Intra- and inter-observer reliability of two different feather scoring methods in Grey parrots (*Psittacus erithacus erithacus*)

Madeleine Bergers

Keywords

Feather damaging behavior
Grey parrots
Feather scoring method by Meehan et al.
Feather scoring method by Van Zeeland
Reliability

Abstract

Feather damaging behavior is one of the most challenging behavioral problems in captive parrots. Several therapeutic options have been proposed and to evaluate the effectiveness of these treatments a reliable and valid way of evaluating feather damaging behavior is needed. The aim of this study was to examine the intra- and inter-observer reliability of the feather scoring method by Meehan and the feather scoring method by Van Zeeland. Photographs of 15 Grey parrots were selected. For the intra-observer reliability 10 parrots were assessed twice by 5 students and 6 specialists. For the inter-observer reliability 10 students and 7 specialists have assessed the photographs of all the 15 parrots. Intraclass correlation coefficients (ICC[2,1] and ICC[2,k]) with 95% confidence intervals, standard error of measurement and the minimum difference to be considered real were calculated. An independent-samples t-test was conducted to compare ICC [2,1] and [2,k] values for the intra- and inter-observer reliability of both scoring methods. The intra- and inter-observer reliability was very high for both scoring methods. The feather scoring method by Van Zeeland showed a significantly higher ICC[2,1] and ICC[2,k] for the intra-observer reliability and a significantly higher ICC[2,1] for the inter-observer reliability. This study demonstrated that the feather scoring method by Van Zeeland showed a significantly higher reliability than the feather scoring method by Meehan.

Introduction

Feather damaging behavior (also referred to as feather picking or feather plucking) is one of the most challenging behavioral problems in captive parrots. An estimated 10% parrots living in captivity show signs of feather damaging behavior (Grindlinger and Ramsay 1991). Feather damaging behavior can be any form of mutilation of the feathers by the beak, including chewing and plucking (Harrison 1986). It has been noted that birds most often self-pluck easily accessible feathers in the regions of the chest, ventral wings and inner thighs. The covert and down feathers are often affected primarily, though damage to the flight and tail feathers has also been described (Nett and Tully Jr. 2003). The total absence of normal feather growth on areas inaccessible to self-plucking excludes feather damaging behavior from the differential diagnosis of feather loss (Bordnick, Thyer et al. 1994). Noninfectious causes, including social, environmental, genetic and neurobiological factors, and infectious causes, including parasites, protozoan, bacterial, viral or fungal agents, have all been mentioned as underlying cause of feather damaging behavior (Mertens 1997; Nett and Tully Jr. 2003; van Zeeland, Spruit et al. 2009). Conversely, feather damaging behavior may develop or persist in the absence of any evident medical cause (Lumeij and Hommers 2008). Several therapeutic options have been proposed, such as changes in housing and environment (interesting toys, foraging opportunities and acoustic stimulation), topical

application of foul tasting substances on the feathers, the use of plastic collars, and pharmacological treatment (Mertens 1997; Cooper and Harrison 2007; Lumeij and Hommers 2008). The effectiveness of these treatments is still unclear and a good method of measuring is desirable.

A reliable and valid way of evaluating feather damaging behavior is needed for monitoring patients and for research on feather damaging behavior. Various methods of evaluating feather damaging behavior have been proposed in the past. Van Hoek and King evaluated the behavior of the bird (Van Hoek and King 1997). This method is, however, very time consuming and therefore hardly feasible. Furthermore, the reliability of such a method is questionable, because feather damaging behavior is often hard to distinguish from normal grooming and may occur at night (Meehan, Millam et al. 2003). Alternatively, a plumage condition scoring method can be applied, by which the condition of the plumage can be determined at any given time. A change in score is an indirect indicator of the severity of feather damaging behavior. This method has been used for feather picking laying hens and was found practical and highly useful (Tauson, Kjaer et al. 2005; Kjaer, Glawatz et al. 2011).

The scoring methods

In recent years a feather scoring method has been developed by Meehan, Millam et al. (2003) to evaluate feather damaging behavior in parrots. This method, which was developed by surveying a colony of Orange-winged Amazons (n=76) and determining the prevalent patterns of feather loss and damage, was specifically developed for observing parrots from a distance. The plumage condition is evaluated using a 10-point scoring scale. Five separate body areas (chest/flank, back, legs, tail and dorsal side of the wings) are taken into account. The subscores of these body areas are combined to form a total score. (Meehan, Millam et al. 2003).

Based on the feather scoring method by Meehan et al. (2003), Van Zeeland et al. (2008) introduced an alternative feather scoring method. Similar to the feather scoring method by Meehan, different body parts are assessed separately. Unlike the feather scoring method by Meehan, the feather scoring method by Van Zeeland includes scoring of the ventral sides of wings (which are commonly involved), and uses an objective, numerical scale for the assessment of each body part, as well as a conversion factor for the relative size of each body part. In the feather scoring method by van Zeeland, two entities are distinguished: first body parts covered by the covert and down feathers (front of the body, back of the body, dorsal and ventral sides of the wings and legs); second the flight feathers (wing: 10 primaries and 10 secondaries, tail: 10 flight feathers also known as rectrices). The different body parts are scored first, using the percentage of down feathers removed and percentage of coverts damaged. These percentages are then multiplied by the relative body surface area of the corresponding body part. The mean surfaces of each body parts have been determined in six grey parrots (*Psittacus erithacus erithacus*) by (Van der Valk 2009) (Table 1). The number of damaged flight feathers is assessed subsequently. There are 50 flight feathers to be assessed with a maximum score of 100 points to be obtained. For each damaged feather 1 or 2 points will be deducted from the total score.

Table 1

Mean relative body surface of the body parts to be assessed in the feather scoring method by Van Zeeland.

Body part	Mean percentage (%) of the total body surface* with SD
Chest/neck/flank	25 ± 1.2
Back	17 ± 1.5
Ventral side of the wings (left and right)	20 ± 0.7
Dorsal side of the wings (left and right)	28 ± 2.2
Legs (left and right)	10 ± 1.2

Note. SD - standard deviation; * excluding surface area of the head and unfeathered parts of the legs.

Though the feather scoring method by Van Zeeland has been developed to improve the feather scoring method by Meehan, the reliability of both feather scoring methods is unknown. The question is whether this alternative method is more reliable and more practical to use than the feather scoring method by Meehan.

The aim of this methodical study was to determine the intra- and inter-observer reliability of the feather scoring method developed by Meehan et al. and the feather scoring method by Van Zeeland in Grey parrots (*Psittacus erithacus erithacus*) showing feather damaging behavior. Reliability is one of the conditions an instrument must meet in order to be clinically applicable. This study had the purpose to investigate which of the two scoring methods gives the most consistent results, both in the context of monitoring patients and research on feather damaging behavior, and to determine which of the scoring methods is most suitable for clinical use.

Methods

Subjects

The chest, back, legs, ventral surface of the wings, dorsal surface of the wings and tail of 24 Grey parrots (*Psittacus erithacus erithacus)* were photographed according to a standardized protocol during examination at the University Clinic for Companion Animals, Utrecht. The background of the photographs was removed using Photoshop CS5 for Windows to prevent bias due to recognition. A total of 15 parrots were selected based on a clear view of all body parts and a wide distribution in extent of feather damaging behavior. The width of the distribution was assessed by application of both feather scoring methods by 2 examiners, that did not participate in further scoring during this study.
10 parrots were randomly selected from the total of 15 parrots for evaluation of intra-observer reliability. All 15 parrots were used for evaluation of inter-observer reliability.

Examiners

Two groups of examiners were composed. 15 experienced examiners (referred to as 'specialists') and 17 unexperienced examiners (referred to a 'students') were asked to participate in this study. The experienced examiners consisted of avian diplomats from the European College of Zoological Medicine ('ECZM') and practicing avian veterinarians. Some of the specialists were familiar with both feather scoring methods. The unexperienced examiners consisted of veterinary students and had little or no experience in the use of feather scoring methods. For the trial of intra-observer reliability, 6 students and 6 specialists were randomly selected. The remaining 11 students and 9 specialists were assigned to participate in the inter-observer reliability trial.

The scoring methods

The feather scoring method developed by Meehan et al. (2003) uses a 10-point scoring method to assess the feather condition of five different body parts: chest / flank, back, legs, tail and wings. These body parts are scored on a scale ranging from 0 to 2 based on the most fitting description (Appendix A). A score of 0 points means all or most feathers are removed and skin damage is present. A score of 2 means all feathers are intact with little or no fraying or breakage (Meehan, Millam et al. 2003).
The feather scoring method by Van Zeeland et al. consists of two parts: the body parts covered by covert and down feathers (front side of the body, back, legs ad dorsal and ventral surfaces of the wings) and the flight feathers (wing: 10 primaries and 10 secondaries, tail: 10-12 feathers) (Appendix B). In the first part, the different body parts are scored using percentages down feathers removed, percentages covert feathers removed and presence of skin damage (yes /no). To calculate the total score, the scores of the different body parts are then multiplied by a conversion factor based on the relative body surface area of the evaluated body part. The second part of the alternative feather scoring method was not implemented in this study, because the assessment of damage to the flight feathers was restricted by insufficient spreading of these feathers on the photographs.

Training

To familiarize themselves with the scoring procedure, the examiners were asked to assess photographs of 3 parrots, prior to starting the trail. These scorings were performed in the same fashion as the following real scorings. The examiners were unaware that this session was not part of the real trial. The 3 series of photos used in the training session were not used in the subsequent sessions and the results of the scoring were not included in the study.

Design

For the intra-observer reliability all examiners (6 students and 6 specialists) were asked to assess photographs of 6 different body parts (front of the body, back of the body, dorsal and ventral sides of the wings, legs and tail) of 10 parrots twice. The 10 parrots were presented in sessions of 5 parrots each, because the assessment of 10 parrots in one session is very time consuming (Fig. 1). A time interval with a minimum of 7 days between the first (T1) and second scoring (T2) was used, in order to prevent recognition of the subject and its subsequent prior evaluation (Streiner and Norman 2003). The time between the two scorings had a maximum of 5 weeks. Random ordering of the subjects body parts was applied before the second scoring sessions to further prevent recognition (Sim and Wright 2005).

For the inter-observer reliability all examiners (11 students and 9 specialists) were asked to assess the photographs of all the 15 parrots, in 3 sessions of 5 parrots each (Fig. 1).

For both the intra- and inter-observer reliability the examiners were not able to revise their previously completed sessions.

The scoring sessions were administered on-line using lime-survey, an Open Source PHP web application to develop and publish surveys, and collect responses. The various components of the feather scoring method by Van Zeeland and the feather scoring method developed by Meehan et al. were converted into 1 to 4 questions for each body part. The questions concerned the best fitting description for the condition of the plumage and skin (feather scoring method by Meehan et al.), assessment of the percentage covert feathers and down feathers removed and the presence of skin damage (feather scoring method by Van Zeeland). Specific questions were asked for each body part (Appendix C). For every scoring session the parrots were renamed parrot number 1 to 5.

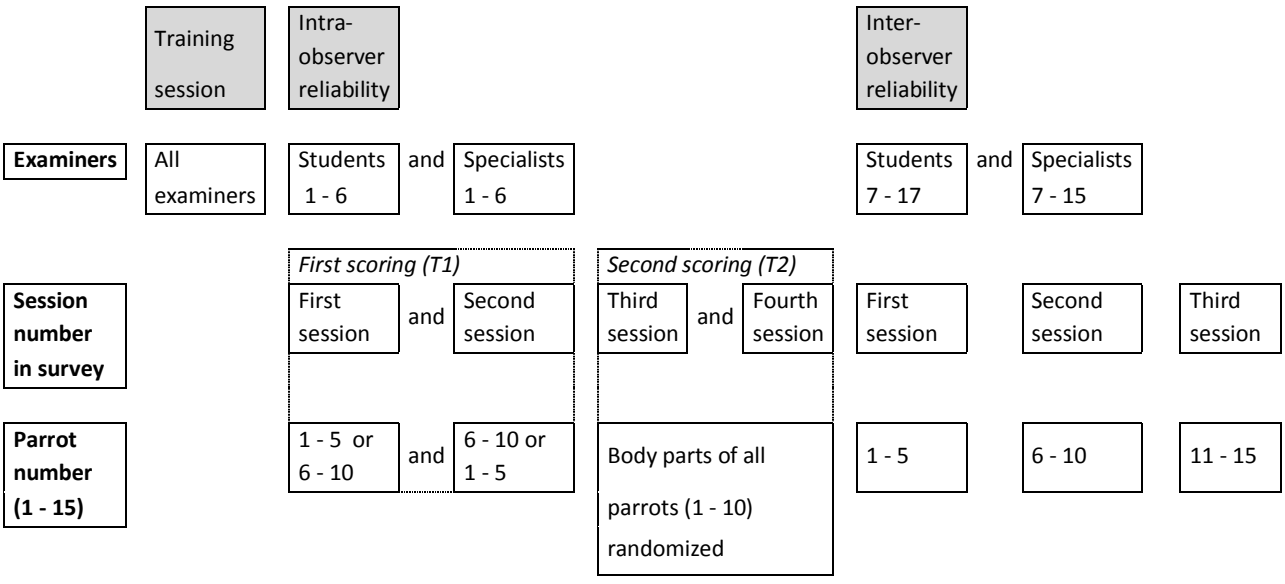

Figure 1. Diagram describing the trials, the examiners, the sessions conducted and the parrots scored in each session

Statistical analysis

A two-way model analysis of variance analysis (ANOVA) was performed for the inter- and intra-observer data. The F ratio was evaluated to compare the amount of systematic variance to the amount of random variance (Weir 2005). All variables were examined for normality using the Shapiro-Wilk test (Field 2009). The test results were non-significant, thus the assumption of normality was not violated.

For the intra-observer and inter-observer reliability of the total feather score of the feather scoring method developed by Meehan et al. (2003), the weighted kappa would be the suitable analysis. The weighted kappa coefficient is an appropriate measure of reliability for ordinal data (Cohen 1968; Fleiss and Cohen 1973; Sim and Wright 2005; Mandrekar 2011). Because the weighted kappa is based on squared disagreement weights, it gives equivalent results to the ICC [2,1] (Fleiss and Cohen 1973; Shrout and Fleiss 1979; Schuster 2004).

Intra-observer and inter-observer relative reliability of the feather scoring method by Van Zeeland was assessed by calculating the Intraclass Correlation Coefficient ('ICC') (Shrout and Fleiss 1979). The application of the ICC is appropriate in cases when the data are measured on a continues scale (Fleiss and Cohen 1973; Eggen and Sanders 1993; Mandrekar 2011).

The ICC [2,1] was chosen because it allows for comparison of the reliability of both the feather scoring method developed by Meehan et al. and the feather scoring method by Van Zeeland.

For analysis of inter- and intra-observer reliability of both feather scoring methods, a two-way random effects model was used, referred to ICC [2,1] and ICC [2,k]. This model is suitable when all subjects are assessed by the same group of examiners, and the examiners are a random sample from a larger population of examiners that one wishes to generalize to other observers within a different population. When one is interested in generalization of the results to individuals, it is useful to determine ICC [2,1]. ICC [2,1] reflects the average agreement between one observer and any other. ICC [2,k] reflects the reliability of the average of a number of observations (also referred to as the reliability of the average rating) and can be generalized to other groups of observers. An increase of the number of observers will increase the ICC [2,k], because a larger sample size improves the estimate of the mean. Thus ICC [2,k] will always be higher than the ICC [2,1] (Shrout and Fleiss 1979).

Because of our interest in the reliability of the individual ratings and the reliability of the mean rating, ICC [2,1] and ICC [2,k] were used respectively (Shrout and Fleiss 1979). An absolute agreement definition was used for both ICC [2,1] and ICC [2,k], meaning that the observer variability was included in the denominator, because of the relevance of the systematic observer variability. This in contrast to the consistency definition, which excludes the observer variability from the denominator (McGraw and Wong 1996). Confidence intervals were set at 95% in all calculations.

An independent-samples t-test was conducted to compare ICC [2,1] and [2,k] values for the intra- and inter-observer reliability (of all the examiners) of both scoring methods. The standard error of measurement for running the t-test was calculated by dividing the difference between the upper en lower confidence interval of the ICC by two times the z-value of 1.96.

Absolute reliability was calculated using the standard error of measurement (SEM). The SEM is an indication of the precision of a individual score and is expressed in the same units as the measurements (Eliasziw, Young et al. 1994; Weir 2005). The SEM was derived from a two-way model ANOVA and calculated as the square root of the residual mean square (Bland and Altman 1996; Weir 2005). The inter-observer SEM for both the alternative feather scoring method and the feather scoring method developed by Meehan et al. was calculated using all scorings in the inter-observer trial, and the intra-observer SEM was calculated from the repeated ratings of 10 parrots for each examiner separately. The SEM for the inter-observer reliability has been used to determine the minimum difference to be considered real ($SEM*1.96*\sqrt{2}$). Any change in a parrot's score, greater than the minimum difference, below or above the previous score, reflects a real difference in 95% of the cases (Weir 2005).

The inter-observer data was plotted to visualize the size, range of differences and scoring distribution. For each parrot the scorings by the 15 examiners were shown, with mean and ±2SD.

The analysis was performed using SPSS (Statistical Package for Social Sciences) version 16.0 for Windows (SPSS, Michigan Avenue, Chicago, IL, USA).

Results

Response rates for the study were 88% (15 of 17) for the students group and 87% (13 of 15) for the specialists group. The main reason for loss of data was non-compliance. 5 students and 6 specialists participated in the intra-observer reliability trial. 10 students and 7 specialists participated in the inter-observer reliability trial.

For the feather scoring method by Meehan et al. the 15 parrots included in this study had a mean score ranging from 1.5 - 8.5 points in the inter-observer reliability trial (Fig. 2). For the intra-observer reliability trial the 10 parrots scored a mean of 2.5 - 8.3 points. In the inter-observer reliability trial of the feather scoring method by Van Zeeland the mean scores of the 15 parrots ranged 24 - 93 points (Fig. 3). For the intra-observer reliability trial the 10 parrots scored a mean of 32 - 86 points. The wide distribution of scores makes the population of parrots suitable for this study.

Table 2

Intra-observer reliability of the total feather score of the feather scoring method by Meehan et al.

Examiner	Mean first scoring	Mean second scoring	Mean difference	ICC [2,1] (95% CI)	ICC [2,k] (95% CI)	SEM	Minimum change
Student 1	5.48	5.15	0.32	0.92 (0.71 - 0.98)	0.96 (0.83 - 0.99)	0.55	1.52
Student 2	5.83	5.85	-0.02	0.97 (0.89 - 0.99)	0.99 (0.94 - 1.00)	0.36	0.99
Student 3	6.00	5.53	0.48	0.90 (0.62 - 0.98)	0.95 (0.76 - 0.99)	0.54	1.49
Student 4	6.28	5.63	0.65	0.83 (0.45 - 0.96)	0.91 (0.62 - 0.98)	0.80	2.21
Student 5	4.75	5.10	-0.35	0.90 (0.64 - 0.97)	0.95 (0.78 - 0.99)	0.50	1.39
Specialist 1	6.83	6.65	0.18	0.92 (0.74 - 0.98)	0.96 (0.85 - 0.99)	0.53	1.48
Specialist 2	5.40	5.60	-0.20	0.85 (0.51 - 0.96)	0.92 (0.68 - 0.98)	0.75	2.09
Specialist 3	5.55	5.63	-0.08	0.90 (0.65 - 0.98)	0.95 (0.79 - 0.99)	0.62	1.71
Specialist 4	6.03	5.78	0.25	0.94 (0.79 - 0.99)	0.97 (0.89 - 0.99)	0.39	1.08
Specialist 5	5.63	5.33	0.30	0.93 (0.75 - 0.98)	0.96 (0.86 - 0.99)	0.47	1.30
Specialist 6	5.13	5.20	-0.08	0.95 (0.80 - 0.99)	0.97 (0.89 - 0.99)	0.46	1.27
All examiners	-	-	-	0.91 (0.89 - 0.93)	0.95 (0.94 - 0.97)	0.54	1.50
All students	-	-	-	0.90 (0.86 - 0.95)	0.95 (0.92 - 0.98)	0.55	1.52
All specialists	-	-	-	0.92 (0.89 - 0.94)	0.96 (0.94 - 0.97)	0.54	1.49

Note. ICC - intraclass correlation coefficient; SEM - standard error of measurement; CI - confidence interval.

Table 3

Inter-observer reliability of the total feather score of the feather scoring method by Meehan et al.

Examiners	ICC [2,1] (95% CI)	ICC [2,k] (95% CI)	SEM	Minimum change
All examiners	0.83 (0.70 - 0.93)	0.99 (0.96 - 1.00)	0.72	1.99
All students	0.82 (0.67 - 0.93)	0.98 (0.95 - 0.99)	0.70	1.94
All specialists	0.82 (0.65 - 0.93)	0.97 (0.93 - 0.99)	0.76	2.09

Note. ICC - intraclass correlation coefficient; SEM - standard error of measurement; CI - confidence interval.

Relative reliability, ICC

Intra-observer reliability and inter-observer reliability for the feather scoring method by Meehan et al. and the feather scoring method by Van Zeeland were high. The results of the intra- and inter-observer reliability for the feather scoring method by Meehan et al. for all the examiners, the students and the specialists are show in Table 2 and Table 3 respectively. For the feather scoring method by Van Zeeland the results are shown in Table 4 and Table 5.

For the feather scoring method by Meehan et al. the intra-observer ICC [2,1] was 0.91 and the intra-observer ICC [2,k] was 0.95. The intra-observer ICC [2,1] for the students was 0.90, the ICC [2,k] for the students was 0.95. The intra-observer ICC [2,1] for the specialists was 0.92, the ICC [2,k] for the specialists was 0.96. The ICC

[2,1] for the first scoring for all the 10 intra-observer reliability trial examiners was 0.81, while the ICC [2,1] for the second scoring was 0.85 (results not shown in tables).
For the feather scoring method by Meehan et al. the inter-observer ICC [2,1] was 0.83 and the inter-observer ICC [2,k] was 0.99. The inter-observer ICC [2,1] for the students was 0.82, the ICC [2,k] for the students was 0.98. The inter-observer ICC [2,1] for the specialists was 0.82, the ICC [2,k] for the specialists was 0.97.

For the feather scoring method by Van Zeeland the intra-observer ICC [2,1] was 0.93, the intra-observer ICC [2,k] was 0.96. The intra-observer ICC [2,1] for the students was 0.91, the ICC [2,k] for the students was 0.95. For the specialists the intra-observer ICC [2,1] was 0.95, the ICC [2,k] was 0.97. The ICC [2,1] for the first scoring for all the 10 intra-observer reliability trial examiners was 0.84, the ICC [2,1] for the second scoring was 0.87 (results not shown in tables).
The inter-observer ICC [2,1] was 0.89, the inter-observer ICC [2,k] was 0.99. For the students the inter-observer ICC [2,1] was 0.88, the ICC [2,k] was 0.99. The inter-observer ICC [2,1] for the specialists was 0.91 and the ICC [2,k] for the specialists was 0.99.

Table 4
Intra-observer reliability of the total feather score of the feather scoring method by Van Zeeland

Examiner	Mean first scoring	Mean second scoring	Mean difference	ICC [2,1] (95% CI)	ICC [2,k] (95% CI)	SEM	Minimum change
Student 1	62.41	61.25	1.16	0.95 (0.81 -0.99)	0.97 (0.89 - 0.99)	4.38	12.14
Student 2	62.89	64.34	-1.44	0.98 (0.94 - 1.00)	0.99 (0.96 - 1.00)	2.59	7.17
Student 3	66.22	61.54	4.69	0.93 (0.61 - 0.99)	0.97 (0.76 - 0.99)	4.01	11.10
Student 4	68.82	62.60	6.22	0.88 (0.44 - 0.97)	0.93 (0.61 - 0.99)	5.32	14.75
Student 5	56.79	59.43	-2.65	0.80 (0.39 - 0.95)	0,89 (0.56 - 0.97)	8.43	23.36
Specialist 1	69.75	70.47	-0.72	0.94 (0.78 - 0.99)	0,97 (0.88 - 0,99)	4.73	13.12
Specialist 2	61.97	61.28	0.69	0.98 (0.93 - 1.00)	0,99 (0.96 - 1.00)	2.59	7.18
Specialist 3	61.74	64.34	-2.60	0.95 (0.79 - 0.99)	0.98 (0.88 - 0.99)	2.85	7.90
Specialist 4	63.59	60.36	3.24	0.92 (0.71 - 0.98)	0.96 (0.83 - 0.99)	4.25	11.77
Specialist 5	56.28	54.21	2.08	0.93 (0.77 - 0.98)	0.97 (0.87 - 0.99)	4.82	13.35
Specialist 6	61.76	60.28	1.48	0.95 (0.83 - 0.99)	0.98 (0.91 - 0.99)	3.59	9.95
All examiners	-	-	-	0.93 (0.90 - 0.96)	0.96 (0.95 - 0.98)	4.32	11.98
All students	-	-	-	0.91 (0.85 - 0.97)	0.95 (0.92 - 1.00)	4.94	13.70
All specialists	-	-	-	0.95 (0.93 - 0.96)	0.97 (0.96 - 0.98)	3.80	10.54

Note. ICC - intraclass correlation coefficient; SEM - standard error of measurement; CI - confidence interval.

Table 5
Inter-observer reliability of the total feather score of the feather scoring method by Van Zeeland

Examiners	ICC [2,1] (95% CI)	ICC [2,k] (95% CI)	SEM	Minimum change
All examiners	0.89 (0.80 - 0.95)	0.99 (0.99 - 1.00)	5.82	16.14
All students	0.88 (0.77 - 0.95)	0.99 (0.97 - 1.00)	5.86	16.25
All specialists	0.91 (0.81 - 0.96)	0.99 (0.97 - 0.99)	5.71	15.84

Note. ICC - intraclass correlation coefficient; SEM - standard error of measurement; CI - confidence interval.

The results of the independent t-tests for comparison of the intra- and inter-observer reliability of all the examiners for both feather scoring methods are show in Table 6 and Table 7 respectively.
A significant difference was found between the feather scoring method by Meehan and the feather scoring method by Van Zeeland in both the ICC[2,1] intra-observer reliability ($t(478)= -17.187$, $p<0.001$) and the ICC[2,k] intra-observer reliability ($t(478)= -13.693$, $p<0.001$).
For the inter-observers reliability a significant difference between the feather scoring method by Meehan and the feather scoring method by Van Zeeland was found in ICC[2,1] ($t(598)= -14.808$, $p<0.001$). Comparison of ICC[2,k] for the inter-observer reliability of all examiners revealed no significant difference $t(598)= 0.000$, ns.

Table 6
Comparison of the intra-observer reliability of all examiners for both feather scoring methods

Intraclass correlation coefficient	Feather scoring method by Meehan	Feather scoring method by Van Zeeland	T-test	
ICC [2,1] (95% CI)	0.91 (0.89 - 0.93)	0.93 (0.90 - 0.96)	*	t= -17.187
ICC [2,k] (95% CI)	0.95 (0.94 - 0.97)	0.96 (0.95 - 0.98)	*	t= -13.693

Note. * P< 0.001.

Table 7
Comparison of the inter-observer reliability of all examiners for both feather scoring methods

Intraclass correlation coefficient	Feather scoring method by Meehan	Feather scoring method by Van Zeeland	T-test	
ICC [2,1] (95% CI)	0.83 (0.70 - 0.93)	0.89 (0.80 - 0.95)	*	t= -14.808
ICC [2,k] (95% CI)	0.99 (0.96 - 1.00)	0.99 (0.99 - 1.00)	**	t= 0.000

Note. * P< 0.001; ** P> 0.05 (non significant)

Absolute reliability, measurement error and minimum change

For the feather scoring method by Meehan et al. the standard error of measurement (SEM) was in the range of 0.36 - 0.80 for the intra-observer reliability. The SEM for the intra-observer reliability for the students ranged from 0.36 - 0.80 and for the specialists from 0.39 - 0.75. The total SEM for the intra-observer reliability was 0.54, 0.55 for the students and 0.54 for the specialists (Table 2).
The total SEM for the inter-observer reliability was 0.72 for the total feather score. The SEM for the inter-observer reliability for the students was 0.70, and for the specialists 0.76 (Table 3).
The minimum change for the inter-observer reliability was 1.99 (Table 3).

The standard error of measurement (SEM) for the intra-observer reliability for the feather scoring method of Van Zeeland was found to be in the range of 2.59 - 8.43. The SEM for the intra-observer data for the students ranged from 2.59 - 8.43 and for the specialists from 2.59 - 4.82. The total SEM for the intra-observer data was 4.32, 4.94 for the students and 3.80 for the specialists (Table 4).
For the inter-observer data the total SEM was 5.82. The SEM for the inter-observer data for the students was 5.86 and for the specialists 5.71 (Table 5).
The minimum change for the inter-observer reliability was 16.14 (Table 5).

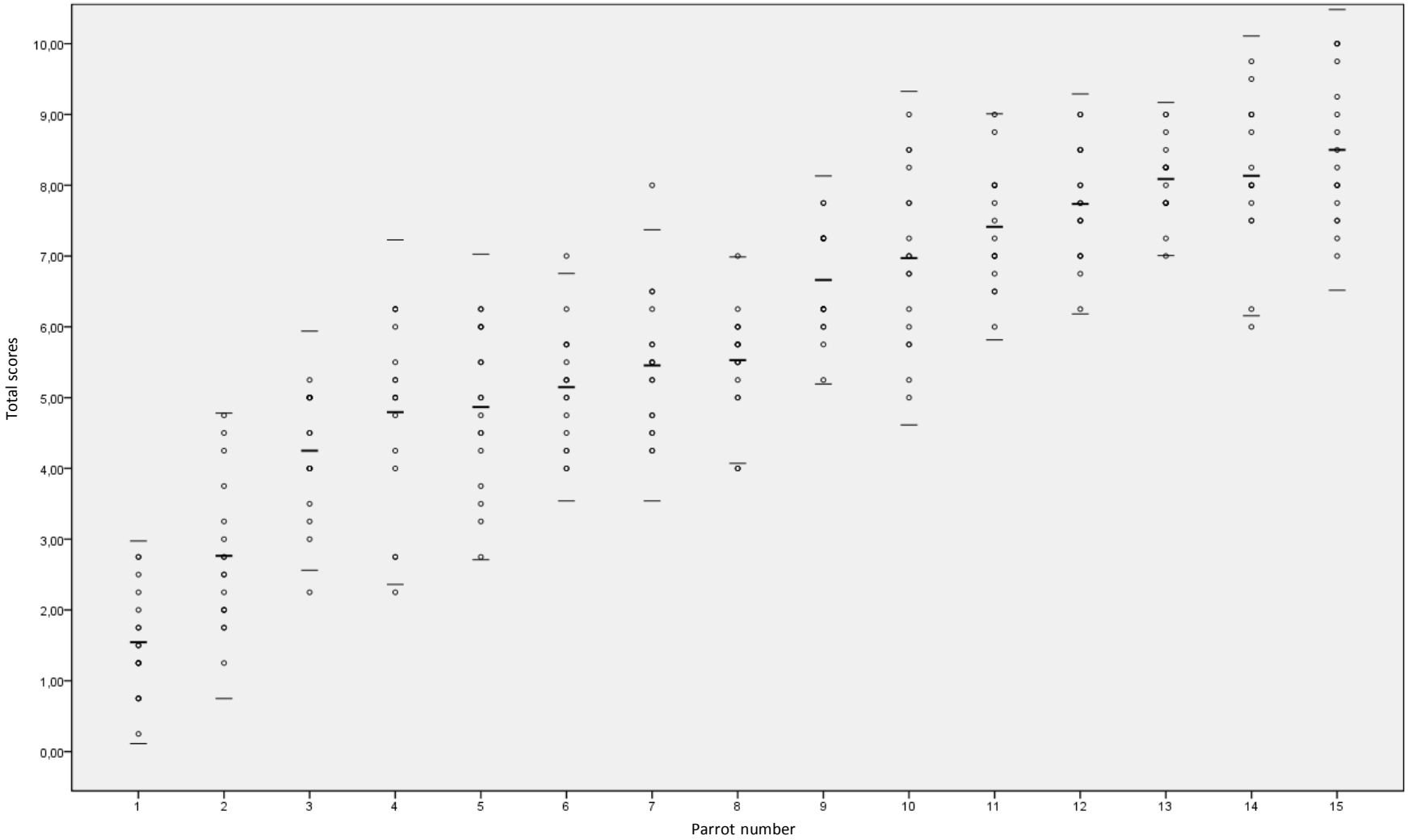

Figure 2. The inter-observer data of the feather scoring method by Meehan et al. For each parrot the scorings by the 17 examiners are shown, with mean (bold line) and ±2SD (solid lines).

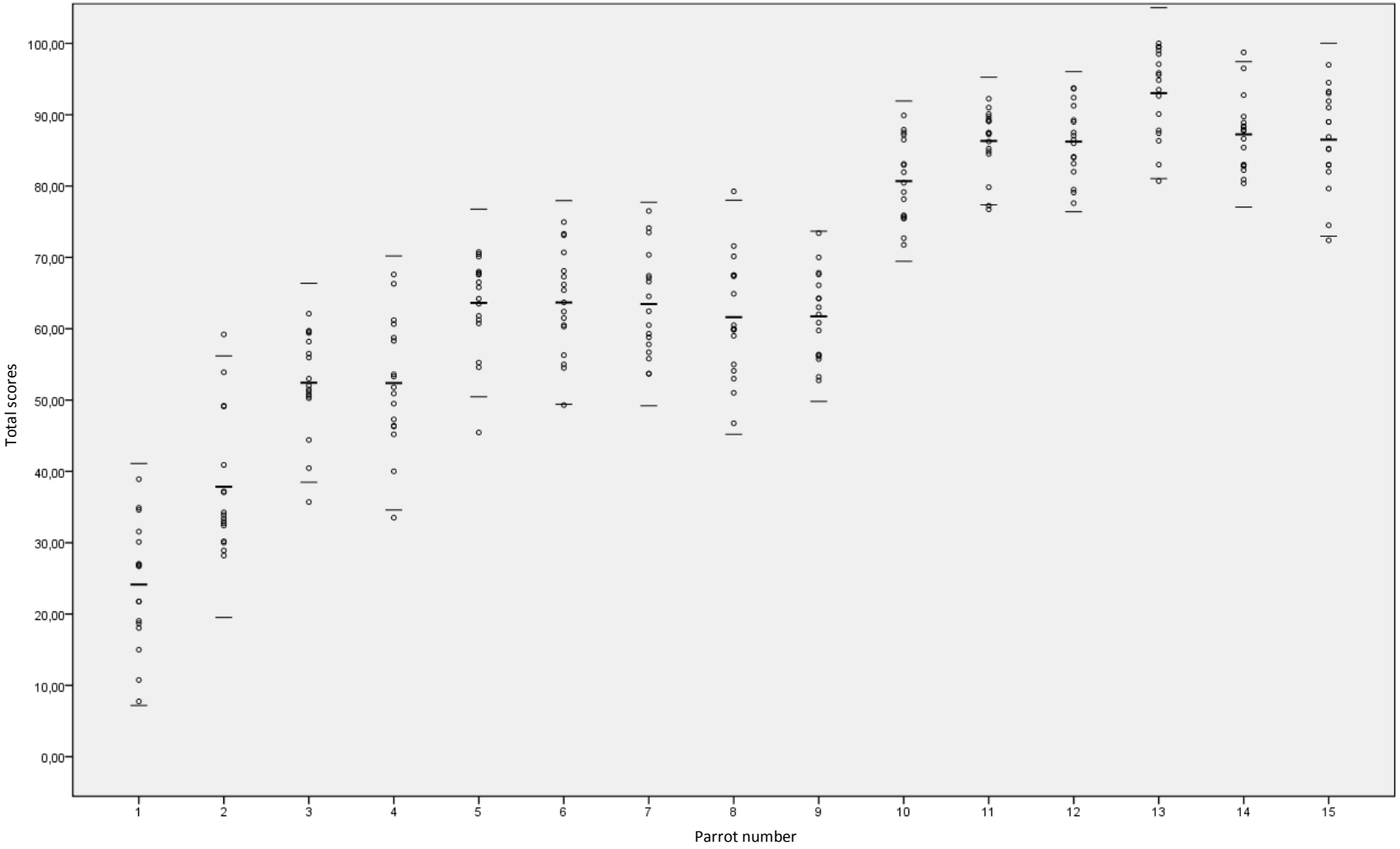

Figure 3. The inter-observer data of the feather scoring method by Van Zeeland. For each parrot the scorings by the 17 examiners are shown, with mean (bold line) and ±2SD (solid lines).

Discussion

The objective of this study was to investigate intra- and inter-observer reliability of the feather scoring method by Meehan et al. and the feather scoring method by Van Zeeland. Reliability was high for both feather scoring methods, with the feather scoring method by Van Zeeland showing a significantly higher ICC[2,1] and ICC[2,k] for the intra-observer reliability for all examiners and a significantly higher ICC[2,1] for the inter-observer reliability for all examiners.

Relative reliability assessed by the intraclass correlation coefficient concerns the consistency of the position or rank of a subject in a group relative to others, i.e. how well can subjects be distinguished from another, despite measurement errors (Weir 2005; de Vet, Terwee et al. 2006). There are no criteria for judging acceptability of intraclass correlation coefficients. The ICC varies depending on which version of the ICC is used and on the between subjects variability (Rankin and Stokes 1998; Weir 2005; Costa-Santos, Bernardes et al. 2011).

A training session was used, because the veterinary students and specialists had different levels of experience using both feather scoring methods and examining parrots in general. Both the students and the specialists showed high reliability for the intra-observer reliability separately and in general, and for the inter-observer reliability in general.

The absolute reliability was assessed by the standard error of measurement (SEM). The SEM estimates how repeated scorings are distributed around the true score. The smaller the SEM, the higher the consistency of two scorings, the more precise the score is (Bland and Altman 1996; Weir 2005).

The SEM for the inter-observer reliability of the total feather score for feather scoring method by Meehan et al. was 0.72. The minimum difference to be considered real would thus be $0.72*1.96*\sqrt{2}$ for 95% of the assessments. This equals 1.99, which constitutes 19.9% of the scale 0 - 10. The SEM for the intra-observer reliability for the total feather score ranged 0.39 - 0.80. The examiners with the lowest and the highest SEM were both students. The standard error of measurement for the inter-observer reliability of the total score for the alternative feather scoring method was 5.82. The minimum difference to be considered real for 95% of the assessments would be 16.14, which constitutes 16.14% of the scale 0 - 100. The SEM for the intra-observer reliability was ranged 2.59 - 8.43. The highest SEM belonged to a student, the lowest to both a student and specialist. The average of the SEM for the students and specialists were 4.94 and 3.80, with an average of 4.32. These SEM values are easier to apply to individual scorings than the ICC numbers and are useful for score interpretation in clinical practice.

The plotting of the inter-observer reliability data was used in addition to the ICC values because neither test alone provides sufficient information about reliability. In comparison with the alternative feather scoring method, for the feather scoring method by Meehan there was a wider distribution of scores given by the examiners for each parrot.

Reliability of the feather scoring method by *Meehan et al.* (2003) has been evaluated by Meehan, Millam et al. (2003) and showed an inter-observer reliability coefficient (Pearson correlation coefficient) of 0.76.

A study on the reliability of the feather scoring method by Meehan by Van der Horst (2007, unpublished) showed an coefficient of variation of 0.15 for the intra-observer reliability and an coefficient of variation of 0.05 for the inter-observer reliability. The outcome of this study may have been biased by a poor quality of the photos and some uncertainty about the criteria. In evaluating the scoring method, it was noted that the ventral sides of the wings are not scored, there are not enough scoring categories for each body part and that all body parts have an equal share in the total score (Van der Horst, 2007, unpublished).

A study performed by (Van der Valk 2009) showed a Cohen's Kappa-coefficient (K_W) of 0.76 for the intra-observer reliability and a K_W of 0.50 for the inter-observer reliability for the feather scoring method by Meehan. For the alternative feather scoring method a K_W of 0.87 for the intra-observer reliability and a K_W of 0.78 for the inter-observer reliability was found. The limitations of this study were the low number of parrots assessed, the absence of parrots with scores in the ranges of 0 - 4 and 7 - 8, insufficient explanation of the difference between any fraying and true feather damaging behavior and limited quality of the photographs (insufficient spreading of the flight feathers).

In this study the various components of the alternative feather scoring method and the feather scoring method developed by Meehan et al. were combined into a group of questions per parrot body part. These questions were repeatedly presented to the examiner in the same order. It is not clear whether scoring using the feather scoring method by Meehan et al. (which was always placed first in the question group) is of influence on the scoring results by the alternative feather scoring method.
In the clinic assessment of real birds takes place. In this study we used photographs with a clear view of all body parts and a wide distribution in extent of feather damaging behavior. Using this method we ensured that the variability of the scoring was not due to variability in a parrot's feather damaging behavior over time or the presentation of the parrot during scoring. However, assessment of the flight feathers was not implemented in this study due to insufficient spreading of the flight feathers on the photographs. Therefore, to evaluate the feather scoring method by Van Zeeland for the flight feathers, it is preferable to use real birds because of more sufficiently spreading of the flight feathers, especially of the tail. 'In vivo' assessment may also be more reliable because of a better visualization around the parrot.
Aside from reliability and validity, feasibility is a determining factor for clinical applicability of any scoring method. The feather scoring method by Van Zeeland uses an objective, numerical scale for the assessment of all the different body parts, which makes scoring faster and more efficient. It is easy to combine the scores for down- and covert feathers into one overall score. The feather scoring method by Meehan lacks scoring of the ventral sides of wings and uses a 10-point descriptive scale. Choosing a description that fits best is less efficient than working with a numerical scale and adequate description is missing occasionally. The feather scoring method by Van Zeeland seems more practical in use than the feather scoring method by Meehan.

Conclusions

This study of intra- and inter-observer reliability of the feather scoring method by Meehan et al. (2003) and the alternative feather scoring method by Van Zeeland (2008), demonstrated a significantly higher reliability of the total feather score in the feather scoring method by Van Zeeland. However, aside from reliability, feasibility is an important factor for clinical applicability. The feather scoring method by Van Zeeland seems to make scoring faster and more efficient due to use of an objective, numerical scale for assessment of all the different body parts.

References

*Bland, J. M. and D. G. Altman (1996). "Measurement error." British Medical Journal **313**(7059): 744.*

*Bordnick, P. S., B. A. Thyer, et al. (1994). "Feather picking disorder and trichotillomania: An avian model of human psychopathology." Journal of Behavior Therapy and Experimental Psychiatry **25**(3): 189-196.*

*Cohen, J. (1968). "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." Psychological Bulletin **70**(4): 213-220.*

*Cooper, J. E. and G. J. Harrison (2007). Dermatology. Avian Medicine: Principles and Application B. W. Ritchie, G. J. Harrison and R. H. Harrison, Wingers Pub.**:** 635-636.*

*Costa-Santos, C., J. Bernardes, et al. (2011). "The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement." Journal of Clinical Epidemiology **64**(3): 264-269.*

*de Vet, H. C., C. B. Terwee, et al. (2006). "When to use agreement versus reliability measures." Journal of Clinical Epidemiology **59**(10): 1033-1039.*

Eggen, T. J. H. M. and P. F. Sanders (1993). Psychometrie in de praktijk, CITO.

*Eliasziw, M., S. L. Young, et al. (1994). "Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example." Physical Therapy **74**(8): 777-788.*

Field, A. P. (2009). Discovering statistics using SPSS. Los Angeles, SAGE Publications.

*Fleiss, J. L. and J. Cohen (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability." Education and Psychological Measurement **33**: 613-619.*

*Grindlinger, H. M. and E. Ramsay (1991). "Compulsive feather picking in birds." Archives of General Psychiatry **48**(9): 857.*

Harrison, G. J. (1986). Disorders of the integument. Philadelphia, WB Saunders Co.

*Kjaer, J. B., H. Glawatz, et al. (2011). "Reducing stress during welfare inspection: validation of a non-intrusive version of the LayWel plumage scoring system for laying hens." British Poultry Science **52**(2): 149 - 154.*

*Lumeij, J. T. and C. J. Hommers (2008). "Foraging 'enrichment' as treatment for pterotillomania." Applied Animal Behaviour Science **111**(1-2): 85-94.*

*Mandrekar, J. N. (2011). "Measures of interrater agreement." Journal of Thoracic Oncology **6**(1): 6-7.*

*McGraw, K. O. and S. P. Wong (1996). "Forming Inferences about Some Intraclass Correlation Coefficients." Psychological methods **1**(1): 30-46.*

*Meehan, C. L., J. R. Millam, et al. (2003). "Foraging opportunity and increased physical complexity both prevent and reduce psychogenic feather picking by young Amazon parrots." Applied Animal Behaviour Science **80**(1): 71-85.*

Mertens, P. A. (1997). "Pharmacological treatment of feather picking in pet birds " Proceedings of the 1st International Conference on Veterinary Behavioural Medicine.

*Nett, C. S. and T. N. Tully Jr. (2003). "Anatomy, clinical presentation and diagnostic approach to feather-picking pet birds." Compendium Continuing Educ. Vet. Practitioner **25**: 206-219.*

*Rankin, G. and M. Stokes (1998). "Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses." Clinical Rehabilitation **12**(3): 187-199.*

*Schuster, C. (2004). "A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales." Educational and Psychological Measurement **64**(2): 243-253.*

*Shrout, P. E. and J. L. Fleiss (1979). "Intraclass correlations: Uses in assessing rater reliability." Psychological Bulletin **86**(2): 420-428.*

*Sim, J. and C. C. Wright (2005). "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements." Physical Therapy **85**(3): 257-268.*

Streiner, D. L. and G. R. Norman (2003). Health measurement scales : a practical guide to their development and use. Oxford ; New York, Oxford University Press.

*Tauson, R., J. Kjaer, et al. (2005). "Applied scoring of integument and health in laying hens." Animal Science Papers and Reports **23**: 153-159.*

Van der Valk, L. (2009). Vergelijking van betrouwbaarheid van twee veerscoresystemen ter bepaling van de mate van verenplukken bij de grijze roodstaart (psittacus erithacus).

*Van Hoek, C. S. and C. E. King (1997). "Causation and influence of environmental enrichment on feather picking of the crimson-bellied conure (Pyrrhura perlata perlata)." Zoo Biology **16**(2): 161-172.*

*van Zeeland, Y. R. A., B. M. Spruit, et al. (2009). "Feather damaging behaviour in parrots: A review with consideration of comparative aspects." Applied Animal Behaviour Science **121**(2): 75-95.*

*Weir, J. P. (2005). "Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM." Journal of Strength and Conditioning Research **19**(1): 231-240.*

Appendix A

The feather scoring system developed by Meehan et al.

Score	Description
a) Scoring system used for chest/flank, back and legs	
0	All or most feathers removed, down removed and skin exposed, evidence of skin or tissue injury
0.25	All or most feathers removed, down removed and skin exposed, no evidence of skin or tissue injury
0.5	All or most feathers removed, some down removed, patches of skin exposed
0.75	All or most feathers removed, down exposed and intact or feathers removed from more than half of the area, some down removed, patches of skin exposed
1.0	Feathers removed from less than half of the area, some down removed and skin exposed
1.25	Feathers removed from more than half of the area, down exposed and intact
1.5	Feathers removed from less than half of the area, down exposed and intact
1.75	Feathers intact with fraying or breakage
2.0	Feathers intact with little or no fraying or breakage
b) Scoring system used for wings	
0	All or most primaries, secondaries and coverts removed, down removed, skin exposed, evidence of skin or tissue injury
0.5	All or most primaries, secondaries and coverts removed, down removed, skin exposed, no evidence of injury
1.0	More than half of coverts removed, down exposed and intact or more than half of primaries and secondaries removed, down exposed and intact
1.5	Fewer than half of coverts removed, down exposed and intact or fewer than half of primaries and secondaries removed, down exposed and intact or primaries and secondaries intact with significant breakage and fraying
2.0	Feathers intact with little or no fraying or breakage
c) Scoring system used for tail	
0	All or most tail feathers removed or broken
1.0	Some tail feathers removed or broken or significant fraying of tail feathers
2.0	Feathers intact with little or no fraying or breakage

Appendix B

The alternative feather-scoring system

Part 1: Score determination table for coverts and down feathers: use for front side of the body, backside of the body, legs, dorsal and ventral side of wings

Down feathers → ↓ Coverts	none of down feathers removed	< 50% of down feathers removed	>50% of down feathers removed	all down feathers removed
All coverts intact	100	85	75	60
Some fraying	95	80	70	55
< 25% of coverts damaged	90	75	65	50
25-50% of coverts damaged	80	65	55	40
50-75% of coverts damaged	70	55	45	30
75-90% of coverts damaged	60	45	35	20
> 90% of coverts damaged	50	35	25	10

NB: In case of skin damage: always score 10 points less than the initial score!
Example: If the initial score for coverts and down was 60 but the bird shows skin lesions on that body part, the final score will be 50

Part 2: Score determination for flight feathers: use for tail and wings

Each wing has 10 primaries and 10 secondaries. The tail has 10 flight feathers, also known as rectrices.

Altogether, there are 50 flight feathers to be assessed. Each feather contributes for 2 points maximum to the total score (i.e. 50 x 2 = 100). For each feather that is damaged, points will be deducted from the total. Accordingly, for each individual feather the following score system applies:

Undamaged feather = 0 points deduction
Feather with signs of fraying and/or breakage with a length of > 50% = 1 point deduction
Feather with signs of fraying and/or breakage with a length of < 50% = 2 points deduction

Appendix C

Feather scoring session #

Welcome to this # session in which you will be asked to assess a total of five parrots.

This # and subsequent sessions are part of a comparative study about the intra- and inter-observer reliability of two different feather scoring systems (i.e. the feather scoring system according to Meehan *et al*. and a novel feather scoring system).

Note 1: In the feather scoring system of Meehan *et al*. the ventral sides of the wings are not assessed.
Note 2: For assessment of the intra- and inter-observer reliability for the novel scoring system only covert and down feathers are included.
For each parrot there is a series of photos, which all show a specific part of the body. The following body parts will be shown and assessed: front side of the body, back, legs, ventral surface of the wings, dorsal surface of the wings, and tail. The photos are arranged in the same order for all parrots.
For each photograph that is shown you will be asked to answer a series of multiple choice questions (max. 4) regarding the condition of the plumage and skin. Please fill in the one answer that you feel is the most appropriate. Based on your answers the total feather score can be calculated for each parrot using both feather scoring systems. You do not have to use any tables or calculate feather scores yourself.
This session will take approximately 35 minutes of your time. Please note that it is possible to save your progress and resume the survey later.
There are 100 questions in this survey.

Parrot # Front side of the body

1 Which of the following descriptions fits best for the condition of the plumage and skin of the body part shown?

- ○ feathers intact with little or no fraying or breakage
- ○ feathers intact with fraying or breakage
- ○ feathers removed from less than half of the area, down exposed and intact
- ○ feathers removed from more than half of the area, down exposed and intact
- ○ feathers removed from less than half of the area, some down removed and skin exposed
- ○ feathers removed from more than half of the area, some down removed, patches of skin exposed
- ○ all or most feathers removed, down exposed and intact
- ○ all or most feathers removed, some down removed, patches of skin exposed
- ○ all or most feathers removed, down removed and skin exposed, no evidence of skin or tissue injury
- ○ all or most feathers removed, down removed and skin exposed, evidence of skin or tissue injury

2 Which category applies best for the covert feathers?

- ○ all coverts intact
- ○ some fraying and/or breakage present (< 10% of feathers damaged)
- ○ < 25% of coverts damaged
- ○ 25-50% of coverts damaged
- ○ 50-75% of coverts damaged
- ○ 75-90% of coverts damaged
- ○ > 90% of coverts damaged

3 Which category applies best for the down feathers?

- ○ none of down feathers removed
- ○ < 50% of down feathers removed
- ○ > 50% of down feathers removed
- ○ all down feathers removed

4 Is the skin damaged?

- ○ Yes
- ○ No

Parrot # Back

5 Which of the following descriptions fits best for the condition of the plumage and skin of the body part shown?

- ○ feathers intact with little or no fraying or breakage
- ○ feathers intact with fraying or breakage
- ○ feathers removed from less than half of the area, down exposed and intact
- ○ feathers removed from more than half of the area, down exposed and intact

○ feathers removed from less than half of the area, some down removed and skin exposed
○ feathers removed from more than half of the area, some down removed, patches of skin exposed
○ all or most feathers removed, down exposed and intact
○ all or most feathers removed, some down removed, patches of skin exposed
○ all or most feathers removed, down removed and skin exposed, no evidence of skin or tissue injury
○ all or most feathers removed, down removed and skin exposed, evidence of skin or tissue injury

6 Which category applies best for the covert feathers?
○ all coverts intact
○ some fraying and/or breakage present (< 10% of feathers damaged)
○ < 25% of coverts damaged
○ 25-50% of coverts damaged
○ 50-75% of coverts damaged
○ 75-90% of coverts damaged
○ > 90% of coverts damaged

7 Which category applies best for the down feathers?
○ none of down feathers removed
○ < 50% of down feathers removed
○ > 50% of down feathers removed
○ all down feathers removed

8 Is the skin damaged?
○ Yes
○ No

Parrot # Legs

9 Which of the following descriptions fits best for the condition of the plumage and skin of the body part shown?
○ feathers intact with little or no fraying or breakage
○ feathers intact with fraying or breakage
○ feathers removed from less than half of the area, down exposed and intact
○ feathers removed from more than half of the area, down exposed and intact
○ feathers removed from less than half of the area, some down removed and skin exposed
○ feathers removed from more than half of the area, some down removed, patches of skin exposed
○ all or most feathers removed, down exposed and intact
○ all or most feathers removed, some down removed, patches of skin exposed
○ all or most feathers removed, down removed and skin exposed, no evidence of skin or tissue injury
○ all or most feathers removed, down removed and skin exposed, evidence of skin or tissue injury

10 Which category applies best for the covert feathers?
○ all coverts intact
○ some fraying and/or breakage present (< 10% of feathers damaged)
○ < 25% of coverts damaged
○ 25-50% of coverts damaged
○ 50-75% of coverts damaged
○ 75-90% of coverts damaged
○ > 90% of coverts damaged

11 Which category applies best for the down feathers?
○ none of down feathers removed
○ < 50% of down feathers removed
○ > 50% of down feathers removed
○ all down feathers removed

12 Is the skin damaged?
○ Yes
○ No

Parrot # Dorsal surface of the wings

13 Which of the following descriptions fits best for the condition of the plumage and skin of the body part shown?
Note: For this description you need to take all feathers (coverts, down feathers, primaries and secondaries) into account.
○ feathers intact with little or no fraying or breakage
○ primaries and secondaries intact with significant breakage and fraying
○ fewer than half of primaries and secondaries removed, down exposed and intact
○ fewer than half of coverts removed, down exposed and intact
○ more than half of primaries and secondaries removed, down exposed and intact
○ more than half of coverts removed, down exposed and intact

○ all or most primaries, secondaries and coverts removed, down removed, skin exposed, no evidence of injury

○ all or most primaries, secondaries and coverts removed, down removed, skin exposed, evidence of skin or tissue injury

14 Which category applies best for the covert feathers?

○ all coverts intact

○ some fraying and/or breakage present (< 10% of feathers damaged)

○ < 25% of coverts damaged

○ 25-50% of coverts damaged

○ 50-75% of coverts damaged

○ 75-90% of coverts damaged

○ > 90% of coverts damaged

15 Which category applies best for the down feathers?

○ none of down feathers removed

○ < 50% of down feathers removed

○ > 50% of down feathers removed

○ all down feathers removed

16 Is the skin damaged?

○ Yes

○ No

Parrot # Ventral surface of the wings

17 Note: You do not need to give a description!
Which category applies best for the covert feathers?

○ all coverts intact

○ some fraying and/or breakage present (< 10% of feathers damaged)

○ < 25% of coverts damaged

○ 25-50% of coverts damaged

○ 50-75% of coverts damaged

○ 75-90% of coverts damaged

○ > 90% of coverts damaged

18 Which category applies best for the down feathers?

○ none of down feathers removed

○ < 50% of down feathers removed

○ > 50% of down feathers removed

○ all down feathers removed

19 Is the skin damaged?

○ Yes

○ No

Parrot # Tail

20 Which of the following descriptions fits best for the tail?

○ feathers intact with little or no fraying or breakage

○ significant fraying of tail feathers

○ some tail feathers removed or broken

○ all or most tail feathers removed or broken