This is the final version of the manuscript sent to journal for publication after acceptance. Note the final printed version might be different from this one due to the editing process.

The full citation: Szabó D., Mills, D. S., Range. F., Virányi, Zs., MIklósi. Á. 2017. Is a local sample internationally representative? Reproducibility of four cognitive tests in family dogs across testing sites and breeds. Animal Cognition, doi: 10.1007/s10071-017-1133-3

Is a local sample internationally representative? Reproducibility of four cognitive tests in family dogs across testing sites and breeds

Authors

Dóra Szabó[a], Daniel S. Mills[b], Friederike Range[c], Zsófia Virányi[c], Ádám Miklósi[a,d]

Affiliations

a Department of Ethology, Eötvös Loránd University, Budapest, Hungary

b University of Lincoln, School of Life Sciences, Lincoln, Lincolnshire, United Kingdom

c Messerli Research Institute, University of Veterinary Medicine Vienna, Medical University of Vienna and University of Vienna, Vienna, Austria

d MTA-ELTE Comparative Ethology Research Group, Budapest, Hungary

* Corresponding author. E-mail: szaboodoora@gmail.com

Tel: +36 1 381 21 79

Fax: +36 1 381 21 80

Keywords: behaviour tests, reproducibility, replicability, dog, cross-country comparison, breed comparison

Abstract

A fundamental precept of the scientific method is reproducibility of methods and results, and there is growing concern over the failure to reproduce significant results. Family dogs have become a favoured species in comparative cognition research, but they may be subject to cognitive differences arising from genetic (breeding lines) or cultural differences (e.g. preferred training methods). Such variation is of concern as it affects the validity and generalisability of experimental results. Despite its importance, this problem has not been specifically addressed to date. Therefore, we aimed to test the influence of three factors on reproducibility: testing site (proximal environment), breed and sex (phenotype). The same experimenter tested cognitive performance by more than two hundred dogs in four experiments. Additionally, dogs' performance in an obedience task administered by the owner. Breed of dog and testing site were found to influence the level of performance only mildly, and only in the means-end experiment and in the obedience task. Our findings demonstrate that by applying the same test protocols on sufficiently large samples, the reported phenomenon in these cognitive tests can be reproduced, but slight differences in performance levels can occur between different samples. Accordingly, we recommend the utilization of well-described protocols supported by video examples of the whole experimental procedure. Findings should focus on the main outcome variables of the experiments, rather than speculating about the general importance of small or secondary performance outcomes which are more susceptible to random or local noise.

Acknowledgement

This research was supported by the European Union and the State of Hungary, co-financed by ESF Research Networking Programme "CompCog": The Evolution of Social Cognition (www.compcog.org) (06-RNP-020), and the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 'National Excellence Program'. Dóra Szabó also

received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 680040), Ádám Miklósi also received support from the Hungarian Academy of Sciences (MTA 01 031), and Friederike Range received financial support from the FWF (project number: P24840-B16). We would like to thank Judit Abdai, Camille Hansart, Beáta Korcsok, Krisztina Kovács and Flóra Szánthó for their help. We are also grateful to all owners and their dogs for their participation in the study.

Introduction

Reproducibility is a fundamental scientific principle of topical importance given the current challenges to the value of science and claims of false discoveries which are later not substantiated; however this topic appears to be addressed rarely in the literature (Casadevall and Fang 2010). A recent large scale study revealed doubts about the reproducibility of experiments carried out in the behaviour sciences (Open Science Collaboration 2015) and serious concern about the inferences being drawn on single laboratory studies have been highlighted in a far reaching publication by Kafkafi et al. (2017). Studies on reproducibility are scarce because they are often not considered novel enough to publish, although scientists agree this information is beneficial for the research community (Moonesinghe et al. 2007). While the issue of reproducibility of animal experiments where behaviour is also analysed (e.g. assessing drug efficacy) has been addressed recently (Crabbe et al. 1999; Wahlsten et al. 2006; Baldini et al. 2013; Tuyttens et al. 2014; Kafkafi et al. 2017), we are not aware of a similar initiative regarding tests within the field of animal cognition.

The terms "reproducibility" and "replicability" are used differently across disciplines, which can lead to confusion. In natural sciences (see Yang et al. 2008; Richter et al. 2009; Casadevall and Fang 2010) the terms are used *sensu* Casadevall and Fang (2010) where "…*reproducibility* refers to a phenomenon that can be predicted to recur even when experimental conditions may vary to some degree". While "*replicability* describes the ability to obtain an identical result when an experiment is performed under precisely identical conditions". Unfortunately, the same terms are used with opposite meaning in the social sciences (e. g. Asendorpf et al. 2013; Klein et al. 2014). Reproducibility closely resembles the concept of *systematic replication* as introduced by Sidman (1960)

The factors influencing the reproducibility of behavioural studies include: (1) proximal environment and (2) phenotype (van der Staay et al. 2010). The first may include any type of

deviation from the original methodology or from the testing environment including the pre-training procedures (if any), or how variables are defined and used. Phenotype refers to differences between the populations being tested such as different strains/lines, age, sex, human-animal relationship or previous experience of the subjects.

The recently raised concern regarding reproducibility (Open Science Collaboration 2015) is especially relevant to canine research because dogs, compared for example to chimpanzees, are tested in significantly higher numbers and in a wider range of laboratories, often in superficially similar tests worldwide. They do not live in the laboratory, and are a very heterogeneous sample in regard to their anatomical features (including surgical alteration), previous experiences and genetic background (Miklósi & Topál, 2013).

Results of previous studies provide evidence that both proximal environment (such as local differences and slight deviations between the applied protocols) and phenotype can influence dogs' performance in cognitive tests. Whether dogs tested in different countries vary in regard to their performance in cognitive tests has not been extensively investigated yet, although cultural differences were shown to affect attachment behaviour in Austrian and Hungarian family dogs (Horn et al. 2013a). Fujita et al. (2012) have also reported differences in performance between Japanese and German dogs in an incidental memory test. However, local (hereafter also used to describe effects at a national level) effects are often incidental and not the primary focus of the study design. The dog-human relationship is also known to affect dogs' performance in cognitive tests (Topál et al. 1997; Horn et al. 2013b), and local variation in dog management practices (e.g. tendency to use food treats in training) has the potential to influence reproducibility, too.

The effect of deviations between the applied protocols was demonstrated in a two-way choice task, where differences in methodology (utilization of a clicker) influenced dogs' behaviour (Pongrácz et al. 2013). Their performance was significantly better in the pointing with clicker

condition than in the standard momentary distal pointing paradigm. While in both cases dogs could choose from two containers (correct choice indicated by the pointing), in the 'pointing with clicker' condition the indicated container was not baited. Instead, when a correct choice was made, the experimenter clicked and delivered the treat into the container.

The effect of dog phenotype was recently demonstrated by Fadel et al. (2016), where the authors reported differences in trait impulsivity between Border Collies and Labrador Retrievers, and between working and show lines within these breeds. The role of previous training experience and breed group regarding dogs' performance in cognitive tests has also been described (Marshall-Pescini et al. 2016). Trained dogs were faster in solving a detour task, while working breeds were performing better in a manipulation task than retriever and herding breeds.

Sex differences in cognitive performance have also been found in some setups, with male dogs performing better when presented with a novel manipulation task (Duranton et al. 2015) and female dogs being more sensitive toward size constancy violation (Müller et al., 2011).

Aim & Hypotheses

We tested the influence of three factors: breed, gender (phenotypic factors) and testing site (proximal factor) on the reproducibility of dog cognition results using a systematic approach. Using the same experimenter and equipment, we compared performance in four cognitive tasks and assessed differences in owner-instructed obedience task, using three comparable dog groups (Border collies, Labrador retrievers; various other breeds based on local availability) of both sexes and neuter statuses (phenotypic features) at three different testing sites (Hungary, Austria and Britain- proximal environments). This study allowed us to calculate the potential magnitude of differences that are not due to experimenter or equipment differences, and their relevance.

If the proximal factor has a significant influence on reproducibility, then we expected general testing site differences irrespective of breed. If breed has a major influence, we expected different performance in the pure breeds (Border collie and Labrador retriever) regardless of testing sites, but no such difference between the mixed breed groups in the different countries because this group consisted of various dog breeds kept as companions in human families. If sex/ neuter status influenced the cognitive performance, these differences were expected to be present at each testing site. An interaction between testing site and breed would suggest, for example, local genetic effects or differences in dog keeping habits. Differences in the obedience test would also reflect differences in dog keeping practices, as these basic obedience tasks measure one aspect of the dog-human relationship, that is, the owner's ability to control the dog in a novel environment instead of dog cognition.

General methods

Subjects

See Table 1 for the number of subjects included in the analysis listed by testing sites and breed. We tested dogs on three testing sites in three different countries: Budapest (Hungary-HU) and Vienna (Austria-AT) are capitals, both with a population of about 1.7 million, while Lincoln (United Kingdom-UK) has an estimated population of 93,000 (2011 census figures applying to defined administrative area). We recruited family dogs over 1 year of age without specific advanced level training from the following three populations: Border collies, Labrador retrievers and any other purebred dogs. We decided to test two popular breeds, which were easily available at all three sites because single-breed groups are genetically more homogenous. The third group (in which several breeds were represented) was targeted to resemble more closely the variability of samples currently being used in family dog studies in different countries. The dogs were required to be motivated by food. The subjects were recruited locally via social media, flyers, and radio advertisements. The testing took place on a single occasion,

and while we tested every dog in all tests, some of them needed to be excluded from one or more tests during data analysis.

Sex of the subjects was balanced across breed groups and testing sites as best as possible. Due to differences in dog keeping practices, the percentage of neutered animals (of both sexes) recruited was higher in the UK sample. We combined sex and reproductive status into a single variable with four categories.

Testing sites

The testing room in Lincoln was secluded, while the other two testing sites were located in laboratories where there was occasional movement and some minor noise in the corridor (Fig 1). For logistical reasons the tests were carried out in one country after another in the following order: Budapest-Vienna-Lincoln-Budapest.

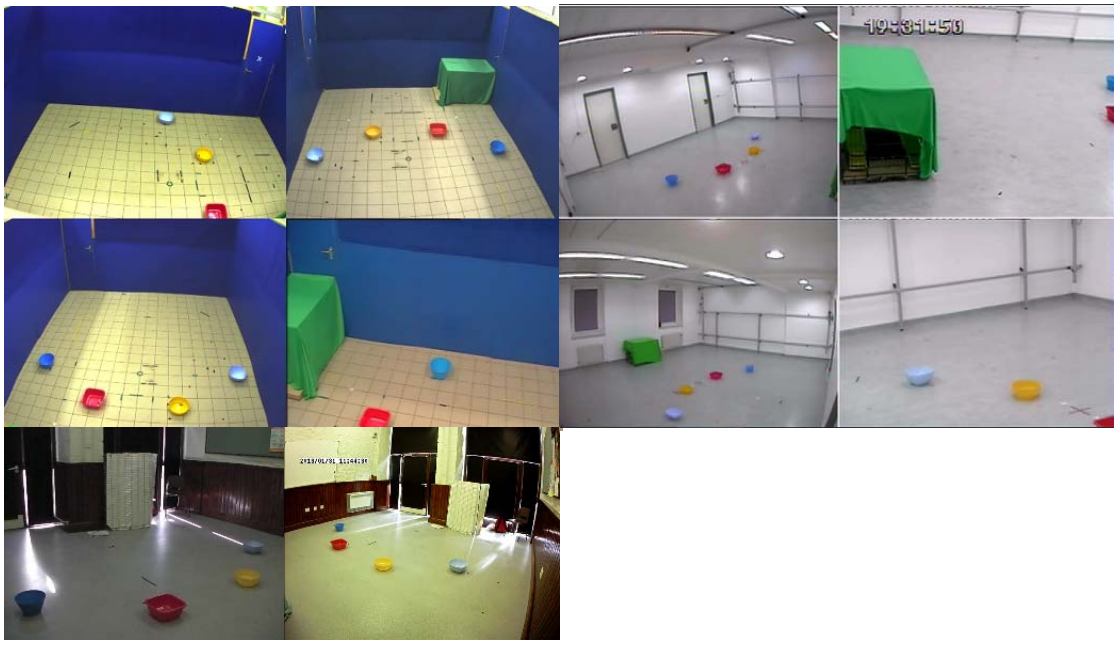

Figure 1 The three testing rooms from the position of recording cameras and room dimensions: Top left Budapest (3.6 m x 4.6 m), Top right Vienna (6 m x 7.2 m), Bottom line Lincoln (5.2 m x 5.9 m)

Test procedures

We looked for tests in the literature that met the followings conditions:

(1) The test could be conducted without extensive pre-training. This was necessary because dogs visited the testing site on a single occasion to avoid dropout.

(2) The test does not last longer than 15 minutes, and is not too exhausting for the dogs.

(3) The test required minimal equipment as it had to be transported from site to site. Since all tests took place in the same room, the setup for each test had to be easy/quick to build up and remove.

(4) The tests overlap or interfere with each other as little as possible. The tests should not rely on the same type of manipulative skills or have the same setup (e.g. two-way choice tasks). The tasks were provided in a fixed order with short breaks in between to standardise and minimize any carryover effect.

(5) The tests cover different facets of dog cognition. We intended to maximize the scope of the gathered behavioural data within the project.

(6) The reported performance of the dogs in the original publication was moderate but above chance at the group level. This was necessary in order to avoid ceiling and floor effects.

Based on these conditions we selected the following tests (described in detail later)

- *Pointing test:* following the human pointing gesture in a communicative situation, to assess behavioural flexibility in a social situation (Brúder 2010)
- *Problem solving test*: solving a problem without demonstration and after witnessing human demonstration of the solution, to assess problem solving abilities in a social learning context (Pongrácz et al. 2012)
- *Means-end test*: pulling out the baited one from two slides (based on visual cues in a support problem task paradigm), to assess physical cognition (Range et al. 2011).
- *Memory test*: choosing from four previously investigated bowls after a 10-minutes long break, to assess memory capacity (Fujita et al. 2012)

- *Obedience test*: testing the dog in a set of basic obedience tasks by the owner, to assess owner control. This test was not intended to measure cognitive abilities as there was no training involved, but was used to assess the level of control the owner possessed over the dog (Fukuzawa et al. 2005), to be able to detect possible differences in dog management practices between the populations

Right before testing, dogs participated in the pre-training phase of the Means-end test, while the presentation phase of the memory test took place before the Obedience test.

The protocols were mainly reproductions of already published studies, sometimes with slight modifications of the original protocol. We decided to use only food rewards (Frolic® Dog Food) to avoid losing subjects that are not both food and toy motivated. Tests were carried out between Augustus 2012 and June 2013.

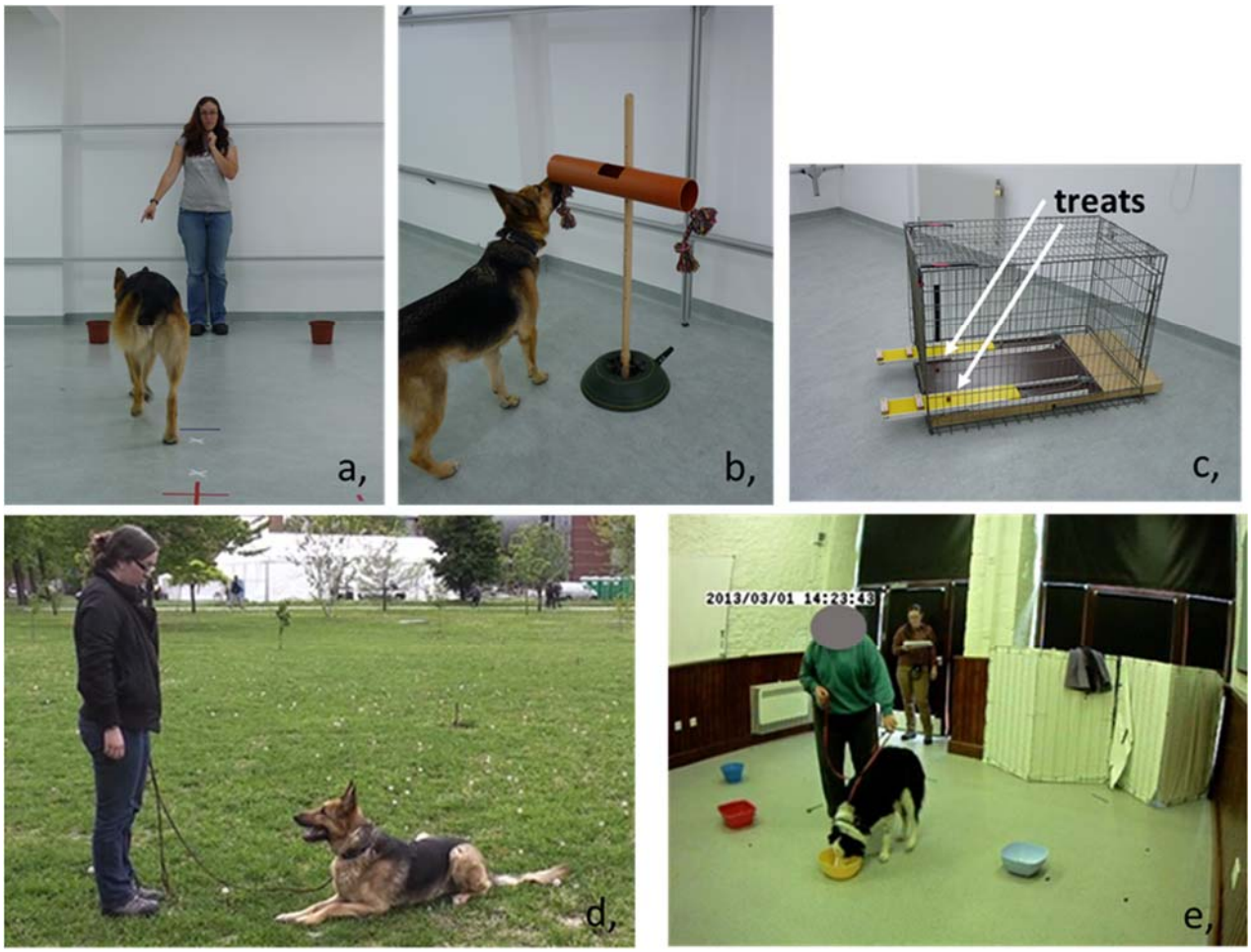

Figure 2 Demonstration of the equipment used during the project. a, Setup of the pointing test b, Setup of the problem solving test c, The apparatus used in the means-end test d, Setup of the obedience task e, Setup of the incidental memory test

Data collection and analysis

The tests were recorded via video cameras, and coded with the coding program Solomon©. Choice proportions are reported as percentages ± standard deviation. The statistical analyses were carried out with SPSS 21 and JASP 0.8.0.1. A priori sensitivity analysis and effect sizes were calculated with G*Power 3.1.9.2.

We coded the same behavioural variables as those coded in the original studies and compared our findings to those, using the same statistical methods where applicable. We used Generalised Linear Models with Poisson distribution and loglinear link to investigate the effects of proximal and phenotype-related factors. Model building was carried out via backward model selection.

The initial factors were the following: testing site, breed, country, sex, testing site x breed. To compare the probability of the null hypothesis (no difference between the samples) and the Bayesian probability of the alternative hypotheses, Bayesian ANOVA was carried with the following fixed factors: testing site, breed, sex, testing site x breed. Any deviations from this procedure are described in the relevant test section.

Inter-observer coding

Four trained coders coded 20% of the videos and Cohen's kappas (linear weighed in case of obedience scoring and unweighted for the rest of the variables) were calculated. This yielded excellent agreements (k>=0.75) between observers in all measured variables (exact values for the individual variables can be found in the appendix).

Ethical approval

The study was approved by the institutional ethics and animal welfare committee at the University of Veterinary Medicine Vienna (11/10/97/2012) and by the School of Life Sciences Ethics Committee at the University of Lincoln, UK (UID COSREC146). According to the Hungarian Animal Protection Act ("1998. évi XXVIII. Törvény", 3. §/9.), which defines experiments on animals, our non-invasive observational study was not considered as an animal experiment and thus did not require approval.

Test 1: Response flexibility in utilising human communicative signals

Method

We used a dynamic, distant pointing test to investigate how accurately dogs follow the experimenter given cues and how flexibly they can use the human pointing gesture. The protocol is based on Brúder (2010). This test investigates two aspects of dogs' cognition: (1) performance in utilizing a simple communicative signal and (2) dogs' ability to shift

(behavioural flexibility in their choice). Dogs are expected to perform above chance level in this task in general, but in this specific design, after having been presented with three consecutive pointing signals in the same direction, a drop in performance is expected at the first pointing in the opposite direction. This is followed by a recovery in performance when these latter signals are repeated. The test consisted of a total of 8 trials: two pre-training trials (see video protocol in appendix) and six test trials. The six test trials occurred in a fixed order (AAABBB), while the direction of the pointing (left or right) was balanced. The dog was held by the collar by the owner 2.5 m from the experimenter. In the pre-training trials, the experimenter put a piece of treat in one container, placed it in front of herself and the dog was allowed to take the treat from the container. During the test trials, the bowls (diameter 18 cm) were 1.5 m from each other, with the experimenter standing 50 cm behind them facing the dog. Before each trial, the experimenter called the dog by its name, established eye contact, then performed the pointing gesture and once she reached a static position (Fig. 2a) the dog was released. If the dog approached the indicated container first, it was allowed to eat the reward, if the dog approached the non-baited container first, the experimenter removed the baited container, the owner called back the dog and the next trial followed. A trial ended when the dog approached one of the bowls within 10 cm.

Measured variables

We coded the total number of correct choices (out of six) as the number of trials in which the dog went to the container signalled by the experimenter. A choice was considered correct if the dog approached the baited bowl (within 10 cm) first.

For analysing the dogs' performance (correct/incorrect choice) after switching from pointing to one side to the other (during Trial 4 & 5), we calculated a Generalised Linear Model with Binomial distribution and logit link.

Results & Discussion

From the three countries, dogs' mean performance in the first trial was 75.7 ± 42.9 %. From a previous dataset (*N*=117, (Brúder 2010) dogs performance was 78.6 ± 41.2 % in this task. We found that neither proximal, nor phenotype-related factors influenced dogs' performance regarding the number of total correct choices (Table 2).

Performance after pointing direction transition

In the fourth trial, when the experimenter first pointed in the other direction, similarly to the original findings (59.8 %), dogs' performance dropped (48.1 %). While dogs in our study performed at chance level, dogs in Brúder (2010) performed above chance. Calculating the effect size revealed that this difference between the two populations/studies was small (η^2 =0.014). Dogs in our study performed above chance level again in Trials 5 and 6 (Fig 3).

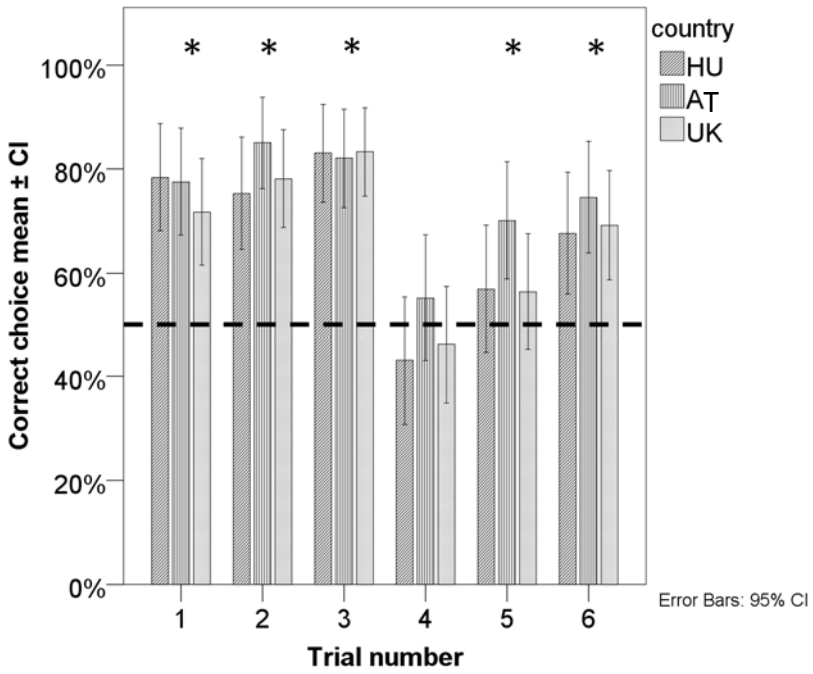

Figure 3 Dogs' performance in the pointing test, * = significance of performance above chance

Dogs' performance in this pointing test was robust; they performed at similar levels on all testing sites, among all three breed groups, regardless of sexual status (Table 2 & Table 3). The

drop in dogs' performance after the transition was also prevalent in every group, which shows that they reacted similarly in such a simple communicative situation.

Test 2. Problem solving before and after demonstration: The tube task

Method

In the tube task, the dogs were provided with a two-action task in an interspecific social learning context. The dogs could obtain a piece of food from a device (Fig. 2b) with two actions: via manipulating the plastic tube or via one of the two ropes attached to the left and right ends of the tube (see video protocol in appendix). Our protocol was based on Pongrácz et al.'s (2012) study, in which the level of success in the control group did not differ from the groups witnessing human demonstration, but when presented with a human demonstrating a rope manipulation, dogs tended to favour the demonstrated action, although they did not routinely follow the demonstrated side. The owner and the dog were 2.5m from the equipment. The height of the tube was adjusted to the height of the dog (the height of the tube could be adjusted between 40 and 100 cm- based on the dogs' height at the withers; 21-30 cm→ 40 cm, 31-40 cm→ 50 cm etc., see Fig. 2b).

In two pre-training trials the dog could witness the experimenter throw a piece of food into the slanted tube so that it fell out immediately at the other end and the dog could collect it. After this, we tested the dog in a control condition in a single trial, where the dog could attempt to extract the treat on its own without experimenter demonstration. After this, in three trials the experimenter demonstrated how the food could be extracted via pulling down the rope (always on the same side for a given subject). After the demonstration, the experimenter put the treat back into the tube, walked to the owner and the next trial began. Except for demonstrations, every manipulation of the tube happened behind an opaque screen, so that the dog could not see how the apparatus was loaded. The trial ended if the dog extracted the treat or after 60

seconds. During the test, the owner was allowed to encourage the dog but not to give instructions or commands to the dog and had to remain in the same position from the start.

Measured variables

We coded the number of successful trials (out of four) in which the dog released the treat from the tube within 60 seconds and the number of trials with successful rope manipulations, when the dog solved the task by manipulating the rope. We also coded on which side of the tube the successful manipulation occurred.

For the number of trials with successful rope manipulations a Generalised Linear Model with negative binomial distribution with log link was built.

Results & Discussion

Dogs' performance in this test was not influenced by either proximal, or phenotype-related factors (Table 4 & Table 5). In our combined sample, success rate was 69.4 ± 46.1 % in the control trial, which is comparable to the success level (72.2 ± 44.8 %) of Pongrácz et al. (2012). In the first trial the proportion of successful rope manipulations was 20.8%, while in Pongrácz et al. (2012) 16.7 % of the dogs succeeded in the first control trial via manipulating the rope. While in Pongrácz et al. (2012), dogs followed the pull demonstration in 43.3 % of cases, our dogs did so in only 22.9 % of the trials.

In our study, dogs always (regardless of testing site, sex, breed, demonstrated side) preferred the left side (one sample binomial, $p<0.05$), while no such preference occurred in Pongrácz et al.'s experiment. One possible explanation for this difference is a change in the procedure: The experimenter took 3 steps backward and remained behind the equipment in Pongrácz et al. (2012), while in our case, the experimenter went around on the left side to go back to the dog's owner. It is probable that the experimenter's movement to the left biased the dogs' attention to the corresponding side.

We reproduced the success rate of the original study, and dogs in our sample were similarly likely to operate the apparatus via a rope during their first encounter without a demonstration even though we used food as a reward, but if they had the chance to first interact with the apparatus without a demonstration, dogs did not follow the method shown in the repeated human demonstrations (manipulating the rope), whereas in the original protocol of Pongrácz et al. (2012) they did.

Test 3: Means-end test

Method

The protocol was based on Range et al. (2011), who measured dogs' performance in a support problem (physical cognition). The apparatus, consisting of two sliding boards, was slightly modified from the original study to test the dogs without the experimenter sitting in front of the dog during the trials (Fig. 2c). The two sliding wooden boards were connected with a 110 cm long string, so that if the dog pulled out one board, the other one was mechanically pulled back into the metal cage (100 x 65 x 65 cm). For the present study, we only used the 'same distance condition' of the original study where the two treats were placed at the same distance from the end of the boards, one on a board and one next to the other board (see video protocol in appendix). Thus, only pulling out the board with the treat on top would be rewarded. In this condition, dogs, as a group, are expected to perform above chance level.

During a pre-training phase, dogs were trained with shaping and positive reinforcement to pull out a baited board. During the pre-training, only a single board was available (the other was pushed back into the cage) but presentation of the boards alternated every time the dog got the treat so that the dogs received an even number of treats from both boards. Pre-training was completed if the dog was able to readily pull out the board without help three times in a row, if the dog lost interest in the task or if it did not learn the task within a twenty-minute session. Dogs that did not reach the learning criterion were excluded from the analysis.

The test consisted of six trials. During the test trials, the owner was seated on a chair 3 meters away from the apparatus. The dog was prevented from seeing the baiting of the apparatus via an opaque screen (at least 100 x 150 cm). After baiting, the experimenter removed the screen, walked back next to the owner and the trial began. The trial ended if (1) the dog pulled out one of the two boards or (2) after 60 seconds. During test trials, the owner was allowed to encourage the dog with his/her voice and gestures, but had to remain seated. When the trial ended, the owner called back the dog and the next trial began. We included only those subjects that made a valid choice in every trials (total number of excluded dogs=90; reasons for exclusion: dog did not reach learning criteria, dog managed to reach the food without pulling out the slide, dog did not make a choice, equipment malfunction)

Measured variables:

We coded the duration of the pre-training (seconds) as the time required to reach the pre-training criterion (pull out the slide without hesitation 3 times in a row). Duration of pre-training was used as a random factor in the statistical models. The number of correct choices was the number of trials in which the dog pulled out the baited slide within 60 seconds. The maximum value was 6.

Results & Discussion

In line with the original study, dogs included in this study (*N*=128) performed above chance level (binomial $P<0.001$ two tailed; 578 out of 768 trials). Dogs' chose the correct slide in 75 ± 18 % of the trials (see detailed information about performance level in Table 6). The influence of the testing site seems inconclusive. Testing site affected the performance of the dogs based on Bayesian analysis but it did not have a significant effect based on the GLM (Table 7). Based on the results of the Bayesian analysis, dogs from AT showed lower performance, but still

performed above chance level (Fig. 4). Interestingly, this lower performance was also closer to the performance level reported in the original study from AT.

We could not reach the planned sample size because many dogs failed to learn how to operate the equipment within the short time frame available to them and we also had to exclude subjects due to equipment malfunction (e.g. the slides got stuck). The rate of exclusion did not differ between testing sites X^2 (2, N = 218) = 0.39, P=0.825, but it differed among breed groups, with a higher dropout rate among the mixed breed dogs X^2 (2, N = 218) = 6.10, P=0.047.

We reproduced the original findings (Range et al. 2011) indicating that the results were robust, with dogs performing significantly above chance level at all testing sites and in every breed/population. The difference in performance levels between testing sites may be a consequence of random effects on the smaller sample size in comparison to the other tests, which highlights the importance of testing at least 15 dogs/group. A larger sample size would be required to test whether there is a real difference between the dogs' performance in this task between testing sites.

Test 4. Incidental memory

Method

This test is an adapted reproduction of Experiment 2 from Fujita et al. (2012). The 'incidental memory' test measures how accurately dogs can recall information in an unexpected memory test. During the presentation phase, the dog (on a leash) is allowed to investigate four bowls (Fig. 2e): an empty one, one containing a pebble and two containing a single piece of food each. The dog is allowed from consuming one and inhibited (via the leash) to consume the other treat, therefore at the end of the presentation phase only a single container had still food in it. After a 10-minute delay, the dog is allowed to choose which bowl to visit (see video protocol in

appendix). Based on Fujita et al. (2012), dogs are expected to remember the location of the remaining treat after the break and go for the container where they left the food.

The bowls were 26 cm in diameter and 10-12 cm in height. We put the bowls 2 m away from the starting point (in Fujita et al. this distance was 1.5 m), while keeping the angle (30 degrees) the same between neighbouring bowls. We decided to increase the distance due to the larger body size of the dogs in our sample. The position of the objects and that from which a treat could be eaten was randomized and told to the owner in advance, so that she could be prepared to prevent the dog from eating the second treat via holding onto the leash. After the presentation phase, the dog and the owner left the room and the experimenter changed the set of containers to a clean one (otherwise identical, but never containing any food). During the 10-minute delay, the owner, the dog, and the experimenter participated in the obedience test. After the delay, they returned to the room and the owner released the dog with a general release command without pointing in any direction. The trial ended after the dog made its second choice (visited the second bowl).

Behavioural variables:

We coded the dogs' first and second choice based on the bowls' content at the end of the presentation phase (where it left the treat, where it had previously consumed the treat, the bowl containing the pebble or the empty container).

Results & Discussion

Dogs chose the container in which they had left a treat significantly above chance level (25%). In our sample from three testing sites, on average 58.5% of the dogs went to the location where they left the food previously (compared to 51.3% in Fujita et al., 2012, for more details see Table 8).

Choice of the container (22.6%) from which the dog had previously eaten (but which was empty at the end of the presentation phase) fell between the German (42.8 %) and Japanese (5.6 %)

results of Fujita et al. (2012) for our dog population. The dogs which made an error during their first choice were more likely to go the container where they have previously eaten (Chi-square test, $P < 0.001$). Of those dogs that did not find the correct location on the first attempt (N=90), 55.6% went there on their second attempt (64.7 % in Fujita et al., 2012).We found that neither proximal, nor phenotype-related factors influenced dogs' performance regarding the measured variables in the test (Table 9).

Test 5. Obedience test

Method

By means of a short behavioural test battery we measured the subject's obedience level (the owner's ability to control the dog with simple commands) outdoors, in an area with moderate disturbance (people occasionally walking by, but no traffic nearby (Fig. 2d & video protocol in appendix). Our aim was to gather information about their training performance and relationship in a relatively objective manner. This part did not assess dog cognition per se as dogs' performance in such a situation most likely depends on their training experience and their handler's skilfulness. We used the following basic obedience tasks: call back, down (3 conditions: only verbal command, only hand signal, both), and stay. Between the tasks the owner was allowed to praise/pet the dog and give treats. The owners were not allowed to hold treats, or touch the dogs during the tasks. The commands were given in a fixed order for all dogs. The dog was on a long leash (5 m) throughout the test, but the owner was free to decide whether (s)he held onto it.

Measured variable

The scoring system was based on Fukuzawa et al. (2005). Each task was evaluated with the same 5-point scale (For a detailed description of the scoring see the appendix). We added additional scores where tasks could be divided into subtasks (e.g. call back and make the dog

sit down) to code the transition as well. The final score was the sum value of the task and transition scores of the five commands.

We scored the dog's performance and summarized the scores received for the different commands (total score, maximum value = 32 points). A Generalised Linear Model with multinomial distribution & cumulative logit link was used.

Results & Discussion

In the GLM, a testing site x breed interaction was revealed (Wald χ^2 (4, N=195) = 10.44, P = 0.034, Table 10, Table 11 and Fig. 5). In contrast, the Bayesian analysis did not support the presence of a testing site x breed interaction and favoured the model including testing site and breed only as separate factors. Border collies achieved higher scores than Labradors and other breeds, and dogs from AT received higher scores than dogs from the UK. Sexual status had no influence on the received obedience score. Obedience score did not influence any other analysed variables including the duration of necessary pre-training for the means-end task.

Although our goal was to recruit dogs without special training, Border collies and their owners may represent a special population, who, regardless of testing site (country of origin), provide some basic training to their dogs, as these dogs are often selected for a range of popular sporting activities like agility. In addition, we found differences between two testing sites, Lincoln and Vienna. One possible explanation is that dogs from a capital need some basic training to live near to traffic and crowded spaces, while this is not necessary in a small city like Lincoln, where owners have more opportunity to exercise their dogs in open fields away from others.

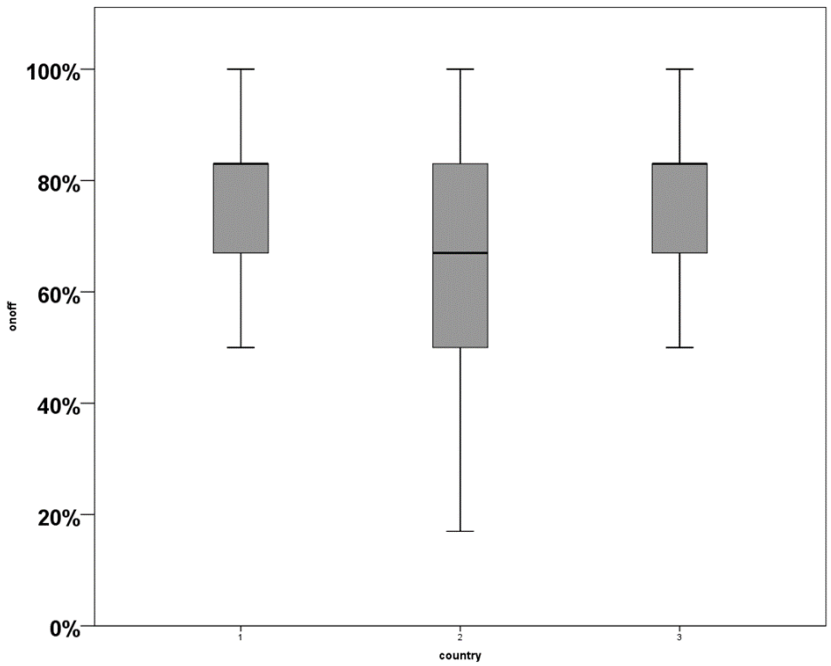

Figure 4 Boxplot of the dogs' performance in the means-end test, * = significance of performance above chance

General discussion

To our knowledge this is the first attempt to specifically measure reproducibility of a range of measures of cognitive-behavioural performance by dogs. We have successfully replicated the main findings of a broad range of cognitive tests (indicating inter-experimenter reliability) across three testing sites (indicating intra-rater and inter-site reliability), using three groups of dogs (indicating inter-subject reliability); indicating that these phenomena are robust and the results are generalizable between geographic regions. Where we did find differences in level of performance among testing sites, these were in the means-end test (where the sample size was smaller than desired) and in the obedience test (which depended on the owner for execution). Our findings do not question the view that certain breeds or even lines (e.g. working vs. show line, Fadel et al. 2016) differ as a population in their behaviour or problem solving performance in some tasks, but indicate these effects may be small. The current population compared only Border collies and Labradors, but the samples were relatively small compared to that of Fadel et al. (2016), and the tasks were not selected with the aim of detecting breed differences.

However, it is worth noting that differences between breeds can be due to genetic, functional, geographic and/or cultural factors (Miklósi 2014) and further work is required to tease out the relative importance of these factors in any discussion of the matter.

Although for most studies we replicated the main findings (whether dogs are able to perform on a similar level as a group in a given condition), there were some minor deviations from the original results, and many of these effects may be due to protocol differences. In the tube task, while we found the level of success and the preferred method (push vs. pull) in the control condition was comparable, dogs in our sample did not copy the demonstrated method, since they were not more likely to perform a pull action following the demonstration. In this case, we deviated significantly from the original protocol (Pongrácz et al. 2012) to make this test suitable for the present project (switching from ball reward to food reward, testing with a within subject design instead of the original between subject design). We also found what appeared to be a local enhancement effect from the experimenter's position during the task. This highlights how small changes in the protocol can have significant effects on the results. It is therefore essential that protocols are fully illustrated so that they can be faithfully reproduced and to this end the use of video demonstration as supplementary material to the methods is invaluable (Kampis et al. 2010; Kaminski et al. 2011; Huber et al. 2012). Videos of all the protocols used here are available in the supplementary information.

The method of pre-training in the means-end task could also have affected the results. Müller et al. (2014) found that with a modified training protocol, Border collies performed at chance level in this test condition. As our aim was to reproduce the dogs' performance from Range et al. (2011) (which we achieved), we did not test whether dogs understood the task from the beginning or learnt the mean-end relations during the pre-training. Nevertheless, as in the original study, we found no effect of the number of pre-training trials on the success rate.

In the memory task, our dogs' performance was between that of the German and Japanese dogs' reported by Fujita et al. (2012). Compared to dogs from Japan, European dogs were more likely to visit the container where they had previously found food (JP 6% below chance level, DE 43 % at chance level, our sample 23 % at chance level). Whether Japanese dogs differ only in this or also in other cognitive aspects from European dogs requires further investigation across a wider range of tasks, and emphasises the need for caution when generalising. This is reinforced by the results of the obedience test, where testing site and breed influenced performance. Training level has been reported to influence performance of dogs both in a physical problem solving task (Marshall-Pescini et al. 2008) and in a food choice task investigating social influence of the owner (Prato-Previde et al. 2008). In our study, we did not have enough dogs with advanced level training to test such effects, as our aim was to focus on effects relating to ordinary family dogs. A short obedience task as used here has the potential to provide objective data about the owner's ability to control the dog in a novel environment. This is more informative in evaluating dogs that have not completed a formal dog training course and makes comparison across dogs with different training background feasible, although it is important to keep in mind that to some extent, this test in its current form also relies on the owner's abilities. This information about the subjects is especially relevant in more sophisticated testing setups, because training level has been shown to influence behaviours which are usually measured in cognitive tasks as dependent variables, such as latency and duration of interaction with the apparatus and performance in a manipulative task (Marshall-Pescini et al. 2016). Thus it may be advisable to test whether the level of training of dogs succeeding in specific training tasks is comparable to the typical family dog population before making any generalisations to the latter (Huber et al. 2013).

In future studies (via close collaboration or by utilizing extensive video protocols, Kampis et al., 2010) behaviour of a large number of dogs from different countries and multiple testing sites could be compared to establish the robustness of other widely used testing protocols. Moreover, another aspect that should be studied to understand reproducibility of cognitive test in dogs is the effect of the experimenter and/or handler since we do not know to what extent the range of unintentional human cues could influence dogs' performance in complex situations.

References

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, Fiedler K, Fiedler S, Funder DC, Kliegl R, Nosek BA, Perugini M, Roberts BW, Schmitt M, Van Aken MAG, Weber H, Wicherts JM (2013) Recommendations for increasing replicability in psychology. Eur J Pers 27:108–119. doi: 10.1002/per.1919

Baldini S, Restani L, Baroncelli L, Coltelli M, Franco R, Cenni MC, Maffei L, Berardi N (2013) Enriched Early Life Experiences Reduce Adult Anxiety-Like Behavior in Rats: A Role for Insulin-Like Growth Factor 1. J Neurosci 33:11715–11723. doi: 10.1523/JNEUROSCI.3541-12.2013

Brúder I (2010) Development of a test battery for evaluating dog personality and investigating the genetic background of personality traits. Eötvös Loránd University

Casadevall A, Fang FC (2010) Reproducible Science. Infect Immun 78:4972–4975. doi: 10.1128/IAI.00908-10

Crabbe JC, Wahlsten D, Dudek BC (1999) Genetics of Mouse Behavior: Interactions with Laboratory Environment. Science 284:1670–1672. doi: 10.1126/science.284.5420.1670

Duranton C, Rödel HG, Bedossa T, Belkhir S (2015) Inverse sex effects on performance of domestic dogs (*Canis familiaris*) in a repeated problem-solving task. J Comp Psychol

129:84–87. doi: 10.1037/a0037825

Fadel FR, Driscoll P, Pilot M, Wright H, Zulch H, Mills D (2016) Differences in Trait Impulsivity Indicate Diversification of Dog Breeds into Working and Show Lines. Sci Rep 6:22162. doi: 10.1038/srep22162

Fujita K, Morisaki A, Takaoka A (2012) Incidental memory in dogs (*Canis familiaris*): adaptive behavioral solution at an unexpected memory test. Anim Cogn 15:1055–1063. doi: 10.1007/s10071-012-0529-3

Fukuzawa M, Mills DS, Cooper JJ (2005) The effect of human command phonetic characteristics on auditory cognition in dogs (*Canis familiaris*). J Comp Psychol 119:117–121.

Horn L, Marshall-Pescini S, Virányi Z, Range F (2013a) Cross-cultural differences in domestic dogs' interactions with humans – preliminary results from ainsworth's strange situation test. J Vet Behav Clin Appl Res 8:e39. doi: 10.1016/j.jveb.2013.04.043

Horn L, Range F, Huber L (2013b) Dogs' attention towards humans depends on their relationship, not only on social familiarity. Anim Cogn 16:435–443. doi: 10.1007/s10071-012-0584-9

Huber L, Racca A, Scaf B, Virányi Z, Range F (2013) Discrimination of familiar human faces in dogs (*Canis familiaris*). Learn Motiv 44:258–269. doi: 10.1016/j.lmot.2013.04.005

Huber L, Range F, Virányi Z (2012) Dogs imitate selectively, not necessarily rationally: Reply to Kaminski et al. (2011). Anim Behav 83:3–5. doi: 10.1016/j.anbehav.2012.03.020

Kafkafi N, Golani I, Jaljuli I, Morgan H, Sarig T, Würbel H, Yaacoby S, Benjamini Y (2017) Addressing reproducibility in single-laboratory phenotyping experiments. Nat Methods 14:462–464. doi: 10.1038/nmeth.4259

Kaminski J, Nitzschner M, Wobber V, Tennie C, Brauer J, Call J, Tomasello M (2011) Do

dogs distinguish rational from irrational acts? Anim Behav 81:195–203. doi: 10.1016/j.anbehav.2010.10.001

Kampis G, Miklosi A, Viranyi Z, Gulyas L (2010) Video Deep Tagging and Data Archiving in the Comparative Mind Database. In: Spink AJ, Grieco F, Krips OE, Loijens L, Noldus L, Zimmerman P (eds) 7Th International Conference on Methods and Techniques in Behavioral Research. Eindhoven, pp 185–188

Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, Bocian K, Brandt MJ, Brooks B, Brumbaugh CC, Cemalcilar Z, Chandler J, Cheong W, Davis WE, Devos T, Eisner M, Frankowska N, Furrow D, Galliani EM, Hasselman F, Hicks JA, Hovermale JF, Hunt SJ, Huntsinger JR, Ijzerman H, John MS, Joy-Gaba JA, Kappes HB, Krueger LE, Kurtz J, Levitan CA, Mallett RK, Morris WL, Nelson AJ, Nier JA, Packard G, Pilati R, Rutchick AM, Schmidt K, Skorinko JL, Smith R, Steiner TG, Storbeck J, Van Swol LM, Thompson D, Van 'T Veer AE, Vaughn LA, Vranka M, Wichman AL, Woodzicka JA, Nosek BA (2014) Investigating variation in replicability: A "many labs" replication project. Soc Psychol 45:142–152. doi: 10.1027/1864-9335/a000178

Marshall-Pescini S, Frazzi C, Valsecchi P (2016) The effect of training and breed group on problem-solving behaviours in dogs. Anim Cogn 19:571–579. doi: 10.1007/s10071-016-0960-y

Marshall-Pescini S, Valsecchi P, Petak I, Accorsi PA, Previde EP (2008) Does training make you smarter? The effects of training on dogs' performance (*Canis familiaris*) in a problem solving task. Behav Processes 78:449–454. doi: 10.1016/j.beproc.2008.02.022

Miklósi Á (2014) Dog behaviour, evolution, and cognition., 2nd edn. Oxford University Press.

Miklósi Á, Topál J (2013) What does it take to become "best friends"? Evolutionary changes in canine social competence. Trends Cogn Sci 17:287–294. doi:

10.1016/j.tics.2013.04.005

Moonesinghe R, Khoury MJ, Janssens ACJW (2007) Most Published Research Findings Are False—But a Little Replication Goes a Long Way. PLoS Med 4:218–221. doi: 10.1371/journal.pmed.0040028

Müller CA, Mayer C, Dorrenberg S, Huber L, Range F (2011) Female but not male dogs respond to a size constancy violation. Biol Lett 7:689–691. doi: 10.1098/rsbl.2011.0287

Müller CA, Riemer S, Virányi Z, Huber L, Range F (2014) Dogs learn to solve the support problem based on perceptual cues. Anim Cogn 17:1071–1080. doi: 10.1007/s10071-014-0739-y

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349:aac4716. doi: 10.1126/science.aac4716

Pongrácz P, Bánhegyi P, Miklósi Á (2012) When rank counts — dominant dogs learn better from a human demonstrator in a two-action test. Behaviour 149:111–132. doi: 10.1163/156853912X629148

Pongrácz P, Gácsi M, Hegedüs D, Péter A, Miklósi A (2013) Test sensitivity is important for detecting variability in pointing comprehension in canines. Anim Cogn 16:721–35. doi: 10.1007/s10071-013-0607-1

Prato-Previde E, Marshall-Pescini S, Valsecchi P (2008) Is your choice my choice? the owners' effect on pet dogs' (*Canis lupus familiaris*) performance in a food choice task. Anim Cogn 11:167–174. doi: 10.1007/s10071-007-0102-7

Range F, Hentrup M, Virányi Z (2011) Dogs are able to solve a means-end task. Anim Cogn 14:575–583. doi: 10.1007/s10071-011-0394-5

Richter SH, Garner JP, Würbel H (2009) Environmental standardization : cure or cause of poor reproducibility in animal experiments ? Nat Methods 6:257–261. doi: 10.1038/NmETH.1312

Sidman M (1960) Tactics of scientifc research: Evaluating experimental data in psychology, vol 5. Basic Books, New York

Topál J, Miklósi Á, Csányi V (1997) Dog-human relationship affects problem solving behavior in the dog. Anthrozoos 10:214–224. doi: 10.2752/089279397787000987

Tuyttens FAM, de Graaf S, Heerkens JLT, Jacobs L, Nalon E, Ott S, Stadig L, Van Laer E, Ampe B (2014) Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? Anim Behav 90:273–280. doi: 10.1016/j.anbehav.2014.02.007

van der Staay JF, Arndt SS, Nordquist RE (2010) The standardization-generalization dilemma: a way out. Genes, Brain Behav 9:849–855. doi: 10.1111/j.1601-183X.2010.00628.x

Wahlsten D, Bachmanov A, Finn DA, Crabbe JC (2006) Stability of inbred mouse strain differences in behavior and brain size between laboratories and across decades. Proc Natl Acad Sci U S A 103:16364–9. doi: 10.1073/pnas.0605342103

Yang H, Harrington CA, Vartanian K, Coldren CD, Hall R, Churchill GA, Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Schindler B, Chourbaji S, Brandwein C, Gass P, Stipdonk N Van, Wolfer DP, Wu H (2008) Randomization in laboratory procedure is key to obtaining reproducible microarray results. PLoS One 3:e3724. doi: 10.1371/journal.pone.0003724

Figure captions:

Fig. 1 The three testing rooms from the position of recording cameras and room dimensions: Top left Budapest (3.6 m x 4.6 m), Top right Vienna (6 m x 7.2 m), Bottom line Lincoln (5.2 m x 5.9 m)

Fig. 2 Demonstration of the equipment used during the project. a, Setup of the pointing test b, Setup of the problem solving test c, The apparatus used in the means-end test d, Setup of the obedience task e, Setup of the incidental memory test

Fig. 3 Dogs' performance in the pointing test, * = significance of performance above chance

Fig. 4 Boxplot of the dogs' performance in the means-end test, * = significance of performance above chance

Fig. 5 Boxplot of the obedience scores by breed groups and countries. BC- Border collie, LR-Labrador retriever, VB-various breeds, HU-Hungary, AT-Austria, UK-United Kingdom

12 **Table 1 Overview of the subjects included in the statistical analysis in each subtest.**

Testing site mean age (in years) ± SD	Sample	SUM included in analysis	Response flexibility (pointing) test	Problem solving (tube) test	Means-end test	Memory test	Obedience test
Budapest (HU) 3.95 ± 2.31 years	Border collie	22	22	22	17	22	17
	Labrador retrievers	21	21	21	14	21	19
	Various breeds	24	22	24	8	24	17
Vienna (AT) 4.88 ± 2.93 years	Border collie	25	23	25	18	25	19
	Labrador retrievers	19	19	19	10	19	17
	Various breeds	27	25	22	12	27	26
Lincoln (UK) 4.95 ± 2.84 years	Border collie	28	27	28	16	28	28
	Labrador retrievers	21	21	21	13	20	21
	Various breeds	31	30	29	20	31	31
Total		218	210	211	128	217	195

13

Table 2 The effect of proximal and phenotypic-related factors on performance in the pointing test. Reported *P* values in case of non-significant factors are the last values before removal from the given model. BF_{10}: The Bayes Factor of H1 against H0. $BF_{Inclusion}$: the change from prior to posterior inclusion odds.

No. of correct choices					
Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$
Breed x Testing site	4	1.802	0.772	0.007	0.019
Breed	2	0.402	0.818	0.084	0.060
Country	2	1.719	0.423	0.469	0.319
Sex	3	1.013	0.798	0.070	0.070
Performance in Trial 4					
Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$
Breed x Testing site	4	4.057	0.398	0.010	0.004
Breed	2	1.330	0.514	0.085	0.058
Country	2	2.126	0.345	0.126	0.085
Sex	3	1.149	0.765	0.037	0.038

18 **Table 3 Dogs' performance in the pointing test**

Success rate in first trial, percentage (mean ± SD)	HU	AT	UK	Original results (Brúder 2010)
Border collie	81.8 ± 38.5 %	78.3 ± 41.2 %	70.4 ± 45.6%	78.6 ± 41.2%
Labrador retrievers	76.2 ± 42.6 %	78.9 ± 40.7 %	76.2 ± 42.6%	(only German
Various breeds	77.3 ± 41.9 %	76.0 ± 42.7 %	70.0 ± 45.3%	shepherd dogs, from HU, *N*=117)
CI	68.2%-88.7%	67.4%-87.9%	61.6%-82.0%	

19

20 **Table 4 Dogs' performance in the problem solving test**

Success rate in first trial, percentage (mean ± SD)	HU	AT	UK	Original results (Pongrácz et al. 2012)
Border collie	59.1 ± 50.3%	84.0 ± 37.4%	78.6 ± 41.8%	72.22 ± 44.8% (from HU, *N*=18, mixed breeds)
Labrador retrievers	81.0 ± 40.2%	63.2 ± 49.6%	66.7 ± 48.3%	
Various breeds	50.0 ± 51.1%	63.6 ± 49.2%	65.5 ± 48.4%	
CI	50.8%-74.6%	60.0%-82.4%	47.1%-83.9%	
No. of successful rope manipulation (mean ± SD)	HU	AT	UK	
Border collie	1.1±1.3	1.3±1.6	0.8±1.1	
Labrador retrievers	1.1±1.5	0.7±1.0	0.9±1.3	
Various breeds	0.4±0.8	0.7±1.0	0.9±1.3	
CI	0.6-1.2	0.6-1.3	0.6-1.2	

21

Table 5 The effect of proximal and phenotypic-related factors on performance in the problem solving test. Reported *P* values in case of non-significant factors are the last values before removal from the given model. BF_{10}: The Bayes Factor of H1 against H0. $BF_{Inclusion}$: the change from prior to posterior inclusion odds.

No. of successful trials					
Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$
Breed x Testing site	4	4.400	0.355	0.063	0.102
Breed	2	4.236	0.120	0.637	0.469
Testing site	2	3.407	0.182	0.347	0.242
Sex	3	0.672	0.880	0.471	0.438
No. of pull manipulations					
Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$
Breed x Testing site	4	4.644	0.326	0.002	0.007
Breed	2	2.878	0.237	0.254	0.616
Testing site	2	0.134	0.935	0.053	0.036
Sex	3	5.740	0.125	0.101	0.095

28 **Table 6 Dogs' performance in the means-end task at three different testing sites (HU, AT, UK)**

Success rate, percentage (mean ± SD)	HU	AT	UK	Original results (Range et al. 2011)
Border collie	77.5 ± 15.5 %	59.3 ± 18.3 %	76.7 ± 12.3 %	66.1±12.7 % from AT (12 trials, *N*=31, mixed breeds)
Labrador retrievers	83.3 ± 11.3 %	78.3 ± 11.3 %	84.6 ± 12.7 %	
Various breeds	85.4 ± 10.9 %	62.5 ± 23.7 %	75.8 ± 19.1 %	
CI	76.0%-86.4%	59.8%-70.2%	73.8%-83.2%	

29

30

Table 7 The effect of proximal and phenotypic-related factors on number of correct choices in the means-end task. Reported *P* values in case of non-significant factors are the last values before removal from the given model. BF_{10}: The Bayes Factor of H1 against H0. $BF_{Inclusion}$: the change from prior to posterior inclusion odds. Significant factors are indicated with bold.

No. of correct choices Factors	df	Wald χ^2	*P* value	BF_{10}	$BF_{inclusion}$	Effect size (η^2)
Breed x Testing site	4	0.931	0.920	15.805	0.500	
Breed	2	1.405	0.495	0.981	0.893	
Testing site	2	4.956	0.084	**59.447**	**40.465**	**0.127 (medium)**
Sex	3	1.502	0.682	0.863	0.628	
Duration of pre-training	1	1.769	0.184	0.143	0.156	

Table 8 Dogs' performance in the memory test at three different testing sites (HU, AT, UK)

Correct choice in first trial, percentage (mean ± SD)	HU	AT	UK	Original results (Fujita et al. 2012)
Border collie	50.0 ± 51.2 %	68.0 ± 47.6%	71.4 ± 46.0%	
Labrador retrievers	52.4 ± 51.2 %	68.4 ± 47.8%	50.0 ± 51.3%	
Various breeds	58.3 ± 48.3 %	44.4 ± 50.0%	61.3 ± 50.1%	51.3 ± 50.0% (from DE and JP, *N*=39)
CI	53%-75%	44%-70%	44%-66%	

Table 9 The effect of proximal and phenotypic-related factors on dogs' performance in the memory test. Reported *P* values in case of non-significant factors are the last values before removal from the given model. BF_{10}: The Bayes Factor of H1 against H0. $BF_{Inclusion}$: the change from prior to posterior inclusion odds.

First choice Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$
Breed x Testing site	4	1.447	0.836	0.001	0.003
Breed	2	0.31	0.822	0.073	0.049
Testing site	2	1.485	0.476	0.189	0.120
Sex	3	0.470	0.925	0.214	0.203

43 **Table 10 Dogs' performance in the obedience test at three different testing sites (HU, AT, UK)**

Obedience score (percentage, mean ± SD)	HU	AT	UK
Border collie	84.7 ± 14.2 %	83.2 ± 15.3 %	85.7 ± 16.4 %
Labrador retrievers	85.4 ± 10.8 %	80.0 ± 9.9 %	68.0 ± 17.5 %
Various breeds	76.1 ± 19.5 %	77.3 ± 21.1 %	65.5 ± 28.0 %
CI	72.7 %-82.6 %	71.6 %-81.14 %	61.4 %-73.2 %

44

Table 11 The effect of proximal and phenotypic-related factors on dogs' performance in the training level test. Reported *P* values in case of non-significant factors are the last values before removal from the given model. BF_{10}: The Bayes Factor of H1 against H0. $BF_{Inclusion}$: the change from prior to posterior inclusion odds. Significant factors are indicated with bold.

Total score Factors	df	Wald χ^2	*P*	BF_{10}	$BF_{inclusion}$	Effect size (η^2)
Breed x Testing site	4	10.440	**0.034**	**34.783**	**1.233**	**0.033 (small)**
Breed+Testing site		-	-	**75.332**	-	-
Breed	2	14.130	**0.001**	**33.892**	**39.562**	**0.057 (small)**
Testing site	2	5.989	**0.050**	**1.562**	**2.209**	**0.040 (small)**
Sex	3	0.548	0.908	0.088	0.157	

Appendix

A priori sensitivity analysis (minimal detectable effect) was carried out with G*Power 3.1.9.2 for a planned sample size of 180 and a power of 0.8 for breed groups and testing sites. The test yielded a sensitivity of 0.052, which meant that our study would be able to detect a medium effect size (η^2=0.06) with a power of 0.8 between testing sites and breeds.

Table 12 Actual sensitivity of the individual subtests regarding testing sites and breed groups.

Sample	Response flexibility (pointing) test	Problem solving (tube) test	Means-end test	Memory test	Obedience test
Total *N*	210	211	128	217	195
Actual sensitivity for Breed/Testing site (η^2; if 1-β= 0.8; α=0.05)	0.046	0.046	0.073	0.042	0.046
Actual sensitivity for Breed x Testing site (η^2; if 1-β= 0.8; α=0.05)	0.055	0.055	0.088	0.055	0.059

58 **Table 13 Interobserver agreement. Cohen's kappa coefficients of the analysed variables.**

Test/Variable	Cohen's κ
Pointing/ no. of correct choices	1
Problem solving test/ no. of successful trials	0.81
Problem solving test/ utilized action	0.82
Means-end test/no. of correct choices	0.76
Memory test/first choice	0.95
Obedience test/total score	0.77

59

60

Detailed scoring of the obedience test:

Three types of down commands in a row (only verbal command, only gesture, both). Calling the dog by name was allowed in every case. The next task was a call back from ca. 5 meters, making the dog sit down within a 1 m radius (performance was scored for each of these two subtasks as well as the transition). The last task was a stay command (for 5 seconds at a distance of 5 meters) after taking a position (sit, lay or stand) of the owner's choice (each of these two subtasks was scored as well as the transition).

Task score

4-complete and instant response to the command

3-complete but delayed response to the command, delay to completion not exceeding 5 s.

2-incomplete response to the command; e.g. the dog may not settle in the sit command

1-a nonspecific response to the command; e.g. the dog orients toward the owner

0-no response within 5 s of the command

Transition score

0- no link -link made only after second command

1- link made after pause / delay

2- link made with no delay

Video examples

https://www.youtube.com/watch?v=aaK6NZYh5QQ

https://www.youtube.com/watch?v=Zlhcb36xr8I

https://www.youtube.com/watch?v=rmVd9UDMHcI

https://www.youtube.com/watch?v=KXhTL1NtYcw

https://www.youtube.com/watch?v=T3wWeLth7S8

Video protocols

Means-end test training phase- http://www.youtube.com/watch?v=C5X_QXyJV_E

Means-end test phase- http://www.youtube.com/watch?v=wzpUSW_fRq0

Pointing test: http://www.youtube.com/watch?v=3i35evlPtiw

Problem solving test: https://www.youtube.com/watch?v=2Kh9PCi5qcY

Memory test: http://www.youtube.com/watch?v=1N316fxBznQ