagnosed by caregivers (eg Pollard *et al* 2017; Malalana *et al* 2019); caregivers' understanding of how observed behaviours relate to an animal's welfare might be limited (eg McBride & Long 2001; Burn 2011); and, importantly, the person responsible for the well-being of an animal may unconsciously downplay, or consciously withhold, evidence of poor welfare.

In our previous comparative studies of parrots, we used owner-reporting of stereotypic behaviours, repetitive behaviours indicative of past or ongoing welfare problems (Mason 2006), as outcome variables (Mellor *et al* 2021). These were feather-damaging behaviour (FDB; self-directed feather chewing and/or plucking: Harrison & Harrison 1986; Orosz 2006; van Zeeland *et al* 2009), other non-feather related oral stereotypic behaviours (eg bar biting and cage chewing) and whole body stereotypic behaviours (eg route tracing, weaving, and head bobbing) (Mellor *et al* 2021). Stereotypic behaviours are more common in birds whose living condi-

tions are impoverished or stressful (eg without enrichment, located near stressors, experience of traumatic events: Meehan *et al* 2003a,b, 2004; Garner *et al* 2006; Greenwell & Montrose 2017), and/or in those unable to perform important natural behaviours (eg being unable to fly: Schmid *et al* 2006; Mellor 2014). FDB and other forms of stereotypic behaviours differ in ecological basis, helping to explain between-species differences in prevalence. Thus, FDB is more prevalent in species whose wild diets require extensive food-handling, such as white cockatoos (*Cacatua alba*) (Mellor *et al* 2021), whereas non-feather-related oral behaviours and movements involving the whole body are more prevalent in intelligent species (as inferred from species-typical brain volume), such as red-shouldered macaws (*Diopsittaca nobilis*) (Mellor *et al* 2021). In our previous work, clear definitions of each type of stereotypic behaviour were provided to owners. However, it remains uncertain how accurately owners might be able to identify and report such behaviours in their birds, and validation is thus required.

Validating owner-reporting of oral and whole-body forms of stereotypic behaviours in pet parrots is challenging, however, because doing so would require extensive international travel to capture all species sampled, intrusion into owners' homes and/or use of recording devices. An unfamiliar researcher making live observations may affect birds' and owners' behaviour (see Carpenter 1934; Martin & Bateson 2007); using recording devices for behavioural data collection involves financial cost and presents some ethical concerns regarding privacy; and regardless of collection method, behavioural observations are time-consuming. FDB reporting, though, might naturally lend itself to quick and easy validation. Performance of FDB can be hard to distinguish from normal preening, especially for the untrained eye, and can take place at night or at other times without an observer present (Meehan *et al* 2003b). However, FDBs' effects may be visible, ie chewed and/or missing feathers and/or skin damage, so the behaviour is normally inferred indirectly using plumage-scoring systems (Meehan *et al* 2003b; van Zeeland *et al* 2013). In our case, this represents opportunity to validate owner-reporting of feather condition visually using photographs of birds (also see van Zeeland *et al* 2013). Photographs are non-invasive, quick, and simple for owners to take and share. Therefore, our focus here is on validation of owner-reporting of feather condition in pet parrots. If owners can accurately determine the presence of feather damage, then we expect to see agreement between owners' reports of damage and those of independent raters assessing photographs of the same birds. With our previous comparative work in mind (Mellor *et al* 2021), we also predict that, if reliable, both owners and raters should report more damage on birds with reported FDB than on those without.

Materials and methods

Ethical approval

Ethical approval for data collection was granted by the University of Bristol's Faculty of Health Sciences Research Ethics Committee (reference number: 32441).

Data collection

To validate owner-reporting of feather condition across psittacine species, we designed a short survey with 13 questions (< 10 min to complete), which was live from May–September 2020 (see Appendix S1 in Supplementary Material for a copy, and Table 1 for details). In addition to filling out the survey, we also requested that owners provide four standardised images of their birds that would allow us to assess feather condition (see example photographs in Appendix S1). For our study, because it was difficult to definitively assess from photographs whether feather damage results from FDB or from other causes, we mainly assessed validation of owner-reporting of *any* feather damage. Requests to participate in our survey were sent out via email to previous participants of a larger ongoing survey that collected information on pet parrots' rearing, living conditions and behaviour (www.parrotsurvey.com) (Mellor *et al* 2021). Of the 1,955 people who submitted survey responses between 1 April 2012 and 1 July 2013, we initially contacted the 1,788 who had followed the correct procedure for identifying their bird's species (from images, to reduce likelihood of owner-bias). Of these, 1,521 were successfully sent an email (ie the email address was still valid). Figure 1 provides a step-by-step schematic showing details of data collection, processing, and analysis.

Data processing and image scoring

We received 259 responses to our current survey (see Figure 1). For 111 of these, owners also sent images of which 110 sets were usable (one folder contained no images). Images from each set were scored independently by two raters who were not parrot specialists, and who were inexperienced in assessing plumage condition but knowledgeable of the subject area (one being the lead author and the other a veterinary student) and available for this project. Raters were blind to the owners' and each other's scores. The raters scored the presence or absence of damaged plumage on various body areas and feather types, did the same regarding skin damage, and scored the overall severity of feather damage (from none [0] to severe [3]; see Table 1). A random number generator was used to select a random 20% of the 110 sets of images for re-scoring, to calculate each rater's intra-observer reliability (ie that *within* a given rater). Afterwards, inter-observer reliability (ie that *between* the two raters) of the two raters was assessed across all 110 sets of scores. For these intra- and inter-observer scores, the raters scored 'Not visible' if a body area and/or feather type was not visible on a given bird's set of images.

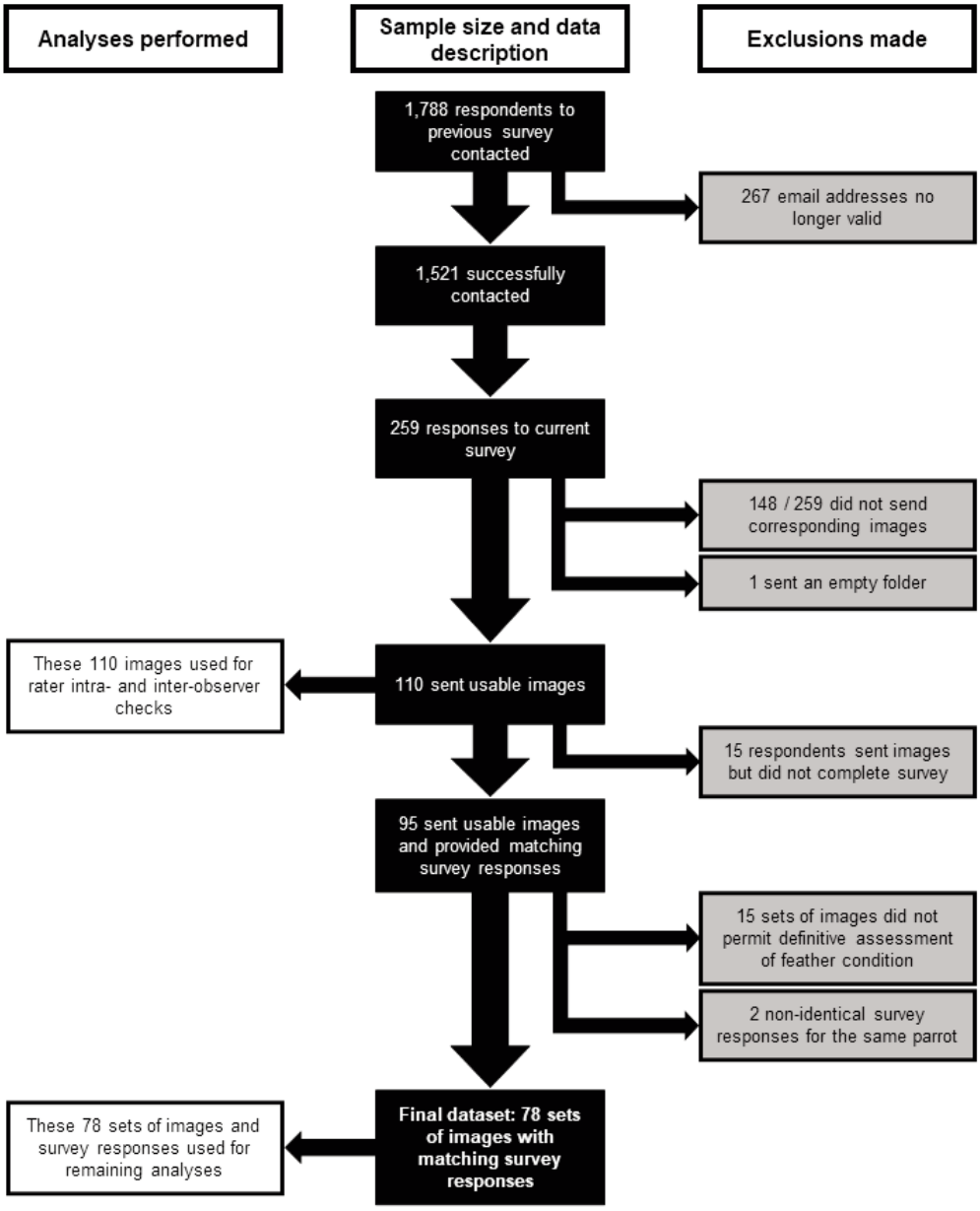
https://doi.org/10.7120/09627286.31.2.001 Published online by Cambridge University Press

Table 1 Details of survey questions and their responses used in the current study. A copy of the entire survey listing all 13 questions is provided in Appendix S1.

Question	Response
Demographics	
Which species is your parrot? Common and/or scientific names acceptable	Free form text box
What sex is your parrot?	Female
	Male
	Uncertain
Information on feather damage?	
Does your parrot currently have damaged plumage, ie damaged and/or missing feathers?	Yes
	No
Which body parts are affected?	Head
	Throat/neck
	Chest
	Back
	Wings: dorsal surface
	Wings: ventral surface
	Tail
	Legs
Which type(s) of feathers are damaged?	Down feathers
	Covert feathers
	Primary and/or secondary flight feathers (remiges, wings)
	Tail feathers (retrices)
	Newly developing feathers (blood feathers)
	Mature feathers
Is there skin damage present?	Yes
	No
Does your parrot pluck, bite or chew its own feathers?	Yes
	No
	No, the parrot is plucked by another bird
	Unknown
How severe is the feather damage caused by your parrot's behaviour? Ordinally ranked 0–3	No damage (0)
	Mild (focal areas, most feathers intact) (1)
	Moderate (patchy distribution, may leave down alone) (2)
	Severe (bird [almost] completely devoid of feathers) (3)

For 15 of the 110 sets of images, owners had not completed the survey, so we excluded these images and scores from further analyses. Additionally, one respondent had submitted duplicate versions of the survey and given different responses, so we discarded theirs too, leaving us with 95 sets of images and corresponding survey responses. Next, we further excluded cases in which images were clearly not taken at the same time (judged by time-stamps or, if these were missing, by image content) and sets that did not permit definitive assessment of feather condition. This

Figure 1



Schematic detailing the workflow regarding data processing and analysis of survey responses and images received from pet parrot owners.

left us with a final dataset of 78 sets of images and their corresponding survey responses for assessment of rater-owner inter-observer reliability (ie *between* each rater and owners). Survey responses made by owners regarding the presence/absence and severity of damaged plumage across the same body areas, feather types and skin damage as mentioned above, were used to calculate inter-observer reliability between each rater and owners. For the rater-owner inter-observer scores, the cells previously scored as 'Not visible' by the raters (see above) were re-coded as 'NA' and any survey responses unanswered by owners were likewise scored as such, to allow for correct comparison across the sets of scores (ie all NAs were excluded from analyses).

As a result of its association with welfare, we also wanted to assess agreement between raters and owners regarding FDB more specifically. Thus, for the subset of 31 parrots whose owners indicated that their bird's feather damage was self-directed (see Table 1), we calculated inter-observer reliability scores between the raters' and owners' scores.

Finally, for the 78 parrots with images and survey responses, we assessed possible bias in overall agreement between each rater and the owners, which might be explained by parrot species identity and/or sex. For these analyses, our outcome variable was, for each individual bird, the proportion of total agreement across all scores between each rater and owners. As some species were repre-

https://doi.org/10.7120/09627286.31.2.001 Published online by Cambridge University Press

sented by single or few animals, we grouped most species into broader, taxonomic groups (see Table S2) to better enable reliable comparisons. Our predictor variables were then 'taxonomic group' and 'sex' (female/male/uncertain) as reported by owners.

Statistical analysis

Statistical analyses were performed in R version 4.0.4 (R Core Team 2021). Accuracy of feather condition reporting was assessed in two ways. First, using the 'irr' package (Gamer *et al* 2019), we calculated intra- and inter-observer reliability scores for raters, as percentage agreement and Cohen's kappa (κ, which accounts for agreement purely by chance) for each body area, for each feather type, for skin damage, and for severity scores (the latter being a weighted kappa, as these data were ordinal). The same procedure was used to calculate inter-observer agreement between the raters' and owners' scores. Second, we compared the raters' scores regarding: (i) proportion of feather and skin damage (0 = no damage, 1 = damage present on all feather types, body areas, and on the skin); (ii) severity scores; and (iii) proportion of scores recorded as 'Not visible' (see Tables S3 and S4). Similarly, for comparisons between the raters' and owners' scores we did (i) and (ii), but excluded any 'NA's (see Tables S3 and S4). As data were non-normally distributed, we used Wilcoxon matched-pairs tests for these comparisons.

To assess whether parrots with owner-reported FDB have more damage than those without, Mann-Whitney *U* tests were used to compare: (i) the proportion; and (ii) severity of feather and skin damage on birds with and without owner-reported FDB, for both the owners' and the independent raters' scores. For these analyses, we excluded all 'NA's (see Tables S3 and S4).

To assess whether the proportion of rater-owner agreement across all 78 birds might be explained by taxonomic grouping and/or sex, as residuals were non-normally distributed, we used beta regression from the 'betareg' package (Cribari-Neto & Zeileis 2010) suitable for bounded data. Beta regression requires that no values be exactly 0 or 1 so, as no values were 0 but some were 1, we took 0.01 off all outcome values. We assessed homogeneity of residuals on diagnostic plots, and potentially influential data-points using Cook's distance. For data-points deemed influential (ie Cook's distance > 0.05) we removed each one in turn and re-ran the model to assess effects. The potential significance of each predictor was assessed using likelihood ratio tests to evaluate changes in model deviance on removing each term in turn (Crawley 2013). When removing a predictor term caused a significant increase in model deviance, we then examined its relationship to the outcome using the terms of the model. Results were considered significant at $P < 0.05$, all *P*-values are two-tailed, and means with standard errors are reported where appropriate.

Results

Frequency tables are provided in Tables S3 and S4, describing a summary of the distribution of scores given by raters and owners. For 31/78 (39.7%) parrots, their owners indicated they had FDB.

Intra-observer reliability

Mean intra-observer reliability across 22 sets of images was 93.3 (± 0.01)% for Rater 1 and 97 ± (0.01)% for Rater 2, and for each agreement was mostly almost perfect to perfect as judged by κ (see Table S5). Therefore, we used each rater's original scores for all inter-observer reliability checks. For both raters there were no significant differences in the proportion of feather damage reported for their first and second sets of scores (Rater 1: $V = 25.50$; $P = 0.31$; Rater 2: $V = 0$; $P = 0.37$), for the proportion of 'Not visible' scores (Rater 1: $V = 6$; $P = 0.37$; Rater 2: $V = 4$; $P = 0.77$), nor for their severity scores (Rater 1: $V = 1$; $P = 0.99$; Rater 2: $V = 1$; $P = 0.99$).

Between-rater inter-observer reliability

When scoring all 110 images received from owners, mean inter-observer reliability agreement between the two raters was 84.1 (± 0.02)%, and mostly moderate to substantial as judged by κ (see Table S6). The raters did not differ in their proportions of feather damage scores ($V = 755.50$; $P = 0.57$), their 'Not visible' scores ($V = 169.50$; $P = 0.62$), nor their severity scores ($V = 314.50$; $P = 0.75$).

Rater-owner inter-observer reliability

Across the full dataset of 78 parrots with survey responses, the mean inter-observer reliability between Rater 1 and owners' scores was 77.6 (± 0.03)%, and agreement was mostly fair to moderate as judged by κ (see Table 7). For Rater 2, mean agreement was 79.1 (± 0.04)%, and fair to substantial (see Table 8). Both raters reported a highly significantly greater proportion of feather damage than did owners (Rater 1: $V = 1{,}257$; $P < 0.001$; Rater 2: $V = 1{,}205.50$; $P < 0.001$), but there were no such differences regarding the severity scores (Rater 1: $V = 423.50$; $P = 0.24$; Rater 2: $V = 297$; $P = 0.30$). Median proportion of feather damage recorded by Rater 1 was 0.29, for Rater 2 was 0.27, whereas for owners it was 0.12.

For the subset of 31 parrots whose owners indicated they have FDB, mean agreement between raters and owners was 69.9 (± 0.05)% for Rater 1 and 72 (± 0.04)% for Rater 2 (Tables 9 and 10). For Rater 1, agreement as judged by κ was mainly slight to moderate, and slight to substantial for Rater 2. Rater 1 tended to report a greater proportion of damage than owners ($V = 326$; $P = 0.06$) and for Rater 2 this was significant ($V = 333.50$; $P = 0.04$), but again for severity scores there were no such differences (Rater 1: $V = 51$; $P = 0.61$; Rater 2: $V = 44$; $P = 0.59$). Median proportion of feather damage reported by Rater 1 was 0.60, by Rater 2, 0.53, while that reported by owners was 0.43.

Table 7 Inter-observer reliability scores calculated between Rater 1 and parrot owners of 78 sets of images that passed quality criteria, given as percentage agreement and Cohen's (κ, agreement between two scores after accounting for agreement purely by chance: Cohen 1960; McHugh 2012).

	Rater 1		
	Agreement	**Cohen's kappa**	**n**
Any feather damage?	67.9%	$\kappa = 0.36, Z = 3.19, P < 0.01$	78
Specific body parts			
Head	94.9%	$\kappa = 0.48, Z = 4.96, P < 0.001$	78
Throat/neck	85.9%	$\kappa = 0.54, Z = 5.18, P < 0.001$	78
Chest	88.5%	$\kappa = 0.72, Z = 6.35, P < 0.001$	78
Back	69.2%	$\kappa = 0.20, Z = 2, P = 0.05$	78
Wings (dorsal surface)	65.4%	$\kappa = 0.31, Z = 3.36, P < 0.001$	78
Wings (ventral surface)	80.0%	$\kappa = 0.58, Z = 2.85, P < 0.01$	20
Tail	76.4%	$\kappa = 0.35, Z = 3.30, P < 0.001$	72
Legs	82.9%	$\kappa = 0.59, Z = 5.33, P < 0.001$	76
Feather types			
Down feathers	87.2%	$\kappa = 0.67, Z = 5.93, P < 0.001$	78
Covert feathers	73.1%	$\kappa = 0.48, Z = 4.96, P < 0.001$	78
Primary/secondary flight feathers	81.6%	$\kappa = 0.40, Z = 3.89, P < 0.001$	76
Tail feathers	76.4%	$\kappa = 0.35, Z = 3.30, P < 0.001$	72
Blood feathers	77.6%	–*	49
Mature feathers	65.5%	$\kappa = 0.33, Z = 2.76, P < 0.01$	58
Other			
Skin damage	94.4%	$\kappa = 0.47, Z = 3.99, P < 0.001$	72
Severity (0–3)	52.6%	$\kappa = 0.43, Z = 5.31, P < 0.001$	78
Mean agreement	77.6%		

κ scores are interpreted as follows: < 0.21 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–0.99 = almost perfect; 1 = perfect. $P < 0.05$ indicates that the rater and owners scores agree more than would be expected by chance. * There was not enough agreement in one level (yes) between the two sets of scores to enable calculation of κ.

Comparisons between parrots with and without owner-reported FDB

For the 31 parrots with owner-reported FDB, their owners scored a highly significantly greater proportion of feather damage than did owners of parrots without reported FDB (n = 47; $W = 1{,}369.50$; $P < 0.001$), as did both raters (Rater 1: $W = 1{,}178$; $P < 0.001$; Rater 2: $W = 1{,}233$; $P < 0.001$). Owner-reported median proportion of feather damage for birds with FDB was 0.43 and without was 0; for Rater 1 values were, respectively, 0.60 and 0; and for Rater 2 they were 0.53 and 0. Severity scores were also highly significantly greater for birds with FDB than those without (owners: $W = 1{,}368.50$; $P < 0.001$; Rater 1: $W = 1{,}161.50$; $P < 0.001$; Rater 2: $W = 1{,}186.50$; $P < 0.001$). The median severity score reported by owners for birds with FDB was 2 and without 0; for Rater 1 values were, respectively, 1 and 0; and for Rater 2, they were 2 and 0.

Potential bias in rater-owner agreement

For Rater 1, removing predictor terms from the full model containing both terms did not significantly increase model deviance (taxonomic grouping: $\chi^2 = 11.72$, df = 7; $P = 0.11$; sex: $\chi^2 = 3.42$, df = 2; $P = 0.18$), and the same was true for Rater 2 (taxonomic grouping: $\chi^2 = 9.38$, df = 7; $P = 0.23$; sex: $\chi^2 = 0.04$, df = 2; $P = 0.98$). Compared to the null model (with no predictor terms), removal of either predictor did not result in a significant increase in model deviance for Rater 1 (taxonomic grouping: $\chi^2 = 10.67$, df = 7; $P = 0.15$; sex: $\chi^2 = 2.37$, df = 2; $P = 0.31$) and, again, the same was found for Rater 2 (taxonomic grouping: $\chi^2 = 9.46$, df = 7; $P = 0.22$; sex: $\chi^2 = 0.12$, df = 2; $P = 0.94$). Therefore, neither taxonomic group nor sex significantly explained the proportion of agreement between raters and owners.

All supplementary material from this paper can be found at https://www.ufaw.org.uk/the-ufaw-journal/31_2_01.

https://doi.org/10.7120/09627286.31.2.001 Published online by Cambridge University Press

Table 8 Inter-observer reliability scores calculated between Rater 2 and parrot owners of 78 sets of images that passed quality criteria, given as percentage agreement and Cohen's (κ, agreement between two scores after accounting for agreement purely by chance: Cohen 1960; McHugh 2012).

	Rater 2		
	Agreement	**Cohen's kappa**	**n**
Any feather damage?	71.8%	$\kappa = 0.43, Z = 3.90, P < 0.001$	78
Specific body parts			
Head	94.9%	$\kappa = 0.48, Z = 4.96, P < 0.001$	78
Throat/neck	75.6%	$\kappa = 0.34, Z = 3.48, P < 0.001$	78
Chest	85.9%	$\kappa = 0.67, Z = 6.02, P < 0.001$	78
Back	78.2%	$\kappa = 0.32, Z = 2.98, P < 0.01$	78
Wings (dorsal surface)	67.9%	$\kappa = 0.34, Z = 3.6, P < 0.001$	78
Wings (ventral surface)	50.0%	$\kappa = 0.06, Z = 0.19, P = 0.85$	8
Tail	89.2%	$\kappa = 0.63, Z = 5.50, P < 0.001$	74
Legs	91.0%	$\kappa = 0.74, Z = 6.56, P < 0.001$	78
Feather types			
Down feathers	88.5%	$\kappa = 0.68, Z = 6.02, P < 0.001$	78
Covert feathers	69.2%	$\kappa = 0.41, Z = 4.32, P < 0.001$	78
Primary/secondary flight feathers	84.0%	$\kappa = 0.45, Z = 4.21\ P < 0.001$	75
Tail feathers	90.5%	$\kappa = 0.66, Z = 5.77, P < 0.001$	74
Blood feathers	83.7%	$\kappa = 0.37, Z = 3.32, P < 0.001$	49
Mature feathers	67.2%	$\kappa = 0.36, Z = 2.94, P < 0.01$	58
Other			
Skin damage	97.2%	$\kappa = 0.65, Z = 5.91, P < 0.001$	72
Severity (0–3)	60.3%	$\kappa = 0.53, Z = 6.47, P < 0.001$	78
Mean agreement	*79.1%*		

κ scores are interpreted as follows: < 0.21 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–0.99 = almost perfect; 1 = perfect. $P < 0.05$ indicates that the rater and owners scores agree more than would be expected by chance.

Discussion

Overall, our study found acceptable levels of agreement between two blinded raters' and owners' scores of the feather condition of 78 pet parrots, as assessed by percentage agreement and using Cohen's kappa. This supports that owner-reports can be reliable (in agreement with Pollard *et al* 2017). As predicted, and reassuringly for our previous work (Mellor *et al* 2021), we also confirmed that owners scored a highly significantly greater proportion of feather damage for parrots with owner-reported FDB than those without, as did both raters. Birds with FDB also had highly significantly greater damage severity scores from both owners and raters. We did find, however, that across the full dataset of 78 parrots, raters scored a highly significantly greater proportion of feather damage than did owners, and a similar pattern (though somewhat less pronounced) was found for the subset of birds with FDB. Additionally, we found slightly less rater-owner agreement for the subset of 31 parrots whose owners indicated they had FDB.

Limitations of our study include variable image quality (eg lighting, focus); the occasional, accidental exclusion of body areas and/or feather types from the shot; and the inability to capture certain body areas in photographs of unrestrained birds, such as the ventral surface of the wing and the area underneath the tail (see also van Zeeland *et al* 2013). This specifically limited the raters' ability to assess damage, hence reducing the efficacy of rater scoring if a bird damaged only these areas. Additionally, the raters were not parrot specialists and, until this study, inexperienced in assessing feather condition. However, our raters' high intra- and inter-observer agreement scores would imply their responses to nonetheless be reliable (regarding reproducibility). We also had some incidences, especially for the subset of parrots with FDB, in which percentage agreement was high, yet κ was low (eg presence of any

Table 9 Inter-observer reliability scores calculated between Rater 1 and parrot owners, on a subset of 31 parrots whose owners report they have feather-damaging behaviour. Agreement is given as percentage and Cohen's kappa (κ, agreement between two scores after accounting for agreement purely by chance: Cohen 1960; McHugh).

	Rater 1		
	Agreement	**Cohen's kappa**	**n**
Any feather damage?	80.6%	κ = –0.06, *Z* = –0.45, *P* = 0.65	31
Specific body parts			
Head	90.3%	κ = 0.37, *Z* = 2.64, *P* = 0.01	31
Throat/neck	77.4%	κ = 0.54, *Z* = 3.39, *P* < 0.001	31
Chest	87.1%	κ = 0.72, *Z* = 4.05, *P* < 0.001	31
Back	41.9%	κ = –0.09, *Z* = –0.57, *P* = 0.57	31
Wings (dorsal surface)	51.6%	κ = 0.11, *Z* = 0.81, *P* = 0.42	31
Wings (ventral surface)	75.0%	κ = 0.39, *Z* = 1.38, *P* = 0.17	8
Tail	75.0%	κ = 0.46, *Z* = 2.65, *P* = 0.01	28
Legs	80.6%	κ = 0.60, *Z* = 3.39, *P* < 0.001	31
Feather types			
Down feathers	80.6%	κ = 0.61, *Z* = 3.43, *P* < 0.001	31
Covert feathers	80.6%	κ = 0.58, *Z* = 3.54, *P* < 0.001	31
Primary/secondary flight feathers	65.5%	κ = 0.27, *Z* = 1.62, *P* = 0.10	29
Tail feathers	75.0%	κ = 0.46, *Z* = 2.65, *P* = 0.01	28
Blood feathers	9.1%	–*	11
Mature feathers	78.9%	κ = –0.08, *Z* = –0.45, *P* = 0.66	19
Other			
Skin damage	86.7%	κ = 0.42, *Z* = 2.32, *P* = 0.02	30
Severity (0–3)	51.6%	κ = 0.33, *Z* = 2.88, *P* < 0.01	31
Mean agreement	*69.9%*		

κ scores are interpreted as follows: < 0.21 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–0.99 = almost perfect; 1 = perfect. *P* < 0.05 indicates that the rater and owners scores agree more than would be expected by chance. * There was not enough agreement in one level (yes) between the two sets of scores to enable calculation of κ.

feather damage and, to a lesser extent, when scoring mature feathers). κ is affected by the prevalence of the outcome under question, meaning that very rare outcomes (for these examples, this being 'no damage') can be associated with very low κ scores, even when agreement is actually high (Viera & Garrett 2005). Kappa likewise under-performed in Giammarino *et al* (2021)'s recent study assessing the performance of several indices of inter-observer reliability for an animal welfare indicator (udder asymmetry of dairy goats). Giammarino *et al* (2021) thus recommended using Bangdiwala's B (Bangdiwala 1985) or Gwet's $\gamma(AC_1)$ (Gwet 2008) rather than κ: a suggestion to bear in mind for future studies. Another limitation of our approach, especially pertinent to welfare, is that photographs only serve as a snapshot in time and may not be representative of a bird's long-term welfare state. This might explain the poor agreement for blood feather scoring for the subset of 31 parrots whose owners reported them as having FDB. If parrots did not have newly growing feathers at the time of being photographed, then the raters would have scored 'no damage', even if the bird did damage them when there (thus likely resulting in disagreement with owners). When applying the approach used here in future work, this issue could be mitigated by regularly taking and storing (standardised) photographs for plumage scoring. Doing so would then create a longitudinal record of a bird's plumage condition representative of its long-term state and facilitate objective identification of changes over time. Despite the current limitations, however, our results do indicate that photographs are a valid, if imperfect, method of independently assessing feather condition (in agreement with Honess *et al* 2005 and van Zeeland *et al* 2013), and also that owner-reporting of FDB seems reliable.

https://doi.org/10.7120/09627286.31.2.001 Published online by Cambridge University Press

Table 10 Inter-observer reliability scores calculated between Rater 2 and parrot owners, on a subset of 31 parrots whose owners report they have feather-damaging behaviour. Agreement is given as percentage and Cohen's kappa (κ, agreement between two scores after accounting for agreement purely by chance: Cohen 1960; McHugh).

	Rater 2		
	Agreement	**Cohen's kappa**	**n**
Any feather damage?	83.9%	$\kappa = -0.05, Z = -0.39, P = 0.70$	31
Specific body parts			
Head	87.1%	$\kappa = 0.30, Z = 2.32, P = 0.02$	31
Throat/neck	64.5%	$\kappa = 0.36, Z = 2.61, P = 0.01$	31
Chest	83.9%	$\kappa = 0.62, Z = 3.47, P < 0.001$	31
Back	61.3%	$\kappa = 0.19, Z = 1.06, P = 0.29$	31
Wings (dorsal surface)	58.1%	$\kappa = 0.21, Z = 1.42, P = 0.16$	31
Wings (ventral surface)	40.0%	$\kappa = -0.36, Z = -0.91, P = 0.36$	5
Tail	82.1%	$\kappa = 0.58, Z = 3.19, P < 0.01$	28
Legs	80.6%	$\kappa = 0.61, Z = 3.39, P < 0.001$	31
Feather types			
Down feathers	80.6%	$\kappa = 0.61, Z = 3.43, P < 0.001$	31
Covert feathers	74.2%	$\kappa = 0.42, Z = 2.87, P < 0.01$	31
Primary/secondary flight feathers	78.6%	$\kappa = 0.48, Z = 2.57, P = 0.10$	28
Tail feathers	85.7%	$\kappa = 0.65, Z = 3.5, P < 0.001$	28
Blood feathers	36.4%	$\kappa = 0.07, Z = 0.64, P = 0.52$	11
Mature feathers	78.9%	$\kappa = -0.09, Z = -0.45, P = 0.66$	19
Other			
Skin damage	93.3%	$\kappa = 0.63, Z = 3.73, P < 0.001$	30
Severity (0–3)	54.8%	$\kappa = 0.38, Z = 3.31, P < 0.001$	31
Mean agreement	72.0%		

κ scores are interpreted as follows: < 0.21 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–0.99 = almost perfect; 1 = perfect. $P < 0.05$ indicates that the rater and owners scores agree more than would be expected by chance.

The greater proportion of feather damage being reported by raters than owners might be explained by owner *under*-reporting and/or rater *over*-reporting. For example, damage might indeed be present, including that resulting from FDB, but as yet undiagnosed by the owners (eg Pollard *et al* 2017; Malalana *et al* 2019). As demonstrated elsewhere, owners vary in the ability to correctly identify issues such as, for instance, obesity (Courcier *et al* 2011; Potter *et al* 2016; Morrison *et al* 2017), and the same could be true for feather condition. While we requested owners report presence of *any* damaged plumage, ie not only that due to FDB, FDB is a well-known parrot behavioural problem, which most owners would likely be aware of even if their bird were unaffected. Therefore, prior knowledge of FDB may have biased owners towards *only* positively reporting damage resulting from FDB and excluding reporting damage and/or loss caused by other means. Raters, in contrast, had no information about the behaviour of each bird prior to it being photographed, which could perhaps lead to raters incorrectly recording damage. Moulting, preening and bathing can all affect feather appearance, as can abrasion against the cage and/or nest-box (indeed this was noted anecdotally by raters as a potential explanation for cases in which feathers had a 'worn' appearance rather than looking chewed or entirely removed). Future studies might mitigate potential rater over-reporting by training raters, and by allowing owners to indicate whether their bird had performed specific, relevant, behaviours immediately prior to it being photographed. Similarly, information about moulting could also be collected. In primates, such as rhesus macaques (*Macaca mulatta*), alopecia (scored from coat quality) and hair-pulling are associated, in that hair-pullers are more likely to have hair loss (Lutz *et al* 2013). However, alopecia does also occur in the *absence* of hair-pulling indicating that, as with feather condition as examined here, alopecia is complex and can be caused by several factors (van Zeeland *et al* 2009; Lutz *et al* 2013).

Part of the rationale underlying our current study was to assess whether owner reports of plumage condition, and of FDB, are valid, as the latter is a key outcome variable in our previous cross-species comparative work (see Mellor *et al* 2021). While owners and raters generally agreed on birds' feather condition here, this agreement was not totally consistent and there seems some indication that, if anything, owners may have under-reported FDB. However, that parrots with FDB were scored as having a greater proportion of feather damage than those without, and had higher severity scores, is in line with what would be expected if owner reports are reliable. Equally reassuring for our comparative work (cf Mellor *et al* 2021), was that we did not find rater-owner agreement to differ between taxonomic groups. In other words, different taxa are equally affected by potential under-reporting of FDB by owners, and so this should not bias any comparative outcome variables.

Animal welfare implications

Considering welfare, it is useful to know that owner reports can be reliable in terms of gaining an understanding of the extent of a welfare-relevant problem. This is especially relevant given the mammalian bias in research focusing on captive wild animal welfare (Rose *et al* 2019). Thus, we encourage the welfare assessment of birds to become more commonplace, both to improve their welfare and caregivers' understanding of their animal's needs. An additional benefit of our work is evidence that photographs, being non-invasive and quick for owners to take and supply, can form a useful part of the diagnostic toolkit of welfare assessment, providing their limitations are considered. Finally, validating owner reports allows us to collect data, with minimal disturbance, from thousands of birds across dozens of countries, to identify species' traits and aspects of husbandry that predispose birds to, or protect them from, welfare problems (Mellor *et al* 2021).

Conclusion

In conclusion, we found that owner-reporting of feather condition of pet parrots is generally reliable, as judged by inter-observer reliability scores between owners and two raters, but not perfect — emphasising the importance of validation. We also found that birds with owner-reported FDB had a much greater proportion of feather damage, and their severity scores were also higher, as scored by owners and raters. Our finding that raters were more likely to report greater proportion of feather damage than owners might be explained by raters wrongly identifying newly preened or bathed feathers as being damaged, by very mild damage being undiagnosed by owners, or by owner-bias in the reporting of damage.

Declaration of interest

None. We confirm that the funder played no role in the study design; collection, analysis and interpretation of data; writing of the paper; and/or decision to submit to *Animal Welfare*.

Acknowledgements

We wish to thank all owners who took time to participate in our survey and provide images of their birds. For providing survey translations, thanks also to Livia Benato, Marco Ramirez Montes De Oca, Evelyn Maniaki, Oceane Schmitt, Gina van Pelt, and Lukas Ziegler. Many thanks to Paul Rose and another referee for valuable suggestions that improved the manuscript. Last, but not least, we thank UFAW for funding EM's postdoctoral position at the time this research was undertaken.

References

Bangdiwala SI 1985 A graphical test for observer agreement. In: Bishop YMM, Fienberg SE and Holland PW (eds) *Proceedings of the 45th International Statistical Institute Meeting* pp 307-308. SpringerLink: Berlin, Germany

Bergman L and Reinisch US 2006 Parrot vocalisation. In: Luescher AU (ed) *Manual of Parrot Behaviour* pp 219-223. Blackwell Publishing: Oxford, UK. https://doi.org/10.1002/9780470344651.ch19

Burn CC 2011 A vicious cycle: A cross-sectional study of canine tail-chasing and human responses to it, using a free video-sharing website. *PLoS One 6*: e26553. https://doi.org/10.1371/journal.pone.0026553

Carpenter CR 1934 A field study of the behavior and social relations of howling monkeys (*Alouatta palliata*). *Comparative Psychology Monographs 10*: 1-168

Courcier EA, Mellor DJ, Thomson RM and Yam PS 2011 A cross-sectional study of the prevalence and risk factors for owner misperception of canine body shape in first opinion practice in Glasgow. *Preventive Veterinary Medicine 102*: 66-74. https://doi.org/10.1016/j.prevetmed.2011.06.010

Crawley MJ 2013 *The R Book*. John Wiley & Sons Ltd: Chichester, UK

Cribari-Neto F and Zeileis A 2010 Beta Regression in R. *Journal of Statistical Software 34*: 1-24. https://doi.org/10.18637/jss.v034.i02

Finnegan SL, Volk HA, Asher L, Daley M and Packer RMA 2020 Investigating the potential for seizure prediction in dogs with idiopathic epilepsy: owner-reported prodromal changes and seizure triggers. *Veterinary Record 187*: 152. https://doi.org/10.1136/vr.105307

Gamer M, Lemon J, Fellows I and Singh P 2019 *irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1*. www.CRAN.R-project.org/package=irr

Garner JP, Meehan CL, Famula TR and Mench JA 2006 Genetic, environmental, and neighbor effects on the severity of stereotypies and feather picking in orange-winged Amazon parrots (*Amazona amazonica*): An epidemiological study. *Applied Animal Behaviour Science 96*: 153-168. https://doi.org/10.1016/j.applanim.2005.09.009

Giammarino M, Mattiello S, Battini M, Quatto P, Battaglini LM, Vieira ACL, Stilwell G and Renna M 2021 Evaluation of inter-observer reliability of animal welfare indicators: Which is the best index to use? *Animals 11*: 1445. https://doi.org/10.3390/ani11051445

Greenwell PJ and Montrose VT 2017 The gray matter: Prevention and reduction of abnormal behavior in companion gray parrots (*Psittacus erithacus*). *Journal of Veterinary Behavior 20*: 44-51. https://doi.org/10.1016/j.jveb.2017.06.005

https://doi.org/10.7120/09627286.31.2.001 Published online by Cambridge University Press

Gwet KL 2008 Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology 61*: 29-48. https://doi.org/10.1348/000711006X126600

Harrison GJ and Harrison LR 1986 *Clinical Avian Medicine and Surgery, Including Aviculture*. Saunders: Philadelphia, US

Honess PE, Gimpel JL, Wolfensohn SE and Mason GJ 2005 Alopecia scoring: The quantitative assessment of hair loss in captive macaques. *Alternatives to Laboratory Animals 33*: 193-206. https://doi.org/10.1177/026119290503300308

Lutz CK, Coleman K, Worlein J and Novak MA 2013 Hair loss and hair-pulling in rhesus macaques (*Macaca mulatta*). *Journal of the American Association for Laboratory Animal Science 52*: 454-457

Malalana F, McGowan TW, Ireland JL, Pinchbeck GL and McGowan CM 2019 Prevalence of owner-reported ocular problems and veterinary ocular findings in a population of horses aged ≥15 years. *Equine Veterinary Journal 51*: 212-217. https://doi.org/10.1111/evj.13005

Martin P and Bateson P 2007 *Measuring Behaviour: An Introductory Guide*. Cambridge University Press: Cambridge, UK. https://doi.org/10.1017/CBO9780511810893

Mason G 2006 Stereotypic behaviour in captive animals: Fundamentals and implications for welfare and beyond, In: Mason G and Rushen J (eds) *Stereotypic Animal Behaviour: Fundamentals and Applications to Welfare Second Edition* pp 325-367. CAB International: Wallingford, UK. https://doi.org/10.1079/9780851990040.0325

McBride SD and Long L 2001 Management of horses showing stereotypic behaviour, owner perception and the implications for welfare. *Veterinary Record 148*: 799-802. https://doi.org/10.1136/vr.148.26.799

Meehan CL, Garner JP and Mench JA 2003a Isosexual pair housing improves the welfare of young Amazon parrots. *Applied Animal Behaviour Science 81*: 73-88. https://doi.org/10.1016/S0168-1591(02)00238-1

Meehan CL, Garner JP and Mench JA 2004 Environmental enrichment and development of cage stereotypy in orange-winged Amazon parrots (*Amazona amazonica*). *Developmental Psychobiology 44*: 209-218. https://doi.org/10.1002/dev.20007

Meehan CL, Millam JR and Mench JA 2003b Foraging opportunity and increased physical complexity both prevent and reduce psychogenic feather picking by young Amazon parrots. *Applied Animal Behaviour Science 80*: 71-85. https://doi.org/10.1016/S0168-1591(02)00192-2

Mellor E 2014 *Can we use biological risk factors for stereotypic behaviour in parrots to predict husbandry risk factors?* MSc Thesis, School of Biological Sciences, Plymouth University, Plymouth, UK

Mellor E, McDonald Kinkaid H and Mason G 2018 Phylogenetic comparative methods: Harnessing the power of species diversity to investigate welfare issues in captive wild animals. *Zoo Biology 37*: 369-388. https://doi.org/10.1002/zoo.21427

Mellor EL, McDonald Kinkaid HK, Mendl MT, Cuthill IC, van Zeeland YRA and Mason GJ 2021 Nature calls: intelligence and natural foraging style predict poor welfare in captive parrots. *Proceedings of the Royal Society B: Biological Sciences 288*: 20211952. https://doi.org/10.1098/rspb.2021.1952

Morrison PK, Harris PA, Maltin CA, Grove-White D, Barfoot CF and Argo CM 2017 Perceptions of obesity and management practices in a UK population of leisure-horse owners and managers. *Journal of Equine Veterinary Science 53*: 19-29. https://doi.org/10.1016/j.jevs.2017.01.006

Müller DW, Lackey LB, Streich WJ, Fickel J, Hatt JM and Clauss M 2011 Mating system, feeding type and ex situ conservation effort determine life expectancy in captive ruminants. *Proceedings of the Royal Society B: Biological Sciences 278*: 2076-2080. https://doi.org/10.1098/rspb.2010.2275

Orosz SE 2006 Diagnostic work-up of suspected behavioural problems. In: Luescher AU (ed) *Manual of Parrot Behavior* pp 195-210. Blackwell Publishing: Oxford. UK. https://doi.org/10.1002/9780470344651.ch17

Pollard D, Wylie CE, Verheyen KLP and Newton JR 2017 Assessment of horse owners' ability to recognise equine laminitis: A cross-sectional study of 93 veterinary diagnosed cases in Great Britain. *Equine Veterinary Journal 49*: 759-766. https://doi.org/10.1111/evj.12704

Pollard D, Wylie CE, Verheyen KLP and Newton JR 2019 Identification of modifiable factors associated with owner-reported equine laminitis in Britain using a web-based cohort study approach. *BMC Veterinary Research 15*: 59. https://doi.org/10.1186/s12917-019-1798-8

Polverino G, Manciocco A and Alleva E 2012 Effects of spatial and social restrictions on the presence of stereotypies in the budgerigar (*Melopsittacus undulatus*): a pilot study. *Ethology Ecology & Evolution 24*: 39-53. https://doi.org/10.1080/03949370.2011.582045

Potter S, Bamford N, Harris P and Bailey S 2016 Prevalence of obesity and owners' perceptions of body condition in pleasure horses and ponies in south-eastern Australia. *Australian Veterinary Journal 94*: 427-432. https://doi.org/10.1111/avj.12506

Potter S, Bamford N, Harris P and Bailey S 2017 Incidence of laminitis and survey of dietary and management practices in pleasure horses and ponies in south-eastern Australia. *Australian Veterinary Journal 95*: 370-374. https://doi.org/10.1111/avj.12635

R Core Team 2021 *R:A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. www.R-project.org/

Rose PE, Brereton JE, Rowden LJ, de Figueiredo RL and Riley LM 2019 What's new from the zoo? An analysis of ten years of zoo-themed research output. *Palgrave Communications 5*: 128. https://doi.org/10.1057/s41599-019-0345-3

Schmid R, Doherr MG and Steiger A 2006 The influence of the breeding method on the behaviour of adult African grey parrots (*Psittacus erithacus*). *Applied Animal Behaviour Science 98*: 293-307. https://doi.org/10.1016/j.applanim.2005.09.002

van Zeeland YRA, Bergers MJ, van der Valk L, Schoemaker NJ, and Lumeij JT 2013 Evaluation of a novel feather scoring system for monitoring feather damaging behaviour in parrots. *The Veterinary Journal 196*: 247-252. https://doi.org/10.1016/j.tvjl.2012.08.020

van Zeeland YRA, Spruit BM, Rodenburg TB, Riedstra B, van Hierden YM, Buitenhuis B, Korte SM and Lumeij JT 2009 Feather damaging behaviour in parrots: A review with consideration of comparative aspects. *Applied Animal Behaviour Science 121*: 75-95. https://doi.org/10.1016/j.applanim.2009.09.006

Viera AJ and Garrett JM 2005 Understanding interobserver agreement: the kappa statistic. *Family Medicine 37*: 360-363