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

Animal Welfare 2022, 31: 319-327
ISSN 0962-7286
doi: 10.7120/09627286.31.3.004

Do you see the same cat that I see? Inter- and intra-observer reliability for Qualitative Behaviour Assessment as temperament indicator in domestic cats

IC Travnik[†‡], DS Machado[†‡] and AC Sant'Anna*[‡§]

[†] Programa de Pós-Graduação em Comportamento e Biologia Animal, Instituto de Ciências Biológicas, Universidade Federal de Juiz de Fora, 36.036-330, Juiz de Fora, Minas Gerais, Brazil
[‡] Núcleo de Estudos em Etologia e Bem-estar Animal, Departamento de Zoologia, Instituto de Ciências Biológicas, Universidade Federal de Juiz de Fora, 36.036-330, Juiz de Fora, Minas Gerais, Brazil
[§] Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil
* Contact for correspondence: aline.santanna@ufjf.edu.br

Abstract

*Qualitative Behaviour Assessment (QBA) is used to assess animals' emotional expressions and its potential for serving as an indicator of temperament has been explored. This method is open to assessors' interpretation and it is therefore necessary to evaluate the observers' reliability for different species and contexts. We aimed to assess the intra- and inter-observer reliability of QBA as an indicator of cat (*Felis catus*) temperament. The QBA was applied by 19 observers with divergent profiles of contact with cats (cat owners vs non-owners) and experience in behavioural assessment (experienced vs inexperienced). Forty-two, 12-min videos were assessed, composed of footage of four behavioural tests: unfamiliar person, novel object, conspecific reaction, and food offering tests. By using Principal Component Analysis, we found three principal components (PC) that were considered the main dimensions of cat temperament. According to Kendall's coefficient of concordance, intra-observer reliability was high to very high in PC1 (0.80–0.90) and moderate to high in PC2 and PC3 (0.50–0.82). Inter-observer reliability for the 19 observers was high in PC1 (0.71) and low in PC2 and PC3 (0.21–0.29). The individual concordances with the gold observer (defined based on greater experience with the QBA) ranged from moderate to high. We concluded that QBA could be a reliable tool to assess cat temperament, given the high values of intra- and inter-observer reliabilities in PC1, which is the dimension that most explains the behavioural variations in the cats' temperament. The same did not occur for PC2 and PC3, showing that reliability varied among the different dimensions and observers.*

Keywords: *animal welfare, behaviour, companion animals, personality, rating method, shelter cats*

Introduction

The use of rating methods for assessing animal behaviours has increased in recent years (Finka *et al* 2019; Fukimoto *et al* 2019, 2020; Salonen *et al* 2019). A widely used type of rating method is the Qualitative Behaviour Assessment (QBA) (Wemelsfelder *et al* 2001). The QBA was developed by Wemelsfelder *et al* (2000) and allows the assessment of animals' emotional expressions, behaviour and patterns of interactions with their environment, holistically and integratively instead of via analysis of discrete and isolated categories of behaviour (Wemelsfelder & Lawrence 2001). Thus, it is possible to identify subtle variations in the animals' body language that are hard to identify using other coding methods (Wemelsfelder *et al* 2001). Additionally, behaviours presented by one or a few animals, which would be disregarded in usual coding methods, can be gathered with QBA. The QBA is based on the use of descriptors to quantify positive or negative mental states on visual analogue scales.

The QBA can be used in one of two ways, from the free choice profile, where an observer uses descriptors chosen by him/her at the moment of the assessment, or the fixed-list, in which the observer uses a list of predefined descriptors (Bokkers *et al* 2012; Phythian *et al* 2013; Diaz-Lundahl *et al* 2019). The latter method has been regarded as valid, mainly for assessments with more practical purposes (Arena *et al* 2019; Travnik & Sant'Anna 2021). The QBA has been applied through video footage, usually making use of short-time videos (1–2 min) (Stubsjøen *et al* 2020), or longer observations of animals in the field (around 20 min) incorporated into welfare assessment protocols (Welfare Quality® 2009; Dwyer *et al* 2015). Recently, the QBA has also been shown to be valid for accessing temperament dimensions using 12-min videos showing inter-individual behavioural differences in animals' reactions to three different stimuli (Travnik & Sant'Anna 2021).

The assessment of consistency among observers for different species and contexts is an important step so that

new methodologies may be widely used (Kaler 2009). Several previous studies analysed inter-observer reliabilities for QBA in different contexts and species, showing divergent results (Walker *et al* 2010; Bokkers *et al* 2012; Phythian *et al* 2013; Clarke *et al* 2016; Diaz-Lundahl *et al* 2019; Stubsjøen *et al* 2020; Ceballos *et al* 2021). Stubsjøen and collaborators (2020) used QBA to assess shelter dogs' styles of behaviour using a fixed-list of descriptors. Twenty-two veterinary and veterinary nurse students evaluated 12 videos, obtaining high inter-observer reliability in PC1 (0.88) and PC2 (0.79) (Stubsjøen *et al* 2020). Walker *et al* (2010) also applied QBA in dogs (*Canis familiaris*), obtaining high inter-observer reliabilities among 18 untrained observers. Other studies reported promising values of inter-observer reliability, ranging from high to very high coefficients in PC1 (ranging from 0.70–0.97), moderate to very high in PC2 (0.45–0.93) (Phythian *et al* 2013; Diaz-Lundahl *et al* 2019; Ceballos *et al* 2021) and moderate to high in PC3 (0.55–0.73) (Ceballos *et al* 2021). In Clarke *et al* (2016), the Kendall's coefficients were calculated for each term of a fixed-list, ranging from 0.37 ('sociable') to 0.64 ('happy'). Finally, Bokkers and collaborators (2012) assessed reliabilities for QBA by using eight experienced and ten inexperienced observers and found that the degrees of concordance varied among the descriptors and dimensions, leading the authors to conclude that QBA has insufficient reliability to be considered as a tool to assess welfare in dairy cattle (Bokkers *et al* 2012).

Studies addressing intra-observer reliability for the QBA are scarce. Ceballos *et al* (2021), reported strong and very strong reliabilities, ranging from, 0.71–0.89 in PC1, 0.74–0.90 in PC2 and 0.63–0.84 in PC3. Diaz-Lundahl *et al* (2019) showed a similar result of Kendall's coefficient of concordance in PC1 (0.89) but a wider range in PC2 (0.45–0.93). Bokkers and collaborators (2012) reported a wide range of intra-observer reliabilities, demonstrating large individual effects in their study, leading the authors to conclude that there was no consistency in scores of individual descriptors within the observers.

Research on temperament in companion animals is mainly focused on dogs (Gartner 2015). For cats (*Felis catus*), there are few papers, mostly focusing on identifying the principal dimensions of temperament. Among these studies, there is a wide variety of methodologies, making their comparison difficult (Gartner & Weiss 2013; Finka *et al* 2019). The first QBA study in domestic cats was recently published (Travnik & Sant'Anna 2021). The QBA outcomes were compared with the dimensions generated by coding methods traditionally used to assess animal temperament, being considered a valid tool to evaluate the temperament of shelter cats (Travnik & Sant'Anna 2021). By using QBA, three main dimensions were described, explaining 76.63% of the total variance, PC1 ranging from 'calm / relaxed / friendly' to 'tense / fearful / alert', PC2 ranging from 'indifferent' to 'agitated/active', and PC3 from 'aggressive' to 'suspicious' (Travnik & Sant'Anna 2021). Nevertheless, the reliabilities of these dimensions obtained with QBA have not previously been investigated in domestic cats. Identifying the principal dimensions of cat temperament through a reliable and feasible methodology could be useful, for instance, for organisations that care for abandoned animals (Fukimoto *et al* 2019). The temperament assessments in shelter cats could be used to develop best management practices, housing adequacy, and the awareness of possible owners who intend to adopt these animals (Gourkow & Fraser 2006; Weiss *et al* 2015; Fukimoto *et al* 2019).

The familiarity of human beings with companion animals could lead to better interpretations of their body language. In this scenario, it is necessary to investigate the performance and consistency of different observers' profiles to apply the QBA to cats. Therefore, this study aimed to assess the intra- and inter-observer reliability of QBA applied by observers with different profiles regarding contact with cats (owners vs non-owners) and experience in quantitative behavioural analysis (experienced vs inexperienced).

Materials and methods

This research was approved by the Animal Ethics Committee at the Federal University of Juiz de Fora, MG, Brazil (protocol no 051/2018). The study was carried out in a private shelter in which there was no adoption or collection of cats for approximately three years, therefore the social groups were considered stable. Forty-two adult (22 females and 20 males), mixed-breed, and neutered cats were studied. The cats were kept in eight differently sized pens averaging 59 m² and each contained both a free-range and an indoor area, with physical enrichment and food and water supplied *ad libitum*. The animals and procedures were the same as previously described in Travnik and Sant'Anna (2021).

First, four behavioural tests were applied to produce videos:

• Unfamiliar person (UP) test to assess the cats' reactions to an unfamiliar human — seven phases with increasing stimulus were used, ranging from phase 1 (with the tester remaining still inside the pen) to phase 7 (with the tester holding the cat's tail firmly for 3 s);

• Novel object (NO) test to assess the cats' responses to novelty by using a toy unfamiliar to the animals — consisted of two phases, one with the train-shaped children's toy placed in the centre of the room turned off for 1 min and the second phase with the object turned on to emit sound and light, for 1 min;

• Conspecific reaction (CR) to evaluate the responses to a stuffed cat positioned in the centre of the pen for 3 min; and

• Food offering (FO) tests to assess the cats' anticipatory behaviours when receiving wet cat food from a human — consisted of two phases; in the first, the tester remained stationary for 3 min standing inside the pen with wet food, and in the second each cat was offered wet food.

The UP and NO tests were performed sequentially and individually for four days. The CR and FO tests were performed in pen groups, one week and 27 days, respectively, after the first two tests. If animals showed any behaviours suggestive of panic, tests were interrupted, as described in Travnik and Sant'Anna (2021).

https://doi.org/10.7120/09627286.31.3.004 Published online by Cambridge University Press

Each animal's video was edited as follows: (i) clipping each test with their respective times; (ii) joining the four tests in a random sequence along the 42 videos; (iii) excluding the tester's voice to ensure observers were not unduly influenced during QBA analysis; (iv) visual indication on cats to be assessed for tests conducted in groups, when more than one animal appeared on the video. Each video lasted approximately 12 min and these longer clips as compared to previous QBA studies (Walker *et al* 2010; Bokkers *et al* 2012; Phythian *et al* 2013; Clarke *et al* 2016; Diaz-Lundahl *et al* 2019) were designed to adapt the QBA method to the temperament assessment, allowing the observer to gather the behavioural expressions of cats in a wider range of contexts (their responses to different stimuli) giving the observer a broader view of each individual behavioural style. The QBA was then applied using a fixed-list of 20 adjectives (active, affectionate, aggressive, agitated, attentive, alert, calm, confident, curious, fearful, friendly, indifferent, nervous, relaxed, sociable, stressed, suspicious, tense, vocal, and greedy) (Travnik & Sant'Anna 2021).

A total of 19 observers aged from 18 to 37 years participated in this study. One of the observers was defined as 'gold' since she had 15 years of experience in quantitative and qualitative behavioural observation and been a cat owner for more than ten years. She had already applied the QBA to different species of domestic animals and had intra-observer reliability greater than 0.90 for QBA. The other 18 observers had no previous experience with QBA and had undergone training in its use via five videos (three of which showing isolated tests with cervids [unknown person or novel object tests], and two with cats obtained in a pilot trial performed previously with animals not participating in this study). Training took 3 h and consisted of a brief introduction to the method and an explanation of the adjectives' meaning, followed by the QBA scoring for the five videos and a subsequent discussion about the meaning of each adjective — the aim being to contribute to better concordance among observers (Grosso *et al* 2016).

Observers consisted of four veterinary students, seven biological sciences students, three animal biology and behaviour postgraduate students, one animal behaviour researcher, and four Human Sciences (linguistics, history, journalism, and psychology) students or professionals. Sixteen were grouped in accordance with their level of experience: (i) experienced (n = 8) — people with a high degree of familiarity with quantitative behavioural analysis; (ii) inexperienced (n = 8) — people with no experience in behavioural analysis. Three of the 19 observers were not included in these groups because they did not fit into either 'inexperienced' or 'experienced' categories (defined as having already conducted quantitative behavioural observations and having completed animal behaviour courses). Observers were then regrouped as per their degree of contact with cats: (iii) cat owners (n = 9) — people who had already been owners of cats; (iv) non-owners (n = 9) — people who had never been owners of cats. One observer was excluded from this category since, despite not being a cat owner, had frequent contact with a cat via a family member. The analyses were also performed using the data of all observers, regardless of their groups (n = 19). For the first QBA assessment, the 42 videos were presented once to the 19 observers who were instructed to watch a maximum of four videos consecutively and no more than 12 video clips (or more than 8 h) on the same day. After every four videos, they were instructed to have a 30-min break and told to interrupt video sessions if feeling tired or inattentive. Inter-observer reliabilities were calculated based on this first assessment.

Ten months on from this first assessment, intra-observer reliability was obtained through the videos being presented again to 13 of the 19 observers. The four Human Sciences observers and two biological sciences students were unable to be included in this due to unavailability.

Data analysis

The temperament dimensions were extracted from the 19 observers' data via Principal Component Analysis (PCA) (Manly 2008). Components were extracted from a correlation matrix without rotation. The principal components with eigenvalues > 1 were retained as the main dimensions of cats' temperament. Terms with loadings ≥ 0.6 were regarded as the main contributions to the dimensions found.

The intra-observer reliability was calculated using Kendall's coefficients of concordance (W) for the 13 observers between the two assessments (n = 42 videos observed twice). The inter-observer reliability was also calculated using Kendall's coefficients of concordance. First, we calculated the concordances of each individual observer with the gold one for the three main dimensions of temperament (PC1 to PC3) (as proposed by Ceballos *et al* 2021). Then Kendall's coefficients within each dimension were calculated for each group (experienced, inexperienced, owners, and non-owners) and for all 19 observers. The W values were interpreted as follows: slight concordance (0.0–0.19); low (0.20–0.39); moderate (0.40–0.69); high (0.70–0.89); and very high (0.90–1.0) (Martin & Bateson 2007).

Results

Characterisation of the cats' temperament

The PCA identified four PC of the cats' temperament based on the QBA scores of the 19 observers. Only 'indifferent' had a loading ≥ 0.6 in PC4, which was considered insufficient to express an interpretable dimension of cat temperament. Thus, the first three components were retained and, together, explained 66.93% of the total variance in the dataset, being considered the principal dimensions of the cats' temperament (Table 1). The PC1 explained 43.29% of the total variance, showing high positive loadings for 'friendly / relaxed / affectionate / confident / curious / calm / greedy / sociable', and negative loadings for 'tense / stressed / fearful / suspicious / nervous', reflecting the valence of cats' behavioural and emotional expressions. PC2 explained 17.13% of the total variance and had only adjectives with high negative loadings: 'attentive / agitated / alert / active', reflecting the level of behavioural and emotional arousal. PC3 explained 6.51% of the variance

Table 1 Loadings of each adjective used in the Qualitative Behaviour Assessment (QBA) for the three principal components (PC) generated in the Principal Component Analysis.

Adjective	PC1	PC2	PC3
Active	0.58	**–0.60**	–0.24
Aggressive	–0.40	–0.10	**0.83**
Calm	**0.73**	0.25	0.05
Affectionate	**0.82**	–0.30	0.10
Tense	**–0.85**	–0.38	–0.04
Relaxed	**0.84**	0.16	0.09
Indifferent	–0.05	0.32	0.18
Curious	**0.74**	–0.46	0.01
Alert	–0.57	**–0.61**	–0.05
Nervous	**–0.66**	–0.37	0.40
Confident	**0.75**	–0.26	0.34
Vocal	0.40	–0.40	0.12
Attentive	–0.24	**–0.67**	0.07
Greedy	**0.67**	–0.47	0.02
Sociable	**0.62**	–0.40	0.07
Stressed	**–0.79**	–0.36	0.16
Fearful	**–0.77**	–0.28	–0.28
Friendly	**0.85**	–0.26	0.08
Agitated	0.21	**–0.64**	–0.29
Suspicious	**–0.77**	–0.38	–0.06
Eigenvalue	8.66	3.43	1.30

Values ≥ 0.6 are highlighted in bold.

and was comprised only of the variable 'aggressive' with a high positive loading ≥ 0.6, identifying animals that reacted aggressively to the stimuli tested.

Intra-observer reliability

The Kendall's coefficient of concordance (*W*) was used to verify the consistency of the observers' view of cats' temperament. In PC1, intra-observer reliability ranged from high (*W* = 0.80 to 0.89) to very high (*W* = 0.91 to 0.94) (Table 2). In PC2 and PC3, the values were moderate (*W* = 0.50 to 0.66) to high (*W* = 0.71 to 0.82). Five of the 19 observers, including the 'gold' one, had very high or high values in all three PC.

In PC1, for the 'experienced' group (n = 7), the average *W* was 0.88 (ranging from 0.80 to 0.91) and for 'inexperienced' (n = 3) was 0.90 (from 0.89 to 0.91), revealing similarity for both groups in PC1 (Table 2). In PC2, the average *W* for 'experienced' was 0.68 (ranging from 0.53 to 0.78) and for 'inexperienced' was 0.70 (from 0.63 to 0.76), also revealing similarity for both groups. In PC3, the average *W* for 'experienced' was 0.64 (from 0.53 to 0.78) and for 'inexperienced' was 0.62 (from 0.54 to 0.73).

Regarding the ownership groups, in PC1 the average *W* for 'owners' (n = 7) was 0.91 (ranging from 0.88 to 0.94) and for 'non-owners' (n = 6) 0.85 (from 0.80 to 0.91), revealing values slightly higher for 'owners' (Table 2). In PC2, the average *W* for 'owners' was also slightly higher (0.70, ranging from 0.57 to 0.82) than for 'non-owners' (0.63, from 0.50 to 0.78). In PC3, similar *W* averages were found for 'owners' (0.65, ranging from 0.53 to 0.78) and 'non-owners' (0.64, ranging from 0.54 to 0.72).

Inter-observer reliability

The inter-observer reliabilities between each observer and the 'gold' were calculated based on the first QBA assessment. In PC1, all the 18 observers had high concordance (≥ 0.70) with the 'gold' being, on average, *W* = 0.84 (Table 3). In PC2, one observer had high concordance, 16 had moderate, and one had low concordance; the average *W* was 0.56. In PC3, eight observers had high, and ten had moderate values, and the average was 0.68 (Table 3). When the reliabilities were calculated for 'all observers' (n = 19 observers and 42 videos), a high value was found in PC1 (*W* = 0.71), whereas in PC2 and PC3, the coefficients were low (*W* = 0.21 and 0.29, respectively).

The reliability of each PC was also calculated within each group of experience and ownership. Similar and high values were found for the four groups in PC1 (Table 4). In PC2, the values of 'experienced' and 'inexperienced' groups were also similar. In contrast, the 'owners' had lower concordance than 'non-owners' (0.17 vs 0.34). In PC3, the coefficients were low, and the values were close for all groups (Table 4).

Discussion

One of the most important characteristics of any behavioural measuring tool for it to be an adequate method is sufficient reliability (Kaler *et al* 2009). Through PCA we obtained three principal dimensions of the cats' temperament, PC1 to PC3. The intra-observer reliability in PC1 showed high and very high concordances, while in PC2 and PC3, the coefficients varied among the observers. Regarding the inter-observer reliability, Kendall's coefficient of concordance (*W*) for all observers was high in PC1 and low in PC2 and PC3. Despite the low concordance in PC2 and PC3, the individual concordances with the gold observer showed moderate (≥ 0.4) to high values (≥ 0.7). The reliabilities for the groups of 'experience' and 'ownership' were similar in PC1 and PC3, while in PC2, the 'owners' had slightly higher intra-observer reliability but lower inter-observer concordance than the non-owners.

The first component extracted from PCA ranged from cats regarded as 'friendly / relaxed / affectionate / confident / curious / calm / greedy / sociable' to cats characterised as 'tense / stressed / fearful / nervous / suspicious.' Through this dimension, the cats could be distinguished based on their style of responses to the stimuli used in the tests (from more positive to negative responses). These results are similar to

https://doi.org/10.7120/09627286.31.3.004 Published online by Cambridge University Press

Table 2 Kendall's coefficients of concordance (*W*) for Qualitative Behaviour Assessment (QBA) intra-observer reliabilities of 13 observers (n = 42 cats observed twice).

Observers	Group		PC1	PC2	PC3
	Experience	Ownership			
1 ('Gold')	Experienced	Owner	**0.92**	**0.76**	**0.78**
2	Experienced	Owner	**0.92**	0.60	0.53
3	Experienced	Owner	**0.88**	0.57	0.56
4	–	Owner	**0.94**	**0.82**	**0.78**
5	Experienced	Owner	**0.89**	**0.76**	0.56
6*	Inexperienced	Owner			
7	Experienced	Non-owner	**0.80**	**0.78**	**0.72**
8	Experienced	Non-owner	**0.84**	**0.77**	**0.71**
9	–	Non-owner	**0.83**	0.52	0.59
10	–	Non-owner	**0.84**	0.50	0.62
11*	Experienced	Non-owner			
12	Experienced	Non-owner	**0.91**	0.53	0.66
13*	Inexperienced	Owner			
14*	Inexperienced	–			
15	Inexperienced	Non-owner	**0.89**	**0.73**	0.54
16	Inexperienced	Owner	**0.91**	0.63	0.60
17	Inexperienced	Owner	**0.91**	**0.76**	**0.73**
18*	Inexperienced	Non-owner			
19*	Inexperienced	Non-owner			

Values ≥ 0.7 are highlighted in bold. * Observers who did not participate in the second video session.

the PC2 found in the study of Arena *et al* (2019), who applied the QBA to dogs, whereby adjectives describing behavioural extremes ranged from 'comfortable / relaxed' to 'anxious / nervous / stressed.' Our results are in agreement with a study applying QBA in sheep (Diaz-Lundahl *et al* 2019), in which the PC1 ranged from 'calm / content / relaxed / friendly' to 'uneasy / vigilant / fearful', reflecting the valence of the emotional expressions, as in the present study. The QBA usually enables grouping the descriptors with positive semantic meanings (positive valence) versus negative meaning (negative valences). From a temperament perspective, it would enable distinguishing the calmer animals from the most fearful ones. The second component (PC2) consisted of the descriptors 'attentive / agitated / alert / active.' Similarly, a study on personality and interactions between domestic cats and their owners (Wedl *et al* 2011), whereby the principal component (PC1) was named 'active', consisted of the descriptors 'curious / active / playful / excitable / vigilant.' In general, both the PC1 of Wedl *et al* (2011) and the PC2 of the present study reflect behaviours indicating more activity, agitation, and attention to stimuli and could express the level of emotional arousal. In turn, the dimension of PC3 was characterised by the adjective 'aggressive.' Several previous papers reported cat temperament to be composed of a dimension characterised as 'aggression' (Arahori *et al* 2017; Ha & Ha 2017; Finka *et al* 2019; Salonen *et al* 2019). In Finka *et al* (2019), the second PC was named 'aggressiveness' since it included agonistic behaviours and a lack of handling tolerance. A correspondent dimension was also found in the study of Arahori *et al* (2017), including the adjectives 'irritable / moody / defiant / dominant / aggressive' in PC3 named 'roughness.'

When we analysed the intra-observer reliability using the Kendall's coefficient of concordance (*W*) we observed sufficient reliabilities in all three components. However, the values were higher PC1 than in PC2 and PC3. This pattern was not reported in some of the previous studies with QBA, in which more similar values were found for all of the components

Table 3 Kendall's coefficient of concordance between each observer and the 'gold' one in the three principal components (PC) (n = 42 cats).

Observers	Group		PC1	PC2	PC3
	Experience	Ownership			
1 ('Gold')	Experienced	Owner	–	–	–
2	Experienced	Owner	**0.93**	0.38	**0.74**
3	Experienced	Owner	**0.88**	0.60	0.49
4	–	Owner	**0.86**	0.48	0.68
5	Experienced	Owner	**0.84**	0.61	0.53
6	Inexperienced	Owner	**0.75**	0.61	0.65
7	Experienced	Non-owner	**0.76**	0.66	**0.75**
8	Experienced	Non-owner	**0.76**	**0.70**	0.63
9	–	Non-owner	**0.80**	0.52	0.62
10	–	Non-owner	**0.82**	0.60	**0.78**
11	Experienced	Non-owner	**0.86**	0.68	**0.76**
12	Experienced	Non-owner	**0.89**	0.53	0.67
13	Inexperienced	Owner	**0.84**	0.68	**0.75**
14	Inexperienced	–	**0.89**	0.58	**0.89**
15	Inexperienced	Non-owner	**0.85**	0.46	**0.75**
16	Inexperienced	Owner	**0.86**	0.49	**0.79**
17	Inexperienced	Owner	**0.86**	0.41	0.68
18	Inexperienced	Non-owner	**0.85**	0.41	0.52
19	Inexperienced	Non-owner	**0.85**	0.62	0.58

Values ≥ 0.7 are highlighted in bold type.

Table 4 Kendall's coefficients of concordance (*W*) for the observers' profile groups ('experienced' vs 'inexperienced' and 'owners' vs 'non-owners') in the three principal components (PC) (n = 42 cats).

Principal components	Experience		Ownership		All observers (n = 19)
	Experienced (n = 8)	Inexperienced (n = 8)	Owners (n = 9)	Non-owners (n = 9)	
PC1	0.71	0.75	0.71	0.74	0.71
PC2	0.27	0.31	0.17	0.34	0.21
PC3	0.37	0.35	0.30	0.32	0.29

(Diaz-Lundahl *et al* 2019; Ceballos *et al* 2021). Ceballos *et al* (2021) found intra-observer reliability ranging from $W = 0.71$ to 0.89 in PC1, $W = 0.74$ to 0.90 in PC2, and $W = 0.63$ to 0.84 in PC3. Diaz-Lundahl *et al* (2019) found intra-observer reliability ranging from $W = 0.89$ to 0.98 in PC1 and $W = 0.45$ to 0.93 in PC2. In these studies, the time interval between the QBA sessions was shorter, ranging from 7 to 15 days, respectively. It is possible to infer that this short interval between assessments could produce higher intra-observer reliability since it is more plausible that the observer had some degree of recollection of the behaviours and scores. In the present study, we used a longer time interval (ten months) to prevent the observer from memorising the behaviours observed. It is reasonable to consider the second assessment based on the observers' perceptions at a second session since they could not rely upon memories of the first assessment.

https://doi.org/10.7120/09627286.31.3.004 Published online by Cambridge University Press

Regarding the effects of groups in the intra-observer agreement, the effect of experience in behavioural assessment was not apparent, but when comparing 'owner' vs 'non-owner', we observed slightly lower intra-observer reliability in PC2 for the 'non-owners.' The group effect was also observed in the study of Diaz-Lundahl *et al* (2019), in which veterinary students had higher reliability in PC2. These findings may highlight the importance of familiarity with the behaviours of the species to perform the QBA. It is also possible to infer that among the non-owners, the interest in cats should be smaller than among the cat owners, leading the non-owners to reduce the attention and focus on the video clips over the observation time.

When we calculated each observer's concordances with the 'gold' using the Kendall's coefficient of concordance (*W*), we observed high and very high concordances in PC1. These results show that the observers were able to distinguish calmer cats from fearful ones, independently from owning cats (or not) or having had experience in behaviour assessment (or not). By calculating Kendall's coefficients in PC1 using the data of all observers and for the groups of experience and ownership, we obtained high concordances that confirm previous literature findings (Phythian *et al* 2013; Diaz-Lundahl *et al* 2019). These results demonstrate that all the selected profiles in this study could discriminate the cats' behavioural expressions in PC1. Bokkers and collaborators (2012), in turn, found slight concordance for PC1 after the first application of QBA by eight experienced observers. In further analysis, after the observers applied the QBA for some time in practical environments, the values increased from slight to low in the experienced group and to moderate in the inexperienced group (Bokkers *et al* 2012). These results were obtained using non-standardised videos (with varying quality, not recorded in a standardised way, showing just part of the herd). The authors, thus, made high-quality standardised videos (the herd recorded from four observation points in the barn) and conducted new analyses based on them. The observers' concordance coefficients increased to moderate. Based on the outcomes by Bokkers *et al* (2012), who showed how the characteristics of the video could influence the interpretation of behaviours, we can infer that the higher values of concordance in PC1 of the present study could be due to the standardisation of the situations the animals were exposed to, and also to the analysis of each animal singly instead of analysing groups of animals. Another possible influence for the high values of reliability in PC1 is the fact that it distinguished the animals' body languages better than PC2 and PC3. The PC1 explained 43.29% of variance in the data (while PC2 explained only 17.13% and PC3, 6.51%), and ranged from positive to negative adjectives. In the study by Arena *et al* (2019), the inter-observer reliability was moderate (0.61) in PC1, which explained 28.3% of the variance, showing high loadings only for positive descriptors. Their PC2 explained 25.9% of the variance, but had descriptors ranging from positive to negative loadings (from relaxed to stressed), which might explain the higher coefficient of concordance (0.80) in this component, similarly to the PC1 of the present study. Thus, it is possible that the observers agreed on components that better distinguish the valence of the expressions (positive vs negative behavioural and emotional expressions).

In PC2, the agreement between each observer with the gold one was moderate for most of them, but low (0.21) values were found when the agreement was calculated among all observers (n = 19). These results may suggest that the behaviours gathered in PC2 were subjected to more divergent interpretations. Perhaps the terms in PC2 were related to more subtle and less evident behavioural elements, such as 'attentive' and 'alert', which may cause divergences among the observers. A different pattern was reported by Phythian and collaborators (2013), in which Kendall's coefficient of concordance in PC2 was closer to the values in PC1. It is worth highlighting that both dimensions (PC1 and PC2) ranged from behaviours with high positive to high negative loadings. The observers analysed the animals' behaviours in 1-min videos and aimed to assess sheep welfare (Phythian *et al* 2013). It also is reasonable to infer that short-time video clips can limit the variety of behaviours expressed within each video and, therefore, might improve concordances by reducing the influence of divergences in observers' interpretations (Phythian *et al* 2013). Perhaps, another possible explanation should be that observers simply could not focus for long enough when asked to watch many 12-min videos. In spite of this possible disadvantage of the longer video clips in terms of observers' agreement, the use of different test situations is desirable to differentiate inter-individual behavioural differences when using QBA to assess temperament.

In PC3, each observer's agreement with the 'gold' one was high for 44.4% of them (8/18) and moderate for 55.6% (10/18) of the observers. This higher concordance in PC3 compared to PC2 could probably be linked to more conspicuous behaviours (aggressive acts such as hissing, slaps, bites), more easily identifiable through subjective assessments of cats' body language. In spite of PC3 getting better concordance between the 'gold' and each observer, the values were also low when we calculated the coefficient for all observers together, similarly to PC2. In PC1, we can clearly identify two opposite expressions, positive vs negative behavioural expressions. On the other hand, the PC2 and PC3 dimensions had only descriptors with high loadings in one behavioural extreme (positive or negative). This characteristic could partially explain the low concordance among the observers. We might infer that for scales ranging from two divergent extremes (opposite behavioural reactions), the agreement between observers can be more easily achieved (Arena *et al* 2019).

Animal welfare implications

Overall, observers were able to identify behavioural extremes ranging from calmer/friendly to more fearful/nervous cats. In practice, these results imply that any observer profile has the potential to be used to analyse the temperaments of cats. In a shelter, the temperament assessment should be used to improve the welfare of cats and raise the chances of successful adoptions. For instance, contact with humans might improve the welfare of friendly and calm cats, while fearful individuals would benefit from physical structures that promote hiding areas (Rochlitz 2000). The temperament of cats will also influence the adoption process (Gourkow & Fraser 2006; Weiss *et al* 2015; Evans *et al* 2019). It is important to inform and advise adopters about their pets' temperament at the time of adoption; this improves the formation of human-animal bonds and reduces the chances of unrealistic expectations by the owners (Weiss *et al* 2015). In order to improve the reliability of PC3 so that it can also be used in shelters, it would suggest training shelter employees about aggressive behaviours in cats, helping to identify and reduce interpretive bias. For future studies, it is advisable to check if, in a practical environment, managers and keepers can perform the QBA to characterise cat temperament and the impact of QBA assessment implementation in the shelter environment.

Conclusion

We concluded that the general view of the cats' temperament was reliable in PC1, which is the dimension that mostly explains the behavioural variations of temperaments. All the observers had very high or high intra- and inter-observer reliability in this dimension. This pattern did not occur in PC2 and PC3, where the reliability varied among the observers. Some observers identified the profiles in PC1 and PC2 with good reliabilities, enabling a practical application in shelters. Considering that the groups of experience and owners did not strongly differ regarding intra- or inter-observer reliabilities, we suggest that different profiles of observers are able to use QBA. Thus, shelter keepers could use the QBA to identify temperament traits in shelter cats, which could generate information with the potential to influence cat welfare positively. These animals would benefit from appropriate handling practices for each type of temperament.

Declaration of interest

None.

Acknowledgements

The authors wish to thank the shelter team, especially Maria José Toledo, for making their facility available to carry out the present research. We also thank the 19 observers who donated their time and participated with dedication. This study is part of ICT's Master's Thesis, prepared for the Graduate Program in Behavior and Animal Biology of the Universidade Federal de Juiz de Fora (UFJF), Brazil. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Arahori M, Chijiiwa H, Takagi S, Bucher B, Abe H, Inoue-Murayama M and Fujita K 2017 Microsatellite polymorphisms adjacent to the oxytocin receptor gene in domestic cats: Association with personality? *Frontiers in Psychology 8*: 2165. https://doi.org/10.3389/fpsyg.2017.02165

Arena L, Wemelsfelder F, Messori S, Ferri N and Barnard SJB 2019 Development of a fixed list of descriptors for the qualitative behavioural assessment of shelter dogs. *PLoS ONE 14(10)*: e0212652. https://doi.org/10.1371/journal.pone.0212652

Bokkers E, de Vries M, Antonissen I and de Boer I 2012 Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. *Animal Welfare 21*: 307-318. https://doi.org/10.7120/09627286.21.3.307

Ceballos MC, Góis KCR, Sant'Anna AC, Wemelsfelder F and da Costa MP 2021 Reliability of qualitative behavior assessment (QBA) versus methods with predefined behavioral categories to evaluate maternal protective behavior in dairy cows. *Applied Animal Behaviour Science* 105263. https://doi.org/10.1016/j.applanim.2021.105263

Clarke T, Pluske JR and Fleming PAJ 2016 Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment. *Applied Animal Behaviour Science 177*: 77-83. https://doi.org/10.1016/j.applanim.2016.01.022

Diaz-Lundahl S, Hellestveit S, Stubsjøen SM, Phythian C, Oppermann Moe R and Muri KJA 2019 Intra-and inter-observer reliability of Qualitative Behaviour Assessments of housed sheep in Norway. *Animals 9*: 569. https://doi.org/10.3390/ani9080569

Dwyer C, Ruiz R, Beltran de Heredia I, Canali E, Barbieri S and Zanella A 2015 *AWIN welfare assessment protocol for sheep.* https://air.unimi.it/bitstream/2434/269114/2/AWINProtocolSheep.pdf

Evans R, Lyons M, Brewer G and Tucci S 2019 The purrfect match: The influence of personality on owner satisfaction with their domestic cat (*Felis silvestris catus*). *Personality Individual Differences 138*: 252-256. https://doi.org/10.1016/j.paid.2018.10.011

Finka LR, Ward J, Farnworth MJ and Mills DS 2019 Owner personality and the wellbeing of their cats share parallels with the parent-child relationship. *PLoS ONE 14:* e0211862. https://doi.org/10.1371/journal.pone.0211862

Fukimoto N, Howat-Rodrigues AB and Mendonça-Furtado OJ 2019 Modified meet your Match® Feline-ality™ validity assessment: An exploratory factor analysis of a sample of domestic cats in a Brazilian shelter. *Applied Animal Behaviour Science 215*: 61-67. https://doi.org/10.1016/j.applanim.2019.03.013

Fukimoto N, Melo D, Palme R, Zanella AJ and Mendonça-Furtado OJ 2020 Are cats less stressed in homes than in shelters? A study of personality and faecal cortisol metabolites. *Applied Animal Behaviour Science* 104919. https://doi.org/10.1016/j.applanim.2019.104919

Gartner MC 2015 Pet personality: A review. *Journal Personality Individual Differences 75*: 102-113. https://doi.org/10.1016/j.paid.2014.10.042

Gartner MC and Weiss A 2013 Personality in felids: a review. *Applied Animal Behaviour Science 144*: 1-13. https://doi.org/10.1016/j.applanim.2012.11.010

Gourkow N and Fraser D 2006 The effect of housing and handling practices on the welfare, behaviour and selection of domestic cats (*Felis silvestris catus*) by adopters in an animal shelter. *Animal Welfare 15*: 371-377

Grosso L, Battini M, Wemelsfelder F, Barbieri S, Minero M, Dalla Costa E and Mattiello S 2016 On-farm Qualitative Behaviour Assessment of dairy goats in different housing conditions. *Applied Animal Behaviour Science 180*: 51-57. https://doi.org/10.1016/j.applanim.2016.04.013

Ha D and Ha J 2017 A subjective domestic cat (*Felis silvestris catus*) temperament assessment results in six independent dimensions. *Behavioural Processes 141*: 351-356. https://doi.org/10.1016/j.beproc.2017.03.012.

Kaler J, Wassink GJ and Green LE 2009 The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *The Veterinary Journal 180*: 189-194. https://doi.org/10.1016/j.tvjl.2007.12.028

Manly JFM 2008 *Métodos Estatísticos Multivariados: Uma Introdução, 3ª Edição*. Bookman: Porto Alegre, Brazil. [Title translation: Multivariate statistical methods: An introduction]

Martin P and Bateson P 2007 *Measuring Behaviour: An Introductory Guide, Third Edition* p 176. Cambridge University Press: Cambridge, UK. https://doi.org/10.1017/CBO9780511810893

Phythian C, Michalopoulou E, Duncan J and Wemelsfelder F 2013 Inter-observer reliability of Qualitative Behavioural Assessments of sheep. *Applied Animal Behaviour Science 144*: 73-79. https://doi.org/10.1016/j.applanim.2012.11.011

Rochlitz I 2000 Feline welfare issues. In: Turner DC and Bateson P (eds) *The Domestic Cat: The Biology of its Behaviour* pp 207-226. Cambridge University Press: Cambridge, UK

Salonen M, Vapalahti K, Tiira K, Mäki-Tanila A and Lohi H 2019 Breed differences of heritable behaviour traits in cats. *Scientific Reports 9*: 7949. https://doi.org/10.1038/s41598-019-44324-x

Stubsjøen SM, Moe RO, Bruland K, Lien T and Muri K 2020 Reliability of observer ratings: Qualitative behaviour assessments of shelter dogs using a fixed list of descriptors. *Veterinary and Animal Science 10*: 100145. https://doi.org/10.1016/j.vas.2020.100145

Travnik IC and Sant'Anna AC 2021 Do you see the same cat that I see? Relationships between Qualitative Behaviour Assessment and indicators traditionally used to assess temperament in domestic cats. *Animal Welfare 30*: 211-223. https://doi.org/10.7120/09627286.30.2.211

Walker J, Dale A, Waran N, Farnworth M, Clarke N and Wemelsfelder F 2010 The assessment of emotional expression in dogs using a free choice profiling methodology. *Animal Welfare 19*: 75-84

Wedl M, Bauer B, Gracey D, Grabmayer C, Spielauer E, Day J and Kotrschal K 2011 Factors influencing the temporal patterns of dyadic behaviours and interactions between domestic cats and their owners. *Behavioural Processes 86*: 58-67. https://doi.org/10.1016/j.beproc.2010.09.001

Weiss E, Gramann S, Drain N, Dolan E and Slater M 2015 Modification of the feline-ality™ assessment and the ability to predict adopted cats' behaviours in their new homes. *Animals 5*: 71-88. https://doi.org/10.3390/ani5010071

Welfare Quality® 2009 *Welfare quality® assessment protocol for cattle*. Welfare Quality® Consortium: Lelystad, The Netherlands. http://www.welfarequalitynetwork.net/media/1088/cattle_protocol_without_veal_calves.pdf

Wemelsfelder F, Hunter EA, Mendl MT and Lawrence AB 2000 The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science 67(3)*: 193-215. https://doi.org/10.1016/S0168-1591(99)00093-3

Wemelsfelder F, Hunter TE, Mendl MT and Lawrence AB 2001 Assessing the 'whole animal': A free choice profiling approach. *Animal Behaviour 62*: 209-220. https://doi.org/10.1006/anbe.2001.1741

Wemelsfelder F and Lawrence AB 2001 Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. *Acta Agriculturae Scandinavica, Section A-Animal Science 51*: 21-25. https://doi.org/10.1080/090647001300004763