


BRIEF COMMUNICATION

Open Access



Can shelter dog observers score behavioural expressions consistently over time?

Solveig Marie Stubsjøen^{1*} , Randi Oppermann Moe², Cicilie Johannessen², Maiken Larsen², Henriette Madsen² and Karianne Muri²

Abstract

A substantial number of dogs live in animal shelters worldwide. Stressors within the shelter environment can compromise their welfare, and scientific evaluations of feasible welfare assessment methods are therefore needed. Qualitative Behaviour Assessment (QBA) is a “whole-animal” approach used to assess welfare by observing animals’ expressive behaviour. To investigate whether observers can score dogs’ behavioural expressions consistently over time, this study replicated and extended previous research, by evaluating intra- and inter-observer reliability of QBA based on video recordings of shelter dogs. In Part I, nine veterinary nurse students received theoretical and practical training, and then scored 12 2 min video recordings of shelter dogs using a fixed list of behavioural descriptors. Three of the students undertook further practice and calibration using direct observations of dog behaviour in a local shelter. In Part II, the videos from Part I were scored by these three observers a second time, 15 months later. QBA data were analysed using principal component analysis (PCA), and reliability was assessed using Kendall’s coefficient of concordance (W). In Part I, the inter-observer reliability was high for both components (0.78 for PC1 and 0.85 for PC2). In Part II, the inter-observer reliability was very high and moderate for PC1 and PC2, respectively (0.90 for PC1 and 0.65 for PC2). The intra-observer reliability was high for both components ($W \geq 0.86$). Our results indicate that the fixed list of behavioural descriptors for shelter dogs can be used reliably when assessing videos, and that observers can score dogs’ behavioural expressions consistently after a break of 15 months following the initial assessment. Nevertheless, the reduction in inter-observer-reliability of PC2 in Part II can indicate that some retraining and calibration may be required to avoid observer drift.

Keywords: Animal welfare, Behaviour, Fixed list of descriptors, Observer reliability, Qualitative behaviour assessment, Shelter dogs

Findings

Animal shelters exist worldwide, and a substantial number of dogs live in shelters for prolonged periods. The aim of shelters is to rehome animals and thereby improving their long-term welfare [1]. To ensure good welfare during their stay in the shelter, dogs should experience more

positive (e.g., pleasure) than negative (e.g., fear, frustration) emotions [2]. There are many potential stressors within the shelter environment that can compromise the dogs’ welfare, such as unfamiliar sounds, smells, routines, and people [3]. Scientific evaluations of feasible welfare assessment methods for shelter dogs are therefore needed. Shelter staff are often under-resourced, and a quick, reliable method to monitor welfare in a potentially large number of dogs over time would therefore be of value.

Qualitative Behavioural Assessment (QBA) is a “whole-animal” approach, where observers assess the animals’

*Correspondence: Solveig-Marie.Stubsjoen@vetinst.no

¹ Department of Animal Health, Welfare and Food Safety, Section for Terrestrial Animal Health and Welfare, Norwegian Veterinary Institute, Elizabeth Stephansens Vei 1, 1433 Ås, Norway
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

expressive behaviour by integrating and summarising the details of behaviour, posture, and movement (body language) in light of the context [1, 2, 4, 5]. The validity of QBA has been supported by studies in various livestock species, demonstrating significant associations with other behavioural and physiological measurements [5]. However, further research is needed to investigate the use of QBA to assess dog welfare. A high level of reliability is an essential requirement for any method used to assess animal welfare [6], and reliability is considered a prerequisite for validity [7]. Observer reliability can be measured within a single observer (intra-observer) and between multiple observers (inter-observer) [8].

In a previous study, Stubsjøen et al. [9] found a high inter-observer reliability when observers used a fixed list of descriptors to assess videos of shelter dogs. To evaluate the robustness of this finding, we replicated the study by using the same fixed list of descriptors and video recordings to reassess inter-observer reliability. In addition, we extended the previous study by assessing the long-term inter- and intra-observer reliability. Our aim was to investigate whether observers can score dogs' behavioural expressions consistently over time.

A group of nine final (3rd) year veterinary nurse students at the Norwegian University of Life Sciences, Faculty of Veterinary Medicine, consented to participate in Part I of the study. The students had not received any training in the use of QBA prior to the introductory presentation given at the first video scoring session. However, three of these students did their final year research project as a part of this project and had been asked to read six scientific papers on QBA of dogs, sheep, and broiler chickens prior to this session [1, 2, 10–13].

On the test day, the nine students were trained using the same procedure as in the previous study [9]. The video recordings of dogs were obtained from a shelter in southern Hungary, in which about 250 dogs were kept [9]. The animals were stray dogs or brought in from private homes for animal welfare reasons. After the introduction (~1 h), the students were first shown three test videos and thereafter encouraged to discuss their interpretation of the dogs' behavioural expressions and to compare their scoring results. Subsequently, the fixed list of 20 qualitative descriptors, which included definitions of each, was used by the observers (Part Ia: n = 9 observers, Part Ib: n = 3 observers) to score the 12 videos [9] (Table 1).

During the following week, the three students who did their final year project on QBA, practiced the method by direct observations of privately owned dogs. 2 weeks after the video scoring, the students applied their QBA skills by scoring during direct observations of dogs in a local shelter. However, the shelter was subsequently

Table 1 Kendall's coefficient of concordance (W) for principal components and individual behavioural terms used by observers in Part I (Part Ia: n = 9 observers, Part Ib: n = 3 observers), and Part II (n = 3 observers) to assess shelter dogs in 12 video clips

Variable	Part Ia W for all observers	Part Ib W for 3 observers	Part II W for 3 observers
PC1	0.78	0.79	0.90
PC2	0.85	0.91	0.65
Content	0.57	0.52	0.37
Uncomfortable	0.58	0.86	0.63
Playful	0.68	0.62	0.64
Depressed	0.67	0.71	0.59
Relaxed	0.59	0.53	0.49
Restless	0.63	0.82	0.51
Alert	0.65	0.70	0.54
Bored	0.53	0.76	0.23
Sociable	0.79	0.78	0.73
Nervous	0.47	0.61	0.53
Expectant	0.73	0.66	0.74
Hesitant	0.61	0.82	0.46
Trustful	0.54	0.56	0.44
Aggressive	0.40	0.46	0.25
Energetic	0.67	0.62	0.70
Frustrated	0.53	0.70	0.39
Curious	0.66	0.79	0.72
Calming	0.32	0.47	0.53
Indifferent	0.54	0.82	0.54
Stressed	0.52	0.78	0.46

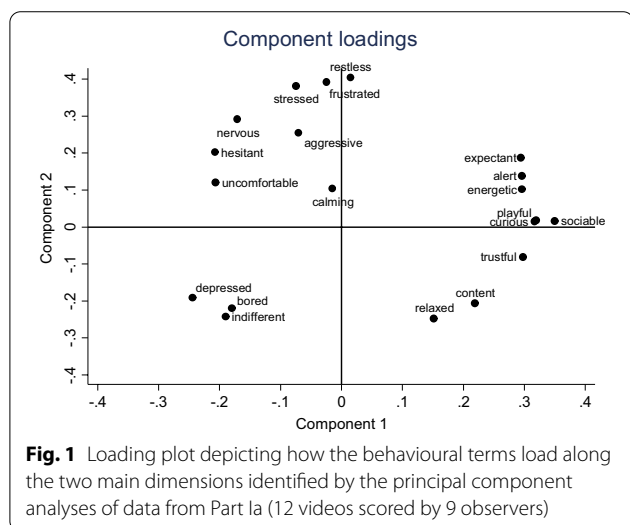
closed for intake of new dogs due to an outbreak of acute haemorrhagic diarrhoea of dogs in Norway [14]. For biosecurity reasons, the scoring sessions could therefore not continue. For the direct observations, the inter-observer reliability was found to be high (0.85 for PC1 and 0.76 for PC2), but due to the low sample size (n = 10), these assessments were considered as practice and calibration.

Fifteen months after the first video scoring session, the three students scored the videos a second time (Part II). The descriptors were not discussed before watching the videos, and no further instructions were given. The same 12 video clips were shown, but the order was changed using random number allocation. Because of strict regulation of social distancing due to the COVID-19 pandemic, it was not possible to score the videos under the same conditions as in the first session (i.e., in the same room). The video-scoring were therefore performed during a Microsoft Teams meeting, where the students watched the videos simultaneously on their computers.

Visual analogue scales (VAS) ranging from *Minimum* to *Maximum* were used to score the intensity of each

behavioural expression. QBA scores were registered by measuring the distance in millimetres between the *Minimum* point of each VAS, to the point where the scale was ticked by the observer. The data were then transferred to a spreadsheet (Microsoft Office Excel® 2010). Statistical analyses were conducted in Stata SE/16.1 (StataCorp, College Station, Texas). The QBA data were analysed using principal component analysis (PCA) with a correlation matrix (no rotation). PCA reveals the underlying structure of the data and reduces the number of variables to a few main components, each comprising correlated behavioural expressions [12]. To assist in the determination of the number of components to retain, we used a combination of the scree plot criterion and Kaiser’s criterion [15]. The components that explained most of the variance in the data were retained (PC1, PC2), and component scores were calculated. Kendall’s coefficient of concordance (*W*) was used to assess observer reliability. Inter-observer reliability was assessed for the component scores as well as the scores for each individual behavioural descriptor. Intra-observer reliability was calculated for the component scores only, with data from the two different time points (Part Ib and Part II). The reliability coefficients were interpreted according to Martin and Bateson [8].

PCA of the data from the nine participants in Part Ia resulted in a two-component solution, explaining 34.8% and 21.9% of the variance, respectively. PC1 ranged from *depressed*, *uncomfortable*, *hesitant*, and *indifferent* to *sociable*, *curious*, *playful*, and *trustful*. PC2 ranged from *relaxed*, *indifferent*, and *content* to *restless*, *frustrated*, and *stressed* (Fig. 1). *W* for the first (PC1) and second (PC2) component were 0.78 and 0.85, respectively, indicating high agreement (Table 1).



PCA of the data from the three observers in Part II resulted in a two-component solution, explaining 30.0 and 20.0% of the variance, respectively. The anchoring points for the principal components were similar to the anchoring points in Part I (Fig. 2), and comparable to the previous study [9]. The first component (PC1) in both parts of the study reflects the dogs’ mood, while the second component (PC2) appears to be related to arousal. *W* for the first component score (PC1) was 0.90, indicating a very high inter-observer agreement. The second component (PC2) had a reliability coefficient of 0.65, indicating moderate agreement (Table 1). The intra-observer agreement was very high for PC1 (*W* > 0.9), and high for PC2 (*W* ≥ 0.86) for all three observers (Table 2).

The high agreement among the observers is in accordance with previous studies [9, 11], however, there were varying levels of reliability of the individual behavioural terms. The nine assessors in Part Ia achieved somewhat lower agreement compared to the three assessors in Part Ib (Table 1). These three students had read six scientific papers on QBA of dogs, sheep, and broiler chickens prior to the scoring session, and this may have improved their understanding of the approach.

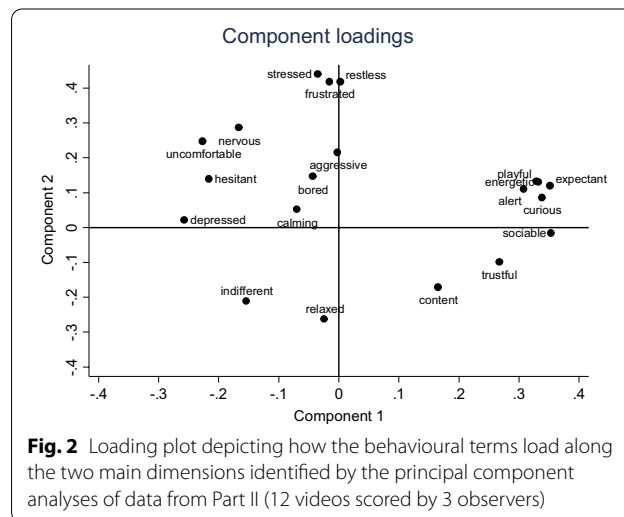


Table 2 Intra-observer agreement for individual observers given as Kendall’s coefficient of concordance (*W*)

Observer	PC1		PC2	
	<i>W</i>	<i>p</i>	<i>W</i>	<i>p</i>
Student 1	0.92	0.04	0.86	0.06
Student 2	0.93	0.04	0.87	0.06
Student 3	0.93	0.04	0.92	0.04

The inter-observer reliability of PC1 and PC2 in Part II were high and moderate, respectively, while the intra-observer reliability was high for both PC1 and PC2. There is a risk of observer drift over time, i.e., the observers unconsciously alter their personal understandings of the descriptors. The observers may have acquired different experiences in the time between the scoring sessions, which may have modified their assessments of the animals' behavioural expressions. Our results suggest that even though the observers were still familiar with the method, and the observer reliability was mainly high, training and calibration sessions are important to avoid observer drift when there is a prolonged period between assessments.

Minero et al. [16] demonstrated that training of observers improves the inter-observer reliability of QBA of donkeys. However, studies addressing the impact of observer training on long-term inter- and intra-observer reliability of QBA assessments appear to be lacking. Bokkers et al. [17] found insufficient observer reliability when experienced observers scored QBA of dairy cattle 9 months after the initial assessments. The authors proposed that one explanation could be that the observers were not trained well enough. The three students participating in Part I and II undertook additional training, practice, and calibration. The combination of video scoring and direct observations associated with training may have been beneficial with regards to the inter- and intra-observer reliability of QBA in this setting.

Our results align with the previous study [9] and suggest that observers can score shelter dogs' behavioural expressions consistently over time using QBA. Nevertheless, the reduced inter-observer reliability of PC2 in Part II indicates that some degree of retraining may be required to maintain good reliability. The results found in Part II of the study highlight that training, practice and calibration may play a role, and future research may shed more light on this.

Acknowledgements

The authors wish to thank the students who participated in the study and the management at the shelters in Hungary and in Oslo.

Prior publication

Data have not been published previously.

Author contributions

SMS secured the funding. SMS and ROM initiated, planned and designed the study and supervised the students. SMS, ROM, CJ, ML and HM collected the data. KM performed the statistical analyses. All authors interpreted the data. SMS drafted the manuscript, and ROM, KM, CJ, ML and HM reviewed and edited the draft. All authors have read and approved the final manuscript.

Funding

The project was funded by Astri og Birger Torsteds legat, Smådyrpraktiserende veterinærers forenings vitenskapelige og faglige fond (Norwegian Small Animal Veterinary Association's Scientific Foundation) and the Norwegian Veterinary Institute's internal funding (Strategic programme on animal welfare).

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study did not involve procedures that necessitate application for ethics approval through the Norwegian Food Safety Authority, as stated in the Regulation on Animal Experimentation (Forskrift om bruk av dyr i forsøk). The dogs were video-recorded or observed directly in their kennels with minimum disturbance of the animals. The study was conducted in accordance with the Ethical Guidelines for the Use of Animals in Research [18]. A written informed consent was obtained from the administration of the shelters prior to the study. All students who assessed the videos gave their informed consent for participation prior to the study.

Consent for publication

The administration of the shelters gave a consent for publication.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Animal Health, Welfare and Food Safety, Section for Terrestrial Animal Health and Welfare, Norwegian Veterinary Institute, Elizabeth Stephansens Vei 1, 1433 Ås, Norway. ²Department of Production Animal Clinical Sciences, Faculty of Veterinary Medicine, Norwegian University of Life Sciences, Elizabeth Stephansens Vei 15, 1433 Ås, Norway.

Received: 25 July 2022 Accepted: 17 November 2022

Published online: 02 December 2022

References

- Walker JK, Dale AR, D'Eath RB, Wemelsfelder F. Qualitative behaviour assessment of dogs in the shelter and home environment and relationship with quantitative behaviour assessment and physiological responses. *Appl Anim Behav Sci.* 2016;184:97–108. <https://doi.org/10.1016/j.applanim.2016.08.012>.
- Arena L, Wemelsfelder F, Messori S, Ferri N, Barnard S. Application of free choice profiling to assess the emotional state of dogs housed in shelter environments. *Appl Anim Behav Sci.* 2017;195:72–9. <https://doi.org/10.1016/j.applanim.2017.06.005>.
- Lamon TK, Slater MR, Moberly HK, Budke CM. Welfare and quality of life assessments for shelter dogs: a scoping review. *Appl Anim Behav Sci.* 2021;244:105490. <https://doi.org/10.1016/j.applanim.2021.105490>.
- Wemelsfelder F. The scientific validity of subjective concepts in models of animal welfare. *Appl Anim Behav Sci.* 1997;53:75–88. [https://doi.org/10.1016/S0168-1591\(96\)01152-5](https://doi.org/10.1016/S0168-1591(96)01152-5).
- Fleming PA, Clarke T, Wickham SL, Stockman CA, Barnes AL, Collins T, Miller DW. The contribution of qualitative behavioural assessment to appraisal of livestock welfare. *Anim Prod Sci.* 2016;56:1569–78. <https://doi.org/10.1071/AN15101>.
- Tuytens FAM, De Graaf S, Heerkens JLT, Jacobs L, Nalon E, Ott S, et al. Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? *Anim Behav.* 2014;90:273–80. <https://doi.org/10.1016/j.anbehav.2014.02.007>.
- Taylor KD, Mills DS. The development and assessment of temperament tests for adult companion dogs. *J Vet Behav.* 2006;1:94–108. <https://doi.org/10.1016/j.jvbeh.2006.09.002>.
- Martin P, Bateson P. *Measuring behaviour: an introductory guide.* 3rd ed. Cambridge: Cambridge University Press; 2007.
- Stubsjøen SM, Moe RO, Bruland K, Lien T, Muri K. Reliability of observer ratings: qualitative behaviour assessments of shelter dogs using a fixed list of descriptors. *Vet Anim Sci.* 2020;10:100145. <https://doi.org/10.1016/j.vas.2020.100145>.

10. Walker J, Dale A, Waran N, Farnworth M, Clarke N, Wemelsfelder F. The assessment of emotional expression in dogs using a free choice profiling methodology. *Anim Welf*. 2010;19:75–84.
11. Arena L, Wemelsfelder F, Messori S, Ferri N, Barnard S. Development of a fixed list of terms for the qualitative behavioural assessment of shelter dogs. *PLoS ONE*. 2019;14:e0212652. <https://doi.org/10.1371/journal.pone.0212652>.
12. Muri K, Stubsjøen SM. Inter-observer reliability of qualitative behavioural assessments (QBA) of housed sheep in Norway using fixed lists of descriptors. *Anim Welf*. 2017;26:427–35. <https://doi.org/10.7120/09627286.26.4.427>.
13. Muri K, Stubsjøen SM, Vasdal G, Moe RO, Granquist EG. Associations between qualitative behaviour assessments and measures of leg health, fear and mortality in Norwegian broiler chicken flocks. *Appl Anim Behav Sci*. 2019;211:47–53. <https://doi.org/10.1016/j.applanim.2018.12.010>.
14. Jørgensen HJ, Valheim M, Sekse C, Bergsjø BA, Wisløff H, Nørstebø SF, et al. An official outbreak investigation of acute haemorrhagic diarrhoea in dogs in Norway points to *Providencia alcalifaciens* as a likely cause. *Animals*. 2021;11:3201. <https://doi.org/10.3390/ani11113201>.
15. Tabachnick BG, Fidell LS, Ullman JB. Using multivariate statistics. 6th ed. Essex: Pearson Education Limited; 2013.
16. Minero M, Dalla Costa E, Dai F, Murray LAM, Canali E, Wemelsfelder F. Use of qualitative behaviour assessment as an indicator of welfare in donkeys. *Appl Anim Behav Sci*. 2016;174:147–53. <https://doi.org/10.1016/j.applanim.2015.10.010>.
17. Bokkers E, De Vries M, Antonissen I, de Boer I. Inter-and intra-observer reliability of experienced and inexperienced observers for the qualitative behaviour assessment in dairy cattle. *Anim Welf*. 2012;21:307–18. <https://doi.org/10.7120/09627286.21.3.307>.
18. The Norwegian National Committee for Research Ethics in Science and Technology (NENT). Ethical Guidelines for the Use of Animals in Research. Oslo. 2018. <https://www.forskningsetikk.no/en/guidelines/science-and-technology/ethical-guidelines-for-the-use-of-animals-in-research/>. Accessed 01.10.2022.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

