Reversal of response to artificial selection on body size in a wild passerine bird

Thomas Kvalnes[1,3], Thor Harald Ringsby[1], Henrik Jensen[1], Ingerid Julie Hagen[1], Bernt Rønning[1], Henrik Pärn[1], Håkon Holand[1], Steinar Engen[2] and Bernt-Erik Sæther[1]

[1]Centre for Biodiversity Dynamics (CBD), Department of Biology, Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway

[2]Centre for Biodiversity Dynamics (CBD), Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway

[3]Kvalnes (corresponding author): thomas.kvalnes@ntnu.no

Ringsby: thor.h.ringsby@ntnu.no

Jensen: henrik.jensen@ntnu.no

Hagen: ingerid.hagen@ntnu.no

Rønning: bernt.ronning@ntnu.no

Pärn: henrik.parn@ntnu.no

Holand: hakon.holand@ntnu.no

Engen: steinaen@math.ntnu.no

Sæther: bernt-erik.sather@ntnu.no

Running head: Artificial selection in the wild

Data archival location: Data will be uploaded to Dryad (datadryad.org) upon article acceptance.

Word count: 7204, Figures: 3, Tables: 6

KEY WORDS: Age structure, gene flow, individual fitness, natural selection, microevolution, *Passer domesticus*

Manuscript number: 17-0109

revised May 10, 2017

Acknowledgements

We are grateful to the numerous people that, over the years, collected data at our field sites; the island inhabitants which made the present study possible through their great hospitality and care; R. Lande for comments and discussions on concepts and analytic methodology; J. Hadfield, J. Reid and J. Conner for excellent comments on a previous version of this manuscript; editor Mohamed Noor, associate editor Derek Roff and two anonymous reviewers for excellent comments that improved the manuscript; and to the Research Council of Norway (project no. 204303 and 221956, and SFF-III 223257/F50), the European Research Council (ERC-2010-AdG 268562) and the Norwegian University of Science and Technology for funding. The authors of the present study have no conflicts of interest to declare. The study design was subject to review by the Norwegian Environment Agency (permits 2001/6427-ARTS/VI/ARE and 2004/1671 ARTS-VI-ID) and complied with the laws of animal welfare in Norway at the time the study was conducted (permits S-2603-01, S-204-8032-1 and S-2007/1482 from the Norwegian Animal Research Authority, and permits from the Bird Ringing Centre at Museum Stavanger, Norway).

Author Contributions

T.K. did the analyses and wrote the paper. T.H.R., H.J. and B.-E.S. came up with the idea and study design. H.J., I.J.H., B.R., H.H. and T.K. did the parentage analyses. S.E. advised the statistical analyses. All authors except S.E. performed fieldwork for the study. All authors contributed to the intellectual content through comments and edits when writing up the manuscript.

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Reversal of response to artificial selection on body size in a wild passerine bird

Journal:	*Evolution*
Manuscript ID	17-0109.R1
Manuscript Type:	Original Article
Keywords:	Age structure, Gene Flow, individual fitness, natural selection, microevolution, Passer domesticus

Reversal of response to artificial selection on body size in a wild passerine bird

Running head: Artificial selection in the wild

Data archival location: Data will be uploaded to Dryad (datadryad.org) upon article acceptance

Word count: 7204, Figures: 3, Tables: 6

KEY WORDS: Age structure, gene flow, individual fitness, natural selection, microevolution, *Passer domesticus*

Manuscript number: 17-0109

revised May 10, 2017

Abstract

A general assumption in quantitative genetics is the existence of an intermediate phenotype with higher mean individual fitness in the average environment than more extreme phenotypes. Here we investigate the evolvability and presence of such a phenotype in wild bird populations from an eleven-year experiment with four years of artificial selection for long and short tarsus length, a proxy for body size. The experiment resulted in strong selection in the imposed directions. However, artificial selection was counteracted by reduced production of recruits in offspring of artificially selected parents. This resulted in weak natural selection against extreme trait values. Significant responses to artificial selection were observed at both the phenotypic and genetic level, followed by a significant return towards pre-experimental means. During artificial selection, the annual observed phenotypic response closely followed the predicted response from quantitative genetic theory ($r_{years} = 0.96$, $r_{cohorts} = 0.56$). The rapid return to pre-experimental means was induced by three interacting mechanisms: selection for an intermediate phenotype, immigration and recombination between selected and unselected individuals. The results of this study demonstrates the evolvability of phenotypes and that selection may favour an intermediate phenotype in wild populations.

Introduction

Natural selection is a key process for adaptation of contemporary wild populations to changing environments (Endler, 1986). Understanding how and when selection shapes phenotypic variation is vital to interpret and understand observable temporal and spatial patterns in fitness related traits and to address evolutionary questions in management (Arnold et al., 2001; Kinnison and Hendry, 2001; Estes and Arnold, 2007; Uyeda et al., 2011; Bell, 2013; Sæther and Engen, 2015). Strong selection has repeatedly been shown to cause rapid adaptation in heritable traits (Endler, 1980; Grant and Grant, 1995; Losos et al., 1997; Reznick et al., 1997; Hendry and Kinnison, 1999; Reznick and Ghalambor, 2001; Darimont et al., 2009; Calsbeek and Cox, 2010). However, most of the time wild populations are subject to weak phenotypic selection while experiencing considerable demographic and environmental stochasticity in individual fitness (Kingsolver et al., 2001; Hereford et al., 2004; Rice, 2008; Coulson et al., 2010; Kingsolver et al., 2012; Engen and Sæther, 2014; Sæther and Engen, 2015; Morrissey, 2016; Hendry, 2017). This creates random variation in individual fitness among individuals and temporal variation in individual fitness among years, which complicates detection of selection on traits and conclusions on their adaptive significance (Arnold et al., 2001; Postma et al., 2007; Haller and Hendry, 2014; Engen et al., 2012; Engen and Sæther, 2014; Sæther and Engen, 2015; Hendry, 2017).

Basic features of phenotypic evolution was described by Simpson (1944), applying theoretical concepts originally provided by Wright (1932), as movements along a n-dimensional adaptive landscape, with variation in fitness for n quantitative traits (Simpson, 1944; Arnold et al., 2001; Hendry, 2017). Lande (1976, 1979) formalized this framework and showed that the evolutionary response to selection on correlated traits, $\mathbf{R}$, can be expressed by a multivariate extension of the breeder's equation $\mathbf{R} = \mathbf{G}\boldsymbol{\beta}$ (Lande, 1979). In this quantitative genetic model, $\mathbf{G}$ is the additive genetic variance-covariance matrix and $\boldsymbol{\beta}$ the vector of selection gradients, i.e. tangents on the adaptive landscape in the direction of higher fitness. The model has been applied successfully to empirical data in animal breeding and laboratory experiments (Hill and Caballero, 1992; Falconer and Mackay, 1996; Lynch and Walsh, 1998; Brakefield, 2003; Conner, 2003). For instance, when artificial selection has been used to explore quantitative genetic constraints (e.g. Beldade et al., 2002; Tigreros and Lewis, 2011; Bolstad et al., 2015) and predictions about rates of adaptive phenotypic evolution (e.g. Lendvai and Levin, 2003; Teuschl et al., 2007). In wild populations, the estimation of selection, genetic parameters and evolutionary responses is more difficult for several reasons. For instance, environmental and demographic stochasticity (Lande et al., 2003; Engen and Sæther, 2014), temporal environmental changes (Merilä et al., 2001), a misidentified target of selection (Price et al., 1988), selection on unmeasured genetically correlated traits (Lande and Arnold, 1983), and gene flow between adjacent populations (Hendry et al., 2001; Hendry, 2017). Accordingly, several empirical studies have reported an apparent lack of correspondence between observed and predicted phenotypic changes in traits under directional selection (Merilä et al., 2001; Brookfield, 2016). Many study populations have overlapping generations with age structure, where an individuals contribution to population growth depends on age-specific components of fecundity and survival (e.g. Reid et al., 2003). Fluctuations in the age distribution of such populations may cause transient phenotypic changes if phenotypes differ between age classes due to previous genetic drift or fluctuating selection (Lande, 1982; Coulson et al., 2003, 2006; Coulson and Tuljapurkar, 2008; Morrissey et al., 2012; Engen et al., 2009, 2011, 2012, 2014). If not accounted for, such temporal changes may conceal responses to actual selection and cause erroneous estimates of selection (Engen et al., 2014).

Selection experiments in the wild have a large potential to reveal novel insights into adaptive evolutionary dynamics, by manipulating the observed link between phenotypes and the environment (Arnold, 1983; Wade and Kalisz, 1990; Conner, 2003; Brakefield, 2003; Reznick and Ghalambor, 2005; Bell, 2008, 2010; Merilä and Hendry, 2014). There are two basic approaches to manipulate selection in natural populations: (1) indirectly by altering biotic or abiotic environmental factors or (2) directly by imposing artificial selection. Both approaches have their advantages; the first offers control over the causal agents of selection (e.g. Endler, 1980; Losos et al., 1997, 2001; Reznick et al., 1997;

Calsbeek and Smith, 2007; Calsbeek and Cox, 2010), while the second offers control over the applied strength of selection. When the main interest is the evolvability of a specific trait or a suite of traits within a population, the second approach is preferable (Wade and Kalisz, 1990; Conner, 2003; Brakefield, 2003; Hansen and Houle, 2004, 2008; Fuller et al., 2005; Bell, 2008, 2010; Merilä and Hendry, 2014). However, to our knowledge, only two artificial selection experiments in wild vertebrate populations have been reported, both on clutch size in birds (Flux and Flux, 1982; Postma et al., 2007). Flux and Flux (1982) artificially selected for large clutch size in starlings *Sturnus vulgaris*. The response was evident when comparing selected to unselected individuals. However, due to high levels of gene flow there was only a marginal response in the population as a whole. In a bidirectional experiment, Postma et al. (2007) artificially selected over eight years for increased and decreased clutch size in two subpopulations of great tit *Parus major*. Despite strong artificial selection, they found no clear evidence of evolutionary change in mean clutch size at the phenotypic level. Large environmentally induced variation in clutch size among years was believed to mask the response.

In the present study, artificial selection on tarsus length was applied in two wild populations of house sparrows *Passer domesticus*, to examine the evolvability of a fitness related trait and the degree to which observed trait values represent an adaptation to prevailing environmental conditions. Tarsus length was selected in opposite directions in the two populations for four subsequent years. Following the artificial selection, the populations were monitored for another seven years. An unmanipulated control population was monitored over the same period. The target of selection, tarsus length, is a heritable trait commonly used as a proxy for structural body size in passerine birds (Jensen et al., 2003, 2008; Rising and Somers, 1989; Senar and Pascual, 1997). The following five objectives were addressed. First, total phenotypic selection was estimated and the contribution from natural selection quantified. Second, variation in individual fitness was compared among individuals with different selective ancestry. Third, annual changes in tarsus length and other phenotypic traits were estimated. Fourth, the additive genetic (co)variance of the traits and the annual change in breeding values were quantified. Finally, observed responses to selection were compared to predictions from quantitative genetic theory.

Material and methods

Study system

The study was conducted in three insular populations of house sparrows in northern Norway. The islands, Hestmannøy ($66°33'N, 12°50'E$), Vega ($65°40'N, 11°55'E$) and Leka ($65°06'N, 11°38'E$), are located along a north-south gradient, separated by 97 (Hestmannøy-Vega) and 54 (Vega-Leka) km of ocean and small islands along the coastline (see map in Hagen et al., 2013). Thus, the geographical distance and the sedentary nature of the house sparrow ensured virtually no migration between the study populations (Altwegg et al., 2000; Tufto et al., 2005; Pärn et al., 2012). All individuals in the populations inhabit dairy farms and human settlements, where they breed in holes and cavities from May until mid-August (Ringsby et al., 1998). The mean generation time of house sparrows in natural populations in this area has been found to be 1.97 years (Stubberud et al., 2017).

In the years 2001-2012, individuals were captured and marked with a unique combination of a numbered metal leg ring from the Ringing Centre at Museum Stavanger and three plastic colour leg rings. Individuals were either followed from the nestling stage or when captured in mist nets during summer (May-August), autumn (late September-October; all populations) or winter (February-March; Leka and Vega). Over 90 % (Hestmannøy) and $\sim$ 90 % (Leka and Vega) of the winter populations were marked at all times during the study. At first capture, a small blood sample (25 μL) was collected, which enabled the construction of a genetic pedigree for each population. Parentage analyses were performed in Cervus 3.0 software with 95 % confidence for parentage assigned (Marshall et al., 1998; Kalinowski et al., 2007), based on genotyping putative parents and offspring for 14 microsatellite markers (Jensen et al., 2004, 2008; Rønning et al., 2016).

The data was organized with pre-breeding census and two age classes: 1 year old and 2+ years old. Hence, annual individual survival was recorded as 1 if an individual in year t was re-sighted (captured or observed) in year $t + 1$ (otherwise 0). Any emigrants from the islands where treated as dead individuals. For each individual, the annual number of offspring produced was recorded as the number of offspring born in year t that survived to year $t + 1$ (i.e. recruits). House sparrows go through a complete post-juvenile and post-breeding moult during autumn, after which ageing based on plumage is not possible. Hence, ageing was either made on individuals marked before the post-juvenile moult during summer or based on an assumption that all full-grown unmarked individuals were born in the most recently completed breeding season. Individuals which we were unable to age, were excluded from the analyses in the year they were marked. In addition, we excluded a few individuals with missing traits and all individuals from one farm at each experimental island, where we did not have access until the final years of the study.

Morphological measurements

Full-grown individuals were measured for tarsus length (± 0.005 mm), body mass (± 0.05 g), wing length (± 0.5 mm), bill length (± 0.005 mm) and bill depth (± 0.005 mm). The measurements were performed by several different fieldworkers. After an initial period of training, each fieldworker measured approximately 30 individuals together with T.H.R or, in some cases, another experienced fieldworker. Then all linear measurements were adjusted according to T.H.R. by adding mean differences when found significant ($P < 0.05$) using paired t-tests. All traits, except tarsus length, display seasonal variation (Anderson, 2006). Hence, only measurements from the main sampling periods were used in the analyses, i.e. summer for the Hestmannøy population and winter for the Leka and Vega populations. Furthermore, within-individual age effects were investigated for body mass, wing length, bill length and bill depth, using an extended data set over the years 1993-2012 at Hestmannøy and 2001-2012 at Leka and Vega. Due to the difference in sampling season, Hestmannøy was analysed separately. Traits were age-standardised by fitting a linear mixed effects model with age and age^2 as explanatory variables, random intercepts with year, cohort and individual identity, and an individual random slope to separate out any between-individual variation (Bates et al., 2015; Schielzeth and Forstmeier, 2009). The significance of each age variable was tested by likelihood ratio tests of nested models (fitted using maximum likelihood). All traits with significant age effects were adjusted to age 1, using predicted values from the model, before individual means were calculated.

Body mass scale with body size, measured as tarsus length, through an allometric relationship $bodymass = b \times bodysize^k$, where k is the allometric exponent (Huxley, 1932). This relationship was linearised for each sex and population separately by log transformation. Residuals from the log-log linear regressions were used as measures of individual body condition in subsequent analyses (Schulte-Hostedde et al., 2005).

Experimental procedure

Each winter of the four years 2002-2005, opposing artificial selection on tarsus length was imposed after a census in the Leka and Vega populations. During the experimental manipulations ~ 90 % of individuals in each population were captured and kept in a large aviary (abandoned cow barn) with *ad libitum* access to food (sunflower seeds, grain feed for cattle, oats and slices of bread), water and perching branches. The ranges in sample sizes during period of artificial selection (2002-2005) were 172-222 (Leka), 155-352 (Vega) and 59-80 (Hestmannøy), while the ranges in the subsequent period (2006-2012) were 89-216 (Leka), 102-330 (Vega) and 104-219 (Hestmannøy). Within each sex, all individuals with tarsi longer (Leka) or shorter (Vega) than the limit of mean ± 0.3 SD were returned to their origins, while the remaining individuals were translocated to populations located at least 70 km from the islands (see also Skjelseth et al., 2007). On average, 56.4 % (Leka) and 62.9 % (Vega) of all captured individuals were removed at each annual episode of artificial selection, such that the artificially selected individuals constituted approximately 78 % of the breeding populations. The whole procedure took between one and two weeks for each population. In the subsequent seven years (2006-2012) on Leka and Vega, the same procedure was followed, except that all individuals were returned to their origin. The Hestmannøy population was used as an unmanipulated control, where individuals were returned directly to the place of capture after banding and measurements. Henceforth, these populations are referred to as *high* (Leka, selected for large body size), *low* (Vega, selected for small body size) and *control*. Each individual in the *high* and *low* populations was assigned a selection category, selected, unselected, intermediate or other, based on whether their parents had been artificially selected (Table 1). Our genetic parentage analyses had a very high probability of assigning a parent to an individual, given that the parent had been sampled. Hence, when no genetic parent had been assigned to an individual, its parents were assumed to not have been artificially selected.

Data analysis

Phenotypic population differences

Differences in phenotype between populations in 2002, before the onset of the experiment, were explored using a multivariate analysis of variance (MANOVA). *Post hoc* tests for each trait were performed by separate analyses of variance (ANOVA). Tukey's range tests were applied to identify which populations differed phenotypically. Pairwise phenotypic correlations are shown in Table S1. Any sexual dimorphism in the traits was accounted for in the models by including sex as a categorical variable.

Analyses of directional selection

Analyses of directional selection were performed for each sex and population separately, and structured into two periods: (1) years 2002-2005 (with artificial selection) and (2) years 2006-2011 (without artificial selection). The demographic framework in the R package *lmf* was applied to analyse selection (Engen et al., 2012). This recently developed framework integrates evolutionary theory with an age-structured model for population dynamics, which accounts for overlapping generations and fluctuating age distribution in the estimation of selection (Engen et al., 2009, 2011, 2012, 2014). The annual absolute fitness of an individual j in age class i was defined by the individual reproductive value (Engen et al., 2009),

$$W_{ij} = J_{ij}v_{i+1} + B_{ij}v_1/2, \tag{1}$$

where J_{ij} is 1 if the individual survives (otherwise 0), B_{ij} is the number of recruits produced and v_{i+1} and v_1 are age-specific reproductive values (Engen et al., 2009; Sæther and Engen, 2015). Defining fitness this way enables correct estimation of an individual's contribution to the total reproductive value next year, by accounting for both survival and reproduction (Engen et al., 2011, 2012; Metcalf and Pavard, 2007; Wilson and Nussey, 2010; Sæther and Engen, 2015). However, additional insights into the selective processes could be obtained by analysing different fitness components separately. This was achieved by defining viability (W_{sij}) and fecundity (W_{fij}) fitness as the first and second additive component in equation 1 (Engen et al., 2011).

The age-specific reproductive values ($\mathbf{v}$), stable age distribution ($\mathbf{u}$) and deterministic multiplicative growth rate (λ) of a population are needed to calculate individual reproductive values and estimate selection gradients. These were obtained from the sex-specific mean projection matrix ($\mathbf{l}$), estimated separately for each population (Table S2) (Caswell, 2001). With two age classes, 1 year old and 2+ years old, $\mathbf{l}$ had age-specific fecundities (f_i) in the first row and age-specific survivals (s_i) in the bottom row. Age-specific fecundities and survivals for each sex and population were estimated as their means across the whole study period (Engen et al., 2011). In these calculations, experimentally removed individuals were excluded in the year they were removed. Then $\mathbf{v}$, $\mathbf{u}$ and λ were estimated as the scaled left and right eigenvector, and the dominant eigenvalue of $\mathbf{l}$ (Table S2) (Caswell, 2001). Eigenvectors were scaled according to $\Sigma u_i = 1$ and $\Sigma v_i u_i = 1$ (Engen et al., 2009). Conditioned on the sex-ratio at birth ($q =$ proportion of females) the growth rate of the male and female segment in each population has to be identical (Engen et al., 2010). Hence, we estimated the growth rate (λ_f) for females and set the growth for males equal to the females by scaling all male fecundities by a constant (c). The constant was estimated by solving the Euler-Lotka equation for the male segment of the population, $c(1-q)\sum_{k=1}^{\infty} \lambda^{-k} l_k m_k = 1$, using Newtons method. Here, $l_k = \prod_{i=1}^{k-1} s_i$, $m_k = f_{i=k}$, $\lambda = \lambda_f$ and in house sparrows the sex ratio at birth does not deviate significantly from 1:1 ($q = 0.5$, Anderson, 2006).

All k traits were centred by the global mean across years prior to analyses. Then directional selection gradients were estimated for each year and age class separately, using multiple regressions of absolute fitness on the trait values (Lande and Arnold, 1983; Engen et al., 2012). Annual selection gradients ($\boldsymbol{\alpha}_t = (\alpha_{0t}, \alpha_{1t}, ..., \alpha_{kt})$) were given as the weighted average of age-specific gradients, $\alpha_{mt} = \Sigma_i u_i \alpha_{imt}$, where $m = (0, 1, ..., k)$ (Engen et al., 2011). Then, assuming no fluctuating selection, the temporal mean selection gradients $\boldsymbol{\alpha} = \mathrm{E}\boldsymbol{\alpha}_t$ were estimated according to procedures in Engen et al. (2012). In addition to estimating the total directional selection (due to artificial and natural selection), we also estimated natural selection separately for the artificially selected individuals. Natural selection was separated into total, fecundity and viability selection. In this model, the growth rate λ is a measure of the expected individual reproductive value (i.e. the mean absolute fitness), with annual estimates given by, $\lambda_t = \Sigma_i u_i \mathrm{E}W_{it}$.

The directional selection coefficients ($\boldsymbol{\alpha}$) were estimated using absolute fitness. Hence, the standard SD-scaled selection gradients ($\boldsymbol{\beta}_{\boldsymbol{\sigma}}$) were calculated by $\boldsymbol{\beta}_{\boldsymbol{\sigma}} = \lambda^{-1}\boldsymbol{\alpha} \odot \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the vector of trait standard deviations (averaged over all years) and $\odot$ denotes element-wise multiplication (Engen et al., 2012). Statistical significance of temporal mean selection gradients was assessed using a multinormal bootstrap procedure for 10000 bootstrap replicates (Engen et al., 2012). 95 % confidence intervals were calculated from the estimated bootstrap distributions.

Demographic and environmental stochasticity, and selection are integral parts in the applied demographic framework for estimating selection. The demographic and environmental variance for the population were estimated as $\sigma_d^2 = \Sigma_i u_i \sigma_{di}^2$, where $\sigma_{di}^2 = \mathrm{E}\mathrm{var}(W_i|z, \varepsilon_t)$ and $\sigma_e^2 \approx \mathrm{var}(\alpha_{0t})$, where α_{0t} is the intercept in year t (Engen et al., 2012).

Variation in individual fitness

The difference in survival and production of recruits among selected, unselected and intermediate individuals (see Table 1) in the years 2003-2012 were analysed using mixed effects logistic and Poisson regression models, fitted using the R package *lme4* (Bates et al., 2015). As the proportion of selected individuals increases over years, an environmental (year) effect could not be estimated directly in the analyses without conflating it with fitness consequences from the experiment. Hence, a year effect (slope) was estimated for each of the two dependent variables with only unselected individuals. Among unselected individuals, there was no significant trend during the years 2003-2009 in recruit production in the *high* population (b_{year} = -0.03±0.03, χ^2 = 1.01, df = 1, P = 0.314), but a slight decrease in the *low* population (b_{year} = -0.07±0.03, χ^2 = 6.94, df = 1, P = 0.008). Survival rates did not show any significant temporal trend across years in unselected individuals in either population (*high*: b_{year} = 0.03±0.05, χ^2 = 0.29, df = 1, P = 0.587, *low*: b_{year} = -0.03±0.05, χ^2 = 0.31, df = 1, P = 0.577).

The significant decrease in recruit production in the *low* population was accounted for in subsequent analyses by fitting it as a covariate with known effect (i.e. offset). In addition, a random intercept associated with individual identity was estimated, age and sex were included to account for differences in survival and fecundity between ages and sexes, and two-way interactions to estimate age- and sex-specific differences among selection categories were included. The significance of the terms of interest were tested using likelihood ratio tests of pairs of nested models fitted with maximum likelihood, where twice the difference in log-likelihood is χ^2-distributed with $df_1 - df_2$ degrees of freedom.

Observed phenotypic change

Annual arithmetic mean phenotypes in age-structured populations are subject to transient temporal fluctuations due to fluctuations in the age distribution and variation in mean phenotype among age classes (Engen et al., 2014). Thus, phenotypic changes in each trait following artificial selection were explored by estimating annual weighted means and 95 % confidence intervals with weights $\mathbf{u}$. The weighting accounted for the effect of fluctuating age distribution on phenotypic means (Engen et al., 2014, 2012). Piecewise regression for each population was used to estimate the change in annual weighted mean phenotype across the years 2002-2012, with a breakpoint in 2006. Sex was included to account for any sexual dimorphism. These rates of responses to selection result from the partial transmission of selection to recruiting individuals and survival of adults, with the final response achieved when all individuals under selection have stopped reproducing.

Corresponding analyses were performed on cohort arithmetic means across the cohorts 2000-2011 with a breakpoint in cohort 2005, to investigate annual changes in recruited offspring separately. These means will be subject to transient temporal fluctuations due to fluctuations in age distribution and age-specific phenotypic means among parents. Each cohort consisted of offspring with two, one or no artificially selected parents (see Table 1). Hence, phenotypic changes across the cohorts 2000-2005 were also analysed separately within selected, intermediate and unselected offspring. Permutation tests were used to test whether slopes were significantly different from zero, and bootstraps were performed to estimate standard errors of the estimated slopes. In both cases 10000 iterations of the models were performed.

Quantitative genetic analyses

Analyses of additive genetic effects included phenotypes from 1141, 1404 and 554 individuals sampled from the *high*, *low* and *control* population over the years 2002-2012. Multivariate Bayesian animal models were constructed with all five traits to estimate additive genetic effects (breeding values), and the **G**-matrices with additive genetic variances and covariances (Lynch and Walsh, 1998; Kruuk, 2004; Hadfield, 2010). As sample sizes did not allow for separate analyses of females and males, models were constructed with sex as a categorical fixed effect. For each trait, phenotypic variation (σ^2_P) was separated into additive genetic variance (σ^2_A), cohort variance (σ^2_C) and residual variance (σ^2_R), such that $\sigma^2_P = \sigma^2_A + \sigma^2_C + \sigma^2_R$. The cohort effect ensured that estimated breeding values were unbiased with respect to any systematic environmental variation in phenotypes (Postma, 2006).

Models were fitted using *MCMCglmm* version 2.22.1 (Hadfield, 2010) with Gaussian distribution and identity link function. Prior to analyses, all traits were standardized by their standard deviation across all individuals to improve model mixing and ease construction of priors. The resulting $\mathbf{G}_{\boldsymbol{\sigma}}$-matrices have heritabilities on the diagonal and genetic correlations in off-diagonal elements. Priors for the fixed effects were the normal distribution with zero mean and large variance (10^{10}), while a parameter expanded prior was used for the variance components by specifying V = $\mathbf{I_5}$, nu = 5, alpha.mu = $\mathbf{0_5}$ and alpha.V = $\mathbf{I_5}\times 100$. Here $\mathbf{I_n}$ is the identity matrix and $\mathbf{0_n}$ is a zero vector with dimensions n. Care was taken to ensure good mixing of the chains and that specified priors did not have exaggerated influence on posterior distributions, by examining the sensitivity of the models to different choices of priors. In the analyses, runs with a burn-in period of 3000 and a thinning interval of 500 ensured low autocorrelation (generally < 0.1) for a total of 1000 independent random samples from the stationary posterior distribution. All estimates are reported as the posterior mode and 95% credibility intervals (CI).

For each trait and population, the temporal change in mean breeding value was analysed across years 2002-2012 and cohorts 2000-2011. Piecewise regression was used with annual weighted mean breeding value (weights $\mathbf{u}$) and a breakpoint in year 2006, or arithmetic cohort mean breeding value and a breakpoint in cohort 2005. To account for uncertainty in the estimated breeding values, these analyses were performed for each realization of the MCMC chain to obtain a full posterior distribution for temporal change (Hadfield et al., 2010). Thus, posterior modes for temporal change could be calculated with credibility intervals to assess whether the changes were significantly different from zero. We also quantified whether estimated slopes differed significantly from slopes expected under genetic drift. This was done by simulating random breeding values down the pedigree for each realization of the $\mathbf{G}_{\boldsymbol{\sigma}}$-matrix in the MCMC chain, using the *rbv* function in the *MCMCglmm* package (Hadfield, 2010). The probability of obtaining a slope of the magnitude observed or larger was then calculated as a two-tailed test using the posterior distribution of the slope under genetic drift.

Response to selection

To assess the agreement between observed phenotypic changes and predictions from quantitative genetic theory, the relationships between annual predicted and observed responses to selection were explored. Because we could only estimate the $\mathbf{G}$-matrix with sexes combined, observed and predicted responses were averaged across sexes. The annual observed phenotypic response to selection was calculated for each trait by subtracting the weighted mean of parents at time t from the weighted mean at time $t+1$, with weights $\mathbf{u}$. At time $t+1$, both adults which survived and recruiting offspring from known parents are included to calculate the weighted mean. To investigate the response in offspring separately, the observed phenotypic response in recruits were calculated by replacing the weighted mean at time $t+1$ by the arithmetic mean of recruiting offspring from known parents. The response in recruits will vary temporally due to fluctuations in the age distribution of parents, and will only capture the partial response because the final response will be achieved when all individuals under selection have stopped reproducing.

The annual predicted phenotypic response to selection ($\mathbf{R_t}$) averaged across females (f) and males (m) was calculated as

$$\mathbf{R_t} = \frac{(\mathbf{G}_{\boldsymbol{\sigma}}\boldsymbol{\beta}_{\boldsymbol{\sigma}\mathbf{tf}}) \odot \boldsymbol{\sigma}_{\mathbf{tf}} + (\mathbf{G}_{\boldsymbol{\sigma}}\boldsymbol{\beta}_{\boldsymbol{\sigma}\mathbf{tm}}) \odot \boldsymbol{\sigma}_{\mathbf{tm}}}{2}, \tag{2}$$

where $\mathbf{G}_{\boldsymbol{\sigma}}$ is the variance-standardized additive variance-covariance matrix, $\boldsymbol{\beta}_{\boldsymbol{\sigma}\mathbf{tf}}$ and $\boldsymbol{\beta}_{\boldsymbol{\sigma}\mathbf{tm}}$ are the vectors of variance-standardized selection gradients, and $\boldsymbol{\sigma}_{\mathbf{tf}}$ and $\boldsymbol{\sigma}_{\mathbf{tm}}$ are the vectors of phenotypic standard deviations. Analyses were performed using the statistical software R version 3.3.3 (R Core Team, 2016).

Results

In 2002, before the onset of artificial selection, there were significant phenotypic differences between the three populations (MANOVA: $F_{10,698} = 20.84$, $P < 0.001$, ANOVAs: all $P < 0.001$, Table S3). Tarsus length was shorter in the *low* population than in the *high* (mean difference = -0.27, $P = 0.004$) and *control* (mean difference = -0.44, $P < 0.001$) population, while the *high* and *control* populations did not differ significantly (mean difference = -0.18, $P = 0.277$).

Phenotypic selection

The artificial selection resulted in strong directional selection towards longer or shorter tarsus in the experimental populations in the years 2002-2005 (Fig. 1). There was no direct artificial selection on the other phenotypic traits (all $P > 0.05$, Table S4). When excluding artificial selection, there was significant directional natural selection on tarsus length towards the pre-experimental phenotypic mean in males of the *low* population (Fig. 1B). When separating natural selection into viability and fecundity selection, only fecundity selection was significant (Fig. 1B). A similar non-significant trend of directional natural selection towards pre-experimental means was also observed in females in the *low* population and in both sexes in the *high* population (Fig. 1A and B). Hence, there was a tendency for natural selection towards phenotypic pre-experimental means (Fig. 1A and B). There was no significant directional natural selection on phenotypic traits in the *control* population over the years 2002-2005 (all $P > 0.05$, Table S4).

During the seven years after the artificial selection ended (2006-2011), there was significant viability selection towards pre-experimental mean tarsus length in females in the *high* population, but the total directional selection was non-significant (Fig. 1C). Instead, there was positive directional selection for longer tarsus in males of the *high* population (Fig. 1D). This was the result of a combined effect of both fecundity and viability selection, as neither component was significant when analysed separately (Fig. 1D). There was no further significant directional natural selection detected in either the *high* or *low* population in the years 2006-2011 (all $P > 0.05$, Table S5).

The demographic variance (σ_d^2) was generally larger in both experimental populations during the period of artificial selection than in the subsequent period (*high*: $\Delta\sigma_d^2$ = -0.18, *low*: $\Delta\sigma_d^2$ = -0.45, Table S6). On average across the populations, the variance in recruit production decreased by 34.0 % and the variance in survival decreased by 4.3 % after completion of the period with artificial selection. Hence, removing individuals from the populations increased the demographic variation in recruit production during the manipulated breeding seasons.

Variation in individual fitness components

Selected and intermediate individuals produced significantly fewer recruits than unselected individuals in the *high* population ($\chi^2 = 9.65$, df = 2, $P = 0.008$, Table 2A). In the *low* population a similar pattern was evident among age 1 individuals (*selection status* × *age*: $\chi^2 = 10.92$, df = 2, $P < 0.001$, Table 2B), where selected individuals produced fewer recruits than unselected individuals. There were no significant differences in survival among individuals in different selection categories (*high*: $\chi^2 = 0.98$, df = 2, $P = 0.613$, *Low*: $\chi^2 = 2.58$, df = 2, $P = 0.275$). Hence, individuals with artificially selected parents appeared to have lower fitness than individuals with unselected parents.

Observed phenotypic change

In the period 2002-2006, the weighted mean tarsus length of both sexes significantly increased in the *high* population (b_{year} = 0.126±0.021, $P < 0.001$, Fig. 2A) and decreased in the *low* population (b_{year} = -0.112±0.020, $P < 0.001$, Fig. 2C). In the *control* population there was no significant change in weighted mean tarsus length during the same period (b_{year} = -0.027±0.028, $P = 0.367$, Fig. 2E). The weighted phenotypic mean of some of the other four traits also changed significantly from 2002 to 2006 in the experimental populations (Table 3).

Across the cohorts 2000-2005, arithmetic mean tarsus length of selected offspring increased significantly in the *high* population (b_{cohort} = 0.167±0.040, $P < 0.001$, Table 4A) and decreased in the *low* population (b_{cohort} = -0.091±0.041, $P = 0.035$, Table 4B).

Such changes were not evident among unselected offspring (*high*: b_{cohort} = 0.008±0.079, P = 0.898, *low*: b_{cohort} = 0.002±0.044, P = 0.964, Table 4). When pooling all offspring, there was still a significant increase in tarsus length across cohorts 2000-2005 in the *high* population (b_{cohort} = 0.099±0.031, P = 0.002, Fig. S1 and Table S7), whereas there was no significant change in the *low* population (b_{cohort} = -0.007±0.024, P = 0.786, Fig. S1 and Table S7). In the *control* population there was no significant change in tarsus length across the same cohorts (b_{cohort} = 0.000±0.038, P = 0.994, Fig. S1 and Table S7).

In the period 2006-2012, there was a significant decrease in weighted mean tarsus length in the *high* population (b_{year} = -0.088±0.013, $P < 0.001$, Fig. 2A). The *low* population displayed a marginally non-significant increase in weighted mean tarsus length over the same period (b_{year} = 0.027±0.013, P = 0.055, Fig. 2C). However, fig. 2C shows that the *low* population reached its pre-experimental weighted mean tarsus length already in 2007. Hence, both populations returned towards their pre-experimental tarsus length following the end of artificial selection. The other four traits generally also returned towards pre-experimental weighted means (Table 3). In the *control* population there was a slight decrease in weighted mean tarsus length over the years 2006-2012 (b_{year} = -0.035±0.016, P = 0.014, Fig. 2E).

Across the cohorts in the same period (2005-2011) there was a significant decrease in arithmetic mean tarsus length in the *high* population (b_{cohort} = -0.073±0.018, $P < 0.001$, Fig. S1 and Table S7). However, there was also a significant decrease in arithmetic mean tarsus length in both the *low* (b_{cohort} = -0.050±0.016, P = 0.002, Fig. S1 and Table S7) and *control* (b_{cohort} = -0.069±0.027, P = 0.003, Fig. S1 and Table S7) population.

Observed genetic change

In all three populations there were significant heritability for tarsus length and the other four traits (Table 5). Furthermore, there were positive genetic correlations between tarsus length and several of the other traits in the *high* and *low* populations (Table 5A and B). A similar pattern was found in the *control* population, but credibility intervals were wide enough to include zero for all genetic correlations (Table 5C).

Over the years 2002-2006, the weighted mean estimated breeding values for tarsus length increased significantly in the *high* population (b_{year} = 0.110, CI = [0.072, 0.152], Fig. 2B and Table 6A) and decreased significantly in the *low* population (b_{year} = -0.103, CI = [-0.137, -0.059], Fig. 2D and Table 6B). These changes were of larger magnitude than expected by genetic drift alone (*high*: $P < 0.001$, *low*: $P = 0.002$, Fig. 2B and D). In the subsequent period (2006-2012), the weighted mean estimated breeding values for tarsus length returned towards their pre-experimental means (*high*: b_{year} = -0.055, CI = [-0.081, -0.026], *low*: b_{year} = 0.044, CI = [0.021, 0.067]). Again the rates of change were larger than expected by genetic drift alone (*high*: $P = 0.037$, *low*: $P = 0.013$, Fig. 2B and D). Correlational change in estimated breeding values for the other traits were not larger than expected from genetic drift alone (Table 6A and B). Similarly, in the *control* population there were no changes in estimated breeding values larger than expected by genetic drift alone (Fig. 2F and Table 6C). Similar results were obtained for the annual changes in cohort arithmetic mean estimated breeding values (Fig. S1 and Table S8).

Observed and predicted response to selection

The observed response to selection closely followed the predicted response during the years of artificial selection ($r_{2002-2005} = 0.96$, Fig. 3A), with a tendency for observed responses to be of larger magnitude than predicted. This observed response include both adults which had survived and offspring that had recruited. Hence, the overshoot of the predicted response was as expected. When limiting the observed response to offspring that recruited, the partial observed response also followed the predicted response, but with larger deviation from the 1:1 line ($r_{2002-2005} = 0.56$, Fig. 3B). In the seven consecutive years with no artificial selection, there was no clear relationship between predicted and observed responses (total: $r_{2006-2011} = 0.15$, only recruits: $r_{2006-2011} = 0.06$).

Discussion

Artificial selection on tarsus length resulted in strong directional selection in opposite directions in two house sparrow populations (Fig. 1). However, individuals with at least one artificially selected parent produced fewer recruits than unselected individuals (Table 2), such that there was a tendency for natural selection to counteract artificial selection (Fig. 1). Still, artificial selection was much stronger than natural selection and resulted in a significant response in tarsus length in both experimental populations (Fig. 2, Tables 3 and 4). The observed phenotypic response during artificial selection closely followed the predicted response according to the multivariate breeder's equation (Fig. 3). Furthermore, the response in breeding values was much larger than expected by genetic drift alone (Fig. 2, Table 6). During the seven years period following the artificial selection, the mean tarsus length and estimated breeding values in the populations gradually returned towards their pre-experimental means (Fig. 2, Tables 3 and 6). Again, the rates of change in breeding values were larger than expected by genetic drift alone (Table 6).

Any finite population may undergo random phenotypic and genetic changes due to genetic drift (and mutation in the long run) (Lande, 1976; Swallow et al., 2009). Replicated selection lines in artificial selection experiments have obvious advantages for estimating the average response and to separate between selection and genetic drift as causes of phenotypic change (Henderson, 1989, 1997; Konarzewski et al., 2005; Swallow et al., 2009). However, in artificial selection experiments in natural populations, adding replicates involves synchronous experiments on additional suitable populations with similar population dynamics and under the same environmental influences. Even if such populations were available, it would represent a considerable increase in logistic effort, which was infeasible in the present system. Instead, we applied a bidirectional design to explore selection for both increased and decreased trait values. The construction of genetic pedigrees allowed us to conduct simulations of change in breeding values under genetic drift. Hence, the probability that the observed changes could have occurred by genetic drift alone could be quantified following Hadfield et al. (2010) and Postma (2006). Although, we were not able to estimate confidence intervals on the average expected responses under replicated experiments, we were still able to exclude genetic drift as an explanation for our results. This approach has previously been applied to observational studies in natural populations. For instance, to distinguish the effects of genetic drift and trophy hunting as causes of temporal change in horn length in bighorn sheep *Ovis canadensis* (Pigeon et al., 2016).

Artificial selection experiments in the wild necessitate capture and tracking of a large proportion of individuals in a population to perform selection and obtain unbiased estimates of responses. Here, a morphological trait was subject to selection by removing individuals with phenotypic values more extreme than a given threshold value. Our effort to capture and include all individuals in the experiment was considerable. Despite this, sampling was still incomplete and approximately 20-25 % of the breeding populations remained unselected each year. Most of the unselected individuals were located in unavailable subpopulations at mainly one farm on each study island. This resulted in a mixture of selected, intermediate and unselected offspring which recruited to the populations. High quality genetic pedigrees allowed us to distinguish between these individuals. Hence, offspring with unselected parents could be excluded to obtain unbiased estimates of responses to artificial selection, and offspring which differed in selective background could be contrasted to explore the variation in each component of individual fitness. A similar use of contrasts was applied in an artificial selection experiment by Flux and Flux (1982) and enabled robust conclusions about the evolutionary dynamics.

Immigrants into the experimental populations originate from distant populations or from the unavailable subpopulations on the study islands. These were pooled together with any few unselected residents as individuals in these two groups could not be distinguished. The focal populations are located distant to other known populations and house sparrows are generally highly sedentary (Anderson, 2006). Previous studies have found that only a small fraction of individuals disperse between populations separated by more than a few kilometres (Altwegg et al., 2000; Tufto et al., 2005; Anderson, 2006; Pärn et al., 2009, 2012). Immigrant house sparrows do not differ morphologically from residents (Altwegg et al., 2000), but immigrant males produce fewer recruits than resident males (Pärn et al., 2009, 2012). Hence, immigrants were likely to mostly originate from the unselected subpopulations, and to have morphological trait values that were randomly distributed around pre-experimental means (see Table S3; see also Holand et al., 2011). Any immigrants from distant populations should not compromise the conclusions on variation in individual fitness, but rather make the analyses more conservative as they might contribute to smooth out fitness differences between selected and unselected individuals.

Mean tarsus length responded to our artificial selection, with significant changes towards more extreme phenotypic and genetic values in both experimental populations (Fig. 2, Tables 3, 4 and 6). Individuals with one or both parents artificially selected (i.e. with tarsus length shifted from the population mean) were shown to produce fewer recruits than unselected individuals (Table 2). However, when combining recruit production and survival into a measure of individual fitness, the natural selection towards pre-experimental means was only significant for males in the *low* population (Fig. 1B). In the seven years after artificial selection, there was no significant natural selection toward pre-experimental means (Fig. 1C and D). This points to the fact that the detectability of a given strength of selection generally is strongly dependent on the magnitude of demographic stochasticity (Hersch and Phillips, 2004; Engen et al., 2012; Engen and Sæther, 2014; Haller and Hendry, 2014). Here, the demographic variance was found to be large during the years of artificial selection (Table S6), compared to previous estimates for house sparrows (Engen et al., 2007; Stubberud et al., 2017) and other small passerines (Sæther et al., 2004). This was probably an effect of translocating individuals, which reduced population size (N) and may have affected the social structure in the populations. Another effect of reducing N was necessarily a reduction of the population density during the breeding season in the two experimental populations. The demographic framework for estimating selection used in this study rest on the simplifying assumption of density-independent vital rates (Engen et al., 2012). A previous study, including the present study populations and other populations from the same area, found no effect of N on population growth (ΔN) during the present study period (Stubberud et al., 2017). Thus, the reduction of N should not have affected our results above the increased random variation in individual fitness among individuals (i.e. increased demographic variance).

When population mean phenotypes are stable over longer time periods, stabilizing selection is a likely explanation (Charlesworth et al., 1982; Estes and Arnold, 2007; Uyeda et al., 2011; Chevin and Haller, 2014; Haller and Hendry, 2014). Stabilizing selection maintains the mean phenotype of fitness related traits at intermediate values of high fitness (Lande, 1976, 1979; Arnold et al., 2001; Kinnison and Hendry, 2001; Sæther and Engen, 2015). In this study, individual fitness was reduced in both directions from the pre-experimental mean tarsus lengths, which suggests that tarsus length was moved away from an adaptive fitness peak (Table 2). However, an alternative explanation may be that tarsus length is constrained by genetic correlations with an unmeasured trait (Lande and Arnold, 1983; Hansen and Houle, 2004, 2008; Morrissey et al., 2010). Then, both traits could be kept from reaching their optimum in a balance of opposing directional selection. This explanation would require that the genetic correlation was so strong that the applied artificial selection also had a large effect on the unmeasured trait. While it is not possible to conclusively exclude an effect of such an unmeasured trait, at least none of the other traits in this study displayed significant changes in breeding values (see Table 6). Still, the expected ubiquitous effect of stabilizing selection is rarely detected in empirical studies of contemporary populations (Kingsolver et al., 2001, 2012). One reason is the low power to detect stabilizing selection in most studies with limited sample size (Kingsolver et al., 2001; Haller and Hendry, 2014), an issue that increases with increasing demographic stochasticity (Engen et al., 2012; Engen and Sæther, 2014; Haller and Hendry, 2014). Stabilizing selection might also be hard to detect due to low phenotypic variance around the peak, as less fit individuals continuously are removed, and the interference of ecological mechanisms, such as competition for resources (Rueffler et al., 2006; Haller and Hendry, 2014). Competition may lead to negative frequency-dependent selection, where intermediate phenotypes experience the largest reduction in fitness (Rueffler et al., 2006; Bolnick and Lau, 2008). Such mechanisms could lead to a flattening of the fitness peak which reduces the possibility for detecting stabilizing selection, or in extreme cases could cause disruptive selection (Haller and Hendry, 2014; Hendry, 2017). Frequency-dependent selection may often arise under parasitism, predation, sexual selection, sexual conflicts or asymmetric resource competition within species (Lande, 1980; Goldberg and Lande, 2006; Hendry, 2017). However, there were no indications of such mechanisms in the present study, where there was a clear reduction in fitness for selected individuals throughout the study period (see Table 2).

A fluctuating environment might constantly induce small random changes in the phenotypic fitness optimum, such that in any year or period of years selection might be directional (Arnold et al., 2001; Lande, 2007). During the period after artificial selection, the directional selection in males of the *high* population was in the same direction as during the artificial selection (Fig. 1D). Intuitively, one might think that artificially enlarged males were at an advantage relative to unselected smaller males. However, *post hoc* tests showed that the reduced recruit production of selected males relative to unselected males was not significantly different between the two periods in the *high* population (2003-2006 vs 2007-2012, *selection status* × *period*: $\chi^2 = 2.31$, df = 2, $P = 0.32$). In addition, the estimated environmental variance was quite large (see Table S6 and previous estimates in Sæther et al., 2004) and after maintaining long tarsus for 2-3 years, phenotypic and genetic values returned towards the pre-experimental means (Fig. 2). Hence, the unexpected positive selection on tarsus length in males was probably due to environmental fluctuations resulting in a brief period with selection for large body size.

The mating of unselected and artificially selected parents produced intermediate individuals with increased mean phenotypic values in the *high* population, but no change in the *low* population (see Table 4). This gene flow between the unselected and selected segment of each population decreased the overall response to artificial selection. Gene flow between wild populations under different selective regimes has repeatedly been suggested as a possible constraint on the phenotypic response in heritable traits (e.g. Slatkin, 1973; Storfer and Sih, 1998; Hendry et al., 2001; Postma and van Noordwijk, 2005; Postma et al., 2007; Rice and Papadopoulos, 2009; Siepielski et al., 2013; Hendry, 2017). Hence, the identification of spatially varying patterns of selection and evolutionary responses in wild unmanipulated populations depends on our ability to distinguish individuals of different origins. Failing to do so could be an important cause of mismatch between expected and observed phenotypic response to selection.

The experimental populations gradually returned towards their pre-experimental mean tarsus length and breeding values after the period of artificial selection ended (Fig. 2). The return in breeding values was faster than expected by genetic drift alone (Table 6). Three interacting mechanisms are believed to be involved in this process: natural selection counteracting the artificial selection, immigration, and recombination between selected and unselected individuals. Provided the recorded strength of natural selection, it would have taken a long time for natural selection alone to restore phenotypes in the populations. Hence, immigration and recombination between selected and unselected individuals were active drivers of changes in phenotypes and breeding values during the period after artificial selection. The expected proportion of the genome in a randomly chosen individual which was inherited from artificially selected ancestors decreased from 0.6-0.7 at the end of artificial selection to c. 0.25 at the end of the study period (Figure S2). Thus, the proportion of individuals that were unselected in each cohort increased towards the end of the study period and there were no selected individuals in the 2011 cohort (see Tables S9 and S10). The change in phenotype may be separated into a selection differential and a transmission term using the Price equation (Price, 1970, 1972; Frank, 2012; Engen et al., 2014; Queller, 2017). In these terms, the transmission term was a large component in the return towards pre-experimental means. Still, the effect of counteracting natural selection was important. Selected individuals produced 35-45 % less recruits than unselected individuals, thus the change in phenotype from the selection differential during artificial selection was reduced (see Table 2). These effects on the phenotypic change might be concealed in age-structured populations, where the final evolutionary response to selection is delayed until the individuals under selection have realized their lifetime reproduction (Hill, 1974; Engen et al., 2014).

Manipulating selection in the wild can yield novel insights into several aspects of evolutionary dynamics in populations under natural conditions. We have demonstrated that strong directional selection on heritable traits produce evolutionary responses in accordance with well-known quantitative genetic models. However, we also illustrate the potential for gene flow to impact the phenotypic trajectory of natural populations under temporal or spatial variation in selection. Perturbing the phenotype away from their natural mean had profound negative fitness consequences. Overall, the results provided indications of a phenotype maintained by selection for an intermediate value subject to environmental variation.

References

Altwegg, R., T. H. Ringsby, and B.-E. Sæther. 2000. Phenotypic correlates and consequences of dispersal in a metapopulation of house sparrows *Passer domesticus*. J. Anim. Ecol. 69:762–770.

Anderson, T. R. 2006. Biology of the ubiquitous house sparrow: From genes to populations. Oxford University Press, Oxford.

Arnold, S. J. 1983. Morphology, performance and fitness. Am. Zool. 23:347–361.

Arnold, S. J., M. E. Pfrender, and A. G. Jones. 2001. The adaptive landscape as a conceptual bridge between micro- and macroevolution. Genetica 112:9–32.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67:1–48.

Beldade, P., K. Koops, and P. M. Brakefield. 2002. Developmental constraints versus flexibility in morphological evolution. Nature 416:844–847.

Bell, G. 2008. Selection: the mechanism of evolution. Oxford biology. 2nd ed. Oxford University Press, Oxford.

———. 2010. Fluctuating selection: the perpetual renewal of adaptation in variable environments. Phil. Trans. R. Soc. B 365:87–97.

———. 2013. Evolutionary rescue and the limits of adaptation. Phil. Trans. R. Soc. B 368:20120080.

Bolnick, D. I. and O. L. Lau. 2008. Predictable patterns of disruptive selection in stickleback in postglacial lakes. Am. Nat. 172:1–11.

Bolstad, G. H., J. A. Cassara, E. Márquez, T. F. Hansen, K. van der Linde, D. Houle, and C. Pélabon. 2015. Complex constraints on allometry revealed by artificial selection on the wing of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 112:13284–13289.

Brakefield, P. M. 2003. Artificial selection and the development of ecologically relevant phenotypes. Ecology 84:1661–1671.

Brookfield, J. F. Y. 2016. Why are estimates of the strength and direction of natural selection from wild populations not congruent with observed rates of phenotypic change? Bioessays 38:927–934.

Calsbeek, R. and R. M. Cox. 2010. Experimentally assessing the relative importance of predation and competition as agents of selection. Nature 465:613–616.

Calsbeek, R. and T. B. Smith. 2007. Probing the adaptive landscape using experimental islands: Density-dependent natural selection on lizard body size. Evolution 61:1052–1061.

Caswell, H. 2001. Matrix population models: Construction, analysis, and interpretation. 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Charlesworth, B., R. Lande, and M. Slatkin. 1982. A neo-Darwinian commentary on macroevolution. Evolution 36:474–498.

Chevin, L. M. and B. C. Haller. 2014. The temporal distribution of directional gradients under selection for an optimum. Evolution 68:3381–3394.

Conner, J. K. 2003. Artificial selection: A powerful tool for ecologists. Ecology 84:1650–1660.

Coulson, T., T. G. Benton, P. Lundberg, S. R. X. Dall, B. E. Kendall, and J. M. Gaillard. 2006. Estimating individual contributions to population growth: evolutionary fitness in ecological time. Proc. R. Soc. B 273:547–555.

Coulson, T., L. E. B. Kruuk, G. Tavecchia, J. M. Pemberton, and T. H. Clutton-Brock. 2003. Estimating selection on neonatal traits in red deer using elasticity path analysis. Evolution 57:2879–2892.

Coulson, T. and S. Tuljapurkar. 2008. The dynamics of a quantitative trait in an age-structured population living in a variable environment. Am. Nat. 172:599–612.

Coulson, T., S. Tuljapurkar, and D. Z. Childs. 2010. Using evolutionary demography to link life history theory, quantitative genetics and population ecology. J. Anim. Ecol. 79:1226–1240.

Darimont, C. T., S. M. Carlson, M. T. Kinnison, P. C. Paquet, T. E. Reimchen, and C. C. Wilmers. 2009. Human predators outpace other agents of trait change in the wild. Proc. Natl. Acad. Sci. USA 106:952–954.

Endler, J. A. 1980. Natural selection on color patterns in *Poecilia reticulata*. Evolution 34:76–91.

———. 1986. Natural selection in the wild. Princeton University Press, Princeton, N.J.

Engen, S., T. Kvalnes, and B.-E. Sæther. 2014. Estimating phenotypic selection in age-structured populations by removing transient fluctuations. Evolution 68:2509–2523.

Engen, S., R. Lande, and B.-E. Sæther. 2011. Evolution of a plastic quantitative trait in an age-structured population in a fluctuating environment. Evolution 65:2893–2906.

Engen, S., R. Lande, B.-E. Sæther, and S. F. Dobson. 2009. Reproductive value and the stochastic demography of age-structured populations. Am. Nat. 174:795–804.

Engen, S., R. Lande, B.-E. Sæther, and P. Gienapp. 2010. Estimating the ratio of effective to actual size of an age-structured population from individual demographic data. J. Evol. Biol. 23:1148–1158.

Engen, S., T. H. Ringsby, B.-E. Sæther, R. Lande, H. Jensen, M. Lillegard, and H. Ellegren. 2007. Effective size of fluctuating populations with two sexes and overlapping generations. Evolution 61:1873–1885.

Engen, S. and B.-E. Sæther. 2014. Evolution in fluctuating environments: Decomposing selection into additive components of the Robertson-Price equation. Evolution 68:854–865.

Engen, S., B.-E. Sæther, T. Kvalnes, and H. Jensen. 2012. Estimating fluctuating selection in age-structured populations. J. Evol. Biol. 25:1487–1499.

Estes, S. and S. J. Arnold. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169:227–244.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Group, Harlow.

Flux, J. E. C. and M. M. Flux. 1982. Artificial selection and gene flow in wild starlings, *Sturnus vulgaris*. Naturwissenschaften 69:96–97.

Frank, S. A. 2012. Natural selection. IV. The Price equation. J. Evol. Biol. 25:1002–1019.

Fuller, R. C., C. F. Baer, and J. Travis. 2005. How and when selection experiments might actually be useful. Integr. Comp. Biol. 45:391–404.

Goldberg, E. E. and R. Lande. 2006. Ecological and reproductive character displacement on an environmental gradient. Evolution 60:1344–1357.

Grant, P. R. and B. R. Grant. 1995. Predicting microevolutionary responses to directional selection on heritable variation. Evolution 49:241–251.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw. 33:1–22.

Hadfield, J. D., A. J. Wilson, D. Garant, B. C. Sheldon, and L. E. B. Kruuk. 2010. The misuse of BLUP in ecology and evolution. Am. Nat. 175:116–125.

Hagen, I. J., A. M. Billing, B. Rønning, S. A. Pedersen, H. Pärn, J. Slate, and H. Jensen. 2013. The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10K SNP-chip for the house sparrow (*Passer domesticus*). Mol. Ecol. Resour. 13:429–439.

Haller, B. C. and A. P. Hendry. 2014. Solving the paradox of stasis: Squashed stabilizing selection and the limits of detection. Evolution 68:483–500.

Hansen, T. F. and D. Houle. 2004. Evolvability, stabilizing selection, and the problem of stasis. Pp. 130–154. *in* M. Pigliucci and K. Preston, eds. Phenotypic integration:

Studying the ecology and evolution of complex phenotypes. Oxford University Press, New York.

———. 2008. Measuring and comparing evolvability and constraint in multivariate characters. J. Evol. Biol. 21:1201–1219.

Henderson, N. D. 1989. Interpreting studies that compare high- and low-selected lines on new characters. Behav. Genet. 19:473–502.

———. 1997. Spurious associations in unreplicated selected lines. Behav. Genet. 27:145–154.

Hendry, A. P. 2017. Eco-evolutionary dynamics. Princeton University Press, New Jersey.

Hendry, A. P., T. Day, and E. B. Taylor. 2001. Population mixing and the adaptive divergence of quantitative traits in discrete populations: A theoretical framework for empirical tests. Evolution 55:459–466.

Hendry, A. P. and M. T. Kinnison. 1999. The pace of modern life: Measuring rates of contemporary microevolution. Evolution 53:1637–1653.

Hereford, J., T. F. Hansen, and D. Houle. 2004. Comparing strengths of directional selection: How strong is strong? Evolution 58:2133–2143.

Hersch, E. I. and P. C. Phillips. 2004. Power and potential bias in field studies of natural selection. Evolution 58:479–485.

Hill, W. G. 1974. Prediction and evaluation of response to selection with overlapping generations. Anim. Prod. 18:117–139.

Hill, W. G. and A. Caballero. 1992. Artificial selection experiments. Annu. Rev. Ecol. Syst. 23:287–310.

Holand, A. M., H. Jensen, J. Tufto, and R. Moe. 2011. Does selection or genetic drift explain geographic differentiation of morphological characters in house sparrows *Passer domesticus*? Genet. Res. 93:367–379.

Huxley, J. S. 1932. Problems of relative growth. Methuen, London.

Jensen, H., I. Steinsland, T. H. Ringsby, and B.-E. Sæther. 2008. Evolutionary dynamics of a sexual ornament in the house sparrow (*Passer domesticus*): The role of indirect selection within and between sexes. Evolution 62:1275–1293.

Jensen, H., B.-E. Sæther, T. H. Ringsby, J. Tufto, S. C. Griffith, and H. Ellegren. 2003. Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (*Passer domesticus*). J. Evol. Biol. 16:1296–1307.

———. 2004. Lifetime reproductive success in relation to morphology in the house sparrow *Passer domesticus*. J. Anim. Ecol. 73:599–611.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall. 2007. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16:1099–1106.

Kingsolver, J. G., S. E. Diamond, A. M. Siepielski, and S. M. Carlson. 2012. Synthetic analyses of phenotypic selection in natural populations: lessons, limitations and future directions. Evol. Ecol. 26:1101–1118.

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gibert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Kinnison, M. T. and A. P. Hendry. 2001. The pace of modern life II: from rates of contemporary microevolution to pattern and process. Genetica 112:145–164.

Konarzewski, M., A. Ksiazek, and I. B. Lapo. 2005. Artificial selection on metabolic rates and related traits in rodents. Integr. Comp. Biol. 45:416–425.

Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the 'animal model'. Phil. Trans. R. Soc. B 359:873–890.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

———. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. Evolution 33:402–416.

———. 1980. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. Evolution 34:292–305.

———. 1982. A quantitative genetic theory of life history evolution. Ecology 63:607–615.

———. 2007. Expected relative fitness and the adaptive topography of fluctuating selection. Evolution 61:1835–1846.

Lande, R. and S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37:1210–1226.

Lande, R., S. Engen, and B.-E. Sæther. 2003. Stochastic population dynamics in ecology and conservation. Oxford University Press, Oxford.

Lendvai, G. and D. A. Levin. 2003. Rapid response to artificial selection on flower size in *Phlox*. Heredity 90:336–342.

Losos, J. B., T. W. Schoener, K. I. Warheit, and D. Creer. 2001. Experimental studies of adaptive differentiation in bahamian *Anolis* lizards. Genetica 112:399–415.

Losos, J. B., K. I. Warheit, and T. W. Schoener. 1997. Adaptive differentiation following experimental island colonization in *Anolis* lizards. Nature 387:70–73.

Lynch, M. and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, Mass.

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7:639–655.

Merilä, J. and A. P. Hendry. 2014. Climate change, adaptation, and phenotypic plasticity: the problem and the evidence. Evol. Appl. 7:1–14.

Merilä, J., B. C. Sheldon, and L. E. B. Kruuk. 2001. Explaining stasis: microevolutionary studies in natural populations. Genetica 112:199–222.

Metcalf, C. J. E. and S. Pavard. 2007. Why evolutionary biologists should be demographers. Trends Ecol. Evol. 22:205–212.

Morrissey, M. B. 2016. Meta-analysis of magnitudes, differences and variation in evolutionary parameters. J. Evol. Biol. 29:1882–1904.

Morrissey, M. B., L. E. B. Kruuk, and A. J. Wilson. 2010. The danger of applying the breeder's equation in observational studies of natural populations. J. Evol. Biol. 23:2277–2288.

Morrissey, M. B., C. A. Walling, A. J. Wilson, J. M. Pemberton, T. H. Clutton-Brock, and L. E. B. Kruuk. 2012. Genetic analysis of life-history constraint and evolution in a wild ungulate population. Am. Nat. 179:E97–E114.

Pigeon, G., M. Festa-Bianchet, D. W. Coltman, and F. Pelletier. 2016. Intense selective hunting leads to artificial evolution in horn size. Evol. Appl. 9:521–530.

Postma, E. 2006. Implications of the difference between true and predicted breeding values for the study of natural selection and micro-evolution. J. Evol. Biol. 19:309–320.

Postma, E. and A. J. van Noordwijk. 2005. Gene flow maintains a large genetic difference in clutch size at a small spatial scale. Nature 433:65–68.

Postma, E., J. Visser, and A. J. Van Noordwijk. 2007. Strong artificial selection in the wild results in predicted small evolutionary change. J. Evol. Biol. 20:1823–1832.

Price, G. R. 1970. Selection and covariance. Nature 227:520–521.

———. 1972. Extension of covariance selection mathematics. Ann. Hum. Genet. 35:485–490.

Price, T., M. Kirkpatrick, and S. J. Arnold. 1988. Directional selection and the evolution of breeding date in birds. Science 240:798–799.

Pärn, H., H. Jensen, T. H. Ringsby, and B.-E. Sæther. 2009. Sex-specific fitness correlates of dispersal in a house sparrow metapopulation. J. Anim. Ecol. 78:1216–1225.

Pärn, H., T. H. Ringsby, H. Jensen, and B.-E. Sæther. 2012. Spatial heterogeneity in the effects of climate and density-dependence on dispersal in a house sparrow metapopulation. Proc. R. Soc. B 279:144–152.

Queller, D. C. 2017. Fundamental theorems of evolution. Am. Nat. 189:345–353.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL `www.R-project.org`.

Reid, J. M., E. M. Bignal, S. Bignal, D. I. McCracken, and P. Monaghan. 2003. Age-specific reproductive performance in red-billed choughs *Pyrrhocorax pyrrhocorax*: patterns and processes in a natural population. J. Anim. Ecol. 72:765–776.

Reznick, D. N. and C. K. Ghalambor. 2001. The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. Genetica 112:183–198.

———. 2005. Selection in nature: Experimental manipulations of natural populations. Integr. Comp. Biol. 45:456–462.

Reznick, D. N., F. H. Shaw, F. H. Rodd, and R. G. Shaw. 1997. Evaluation of the rate of evolution in natural populations of guppies (*Poecilia reticulata*). Science 275:1934–1937.

Rice, S. H. 2008. A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution. BMC Evol. Biol. 8.

Rice, S. H. and A. Papadopoulos. 2009. Evolution with stochastic fitness and stochastic migration. PloS ONE 4:e7130.

Ringsby, T. H., B.-E. Sæther, and E. J. Solberg. 1998. Factors affecting juvenile survival in house sparrow *Passer domesticus*. J. Avian Biol. 29:241–247.

Rising, J. D. and K. M. Somers. 1989. The measurement of overall body size in birds. Auk 106:666–674.

Rueffler, C., T. J. M. Van Dooren, O. Leimar, and P. A. Abrams. 2006. Disruptive selection and then what? Trends Ecol. Evol. 21:238–245.

Rønning, B., J. Broggi, C. Bech, B. Moe, T. H. Ringsby, H. Pärn, I. J. Hagen, B.-E. Sæther, and H. Jensen. 2016. Is basal metabolic rate associated with recruit production and survival in free-living house sparrows? Func. Ecol. 30:1140–1148.

Schielzeth, H. and W. Forstmeier. 2009. Conclusions beyond support: overconfident estimates in mixed models. Behav. Ecol. 20:416–420.

Schulte-Hostedde, A. I., B. Zinner, J. S. Millar, and G. J. Hickling. 2005. Restitution of mass-size residuals: Validating body condition indices. Ecology 86:155–163.

Senar, J. C. and J. Pascual. 1997. Keel and tarsus length may provide a good predictor of avian body size. Ardea 85:269–274.

Siepielski, A. M., K. M. Gotanda, M. B. Morrissey, S. E. Diamond, J. D. DiBattista, and S. M. Carlson. 2013. The spatial patterns of directional phenotypic selection. Ecol. Lett. 16:1382–1392.

Simpson, G. 1944. Tempo and mode in evolution. Columbia University Press, New York.

Skjelseth, S., T. H. Ringsby, J. Tufto, H. Jensen, and B.-E. Sæther. 2007. Dispersal of introduced house sparrows *Passer domesticus*: an experiment. Proc. R. Soc. B 274:1763–1771.

Slatkin, M. 1973. Gene flow and selection in a cline. Genetics 75:733–756.

Storfer, A. and A. Sih. 1998. Gene flow and ineffective antipredator behavior in a stream-breeding salamander. Evolution 52:558–565.

Stubberud, M. W., A. M. Myhre, H. Holand, T. Kvalnes, T. H. Ringsby, B.-E. Sæther, and H. Jensen. 2017. Sensitivity analysis of effective population size to demographic parameters in house sparrow populations. Mol. Ecol. 26:2449–2465.

Swallow, J., J. Hayes, K. Pawel, and T. J. Garland. 2009. Selection experiments and experimental evolution of performance and physiology. *in* T. J. Garland and M. Rose, eds. Experimental evolution: Concepts, methods and applications of selection experiments. University of California Press, Berkeley and Los Angeles, California.

Sæther, B.-E. and S. Engen. 2015. The concept of fitness in fluctuating environments. Trends Ecol. Evol. 30:273–281.

Sæther, B.-E., S. Engen, A. P. Møller, H. Weimerskirch, M. E. Visser, W. Fiedler, E. Matthysen, M. M. Lambrechts, A. Badyaev, P. H. Becker, J. E. Brommer, D. Bukacinski, M. Bukacinska, H. Christensen, J. Dickinson, C. du Feu, F. R. Gehlbach, D. Heg, H. Hotker, J. Merilä, J. T. Nielsen, W. Rendell, R. J. Robertson, D. L. Thomson, J. Torok, and P. Van Hecke. 2004. Life-history variation predicts the effects of demographic stochasticity on avian population dynamics. Am. Nat. 164:793–802.

Teuschl, Y., C. Reim, and W. U. Blanckenhorn. 2007. Correlated responses to artificial body size selection in growth, development, phenotypic plasticity and juvenile viability in yellow dung flies. J. Evol. Biol. 20:87–103.

Tigreros, N. and S. M. Lewis. 2011. Direct and correlated responses to artificial selection on sexual size dimorphism in the flour beetle, *Tribolium castaneum*. J. Evol. Biol. 24:835–842.

Tufto, J., T. H. Ringsby, A. A. Dhondt, F. Adriaensen, and E. Matthysen. 2005. A parametric model for estimation of dispersal patterns applied to five passerine spatially structured populations. Am. Nat. 165:E13–E26.

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proc. Natl. Acad. Sci. USA 108:15908–15913.

Wade, M. J. and S. Kalisz. 1990. The causes of natural selection. Evolution 44:1947–1955.

Wilson, A. J. and D. H. Nussey. 2010. What is individual quality? An evolutionary perspective. Trends Ecol. Evol. 25:207–214.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc. VI Inter. Congr. Genet. 1:356–366.

947 # Tables

Table 1: Individual selection category based on whether both, one or none of the parents had been artificially selected in two house sparrow populations in Norway. The populations were subject to artificial selection for long or short tarsus.

Selection category	Description
Selected	Both parents artificially selected
Unselected	No parent artificially selected
Intermediate	One parent artificially selected
Other	All other individuals

Table 2: Parameter estimates and 95% confidence intervals for models explaning the production of recruits over the years 2003-2012 in two house sparrow populations in Norway. The populations were subjected to artificial selection for long (*high*) or short (*low*) tarsus. The selection categories were unselected, intermediate and selected (see Table 1). Estimates are given relative to unselected females of age 1 (Intercept). Generalized mixed effects models were fitted with a Poisson error structure and a log link function. Models were fitted with a random intercept for individual identity.

		Confidence interval	
	Estimate	Lower	Upper
(A) *High*			
Intercept	-0.23	-0.46	0.00
Selection category			
Selected	-0.43	-0.69	-0.16
Intermediate	-0.16	-0.43	0.10
Age 2	0.28	0.09	0.47
Male	-0.10	-0.32	0.12
(B) *Low*			
Intercept	-0.05	-0.31	0.20
Selection category			
Selected	-0.58	-0.96	-0.21
Intermediate	-0.35	-0.70	-0.01
Age 2	-0.00	-0.28	0.27
Male	0.06	-0.17	0.30
Sel.status × age			
Selected × age 2	0.77	0.30	1.24
Intermediate × age 2	0.10	-0.40	0.59

Table 3: Annual phenotypic change (slope±SE) in weighted means in three house sparrow populations in Norway. Two of the populations were subjected to artificial selection for longer (*high*) or shorter (*low*) tarsus in the years 2002-2005. In the period 2006-2012 the populations were monitored with no further manipulations. Permutation tests with 10 000 iterations were used to assess the significance of the estimated annual changes. Annual changes were estimated using linear regression, accounting for mean differences between sexes in phenotypes. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

	Years	
	2002-2006	2006-2012
(A) *High*		
Tarsus length	0.126±0.021***	-0.088±0.013***
Wing length	0.211±0.044***	-0.003±0.023
Body condition	0.009±0.001***	-0.003±0.001***
Bill length	0.042±0.015**	-0.039±0.009***
Bill depth	0.015±0.007*	-0.007±0.004
(B) *Low*		
Tarsus length	-0.112±0.020***	0.027±0.013
Wing length	-0.027±0.048	0.129±0.028***
Body condition	0.016±0.002***	-0.003±0.001***
Bill length	0.017±0.014	-0.009±0.008
Bill depth	-0.025±0.006***	0.008±0.005
(C) *Control*		
Tarsus length	-0.027±0.028	-0.035±0.016*
Wing length	-0.085±0.063	0.029±0.031
Body condition	0.005±0.002	-0.002±0.001
Bill length	0.016±0.022	-0.004±0.010
Bill depth	-0.010±0.010	-0.004±0.005

Table 4: Cohort phenotypic change (slope ± SE) in arithmetic mean over the cohorts 2000-2005, for each of three selection categories in two house sparrow populations in Norway. Artificial selection on tarsus length was performed for longer (*high*) or shorter (*low*) tarsus on the pre-breeding populations in the years 2002-2005. Any sexual dimorphism in the traits was accounted for by including sex in the models. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

	Selection category		
	Selected	Intermediate	Unselected
(A) *High*			
Tarsus length	0.167±0.040***	0.127±0.047*	0.008±0.079
Wing length	0.178±0.074*	0.170±0.070	0.271±0.128
Body condition	0.012±0.002***	0.004±0.003	0.008±0.003
Bill length	-0.008±0.028	0.063±0.039*	0.061±0.034
Bill depth	0.020±0.012	0.002±0.014	0.029±0.018
(B) *Low*			
Tarsus length	-0.091±0.041*	0.057±0.042	0.002±0.044
Wing length	-0.068±0.070	-0.049±0.075	0.093±0.107
Body condition	0.018±0.004***	0.016±0.004***	0.014±0.003***
Bill length	0.056±0.024*	0.048±0.019	0.055±0.027
Bill depth	-0.023±0.015	-0.010±0.013	-0.024±0.017

Table 5: The $\mathbf{G}_{\boldsymbol{\sigma}}$-matrix for three house sparrow populations (*high*, *low* and *control*) in the years 2002-2012 in Norway. The final cohort included in the analyses was 2011. Two of the populations were subjected to artificial selection for longer (*high*) or shorter (*low*) tarsus in the years 2002-2005. Posterior modes with 95% credibility intervals are given. All traits were SD-standardised prior to analyses, such that the matrices have heritabilities on the diagonal and genetic correlations in the off-diagonal elements.

	Tarsus length	Wing length	Body condition	Bill length	Bill depth
(A) *High*					
Tarsus length	0.396 (0.281,0.542)	0.144 (0.068,0.248)	0.064 (-0.035,0.146)	0.097 (-0.002,0.188)	0.156 (0.072,0.277)
Wing length		0.315 (0.225,0.447)	0.097 (0.031,0.190)	0.062 (-0.010,0.160)	0.106 (-0.008,0.189)
Body condition			0.408 (0.286,0.544)	0.125 (0.017,0.204)	0.115 (0.018,0.227)
Bill length				0.625 (0.469,0.734)	0.215 (0.112,0.321)
Bill depth					0.442 (0.314,0.626)
(B) *Low*					
Tarsus length	0.313 (0.229,0.436)	0.137 (0.080,0.219)	0.013 (-0.057,0.103)	0.120 (0.059,0.230)	0.113 (0.027,0.192)
Wing length		0.333 (0.243,0.412)	0.061 (-0.006,0.148)	0.075 (0.020,0.163)	0.078 (0.004,0.154)
Body condition			0.402 (0.272,0.524)	0.057 (-0.021,0.155)	0.170 (0.071,0.260)
Bill length				0.391 (0.246,0.506)	0.145 (0.049,0.233)
Bill depth					0.418 (0.310,0.578)
(C) *Control*					
Tarsus length	0.416 (0.260,0.625)	0.076 (-0.025,0.181)	-0.041 (-0.142,0.082)	0.094 (-0.034,0.234)	-0.059 (-0.163,0.089)
Wing length		0.289 (0.191,0.409)	0.033 (-0.039,0.141)	0.060 (-0.033,0.180)	0.037 (-0.047,0.150)
Body condition			0.154 (0.012,0.326)	-0.003 (-0.152,0.095)	0.086 (-0.015,0.215)
Bill length				0.458 (0.251,0.674)	0.097 (-0.054,0.227)
Bill depth					0.409 (0.225,0.626)

Table 6: Annual change in the weighted mean estimated breeding values for three house sparrow populations in Norway. Two of the populations were subjected to artificial selection for longer (*high*) or shorter (*low*) tarsus in the years 2002-2005. Stars indicates if the estimated changes are larger than expected by genetic drift alone. Posterior modes with 95% credibility intervals are given. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

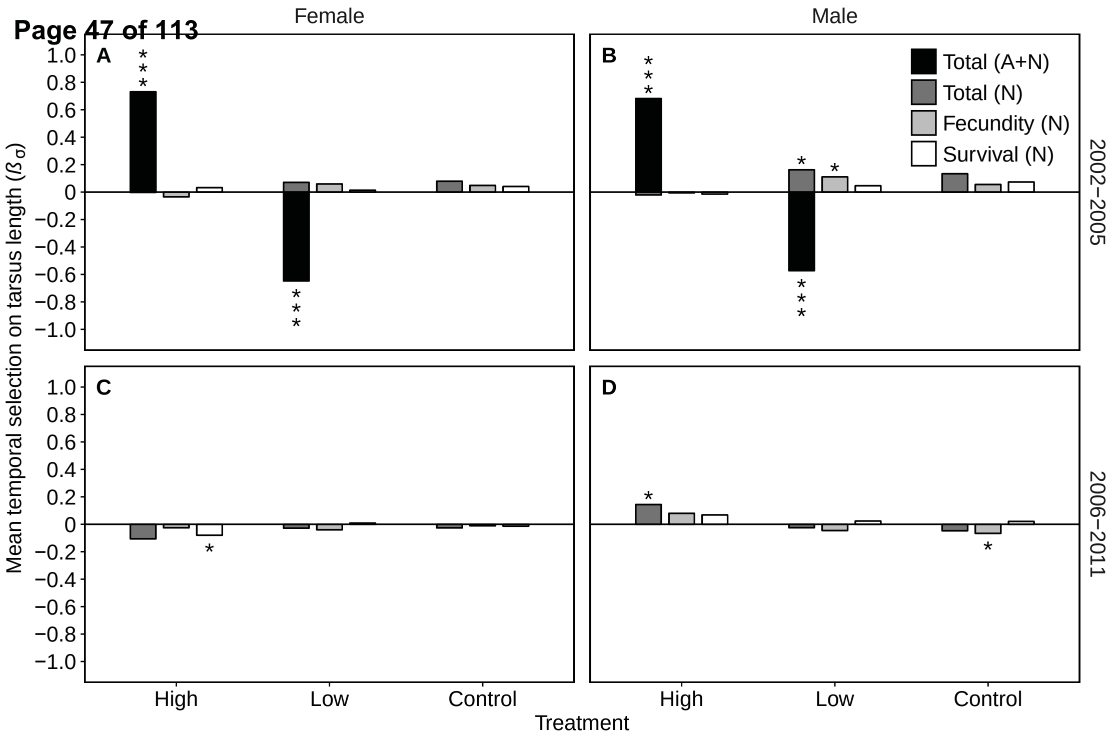
	Years	
	2002-2006	2006-2012
(A) *High*		
Tarsus length	0.110 (0.072, 0.152)***	-0.055 (-0.081, -0.026)*
Wing length	0.027 (-0.005, 0.064)	0.000 (-0.021, 0.026)
Body condition	0.050 (0.003, 0.083)	-0.008 (-0.037, 0.018)
Bill length	0.028 (-0.012, 0.070)	-0.013 (-0.043, 0.017)
Bill depth	0.035 (-0.002, 0.078)	-0.014 (-0.041, 0.012)
(B) *Low*		
Tarsus length	-0.103 (-0.137, -0.059)***	0.044 (0.021, 0.067)*
Wing length	-0.025 (-0.059, 0.005)	0.034 (0.016, 0.057)
Body condition	0.038 (0.000, 0.080)	-0.024 (-0.048, 0.001)
Bill length	-0.019 (-0.059, 0.019)	0.011 (-0.015, 0.034)
Bill depth	-0.019 (-0.063, 0.017)	0.003 (-0.023, 0.027)
(C) *Control*		
Tarsus length	0.003 (-0.051, 0.036)	-0.019 (-0.043, 0.019)
Wing length	-0.002 (-0.039, 0.033)	0.021 (0.001, 0.044)
Body condition	0.005 (-0.022, 0.044)	0.001 (-0.022, 0.019)
Bill length	0.011 (-0.030, 0.055)	0.006 (-0.020, 0.035)
Bill depth	-0.001 (-0.041, 0.045)	0.016 (-0.014, 0.044)

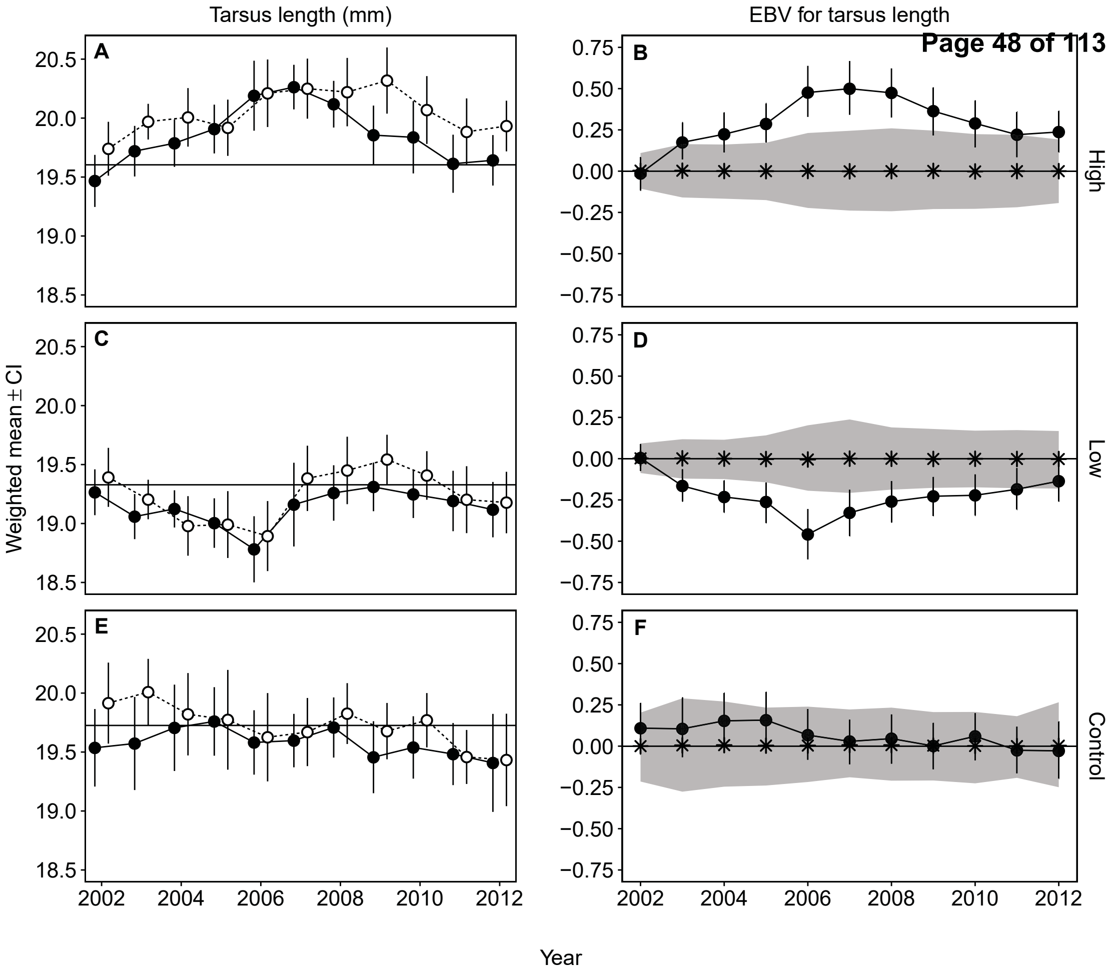
Figure legends

Figure 1: Temporal mean SD-scaled directional selection gradients (β_σ) over the periods 2002-2005 and 2006-2011 for female and male house sparrows in each of three populations (*high*, *low* and *control*) in Norway. In the first period (**A**, **B**), two of the populations were subjected to artificial selection for long (*high*) or short (*low*) tarsus. In the subsequent period (**C**, **D**), all three populations were monitored with no artificial manipulations of the distribution of phenotypes. Selection was estimated including both artificial (A) and natural (N) selection, and natural selection was further decomposed into viability and fecundity selection. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Figure 2: Annual estimates (weighted mean ± 95% CI) of phenotypic (**A**, **C**, **E**) and genetic (**B**, **D**, **F**, estimated breeding value [EBV]) tarsus length (mm) in three house sparrow populations (*high* [**A**, **B**], *low* [**C**, **D**] and *control* [**E**, **F**]) in Norway. The *high* and *low* populations were subjected to artificial selection for longer (*high*) or shorter (*low*) tarsus before the breeding seasons in the years 2002-2005. Males (open circles, dashed lines) and females (solid circles and lines) were analysed together in the animal models, including sex as a fixed effect. EBV is are shown with solid circles and lines, while the stars and shaded areas are the expected EBV with 95% credibility intervals simulated under genetic drift alone. The horizontal lines in the left panels (**A, C, E**) are the mean tarsus length for each population across sexes in 2002.

Figure 3: Predicted and observed response to selection in two house sparrow populations in Norway. The populations were subjected to artificial selection for long (*high*) and short (*low*) tarsus in the years 2002-2005. The annual responses are averaged across sexes as sample sizes did not allow sex-specific **G**-matrices. During the period 2006-2011, populations were monitored without additional manipulations. (**A**) The complete annual response, which includes both survival of adults and recruitment of new individuals from known parents. (**B**) The partial annual response, includes only recruitment of new individuals from known parents. Unselected and intermediate individuals were excluded to estimate the observed responses.


Female
Male
A
B
C
D
Total (A+N)
Total (N)
Fecundity (N)
Survival (N)
2002–2005
2006–2011
Mean temporal selection on tarsus length (β_σ)
1.0
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1.0
***
*
High
Low
Control
Treatment


Tarsus length (mm)
EBV for tarsus length
A
B
C
D
E
F
High
Low
Control
Weighted mean ± CI
Year
20.5
20.0
19.5
19.0
18.5
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
2002
2004
2006
2008
2010
2012