



Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds

Stéphanie M. van den Berg^{a,*}, Henri C.M. Heuven^{b,1}, Linda van den Berg^{c,2}, Deborah L. Duffy^{d,3}, James A. Serpell^{d,3}

^a University of Twente, Faculty of Behavioural Sciences, Department of Research Methodology, Measurement and Data Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands

^b Utrecht University, Faculty of Veterinary Medicine, Department of Clinical Sciences of Companion Animals, P.O. Box 80163, 3508 TD Utrecht, The Netherlands

^c Leiden University Medical Center, Department of Pediatrics and Department of Human Genetics, S4-P, P.O. Box 9600, 2300 RC Leiden, The Netherlands

^d University of Pennsylvania, School of Veterinary Medicine, Department of Clinical Studies VHUP, Philadelphia, PA 19104-6010, USA

ARTICLE INFO

Article history:

Accepted 11 February 2010

Available online 15 March 2010

Keywords:

Item Response Theory

IRT

Rasch

C-BARQ

Canine

Aggression

ABSTRACT

Most test development in animal behaviour research is based on classical test theory. Modern test theory, also known as item-response theory, has however a number of advantages. Here, item-response theory was applied to confirm the stranger-directed aggression subscale of the Canine Behavioral Assessment and Research Questionnaire (C-BARQ). The sample consisted of a total of 1000 German Shepherds, Golden Retrievers and Labrador Retrievers. Results showed a good fit for a 1-parameter (Rasch) model. Although there were statistically significant violations of the measurement invariance assumption across breeds, the violations were only minor and without practical consequence. It is therefore concluded that the subscale consisting of 10 items can be used to quantify stranger-directed aggression in these three breeds, and that scores of different dogs can be compared meaningfully regardless of the dogs' sex, breed or neutering status.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Dogs are bred and trained to work in a variety of settings, including police work, drug detection, and guiding the blind. Behaviour is a crucial determinant of working dog performance. For instance, most guide dog failures are caused by behaviour problems such as fearfulness, distractibility, and aggression (Goddard and Beilharz, 1983).

These behavioural traits are heritable and they respond to selection (Trut, 2001; Gulevich et al., 2004; Courreau and Langlois, 2005; Sætre et al., 2006). Therefore, behavioural measures are useful to predict later success, to determine which dogs should enter the training program, and to select dogs for breeding.

Several tools have been developed to measure canine behaviour. Some are based on direct observation of the dog's response to several test situations (e.g. Svartberg and Forkman, 2002). Other methods focus on the assessment of day-to-day behaviour using a questionnaire for the dog owner. Both have advantages and disadvantages (see e.g. van den Berg et al., 2006). A widely used questionnaire is the canine behavioural assessment and research questionnaire (C-BARQ). C-BARQ contains items regarding aggression, fear and anxiety, trainability, excitability, separation-related behaviour, attachment,

* Corresponding author. Tel.: +31 53 489 2422; fax: +31 53 489 4239.

E-mail addresses: stephanie.vandenberg@utwente.nl

(S.M. van den Berg), h.c.m.heuven@uu.nl (H.C.M. Heuven),

l.van.den.berg@lumc.nl (L. van den Berg), duffydl@vet.upenn.edu

(D.L. Duffy), serpell@vet.upenn.edu (J.A. Serpell).

¹ Tel.: +31 30 2539782.

² Tel.: +31 71 526 9400.

³ Tel.: +1 21 5898 1004.

attention-seeking, and chasing. The items are based on clinical signs of behaviour problems (Hsu and Serpell, 2003; Serpell and Hsu, 2005; Duffy et al., 2008). The items describe everyday situations and the owner is asked to rate the dog's behaviour in the recent past on 5-point rating scales. A factor analysis of the C-BARQ items with varimax rotation revealed 11 factors (Hsu and Serpell, 2003). The factor with the highest Eigenvalue could be interpreted as stranger-directed aggression (SDA). The SDA factor was composed of ten items with practically identical factor loadings and a Cronbach's α of 0.93 (Hsu and Serpell, 2003). SDA scores were significantly higher in dogs that were diagnosed as aggressive to strangers by behaviour experts (Hsu and Serpell, 2003). The SDA subscale therefore seems a promising tool to measure individual differences in aggression towards strangers. For instance, it has been used in a study on the genetics of aggression in Golden Retrievers (van den Berg et al., 2008).

The SDA subscale was constructed based on classical test theory (i.e. factor analytic methods). There are a number of concerns regarding the application of classical test theory to the C-BARQ data. First, the assumptions underlying the factor analytic method are not met. The factor analytic method assumes a normally distributed latent variable that has a linear relationship with the observed item data, with error terms that are also normally distributed. The normality assumption is violated in the case of the C-BARQ: the distributions of item scores in the general dog population are highly skewed, with most dogs showing minimum scores (van den Berg et al., 2006). This is frequently observed with assessment tools for behaviour problems (Reise and Waller, 2009). In addition, a linear relationship between latent variable and item score may not be expected because the C-BARQ consists of ordinal data.

Second, it is unknown whether the same subscale can be used to quantify individual differences *within* different dog breeds. The factor analysis and the high coefficient for internal consistency was based on data from a wide variety of dog breeds, with large differences in average scores across breeds (Duffy et al., 2008). The subscale may perform worse when assessing aggression within breeds. A third concern is that the subscale may function differently across different breeds. There may not be only quantitative differences between breeds, but also qualitative differences. Dogs of a particular breed may for instance be particularly prone to bark, irrespective of whether it is related to aggression, fear or excitement. Such phenomena may cause bias when trying to compare two dogs of different breeds based on an averaged score on a set of items. Two averaged scores can only be compared when the measurement models are identical across the two groups. This assumption of equal measurement models is referred to as measurement invariance (Mellenbergh, 1989). Note that measurement invariance is more stringent than simply finding similar factor structures across breeds (cf. van den Berg et al., 2006). In the context of factor analysis, measurement invariance holds when in addition to structural similarity, both factor loadings and intercepts (means) are equal across subgroups.

The research presented here applies modern test theory to address the above concerns. Modern test theory, or Item Response Theory (IRT), supplies models specifically suitable for categorical and ordinal item data. To our knowledge, IRT has not been applied to animal behaviour before. Briefly, IRT models the observed data probabilistically, where the probability of a certain response on a particular item depends on a person parameter and one or more item parameters. For a first introduction to IRT, see for example Embretson and Reise (2000). Its main practical advantages, apart from the distributional assumptions about item data, include that scores from different test versions can be compared (e.g. in longitudinal research) and that when item data is missing at random, a score can still be estimated and compared with scores from dogs with complete data. The IRT framework is also suitable to test for measurement invariance: can the aggression subscale be used to quantify individual differences not only within separate dog breeds but also across breeds? The objective of this paper is therefore two-fold: (i) to determine whether the subscale for SDA can be confirmed using modern test theory (i.e. to determine whether IRT supports the hypothesis of a single trait underlying the item responses) and (ii) to test whether the subscale shows strict measurement invariance across breeds.

2. Material and methods

2.1. Measures

Behavioral data on stranger-directed aggression were collected from dog owners using the Canine Behavioral Assessment and Research Questionnaire (C-BARQ™). Details of this survey, its validation and reliability, have been described elsewhere (Hsu and Serpell, 2003; Serpell and Hsu, 2005; Duffy et al., 2008). Briefly, the C-BARQ is a standardized questionnaire designed to assess the prevalence and severity of behaviour problems in dogs. It consists of 101 items that ask owners to indicate how their dogs have responded "in the recent past" to a variety of common events and stimuli using a series of 0–4 rating scales. The wording of these rating scales differed depending on the questions: either in frequency terms (0 = never, 1 = seldom, 2 = sometimes, 3 = usually, and 4 = always) or in qualitative terms (0 = no signs of the behaviour, 1–3 = mild to moderate signs of the behaviour, and 4 = severe signs of the behaviour). For these rating scales, a brief explanation was included describing the behavioural signs involved in the behaviour (e.g., "Typical signs of moderate aggression in dogs include barking, growling, and baring teeth. More serious aggression generally involves snapping, lunging, biting, or attempting to bite.") Sixty-eight of these items were condensed by factor analysis into 11 distinct subscales (Hsu and Serpell, 2003). The stranger-directed aggression factor (SDA) comprises the following 10 questionnaire items: describe the extent to which you observe aggressive behaviour (1) when the dog is approached directly by an unfamiliar male adult while being walked or exercised on a leash; (2) when approached directly by an unfamiliar female adult while being walked or exercised on a leash; (3) when approached directly by an unfamiliar child while

being walked or exercised on a leash; (4) toward unfamiliar persons approaching the dog while it is in the owner's car; (5) when an unfamiliar person approaches the owner or a member of the owner's family at home; (6) when an unfamiliar person approaches the owner or a member of the owner's family away from home; (7) when mailmen or other delivery workers approach the home; (8) when strangers walk past the home while the dog is in the yard/garden; (9) when joggers, cyclists, roller skaters, or skateboarders pass the home while the dog is in the yard/garden; and (10) toward unfamiliar persons visiting the home.

2.2. Sample

The C-BARQ data were collected via an online version of the questionnaire (<http://www.cbarq.org>) that became freely available to pet owners in April 2006. The online survey was advertised via an article in the newsmagazine of the Veterinary Hospital of the University of Pennsylvania, USA (<http://www.vet.upenn.edu/bellwether/v64/article10.shtml>) and by notices sent to Philadelphia-area veterinary clinics, and the top 20 USA breed clubs based on AKC registrations. Availability of the survey then spread via word of mouth. Here we used data available on Golden Retrievers, Labrador Retrievers and German Shepherds. Since these are very popular breeds, there were enough data for an analysis with sufficient statistical power. Only dogs for which we had complete data on the 10 items for stranger-directed aggression were included in the analyses ($N = 1000$).

2.3. Statistical analysis

The data were analysed using models based on IRT. A simple IRT model is the Rasch model. It models dichotomous (1/0, YES/NO) item data with a probabilistic model where the probability of a positive response on a particular item is a logistic function of the latent score θ of the individual being tested and an item parameter β that indicates how prevalent a positive response is in the general population. This β parameter for an item is the IRT analogue of the mean (intercept) parameter in the classical factor analysis approach. It is generally known as the difficulty parameter. The Rasch model has no equivalent for the factor loading as it assumes that all items have the same correlation to the underlying trait θ . It also assumes local independence: given an individual i 's latent score θ_i and the item parameter β_j for item j , the probability of a positive response on that item j by that individual i is independent of all parameters of other items and other individuals' θ

scores.

$$p(\text{YES}_{ij}|\theta, \beta) = \frac{1}{1 + e^{\beta_j - \theta_i}} \quad (1)$$

In the Rasch model, all individual θ and β parameters are fixed effects that can be estimated without any assumption about their distribution, although it is possible to include such an assumption. The above simple Rasch model is unidimensional (individuals only have one latent score θ_i), so that if it fits the observed data, this supports the hypothesis of a unidimensional trait underlying the item responses. If the Rasch model does not fit, this could be an indication that the items are related to two or more independent traits. It is also possible that the data would fit the model better if one would allow for unequal correlations between the latent variable and the various items. Such a 2-parameter model can be parameterized as follows:

$$p(\text{YES}_{ij}|\theta, \alpha, \beta) = \frac{1}{1 + e^{\alpha_j(\beta_j - \theta_i)}} \quad (2)$$

For large values of α , the probability of a positive response increases faster with increasing θ than for smaller values of α . In other words, a large α indicates a high loading on the latent variable θ . Generally, parameter α is denoted as the discrimination parameter. To identify a 2-parameter model it is necessary to make an assumption about the shape of the distribution of the individual θ s and to fix the scale by choosing a variance. Here we used the assumption of a normal distribution for both kinds of models and unit variance for the 2-parameter model.

Both the Rasch model and the 2-parameter model can be extended to be able to deal with polytomous data, such as 5-point rating scales as we have for the C-BARQ. However, since frequencies for all categories other than 0 were low for the three breeds, the item data were dichotomized by collapsing the categories 1 through 4. The Rasch model and the 2-parameter models were tested on all data at once in order to test for measurement invariance across breeds. In IRT, a violation of this assumption is termed differential item functioning or DIF. Next, analyses were run on each breed separately in order to examine item fit statistics and to test measurement invariance across sex and neutering status. All IRT analyses were run using the MIRT software package, developed at the Department of Research Methodology, Measurement and Data Analysis at the University of Twente.

DIF, or violation of the measurement invariance assumption, was tested using Lagrange multiplier (LM) test statistics that are based on differences between observed and expected item scores between subgroups. This is equivalent to testing the hypothesis that item parameters are equal across subgroups (Glas, 1998). These LM

Table 1
Mean age, sex and neutering status per breed.

	Mean age in weeks (SD)	Female/male ratio	Not neutered/neutered ratio
Golden Retrievers	325 (179)	110/114	45/179
Labrador Retrievers	316 (167)	220/223	81/362
German Shepherds	305 (167)	174/159	101/232

Table 2

Parameter estimates (SE) and Lagrange multiplier tests for DIF across breeds based on the 2-parameter model.

Item	α parameter	β parameter	LM	df	<i>p</i>	Abs. dif
1	3.427 (0.326)	2.370 (0.240)	0.19	2	.91	0.00
2	2.655 (0.277)	3.065 (0.264)	10.71	2	<.01*	0.02
3	2.145 (0.192)	0.468 (0.129)	9.22	2	.01*	0.03
4	2.454 (0.226)	0.536 (0.138)	3.56	2	.17	0.02
5	2.926 (0.284)	1.598 (0.187)	5.27	2	.07	0.02
6	2.694 (0.245)	−0.992 (0.158)	0.69	2	.71	0.01
7	2.467 (0.224)	−1.191 (0.164)	11.43	2	<.01*	0.03
8	3.321 (0.368)	2.817 (0.293)	0.93	2	.63	0.01
9	2.161 (0.204)	−0.421 (0.135)	1.69	2	.43	0.01
10	2.422 (0.223)	1.208 (0.157)	3.90	2	.14	0.02

Note: Abs. dif is the average mean deviation between the expected and the observed probability to be in a particular response category, aggregated over a number of levels of the total test score.

* Statistically significant at an alpha of 0.05.

test statistics are asymptotically χ^2 -distributed. In case test statistics could not be computed with MIRT, the eRm package was used (Mair and Hatzinger, 2007).

3. Results

There were complete data on 224 Golden Retrievers, 443 Labrador Retrievers, and 333 German Shepherds as shown in Table 1. Sex and neutered ratios were similar across breeds.

A 2-parameter model was fitted first, for all breeds together, allowing only for any breed differences in average aggression and variance of scores (with a 2-parameter model, only one of the breed variances needs to be fixed). Tables 2 and 3 show estimates for the model parameters. Table 2 shows results for the assumption of measurement invariance, i.e. if the IRT model parameters for the items are different among the three breeds and if breeds differ with respect to average score and variance or both. For three of the ten items there was a significant violation of the measurement invariance assumption. However, these violations were only minor: absolute differences (abs. dif) between expected and observed means were all well below 0.10. The 2-parameter model therefore seems to fit the data reasonably well. Table 3 shows that German Shepherds have significantly higher scores for aggression than the Labradors, and that the Labradors in turn have significantly higher scores than the Golden Retrievers. In an absolute sense, the variances are more or less equal across breeds, although statistically Labradors show significantly more variation than the German Shepherds and the Golden Retrievers. The α parameters (discrimination parameters, comparable to factor loadings) were all very similar (Table 2), which suggests that a Rasch model would fit the data probably equally well.

Table 3Means and standard deviations of θ parameters based on the 2-parameter model.

	Mean (SE)	Standard deviation
German Shepherds	0 (fixed)	1 (fixed)
Golden Retrievers	−0.875 (0.112)	1.057 (0.099)
Labrador Retrievers	−0.556 (0.090)	1.299 (0.091)

Table 4

Parameter estimates (SE) and Lagrange multiplier tests for DIF across breeds based on the 1-parameter model.

Item	β parameter	LM	df	<i>p</i>	Abs. dif
1	1.955 (0.168)	0.64	2	0.73	0.00
2	3.023 (0.180)	9.22	2	0.01*	0.02
3	0.483 (0.145)	6.66	2	0.04*	0.02
4	0.540 (0.143)	3.01	2	0.22	0.01
5	1.476 (0.161)	5.07	2	0.08	0.02
6	−0.945 (0.145)	0.99	2	0.61	0.01
7	−1.225 (0.155)	10.58	2	0.01*	0.03
8	2.373 (0.163)	1.28	2	0.53	0.01
9	−0.510 (0.148)	2.95	2	0.23	0.02
10	1.244 (0.154)	3.74	2	0.15	0.02

* Statistically significant at an alpha of 0.05.

Table 4 shows results for a 1-parameter (Rasch) model. Again, there does not appear to be any serious violation of the measurement invariance assumption across breeds. Three out of 10 items showed significant differential item functioning, but the deviations from expected were only small. Largest DIF was observed for item 7. A closer analysis revealed that German shepherds, although they have higher scores in general, when correcting for total score, have relatively lower scores on item 7 (results not shown). Conversely, while Golden Retrievers have generally lower total scores, they more often show aggressive behaviour on this item than the other two breeds when adjusting for total score (Table 5).

Next, the 1-parameter model was fitted separately for each breed. Table 6 shows Lagrange multiplier fit statistics for the Labradors and German Shepherds. Response category frequencies were too low for the Golden Retrievers to obtain fit statistics. A slightly different test was therefore used in order to assess overall model fit for the Golden Retrievers. This test compares model parameters for the 50% lowest scoring dogs (based on the sum score), and the 50% highest scoring dogs per breed (the LM test in the eRm

Table 5Means and standard deviations of individual θ parameters based on the 1-parameter model.

	Mean (SE)	Standard deviation
German Shepherds	0 (fixed)	2.658 (0.147)
Golden Retrievers	−2.206 (0.253)	2.666 (0.217)
Labrador Retrievers	−1.402 (0.211)	3.329 (0.185)

Table 6

Item fit statistics for two breeds separately based on a 1-parameter model.

Item	Labradors				German Shepherds			
	LM statistic	df	<i>p</i>	Abs. dif	LM statistic	df	<i>p</i>	Abs. dif
1	2.83	1	0.09	0.02	6.26	1	0.01*	0.04
2	NA	–	–	0.00	0.00	1	0.97	0.00
3	1.59	1	0.21	0.02	3.44	1	0.06	0.04
4	0.01	1	0.94	0.00	5.05	1	0.02*	0.04
5	0.36	1	0.55	0.01	1.53	1	0.22	0.02
6	2.54	1	0.11	0.02	1.13	2	0.57	0.02
7	0.12	1	0.73	0.00	3.18	2	0.20	0.03
8	NA	–	–	0.02	2.37	1	0.12	0.02
9	3.22	1	0.07	0.03	5.91	1	0.02*	0.04
10	1.13	1	0.29	0.01	1.02	1	0.31	0.02

Note: NA = Not available due to low response frequencies.

* Statistically significant at an alpha of 0.05.

package by Mair and Hatzinger, 2007) and they were not significantly different, LR = 2.94, df = 6, $p = .82$.

Separate analyses were done to check for measurement invariance across sex and neutering status. For the Labradors, none of the items showed differential item functioning across sex; the lowest p -value was .19 for item 2, with an absolute difference of 0.02. Across neutering status, lowest p -value was .05 again for item 2, with an absolute difference of 0.02.

For the German shepherds, lowest p -value for differential item functioning concerning sex was .14 for item 9, with an absolute difference of 0.03. For neutering status, only one out of 10 items showed a significant violation of the measurement invariance assumption for neutering status, item 3, at a p -value of .03 with an absolute difference of 0.05.

For Golden Retrievers, lowest p -value for differential item functioning concerning sex was .07 for item 4, with an absolute difference of 0.03. For neutering status, one item showed a significant violation of the measurement invariance assumption for neutering status, item 7, at a p -value of <.01 with an absolute difference of 0.07.

4. Discussion

Stranger-directed aggression was the factor extracted from the C-BARQ with the highest Eigenvalue and this subscale showed a Cronbach's α of .93 (Hsu and Serpell, 2003). These results were based on classical test theory, and not taking into account the non-normality of the item distributions. Here we confirmed the quality of that subscale using modern test theory. The dichotomized data fitted the Rasch model well. This 1-parameter IRT model is more suitable for assessing the quality of a scale than factor analysis since it is able to handle non-continuous data. The good fit of the Rasch model justifies the use of a summed score or an averaged score as a measure for stranger-directed aggression (after dichotomization).

Here we chose to dichotomize the data prior to analysis. This was done because the response frequencies for categories 1 and higher were very low. Although it is easy to apply a more elaborate model that would be able to handle more than 2 categories, we chose not to do this, since – due to the low response frequencies – estimates would be

very unreliable. It is however unfortunate to neglect information like this. A possible solution would be to alter the phrasing of the response category labels in order to get a more uniform spread across categories, although it should be noted that the three breeds analysed here showed relatively low scores for this subscale compared to other breeds (Duffy et al., 2008). Making the questionnaire more suitable for these three breeds would automatically render it less suitable for other breeds.

The sum score or average score on these ten items can be used to compare dogs of different breeds, at least across German Shepherds, Golden Retrievers and Labrador Retrievers. In other words, the sum score can be used to determine which dog is most aggressive towards strangers, regardless of breed. This is also true for comparison across sex and neutering status. This was shown by testing the measurement invariance assumption. In IRT, a violation of this assumption is termed differential item functioning or DIF. Generally, there were very few indications of DIF. Across breeds there was statistically significant DIF, particularly for item 7 which pertains to aggressive behaviour in a situation where a delivery worker approaches the home. German Shepherds and some other sheepdogs have an established reputation for being more protective of their owners compared to most dogs that are typically more protective of their homes or territories. Here, this effect was confirmed: corrected for overall aggression, German Shepherds are less likely to show aggressive behaviour to such a stranger approaching the home than Labradors and Golden Retrievers. The size of this effect, although significant in a statistical sense, was small enough not to cause problems when measuring behaviour with the subscale including this item.

Measurement invariance was also checked across sex and neutering status. The most serious violation was for a comparison across neutered and non-neutered Golden Retrievers for item 7. But again, the degree of violation was not so serious as to discourage using the sum score for comparison across neutered and non-neutered Golden Retrievers.

Whether the C-BARQ subscale for stranger-directed aggression is also suitable for other dog breeds remains to be investigated. But given that the scale has already good measurement properties for dogs that generally score

low on aggression, with very little variability on the rating scales, we expect that the quality will even be higher in more aggressive breeds, because then the rating scales provide more information and no dichotomization will be needed.

5. Conclusions

Summarizing, the stranger-directed subscale of the C-BARQ is a reliable tool to assess stranger-directed aggression in dogs. It can be used to identify the best and worst performing dogs regardless of sex, neutering or breed (at least regarding the three most common breeds studied here). Combined with other information, C-BARQ sum scores after dichotomization can be used to select the animals for training and/or breeding programs.

Here item-response theory was used to assess measurement properties of a subscale of the C-BARQ. Apart from being suitable for categorical and ordinal data (as one most often has with animal behaviour studies), IRT has a number of other advantages that may be of interest to researchers in animal behaviour. For instance, IRT provides a framework that is able to handle missing data very effectively. This is due to a nice property of IRT: θ s can be estimated regardless of what items are included in a scale, as long as they measure the same trait. This allows, for instance, combining data from animals that were tested using different versions of a questionnaire (or behavioural test). For an overview of the application of IRT in clinical studies with their particular problems of for instance skewed distributions, see [Reise and Waller \(2009\)](#). IRT also has advantages particularly useful for genetic studies, see for example [van den Berg et al. \(2007\)](#) ([van Leeuwen et al., 2008](#)). Testing for measurement invariance is easily done within the IRT framework. Measurement invariance is, for example, very important when one wants to take a two-stage approach in gene mapping ([Karlsson et al., 2007](#)): first finding a broad genomic region using dogs of one breed (taking advantage of extended linkage disequilibrium across the genome in one breed), and next fine-mapping using dogs of different breeds, taking advantage of the large inter-breed genomic differences. Phenotypic data from different breeds can only be combined when one can show that the scores can be compared across groups. IRT provides the framework that makes this possible.

References

- Courreau, J.-F., Langlois, B., 2005. Genetic parameters and environmental effects which characterise the defence ability of the Belgian shepherd dog. *Appl. Anim. Behav. Sci.* 91, 233–245.
- Duffy, D.L., Hsu, Y., Serpell, J.A., 2008. Breed differences in canine aggression. *Appl. Anim. Behav. Sci.* 114, 441–460.
- Embretson, E.S., Reise, S.P., 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum, Mahwah, NJ.
- Glas, C.A.W., 1998. Detection of differential item functioning using Lagrange multiplier tests. *Stat. Sin.* 8, 647–667.
- Goddard, M.E., Beilharz, R.G., 1983. Genetics of traits which determine the suitability of dogs as guide-dogs for the blind. *Appl. Anim. Ethol.* 9, 299–315.
- Gulevich, R.G., Oskina, I.N., Shikhevich, S.G., Fedorova, E.V., Trut, L.N., 2004. Effect of selection for behavior on pituitary–adrenal axis and proopiomelanocortin gene expression in silver foxes (*Vulpes vulpes*). *Physiol. Behav.* 82, 513–518.
- Hsu, Y., Serpell, J.A., 2003. Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *J. Am. Vet. Med. Assoc.* 223, 1293–1300.
- Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H.C., Zody, M.C., Anderson, N., Biagi, T.M., Patterson, N., Pielberg, G.R., Kulbokas, E.J., Comstock, K.E., Keller, E.T., Mesirov, J.P., von Euler, H., Kampe, O., Hedhammar, A., Lander, E.S., Andersson, G., Andersson, L., Lindblad-Toh, K., 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat. Genet.* 39, 1321–1328.
- Mair, P., Hatzinger, R., 2007. Extended Rasch modeling: The eRm package for the application of IRT models in R. *J. Stat. Software* 20, 1–20.
- Mellenbergh, G.J., 1989. Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143.
- Reise, S.P., Waller, N.G., 2009. Item response theory and clinical measurement. *Ann. Rev. Clin. Psychol.* 5, 27–48.
- Sætre, P., Strandberg, E., Sundgren, P.-E., Pettersson, U., Jazin, E., Bergström, T.F., 2006. The genetic contribution to canine personality. *Genes Brain Behav.* 5, 240–248.
- Serpell, J.A., Hsu, Y., 2005. Effects of breed, sex, and neuter status on trainability in dogs. *Anthrozoös* 18, 196–207.
- Svartberg, K., Forkman, B., 2002. Personality traits in the domestic dog (*Canis familiaris*). *Appl. Anim. Behav. Sci.* 79, 133–155.
- Trut, L.N., 2001. Experimental studies of early canid domestication. In: *Ruvinsky, A., Sampson, J. (Eds.), The Genetics of the Dog*. CABI Publishing, Oxon, NY, pp. 15–41.
- van den Berg, L., Schilder, M.B.H., de Vries, H., Leegwater, P.A.J., van Oost, B.A., 2006. Phenotyping of aggressive behavior in golden retriever dogs with a questionnaire. *Behav. Genet.* 36, 882–902.
- van den Berg, S.M., Glas, C.A.W., Boomsma, D.I., 2007. Variance decomposition using an IRT measurement model. *Behav. Genet.* 37, 604–616.
- van den Berg, L., Vos-Loohuis, M., Schilder, M.B.H., van Oost, B.A., Hazewinkel, H.A.W., Wade, C.M., Karlsson, E.K., Lindblad-Toh, K., Liinamo, A.-E., Leegwater, P.A.J., 2008. Evaluation of the serotonergic genes *htr1A*, *htr1B*, *htr2A*, and *slc6A4* in aggressive behavior of golden retriever dogs. *Behav. Genet.* 38, 55–66.
- van Leeuwen, M., van den Berg, S.M., Boomsma, D.I., 2008. A twin-family study of general IQ. *Learn. Ind. Differ.* 18, 76–88.