Animal Behaviour xxx (xxxx) xxx

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Performance on tests of cognitive ability is not repeatable across years in a songbird

Jill A. Soha [a, *], Susan Peters [a], Rindy C. Anderson [b], William A. Searcy [c], Stephen Nowicki [a]

[a] Department of Biology, Duke University, Durham, NC, U.S.A.
[b] Department of Biological Sciences, Florida Atlantic University, Boca Raton, FL, U.S.A.
[c] Department of Biology, University of Miami, Coral Gables, FL, U.S.A.

ARTICLE INFO

Article history:
Received 12 March 2019
Initial acceptance 18 April 2019
Final acceptance 5 August 2019
Available online xxx
MS. number: A19-00187R2

Keywords:
avian cognition
cognition–behaviour correlation
cognitive repeatability
inhibitory control
learning
song repertoire size
song sparrow

Studies of the cognitive abilities of animals aim to help us understand how they communicate, obtain resources, avoid danger and otherwise thrive in a given environment. But to what extent is cognitive ability a fixed trait in individuals? And can we answer this question by measuring performance on tests of cognitive ability? We tested the same 18 male song sparrows, *Melospiza melodia*, once yearly, across three consecutive years, with tests of four putative cognitive traits and a test of neophobia. We also tested 19 females twice, once in the first year and once in the third year. All birds were hand-reared and tested in the laboratory. Analyses of both data sets indicate repeatability of neophobia but not of performance on the cognitive tests. In addition, correlations among cognitive performance, neophobia and song quality that were observed in the first year were not observed in subsequent rounds of testing. These results suggest that cognitive ability is not a fixed trait in individuals, or that the tests used do not accurately measure cognitive ability, or both. Conclusions drawn from a single round of cognitive tests should therefore be interpreted with caution in this species and in any species in which repeatability has not been verified.

Tests of animal cognitive abilities are used to address a variety of questions in behavioural ecology. For example, cognitive tests conducted in the wild have addressed questions such as the correlation between cognitive performance and reproductive success (Cauchard, Boogert, Lefebvre, Dubois, & Doligez, 2013) and the relationship between cognition and social group size (Ashton, Ridley, Edwards, & Thornton, 2018). Laboratory tests have investigated the relationship between cognition and personality (Brust, Wuerz, & Krüger, 2013; Dugatkin & Alfieri, 2003; Guillette, Hahn, Hoeschele, Przyslupski, & Sturdy, 2015), the extent to which individually distinct mating displays or ornaments signal overall cognitive ability (Keagy, Savard, & Borgia, 2012; Mateos-Gonzales, 2011) and the factors that shape the evolution of cognition itself (MacLean et al., 2014; Pravosudov & Roth, 2013).

It is widely recognized that factors other than cognitive ability can affect individual performance on the types of tests used in such studies (Morand-Ferron, Cole, & Quinn, 2016; Rowe & Healy, 2014). Despite this problem, it has not been standard practice to measure consistency of performance on cognitive tests in animals. Cauchoix et al. (2018) used mostly unpublished data from 25 species, including humans, to assess two types of repeatability in cognitive performance: temporal repeatability (consistency in performance over time on the same task) and contextual repeatability (similarity in performance on different tasks designed to test the same cognitive trait). The authors found that repeatability varied considerably across species and was affected by the type of performance metric used (e.g. percentage correct, latency, trials to reach criterion). In the case of contextual repeatability, they also found a bias towards publication of significant repeatability values, although the number of publications reporting any repeatability values for animal cognitive performance is small (Cauchoix et al., 2018, included only six). Thus, it remains important to measure and report repeatability of cognitive test performance in animals.

Measuring repeatability is particularly important when the question of interest is how individual variation in cognitive ability correlates with variation in another trait. In songbirds, for example, correlations between general cognitive ability and aspects of male song could arise during early life due to the effects of stress on brain development, and these correlations would mean that song could function as an acoustic signal of cognitive ability across domains (Peters, Searcy, & Nowicki, 2014). In cases where the trait of interest

* Correspondence: J.A. Soha, Department of Biology, Duke University, Durham, NC, 27708, U.S.A.
E-mail address: jill.soha@duke.edu (J. A. Soha).

https://doi.org/10.1016/j.anbehav.2019.09.020
0003-3472/

is fixed throughout an animal's life, such as repertoire size in songbirds that learn songs only when young, performance on cognitive tests should likewise be consistent over time (i.e. repeatable) if two conditions hold: (1) cognitive ability is correlated with the trait of interest and (2) the cognitive tests used accurately assess cognitive ability.

Boogert, Giraldeau, and Lefebvre (2008) found a positive correlation between song complexity and performance on a novel foraging task in zebra finches, *Taeniopygia guttata*, suggesting that a link between song quality and other cognitive traits might help explain female selection for complex song. However, studies with male song sparrows, *Melospiza melodia*, have found that neither repertoire size nor the imitation accuracy of learned songs are consistently correlated with performance on cognitive tests (Anderson et al., 2017; Boogert, Anderson, Peters, Searcy, & Nowicki, 2011; Sewall, Soha, Peters, & Nowicki, 2013). Similarly, correlations have not been found between song quality and cognitive performance in swamp sparrows, *Melospiza georgiana* (DuBois, Nowicki, Peters, Rivera-Cáceres, & Searcy, 2018) or between song repertoire size and performance on two cognitive tests in a food-caching species, the New Zealand robin, *Petroica longipes* (MacKinlay & Shaw, 2019). The question remains, however, how accurately these tests measure cognitive ability. If other factors also influence performance on the tests, as Rowe and Healy (2014) have emphasized is likely, these tests might provide only rough approximations of cognitive ability. In this case, performance might not be repeatable from one round of testing to the next, and results from any one study should be considered in this context.

Here we test for temporal repeatability of performance on four cognitive tests in song sparrows. These tests were designed to measure individual abilities in associative learning (colour association and colour reversal), spatial learning and inhibitory control (detour reaching). We expected that all birds would improve to some extent from the first year to the second, but beyond that, would higher-performing individuals in the first year remain higher-performing in subsequent years? Previous studies indicate that intelligence in sparrows is modular rather than general, such that performance on a test in one cognitive domain is not necessarily correlated with performance in another domain (see Searcy & Nowicki, 2019). Given this modularity, we assess repeatability for each task individually.

We also test for correlations among cognitive performance measures and neophobia, and — in males — between measures of song quality and cognitive performance, to determine whether correlations observed in the first year of testing in these birds (Anderson et al., 2017) also held in subsequent years. Finally, we examine whether song quality is correlated with average performance on the cognitive tests across all 3 years of testing in males, to assess whether fluctuations in performance around individual means might obscure, in any single round of testing, a real correlation.

METHODS

Subjects and Song Analysis

The 18 male and 19 female song sparrows in this study were collected as nestlings 3–6 days after hatching in Crawford County, Pennsylvania, U.S.A., in May 2013. Details of their hand-raising and song tutoring are given in Anderson et al. (2017). Briefly, all birds were tutored as a group for 12 weeks beginning at 10 days after hatching. The tutor songs were 32 distinct song sparrow songs recorded in 2009 and 2010 at the same sites from which the nestlings were collected. The male subjects were then recorded for 5 weeks, beginning at about 11 months. Recording 200 songs usually suffices to capture the entire repertoire of a wild song sparrow (Searcy, McArthur, & Yasukawa, 1985), and here we recorded an average of 1578 songs per male (range 254–2907). Repertoire size was ascertained by visual analysis of spectrograms in Syrinx (Burt, Campbell, & Beecher, 2001). The most common variant of each song type was identified and compared to the tutor songs: the proportion of notes in these variants that were copied from tutor models was assessed visually, and copy accuracy was measured using spectrogram cross-correlations of the copied notes versus their tutor models in Signal for Windows v.4 (see Anderson et al., 2017 for details).

Cognitive Tests

All tests were done on adults, and all subjects were approximately the same age. Male subjects were tested three times, all at Duke University. The first tests began in June 2014, the second in July 2015, and the third in March 2016. Females were tested twice, once at Duke University beginning in March 2014 and again at Florida Atlantic University beginning in March 2016. Birds were maintained on a natural photoperiod throughout. Song sparrows are known to live as long as 11 years in the wild (USGS, 2017), so all of our tests were done before any effects of senescence are likely.

Tests were done every weekday afternoon, after subjects had been deprived of food for 5 h to ensure that they were motivated to work for food rewards (mealworms). Ad libitum food was restored immediately after testing was completed. On each day of testing, six trials were conducted per bird, with a minimum of 20 min between trials for any individual bird and a maximum of nine birds tested per day. All trials were 2 min or less in duration. Each bird was tested in its home cage ($46 \times 22 \times 26$ cm), while visually but not acoustically isolated from other birds. Trials were scored by an observer seated outside of the testing room and watching via streaming video. Video streams were also saved for review in any cases of uncertainty in the real-time scoring. Trials began when the observer closed the door of the testing room and immediately started a timer, and ended when the bird completed the task or after 2 min, whichever occurred first.

Birds were given first a test of neophobia and then a series of cognitive tests modified from Boogert et al. (2011) and Sewall et al. (2013). These are described briefly below (see Anderson et al., 2017, for additional details and images of the test equipment). Upon completing each task by meeting the task-specific criterion, birds progressed immediately to the next task; all birds progressed through the tasks in the sequence given below. The number of trials required to reach the criterion in each task was used as the performance score.

Neophobia

Neophobia was measured as the time (in minutes) taken by the bird to remove a mealworm from a foraging grid for the first time in each year of testing. If this took longer than 6 h total (across 3 days of testing), the foraging grid was left in the cage overnight along with the familiar seed cup; this was necessary for one bird only. Foraging grids were $13.5 \times 9 \times 2.5$ cm blocks of plastic containing six wells 1.3 cm in diameter and 0.8 cm deep. These grids were novel in the first year but familiar (from previous tests) in subsequent years.

Novel foraging

Birds next had to learn to remove lids from the foraging grid wells. Mealworms were placed into four of the six wells in the grid, one per well, and these baited wells were covered with 2.5 cm diameter plastic discs. A successful trial was one in which a bird obtained at least two mealworms within 2 min. In the first year,

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020

training was done in five stages with the lids covering progressively more of each well, and birds completed each stage upon succeeding in four of five consecutive trials. Birds did not regress to earlier shaping stages if they failed multiple trials in a row; rather, the baited foraging grids were left in the cage and the bird had at least 20 min to remove the mealworms before the next trial began. This proved sufficient for shaping. In subsequent years, we shortened the procedure to four stages for those birds that ate within 2 min in the neophobia test, and for all birds we changed the pass criterion to simply obtaining two mealworms once within 2 min (i.e. succeeding once) at each stage. Although this novel foraging task was included as a cognitive test in the initial analysis of song learning and cognitive abilities (Anderson et al., 2017), we do not include it in our repeatability analysis because, once experienced, the task is no longer 'novel' on retesting. Rather, we used this procedure as refresher training, enabling us to verify that birds remembered how to forage from the testing grid.

Task 1: colour association

In this task, birds had to learn to associate one of two colours with a food reward. Two foraging grids were used, for a total of 12 available wells. Four wells were baited with mealworms and covered with plastic lids of one colour. Four empty wells were covered with plastic lids of a second colour, and four were left uncovered. Positions of baited wells and empty covered wells were arbitrarily chosen in each trial. The first tests, in 2014, used yellow and blue lids. The second tests (males in 2015 and females in 2016) used green and red lids. The third tests (males in 2016) used black and white lids. Within each year, the rewarded colour and the unrewarded colour were balanced across subjects. The prepared foraging grids were placed together in the cage, and to complete this task, a bird had to do either of the following in six of seven consecutive trials: (1) remove at least two lids of the rewarded colour and none of the unrewarded colour or (2) remove all four lids of the rewarded colour before removing any lids of the unrewarded colour.

Task 2: colour reversal

This task was the same as the colour association task except that the rewarded colour and the unrewarded colour were switched. Each bird therefore had to extinguish the previously learned association and learn the new one. The criteria for completion of this task were the same as for the colour association task.

Task 3: spatial learning

The spatial learning task required birds to learn the location of a predictable food reward. Small (6.4 cm square) blocks, made from the same material as the foraging grid but each containing only a single well, were used. In six preliminary trials, one baited block was placed in each corner of the home cage. The wells in these blocks were covered with lids of the colour rewarded in the previous task. For the spatial task itself, all four wells were covered but only one corner was rewarded. The location of the rewarded corner was semirandomly chosen for each subject each year, avoiding any corner for which the bird had shown a positive (or negative) preference by visiting it first (or last) in at least half of the preliminary trials; this procedure occasionally resulted in the same corner being rewarded in successive years. To complete this task, a bird had to remove the lid on the baited block first in six of seven consecutive trials. In a single probe trial after completion, the baited block was placed in a different corner to verify, based on the first corner visited, that birds had learned the rewarded location and were not detecting the mealworm using sensory cues.

Task 4: detour reaching

In this task, birds had to learn to remove a mealworm from a horizontally oriented plastic cylinder (4 cm diameter, 5 cm length) through one of the ends instead of pecking the side of the cylinder in an attempt at more direct access. An opaque black cylinder was used first, in which the mealworm was visible only through the ends. After the bird removed the mealworm from this cylinder without pecking its side in four of five consecutive trials, a clear cylinder was used. A bird completed this task upon removing the mealworm from the clear cylinder without pecking its side in six of seven consecutive trials. The total number of trials run using the clear cylinder constituted the performance score.

Repeatability Analysis

We assessed repeatability of neophobia and the cognitive task performance scores in two ways. First, we used Spearman's rank correlation tests to compare data from consecutive years of testing. If performance is highly repeatable, this nonparametric test should suffice. Second, we used general linear mixed models (GLMMs) to analyse the cognitive task scores. Nakagawa and Schielzeth (2010) recommend this approach for assessing repeatability of non-Gaussian data. We used the R package 'rptR' by Stoffel, Nakagawa, and Schielzeth (2017). Within rptR, we used 'rptPoisson' (for count data) with log link on nontransformed data. This method uses parametric bootstrapping and Bayesian methods to estimate confidence intervals and randomization methods for significance testing. We ran a separate GLMM for each behavioural task. In all models, bird identity was included as a random effect, and fixed effects were year (Biro & Stamps, 2015) and date-within-year. Date-within-year referred to the first date of testing on each task, and was coded as the number of days since 1 January, scaled down by dividing by 100. We analysed male and female data separately because of the differences in testing schedules and locations, described above. We also present the results of pooling male and female data, with year, date-within-year and sex as fixed effects. For the pooled analyses, year was coded as 1, 2 or 3 for males and as 1 or 3 for females (because they were tested in the first and third years of the study; coding year as 1 or 2 for females did not alter the results).

Correlations Among Tasks and Song Features

In the first set of tests on these birds (i.e. the first year), Anderson et al. (2017) found that performance on colour association was correlated with performance on colour reversal, and performance on the detour-reaching task was negatively correlated with neophobia. To ascertain whether these correlations remained similar from year to year, we first redid these first-year analyses separately for males and females, and then examined the same pairings within each subsequent year (the second and third years for males, and the third year for females), using Spearman's rank correlation tests.

We also assessed correlations between cognitive performance and song quality in males, in data from the second and third rounds of testing, using the same approach that Anderson et al. (2017) used on data from the first round of tests. First, we used Spearman's rank correlation tests to assess correlations between each song measure (repertoire size, percentage of notes copied, copy accuracy) and performance on each cognitive task. We then used GLMMs (using the R package 'lme4') to model the association of each song measure with cognitive task performance, with random intercepts for individual bird, nest of origin and particular task. These models were the same as those used by Anderson et al. (2017) except for two differences: first, we did not include novel foraging as a cognitive task, and second, we removed the effect of neophobia from the model after verifying that this was insignificant in the first year. For each song

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020

measure, we ran a separate model for each year, and we also ran a fourth model using the average performance scores across all 3 years. For all tests, unadjusted *P* values are reported.

Ethical Note

All procedures were approved by the Institutional Animal Care and Use Committees of either Duke University (Protocol A032-14-02) or Florida Atlantic University (Protocol A15-28) and followed the Guidelines for the treatment of animals in behavioural research of the Animal Behavior Society and the Association for the Study of Animal Behaviour. We carefully observed all trials, and birds exhibited no signs of stress during testing. In housing and caring for birds between trials, we worked to maximize the health and longevity of the birds by causing as little stress as possible.

RESULTS

Summary of Performance Across Years

Figs. 1, 2 and 3 show neophobia scores and cognitive task performance for each round of testing. In some cases, average performance scores improved with repeated testing, but we observed considerable variation. At least half of the males performed worse in the second year than in the first on three of the four cognitive tasks (all but detour reaching), and a third or more of the females did worse in their second year of testing on all four tests. Within each task, some birds improved and some got worse from one year to the next (Fig. 4 illustrates this for the spatial learning task). Every bird performed worse on at least one task in the second round of tests relative to the first.

Repeatability of Cognitive Performance

Spearman's tests revealed little consistency in the rankings of birds, by either neophobia or performance scores, across consecutive years. In females, no rankings were similar across years. In males, neophobia rankings were similar across years (years 1 versus 2: $r_S = 0.472$, $P = 0.048$; years 2 versus 3: $r_S = 0.656$, $P = 0.003$). On the cognitive tasks, however, males were not ranked similarly in performance from one year to the next, and on the colour reversal task, male rankings were negatively associated between years 2 and 3 ($r_S = -0.496$, $P = 0.043$).

GLMM analyses (Table 1) revealed significant repeatability in neophobia for females ($P = 0.011$) and males ($P = 0.005$), with the 95% CI excluding zero in both cases. The only cognitive task with a significantly repeatable performance score, according to this approach, was colour reversal in females ($P = 0.04$), but here the 95% CI included zero. Pooling the male and female data resulted in narrower 95% CIs but also tended to lower the value of *R* and did not yield significant repeatability in any of the cognitive task scores (Table 1).

Correlations Among Cognition, Neophobia and Song Quality

We confirmed the previously reported correlation between colour association and colour reversal in the first year, in both females ($r_S = 0.486$, $P = 0.048$) and males ($r_S = 0.647$, $P < 0.005$), but found no correlation between the two in later rounds of testing for either sex. In females, detour reaching was not correlated with neophobia in either year of testing. In males, the negative correlation between detour reaching and neophobia existed only in the first year ($r_S = -0.492$, $P = 0.038$). In this latter test, our values of r_S and *P* for the first year differed slightly from those obtained by Anderson et al. (2017) because our analysis excluded one male who died after the first year of testing (the number of males in Anderson et al., 2017, was 19).

Spearman's tests revealed no correlations between song quality and cognitive task performance in the first or third years of testing. For the first year of testing, this result replicates Anderson et al. (2017) but with one less bird. In the second year of testing, copy accuracy was positively correlated with colour association ($r_S = 0.507$, $P = 0.032$) and negatively correlated with colour reversal ($r_S = -0.507$, $P = 0.032$), but no other correlations existed. (Note that the positive correlation reported here means that better copy accuracy was correlated with worse performance on the colour association task, because higher scores on the cognitive tasks reflect lower performance. The negative correlation means that better copy accuracy was correlated with better performance on the colour reversal task.) Copy accuracy was negatively correlated with 3-year average colour reversal scores ($r_S = -0.616$, $P = 0.006$), but this was the only correlation between a 3-year average cognitive test score and a song quality measure.

GLMMs indicated associations between song quality and cognitive task performance for three of the four tasks in the first year but different patterns of association in the second and third years (Table 2; complete model results in Supplementary Table S1). Of the 12 combinations of song measures ($N = 3$) and cognitive tasks ($N = 4$), none exhibited a consistent association across all 3 years. Repertoire size and percentage copied were positively associated with the 3-year average score on detour reaching, and negatively associated with the average scores on colour reversal and spatial learning. Copy accuracy was not correlated with any of the 3-year average task scores (Table 2).

DISCUSSION

We found low temporal repeatability (sensu Cauchoix et al., 2018) of performance on tests of cognitive ability in song sparrows tested in multiple years. Performance on these tests might be expected to improve over time, and it did in some cases (Figs. 1, 2 and 3). This improvement was not consistent, however (Fig. 4). In contrast,

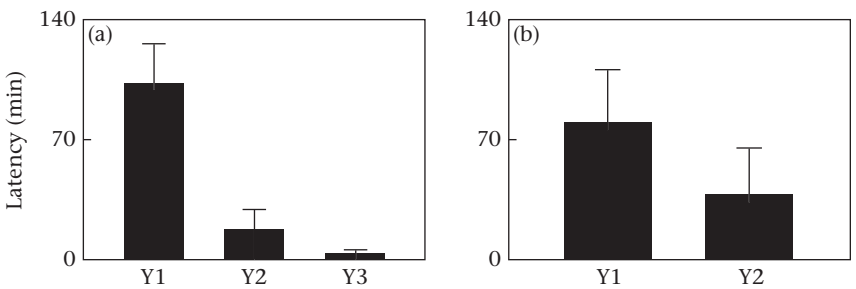

Figure 1. Neophobia in repeated rounds of testing in song sparrows, measured as the number of minutes elapsed before subjects ate from a novel feeding grid. (a) Males ($N = 18$) were tested yearly three times (Y1 = 2014, Y2 = 2015, Y3 = 2016) and (b) females ($N = 19$) were tested twice (Y1 = 2014, Y2 = 2016). Means ± SE across individuals are shown.

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020

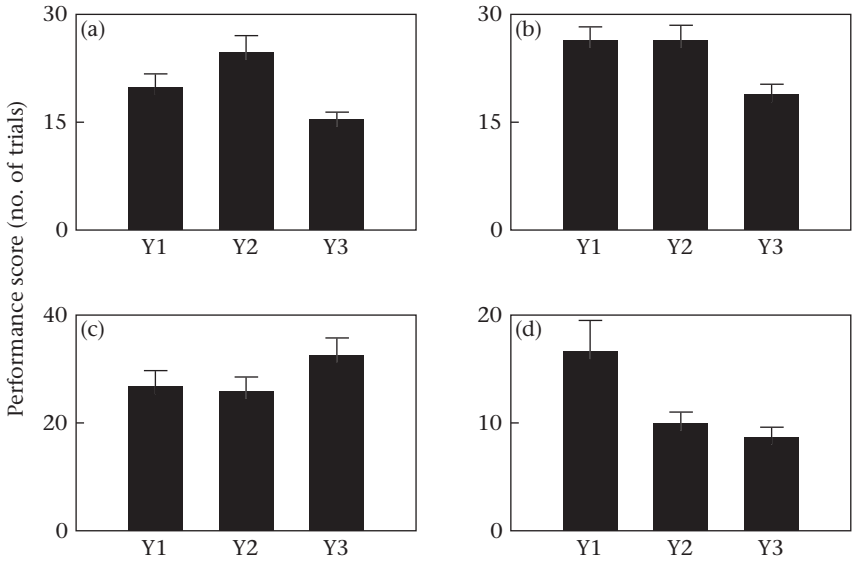

Figure 2. Performance on four cognitive tasks in repeated rounds of testing in male song sparrows ($N = 18$): (a) colour association, (b) colour reversal, (c) spatial learning and (d) detour reaching. Means ± SE across individuals are shown. Years of testing are Y1 = 2014, Y2 = 2015, Y3 = 2016. Scores represent the number of trials required for a bird to succeed at a task, so improvement in performance is indicated by a shorter bar in the second or third year.

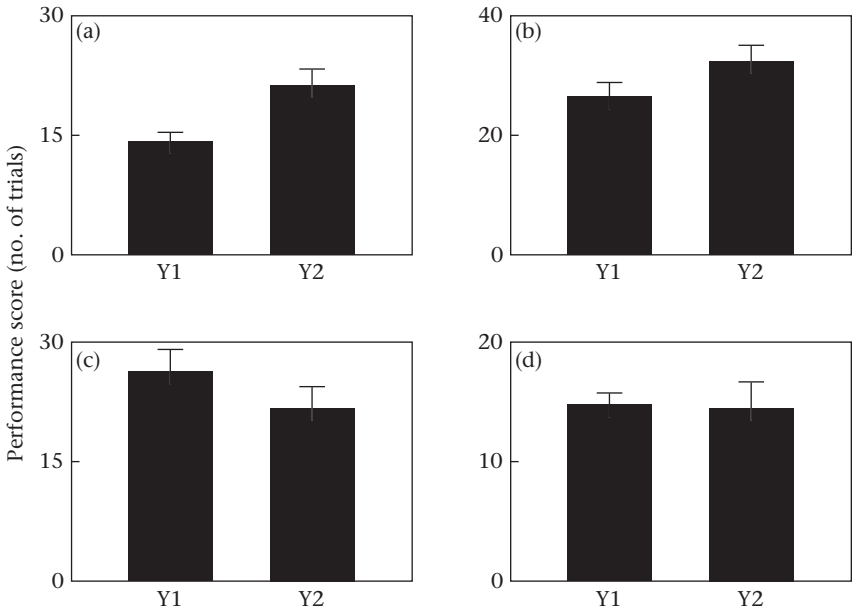

Figure 3. Performance on four cognitive tasks in repeated rounds of testing in female song sparrows ($N = 19$): (a) colour association, (b) colour reversal, (c) spatial learning and (d) detour reaching. Means ± SE across individuals are shown. Years of testing are Y1 = 2014, Y2 = 2016. Scores represent the number of trials required for a bird to succeed at a task, so improvement in performance is indicated by a shorter bar in the second year.

neophobia was repeatable when assayed at the same time intervals. This neophobia result is qualitatively in agreement with previous studies that documented repeatability of behaviour (Bell, Hankison, & Laskowski, 2009) and with findings that, in song sparrows, aggressive approach and signalling behaviours are individually consistent (Akçay, Campbell, & Beecher, 2014; Hyman, Hughes, Searcy, & Nowicki, 2004; Nowicki, Searcy, Krueger, & Hughes, 2002). Our neophobia result also confirms that our sample size and testing schedule were sufficient to detect repeatability, at least for R values above 0.25. Across four cognitive tasks, however, the average value of R (from unpooled analyses) in our study was only 0.13 (range 0–0.41; Table 1). Pooling the data from males and females did not yield significant repeatability and tended to decrease R values, but larger sample sizes might yield different results.

The low repeatability values we found are typical of studies of cognitive performance in nonhuman animals. In their recent meta-analysis, Cauchoix et al. (2018) assessed repeatability of cognitive performance using 44 data sets from insects, molluscs, reptiles, birds and mammals (including humans). The average value reported in that study for 'temporal repeatability adjusted for test order', the measure most similar to the one we used, was 0.15. This average value is inflated to some degree by the inclusion in its calculation of a number of repeatability estimates from studies of humans, which in general yield substantially higher repeatabilities

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020

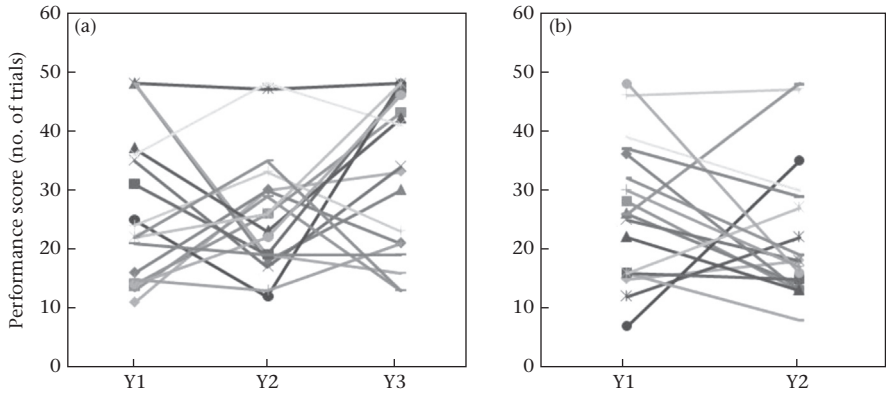

Figure 4. Variation among individual song sparrows in performance on the spatial learning task across (a) three rounds of testing in males (Y1 = 2014, Y2 = 2015, Y3 = 2016) and (b) two rounds of testing in females (Y1 = 2014, Y2 = 2016). Lines connect data points within individual birds across years.

Table 1
Repeatability of neophobia measures and cognitive test scores, across years, in song sparrow males (tested three times) and females (tested twice)

	Males	Females	Pooled
Neophobia	0.26 (0.01–0.57) $\boldsymbol{P = 0.005}$	0.35 (0.08–0.79) $\boldsymbol{P = 0.011}$	0.20 (0.05–0.42) $\boldsymbol{P = 0.004}$
Colour association	0.22 (0–0.50) $P = 0.073$	0 (0–0.46) $P = 0.500$	0.05 (0–0.32) $P = 0.324$
Colour reversal	0 (0–0.29) $P = 1.00$	0.41 (0–0.72) $\boldsymbol{P = 0.040}$	0 (0–0.25) $P = 1.00$
Spatial learning	0.08 (0–0.37) $P = 0.280$	0.08 (0–0.54) $P = 0.359$	0.08 (0–0.32) $P = 0.255$
Detour reaching	0.11 (0–0.40) $P = 0.220$	0.17 (0–0.61) $P = 0.230$	0.06 (0–0.31) $P = 0.295$

Repeatability values (R) are followed by 95% confidence intervals (in parentheses) and P values. All values were obtained by GLMM analyses using the R package rptR. Significant P values are indicated in bold.

for cognitive measures than do studies of other species (Cauchoix et al., 2018). At a more detailed level, Table S1 in Cauchoix et al. (2018) lists 27 individual values of R_n (temporal repeatability adjusted for test order) obtained from nonhuman animals. The individual nonpooled R values in our study, constituting eight values from two sexes and four tasks, do not differ significantly from those 27 values (Mann–Whitney U test: $W = 84$, $N_1 = 27$, $N_2 = 8$, $P = 0.36$). Our results differ from those of Cauchoix et al. (2018) in the distribution of 95% confidence intervals: all eight of our R values had associated 95% confidence intervals that included zero, whereas only half (13/27) of those listed by Cauchoix et al. (2018) did. However, this latter proportion increases to 7/10 when only the avian data sets analysed by Cauchoix et al. (2018) are considered.

Cauchoix et al. (2018) examined a number of factors that they thought might explain variation in repeatability across the studies in their meta-analysis. The strongest association they found was with the type of cognitive measure used; in their analysis, the type of measure we employed — the number of trials needed to reach a cognitive criterion — yielded higher-than-average repeatabilities. Laboratory-reared subjects, such as we used, tended to give low repeatabilities, but that effect was not significant. Testing animals in the laboratory, as we did, tended to give higher repeatability than testing in the wild, but that effect was also not significant. The duration of the interval between successive tests had no association with repeatability. As Cauchoix et al. (2018) point out, however, their sample size was limited and thus all of these results should be interpreted with caution. These aspects of our study might still partially explain the low repeatabilities we found. In particular, our study subjects were raised and maintained in a less cognitively demanding environment than that to which the species is adapted. It is possible that laboratory-reared nondomestic animals do not develop or consistently use their full cognitive capacity, and this might negatively affect repeatability of performance on cognitive tasks.

Within each sex, the birds in this study all lived in the same environment and had access to the same types and quantities of resources during and between testing. This limits the amount of variation in individual experience, a factor that might affect performance on cognitive tests (Rowe & Healy, 2014). However, it is possible that birds experienced different levels of distraction by other birds during each round of testing, either by specific individuals or by the overall number of individuals in the room, which was not always the same. The other birds in the room were audible but not visible to subjects. This set of other birds differed each year, during a given task for a given subject, and in some cases changed part-way through a task. Distraction by other birds might therefore have caused idiosyncratic effects (i.e. on the performance of some individuals on some tasks in some years).

Another potential confounding factor is variation in energetic state, which presumably determines the birds' motivation to complete a task. We attempted to minimize variation in motivation by removing food 5 h prior to the first test each day so that all birds were hungry when testing began. We then tested motivation after the last trial each day by measuring the latency to feed from a familiar seed container. For each bird in each year, the average latency was under 15 s. However, we agree with Rowe and Healy

Table 2
Results of GLMMs testing associations between song measures and cognitive task scores, showing inconsistency across years

	Repertoire size	Percentage copied	Copy accuracy
Colour association	0 / + / 0 // 0	0 / 0 / 0 // 0	0 / + / 0 // 0
Colour reversal	− / 0 / 0 // −	− / 0 / 0 // −	+ / − / 0 // 0
Spatial learning	− / 0 / + // −	− / 0 / + // −	+ / 0 / − // 0
Detour reaching	+ / 0 / − // +	+ / 0 / − // +	− / 0 / + // 0

Four models were run per song measure, each model including scores from all four cognitive tasks. Results of these models are shown as follows: year 1 / year 2 / year 3 // 3-year average. + = positive association; − = negative association; 0 = no association ($\alpha = 0.05$).

Table 3
Post hoc tests for effects of lid colour and spatial memory on task performance

	M-2014	M-2015	M-2016	F-2014	F-2016
Colour association					
Colour 1	B(9): 19.8 ± 1.9	G(9): 23.8 ± 3.8	K(9):14.7 ± 1.6	B(10): 13.8 ± 1.8	G(9): 16.3 ± 1.5
Colour 2	Y(10): 19.6 ± 3.2	R(9): 25.4 ± 3.2	W(9):16.1 ± 1.3	Y(7): 14.9 ± 1.5	R(10):26.0 ± 2.5
	$W = 48$, $P = 0.84$	$W = 35$, $P = 0.66$	$W = 32$, $P = 0.48$	$W = 27.5$, $P = 0.49$	$W = 13$, $\mathbf{P = 0.01}$
Colour reversal					
Colour 1	B(10): 24.6 ± 2.6	G(9): 27.0 ± 2.9	K(8): 17.6 ± 1.9	B(9): 28.4 ± 3.6	G(10): 31.7 ± 3.5
Colour 2	Y(9): 28.8 ± 1.6	R(9): 25.9 ± 2.7	W(9): 19.8 ± 2.5	Y(10):24.8 ± 2.3	R(9): 33.1 ± 3.8
	$W = 63$, $P = 0.15$	$W = 39.5$, $P = 0.96$	$W = 40.5$, $P = 0.70$	$W = 38$, $P = 0.60$	$W = 52$, $P = 0.59$
Spatial task preliminary trials					
First visits		7/18 ($P = 0.18$)	3/18 ($P = 0.59$)		9/18 ($\mathbf{P = 0.02}$)
Bias		5/18 ($P = 0.79$)	6/18 ($P = 0.42$)		5/18 ($P = 0.79$)
Spatial task improvement					
Same corner		4.2 ± 6.7 (4)	12.7 ± 6.6 (3)		6.5 ± 7.2 (6)
New corner		−0.5 ± 4.3 (14)	−10.5 ± 3.7 (15)		3.0 ± 4.0 (12)
		$W = 19$, $P = 0.37$	$W = 4$, $\mathbf{P = 0.03}$		$W = 25$, $P = 0.32$

Columns include data for each sex per year of testing. For colour association and colour reversal, mean ± SE scores are given for birds grouped by rewarded colour (B = blue, Y = yellow, G = green, R = red, K = black, W = white), number of birds is indicated in parentheses, and test statistics and *P* values are from Wilcoxon–Mann–Whitney two-sample tests. For the spatial task, the proportion of birds that first visited the previously rewarded corner in the initial preliminary trial and the proportion showing bias towards the previously rewarded corner (by visiting it first in at least half of the preliminary trials) are given, with *P* values from binomial tests. Mean ± SE improvement in spatial task scores over the previous year are then given for birds grouped by whether the same corner or a new corner was rewarded; number of birds is indicated in parentheses, and test statistics and *P* values are from Wilcoxon–Mann–Whitney two-sample tests. Significant *P* values are indicated in bold.

(2014) that motivation is difficult if not impossible to control completely. In cognitive tests with humans, typically no external motivators are used, as humans are assumed to be internally motivated. This might partially explain why repeatability is higher on cognitive tests in humans than in nonhuman animals. In actuality, use of material incentives affects performance on intelligence tests in humans, and controlling for motivation weakens the association between intelligence scores and life outcomes (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011).

Two other possible sources of variation in performance in our study are the date of testing and the colours of the plastic lids used each year. Birds were on a natural photoperiod, and normal seasonal changes in anatomy and physiology might affect their performance on cognitive tasks. For each bird, the dates of testing differed across years because we began testing on different dates and randomized the order in which individuals were tested. We accounted for this in the repeatability analysis (GLMM) by including date-within-year as a factor. More generally, testing of males began during the breeding season in the first 2 years and a few weeks before the breeding season in the third year; despite this, our Spearman's tests did not reveal higher repeatability between the first 2 years than between years 2 and 3.

The colours of the plastic lids used were intentionally changed each year, in order to test colour association and reversal independently of memory of the colours used in the previous year. Within each year, all birds were tested on the same pair of colours, but half of the birds were initially rewarded on one colour and half on the other. This enabled us to conduct post hoc tests for general effects of colour (Table 3). We found no effect of rewarded colour on performance on either the colour association or the colour reversal tasks in males. Females did better on the colour association task when green was rewarded over red (Table 3), which might have affected repeatability on the colour association task in this sex. Rewarded colour did not affect performance on the colour reversal task in females, however. Idiosyncratic effects on performance are also possible; for example, if individual differences in cue salience occur (Rowe & Healy, 2014), one bird might more readily associate food with yellow (versus blue, in the first year) than with red (versus green, in the second year), and another bird might do the opposite. In addition, perceived relative similarity of a given colour to either of the colours used the previous year might affect individual performance. Despite these possibilities, colour reversal was the only cognitive task for which we found any evidence of repeatability (in females). It is possible that repeatability on the colour tasks would have been higher if all birds had been tested with the same colours in the same sequence across years. However, the detour-reaching task used no coloured lids, nor other equipment that differed across years, and repeatability was no higher on this task.

Beyond the potential confounding factors discussed above, it is important to recognize that the nature of a cognitive task changes on repeated performance by the same individual, and to consider the extent to which this might affect repeatability. When cues change between runs, as in our colour association and colour reversal tests, familiarity with the structure of the test might affect performance in later rounds. For tasks in which the correct solution can change between rounds of testing, as in our spatial learning task, memory of the first solution might interfere with learning of the second solution. In post hoc analyses, we found that females tended to first visit the previously rewarded location in the initial preliminary trial, but neither sex was biased towards this location across all six preliminary trials (Table 3). Given this apparent lack of bias, the same location was rewarded in consecutive tests for a few birds, and this might have improved task performance (Table 3). Even in tests administered more consistently each time, memory can contribute to performance in later runs, as suggested by the improvement in performance on the detour-reaching task by males in the second year. Such tests can therefore morph from tests of trial-and-error learning to tests of long-term memory. It is therefore difficult to measure exactly the same cognitive trait twice in the same individual. To the extent that different abilities underlie performance on initial versus later repetitions of a task, differences in cognitive styles (Sih & Del Giudice, 2012) could result in variation among individuals in relative performance on initial versus later repetitions of the same task, yielding low temporal repeatability.

As our results demonstrate, low repeatability of cognitive performance means that conclusions about correlations between cognitive abilities and other traits should be treated with caution if these are based only on tests done once. Anderson et al. (2017) found correlations between song quality and certain cognitive test scores using GLMMs, but these were inconsistent with the results of previous studies (Boogert et al., 2011; Sewall et al., 2013). We found that these same birds, tested 1 and 2 years later, yielded different results (Table 2). Studies investigating how cognitive

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020

ability relates to other traits should include repeated tests of cognitive performance. Similar arguments are made by Thornton, Isden, and Madden (2014) for studies linking cognition to fitness and by Griffin, Guillette, and Healy (2015) for those linking cognition to personality.

Conclusions about correlations between multiple cognitive traits should also not be made based on a single set of tests. Performance scores on the colour association and colour reversal tasks were correlated in the first year of testing (Anderson et al., 2017), but not in our subsequent tests. This latter finding is consistent with the idea that songbird cognition is highly modular (see Searcy & Nowicki, 2019); it could be that performance scores on the colour association and colour reversal tests were uncorrelated because the neural mechanisms underlying the abilities measured in these tests are more independent than previously thought. In pheasants, van Horik, Langley, Whiteside, Laker, and Madden (2018) found intra-individual variation in performance even across multiple assays supposedly testing the same cognitive domain. However, the low repeatability across repetitions of the exact same test (e.g. the detour-reaching test) in our study indicates that, as discussed above, factors other than task identity and individual ability affect performance, and these could cause variation across tests within the same domain.

Repeated testing and assessment of average performance might be a useful approach in future studies of relationships between cognitive ability and other traits. Prior to the repeated testing done in our study, it remained possible that features of song sparrow song could signal other cognitive abilities but that other effects on task performance obscured this relationship in the initial round of testing. The three earlier studies in song sparrows (Anderson et al., 2017; Boogert et al., 2011; Sewall et al., 2013) disagree with one another in certain findings, and these discrepancies might largely be explained by the effects on task performance of factors other than cognitive ability. Despite the variation introduced by these factors, our GLMM results suggest that average cognitive performance over repeated trials is correlated with some song features (Table 2). However, the direction of these correlations is not consistent. This reinforces our previous conclusion (Anderson et al., 2017) that, in the song sparrow, song and other cognitive abilities are not positively correlated overall.

Acknowledgments

We thank Sabah Ali, Caitlin Cantrell, Nali Gillespie and Philippa Tanford for help with hand-rearing and testing and Santiago Olivella, Casey Klofstad and Matthew Zipple for help with the statistical analyses. This work was supported by National Science Foundation Grants to S.N. (IOS-1144991) and W.A.S. (IOS-1144995).

Supplementary Material

Supplementary material associated with this article is available, in the online version, at https://doi.org/10.1016/j.anbehav.2019.09.020.

References

Akçay, Ç., Campbell, S. E., & Beecher, M. D. (2014). Individual differences affect honest signalling in a songbird. *Proceedings of the Royal Society B: Biological Sciences, 281*, 20132496.

Anderson, R. C., Searcy, W. A., Peters, S., Hughes, M., DuBois, A. L., & Nowicki, S. (2017). Song learning and cognitive ability are not consistently related in a songbird. *Animal Cognition, 20*, 309–320.

Ashton, B. J., Ridley, A. R., Edwards, E. K., & Thornton, A. (2018). Cognitive performance is linked to group size and affects fitness in Australian magpies. *Nature, 554*, 364–367.

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: A meta-analysis. *Animal Behaviour, 77*, 771–783.

Biro, P. A., & Stamps, J. A. (2015). Using repeatability to study physiological and behavioural traits: Ignore time-related change at your peril. *Animal Behaviour, 105*, 223–230.

Boogert, N. J., Giraldeau, L.-A., & Lefebvre, L. (2008). Song complexity correlates with learning ability in zebra finch males. *Animal Behaviour, 76*, 1735–1741.

Boogert, N. J., Anderson, R. C., Peters, S., Searcy, W. A., & Nowicki, S. (2011). Song repertoire size in male song sparrows correlates with detour reaching, but not with other cognitive measures. *Animal Behaviour, 81*, 1209–1216.

Brust, V., Wuerz, Y., & Krüger, O. (2013). Behavioural flexibility and personality in zebra finches. *Ethology, 119*, 559–569.

Burt, J. M., Campbell, S. E., & Beecher, M. D. (2001). Song type matching as threat: A test using interactive playback. *Animal Behaviour, 62*, 1163–1170.

Cauchard, L., Boogert, N. J., Lefebvre, L., Dubois, F., & Doligez, B. (2013). Problem-solving performance is correlated with reproductive success in a wild bird population. *Animal Behaviour, 85*, 19–26.

Cauchoix, M., Chow, P. K. Y., van Horik, J. O., Atance, C. M., Barbeau, E. J., Barragan-Jason, G., et al. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*, 20170281.

DuBois, A. L., Nowicki, S., Peters, S., Rivera-Cáceres, K. D., & Searcy, W. A. (2018). Song is not a reliable signal of general cognitive ability in a songbird. *Animal Behaviour, 137*, 205–213.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 7716–7720.

Dugatkin, L. A., & Alfieri, M. S. (2003). Boldness, behavioral inhibition and learning. *Ethology. Ecology & Evolution, 15*, 43–49.

Griffin, A. S., Guillette, L. M., & Healy, S. D. (2015). Cognition and personality: An analysis of an emerging field. *Trends in Ecology & Evolution, 30*, 207–214.

Guillette, L. M., Hahn, A. H., Hoeschele, M., Przyslupski, A.-M., & Sturdy, C. B. (2015). Individual differences in learning speed, performance accuracy and exploratory behaviour in black-capped chickadees. *Animal Cognition, 18*, 165–178.

Hyman, J., Hughes, M., Searcy, W. A., & Nowicki, S. (2004). Individual variation in the strength of territory defense in male song sparrows: Correlates of age, territory tenure, and neighbor aggressiveness. *Behaviour, 141*, 15–27.

Keagy, J., Savard, J.-F., & Borgia, G. (2012). Cognitive ability and the evolution of multiple behavioral display traits. *Behavioral Ecology, 23*, 448–456.

MacKinlay, R. D., & Shaw, R. C. (2019). Male New Zealand robin (*Petroica longipes*) song repertoire size does not correlate with cognitive performance in the wild. *Intelligence, 74*, 25–33.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., et al. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America, 111*, E2140–E2148.

Mateos-Gonzalez, F., Queseda, J., & Senar, J. C. (2011). Sexy birds are superior at solving a foraging problem. *Biology Letters, 7*, 668–669.

Morand-Ferron, J., Cole, E. F., & Quinn, J. L. (2016). Studying the evolutionary ecology of cognition in the wild: A review of practical and conceptual challenges. *Biological Reviews, 91*, 367–389.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews, 85*, 935–956.

Nowicki, S., Searcy, W. A., Krueger, T., & Hughes, M. (2002). Individual variation in response to simulated territorial challenge among territory-holding song sparrows. *Journal of Avian Biology, 33*, 253–259.

Peters, S., Searcy, W. A., & Nowicki, S. (2014). Developmental stress, song learning, and cognition. *Integrative and Comparative Biology, 54*, 555–567.

Pravosudov, V. V., Roth, T. C., & II. (2013). Cognitive ecology of food hoarding: The evolution of spatial memory and the hippocampus. *Annual Review of Ecology, Evolution and Systematics, 44*, 173–193.

Rowe, C., & Healy, S. D. (2014). Measuring variation in cognition. *Behavioral Ecology, 25*, 1287–1292.

Searcy, W. A., & Nowicki, S. (2019). Birdsong learning, avian cognition, and the evolution of language. *Animal Behaviour, 151*, 217–227.

Searcy, W. A., McArthur, P. D., & Yasukawa, K. (1985). Song repertoire size and male quality in song sparrows. *Condor, 87*, 222–228.

Sewall, K. B., Soha, J. A., Peters, S., & Nowicki, S. (2013). Potential trade-off between vocal ornamentation and spatial ability in a songbird. *Biology Letters, 9*, 20130344.

Sih, A., & Del Giudice, M. (2012). Linking behavioural syndromes and cognition: A behavioural ecology perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 2762–2772.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution, 8*, 1639–1644.

Thornton, A., Isden, J., & Madden, J. R. (2014). Toward wild psychometrics: Linking individual cognitive differences to fitness. *Behavioral Ecology, 25*, 1299–1301.

USGS (United States Geological Survey). (2017). Longevity records of North American birds. https://www.pwrc.usgs.gov/BBL/longevity/Longevity_main.cfm.

van Horik, J. O., Langley, E. J. G., Whiteside, M. A., Laker, P. R., & Madden, J. R. (2018). Intra-individual variation in performance on novel variants of similar tasks influences single factor explanations of general cognitive processes. *Royal Society Open Science, 5*, 171919.

Please cite this article in press as: Soha, J. A., et al., Performance on tests of cognitive ability is not repeatable across years in a songbird, Animal Behaviour (2019), https://doi.org/10.1016/j.anbehav.2019.09.020