Quantifying evolution in wild populations

Jip J.C. Ramakers

Propositions

1. Answering complex quantitative genetic questions is rarely possible in long-term studies of wild populations as these studies are simply not designed for it.
   (this thesis)

2. Field experiments can only allow us to draw conclusions about the natural world when carried out over a sufficient length of time to cover the whole range of environmental conditions.
   (this thesis)

3. Providing p-values to substantiate the importance of a scientific claim is nonsense if not accompanied by effect sizes and imprecision estimates.

4. The higher the impact factor of a journal, the more suspicious one should be of its content.

5. Science education should convey that scientific philosophy is incompatible with religious beliefs, but should do so without becoming dogmatic itself.

6. It is morally inconsistent to strongly support anti-abortion movements while not equally supporting fights for the lives of innocent animals in industry.

Propositions belonging to the thesis entitled:

"Predicting evolution in wild populations"

Jip J.C. Ramakers

Wageningen, 8th May 2019

Quantifying evolution in wild populations

Jip J.C. Ramakers

Thesis committee

Promotor

Prof. Dr Marcel E. Visser

Special professor Ecological Genetics

Wageningen University & Research /

Head of department, Department of Animal Ecology

Netherlands Institute of Ecology (NIOO-KNAW), Wageningen

Co-promotor

Dr Phillip Gienapp

Postdoctoral researcher, Department of Animal Ecology

Netherlands Institute of Ecology (NIOO-KNAW), Wageningen

Other members

Dr Martine E. Maan, University of Groningen

Prof. Dr Bas J. Zwaan, Wageningen University & Research

Dr Wouter Halfwerk, VU Amsterdam

Dr Erik Postma, University of Exeter, UK

This research was conducted under the auspices of the Wageningen Institute of Animal Sciences (WIAS) graduate school.

Quantifying evolution in wild populations

Jip Jacques Claudia Ramakers

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus

Prof. Dr A.J.P. Mol,

in the presence of the

Thesis committee appointed by the Academic Board,

to be defended in public

on Wednesday 8 May 2019

at 4 p.m. in the Aula

Jip J.C. Ramakers
Quantifying evolution in wild populations,
278 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summary in English and in Dutch

ISBN: 978-94-6343-920-6
DOI: https://doi.org/10.18174/472942

Table of contents

Chapter 1

General introduction

"*Adapt or perish, now as ever, is Nature's inexorable imperative*" (H.G. Wells in *Mind at the End of its Tether*, 1945). Although Wells' counsel particularly pertained to changes in the sexual standards and dogmas in modern human societies, it could easily apply to all life on our planet. Life evolves (Darwin 1859) and has done so since its origin. In its most simplistic definition, adaptive evolution is the outcome of natural or sexual selection on (genetically) heritable traits (that is, leaving aside mutation, dispersal or random genetic drift). Already Darwin realised that most scope for macro-evolutionary processes is present during episodes of strong environmental change. As an example, during the Great Oxygenation event approximately 2.3 billion years ago, aerobic respiration evolved in clades of unicellular organisms in response to the massive rise in oxygen levels by cyanobacteria during photosynthesis (Soo et al. 2017). Mammalian life forms radiated explosively during Early to Middle Jurassic and at the end of the Mesozoic Era as novel ecological niches previously vacated by dinosaurs became available (Archibald 2011; Close et al. 2015). Such examples of 'macroevolution', i.e. above the level of the species, typically occur at 'evolutionary timescales' (i.e. spanning many thousands to millions of years) and can only be studied retrospectively using fossil records or phylogenetic 'trees of life' (Pace 2009). Importantly, evolutionary forces also operate at 'ecological timescales' (decades to centuries)—and can be studied 'live in action' (although the distinction between the two timescales is not necessarily clear; Hendry and Kinnison 1999). A well-known example of rapid evolution in the wild is the morphological changes in a translocated population of a wall lizard (*Podarcis sicula*); within 36 years, the introduced islet population showed distinct head and digestive system morphology and a concordant dietary shift compared to the source population on a nearby islet (Herrel et al. 2008). Even faster evolutionary change can be studied in laboratory conditions in small (often unicellular) organisms with generation times as short as a couple of hours (e.g. Lenski et al. 1991). The study of evolutionary processes (adaptation, selection and (changes in) genetic variation) has become a central topic in evolutionary ecology in the context of global environmental change (Parmesan 2006; Reusch and Wood 2007; Merilä and Hendry 2014).

Studying evolution in the wild: concepts and tools

Variation is everywhere

In evolutionary ecology, it is of particular interest to understand how organisms interact with their environment and how this environment can shape evolutionary processes (i.e. are drivers of selection)—and vice versa (eco-evolutionary dynamics; Hendry 2017). Understanding these eco-evolutionary dynamics is important to infer rates of (micro)evolutionary change, which are necessary, for instance, to make predictions about the consequences of human disturbance on population viability and persistence (Hendry and Kinnison 1999). Key to the study of evolution is variation in phenotypes and how this mediated by genetic effects (Lynch and Walsh 1998). Genetic variation underlying

Box 1.1. *Quantitative genetics, variance components and selection*

In its most simplistic form, the phenotype (*z*) of a given individual (*i*) is composed of a genetic and environmental component:

$$z_i = g_i + e_i.$$

Similarly, at the population level, phenotypic variation can be partitioned into a genetic and non-genetic component:

$$V_z = V_G + V_E$$

(Lynch and Walsh 1998). In words, this means that phenotypic variation at the population level is the sum of genetic variation and 'environmental' variation. V_G is in fact a composite of additive genetic variance (V_A: variance attributable to breeding values, i.e. the additive effect of independent loci), dominance variance (V_D: variance attributable to deviations from the breeding value due to within-locus allelic interactions) and epistatic variance (V_I: variance attributable to epistatic interactions between loci). When studying the short-term evolutionary potential of the population, we are mainly concerned with the additive effect of different alleles (V_A). The environmental variance, V_E, in turn is a composite of general environmental variance ($V_{E,g}$: among-individual variance) and specific environmental variance ($V_{E,S}$: within-individual variance). The former, $V_{E,g}$, is often denoted as 'permanent environment' variance (V_{PE}) and represents the variation between individuals that cannot be attributed to (additive) genetic effects but rather (unmeasured) environmental effects that are constant across repeated measures of the individual; the latter, $V_{E,S}$, is often denoted as residual variance (V_R). In most practical situations, then, phenotypic variance is partitioned as

$$V_z = V_A + V_{PE} + V_R$$

(note that V_D and V_I are usually ignored for simplicity). This is because for short-term evolutionary change, we are mainly concerned with the additive effect of genes as the hereditary units of transgenerational transmission. The amount of genetic variation relevant for selection is then expressed as the proportion of additive genetic information relative to total phenotypic variation:

$$h^2 = V_A/V_z,$$

where h^2 is the heritability of the trait. The heritability is used in animal and plant breeding to predict the evolutionary response to selection, using the 'breeder's equation':

$$R = h^2 s,$$

where *R* is the change in the mean trait value from one generation to the next and *s* is the selection differential, i.e. the covariance between relative fitness and the trait value, also expressed as the difference in the mean phenotype between the parental and offspring generation (Falconer and Mackay 1996; Lynch and Walsh 1998). In its multivariate form (Lande 1979; Lande and Arnold 1983), i.e. when estimating the expected response in two (correlated traits), this can be rewritten as

$$\Delta z = \boldsymbol{G}\beta,$$

continued

***Box 1.1** (continued)*

… where Δz is a vector of responses in the traits, ***G*** is the genetic variance-covariance matrix of the traits, and β is the selection gradient, i.e. the partial regression coefficient of fitness on a trait.
Variance can be partitioned statistically by measuring phenotypes in carefully designed full-sib/half-sib experiments (Lynch and Walsh 1998). Alternatively, when pedigree information is available, variance can be partitioned using so-called 'animal models' (Henderson 1988; Kruuk 2004), i.e. mixed-effects models that allow for the inclusion of relatedness matrices (pedigrees). See Lynch and Walsh (1998) and Falconer and Mackay (1996) for a complete account for quantitative genetic approaches in animal and plant breeding. Importantly, the application of the breeder's equation to predict evolution in natural populations has been discommended on grounds of uncertainty of having fitness effects of correlated (unmeasured) traits appropriately accounted for (Morrissey et al. 2010).

phenotypic variation has been found to be omnipresent in nature across life history, behaviour, morphology, and physiology (Postma 2014).

To understand how evolutionary dynamics operate at the population level, we first need a thorough understanding of phenotypic variation among as well as within individuals (or genotypes). The field of quantitative genetics, which is specifically concerned with understanding the genetics of quantitative, polygenic traits, has long been used in animal and plant breeding to understand the consequences of selection on not only mean, but also variation in quantitative trait values (Falconer and Mackay 1996; Lynch and Walsh 1998). It is therefore of particular use in studying selection responses in breeding programmes as it allows population-level phenotypic variation to be partitioned into heritable and non-heritable components (Box 1.1; Falconer and Mackay 1996; Lynch and Walsh 1998). With the increased number of long-term, wild population studies of individually marked animals, the field of quantitative genetics is increasingly applied in studies of evolution in wild populations (Charmantier et al. 2014).

Aside from genetic and environmental differences *among* individuals, evolutionary and behavioural ecologists alike are increasingly aware of and interested in variation *within* individuals (Piersma and Drent 2003; Dingemanse et al. 2010; Westneat et al. 2015). For example, animals are believed to exhibit 'behavioural syndromes' or personality, expressing similar behaviour across time and contexts (Réale et al. 2007); different individuals have different personalities, and behaviour thus varies between individuals. However, both behavioural and life-history traits are known to be often phenotypically plastic (Scheiner 1993; Schlichting and Pigliucci 1998; Pigliucci 2001), meaning that phenotypes respond to fluctuating environmental conditions; the same individual thus expresses (within-individual) variation in its phenotypes (Box 1.2).

The importance of understanding environmentally induced variation in phenotypes

When we study ecological and evolutionary processes in animals in the wild, we need to understand why and how the environment shapes variation in phenotypes. The environment can shape the phenotype of a particular individual in a reversible way in

(labile) traits that are phenotypically plastic with respect to that environment (Box 1.2). Phenotypic plasticity is generally thought of as an adaptive mechanism to cope with varying phenotypic optima in fluctuating environments (Scheiner 1993). In avian timing of breeding, for example, the optimal laying date varies from year to year because food availability also varies from year to year (Visser and Both 2005; Verhulst and Nilsson 2008). The optimal timing can be predicted by birds with a reasonable accuracy because the phenology of the food (e.g. emergence of insect prey) is driven by temperature (Embree 1970; Danks 1987). The temperatures in the part of the reproductive season that drive insect emergence correlate well with the temperatures in the part of the season that drive the onset of egg laying in birds, making temperature a reliable environmental cue (Gienapp et al. 2005; Schaper et al. 2012; Gienapp et al. 2014).

At this point I need to make the side note that phenotypic plasticity is a broad term that extends beyond the simple (adaptive) case that we have discussed here (Box 1.2). Phenotypic responses to a variable environment may be maladaptive in certain situations if, for example, individuals get exposed to novel, 'non-preferred' conditions because this novel environment impairs development and/or homeostasis (Ghalambor et al. 2007). In this view, phenotypic plasticity in seasonally breeding birds may be viewed as a result of the physical or physiological constraints of breeding in cold conditions in early spring (Perrins 1970; Stevenson and Bryant 2000), rather than an adaptive response to temperature fluctuations, although this notion is generally not well supported (Visser and Both 2005; Charmantier et al. 2008; Visser 2008). Also, there may be several plasticity-like mechanisms that simultaneously determine the phenotype. For example, female great tits (*Parus major*) will change their laying date in response to experiences in the previous seasons (Grieco et al. 2002; Gienapp and Visser 2006), and maternal (nongenetic) effects may determine the offspring's development with lasting effects on their phenotype as adults (Mousseau and Fox 1998; Räsänen and Kruuk 2007). Chapter 8 explores this special case of 'developmental plasticity' induced by maternal effects.

Understanding how phenotypic plasticity, as well as its (additive) genetic underpinnings, operates within populations is crucial in evolutionary ecology studies because it determines the precision with which patterns of individual consistency in behaviour or life history (permanent-environment effects or repeatability, i.e. the relative contribution of between-individual effects to the total phenotypic variability (Lessells and Boag 1987)) can be detected (Dingemanse et al. 2010; Van de Pol et al. 2016; Gienapp 2018; see page 13). If one, for example, were to estimate between-individual variation in avian

Box 1.2. *Phenotypic plasticity*

Phenotypic plasticity in labile traits describes the situation in which an individual (or genotype) expresses different phenotypes as a function of the environment (Schlichting and Pigliucci 1998; Pigliucci 2001). Adaptive plasticity arises because individuals can track phenotypic optima varying with environments, thereby maximising fitness across environments (Scheiner 1993; Ghalambor et al. 2007). The function describing phenotypic plasticity is called a reaction norm (Woltereck 1909; Scheiner 1993). In its most simplistic form, this reaction norm is a linear function, described by …

Box 1.2 *(continued)*

… an intercept or elevation (i.e. the trait value in the average environment) and a slope (the sensitivity of the trait to the environment) (Fig. B1.2.1). For example, many phenological traits, such as the timing of flowering or breeding, respond plastically to temperatures, with warmer springs leading to earlier phenological events (e.g. Nussey et al. 2005c; Brommer et al. 2008; Charmantier et al. 2008; Phillimore et al. 2010; 2012) (Figs. B1.2.1b and c):

$$z = a + bx,$$

where *z* is the trait, *a* and *b* are the intercept and slope of the reaction norm, respectively, and *x* is the (mean-centred) environment. In the phenology example, *b* is a negative number. This linear function is often assumed to adequately describe the thermal reaction norms of phenological traits, but may be over-simplistic in other contexts (e.g. Brommer et al. 2012; Carter et al. 2017), in which case some higher polynomial (e.g. a quadratic term) may be more suitable. This requires, however, (many) more than two observations per individual, a requirement seldom met in typical empirical (natural) systems.

Reaction norms of different individuals can run in parallel (Fig. B1.2.1a and b); this means that different individuals exhibit the same degree of plasticity with respect to the environment. In addition, the vertical position of each individual's reaction norm relative to that of its conspecific reveals the similarity in performance across all environments (in the examples of panels **a** and **b**, one individual (dotted line) consistently has a higher trait value than the other (solid line), regardless of the degree of plasticity). If reaction norms do not run in parallel (Fig. B1.2.1c), this indicates—in this particular case—that one of the individuals is less responsive to the environment than the other (sometimes leading to crossing reaction norms). As a result, the variation in phenotypes along the vertical axis differs among environments, a process known as individual-by-environment interaction or I×E (note that I×E can take many more shapes than merely the depiction in panel **c**). If I×E has a genetic basis, we term this a gene-by-environment interaction or G×E. The presence of G×E means that genetic trait variation is not constant across environments, which may have consequences for the ability of a population to respond genetically to selection (Merilä et al. 2001b; Turelli and Barton 2004; Kokko and Heubel 2008).

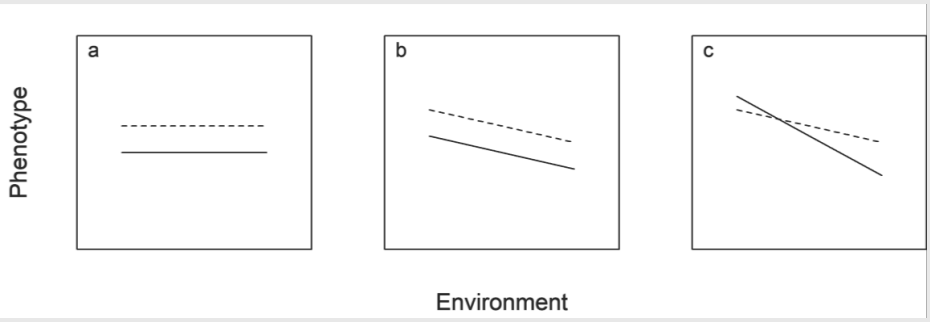


Figure B1.2.1*. Three (non-exhaustive) scenarios for phenotypic plasticity described by linear reaction norms; the dotted and dashed lines represent different individuals/genotypes. In* ***a****, neither individual responds plastically to the environment. In* ***b****, both individuals respond plastically to the environment to a similar degree (i.e. in this particular case the trait value decreases as the environmental value increases). In both* ***a*** *and* ***b****, individuals differ in their mean response across environments (difference in the intercepts). In* ***c****, both individuals respond plastically to the environment, but differ in their sensitivity (I×E) as well as the intercept (height) of the curve.*

timing of breeding without accounting for the effect of temperature, the relative contribution of between-individual to the total phenotypic variation would be strongly downwardly biased. This is because temperature-induced variation in laying dates would be 'consumed' by the residual variance component, reducing the relative contribution of the permanent environmental (and additive genetic) variance (Box 1.1). This has direct implications for the reliability of estimates of 'evolutionary potential' or 'evolvability' of a population (Houle 1992; Hansen and Houle 2004) and may lead to estimated responses to selection that do not concur with observations in the wild (Merilä et al. 2001b; see below). In the laying dates example, including a fixed effect of temperature in the model, i.e. fitting a reaction norm for each female, may reveal that although there is between-individual variation in laying dates, each individual also responds plastically to temperature, accounting for a substantial amount of the total phenotypic variation (Fig. B1.2.1b; Chapter 10). In addition to getting more accurate estimates of between-individual variation, recognising the role of the environment in shaping the phenotype through phenotypic plasticity helps explain how many populations have adjusted their phenology in response to global warming, as a population-wide shift in phenotype may sometimes mistakenly be interpreted as an evolutionary response (Gienapp et al. 2008; Merilä 2012; Charmantier and Gienapp 2014; Merilä and Hendry 2014).

Contemporary climate change as a driver of selection

Much like past major episodes of selection, the natural world is currently subjected to a plethora of selective pressures in the Anthropocene era (Rockström et al. 2009; Steffen et al. 2015; Scheffers et al. 2016), with humans now being considered the largest driver of evolutionary change in natural populations (Palumbi 2001). Among the anthropogenic drivers, habitat loss and climate change through greenhouse gas emissions are expected to have the strongest impact on biodiversity loss. A changing climate is projected to be accompanied by higher mean temperatures as well as more frequent extremes in temperatures (Beniston et al. 2007; Field et al. 2014). This puts selective pressures on populations that need to cope with this by adapting locally or by dispersing to more favourable habitat (Parmesan 2006).

One of the most recorded effects of climate change is the change in the timing of phenological events such as migration and reproduction (Parmesan and Yohe 2003; Root et al. 2003). In most cases, across taxa and ecosystems, increased temperatures in recent decades have led to an advancement in phenology. Most notably, this increase has been strongest in higher trophic levels (Thackeray et al. 2010; Thackeray et al. 2016), creating a mismatch between consumer and resource phenology in a wide range of study systems (Kharouba et al. 2018). Birds, for example, need to change their timing of breeding to maintain synchrony with the phenology of their food (e.g. plants or invertebrates). Many avian species, however, have not been able to keep up with the change in the phenology of their food (e.g. Visser et al. 1998; Both and Visser 2001; Thomas et al. 2001; Both et al. 2006; Nielsen and Møller 2006; Schultz et al. 2009; McKinnon et al. 2012). Chapter 2 in this

thesis deals in more detail with the reported effects of climate change on bird biology in general.

How does climate change lead to selection on phenology? To comprehend this, we need to understand the relationship between the timing of the phenological trait of interest (considering here the onset of egg laying) and the time when natural selection acts, i.e. the time when the nestlings' food requirements are highest. The former we refer to as the environment of decision-making (1), and the latter as the environment of selection (2) (Visser et al. 2010). Both environments, as described above, are affected by spring temperatures, but in a different manner (Visser et al. 2006; Gienapp et al. 2014). In great tits, females need to time their reproduction about a month ahead of the time the main food source for their nestlings—mainly caterpillars—abound (Visser et al. 2004a; Chevin et al. 2015). Successful reproduction can therefore only be achieved if the onset of laying was well timed (Visser et al. 2006; Visser and Gienapp in press), but simulations have shown that if one of the two environments (1 and 2) changes due to climate change, this will put selective pressure on consumer phenology (Gienapp et al. 2014). This is partly because the relationship between environment (1) and (2) is disrupted and temperatures in environment (1) can no longer be used as a reliable cue to predict conditions in environment (2) (with the additional, more fundamental reason that temperature cues are never perfectly reliable (Gienapp et al. 2014)). Chapter 10 deals in detail with the selective pressures on reaction norms.

The challenges of predicting evolution in the wild

When the breeder's equation fails

When (additive) genetic variation in a trait in the population and selection acting on this variation are quantified, we can predict the evolutionary response to selection from one generation to the next (Falconer and Mackay 1996; Lynch and Walsh 1998). Traditionally, in animal breeding, this is achieved using the breeder's equation (Box 1.1). In essence, the selection differential (S) can be considered the strength of selection, as it is the difference in the mean trait value between the generations (i.e. the original and the selected population). To calculate the response (R), however, S has to be multiplied by the heritability of the trait (h^2), as only part of the observed phenotypic variation can be attributed to additive genetic effects. This equation yields fair estimates of selection responses in animal breeding because not only can h^2 be quantified with reasonable accuracy in controlled conditions (Weigensberg and Roff 1996), selection (S) is always exactly known (the animal breeder determines who breeds with whom and how many offspring result from this). In the wild, however, the use of the breeder's equation, in its univariate or multivariate form (Box 1.1), is generally not recommended (Morrissey et al. 2010). This is because its application relies on strong assumptions about the causation of variation in fitness. We may estimate the phenotypic selection differential (S) for a given trait at a given time, e.g. laying date in wild birds in a particular breeding season, and conclude that early-breeding birds have a selective advantage. This conclusion, however,

is potentially flawed if the underlying assumptions are violated (Morrissey et al. 2010). There may well be a second trait correlated with fitness, e.g. clutch size or nutritional state, which is causing a covariance between the focal trait and fitness. To circumvent this issue, it has been proposed to estimate selection at the genetic, rather than the phenotypic, level (Rausher 1992; Stinchcombe et al. 2002), using the Robertson–Price Identity, or Secondary Theorem of Selection, for the additive genetic covariance between trait value and relative fitness in a simple quantitative genetic model (Robertson 1966; Price 1970). This approach is, however, far more data-hungry and hence often unfeasible when one is interested in annual estimates of selection (see discussion in Chapter 9). Reassuringly, phenotypic estimates of selection are not necessarily inferior to genetic estimates in every situation (Morrissey and Ferguson 2011; Reed et al. 2016b).

Inaccurate estimates of selection will lead to biased estimated responses to selection (Kruuk et al. 2003). The same is obviously true if the other component of the breeder's equation, heritability, is inaccurate. Phenotypic plasticity—as one of the several potential causes for apparent 'evolutionary stasis' (Merilä et al. 2001b)—may have a genetic basis (Box 1.2); non-parallel reaction norms will therefore lead to changes in genetic variation across the environmental gradient (Scheiner 1993). Environmental heterogeneity in additive genetic variance may lead to variability in true heritability if the other (environmental) sources of variation do not change at an equal rate. Failure to take these changes in (additive) genetic variation into account may lead to biased estimates of selection response (Hoffman and Merilä 1999). In Chapter 9, we expand on the issue of having selection and genetic variation (co)vary with the environment. Besides the environmental variability of genetic variance (and of selection: Wade and Kalisz 1990; Siepielski et al. 2009, 2013) there are numerous reasons why h^2 times s does not equal the observed evolutionary responses (R). I refer the reader to more comprehensive work for more information on this subject (e.g. Merilä et al. 2001b; Hansen and Houle 2004; Postma 2006; Morrissey et al. 2010).

The difficulties of interpreting phenotypic changes in the wild

In the face of global climate change, it is imperative to understand how populations will cope with these changing conditions (Visser and Both 2005; Tylianakis et al. 2008; Visser 2008). It starts, however, with the need for a clear picture of the underlying processes affecting putative responses, which Merilä and Hendry (2014) summarised in three categories. First, it has proven to be difficult in many long-term population studies to separate observed phenotypic responses (through phenotypic plasticity) from genetic change (Gienapp et al. 2008; Merilä 2012). Many phenological traits, for example, exhibit strong plasticity (see Box 1.2); when the climate warms and selection favours individuals to time their events earlier in the season, phenotypic plasticity often largely accommodates the observed population-level phenological changes (Gienapp et al. 2008). We have discussed above that such phenological adaptations will not be sufficient if the environment of selection changes at such a rate that the cue environment is no longer reliable (Visser 2008; Gienapp et al. 2014), leading to selection on the consumer reaction norm (cf. Chapter 10).

Second, if there is a phenotypic response to climate change (independent of whether this response has a (partly) genetic basis) it is not always clear whether the response is adaptive (i.e. leading a phenotypic shift towards the optimum). For example, an alpine snow vole (*Chionomys nivalis*) population has seen an increase in body mass in the past decade or so, whereas breeding values for body mass have decreased (Bonnet et al. 2017). Although over-winter survival was highest for heavy individuals, suggesting phenotypic selection for higher mass, the predicted survival from selection in recent short-summer years—when late-born juveniles are still growing when the snow-free season ends—in fact decreased with mass. This counterintuitive pattern could be explained as an adaptive response to viability selection in these juveniles for faster development into smaller adults (Bonnet et al. 2017). Importantly, Bonnet et al. (2017) showed that both evolution and selection (at the genetic level) opposed the direction of selection at the phenotypic level, warranting due care when inferring selection and adaptive change in natural populations. Several complementary ways exist that can be deployed to reliably infer the adaptiveness of a response (Merilä and Hendry 2014). In short, one can (1) perform reciprocal transplants and assess the fitness of 'novel' genotypes in simulated novel or past environments; (2) retrospectively estimate phenotypic selection and assess whether the change in mean trait value occurred in the expected direction; (3) estimate genetic selection (see previous section); (4) comparison to predictions of null-models of evolution (e.g. genetic drift) and (5) Q_{ST}–F_{ST} comparisons, i.e. comparison of the divergence in quantitative traits (Q_{ST}) with divergence in neutral molecular markers (F_{ST}) (Leinonen et al. 2013). Naturally, each of these methods in and of itself can seldom provide a reliable picture of whether phenotypic or genetic change is adaptive, and a combination of methods is therefore preferred (Merilä and Hendry 2014). Most importantly, perhaps, the researcher should know their study species and its ecology well and use common sense. In some bird species, for example, we know a fair deal about the ecological interactions with prey (Chapter 5), phenotypic and genetic correlations between life-history traits (e.g. laying date and clutch size; Sheldon et al. 2003; Postma 2005), and the adaptive landscape within which selection on laying date and clutch size operates (Chevin et al. 2015; Gamelon et al. 2018).

Lastly and briefly, changes in phenotypes and concomitant selective pressures are sometimes attributed to climate change where in reality a different actor is operating, such as habitat degradation, overexploitation (e.g. fishing), pollution, and more (see Merilä and Hendry 2014). Identifying the right environmental driver of selection is therefore crucial not only for targeted conservation measures, but also accurate evolutionary predictions. This is, however, beyond the scope of this thesis.

This thesis

General aims of this thesis

In this thesis, I use a combination of field experiments and state-of-the-art statistical modelling approaches to explore the evolutionary potential in wild, vertebrate

populations. The ecology and evolution literature generates a multitude of hypotheses, for example as to which environmental factors drive selection, how genetic variation varies with the environment, and which other (ecological) factors determine the accuracy of predictions of evolution in the wild (e.g. Merilä et al. 2001b). Outstanding questions posed in this context more generally could revolve around the question of which of these factors put(s) a (apparent) constraint on adaptation, whether we can identify it (or them) and how we can improve our (statistical) methodology to generate better evolutionary predictions.

Using the great tit (*Parus major*) as my main modelling system (Box 1.3), I aim to answer a broad range of questions all tightly revolving around the central question: **How are wild populations coping with environmental change and which ecological processes affect the rate of genetic adaptation?** A central approach to answering this question is quantifying the parameters that affect evolutionary dynamics (selection, plasticity, genetic variation) and predict quantitatively the course of microevolution against the backdrop of climate change. Although the great tit population of the Hoge Veluwe (Box 1.3) will be the main point of focus in this thesis, I will occasionally depart from this species and population to explain general concepts, simulate eco-evolutionary processes, and in one case use a multi-taxa approach to answer eco-evolutionary questions beyond single study systems.

Rather than attempting to provide an all-inclusive, clear-cut answer to the general question posed above, which will be impossible, I used different, complementary approaches to unravel the processes underlying (genetic) adaptation to novel environmental conditions. To this end, this thesis is divided into three main parts: **Part I** gives an overview of what we know about the ecological and biological consequences of climate change on birds. **Part II** explores, using experimental and observational approaches, the fitness consequences of reproductive timing in great tits. An **Intermezzo** will focus on some methodological aspects in study of ecology and evolution. Finally, **Part III** combines quantitative genetic and other statistical and simulation approaches to unravel patterns of adaptation in wild populations more generally.

Thesis outline

Part I. In **Chapter 2** of this thesis, I explore the ecological and biological consequences of climate change in birds. In an extensive (but non-exhaustive) review, we outline the known causes and consequences revolving around the geographical distribution of birds, their phenology (breeding time and migration), morphology and demography under a changing climate. We conclude with an outline of possible impacts of climate change in the (near) future.

Part II. Climate change is affecting phenology of consumer and prey at different rates (Thackeray et al. 2010, 2016; Kharouba et al. 2018) but we still know little about the mechanism underlying seasonal timing of reproduction and how this timing is constrained by external (a)biotic factors. To this end, Part II of this thesis aims to

Box 1.3. *Main study system: the great tit (*Parus major *L.)*

General research procedures

The majority of the work in this thesis takes place in the Hoge Veluwe National Park (52°02'07" N, 5°51'32" E, central Netherlands). The 171-ha study area consists of mixed stands of deciduous and coniferous woodlands, surrounded by a matrix of suitable habitat facilitating from and into the study area. The main species that make up the deciduous forest stands are pedunculate oak (*Quercus robur*), red oak (*Quercus rubra*), beech (*Fagus sylvatica*), and birch (*Betula sp.*). The coniferous stands consist of scots pine (*Pinus sylvestris*), spruce (*Picea abies*), Douglas fir (*Pseudotsuga menziesii*) and larch (*Larix* sp.). Situated on poor sandy soils, undergrowth in the Hoge Veluwe area is typical for nutrient-poor sandy systems, including heather (*Calluna vulgaris*) on the poorest soils and purple moor-grass (*Molinia caerulea*) on the richer parts. The forested area of the entire Hoge Veluwe park is interspersed with heathlands and sand dunes. Although the breeding bird populations have been monitored continuously since 1955, it is since 1973 that *c.* 400 nest boxes (+ *c.* 50 extra for smaller passerines) are permanently available. With few natural nesting cavities present in the study area, several hole-breeding species readily accept these boxes, leading to annual densities of roughly 120 pairs of great tit, 85 pairs of blue tit (*Cyanistes caeruleus*), 90 pairs of pied flycatcher (*Ficedula hypoleuca*), 15 pairs of nuthatch (*Sitta europaea*), and 1–2 pairs of coal tit (*Periparus ater*). The laying date and clutch size and the number of fledged chicks is recorded for each breeding pair and both the parents and their nestlings are equipped with a leg ring with a unique identifier. This allows us to follow individuals over their lifetime and to construct a 'social' pedigree (as opposed to a genetic pedigree, since a significant portion of offspring in an average brood is sired by a different male (Van Oers et al. 2008; Brommer et al. 2010).

The great tit as central species in ecology and evolution

The great tit (*P. major*) has been the main study species of many ecological, evolutionary, and behavioural studies, so a lot is known about its biology. Since the Hoge Veluwe population, as well as other populations within (e.g. Vlieland, Liesbosch, Oosterhout; Van Balen 1973; Postma 2005) and outside the Netherlands (e.g. Wytham Woods, UK; Cresswell and McCleery 2003; Charmantier et al. 2008) have been monitored for several decades, they provide an invaluable source of data to answer outstanding questions in ecology and evolution (Clutton-Brock and Sheldon 2010). The central focus of research in the Hoge Veluwe population has been to understand variation in seasonal timing of reproduction (laying date of the first egg) and its proximate and ultimate causes (e.g. Van Balen 1973), and more recently, its evolutionary potential within the context of global climate change (Visser et al. 1998; Gienapp et al. 2006; Visser et al. 2006; this thesis). Great tits rely strongly on the abundance of Lepidopteran caterpillars to raise their offspring (Lack 1950; Betts 1955; Royama 1966; Van Balen 1973). Two main prey species are the winter moth (*Operopthera brumata*) and oak-leaf roller (*Tortrix viridana*), with penduncate oak as one of their main hosts. The abundance of these caterpillars is sampled every year, starting late April and ending early-to-mid June, using frass nets deployed underneath several oak trees across the study area (see Visser et al. 2006 for details). The reliance on caterpillars, as well as the obligate cavity-breeding behaviour, render old oak …

Figure B1.3.1. *Female incubating great tit*

Box 1.3 *(continued)*

… woodlands the typical and most suited habitat for the great tit, although the species has adapted well to other, less-than-ideal habitats (Gosler 1993).

Laying date and clutch size are **heritable traits** (Van Noordwijk et al. 1981; Sheldon et al. 2003; Postma and van Noordwijk 2005b) and have therefore the potential to respond to selection. Importantly, as described in previous sections, both traits can be expressed several times within the lifetime of an individual and is responsive to an environmental cue (i.e. temperature—and photoperiod—for laying date and breeding density or food availability for clutch size (e.g. Both et al. 2000)). Despite its relatively short generation time of about two years (Garant et al. 2004a; Kvist et al. 2007) and its concomitant short lifespan (about half of the great tits in the Hoge Veluwe population only breeds once in their lifetime), it has been a central topic in studies of phenotypic plasticity (Nussey et al. 2005c; Charmantier et al. 2008; Husby et al. 2010, 2011; Chapter 10). Like other long-term monitored populations, it has a central place in the prediction of evolution in response to climate change (for some examples including other species than great tit, see e.g. Visser et al. 1998, 2006; Réale et al. 2003b; Both et al. 2004b; Brommer et al. 2005, 2008; Gienapp et al. 2006; Charmantier et al. 2008; Plard et al. 2014).

Climate change as an environmental driver of selection in great tits

Climate change is altering **selection** pressures on great tit phenology via a **phenological mismatch** (Visser et al. 1998; Reed et al. 2013b). Compared to the 1970s and early ’80s, caterpillar phenology (egg hatch date of the winter moth) than the rate at which the timing of bud burst of the oaks advanced (Visser and Holleman 2001). This fast change in caterpillar hatch date, which led to a concomitant shift in the peak date of caterpillar availability, was not tracked fast enough by great tits, who now started laying eggs increasingly late with respect to this peak date in food abundance (Visser et al. 1998). This ‘mismatch’ with food abundance has severe fitness consequences, as offspring from mismatched parents are in poorer physical shape and have a lower recruitment probability than those from well-matched parents (Visser et al. 2006; Reed et al. 2013b). As mismatch increases due to global warming, selection for earlier egg-laying intensifies (Fig. B1.3.2). Models of extreme climate scenarios predict that sustained direction selection without concomitant evolutionary response may have long-term population consequences (Gienapp et al. 2013a; Reed et al. 2013a).

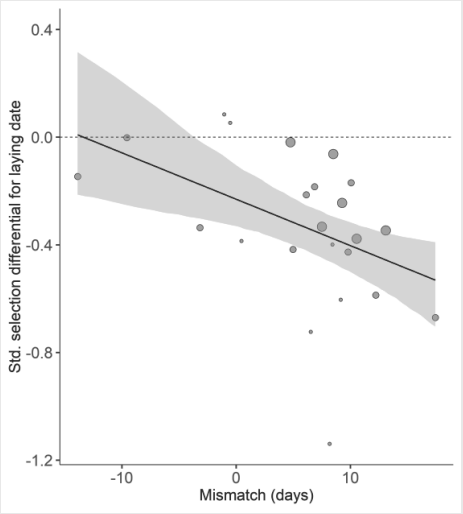


Figure B1.3.2. *Standardized selection differentials (based on the number of recruiting offspring) for laying date against mismatch between peak dates in food demand and caterpillar availability in the Hoge Veluwe. Negative mismatch means birds bred too early; positive values indicate that birds bred too late. Shown are 23 years between 1994 and 2017 (i.e. years that had a complete caterpillar distribution; see Chapter 5). Line and shading are estimates and 95% bootstrapped confidence interval of a regression weighted by the number of recruits (coefficient: –0.015 [–0.032, –0.005]; r^2 = 0.181 [0.016, 0.418]). Symbol sizes, small: ≤ 20 recruits; medium; > 20 and 40 recruits; large: > 40 recruits.*

understand why birds are not advancing their reproduction in response to selection for earlier breeding (Visser et al. 1998; Gienapp et al. 2006). In **Chapter 3**, we specifically aim to test the *constraints* hypothesis of reproduction posed by Visser et al. (2012) (i.e. that fitness costs of breeding too early prevent a response to selection for earlier breeding) in great tits by means of a supplemental feeding experiment. Food supplementation prior to egg-laying is known to advance laying date in a variety of contexts (Verhulst and Nilsson 2008; Ruffino et al. 2014). When (pre-laying) breeding conditions are poor, then, we would expect birds that breed too early to pay a fitness cost once food supplementation ceases at the start of laying. Individual females vary in quality, however, and some of them may be genetically predisposed to breed earlier and in that way suffer less major fitness costs or be less sensitive to food supplementation. In this experiment, we thus combine an experimental feeding approach with quantitative genetics to obtain a comprehensive insight into the proximate causes of laying date.

Despite the usefulness of feeding experiments, they do have a major disadvantage, namely that they change the physical condition of the females, with potential carry-over effects to subsequent breeding stages (Verhulst and Nilsson 2008). The ideal experiment, therefore, would be to manipulate laying 'cleanly' and subsequently assess fitness costs. One decent approach would be to genetically manipulate birds into breeding either earlier or later through strong artificial selection on (genomic breeding values for) laying date and assess the causal fitness consequences of laying date in the wild. Such a selection experiment is being undertaken at the moment. Briefly, about 2000 birds from the Hoge Veluwe population were genotyped and their genomic breeding values for laying date calculated; based on these breeding values, extreme genotypes were selected and mated in aviaries to produce extreme-genotype offspring (details are given in Gienapp et al. (2019) and Verhagen et al. (in review)). The aim of **Chapter 4** is, first, to assess the immediate fitness consequences of genomic selection on laying date in foster-reared selection-line offspring in the wild. Eggs produced by birds in aviaries from the selection lines were experimentally fostered with wild breeding pairs and the fitness of these offspring were subsequently monitored. Second, we monitored the recruiting offspring from the eggs of the selection lines that were taken to the wild to determine their realised laying date in the wild and understand the fitness effects resulting from this. Chapter 4 provides the first, tentative results of this experiment, as it is still ongoing.

Ultimately, breeding success in great tits is determined by the match with caterpillar phenology (Visser et al. 2006; Reed et al. 2013b), but the best way to describe phenological match is still a matter of debate. It has been suggested that rather than looking solely at the match between peak dates in consumer and prey phenology, greater care must be taken to incorporate the full phenological distributions to accurately estimate ecological interactions between trophic levels and hence the evolutionary and demographic consequences of climate change (Miller-Rushing et al. 2010; Lindén 2018). So far, studies that have incorporated this comprehensive measure of synchrony are rare, most likely due to the lack of appropriate data. In **Chapter 5** we uniquely make use of long-term data on great tit and caterpillar phenology to quantify phenological synchrony in the two ways outlined above. We specifically aim to test whether the more complex measure of synchrony (i.e. the degree of overlap between the distributions) outperforms the simpler

version of the match in peak dates, and give recommendations for researchers of long-term study populations as to how to go about studying phenological mismatch in species in similar, highly seasonal environments.

Intermezzo. To push the science of ecology and evolution forward, we continuously need to seek for better concepts and better tools. One of the novel developments in the field is the use of open data to (1) conduct data-driven meta-analysis (as opposed to traditional meta-analysis on reported effect sizes) and to (2) answer novel outstanding questions with available data that were not originally collected for the purpose of these questions (Whitlock et al. 2010; Hampton et al. 2013). In **Chapter 6**, we present a how-to paper in which we combine both aspects and explain how data freely available from online data repositories can be used to answer novel biological questions in a meta-analytic framework. We make us of a previously published guideline for how to find open data (Culina et al. 2018) and outline all procedures that need to be undertaken to successfully perform a meta-analysis. We make us of a concrete example, which we will address in Chapter 9.

One major tool in ecology and evolution, and one recurring in this thesis, is the random regression model to infer among-individual variation in plasticity (I×E and/or G×E; Nussey et al. 2007; Dingemanse and Dochtermann 2013). A powerful tool, it can sometimes give misleading results because of an inherent property in many traits sometimes overlooked: heterogeneity in residual variance. As we will see later in this thesis, in at least the particular case of the great tit, comparison between studies and populations reveals a lack of consistency with respect to the reported I×E and G×E, which—at the Hoge Veluwe—can be mostly attributed to the way residual variance was treated in the random regression model. As there seems to be no consensus on how residual variance in random regression models should be treated, and students of ecology and evolution may not be fully aware of its impact, we conducted a simulation study in **Chapter 7** to specifically test the effect of heterogeneity in residual variance on estimates of I×E and the statistical 'power' to detect it.

Part III. In the final part of this thesis, we use quantitative genetic methods to make evolutionary predictions in wild populations. In **Chapter 8**, we depart from laying date and use clutch size as the focal life-history trait. Clutch size is a heritable trait that has a strong environmental component, most notably breeding-pair densities and concomitant competition for food (e.g. Both et al. 2000). Constraints in food availability may cause nestlings from larger broods to attain a poorer physical condition and, if this carries over to the adult stage, thus render their own clutch size smaller. In this chapter, we explore the ecological mechanisms underlying this *negative maternal effect* and use an individual-based model to test whether this effect has the potential to slow down or speed up the rate of (phenotypic and genetic) adaptation to novel environmental conditions.

Quantitative genetic analysis is a useful tool to predict evolutionary change, but has been proved to be difficult in the wild (see previous sections). This may sometimes be

inferred as evolutionary stasis (Merilä et al. 2001b), but the reason for differences in expected and observed evolutionary responses may in fact be rooted in the failure to recognize the presence of an underlying environmental coupling between heritability and selection (Hoffman and Merilä 1999; Wood and Brodie III 2016). In **Chapter 9** we specifically address the question of whether (1) such an environmental coupling is ubiquitous in wild, vertebrate populations and (2) whether this affects the rate of expected evolutionary change. We do this using a unique open data approach (see Chapter 6), collecting phenotypic data and pedigrees from a range of (mainly avian) species and populations and a variety of traits. The aim of this chapter, with its broad spectrum of species and traits, is to contribute to the general debate about the role of an environmental coupling between genetic variation and selection as a putative force for constraints in adaptation.

Phenotypic plasticity is an important mechanism by which organisms can respond to environmental conditions, but the question remains whether this plasticity is sufficient to keep track of a directionally changing environment (Visser 2008) and whether *evolutionary rescue* is necessary to safeguard populations from extinction (Carlson et al. 2014). In the final research chapter, **Chapter 10**, we investigate whether selection acts on either component (elevation and slope) of the thermal laying-date reaction norm in great tits. We aim to understand (1) how individuals differ in their reaction norm and whether this difference has a genetic basis, (2) whether there is selection on the reaction norm, rather than the trait (laying date) in a given environment, and (3) whether selection on the reaction norm has led to evolutionary change under a changing climate. In this chapter we use an integrated approach where we identify the underlying source of selection (i.e. changes in the timing of maximal caterpillar biomass reaction norm) and use rigorous, hitherto little-used statistical tools to quantitatively predict the evolutionary potential of a key life-history trait in a reaction norm context.

In **Chapter 11** I synthesise the results and patterns found in this thesis in a general discussion.

PART I

General patterns

Chapter 2

Climate change impacts: birds

Barbara M. Tomotani, Jip J.C. Ramakers & Phillip Gienapp

ABSTRACT

Climate change can affect populations and species in various ways. Rising temperatures can shift geographical distributions and lead to (phenotypic or genetic) changes in traits, mostly phenology, which may affect demography. Most of these effects are well documented in birds. For example, the distribution of species has shifted polewards, and birds are nowadays breeding or migrating earlier. An important aspect of the observed phenological changes is whether species are thereby able to maintain synchrony with phenological changes in their environment, e.g. the phenology of their prey species. Disrupted synchrony, for example between predator and prey, can lead to reduced reproductive success or survival, which can negatively affect demography. Evidence for this happening in birds is – so far – limited but theoretical models predict that extinction risks could arise through insufficient adaptation to such phenological mismatches.

eLS, John Wiley & Sons, Ltd: Chichester (2016)

Introduction

Over the past 100 years the global climate has warmed considerably, mainly from the 1980s onwards. This increase in temperature is not globally uniform but differs between regions and within seasons. For example, winter temperatures have increased more than summer temperatures, and temperatures in the northern hemisphere have increased more than in the southern hemisphere (Walther et al. 2002). This spatial and temporal heterogeneity can have important consequences for species and populations. Numerous studies covering a wide range of taxa have shown biological responses to global warming and a 'coherent fingerprint of climate change' is visible (Parmesan and Yohe 2003; Root et al. 2003; Parmesan 2006). For example, insects, birds and even fish have extended their geographical distribution poleward because the geographic distribution of their 'bioclimatic envelopes' shifted.

Climate can affect any species in two fundamentally different ways. First, ambient temperature can directly affect the organism itself. The rates of cellular processes are temperature dependent: a temperature increase of 10°C doubles it. Therefore all physiological processes in ectotherm organisms, as e.g. insects, fish or reptiles, are strongly dependent on ambient temperature. Endotherm organisms, i.e. mammals and birds, keep their body temperature constant and consequently their physiological processes are independent of the ambient temperature. However, to achieve this they have to spend energy on thermoregulation, which can be substantial under extreme conditions.

Second, ambient temperatures affect the organism's biotic environment by effects on interacting species, i.e. predators, prey or competitors. Rising temperatures can disrupt the phenological synchrony between species, for example between the time when great tits, an insectivorous passerine, breed and need abundant prey to raise their large broods and the time when this prey is most abundant (Visser et al. 1998; Visser et al. 2006). Such indirect climatic effects can also be more complex and have a more dramatic impact: the regularly occurring El Niño atmospheric phenomenon causes a shift in the cold Humboldt Current in the Pacific Ocean. This current brings nutrients to surface waters where they sustain rich algae growth, which in turn sustains abundant fish populations. During an El Niño the fish populations crash, which leads to complete breeding failure and even increased adult mortality among seabirds along the West-Coast of North- and South-America (Barbraud and Weimerskirch 2003).

Birds are generally well studied and many bird populations have been monitored for a long time, sometimes even for more than half a century. This presents a unique opportunity to study the impact of climate change since it is possible to combine extensive data sets with comparably good knowledge about relevant biological effects and mechanisms.

Observed impacts of climate change on birds

Geographical distributions

The current distribution of bird species can be mapped as a function of their environment using so-called 'bioclimatic envelopes' (e.g. Howard et al. 2015). Under global warming these 'envelopes' are expected to shift, leading to latitudinal or elevational shifts in species' distributional ranges (e.g. Huntley et al. 2008). Various bird species in Great Britain and North America, for example, have shifted their northern range margins toward higher latitudes, with southern birds moving at a rate of 0.95 and 2.35 km/year, respectively (Thomas and Lennon 1999; Hitch and Leberg 2007). Not every species, however, is expected to shift its distribution at the necessary rate expected from changing abiotic conditions. For example, migratory birds are likely to suffer increased competition for resources with resident birds under increased winter temperatures, as these more benign winter conditions increase the survival probability of residents and may enhance their dispersal and colonisation of new sites (Schaefer et al. 2008). Conversely, if temperatures and resource availability increase in spring, migrants may benefit because they can colonize new breeding sites previously too cold or resource-limited. As another example, montane birds, confined to mountains, may be inhibited in their dispersal abilities and therefore be susceptible to extinction due to global warming (Sekercioglu et al. 2008). Species may further be limited in their dispersal ability if this will result in decoupling of crucial trophic interactions, for example, if birds' dispersal abilities exceed that of their resources (Van der Putten et al. 2010). Ultimately, a species' propensity to change its distributional range over decadal scales will depend, amongst others, on the life history of the species, average climate conditions, geographical context, and human land-use practices (Bradshaw et al. 2014; Lehikoinen and Virkkala 2016). See also DOI: 10.1038/npg.els.0003238

Phenology: breeding time

One of the first reported impacts of climate change on avian biology was the advancement in breeding time (Crick et al. 1997; Dunn and Winkler 1999). Avian breeding time is strongly plastic in response to ambient temperatures with birds breeding early under warmer temperatures. As climate change has increased spring temperatures in recent decades, avian breeding time has advanced along with it, with the magnitude of the response differing between species or populations within a species (Both et al. 2004a; Torti and Dunn 2005).

One exceptionally well-studied example of changes in avian phenology and arising mismatches between trophic levels comes from a Dutch long-term study on great tits (Visser et al. 1998; Visser et al. 2006). The great tit is the secondary consumer in the great tit–winter moth–oak food chain. The phenology of great tits needs to be well timed with

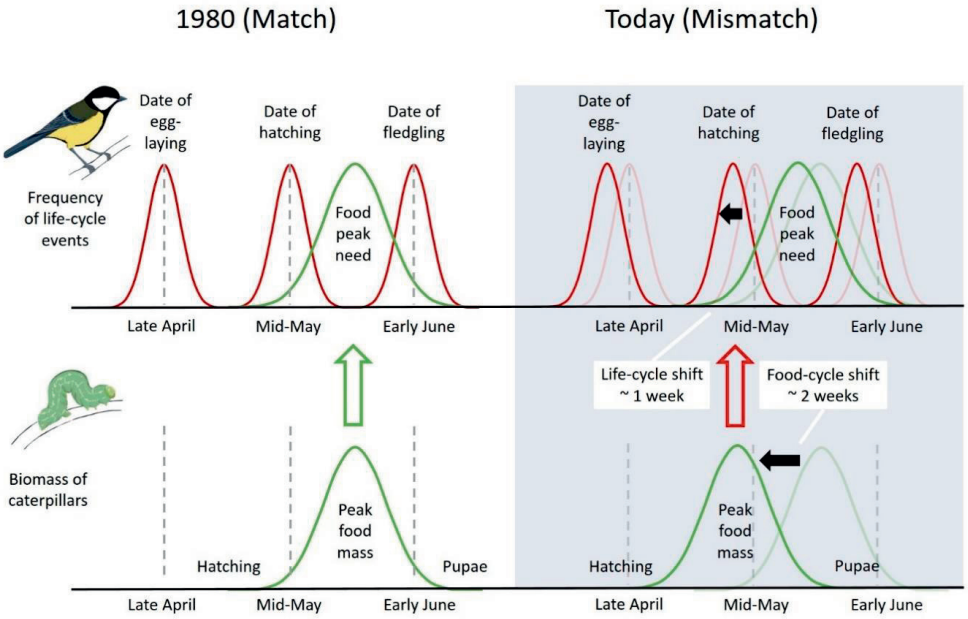


***Figure 2.1**. Schematic representation of the great tit–caterpillar phenology in the Hoge Veluwe. Three important life-history events in the great tit reproduction cycle are denoted with the red line and nestling peak food need with a green line in the top half of the schematic; the caterpillar phenology is indicated with the green line in the lower half. Before the effects of climate change were apparent, peak food demands and availability coincided (left-hand side of the schematic); due to increasingly warmer springs, the caterpillar biomass peak has advanced by ~2 weeks, whereas the timing of nestling peak food need has advanced at a slower rate, leading to 'phenological mismatch' (right-hand side).*

that of lower trophic levels as the birds rely heavily on caterpillars as the main food source for their nestlings. Caterpillars of the winter moth (*Operophtera brumata*) hatch in April and have to feed on the fresh leaves of oaks, as they grow less rapidly and attain lower fecundity if they feed on older leaves (Van Asch and Visser 2007). They develop through all instar stages in May to pupate in late May or June. These caterpillars thus show a clear biomass peak generally around mid-May but depending on spring temperature this can vary by about two weeks. Great tits use this biomass peak to provision their nestlings, whose energy demands are highest when they are nine to twelve days old (Visser et al. 2006; Both 2010b) and brood success is highest when this energy demand coincides with biomass peak date (Visser et al. 2006; Reed et al. 2013b; see left-hand side in Fig. 2.1). Due to warming springs, however, the caterpillar biomass peak has advanced by about two weeks, whereas the time of peak nestling energy demand has advanced by about five days (Visser et al. 1998; Visser et al. 2006; see right-hand side of Fig. 2.1). Consequently, there is now asynchrony between nestling food demand and food availability.

Why is there a mismatch between the caterpillar and great tit phenology? The lack of an adequate response to warming springs has been related to two, mutually non-exclusive theories (Visser et al. 2012). First, birds may be constrained in the advancement of their reproduction in early spring simply because they cannot obtain enough resources to meet the physiological demands associated with egg production (the *constraint* hypothesis). This means that birds either cannot lay eggs in early spring or, if they can, they will incur high survival costs. This hypothesis is unfortunately very difficult to test as it requires experimentally advancing breeding time without providing the birds with additional resources (which would lift the very constraint one aims to test; Verhulst and Nilsson 2008; but see Gienapp et al. 2006).

A second hypothesis states that the cues that birds use to time their reproduction are no longer adaptive (the *cue* hypothesis). Great tits have to plan their reproduction about a month in advance of the caterpillar biomass peak. By the time this peak occurs, the birds should have built the nest, laid and incubated the entire clutch of eggs, and raised the chicks up until the moment when their energy requirements are highest. Thus, the birds make their reproductive decision (i.e., when to lay eggs) in a different environment to where selection (i.e., nestling survival) takes place (Visser et al. 2004b). The phenology of both caterpillars and great tits depends on temperature. However, due to imperfect cue reliability, consumer (e.g. great tit) phenology tends to be always less plastic than the resource (e.g. caterpillar) phenology; this means that even under homogeneous environmental change phenotypic plasticity of the consumer phenology will be insufficient and this will inevitably lead to selection on consumer phenology (Gienapp et al. 2014). Ultimately, therefore, birds will only be able to keep their phenology in synchrony with that of their prey through a genetic shift (advance) in their average breeding time (See Plastic versus Genetic Changes). Not all species or populations have become maladapted to these novel environmental conditions. Great tits in UK forest systems, for example, now breed too early but maintain their synchrony with the food peak by increasing the incubation period (Charmantier et al. 2008).

Phenology: migration

The annual cycle of a temperate-zone migrant, which comprise the most studied species, can be divided into four main phases: a) "wintering", when no breeding activity occurs, b) spring (or vernal) migration, the movement from the wintering to the breeding grounds, c) breeding, and d) autumn migration, the movement from the breeding to the wintering grounds. A number of studies focus on the spring migration and, more specifically, on the arrival time of migratory birds, particularly passerines (Both and Visser 2001; Ahola et al. 2004; Kristensen et al. 2015). Similarly to the breeding stage, there is an optimal time to migrate and arrive at the breeding sites (Jonzen et al. 2007; Alerstam 2011). On one hand, arriving too early can be costly when environmental conditions are still harsh or unpredictable. On the other hand, late arriving individuals can face stronger competition

for mates or territories and may also experience reduced reproductive success due to rapid decline of resources in summer. For example, in 1996 five days of exceptionally cold and rainy weather during the main arrival time caused mortality of about 50-70% in North American Cliff Swallows (Brown and Brown 2000). Such cold spells happen regularly during the arrival time of this species imposing a high cost of early arrival. However, selection for early arrival seems to be the general case in migratory birds (Bêty et al. 2004; Smith and Moore 2005; Rubolini et al. 2010; Gienapp and Bregnballe 2012; Arnaud et al. 2013).

Since the optimal arrival time varies among years depending on the progress of spring, avian migration time also shows phenotypic plasticity similar to avian breeding time. It has been shown that birds adjust timing of their migration to climate and arrive earlier in warmer springs and after milder winters. So an expected effect of climate change on bird migration would be the earlier arrival of migrants to their breeding grounds (Walther et al. 2002).

However, the observed pattern was not uniform and while advancements were reported for some species (Marra et al. 2005), in others there was very little change (Both and Visser 2001). For example, long- and short-distance migrants could differ in the degree of phenotypic plasticity expressed in arrival time. Short-distance migrants are likely to show more flexible responses since climatic conditions at their wintering areas are more closely related to the ultimately important conditions at the breeding areas. Thus, more reliable cues may be available for them than for long-distance migrants. For long-distance migrants (e.g. those wintering south of the Sahara) climatic conditions at the wintering areas correlate less closely with climatic conditions at the breeding areas. They are therefore supposed to rely mainly on internal rhythms and photoperiod to time their departure from the wintering areas (Gwinner 1996). This would mean that their departure time is more or less constant among years, which in turn means that these species may be too inflexible to adjust to climate change. Alternatively, by adjusting their migration speed to environmental conditions en route would be a way for long-distance migrants to express some degree of phenotypic plasticity and be able to adjust their arrival time accordingly (Both 2010a). Another way of compensating would be the shortening of migration distances observed in some species. This not only reduces the distance needed to be covered by the migrant but also potentially allow cues to be more correlated and predictable (Visser et al. 2009b).

A meta-analysis combining data on 249 species from 18 studies found that arrival time of migratory birds has advanced, with birds arriving earlier after milder winters and in warmer springs (Gienapp et al. 2007). However, in this case no clear differences between European long- and short-distance migrants were found, which indicates that also long-distance migrants have been able to respond to climate change. The remaining question is whether the observed advancements in migration time match the supposed shifts in the optimal arrival time. Unfortunately, almost no study has reported data on reproductive

success or survival in relation to arrival date, mainly due to the difficulties in getting such information for most species.

The majority of studies on climate change effects on birds have been conducted on passerines. However, similar patterns are observed in other avian orders, with variable responses according to species characteristics. For example, earlier arrival time at the breeding grounds in Iceland has been reported for several non-passerine taxa (gulls, waders, geese; Gunnarsson and Tomasson 2011). Arctic-breeding geese are an interesting example as they present marked differences from passerines: their northward migration is related to the timing of vegetation growth, their main food source, and if a bird is able to fatten enough during migration it will also be able to breed earlier upon arrival (Van der Graaf et al. 2006; Van der Jeugd et al. 2009). Similarly to passerines, geese such as the brent goose seem to also suffer from mismatches: climate-related changes in timing of vegetation growth mean the birds are now late in relation to their food, which means less fattening opportunities for themselves or their offspring (Clausen and Clausen 2013). Analysis of long-term datasets also shows that some populations of the barnacle goose shortened their migration distance and/or became resident in the temperate wintering location. This change in breeding location, however, may also make the birds more mismatched, as their chicks are born too late in relation to the onset of vegetation growth in the new location (Van der Jeugd et al. 2009).

The complex annual cycle of migrants that makes the animals experience climatic conditions from several portions of the world brings the challenging task of assessing the impact of climate change with respect to their entire annual cycle (Marra et al. 2005; Visser and Both 2005). For example, some studies argue that conditions experienced in the wintering grounds or during migration can be even more important than those experienced at the breeding environment to explain fitness differences or population declines (Small-Lorenz et al. 2013).

Most studies focus on the (spring) migration from the wintering to the breeding grounds, mostly based on arrival dates at the breeding ground or passage dates at a given point close to the breeding grounds. Departure dates from the breeding (Europe) to the wintering grounds (Africa) are also changing. The pattern, however, differs between short and long distance migrants: while advancements have been observed for long distance migrants, delays have been observed in short distance migrants (Jenni and Kery 2003).

The recent development and miniaturization of better tracking devices now also allows following individuals of small species, such as most passerine long-distance migrants, throughout the annual cycle (Stutchbury et al. 2009; Bridge et al. 2011; McKinnon et al. 2013). This will enable much more detailed studies on timing, movements, and site selection of migratory birds and thereby, hopefully, shed new light on our understanding of the impacts of climate change on migratory birds. See also DOI: 10.1002/9780470015902.a0005450.pub2.

Morphology

Temperature is known to correlate with body size of different groups, with paleontological (Smith et al. 2009) and recent (Sheridan and Bickford 2011) evidence that warmer temperatures correlate with smaller body sizes. For homeotherms, Bergmann`s rule predicts that body size is adjusted to climate and animals would be larger in higher latitudes (i.e. colder climates) as an energetic adaptation to the colder temperatures. Although originally proposed for different species inhabiting distinct latitudes, it was later also applied to differences among populations of the same species (Salewski et al. 2010). If higher temperatures predict smaller body size, then climate change should lead to smaller body size in birds and mammals (Sheridan and Bickford 2011).

Although a number of studies have reported changes in avian body size correlated to temperature changes, the relation to temperature may, however, be indirect and a true adaptive response to changing climate is still lacking (Teplitsky and Millien 2014). This was investigated in depth in at least two species: red-billed gulls (Teplitsky et al. 2008) and great tits (Husby et al. 2011a). In both cases, a plastic rather than microevolutionary body size adjustment was detected (see Plastic versus genetic changes). The correlation between body size and temperature is particularly difficult, since body size is also affected by other factors than heat-dissipation capability that also co-vary with temperature as, e.g., food availability, which may determine the growth at young age and also resistance to starvation (Teplitsky et al. 2008; Sheridan and Bickford 2011; Teplitsky and Millien 2014).

Demography

Climate change can affect reproductive success and survival directly or indirectly. Extreme weather events can have strong direct effects; for example, unexpected cold spells can induce mass mortality in migrants as pointed out above. Most effects, however, will work indirectly by changing abundances or synchrony of interacting species. For example, winter climate in Antarctica affects krill (*Euphausia superba*) abundance, which in turn affects reproductive success and thereby population numbers in Adélie and chinstrap penguins (Trivelpiece et al. 2011).

Climate change has also disrupted the synchrony between interacting species, which could, for example, mean that a predator now does not encounter the maximum prey abundance anymore because the phenology of the predator has advanced less (or more) than that of its prey. Such disruptions have often been reported and are also predicted to be common (Gienapp et al. 2014). The demographic consequences of disrupted synchrony between predator and prey have been well studied in great tits and caterpillars. The disrupted synchrony has led – as expected – to selection on the birds' breeding time (Visser et al. 1998). The corresponding reduced reproductive success (the 'demographic load of selection') can drive populations to extinction if selection becomes too strong or the rate

of evolutionary change of the population is too small (Bürger and Lynch 1994; Lynch and Lande 1998).

Using a theoretical modelling approach, Gienapp et al. (2013a) could show that climate change would increase the mismatch between the great tits and their caterpillar prey too strongly for the birds to adapt by microevolution and that this would lead to a non-negligible extinction risk. However, the expected negative relationship between the strength of selection and population growth rate has not been observed in this population due to density-dependent juvenile winter survival (Reed et al. 2013b). Incorporating this effect of density-dependence into the theoretical model by Gienapp et al. (2013a) showed that density-dependence can buffer populations against reduced reproductive success due to disrupted synchrony and thereby reduce extinction risk (Reed et al. 2015).

Plastic versus genetic changes

As described above, changes in the phenology of birds have frequently been observed. Such consistent trends over time could be the results of phenotypic plasticity or an evolutionary response to selection. Phenological traits show large year-to-year fluctuations mostly driven by ambient temperature and the observed advancements could hence be a phenotypically plastic response to increasing temperatures. As also pointed out above, climate change is likely to lead to selection on phenology (Gienapp et al. 2014) and these observed changes could therefore also be an evolutionary response to this selection. Disentangling phenotypic from genetic changes is difficult when a genetic change cannot be directly tested because no suitable molecular genetic markers are available. One possibility is to predict 'breeding values' of individuals using quantitative genetic approaches (Wilson et al. 2010) but these require a known pedigree. These fairly high demands on data quality are likely the explanation why, so far, no study reported a genetic change in phenology in response to climate change (Charmantier and Gienapp 2014). This lacking evidence does, however, not mean that populations will not be able to adapt to climate change by phenotypic plasticity but only that we have no suitable data or methods to show this. See also DOI: 10.1002/9780470015902.a0022545 and DOI: 10.1038/npg.els.0001789.

Possible impacts in the future

Global temperatures are predicted to rise at least by 1.5°C until 2100, with the most extreme scenario predicting increases by from 3.5 to 6°C (Field et al. 2014). Consequently, the whole biosphere will be confronted with on-going climate change and observed changes are very likely to continue or increase. There are three ways for populations to

survive: 1) they can evade by dispersing to suitable habitats elsewhere, where climatic conditions are still favourable; 2) they can stay put and adjust to the changed conditions by means of phenotypic plasticity without altering their genetic constitution; 3) they can adapt to the changed conditions by means of genetic changes through the process of evolution. Of course, it is also possible – and even likely – that all three processes happen simultaneously and the relative importance of these three different ways to cope with climate change depends on the time scale considered, the species' life history, the rate and the extent of (predicted) climate change, the availability of alternative habitats, and the dispersal ability of the species.

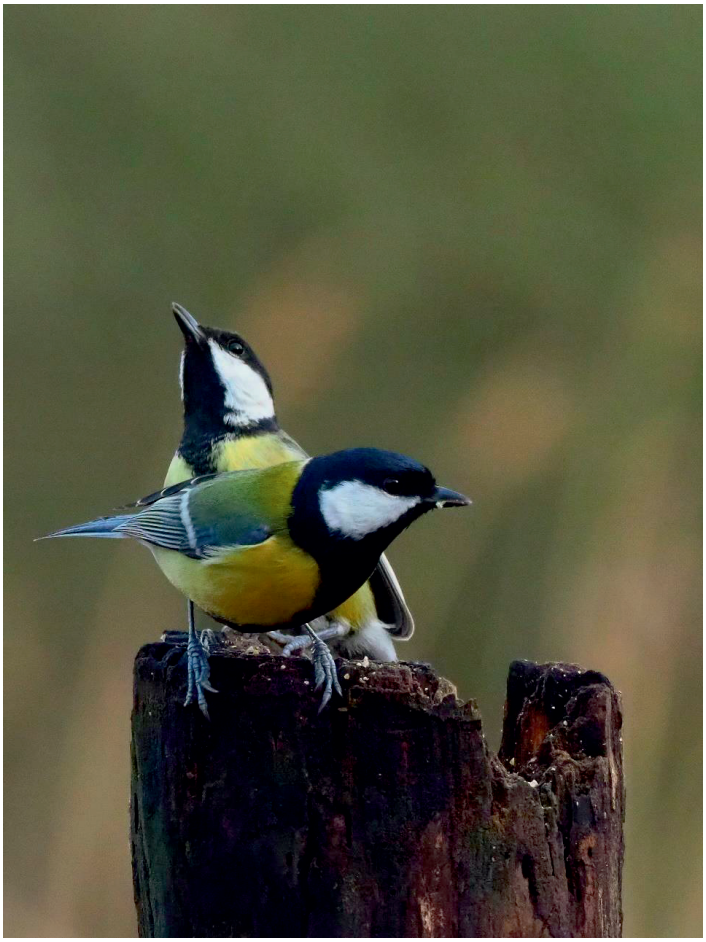
Moving along with the shifting 'bioclimatic envelope' is obviously only possible for a species if suitable habitat is available and the dispersal ability of the species is sufficient. In this respect habitat degradation and loss become doubly relevant since they not only directly threaten population persistence but may also cut off populations from suitable habitat elsewhere. Thomas and co-workers (2004) used the observed 'bioclimatic envelopes' and climate change projections to predict the future distributions of a number of species. Then they used these predicted distributions to assess the likelihood for extinction and found that many species in mountain habitats in the tropics are prone for extinction. The rising temperatures will simply shift the suitable habitat as defined by the 'bioclimatic envelopes' to the mountain tops until no suitable habitat will be left. Since the tropics are biodiversity 'hot spots' this process is predicted to lead to a loss of about 25 % of global biodiversity. One important assumption behind these models is however, that whole species communities will be able to shift at the same rate and that climate zones simply will move northwards without further changes. Unfortunately, both assumptions are unlikely to be true. First, the dispersal ability of species can differ substantially. For example, distances of natal dispersal, i.e. dispersal to the place of (first) breeding after independence, typically range from hundreds of meters to a few kilometres in small passerines as Great Tits, but large predatory birds as Goshawks easily cover tens of kilometres. Second, climate change induced warming trends differ between seasons (Easterling et al. 1997) and regions (Høgda et al. 2001; Giorgi and Lionello 2008). Consequently, even species with a good dispersal ability that could track their 'bioclimatic envelope' will (very) likely face a change in their biotic and abiotic environment even after shifting along with their bioclimatic envelope.

Phenotypic plasticity generally enables populations or species to cope with novel or changed environments. As pointed out above, the current phenotypic plasticity will, however, unlikely to allow perfect tracking of shift in the species' biotic environment (Gienapp et al. 2014). Consequently, neither 'evasion' nor phenotypic plasticity alone seems to be sufficient mechanisms: adaptation by microevolutionary change(s) is necessary to cope with climate change. Unfortunately, while the evidence for climate change-induced changes in wild populations is indisputable, it is mostly unclear whether these changes are phenotypic plastic responses or microevolution (see Plastic versus Genetic changes). Our general understanding of microevolutionary adaptation to climate

change has hence not advanced very much from the statement by Holt (1990) that "There is almost no species for which we know enough relevant ecology, physiology and genetics to predict its evolutionary response to climate change".

PART II

Fitness consequences of reproductive timing



Chapter 3

Testing constraints on the timing of breeding in great tits (*Parus major*) by manipulating pre-breeding food availability

Jip J.C. Ramakers, Phillip Gienapp & Marcel E. Visser

ABSTRACT

Timing of reproduction has crucial fitness consequences for seasonal breeders. In some bird species, temperature rises due to global warming have led to a mismatch between timing of breeding and food abundance, leading to apparent selection for earlier breeding. Yet evidence of a response to selection is rare. One explanation for this is that physiological (food-related) constraints experienced by birds breeding (too) early result in fitness loss, for example due to decreased survival under these harsh conditions. We tested this hypothesis in free-living great tits (*Parus major*) through supplementary feeding prior to egg laying, where females were fed until the first egg (partly fed), fed up until clutch completion (fully fed), or not fed (control). We predicted that fed birds would advance their laying date compared to control birds, and that partly fed birds would suffer reduced fitness compared to both fully fed and control birds as they were manipulated into laying at times where egg production is still too costly to outweigh the fitness benefits of having chicks earlier in the nest. We also predicted that this effect would be mitigated by genotype (predicted breeding values (PBVs) for laying date) since the effect of supplemental feeding was expected to be less strong for with early breeding values. We found that treatment did not affect laying date, nor did it interact with PBV. We found no effect of treatment or PBVs on brood size or chick mass, but control females fledged on average fewer young and had higher brood failures. Lastly, treatment had no carryover effect on feeding activity in the chick-rearing phase. The results suggest that supplemental feeding was ineffective at advancing breeding time and, consequently, that birds are not constrained in the timing of their egg production, although this may have been a direct result of the timing between the onset of feeding and egg-laying. However, both full and partly feeding conferred a fitness advantage in an overall poor breeding season showing that the additional food was affecting the birds in other aspects.

Manuscript submitted

Introduction

Timing of life-history events like reproduction is crucial for fitness, especially when such events are seasonal and highly dependent on the timing of other species in the ecosystem. Seasonally breeding birds face a trade-off between offspring survival value on the one hand, which is achieved by breeding early because of a higher availability of food available to offspring, and their own survival on the other hand, which is benefited by breeding later to allow for sufficient attainment of body condition as environmental conditions improve over the breeding season (see Rowe et al. 1994; Lof et al. 2012). To ensure the maximum amount of food available for their offspring, birds need to time their reproduction so as to match the food peak with the time at which nestling demands are highest (Durant et al. 2007; Reed et al. 2013b). Increasing spring temperatures due to climate change have induced advances in the timing of the food peak maximum but less so in the timing of the resource needs of secondary consumers, resulting in increased mismatches between food availability and resource needs in recent decades (Stenseth and Mysterud 2002b; Visser et al. 2004a; Durant et al. 2007; Thackeray et al. 2010; Thackeray et al. 2016).

To restore the match, birds should advance their breeding at the same rate as the food they use for their offspring (Visser et al. 2010). Phenotypic plasticity (Schlichting and Pigliucci 1998; Pigliucci 2001) is an important mechanism by which birds can adjust their laying date to fluctuating temperatures between years (Brommer et al. 2005; Nussey et al. 2005b; Charmantier et al. 2008). For the great tit (*Parus major*) in the Hoge Veluwe population, however, we know this is not sufficient to restore the match (Visser et al. 2006; Chapter 10); these birds need to genetically alter their mean response to temperature (Visser 2008; Carlson et al. 2014; Chapter 10). Genetically based adaptations, however, have been proven difficult to detect in wild populations in general (Gienapp et al. 2008; Merilä 2012; Gienapp and Brommer 2014; Merilä and Hendry 2014).

One hypothesis as to why some populations do not advance their mean laying date as much as their food relates to the *constraints* hypothesis originally posited by Perrins (1970). As early spring (pre-laying) temperatures have not changed as much as the (post-)laying temperatures, advancing breeding may simply not be possible because females would risk death from costly egg production (Stevenson and Bryant 2000; Visser and Lessells 2001) in harsh conditions with a scarcity of animal protein. Birds that do not breed earlier may therefore be 'adaptively mismatched', as the fitness benefits of breeding early are outweighed by the benefits of breeding at a time when a sufficient body condition has been attained (Lof et al. 2012; Visser et al. 2012). Studies show that alleviating the harsh conditions through experimental food supplementation will improve laying conditions for females but ultimately not increase their fitness when incubation or chick-rearing conditions are still unfavourable upon termination of the food supplementation (Nilsson 1994; Nager et al. 1997; Ramsay and Houston 1997; cf. Harrison et al. 2010). These studies, however, are likely to be inconclusive for several reasons. First, they cannot distinguish between the effects of reducing the fitness costs of laying and the effect of receiving a threshold cue (e.g. amount of food), as posited under the alternative, *cue* hypothesis

(Visser et al. 2012). Second, they provide no direct comparison between birds that encounter continuous advantageous food conditions and birds that encounter such conditions only up to the onset of laying (cf. Nilsson 1994). A third potential drawback is that different genotypes may respond differently to supplemental feeding. For example, the earliest genotypes may advance their laying date in response to extra food to a lesser extent than later genotypes. This would create an interaction between food availability and genotype and this effect may carry-over to subsequent breeding stages. Therefore, particularly small effects of supplemental feeding on laying date and subsequent breeding success may not be detected if the genotype of the focal bird is not accounted for in the analysis.

Fitness costs of breeding too early may arise through reduced survival or nest abandonment during the nest-building or egg-laying phase, but may also carry over to the chick-feeding stage as a delayed cost, since chick provisioning is energetically costly (Te Marvelde et al. 2011). Energetic costs of provisioning increase as food availability decreases (Te Marvelde et al. 2011) and therefore, reduction in body condition incurred from breeding too early may be partly compensated by reducing chick-provisioning rates. It may therefore be insightful to measure chick-provisioning rates as it may explain why broods may ultimately fail.

In this study, we aimed to explore the constraints hypothesis on the timing of breeding in a free-living population of great tits, using experimental manipulation of the birds' food availability and estimation of their breeding value for laying date. Female great tits were assigned to one of three groups: (1) fed until first egg is laid, (2) fed until the clutch is complete, and (3) an unfed control. Potential effects on local survival and other components of fitness were compared among groups and related to their genotype (predicted breeding values or PBVs). Specific questions addressed were: (*RQ1*) What is the interactive effect of (pre-)laying food supplementation and PBV on laying date? (*RQ2*) How is female great tits' breeding success affected by (pre-)laying food supplementation? (*RQ3*) Does (pre-)laying food supplementation affect workload during chick provisioning? We expected (*RQ1*) that supplemental feeding prior to egg laying would advance laying date in both fed groups compared to the control group, but that this effect may wane in birds with low PBVs (Fig. 3.1a). As a result of laying too early when food supplementation ceased, we expected (*RQ2*) that the reproductive success of the partly fed group would decline more steeply with PBVthan the control group, whereas in the fully fed group all birds would do equally well (Fig. 3.1b). Finally, with respect to provisioning rates during the chick phase (*RQ3*), we envisaged two scenarios. As food conditions improve over the breeding season, the partly fed group may increase provisioning rates during the late chick phase to compensate for losses in brood vitality caused by breeding too early. More likely, however, provisioning rates may decrease because the female (and not the male, since he does not produce and incubate the eggs) is in poor shape. Increased provisioning rates may then be reflective of either improved food availability, since there is more food available, or reduced availability, necessitating more frequent foraging trips to reach the energy demands by the chicks.

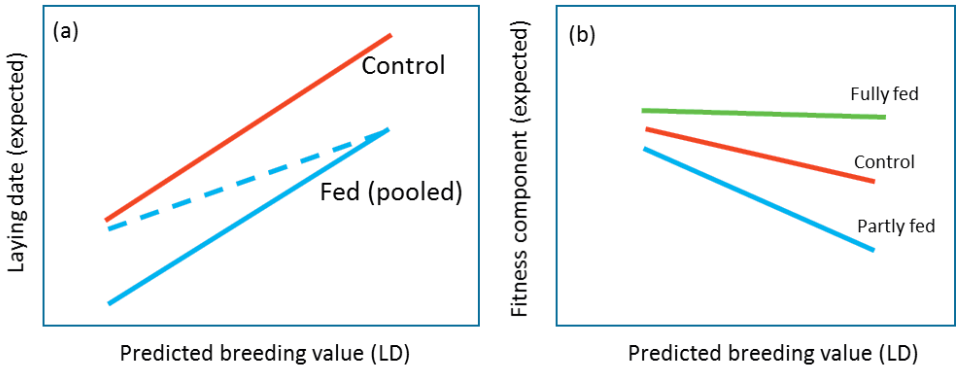


Figure 3.1. *Schematic of expected outcomes of supplemental feeding. (**a**) Birds that were fed prior to laying will advance their laying date, but possibly only birds with a high predicted breeding value for laying date (dashed line). (**b**) Fitness (reproductive success) was expected to decrease with increasing breeding value for laying date, and but this effect was expected to be strong in the partly fed group and weaker in the fully fed group.*

Methods

Field procedures

The study took place at the Hoge Veluwe National Park in the Netherlands (52°02′07″ N, 5°51′32″ E). The 171-ha study area consists of mixed pine–deciduous woodland on poor sandy soils with ~400 nest boxes. The reproductive biology of great tits has been studied here since 1955, with systematically recorded family relationships since 1973. In early March 2015, an evening nest-box check was carried out and all female great tits captured for which pedigree information was available (i.e., excluding immigrants without known breeding phenotypes) were fitted with a leg ring carrying a passive integrated transponder (PIT). This was done to identify individuals that could be included in our experiment and to be able to track down these individuals from the start of nest building later that season. In total, we equipped 53 females with a PIT.

From mid-March onward, nest-building activity was monitored twice a week; once nesting material in a given nest box was found covering the bottom by ≥ 50%, a transponder reader (Dorset ID, Aalten, The Netherlands), set to record birds entering or exiting the nest box with a 1-s interval, was fitted around the entrance of the nest box. These boxes were then visited daily to monitor building progress and determine the 'owner' of the box; if nest building had progressed but no readings had been recorded, we assumed that the box was not being used by a PIT-fitted bird and removed the transponder reader. Once the bird had been identified, it was assigned to one of three experimental groups through random block allocation in three consecutive birds: fed until first egg was laid (the *partly fed* group), fed until the clutch was complete (the *fully fed* group), and not fed (the *control* group). Supplementary feeding started as soon as birds

were assigned to their respective groups. Fed birds were provided daily with ~10 g of mealworms in small, transparent feeding trays suspended within the nest box. This amount should correspond roughly to 100% of their daily energy expenditure, assuming a rough average of 90 kJ d^{-1} throughout the reproductive phase (Te Marvelde et al. 2011; te Marvelde et al. 2012b; Williams 2012) and assuming an energetic content of mealworms of 8.6 kJ g^{-1} (Finke 2002). All birds eventually consumed the mealworms, but we could not quantify the amount eaten by each female reliably as other birds occasionally fed from the mealworms (personal observation). Transponder readers were removed upon the onset of incubation.

All eggs were individually numbered and weighed on the day they were laid. Hatch date was determined by checking the nest daily from the 12th day of incubation onward. Chicks were weighed and ringed at age = 7d (day 0 = hatch date). On day 8, both parents were captured using a spring-loaded trap. Standard biometric measurements (weight, tarsus length, and length of third primary) and blood samples (for a long-term pedigree dataset) were taken and the males were fitted with a PIT. A transponder was installed at the nest box to record feeding activity of both parents up until day 15. On day 15, biometrics and blood samples were taken from the chicks. Nests were checked from a week afterwards to assess the number of fledged young.

Predicting breeding values

For all non-immigrant birds, we estimated their predicted breeding value (PBV) for laying date using our long-term (1973–2014) great tit breeding database for the Hoge Veluwe study population (including only first, unmanipulated clutches in each breeding season). We built an 'animal model' (Henderson 1988; Kruuk 2004) through REML estimation using ASReml-R (Butler et al. 2009; Gilmour et al. 2009). Laying date was the response variable, female age (first-year breeder or older) and year (as a factor) were the fixed effects, and female identity (permanent environment) and a social pedigree were the random effects, following previous model exercises in our population (Husby et al. 2011b; Reed et al. 2016b). We obtained individual point estimates (best linear unbiased predictions: BLUPs) from this model as estimated, 'genetic' deviations from the population-average laying date. We are fully aware that PBVs can come with substantial and potentially non-random prediction errors (Hadfield et al. 2010) but including the additive genetic component directly in our models, by fitting an 'animal model', would have been impractical due to too small sample sizes of this experiment for quantitative genetic analyses.

Statistical analysis

All analyses were done in R 3.3.1 (https://cran.r-project.org/). Throughout, we relied on bootstrapping methods to identify the best models and simulated confidence intervals (CIs) or to calculate *p*-values for likelihood-ratio tests (mixed-effects models). For (generalized) linear models ((G)LMs), we used the package 'boot.stepAIC' (Rizopoulus 2009) to perform an iterative, backward stepwise model selection based on the Akaike

Information Criterion (AIC). This procedure uses sampling with replacement from rows in the data to refit the (G)LM and perform a stepwise AIC selection in each iteration (n = 1000). We identified variables that were selected in > 90% of the iterations as 'candidate' variables. Although this is an arbitrary number, it allows us to explore a set of likely important variables. For models that required inclusion of a random term, we fitted (generalized) linear mixed-effects models ((G)LMMs) in package 'lme4' (Bates et al. 2018) using Maximum Likelihood estimation to allow for calculation of the log-likelihood. We performed a likelihood-ratio test for competing models (i.e. differing in their fixed effects) by parametric bootstrapping (n = 1000) to obtain robust p-values and 95% CIs. From candidate models (i.e. those selected at a rate of > 0.9 in the stepwise procedure or those based on bootstrapped p-values), we simulated posterior estimates β (n_{sim} = 1000) of each variable using the *sim* function in the package 'arm' (Gelman et al. 2016). This procedure simulates the residual standard deviation σ through random draw from the χ^2 distribution and, based on that draw, simulates β coefficients from a multivariate normal distribution with mean $\hat{\beta}$ (i.e. β predicted from the model) and variance matrix $\sigma^2 V_\beta$ (see pp. 142–143 in Gelman and Hill (2006) for details). Note that these simulations are not intended to assess statistical significance, but merely to provide reliable measures of uncertainty.

RQ1: To test the combined effect of treatment and PBV on laying date (RQ1), data were restricted to females with a known PBV and those that started laying at least five days after the treatment allocation (to cover the window between decision making and actual egg laying; C: n = 9; PF: n = 10; FF: n = 10). The reason for the exclusion of data was that, since it takes around five days from the decision to start laying and the actual laying date (Williams 2012), supplemental feeding within these four days would likely not have affected the decision making of the female. Both factors and their interaction term were added as fixed terms in a linear model (LM), in addition to the date of treatment allocation to account for variance in exposure to the treatment.

RQ2: To test the effect of supplemental feeding on fitness, measures of brood success were analysed using (generalized) linear models ((G)LMs) or (generalized) linear mixed-effects models ((G)LMMs). We analysed the effect of treatment×PBV, as well as observed laying date as a covariate, on clutch size (LM with Gaussian errors) and fledging success, analysed in two ways: the probability of fledging at least one chick (GLM with binomial errors with a logit link) and the proportion of fledged chicks from the total original clutch size (GLMM with binomial errors and an observation-level random term to account for overdispersion). As we may expect cascading effects of supplementary feeding on the viability of eggs, we would expect chick weight and growth rates to differ among treatments. We therefore modelled the log-transformed difference in chick weight between d7 and d15 in interaction with treatment in an LMM with brood ID as a random effect. Sample sizes for each of these analyses are given in the results section.

RQ3: To test for a carry-over effect of supplementary feeding (or a lack thereof) on food provisioning in the late nestling phase, we recorded feeding activity from day 8 to 15 post-hatching; the first and the last day with recordings were removed as they provide

incomplete feeding records. The number of nest-box visits of each parent was determined by plotting a frequency histogram of the time difference between consecutive recordings; all data before the end of the first peak (at 17 s intervals) were considered non-independent visits and were therefore discarded. This method was validated using visit data from pied flycatchers (*Ficedula hypoleuca*) in the same study area and using the same equipment for which also video recordings were available (Tomotani et al. 2017). In that study, the cutoff point visually determined in the frequency histogram was the one with the strongest correlation with video-recorded visit rates for both males and females (r^2 = 0.91). We predicted daily feeding activity for both males and females first by constructing an LMM with treatment, sampling day (individual-centred), sex, and their three-way interaction as fixed effects; day number×parent ID (nested within brood ID) was added as a random slope effect to reduce heteroskedasticity in the residuals. As the number of chicks in the nest may a better predictor for parental feeding activity, we then replaced day number in the fixed and random effects structure with the estimated number of chicks (individual-centred). Because we had no daily count of the number of chicks, we estimated it through linear interpolation between the number of chicks present on the first and the last recording day (i.e. on day 15 or earlier when parents abandoned the nest prematurely). The response variable, feeding rate, was log-transformed before analysis to improve normality of residuals and to relax the assumption that, in case of an interaction between treatment and sex, the effect of one in presence of the other is multiplicative.

Results

Effects of food supplementation on laying date

Supplementary feeding prior to egg-laying did not affect laying date, nor did it interact with PBV (appearance rate in bootstrapped AIC model selection procedure: treatment×PBV: 0.40, treatment: 0.52, start date of treatment: 0.71). Laying date, as expected, was affected by PBV (appearance rate: 0.99; posterior estimate of slope [95% CI]: 1.79 [0.43, 3.09]).

Effect of food supplementation on breeding success

Treatment did not affect clutch sizes; that is, receiving food prior to laying and up to incubation did not result in an increased clutch size relative to not receiving any food (GLM: appearance rate for all variables < 0.45; n = 13 (C), 11 (PF), 12 (FF); Fig. 3.2a). Treatment did not affect chick weight at day 7 (LMM: bootstrapped likelihood-ratio test: laying date, p_{boot} = 0.47; treatment, p_{boot} = 0.31; $n_{chicks/nest}$ = 55/9 (C), 70/11 (PF), 63/11 (FF)) or day 15 (laying date, p_{boot} = 0.028, treatment, p_{boot} = 0.16; $n_{chicks/nest}$ = 15/4 (C), 42/9 (PF), 47/10 (FF); Fig. 3.2b). Interestingly, only 3 out of 13 nests from the C group reached fledgling, compared to 8 out of 11 (PF) and 8 out of 12 (FF) from the fed birds (GLM: variable appearance rate: treatment×PBV: 0.44; laying date: 0.26, PBV: 0.56, treatment: 0.91; Fig. 3.2c). Likewise, the proportion of fledged chicks from the total clutch size was lowest for

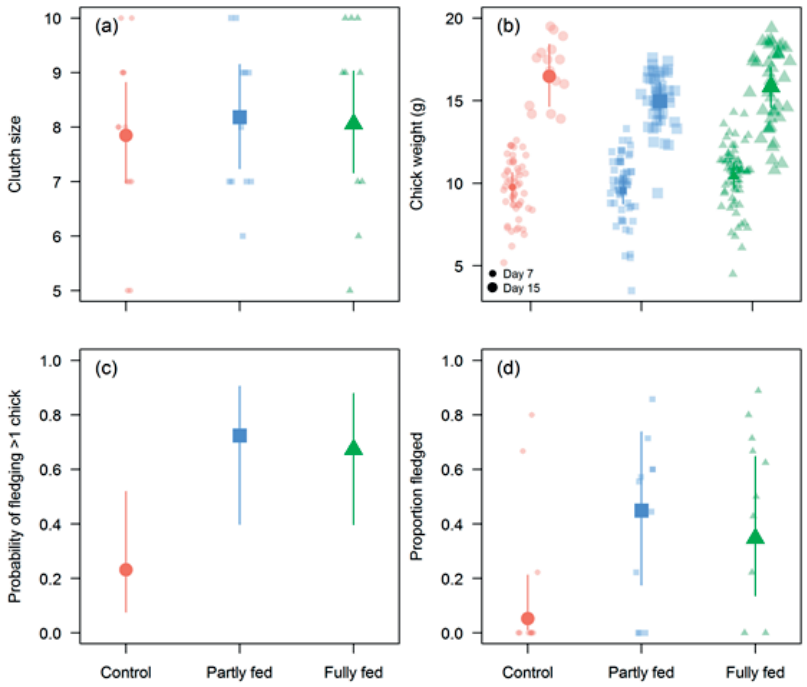


Figure 3.2. *Different fitness proxies (posterior medians and 95% CIs resulting from simulation of coefficients) as a result of supplemental feeding of female great tits. (**a**) Clutch size; (**b**) chick weight at day 7 and day 15; (**c**) probability of fledging at least one chick; (**d**) proportion of chicks fledged relative to the total clutch size.*

the control group compared to the two fed groups (GLMM, bootstrapped likelihood-ratio test: treatment×PBV, $p_{boot} = 0.78$; PBV, $p_{boot} = 0.61$; laying date, $p_{boot} = 0.27$; treatment, $p_{boot} = 0.013$), with no difference between the two fed groups (Fig. 3d).

Effect of food supplementation on parental workload

The parents' visiting rates during chick feeding (d9–d14) was not affected by treatment (LMM: $p_{boot} > 0.27$; $n_{broods} = 5$ (C), 10 (PF), 9 (FF)). The number of visits was, however, related to both the age of the chicks (day number) and the number of chicks estimated to be present in the nest (estimated from different models) in a sex-dependent manner (chick age×sex: $p_{boot} = 0.002$; number of chicks×sex: $p_{boot} = 0.003$). In both models (Fig. 3.3a and b), male feeding activity declined with chick age and increased with chick number (slope, on a log scale, for chick age: –0.05 [–0.07, –0.02]; for number of chicks: 0.13 [0.05, 0.22]), whereas female feeding activity was unaffected by either (chick age: 0.01 [–0.01, 0.03]; number of chicks: 0.02 [–0.07, 0.11]). A model with a threey-way interaction between chick age, sex and treatment was not supported ($p_{boot} = 0.25$). There was a marginally significant

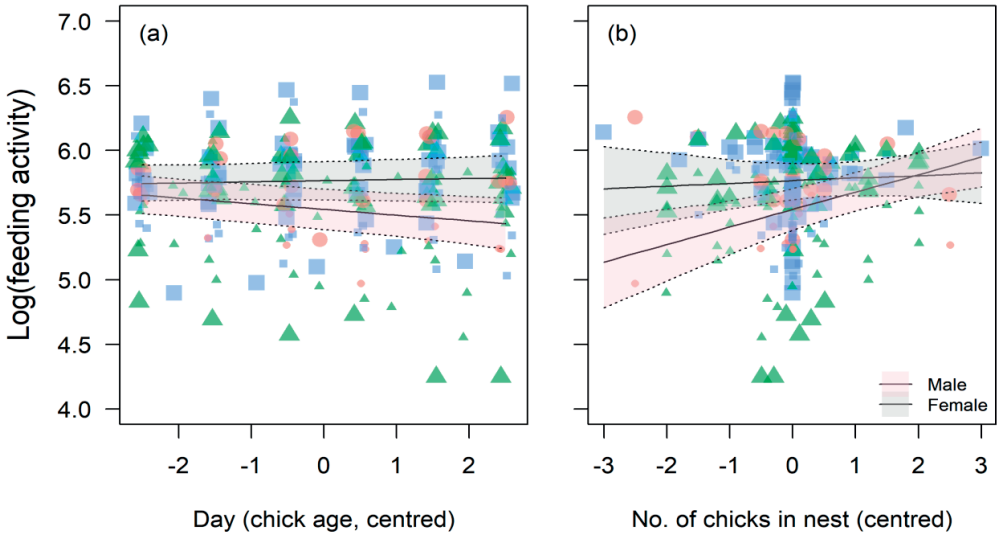


Figure 3.3. *Daily feeding activity (log transformed) of male and female great tits against centred chick age (**a**) and the centred estimated number of chicks in the nest (**b**). Lines are posterior estimates (+ 95 CIs) resulting from simulating coefficients of a LMM with Day×Sex (**a**) and Chicks×Sex (**b**) as fixed effects, and Day×Individual ID (**a**) and Chicks×Individual ID (**b**) (both nested within brood ID) as random effects. Data points represent daily, individual feeding frequencies (large symbols: females; small symbols: males). Different symbols and their colour denote the different feeding treatments (red circles: control; blue squares: partly fed; green triangles: fully fed).*

interaction between number of chicks, sex and treatment (p_{boot} = 0.065), but this did not lead to different slopes of feeding activity against the number of chicks between treatments in either sex (slopes difference males, PF: –0.019 [–0.354, 0.299]; FF: –0.044 [–0.328, 0.231]; slpe difference females, PF: –0.001 [–0.145, 0.177]; FF: 0.075 [–0.080, 0.233]).

Discussion

We used supplemental feeding to test whether food availability prior to egg-laying can act as a constraint in the advancement of reproduction in a seasonally reproducing passerine bird. Great tits that were fed inside their nest boxes only prior to laying were expected to advance their egg-laying date compared to control birds and thereby pay fitness cost once these benign food conditions were removed (i.e. when the clutch still had to be laid and incubated). We also expected that this fitness cost would be lessened in birds with an early genotype (PBV). We found that treatment did not affect laying date and could not verify the prediction that birds manipulated into laying earlier would pay fitness costs. In fact, birds that received any food (either prior to laying or up to and including incubation) had a higher reproductive success than birds that did not receive any food at all, meaning that partial supplemental feeding increased, rather than reduced, the fitness of the brood. This

could indicate that many birds were constrained to successfully reproduce due to overall poor food conditions but that this constraint was not related to the timing of egg-laying.

Demonstrating any effect of supplementary feeding on breeding success was conditional on laying date being shifted as a result of supplementary feeding in the first place. Achieving this proved to be logistically challenging because it required the identification of females well before the onset of laying to be able to assign them to a treatment. Although we captured females in their night roosts in early March (to provide them with a PIT), females rarely use their exact same roosting box as their breeding nest box, so we had to rely on nest-building activity (automatically recorded using transponder readers) to identify where birds would be likely to breed. This, however, limited the amount of time we had between treatment allocation and the onset of laying. We allocated treatments as soon as we knew which female was building a nest, which in the bulk of the population occurred relatively close to actual laying, as temperatures rose rapidly in the week prior to the first egg. Food supplementation therefore started likely too close to the onset of laying to have an effect (mean ± SD days between treatment allocation and laying: 9.36 ± 7.83 days; median 7 days), seeing as the decision to start laying is made around five days prior (Williams 2012). Alternatively, if birds were genuinely not constrained by food availability early in the season, we would not expect any advance in laying date to take place in the first place, although advancement is a quite general outcome of supplemental feeding experiments (e.g. Verhulst and Nilsson 2008; Ruffino et al. 2014). Nevertheless, even if laying date was significantly affected by our treatment, the effect size would necessarily have to be small, too small to affect subsequent reproductive performance in the predicted direction.

Our prediction was that, if providing supplemental food during egg production lifted energetic constraints, cessation of supplemental feeding upon the start of incubation in the 'partly fed' group would lead to reproductive costs (with potential nest desertions) in females with high breeding values for laying date. Contrary to this prediction, birds in the control group (regardless of their PBV for laying date) had lower reproductive success, i.e. abandoned nests more often and fledged—proportionately—the fewest offspring (Fig. 3.2c and d). Supplemental feeding is known to be of particular benefit when conditions are poor (Ruffino et al. 2014). Although we have no sufficient data as of yet to quantify food conditions during the laying period, the availability of caterpillars later that season (during the chick-feeding stage) was exceptionally poor (see Chapter 5). The lower reproductive success in females in the control group was then most likely a result of overall poor breeding conditions, and any additional food (whether up until the onset of laying or up until completion of the clutch) would confer a benefit to females, potentially by compensating energy expenditure associated with food-searching bouts during the chick-feeding stage (e.g. Grieco 2002; Te Marvelde et al. 2011).

Supplemental feeding did not affect feeding frequencies in the parents. However, in contrast to females, males appeared to decrease feeding activity as the estimated number of chicks in the nest declined (Fig. 3.2b). Although it is not certain why, males have been shown to be more likely than females to decrease provisioning efforts or even abandon nests under unfavourable conditions (Sasvari 1986; Sanz et al. 2000; Griggio and Pilastro

2007). Although sample sizes are too low to officially test, the observed decrease in breeding activity in males seems to be not only a between-brood but also a within-brood effect. This means that males may anticipate brood failure as the number of chicks in their nests declines and therefore steadily reduce their feeding efforts. We need to interpret the lack of a treatment effect with caution since only 3 out of 13 control broods reached fledging. However, the near-significant three-way interaction between treatment, sex and number of chicks (p_{boot} = 0.065) is suggestive of such an effect. The less steep slopes estimated for the PF and FF groups compared to the controls could reflect the fact that supplemental feeding (irrespective of the feeding regime) indeed conferred a benefit to females (and potentially males) and negated the energetic costs associated with breeding in an overall poor environment by enabling (or motivating) them to feed more frequently regardless of the number of chicks in the nest.

The timing of breeding of organisms in seasonal environments has been a topic of great interest among (evolutionary) ecologists (Durant et al. 2007; Visser 2008; Both et al. 2009b; Thackeray et al. 2010; Kharouba et al. 2018). Whether a population's capacity to adapt to climate change is limited by energetic constraints (Perrins 1970; Stevenson and Bryant 2000) or lack of shift in the cues that are used to time reproduction (see Visser et al. 2012), or a combination of both, remains a relevant question. To test this, we first need rigorous experiments to unveil the causal effects of timing of breeding on parameters of fitness, but achieving this is not straightforward (Verhulst and Nilsson 2008). For this, experiments are required that are able to manipulate the timing of breeding (e.g. avian laying date) without carry-over effects on e.g. body condition during subsequent reproductive stages (Verhulst and Nilsson 2008), which have been shown to be associated with high energetic and fitness costs (Visser and Lessells 2001; Te Marvelde et al. 2011). A potentially useful way is to create experimental strains of birds genomically selected for their laying date, release them into the wild, and study their reproductive behaviour. An experiment of this kind is currently in progress (Verhagen et al. in review) and the first, tentative results from the wild are presented in Chapter 4. Feeding experiments like the one discussed in this chapter and elsewhere (e.g. Verhulst and Nilsson 2008) may provide useful insights into the causal effect of food on fitness, but only when carried out over multiple seasons and hence environments; yet still there will be potential carry-over effects that cannot be accounted for. For now, therefore, the answer to the question posed in this paper remains an elusive one.

Acknowledgements

This work would not have been possible without the board of the Hoge Veluwe park, who granted us access to the forest to conduct the experiment. Jan Visser and Louis Vernooij carefully managed the long-term database. We are very grateful to Cynthia Lange, Henri Bouwmeester and Emma Rietveld for their invaluable help in collecting the data. This work was funded in part by an ERC Advanced Grant (339092 - E-Response) to M.E.V.

Chapter 4

Offspring viability and laying dates in the wild of great tits selected for seasonal timing of reproduction

Jip J.C. Ramakers, Phillip Gienapp & Marcel E. Visser

ABSTRACT

Seasonal timing of avian reproduction is an important determinant of reproductive success. Climate warming has led to directional selection on timing in great tits (*Parus major*), but the (phenotypic) response has been small. Two (non-mutually exclusive) hypotheses state that (1) birds are constrained to breed earlier due to high fitness costs, and (2) birds are missing the relevant cues to start breeding earlier. To distinguish between both hypotheses we need to manipulate egg-laying date in the wild to estimate the causal effect of laying date on reproductive success and whether this reproductive success is increased or reduced for birds manipulated to breed earlier than the natural population. We manipulated laying date by releasing birds from selection lines for early and late reproduction (through genomic selection) into the wild. Here, we report the first results of the effects of this selection on early-life fitness in the offspring and on their subsequent laying dates as adults when they recruit into the breeding population. Great tits from early and late selection lines produced eggs in aviaries, which were fostered in our wild study population using a randomized clutch-swapping experiment in 2017 and 2018. Eggs from the early and late line did not differ in quality: neither fledgling success nor fledgling weight differed between these lines. However, both fledging success and fledgling weight differed between the foster nests and control (unmanipulated) wild nests. These differences were, however, likely too small to be biologically meaningful, meaning that offspring from selection lines do not differ in early-life fitness from birds from the natural population. Only 11 nestlings in 2017 recruited to the breeding population in 2018, of which five were female (two from the early and three from the late selection line). Although we could not statistically test the difference in mean realised laying dates between lines, early-line females bred earlier than the late-line females, following expectations. We conclude that offspring originating from the selection lines had a similar start of their lives compared to the natural population, and tentatively conclude that genomic selection for extreme laying dates resulted in the expected phenotypes. Multiple years of data collection are, however, necessary to draw conclusions as to the causal effect of laying date on reproductive success, as well as which mechanism(s) constrain(s) birds in advancing their timing of breeding.

Unpublished results

Introduction

Timing of reproduction is key for reproductive success in seasonally breeding birds. Generally, late reproducers have lower reproductive success, either because they lay smaller clutches or have fewer fledglings due to deteriorating food conditions (Rowe et al. 1994; Dalhaug et al. 1996; Winkler and Allen 1996). Breeding too early, on the other hand, may also bear fitness consequences due to increased costs of laying (and incubating) under harsh conditions (Rowe et al. 1994; Bêty et al. 2003) (see below). Seasonally breeding organisms are thus under selective pressure to breed at the right time to maximise their lifetime reproductive success (Rowe et al. 1994; Chevin et al. 2015; Gamelon et al. 2018).

As an important ecological driver, climate change is altering environmental conditions globally (Root et al. 2003; Parmesan 2006). One important ecological effect is on phenology (Chapter 2), with unequal shifts between trophic levels leading to a mismatch between consumer and resource phenology (e.g. Thackeray et al. 2010; Thackeray et al. 2016; Kharouba et al. 2018). In seasonal birds, this has led to the disruption of the needs of offspring (nestlings) and their main (often invertebrate) prey (Both and Visser 2001; Thomas et al. 2001; Both et al. 2006) and concomitant (apparent) selection for earlier breeding (Visser et al. 1998; Reed et al. 2013b; Marrot et al. 2018). Although many populations advance their laying date under warmer conditions through phenotypic plasticity (Crick et al. 1997; Cresswell and McCleery 2003; Nussey et al. 2005b; Porlier et al. 2012), evidence for an adaptive genetic change in laying date (i.e. an advancement in the reaction norm describing phenotypic plasticity; see Chapter 10) remains scarce (Charmantier and Gienapp 2014; Merilä and Hendry 2014). Such an advancement would be essential for population persistence under increasing mismatch due to further global warming, because the current level of plasticity is not sufficient to bring the population to the new optima (Visser 2008).

One hypothesis as to why some populations do not advance their mean laying date in the first place relates to the availability of food in the pre-laying season (Perrins 1970; see also Chapter 3). Advancing breeding may simply not be possible because conditions are too harsh to survive costly egg production (Stevenson and Bryant 2000; Visser and Lessells 2001). This *constraints* hypothesis therefore suggests that birds that do not advance breeding may be 'adaptively mismatched', as the fitness benefits of breeding at a time when a sufficient body condition has been attained may outweigh the benefits of breeding in synchrony with the food peak if this means they have to breed under adverse conditions (Lof et al. 2012; Visser et al. 2012). A second hypothesis, the *cues* hypothesis, states that birds cannot breed adaptively earlier because the essential cues to do so have become inaccurate due to climate change (Visser et al. 2012). For example, if the temperatures later in spring that drive the timing of food availability—which is when selection via nestling survival takes place (Visser et al. 2006; see also Chapter 5)—increase consistently, but those in early spring that determine the onset of laying do not, the mismatch between predator and prey phenology will increase. Genetic adaptation is necessary to adapt to these novel conditions (Visser 2008), either by changing the trade-off between early laying and costs of reproduction (allowing birds to somehow better cope with harsh conditions) or by

adjusting cue sensitivity. Whichever the route adaptation should or will take, it will be a slow process (Gienapp et al. 2006; Charmantier and Gienapp 2014; Chapter 10).

To gain insight into the likely processes underlying adaptation in breeding time to novel environmental conditions, we first need a thorough understanding of how the timing of breeding determines reproductive success. Many food-supplementation experiments have been conducted in avian populations during the pre-laying phase, and the general outcome of these studies is that supplemental feeding has the potential to advance laying (Ruffino et al. 2014). Importantly, however, supplemental feeding changes the physiological state of an individual (Verhulst and Nilsson 2008) and therefore the effect of advancing laying on reproductive success is not straightforwardly decoupled from the effect of changing the physical condition of females (see also Chapter 3). A clean manipulation of laying date has been undertaken In Dutch great tits, where the expectations of the birds in the current year were manipulated through food supplementation during the chick-feeding phase in the previous year (Gienapp and Visser 2006). Although this experiment effectively advanced laying date in one of the two studied populations, sample sizes were too low to draw safe conclusions about the fitness effects of advancing. Leptin implantation experiments to manipulate females' perceived body condition (Te Marvelde and Visser 2012) and photoperiodic manipulation to stimulate follicle growth (Te Marvelde et al. 2012a) were not successful at manipulating laying dates in the wild. A clean, hitherto untested method to cleanly manipulate laying date is to artificially select on laying date using genomic selection tools, and use their progeny to measure phenotypes (laying date) and reproductive success in the wild. This is possible because laying date is a heritable (polygenic) trait (Van der Jeugd and McCleery 2002; Gienapp et al. 2006). Such a selection experiment was recently completed in great tits (*Parus major*) originating from the Dutch Hoge Veluwe population (see Gienapp et al. 2019; Verhagen et al. in review for details of this selection experiment). Female offspring of birds that were selected to breed either early or late in outdoor aviaries exhibited extreme genomic breeding values and concomitant different phenotypes between early- and late-line birds in captivity (Verhagen et al. in review). When offspring from these lines are released into the wild and survive to breed, they will thus likely exhibit extreme phenotypes, providing a 'clean' test of the causal effect of laying date on reproductive success (but see Results and discussion for potential caveats).

In this study we give an overview of the first tentative results of the effects of the selection experiment on parameters of fitness and phenotypes (laying dates) in the wild. Eggs from the F3 generation of both the early and late selection lines, which constitute the most extreme genotypes from the selected population, were translocated and fostered with breeding pairs in the wild in the breeding seasons of 2017 and 2018. These eggs were incubated and raised by foster parents, and a portion of the fledged chicks were expected to recruit into the breeding population the following year. We therefore have data on 'early-life fitness' (i.e. the proportion of eggs from the selection lines ending up as fledged chicks) for both years, and have phenotypic data (laying date) and reproductive success for 2018 (i.e. recruits from the 2017 eggs). The objectives of this (ongoing) study are to test (i) whether eggs from selection lines have different fitness compared to eggs from the wild population during early life (up to fledging) and whether there are between-line

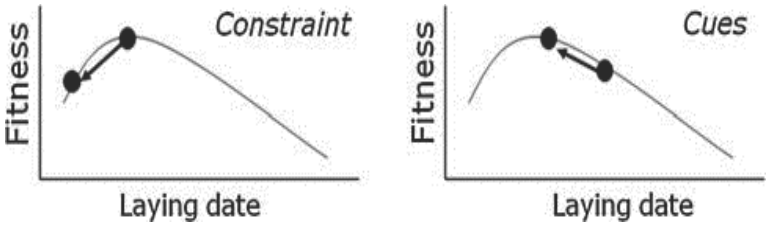


***Figure 4.1**. Schematic of the prediction of the effect of experimentally advancing laying date (arrow) on the fitness of individual females. If physiological constraints inhibit advancement in natural conditions, an experimental advancement will decrease fitness; if the lack of appropriate cues inhibit advancement, an experimental shift will increase fitness. Schematic from Visser et al. (2012).*

differences in this respect, and (ii) whether recruits from the two lines had distinctly different phenotypes and reproductive success when breeding in 2018. With respect to the first objective, we expected that the selection experiment had no effect on the 'vitality' of the eggs and that experimental and wild eggs and fledglings had an equal start in their lives. With respect to the second objective, we first predicted that early-line recruits exhibited laying dates at the lower extreme end of the laying date distribution, whereas we expected the opposite for late-line recruits. Second, among early-line recruits, we had two different expectation with respect to reproductive success, related to the *constraints* and *cues* hypothesis (see Visser et al. 2012). If the *constraints* hypothesis was true, birds that were 'genetically forced' to breed early would pay a fitness cost (Fig. 4.1, left). That is, if conditions were harsh (cold, little food) they would be more prone to desert or die in the process of nest-building or egg-laying. To this end, we monitored nest-building using automated transponder readers to track down recruited females (who carried a transponder, see Methods) during the pre-laying period. If the *cues* hypothesis was true, we may expect an increase in fitness as birds advanced their laying date, because they would be in better synchrony with the food peak later that season (Fig. 4.1, right). Ultimately, however, we will need multiple years of observations (with concomitant variation in environmental conditions) before we can answer the question of which of the two hypotheses most likely explains the observed lack of a shift in laying dates in great tits.

Methods

Selection experiment aviaries

A full and detailed description of the selection experiment is given in Verhagen et al. (in review) and Gienapp et al. (2019). Briefly, nestlings from a wild population of great tits at the Hoge Veluwe national park, central Netherlands (52°02′07″ N 5°51′32″ E), were collected in 2014 from the extreme ends of the laying-date distribution from parents (P, parental generation) with low and high pedigree-derived breeding values for laying date

(i.e. 'early' and 'late' birds). The nestlings (F1 generation) were hand-raised at the institute and served as the parental generation for the next generation (F2; 2015), who in turn served as the parental generation for F3 (2016), which in turn produced F4 (2017 and 2018). Nestlings from each generation, as well as ~2000 birds from the original population, were genotyped on a 650 kSNP chip, which allowed the estimation of genomic breeding values (GEBV). These GEBVs were used to select amongst the most extreme genotypes with respect to laying date in each generation, which were paired up disassortatively within selection lines (early vs. late) to maintain genetic variation. Every breeding season, we had approximately 100 breeding pairs in outdoor and climate-controlled aviaries. Selection on GEBVs was moderately strong (Kingsolver et al. 2001), with standardized selection differentials ranging from –0.554 to –0.703 in the early line from the parental through to the F2 generation, and from 0.528 to 0.658 in the late line (Verhagen et al. in review).

Field sites

Fieldwork mainly took place at the Hoge Veluwe national park (HV). The 171-ha study area consists of a mixture of coniferous and deciduous woodland and has ~400 nest boxes with ~100–150 great tit breeding pairs each year. The remainder of the fieldwork was done on a great tit population in a nearby deciduous woodland in Bennekom (BE; 52°00'02" N 5°41'30" E), which has ~200 boxes and is regularly used as foster population for incubation and chick rearing.

General field experimental procedures

In the spring of 2017 and 2018, eggs from the selection lines were brought to the wild to study (i) the effect of the selection experiment on early life fitness and (ii) the reproductive consequences of early or late breeding in recruiting females.

In both years, breeding activity in both selection lines was monitored in the aviaries from March onward. Breeding pairs (F3, producing the F4 generation) were provided with moss and hairs, and nest-building activity was monitored initially, weekly, then twice per week and more frequently as nest building progressed. Eggs were collected daily from the nest boxes in the aviaries and individually marked. Simultaneously, nest-building and egg-laying activity was monitored in both field sites. Eggs from aviaries were brought to the wild as soon as complete clutches were available. When possible, eggs from the two selection lines (early and late) were combined into a single clutch such that a given clutch contained an equal number of early and late eggs. However, when eggs from only one of the lines were available (for example, early in the breeding season most available eggs came from the early line, whereas in the tail of the breeding season mostly late-line eggs were available), we created single-line clutches (24% and 6% of the clutches in 2017 and 2018, respectively). The composite clutches were fostered with wild breeding pairs in BE, whose original clutch was removed (and whose eggs were used in another project on egg yolk hormones). BE foster parents incubated the eggs for a minimum of five days, after which we could determine the viability of eggs (i.e. through visibly developing blood vessels) before moving them to their final nest box in HV.

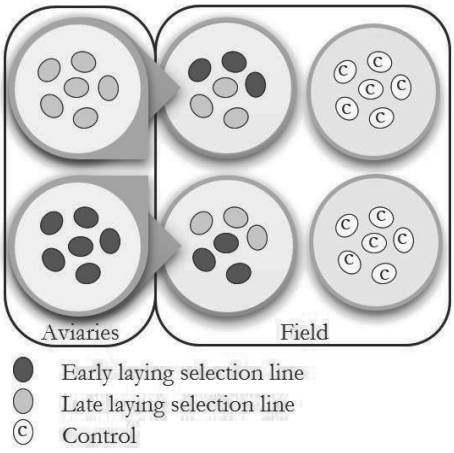


***Figure 4.2**. Schematic of the experimental design. Eggs from the two selection lines (in aviaries) were made into composite clutches, which were temporarily fostered with breeding pairs in BE (not shown) and subsequently fostered with breeding pairs in the Hoge Veluwe. Drawing by M. Swinkels.*

Fostering of F4 clutches from the selection lines with breeding pairs in HV was done experimentally in 2017 and 2018, with slight differences between the two years. In 2017, incubating females at HV were randomly allocated to one of four experimental groups, consisting of two experimental nests and two control nests (Fig. 4.2). Grouping was done in blocks of four to ensure each treatment was equally represented throughout the breeding season. When a brood had been incubated for a minimum of five days, eggs were removed and replaced with an approximately equally sized composite clutch from the BE foster pairs, which also had been incubated for five days. The reason we swapped after HV females had been incubating for a minimum of five days was that birds were prone to nest desertion at very early stages of incubation (personal observations in 2016). The control nests in HV were not swapped and left alone, aside from standard fieldwork procedures (see below; no swaps were performed between control nests because these nests were to remain undisturbed for the long-term population study). In 2017, 43 clutches were fostered in HV.

In 2018, the procedure in HV was different. Again, broods were allocated to one of four treatments (Fig. 4.2), but this allocation started as soon as the female had laid the third egg. Nests were checked daily to ensure the date of the third egg could be reliably determined. Upon finding the third egg, four dummy eggs were placed into the two treatment nests; this was done to advance the onset of incubation to increase synchrony with the aviary population. As in 2017, clutches were removed and replaced with composite clutches from BE after at least five days of incubation; the eggs that came from foster parents in BE had been incubated > 5 (usually 8–10 days), to further advance the hatch date of the fostered clutch in HV, for the same reason. This resulted in an advancement of 1.7 (0.2, 3.2 (95% CI)) days. In 2018, 53 clutches were fostered in HV.

Birds that fledged from experimental nests in 2017 and recruited into the breeding population were tracked from mid-March 2018 onward to monitor breeding activity, daily energy expenditure (still ongoing so not reported here) and reproductive success (see below for details). Because fledglings from experimental nests had received a passive integrated transponder (PIT) tag (see below), we were able to track the nesting behaviour

of these birds by placing automated transponder readers (DorsetID, Aalten, the Netherlands) at each nest box with nest-building activity, mounting the antenna around the entrance hole. Using activity records and the progression of nest material across time, we were able to identify whether a given nest belonged to one of our recruited experimental birds. If this was the case, this nest was exempt from the treatment allocation described above. The reason why we started monitoring activity at the time of nest-building and did not wait until we could identify the breeding female on the nest was that females from the early line that bred too early may potentially abandon the nest or die in the process of building or egg-laying. We needed to identify these birds as they would represent a non-random portion of the breeding population and as we were specifically interested in the fitness effects of breeding (too) early.

Brood swap experiment: measuring morphology, fledging and fitness

Once clutches in experimental nests had been swapped, experimental and control broods alike were subjected to standard field procedures. Hatch date of eggs was checked daily around the expected date of hatching. At day 8 after hatching (day 0 = hatch date), the (foster) parents were captured and identified at the nest and nestlings were ringed. On day 15, i.e. close to fledging, nestlings were blood-sampled, weighed (to the nearest 0.1 g) and measured (length of third primary (P3) and tarsus). Each nestling originating from the selection lines (all chicks in experimental nests) was given a PIT tag to allow studying fledging behaviour and tracking of individual birds in subsequent breeding seasons. The number of fledged chicks was determined for both experimental and control nests.

Identification of genetic parents fostered nestlings

Blood samples taken from the nestlings were used to identify the aviary breeding pair that produced the nestling. DNA was extracted from a Queens buffer using the FavorPrep 96-well Genomic DNA Kit centrifugation process (Favorgen Biotech Corp., 2009). PCR was executed using the protocol as described in Saladin et al. (2003). To dilute the PCR-product, 190 µl MilliQ was added to the PCR-plates after PCR and 2 µl was transferred to an ABI-plate. 9 µl of a LIZ-Hi-Di mixture (9 µl LIZ-sizer + 1ml Hi-Di Formamide) was added to each sample and the plates were analyzed with the ABI 3130. Microsatellites, segments of repetitive DNA sequence with a high mutation rate, were marked on five different loci and scored with the help of GeneMapper 5®, to later determine the genetic parents based on similar microsatellites.

Data analysis: brood swap experiment

To test the effect of the selection lines on parameters of brood fitness (i.e. combined offspring vitality) we tested the effect of (i) the brood swapping per se (i.e. the between-brood difference between control and experimental nests) and (ii) the effect of selection lines (i.e. the within-brood difference between 'early' and 'late' chicks). Because not all eggs could be retrieved, or because dead chicks were removed by the parents, we could

not estimate fitness components for each individual egg; we therefore viewed fitness as a property of the brood as a whole (between-brood effect) or of each selection line (within-brood effect). The fitness parameters we studied were (1) the proportion of chicks fledged relative to the total clutch size and (2) the total number of fledged chicks. We furthermore investigated variation in (3) fledgling weight as an indicator of offspring vitality (see Chapter 8). Each trait was investigated between control and experimental nests (treatment 'experiment') or between lines within experimental broods (treatment 'selection line').

For each of the analyses, we constructed linear models (LMs) or generalized linear models (GLMs) with the following error structures: (1) binomial with logit link and (2) Poisson with log link and (3) Gaussian with identity link. In the analysis comparing between treatments (selection lines) within broods, we added a random effect of 'brood', with treatment nested within it. In analysis (2) between broods, but not within broods, the number of fledged chicks was underdispersed, so we fitted a Generalized Poisson GLM (package 'VGAM'; Yee 2010). In analysis (3) we always included a random effect of brood because the individual chick was the level of observation. Aviary of origin or maternal identity never explained variation in the traits of interest, so both random factors were dropped from analysis. Besides treatment, we included the covariates hatch date (to capture seasonal trends) and final (i.e. after swapping) clutch size. Year (2017 or 2018) was added as a two-level factor. In all models, we fitted an interaction between hatch date and year.

We used Akaike's Information Criterion corrected for small sample sizes (AICc; Burnham and Anderson 2002) to compare six models: (i) the base model excluding treatment (i.e. experiment (between broods) or line (within broods)), (ii) a model with treatment, (iii) a model with an interaction between treatment and hatch date and (iv) between treatment and year, (v) a model containing both interactions and (vi) and a model containing the three-way interaction between year, treatment and hatch date. The most parsimonious model within 2 ΔAICc units from the model with the lowest AICc value was considered most plausible (but see discussions in Richards 2005, 2008; Burnham et al. 2011 for caveats of using this threshold). Uncertainty in model estimates was estimated through bootstrapping the 95% confidence intervals with 1000 iterations.

Finally, for 2018 only, we tested whether the number of fledglings (generalized Poisson) and fledgling weight in experimental vs. control nests showed a non-linear relationship with hatching date as a result of the experimental advancement of the start of incubation. Clutch size (i.e. the number of eggs a female laid, irrespective of the added dummy eggs) and final clutch size (i.e. the net number of eggs in the nest after egg-swapping) correlated strongly ($r^2 = 0.64$), so we only tested final clutch size as an additional covariate, besides experiment and hatch date. We then added a quadratic term of hatch date and an interaction between experiment and hatch date and [hatch date]^2.

Results and discussion

Egg-swapping experiment (2017 and 2018)

We had data for 38 experimental nests in 2017 and 42 in 2018, with 86 and 42 nests marked as control, respectively. The total number eggs brought to the wild was 389 in 2017 and 461 in 2018 (including some clutches that were deserted by the foster parents after swapping). Of these, 207 and 315 chicks, respectively, survived until fledging.

Fledgling success. Since we did not have detailed data on the number of eggs that hatched (but a proxy, i.e. how many chicks alive at day 8), we analysed the number of nestlings that fledged in proportion to the final clutch size. The proportion of fledged chicks (PFC) depended on both year and hatch date. AIC_c model selection revealed that treatment (control vs. experimental) in the between-nest analysis interacted with both year and hatch date (Table S4.1a). That is, PFC in experimental relative to control nests differed between years, and the dependence on hatch date differed between treatments (Fig. 4.3). In 2017, PFC in the control group dependent only very weakly on hatch date (slope control nests: 0.001 [–0.018, 0.018]), whereas the slope for experimental nests deviated from this (slope deviation: –0.035 [–0.067, –0.004]). In 2018 there was no statistically discernible difference in slopes between treatments (slope control nests: –0.010 [–0.135, 0.109]; deviation slope experimental nests –0.141 [–0.342, 0.046]). A similar pattern was found for the absolute number of fledglings, although the effect of treatment did not interact with year (Table S4.1b). It should be noted, however, that some nests in 2017 were fostered in very late

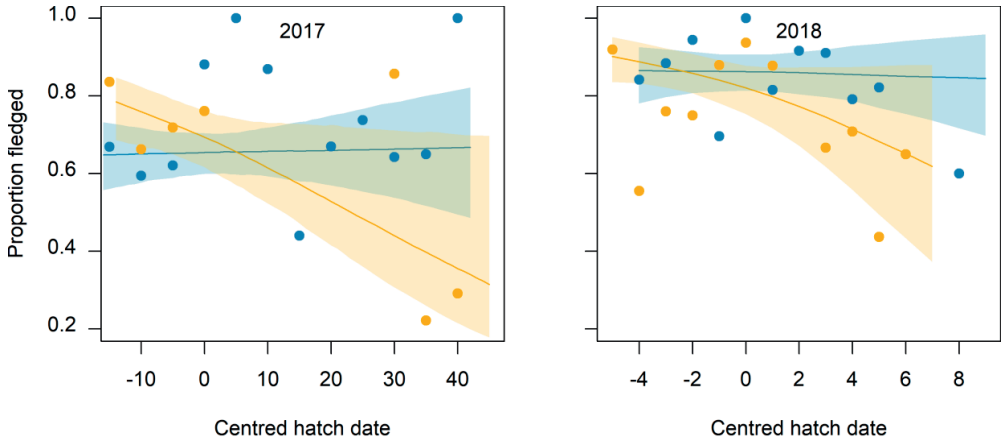


***Figure 4.3**. Proportion of chicks (of total clutch size) fledged, as a function of mean-centred hatch date in 2017 and 2018 in the between-brood comparison. Lines and shading are model predictions and 95% bootstrapped CIs (blue: control nests; orange: experimental (swapped) nests). Points are mean proportions, binned by 5 days in 2017 and by 1 day in 2018, for visual purposes.*

broods (perhaps replacement clutches from birds with a failed first), where any disturbance at the nest on top of the swapping in some cases caused the foster parents to abandon the nests prematurely; this may have exacerbated the negative trend with hatch date in the experimental group and therefore 'pull' the line in Fig. 4.3 down.

Within nests (between selection lines) we found no main effect of treatment (late vs. early line), nor any interactions with year or hatch date for PFC or the number of fledglings (Table S4.2).

The intended advancement in hatch date in 2018 as a result of adding the dummy eggs during the laying phase in experimental nests did not result in fewer fledglings in the earliest broods (Table S4.3a).

Fledgling weight. The AIC_c comparison of mixed-effects models of fledgling weight revealed, aside from an interaction between hatch date and year, an additional interaction between experiment (swapped vs. control nests) and hatch date in the between-nest analysis (Table S4.1c). That is, both the mean fledgling weight across hatch dates (intercepts) and the slopes of weight against hatch date differed between years and treatment (Fig. 4.4). The mean fledgling weight for the control group in 2017, after correcting for the number of hatched chicks, was 15.30 g, whereas the mean for the experimental group was 0.59 g higher (95% CI: 0.300, 0.806); for 2018 this was 17.10 g for the control group, with the experimental group having on average a 0.60 g lower weight (95% CI: –0.791, –0.402). Thus, average weights differed between the treatments in a reversed order between the two years. The slopes of weight against hatch date in 2017 and 2018 were –0.011 (–0.025, –0.002) and –0.202 [–0.231, –0.174), respectively (Fig. 4.4). Within nests, i.e. between selection lines, there was no difference in weights (Table S4.2c). Again,

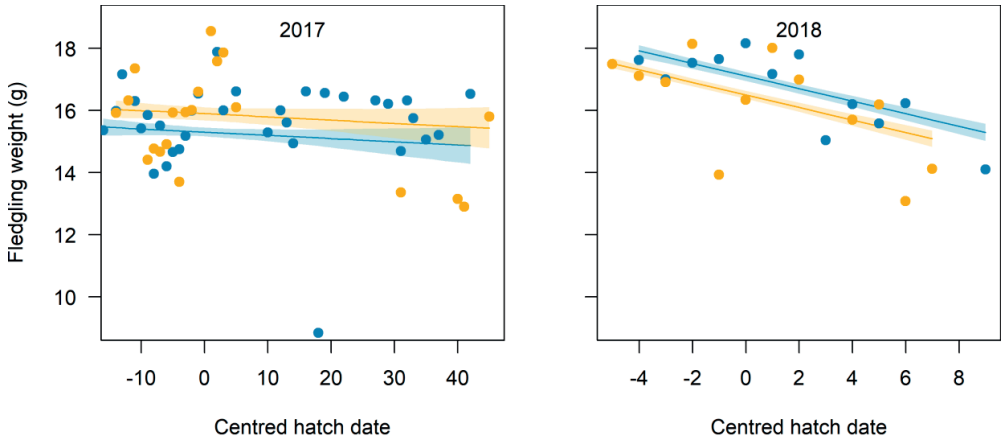


***Figure 4.4**. Fledgling weight (at day 15) as a function of mean-centred hatch date in the between-brood comparison in 2017 and 2018. Lines and shading are mixed-model estimates and their bootstrapped 95% CIs (blue: control nests; orange: experimental (swapped) nests); data points are average weights for each hatch date.*

the intended advancement in hatch dates in 2018 by adding dummy eggs to the experimental broods did not result in low-weight offspring in the earliest broods (Table S4.3b).

So, do offspring from the selection lines have a 'bad start' compared to the wild population? The most important result from the genomic selection experiment is that selection-line offspring do not differ in early-life fitness. That is, offspring from early lines are not likely to have a higher or lower recruitment probability than those from the late line, nor are they more likely to have (negative) carry-over effects to their first breeding season. Experimental eggs/chicks in general, however, appear to perform slightly differently than those from the natural population; that is, experimental broods had a stronger decline in the proportion of fledged chicks with later hatch dates compared to control nests and fledgling weights differed between these groups and between years. Two possible explanations may underlie these results.

First, there may be an inherent difference between the aviary-bred and 'wild' eggs, e.g. due to effects related to loss of genetic variation in the selection process. We believe, however, this is unlikely since birds in the aviaries were selectively bred for only four generations and paired up disassortatively with respect to breeding values for laying date within each selection line (Verhagen et al. in review), thereby maintaining genetic variation as much as possible. Moreover, the effects of treatment (control vs. experimental nests) on PFC was rather small (absent in 2018; Fig. 4.3) and fledgling weight showed a reversed pattern (Fig. 4.4) between years, indicating that whatever the mechanism behind these effects, it was not consistent (also note that for the most part the same F3 pairs were used to produce the eggs in both years).

Second, and perhaps more likely, the difference between experimental and control nests may lie in the post-egg-laying experimental procedures, although it is difficult to pinpoint which event or chain of events was responsible. Eggs were kept on a turning device at room temperature—sometimes for more than a week—before they were moved to foster parents in BE and before they were translocated to HV after ≥ 5 days of incubation. All factors may play a role in a determining the egg viability—and therefore hatching success. We do not have exact data on hatching success per nest, but if we use the number of chicks in the nest at day 8 as a (rather coarse) proxy for the number of hatched chicks, the proportion of hatched eggs relative to the clutch size in a binomial model is explained by an interaction between treatment and year ($\Delta AIC_c = 5.72$); that is, experimental (i.e. swapped) nests have a higher hatching success in 2017 but not in 2018 (Fig. S4.1). Although the cause for this difference is unclear, it has been suggested that exposing incubated eggs to suboptimal ambient temperature may impair embryonic development (Williams 2012). This effect should however only become apparent after at least a day's worth (or more) of exposure to such temperatures (Veiga 1992; Arnold 1993). Most likely, experiment-induced effects on PFC or fledgling weight occur between egg-laying and the first movement to the foster parents. The likelihood of embryonic development strongly decreased after the first week (personal observation); therefore, some eggs may not develop in the first five days in the foster nest and therefore only viable eggs ended up being moved to HV. This quality check did not take place with the 'wild' (control) eggs

and therefore a disproportionately high number of eggs may fail to hatch in these broods. Additionally, the hard selection on egg viability from the BE foster nests may have decreased the final clutch size in HV, making the per-egg odds of survival to fledging better due to decreased competition for food (Both et al. 2000; see Chapter 8). This was probably less the case in 2018, when the final clutch size was generally larger than the wild female's original clutch size, potentially explaining the lower average fledgling weight in this year (Fig. 4.4).

From the birds that fledged, weights differed between experimental and control nests and between years (Fig. 4.4). Fledgling weight is known to affect survival recruitment probability (Chapter 8), but the difference between treatments in both years is only about half a gram. This may seem a lot given the average weight of an adult great tit (~18 g), but overall recruitment probability in our population is generally low (~10%) and stochastic in nature. It remains therefore too early to conclude whether offspring resulting from the selection lines have different survival prospects than those from the wild population.

Recruited birds in 2018

From the selection-line fledglings in 2017, 11 birds (5.3%, vs. 11.4% of the natural population) recruited into the 2018 breeding population: two females and four males from the early selection line, and three females and two males from the late line. The recruited females, expectedly, laid clutches at times that matched the selection line they originated from; they were, however, not at the extreme end of the laying date distribution (Fig. 4.5). Sample sizes were too small to do statistics on them but the histogram in Fig. 4.5 shows that the two female recruits from the early line bred five days earlier than the first late-line female. The pattern for male recruits (that is, the laying date of the female wild partners of these males) is less clear, but not surprising given that laying date is determined by the female (Caro et al. 2009). That we did not observe both early and late females at the respective extremes of the distribution seems counterintuitive, but should in fact not be surprising given the small sample sizes. Our efforts to track down the females during the nest-building phase means that we were most likely able to identify all recruited birds within HV, and all identified recruits were observed breeding. Therefore, selective disappearance of birds breeding too early was not a likely reason for the lack of extremely early laying dates in the population distribution (Fig. 4.5).

In terms of the fitness consequences of laying date, the results are too premature to draw conclusions. In Fig. 4.6, the distribution of fledglings of nests in which at least one chick fledged are depicted. Statistical analysis notwithstanding, a first glance reveals that there is no discernible pattern in the data. Note that all tracked recruits eventually started breeding and only two of these broods were omitted from Fig. 4.6 because they were deserted due to disturbance (measurement of daily energy expenditure; not reported here). There were hence no reasons to either accept or reject either hypothesis laid out in the introduction. Since reproductive success usually declines with the progressing season (Rowe et al. 1994), we would expect birds breeding early in the season to have a higher reproductive success due to better chick-provisioning conditions (see also Chapter 5). In

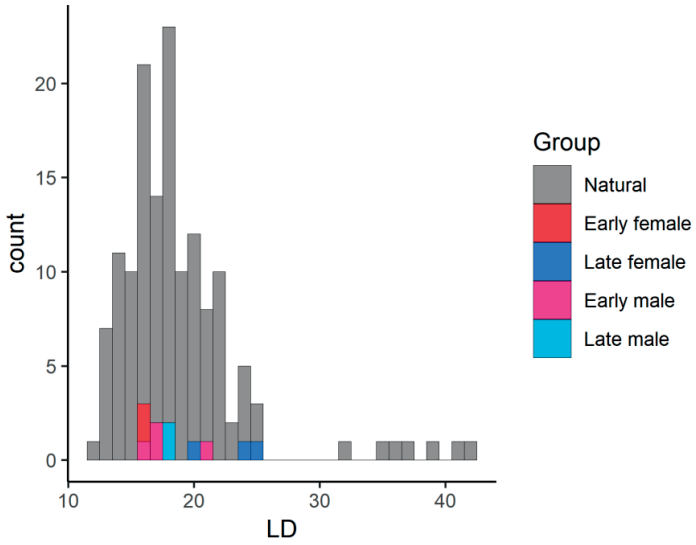


Figure 4.5. *Frequency plot of first-clutch laying dates (LD) of female great tits at the Hoge Veluwe in 2018 (i.e. the first year in which recruits from the selection lines were recorded). Different colours denote the background of each female (red: early selection line; pink; females mated to males from the early line; dark blue: late line; light blue: females mated to males from the late line; grey: wild population).*

2018 in the HV, there was no distinct time trend (linear model of number of fledglings against laying date for broods that had ≥ 1 fledgling: laying date: –0.035 [–0.070, 0.005]; (laying date)^2: –0.006 [–0.017, 0.008]). Therefore, even if the recruits of the selection lines had been at the extreme ends of the laying date distribution, this would not have given them any reproductive advantage (although, granted, the number of fledglings is an

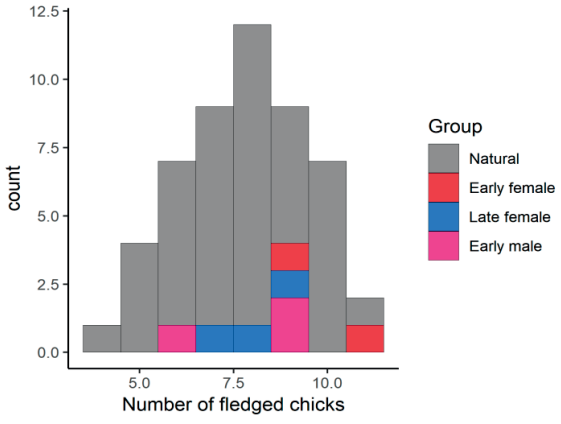


Figure 4.6*. Frequency plot of the number of fledglings of non-manipulated nests in the Hoge Veluwe in 2018 (i.e. the first year in which recruits from the selection lines were recorded). Different colours denote the background of each female (red: early selection line; pink; females mated to males from the early line; blue: late line grey: wild population). Note: two broods from selection-line recruits were prematurely deserted due to disturbance and are not shown here.*

incomplete measure of fitness, further precluding us from drawing definitive conclusions).

Concluding remarks

To be able to understand how laying date affects reproductive success, we need experiments that cleanly manipulate the timing of breeding. Genomic selection on laying date based on genetic markers (Gienapp et al. 2019) can provide a promising step to achieving this because it provides an accurate and efficient way of manipulating birds at the genome level in a mere few generations. It is important to note, however, that selecting on additive genetic variation cannot create more extreme genotypes than the parents. This means that he most extreme genotypes emerging from the selection lines will be within the natural boundaries of the population with respect to their breeding values for laying date, and are therefore not likely to have more extreme phenotypes in any environment compared to the most extreme genotypes in the natural population.

Possibly, selecting for laying date means selecting for other traits as well, for example because certain loci are in linkage disequilibrium and because variation in life-history traits is likely to be caused by many loci of small effects spread across the genome (Santure et al. 2013). In the wild, for example, laying date is associated with clutch size, with early layers having a larger clutch (e.g. Dalhaug et al. 1996; Winkler and Allen 1996), which has been postulated to be an adaptive response to anticipated chick-rearing conditions later in the season (e.g. Winkler and Allen 1996) or with variation in the quality of females laying at different times (e.g. Christians et al. 2001). If there is a genetic basis for this association between laying date and clutch size (Sheldon et al. 2003; but see Postma 2005) or female quality, the genomic selection on timing might cause a correlated response in these other traits. This is, however, difficult to measure; from three generations of egg-laying in the aviaries during the selection experiment (Verhagen et al. in review), clutch size could not be measured because some birds would lay as many as 30 eggs in *ad libitum* food conditions.

Crucially, therefore, we depend strongly on many years of data collection on many individual recruits from the selection lines to be able to draw conclusions about (1) the causal effect of laying date on reproductive success and (2) whether selection for extremely early laying improves or deteriorates fitness. In particular for the latter objective, we need several years of observations from both 'bad' and 'benign' springs. The *constraints* hypothesis, which states that birds pushed forward in time by too much will pay fitness costs, can only be tested if conditions in early spring are genuinely harsh; if not, any advancement in laying date may have the opposite effect, i.e. that they are simply better matched with the caterpillar peak without having to pay survival (or other) costs early in the season. Additional field manipulations among the earliest of the recruits—such as food supplementation in early spring (see Chapter 3)—may help push birds forward in time even more, but this comes with the obvious disadvantage of changing the physical state of the females, hence possibly biasing estimates of reproductive success (Verhulst and Nilsson 2008). The current study, which is being continued at this moment, can therefore

only be successful if we continue to breed birds in captivity and foster their eggs in the wild.

Acknowledgements

We thank the board of the Hoge Veluwe National Park to allow us to do fieldwork, and thank Mees Swinkels, Imre Schuitemaker, Mireille van Bentvelzen, Maartje van Deventer, Martijn Burgman, Sylvana Harmsen and Melanie Lindner for their invaluable contribution in collecting the field data and the DNA analysis. We are grateful to Barbara Tomotani and Irene Verhagen, as well as the animal caretakers, for their invaluable help managing the aviaries. This work is supported by an ERC Advanced Grant (339092 - E-Response) to M.E.V.

Supplementary Information S4

Table S4.1. *AIC_c comparison of competing models explaining variation in (a) the proportion of chicks fledged, (b) the number of fledged chicks, and (c) individual fledgling weights* ***between*** *nests (i.e. experimental vs. control nests).*

Model	AIC_c	ΔAIC_c
(a) Proportion of chicks fledged (relative to clutch size)		
HD + BS + BY + HD:BY	866.6	14.9
Exp + HD + BS + BY + HD:BY	868.2	16.5
Exp + HD + BS + BY + HD:BY + Exp:BY	863.7	12.0
Exp + HD + BS + BY + HD:BY + Exp:HD	856.4	4.7
Exp + HD + BS + BY + HD:BY + Exp:BY + Exp:HD	**852.4**	**0.7**
Exp + HD + BS + BY + HD:BY + Exp:BY + Exp:HD + Exp:BY:HD	851.7	0
(b) No. of fledged chicks (Generalized Poisson)		
HD + BS + BY + HD:BY	887.0	4.2
Exp + HD + BS + BY + HD:BY	889.0	6.3
Exp + HD + BS + BY + HD:BY + Exp:BY	889.8	7.1
Exp + HD + BS + BY + HD:BY + Exp:HD	**882.7**	**0**
Exp + HD + BS + BY + HD:BY + Exp:BY + Exp:HD	883.8	1.1
Exp + HD + BS + BY + HD:BY + Exp:BY + Exp:HD + Exp:BY:HD	885.6	2.8
(c) Fledgling weight		
BY + NH + HD + BY:HD	4921.8	2.6
BY + NH + HD + BY:HD + Exp	4924.9	5.7
BY + NH + HD + BY:HD + Exp + Exp:HD	4930.7	11.4
BY + NH + HD + BY:HD + Exp + Exp:BY	**4919.2**	**0**
BY + NH + HD + BY:HD + Exp + Exp:HD + Exp:BY	4924.8	5.5
BY + NH + HD + BY:HD + Exp + Exp:HD + Exp:BY + Exp:HD :BY	4929.5	10.3

Notes: HD = hatch date; BS = brood (clutch) size; BY = brood year; Exp = experiment (swapped vs. control nests).
Best models are marked in bold.
The models under (*c*) contained a random effect of 'brood ID'.

Table S4.2. *AIC_c comparison of competing models explaining variation in (a) the proportion of chicks fledged, (b) the number of fledged chicks, and (c) individual fledgling weights* ***within*** *nests (i.e. early vs. late selection line).*

Model	AIC_c	ΔAIC_c
(a) Proportion of chicks fledged (relative to clutch size)		
HD + BY	**528.5**	**0**
HD + BY + LI	530.0	1.5
HD + BY + LI + LI:HD	532.1	3.6
HD + BY + LI + LI:BY	531.6	3.1
HD + BY + LI + LI:HD + LI:BY	533.7	5.2
HD + BY + LI + LI:HD + LI:BY + LI:HD:BY	538.1	9.6
(b) No. of fledged chicks (Poisson)		
HD + BY	**741.7**	**0**
HD + BY + LI	745.2	3.5
HD + BY + LI + LI:HD	743.2	1.5
HD + BY + LI + LI:BY	746.0	4.3
HD + BY + LI + LI:HD + LI:BY	744.5	2.8
HD + BY + LI + LI:HD + LI:BY + LI:HD:BY	748.6	6.9
(c) Fledgling weight		
BY + NH + HD + BY:HD	**1784.5**	**0**
BY + NH + HD + BY:HD + LI	1786.7	2.2
BY + NH + HD + BY:HD + LI + LI:HD	1794.9	10.4
BY + NH + HD + BY:HD + LI + LI:BY	1789.5	5.0
BY + NH + HD + BY:HD + LI + LI:HD + LI:BY	1797.8	13.3
BY + NH + HD + BY:HD + LI + LI:HD + LI:BY + LI:HD :BY	1804.0	19.5

Notes: HD = hatch date; BS = brood (clutch) size; NH = number of eggs hatched; BY = brood year; LI = selection line (early vs. late).
Best models are marked in bold.
All models contained a random effect of 'brood ID', with treatment (selection line) nested within.

Table S4.3. *Results of the* ***between-brood*** *analysis for 2018 only, testing the quadratic effect of hatch date in experimental (swapped) nests.*

Model	AIC_c	ΔAIC_c
(a) Number of fledglings		
CS + HD + Exp	**353.35**	**1.11**
CS + HD + Exp + HD2	354.02	1.78
CS + HD + Exp + HD2 + Exp:HD	352.23	0
CS + HD + Exp + HD2 + Exp:HD + Exp:HD2	354.67	2.44
(b) Fledgling weight		
CS + HD + Exp	**2231**	**0**
CS + HD + Exp + HD2	2234.45	3.45
CS + HD + Exp + HD2 + Exp:HD	2234.557	3.557
CS + HD + Exp + HD2 + Exp:HD + Exp:HD2	2242.224	11.224

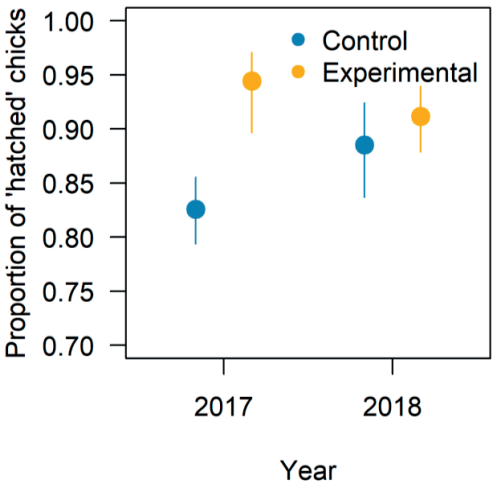


Figure S4.1*. Proportion of hatched eggs (95% bootstrapped CIs) for experimental and control nests in the between-nest comparison. The number of hatched eggs is a proxy (number chicks alive at day 8) because we have no complete tally of the actual hatched eggs.*

Chapter 5

Comparing two measures of phenological synchrony in a predator–prey interaction: simpler is better

Jip J.C. Ramakers, Phillip Gienapp & Marcel E. Visser

Abstract

Recent decades have seen an increasing interest in the impact of global warming on mismatch between consumer and resource phenology. Most studies have focussed on the temporal synchrony between the dates of peak consumer demands and peak resource availability (match in dates; MD). However, Lindén [(2018) *PNAS* 115(20):5057–5059] argued that a rigorous definition of phenological synchrony should take into account the shape and height of the temporal phenological distributions and describe phenological synchrony as the degree of overlap between them (match in overlap; MO). We tested whether phenological synchrony is better described by MD or MO using 24 years of breeding data of the great tit (*Parus major*) and the main food source for its nestlings, caterpillars. We estimated caterpillar availability and nest-level food requirements on a daily basis throughout the breeding season to determine MO. MO and MD correlated strongly: years with high matching between peak dates showed the highest degree of matching in overlap. However, offspring recruitment probability, a key demographic parameter, correlated strongly with MD but weakly with MO. Furthermore, we identified MD, and not MO, as a driver for selection on egg-laying date. Thus, temporal match in peak dates has better explanatory power than the overlap between the phenological distributions. We argue this is because, unlike MD, quantifying MO is not straightforward and has to be based on non-trivial assumptions. We conclude that a detailed, season-wide description of resource availability is not always essential—or even possible—to describe important demographic processes in wild populations.

Manuscript submitted

Introduction

Organisms in seasonal environments, where the phenology of resource abundance varies from year to year, need to adjust their timing of reproduction to match this variation to ensure successful reproduction (Lepage et al. 1998; Siikamäki 1998; Verboven and Visser 1998; Kokko 1999; Réale et al. 2003a; Smith and Moore 2005; Plard et al. 2014; Reid et al. 2018). Recent decades have seen a growing interest among biologists in the effect of climate warming on changes in phenology (Visser et al. 1998; Parmesan and Yohe 2003; Both et al. 2004a; Durant et al. 2007; Visser 2008; Singer and Parmesan 2010; Dunn and Moller 2014; Plard et al. 2014). Typically, warming springs lead to an advancement in phenological events and these advancements occur at different rates between different trophic levels (Thackeray et al. 2010; Thackeray et al. 2016; Kharouba et al. 2018). The unequal shift in phenology between consumers and their resources, referred to as 'phenological mismatch' (Cushing 1990; Stenseth and Mysterud 2002a; Durant et al. 2007), has in some cases been linked to directional selection on consumer phenology (Visser et al. 1998; Reed et al. 2013b; Marrot et al. 2018) and negative effects on consumer demography (Plard et al. 2014).

In a recent response to a large-scale meta-analysis on climate change-driven phenological mismatch (Kharouba et al. 2018), Lindén (2018) argued that to better understand the demographic processes mediated by phenological mismatches, a clear and rigorous definition of phenological synchrony is needed. This synchrony between consumer and resource phenology can be described as the difference between the dates when the phenological distributions of consumer and resource peak (match in dates; MD). Most studies have used this match in peak dates as a proxy to study phenological synchrony (Visser et al. 1998; Thackeray et al. 2010; Reed et al. 2013b; Kharouba et al. 2018). A number of publications (Durant et al. 2005; Durant et al. 2007; Miller-Rushing et al. 2010; Lindén 2018), however, have suggested that a better measure from the consumer's perspective would be the 'area of overlap' under the intersecting distributions of consumer and resource phenology (match in overlap, MO). The key argument is that resources may be plentiful even when peak dates are out of synchrony when the resource peak is either high (years with plenty of food) or wide (Miller-Rushing et al. 2010; Lindén 2018). Conversely, even if peak dates in phenologies were well matched, overall low resource availability would reduce consumer fitness (Cushing 1969). Although these two measures of phenological synchrony will often be highly correlated (Miller-Rushing et al. 2010; Lindén 2018), it is of interest to test which of them is most relevant for demographic and evolutionary processes. One important caveat is that estimating absolute food availability to the consumer requires important assumptions that are difficult—if not impossible—to verify. For example, great tits (*Parus major*) are highly dependent on ephemeral abundances of caterpillars (Lepidoptera) to feed their offspring in some regions (Lack 1950; Betts 1955; Royama 1970; Van Balen 1973). Even if one were able to reliably estimate the total amount of caterpillars available in a given area at a given point in time, the net amount of food available to the individual nestlings would depend strongly on factors such as the density of the breeding population (through competition) and the

spatiotemporal distribution of prey, both in size and numbers, affecting the search time and radius of the parents. Simply quantifying overlap between resource and demand assumes that what is available can be effectively used by the consumer, an assumption that may not be true (Pyke et al. 1977). Moreover, it assumes that we can estimate reliably the day-to-day required amount of caterpillar biomass, which in reality will be context dependent (Royama 1966; O'Connor 1975; Mertens 1977). MD, on the other hand, is free of such assumptions as it only requires an estimation of the date at which energy requirements are highest (in great tits around day 10 post-hatching (Keller and Van Noordwijk 1994; Mols et al. 2005)) and the date at which caterpillars are likely to be most abundant.

Here, we tested which of the two quantifications of phenological synchrony—the match of peak dates and the phenological overlap—correlated better with selection and recruitment in a wild population of great tits (*Parus major*). Great tits in this population are strongly (albeit not exclusively) dependent on caterpillars (mainly *Operopthera brumata* and *Tortrix viridana*) to raise their offspring (Van Balen 1973), which are available to them over a span of a few weeks during the breeding season. Egg-laying date in this population is under increased directional selection due to climate warming, which has been linked to the decreased temporal synchrony with caterpillar abundance (Visser et al. 2006; Reed et al. 2013b). We used our long-term (24 years) data to construct a daily food-availability and food-requirement profile throughout the breeding season to estimate the overlap between the distributions (MO) as well as the temporal match of peak dates in phenology (MD). We compared models containing either or both of the metrics of phenological synchrony to test their importance in predicting (i) the recruitment probability of great tit nestlings and (ii) selection on egg-laying date of the mothers. We discuss important limitations of constructing food-availability and food-requirement distributions as well as the appropriateness of using either measure of phenological synchrony to describe ecological interactions between trophic levels.

Results and discussion

We quantified food availability throughout the season for 24 years in the Hoge Veluwe National Park, the Netherlands, by collecting caterpillar droppings on multiple days during the breeding season, which we used to calculate the biomass of caterpillars (g m^{-2}) on each measurement day. Similarly, we also estimated the biomass in caterpillars required by great tit nestlings between 10 and 15 days old and summed these requirement per nestling over all nests to estimate the required caterpillar biomass for all dates throughout the season. We used these data to estimate the phenological overlap between food availability and requirements. Because availability and requirements were on different scales (g m^{-2} vs g, respectively), we rescaled both such that the total area under each curve equalled 1 (Fig. S5.1). The intersection of the food-abundance distribution and the distribution of the requirements of all nestling in the population is now the proportional area of overlap at the population level (MO_P; see Methods and the discussion

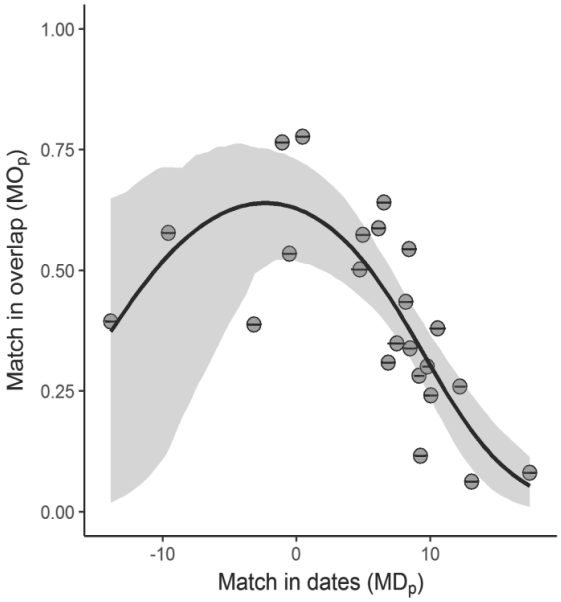


Figure 5.1. *Coefficient of match in overlap (MO) against the match in dates (MD) in phenology at the population level (denoted by subscript p) in great tits. MO_p is the proportion of the food requirement distribution overlapping with the food availability curve. MD_p is the difference between the average egg-laying date + 33 days and the peak date of caterpillar biomass, where positive values indicate that the population on average bred too late relative to the food peak, and negative values indicate that it bred too early (0 = perfect match). Line and shading represent estimates and 95% bootstrapped CI from a beta-regression model, accounting for standard errors in MD_p (horizontal lines).*

below for important caveats). At the individual brood level, MO_b was defined as the within-season-standardized amount of food available to chicks from day 10 to 15 post-hatching for a specific brood.

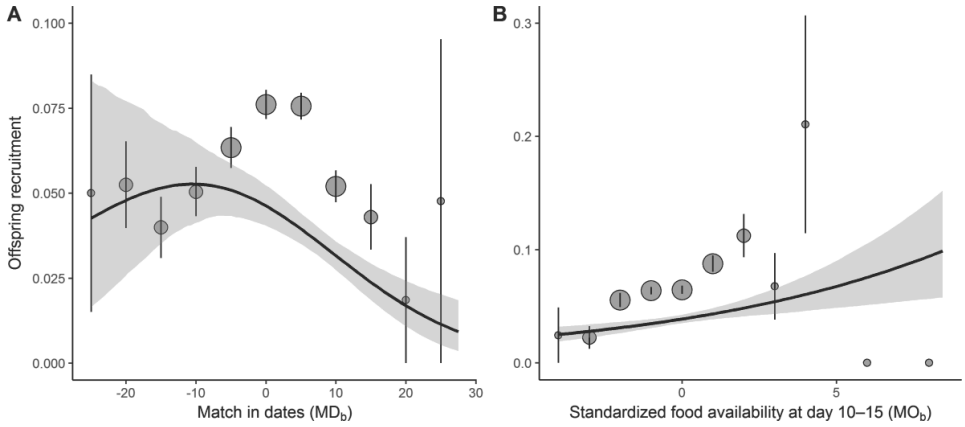


Figure 5.2. *Nestling survival to the next breeding season as a function of (**A**) MD_b (i.e. the date on which nestlings are 10 days old minus the peak date of caterpillar biomass) and (**B**) MO_b (i.e. food availability to 10–15-d-old nestlings, standardized across broods within a season). Points are binned raw means with their standard errors, plotted for visual purposes only, with symbol sizes corresponding to sample sizes (small: ≤ 100 nestlings; medium: < 100 and ≤ 1000 nestlings; large: > 1000 nestlings). The prediction lines and 95% bootstrapped CIs (shadings) were derived from the 2nd (**A**) and 3rd (**B**) model in Table 5.1a, keeping other variables constant at their means. Note the different scaling on the y-axes.*

The proportional phenological overlap between the curves (MO_p) correlated non-linearly with match in peak dates in phenologies (MD_p; average egg-laying date + 33 days minus the caterpillar biomass peak date) (Fig. 5.1; beta-regression coefficients [bootstrapped 95% CI]: MD_p: –0.045 [–0.110, 0.027]; MD_p^2: –0.007 [–0.013, –0.002]; pseudo r^2 = 0.59 [0.32, 0.80]). That is, the temporal, proportional overlap between food need and food availability was largest in years when the date of the peak food requirements (i.e. the average date across broods at which nestlings were 10 days old, when demand is highest (Keller and Van Noordwijk 1994; Mols et al. 2005)) was well matched with the date of peak caterpillar availability. Therefore, we would predict that MD and MO drive offspring recruitment and selection on breeding time to a similar degree.

We tested whether the survival of a nestling to recruitment (i.e. being found as a breeding bird in later years) correlated with either or both of the two measures of phenological synchrony: (i) the degree of matching between the date when chicks were 10 days old and the peak date in caterpillar biomass (MD at the brood level: MD_b) and (ii) MO_b. We constructed Generalized Linear Mixed Models (GLMMs, with binomial errors) that included either MD_b or MO_b, or both, with other important determinants of recruitment (breeding density and standardized fledgling weight (Reed et al. 2013b; Chapter 8)) as fixed effects, and random effects of year, mother and brood ID nested within mother. The best model explaining variation in offspring recruitment probability contained MD_b, including its quadratic term, but not MO_b (Table 5.1a). Offspring

Table 5.1. *Comparison of models containing the two metrics of phenological synchrony (MD and MO) explaining variation in (a) P. major nestling survival to recruitment (GLMMs; n = 14535 nestlings from 2009 broods) and in (b) standardized selection differentials for P. major egg-laying date (linear models weighted by the number of recruits; n = 23 years).*

Model terms	AIC_c	ΔAIC_c
(a) Offspring recruitment probability		
wt + wt^2 + dens + MD_b	6579.96	2.72
wt + wt^2 + dens + MD_b + MD_b^2	6577.24	0
wt + wt^2 + dens + MO_b	6588.89	11.65
wt + wt^2 + dens + MD_b + MO_b	6580.90	3.66
wt + wt^2 + dens + MD_b + MD_b^2 + MO_b	6579.24	2.00
(b) Standardized selection differential		
MD_p	2.28	0
MO_p	4.88	2.60
MO_p + HCP	7.84	5.56
MD_p + MO_p	5.12	2.84
MD_p + MO_p + HCP	8.38	6.10

Notes: *(a)* wt = standardized fledgling weight; dens = breeding-pair density; MD_b = brood-level phenological match in dates; MO_b = standardized food availability to a nest (day 10–15), as a proxy for brood-level match in overlap. Random effects were year, mother and brood ID (nested within mother).

(b) MD_p: population-level phenological match in dates; MO_p = population-level phenological match in overlap; HCP = height of the caterpillar peak.

recruitment was highest when broods with 10-d-old nestlings were close to matching with the peak date of caterpillar availability (Fig. 5.2a; estimate MD_b [bootstrapped 95% CI]: –0.026 [–0.042, –0.011]; MD_b^2: –0.001 [–0.003, –0.0003]; see also Visser et al. (2006)). Recruitment probability correlated significantly positively with MO_b in a model that did not contain MD_b (Fig. 5.2b; 0.115 [0.053, 0.177]), but this model performed substantially worse than the best model that contained MD_b and MD_b^2 ($\Delta AIC_c = 11.65$).

Since food availability determines offspring recruitment probability (see above; e.g. Durant et al. 2005; Toupoint et al. 2012; Reed et al. 2013b), reproductive success should decline with breeding time if the population breeds too late in relation to caterpillar phenology (and increase if it breeds too early). We estimated standardized selection differentials for egg-laying date for each year and used them as a response variable in a weighted linear regression model to test the performance of MD_p and MO_p (signed to match the direction of MD_p), whilst also fitting the height of the caterpillar distribution to capture a relevant dimension of its original distribution. The best model contained MD_p but not MO_p ($\Delta AIC_c = 2.60$ to 5.56; Table 5.1b). Selection for earlier breeding intensified significantly (became more negative) at higher values of MD_p (i.e. larger mismatch of peaks; Fig. 5.3a; estimate [bootstrapped 95% CI]: –0.015 [–0.032, –0.005]; $r^2 = 0.181$ [0.016, 0.418]). MO_p, on the other hand, correlated only weakly with selection differentials (Fig. 5.3b; estimate: –0.186 [–0.392, 0.099]; $r^2 = 0.083$ [0.000, 0.370]; note that including the height of the caterpillar distribution worsened model fit: $\Delta AIC_c = 2.96$).

Our results show that the phenological synchrony of food availability and food requirements can be better estimated as the differences in days between the mean phenologies (MD) than as the relative degree of overlap of these two distributions (MO), even though MD and MO correlated with one another in a predictable fashion, both at the population level (Fig. 5.1) and the brood level (through nest-level food availability; Fig.

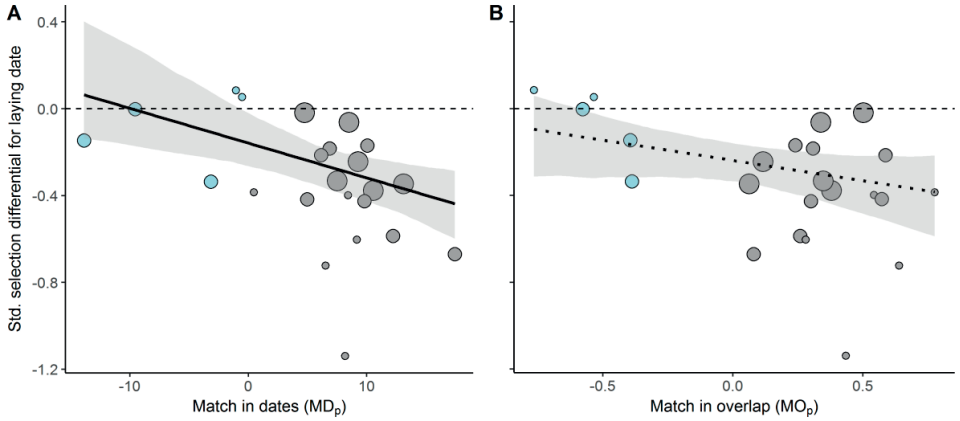


Figure 5.3. *Standardized selection differentials for egg-laying date as a function of (**A**) MD_p (0 = perfect match) and (**B**) MO_p, signed to match the direction of MD_p (years with negative MD_p are coloured in blue). Symbol sizes indicate the number of recruits in each year (small: ≤ 20 recruits; medium: > 20 and ≤ 40 recruits; large: > 40 recruits). Lines and shadings are estimates and 95% bootstrapped CIs from regression models weighted by the number of recruits.*

S5.2). One important reason why MD performed better than MO may be due to the logistical difficulty inherent to estimating food availability; after all, to get accurate estimates of phenological overlap between predator and prey (Lindén 2018), sufficient knowledge of resource availability (e.g. total number of prey, their size, and the spatiotemporal distribution of both) is required but this will be challenging in natural systems for various reasons. For example, to construct the food abundance throughout the entire breeding season we needed to extrapolate the shape of the distribution outside the measuring period, when the values at either the first or the last measurement were > 0 (see Methods for how these data were treated). Similarly, to construct a food-requirement distribution, we had to make assumptions about age-specific energy requirements and food intake rates in great tit nestlings, which may vary with context (Royama 1966; O'Connor 1975; Mertens 1977). Even if we assume that we managed to estimate both distributions with reasonable accuracy, we had to transform them both to get them on the same scale. This means that our measure of MO was now not an absolute measure of overlap (MO_P), which has been argued to matter most in consumer–resource interactions (Durant et al. 2007; Miller-Rushing et al. 2010; Lindén 2018). We attempted to get around this problem in our analysis of selection on egg-laying date by fitting the absolute maximum height of the caterpillar distribution (as a proxy for total biomass in the season) to correct for the lack of dimension in MO_P. In our analysis of offspring recruitment, we standardized food availability across broods such that it became a measure of what was available relative to other broods in that year. The latter approach (MO_b) may appear more useful than MO_P since it provided a more direct measure of overlap and it indeed correlated with offspring recruitment probability in the predicted direction (Fig. 5.2b). In the model comparison, MO_b was nevertheless statistically outperformed by temporal match of peaks (MD_b). Importantly, while MO_b does not suffer from the same scaling issue as MO_P, it still assumes that whatever amount of food is available on a given day will be effectively available to the great tit nestlings, ignoring issues of parental search and travelling time. We therefore show that MD in our system is a better quantification of phenological synchrony than MO.

Our findings echo previous work that highlight match of peak dates in phenology as an important factor influencing mother and offspring fitness (Vatka et al. 2011; Reed et al. 2013b; Chevin et al. 2015). We do not argue that this will necessarily be true in all study systems: in species that are not highly dependent on a single food type, or whose food does not exhibit a well-defined, seasonal distribution, demographic processes will either depend more strongly on MO or on neither MD nor MO (Durant et al. 2005; Dunn et al. 2011). However, studies reporting fitness and demographic consequences in this context so far have generally used (proxies of) MD to quantify phenological mismatch and reported reduced fitness in years when temporal mismatch was high (Plard et al. 2014; Regular et al. 2014; Arlt and Pärt 2017; Marrot et al. 2018). Durant et al. (2005), on the other hand, quantified effects of MD and food abundance on population indices of reproductive success in three study systems and found that in two of them food availability was a better predictor than MD. In one of these two systems (Soay sheep *Ovis aries*), however, food (i.e. vegetation, indicated by integrated NDVI) was weakly seasonal, whereas in the other system (Atlantic puffins *Fratercula arctica* and herring *Clupea harengus*) an incomplete

measure of fitness (i.e. the number of fledged chicks) was used (Durant et al. 2005), making these studies not totally comparable to ours. There is hence no *a priori* expectation as of yet that consumer–prey interactions in other highly seasonal environments should be critically different from that of the great tits reported here.

Lindén (2018) ends his commentary with the recommendation that instead of focussing solely on temporal, phenological synchrony (e.g. of peak dates) to describe ecological interactions between trophic levels, we should also incorporate information on abundances across the season. While we agree with the underlying logic, we show that phenological match in peaks is in fact a reliable proxy describing demographic processes in a system in which the consumer is strongly dependent on highly ephemeral prey. The important advantage of using MD to quantify phenological synchrony is that it requires a comparatively straightforward way of collecting data that, in any case, will be more accurate than any approximation of absolute food availability. This is because MD 'only' requires sampling food (e.g. per unit area) at regular time intervals, preferably across multiple sites within the study area, spanning a wide enough range to be able to estimate when abundance peaks. As we have shown here, we can attempt to develop proxies of phenological overlap (MO) but our expectation is that in many contexts MD will be a more effective and less biased measure of phenological synchrony.

We encourage other researchers of long-term population studies of species highly dependent on an ephemeral food source to start collecting the data necessary to quantify MD. It is these long-term data that will enable us to understand the long-term population consequences of phenological mismatch under a changing environment (Visser 2008; Clutton-Brock and Sheldon 2010).

Methods

Data collection

We made use of 24 years (1994–2018, excluding 1997; see *Estimating food availability and food requirement* for justification) of data on caterpillar availability and great tit (*Parus major*) breeding data at the Hoge Veluwe National Park (HV; 52°02′07″ N, 5°51′32″ E, central Netherlands). In this area, approximately 400 nest boxes are available for great tits and other hole-breeding passerines to nest, and the whole reproductive cycle from egg laying to fledging of chicks is monitored. Adults are captured at the nest and identified by means of aluminium leg bands during the chick-provisioning stage. Chicks are banded and weighed on day 15 post-hatching, which is close to the date of fledging.

During the breeding season, the caterpillar biomass are estimated by putting up two frass nets (cheese cloths) underneath 15 pedunculate oak (*Quercus robur*) trees spread across the 171-ha study area (see Visser et al. 2006 for details). These nets capture the droppings (frass) of caterpillars (mostly winter moth (*Operopthera brumata*) and oak leaf roller (*Tortrix viridana*)) present in the trees. Nets are usually deployed from mid-April to mid-June, and sampled every 3–4 days. Caterpillar droppings are collected, dried at 60°C

for 24h, and sorted (i.e. debris removed). The dried droppings are then used to calculate the caterpillar biomass whilst correcting for daily temperatures (which affect caterpillar growth) using the equation in Tinbergen and Dietz (1994), which correlates well with biomass obtained from branch samples (Visser et al. 2006). Biomass is first averaged per tree and then across sampling trees to get grams of biomass per square meter for the date which falls in the middle of the sampling days.

Estimating food availability and food requirement

To estimate caterpillar biomass on a daily basis, we used a smooth-spline technique with maximal degrees of freedom to interpolate biomass between measuring days. With this method, biomass outside the measuring period is predicted as a linear function, adopting the slope estimated from the last (or first, depending on the side of the curve) interpolation point. In most years, predicted biomass would therefore linearly decline toward zero. In some years, however, the slope at the last or first interpolation point was slightly positive, leading to an upward prediction of caterpillar biomass at the both ends of the food curve; if this was the case, we arbitrarily set biomass beyond the first or last measuring point to zero. We believe this is a reasonable approach, since in most years the frass sampling scheme started and ended when apparent biomass was (close to) zero. An exception was 1997, where sampling started when caterpillar biomass was clearly on the rise, so we discarded this year from our analyses.

To estimate nestling food requirements we summed up the needs of every great tit nestling from first broods from age 10 to 15 days post-hatching. We chose this period for two reasons. First, energy requirements and intakes are highest from around day 10 onward (Royama 1966; Keller and Van Noordwijk 1994; Mols et al. 2005). Second, since much of nestling mortality takes place within the first week after hatching (e.g. Nur 1984), restricting the dataset to day 10–15 gives us confidence that brood size on day 15 (when they are banded and measured) in most cases accurately reflects the number nestlings present from day 10 to 15. We used the observed, age-specific energy intake as estimated by Mols et al. (2005) and Royama (1966) as a proxy for required energy intake from day 10 to 15 (kJ nestling^{-1} day^{-1}; see Figure 1 in Mols et al. (2005)). Note that other factors than age (e.g. ambient temperature) may affect metabolic rates and hence the required energy intake (Royama 1966; O'Connor 1975; Mertens 1977), but we assume here that these factors average out in the estimates derived from Mols et al. (2005) and Royama (1966). We divided the required energy intake by the energy content of caterpillars (21.4 kJ g^{-1} dry weight (Bell 1990)) to get the dry biomass of caterpillar required per nestling per day. Assuming 80% wet mass in caterpillars (Bell 1990), we multiplied the dry biomass by 5 to get the total required biomass, which amounted to 3.97, 4.21, 4.37, 4.49, 4.51 and 4.51 g nestling^{-1} day^{-1} from day 10 to 15, respectively. This agrees reasonably well with the estimated mean caterpillar intake of 4.66 g nestling^{-1} day^{-1} in great tit broods with nine nestlings found by Gibb and Betts (1963). Daily estimates of food requirements were summed across broods to create a food requirement distribution for all great tit nestlings in the study area.

One definition of phenological match is the degree of overlap between the food requirement and availability distributions (Durant et al. 2007; Miller-Rushing et al. 2010; Lindén 2018). The idea behind it is that even when peak dates differ, the population may not be mistimed because food is still plentiful. However, food availability and requirements are on a different scale (g m^{-2} vs g, respectively).We therefore scaled both food availability and requirement as a proportion of the total in a given season, such that the integral under each curve equalled 1 (Fig. S5.1). Relative overlap was then determined as the integral of the overlapping area (Miller-Rushing et al. 2010) (see Results and Discussion for issues with this approach). In subsequent analysis (selection on laying date; see below), we corrected for the loss of dimension using the absolute maximum height of the caterpillar distribution.

Analysis

We compared the performance of the two main measures of phenological match—i.e. the temporal synchrony in days between the peak dates of the food needs and the food availability curves (or MD) and the amount of overlap between the food availability and requirement curves (MO) in explaining (*a*) recruitment probability and (*b*) the strength of selection.

(a) Offspring recruitment probability. We fitted a generalized linear mixed-effects model (GLMM, package 'lme4' (Bolker et al. 2009; Bates et al. 2018)) with a binomial error structure to model nestling recruitment (survival to breed in the next year) for each brood that had nestlings on days 10–15 (n = 14535 nestlings from 2009 broods, excluding broods that failed during egg-laying, incubation or early nestling stages, and excluding the year 2018 for the lack of recruitment data). We fitted five different models to assess the relative importance of MD and MO, whilst controlling for the density of breeding pairs (Reed et al. 2013b) and the linear and squared terms of fledgling weight, standardized across broods within a season (recruitment probability increases with fledgling weight but falls in the heaviest fledglings; Chapter 8): (i) "+ MD_b" (i.e. brood-level MD: the difference between the date at which the chick are 10d old and the peak date in caterpillar biomass); (ii) "+ MD_b + MD_b^2" (because we would expect recruitment probability to peak around MD ~ 0); (iii) "+ MO_b" (i.e. brood-level MO: the total amount of food available to a given brood from day 10 to 15, standardized across broods within a season); (iv) "+ MD_b + MO_b"; (v) "+ MD_b + MD_b^2 + MO_b". Random effects were year, female and brood identity (nested within the female's identity, to account for multiple breeding attempts by the same female across years). Variance inflation factors (VIFs) confirmed that multicollinearity was not an issue in our data (VIF ≤ 1.17). Since we fitted models with similar degrees of freedom, we compared them using Akaike's Information Criterion corrected for small samples (AIC_c) to assess whether MD_b outperformed MO_b or vice versa (models within 2 AIC_c units from the top-ranked one were considered competitive (Burnham and Anderson 2002)). To assess 'significance' of MD_b and MO_b, we obtained the estimates from the most parsimonious model containing the variable of interest and calculated bias-corrected and accelerated (BC_a) 95% confidence intervals (DiCiccio and Efron 1996) with 1000 iterations.

(b) Selection on egg-laying date. To test the effect of MD and MO on selection on egg-laying date, we first estimated annual selection differentials. We used all females' first broods that were not experimentally manipulated (n = 2054 broods from 23 years, excluding 2018 for the lack of recruitment data). Standardized selection differentials (s') were estimated as the covariance of individual relative fitness (w, the number of recruits divided by the mean number of recruits across females) with the standardized egg-laying date (z): $s' = cov(w, z)/\sigma_z$ (Lande and Arnold 1983).

With the standardized selection differential s' in place, we fitted a linear model (n = 23 years) to assess the relative importance of population-level MD (MD_P) and MO (MO_P) on s', weighting data points by the number of recruits produced in a given year to account for the uncertainty in the estimation of s'. We also fitted height of the caterpillar distribution at its peak (HCP) because the expectation was that the effect of the overlap coefficient (MO_P) needed to be corrected for the loss of dimension in the rescaling process. We compared five models: (i) "MD_P"; (ii) "MO_P"; (iii) "HCP + MO_P"; (iv) "MD_P + MO_P; (v) "MD_P + MO_P + HCP". MD_P is different from MD_b in the GLMM in that it is the estimated, rather than the observed, MD. This is because females make the decision to start egg laying approximately a month before nestling demands peak (Visser et al. 2004a); some nests may fail well before that time, precisely because they mistimed their reproduction. We defined population MD_P as the population-mean laying date in that year plus 33 days (see ref. Chevin et al. 2015) minus the caterpillar peak date, where negative and positive values of MD_P indicate that the population bred on average too early or too late, respectively, with respect to the peak date of caterpillar biomass. The overlap coefficient (MO_P) was signed such that it matched the sign of MD_P. The reason we signed MO_P is that it should matter for the selection differential whether overlap was in a positive or negative direction. The relative importance of both metrics was judged using AIC_c as above. Significance of MD_P and MO_P was assessed using the bootstrapped 95% CIs as above (i.e. BC_a, 1000 iterations).

Acknowledgements

We thank the board of the Hoge Veluwe National Park for allowing us to do fieldwork in all these years, and the countless number of assistants, students and volunteers for their help in collecting the data. Jan Visser and Louis Vernooij carefully managed the long-term database. This work is supported by an ERC Advanced Grant (339092 - E-Response) to M.E.V.

Supplementary Information S5

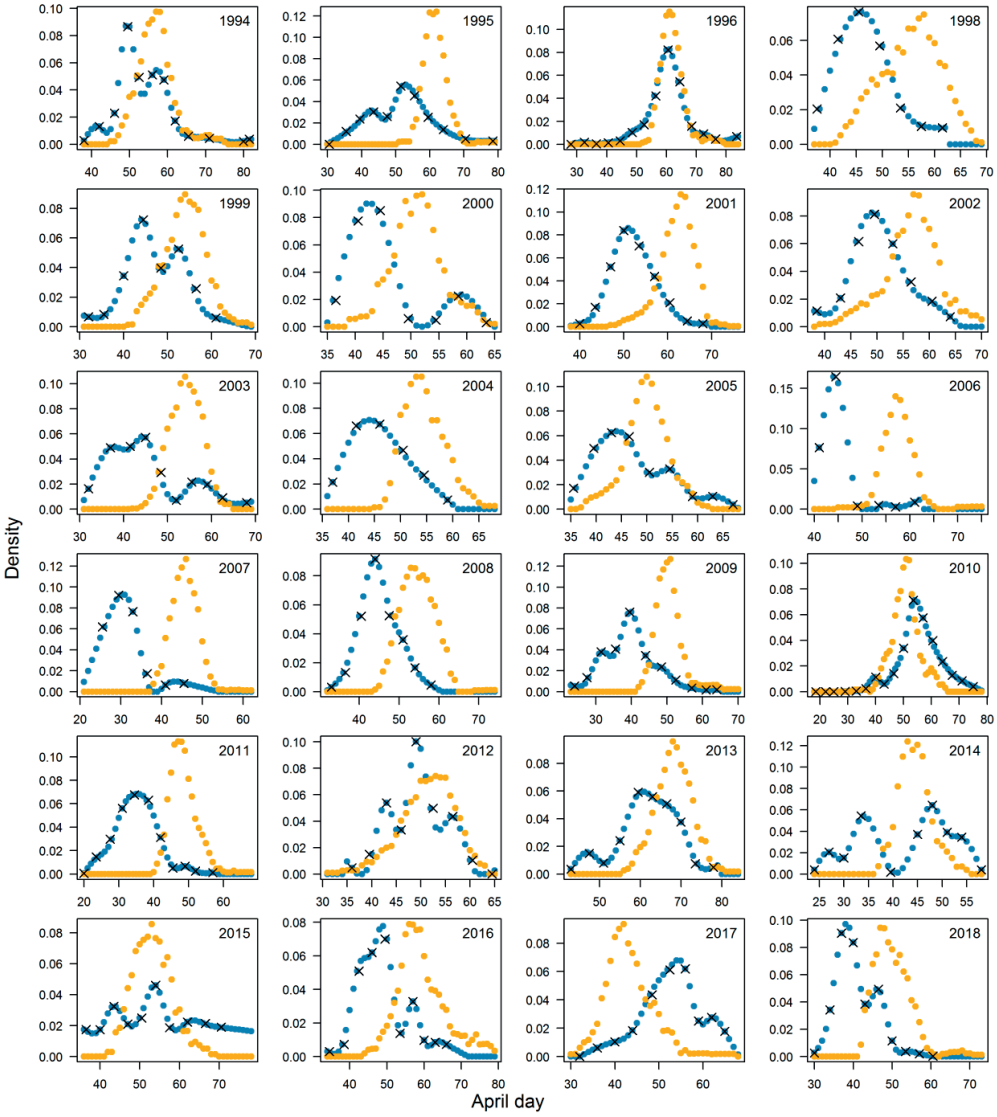


Figure S5.1. *Relative availability of caterpillars (blue dots) and food requirements of Parus major nestlings (orange dots) throughout the breeding season in 24 years at the Hoge Veluwe National Park. Values are scaled such that the area under the curve equals 1 for both curves (see main text for details). Black crosses indicate actual caterpillar sampling days.*

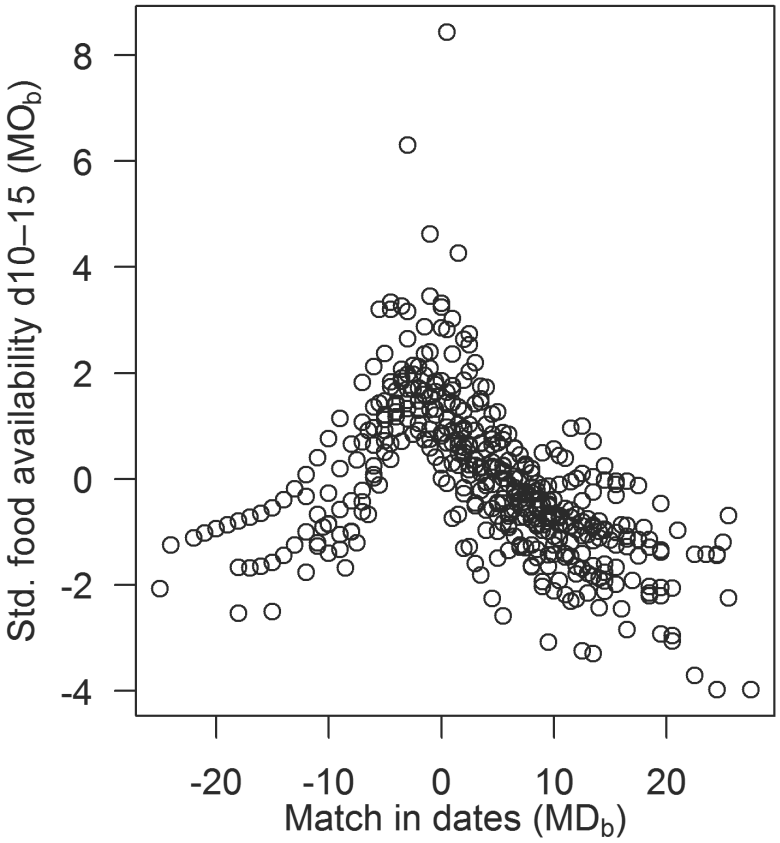


Figure S5.2. *Brood-level, within-season-standardized food availability from nestling day 10 to 15 (i.e. brood-level MO) plotted against the brood-level match in peak dates (caterpillar peak date – date at chick age 10d). Broods that were well matched with the food peak ($MD_b = 0$) had access to most food relative to other broods that were either too early ($MD_b < 0$) or too late ($MD_b > 0$).*

INTERMEZZO

Methodologies in ecology and evolution

Chapter 6

Quantifying individual variation in reaction norms: mind the residual

Jip J.C. Ramakers, Marcel E. Visser & Phillip Gienapp

ABSTRACT

The study of phenotypic plasticity is a central topic in ecology and evolution. Individuals may differ in the degree of plasticity to the environment at the phenotypic (individual-by-environment interaction or I×E) or at the genetic level (G×E), which has implications for the capacity of populations to respond to selection. The number of studies of plasticity in behavioural or life-history traits from wild populations is increasing with the advent of powerful random regression models (RRMs) available to ecologists. Evidence for the presence of I×E or G×E is however mixed, differing between species, populations, and even between studies on the same population. One important source of discrepancies between studies lies in the treatment of heteroscedasticity in residual variance. To date, there seems to be no collective appreciation of its influence on the estimation of I×E and G×E or a consensus on how to best approach it among ecologists. To this end, we performed RRMs with differing residual variance structures on simulated data with varying degrees of heteroscedasticity, sample size and environmental variability to test under which scenarios RRMs would be able to correctly identify I×E. The chosen residual structure in the RRMs affected the precision of estimates of I×E as well as the probability of statistically detecting it, with substantial overestimation and lack of precision when sample size, environmental variability and variance in I×E were small. We show that model comparison using information criteria (e.g. AIC or DIC) can be used to choose the best residual structure, and reinforce this point by analysis of real data of two study populations of great tits (*Parus major*). We stress, however, that small sample sizes can be problematic and RRMs with heterogeneous residual variances on such data may be prone to overfitting. We provide a set of guidelines that can be used by ecologists studying of I×E (and G×E) that, ultimately, will hopefully lead to a reduction in bias in the literature.

Manuscript submitted

Introduction

Behavioural and evolutionary ecologists have long been interested in studying within-individual variation in animal behaviour and life history (Piersma and Drent 2003; Dingemanse et al. 2010). For example, the amount of parental care may be altered by offspring needs and explorative behaviour may depend on the time of day (Dingemanse et al. 2010). Similarly, life-history decisions such as clutch or litter size and timing of reproduction are responsive to the environment, e.g. food availability or local temperatures (Both et al. 2000; Réale et al. 2003b; Brommer et al. 2008; Chapter 10). Many labile traits are thus phenotypically plastic (Schlichting and Pigliucci 1998; Pigliucci 2001), and this plasticity is thought to be adaptive in many contexts because it allows organisms to track phenotypic optima that vary with the environment (Scheiner 1993).

Phenotypic plasticity can be described by reaction norms (Woltereck 1909), that is, the function describing the phenotypic response the environment. Often, these reaction norms are (assumed to be) linear, meaning that they can be described by an intercept or elevation (i.e. the trait value in the average environment) and a slope (i.e. the sensitivity of the trait to the environment), although in certain cases the phenotypic response to the environment may be nonlinear and hence have a reaction norm with an additional component related to 'curvature' (Morrissey and Liefting 2016). Animals may differ consistently from their conspecifics in their mean behaviour (reaction norm elevation) across contexts ('behavioural syndromes' or 'animal personality'; e.g. Réale and Dingemanse 2010; Dall et al. 2012), but they may also differ in the degree of phenotypic plasticity (individual-by-environment interactions or I×E, caused by differences in slopes between individuals), leading to changing variances in phenotypic expressions across the environmental gradient (Nussey et al. 2007). When these variances have a genetic basis (gene-by-environment interactions or G×E) this may impact on how populations can respond evolutionarily to environmental change (Merilä et al. 2001b; Turelli and Barton 2004; Kokko and Heubel 2008; Wood and Brodie III 2016; but see Chapter 9). It is hence important to study variation in reaction norms to understand ecological and evolutionary processes in wild populations (Piersma and Drent 2003; Dingemanse et al. 2010).

Mixed-modelling approaches have been advocated as powerful tools to study individual (or genetic) sources of phenotypic variation in natural populations (Nussey et al. 2007; Bolker et al. 2009; Van de Pol and Wright 2009; Wilson et al. 2010; Dingemanse and Dochtermann 2013). Random regression models (RRMs) are a special case of mixed-effects models that allow different individuals to have different intercepts (trait value in the average environment) as well as different slopes (estimate of the regression of the phenotype against the environment) of the reaction norm (Nussey et al. 2007; Dingemanse and Dochtermann 2013). Phenotypic variance in a particular environment can be partitioned into a component attributable to variance in intercepts and slopes (Morrissey and Liefting 2016). RRMs can be further extended to include an additive genetic effect (e.g. via a pedigree; Henderson 1988; Kruuk 2004) in a so-called 'random regression animal model' (RRAM), allowing one to partition I×E into a permanent-environment (i.e. phenotypic) component (PE×E) and an additive genetic component (G×E). These methods

have been widely used in the evolutionary literature to study the evolutionary potential of a variety of behavioural and life-history traits (see Gienapp and Brommer (2014) and appendix S1 in Van de Pol (2012) for relevant overviews).

The advantages of RRM/RRAM notwithstanding, there are several issues that can lead to misleading conclusions when modelling variation in plasticity (here for simplicity referring to I×E, as opposed to PE×E or G×E). First, in this non-exhaustive list, statistically uncovering I×E greatly hinges on the sampling design of the study and the total sample size (Martin et al. 2011; Van de Pol 2012). Van de Pol (2012) showed that studies reporting evidence for I×E (mostly life-history traits) generally had large sample sizes (typically > 1000), whereas those that did not report evidence for I×E (mostly behavioural traits) had low sample sizes (typically a few hundred). This means that there is quite likely a bias in the behavioural literature towards a lack of I×E attributable to the lack of data. Second, I×E can only be statistically detected if an appropriate (approximation of the) environmental covariate (the 'cue' affecting the phenotype) is studied (Gienapp 2018). For example, Charmantier et al. (2008) did not report I×E for phenology in response to temperature ('warmth sum') in a UK great tit (*Parus major*); however, a reanalysis of the same dataset with a different covariate (mean spring temperature) revealed the presence of I×E (Husby et al. 2010). In the absence of a known environmental driver, phenotypic means have been shown to perform well as a proxy for the environment in the animal- and plant-breeding literature (Lynch and Walsh 1998), and Gienapp (2018) showed by simulation that this method can be deployed effectively in ecology and evolution as well (see Chapter 9 for a practical application). Third, environmental trends in phenotypic variance may be caused by heterogeneity in residual variance and not by I×E. For example, Husby et al. (2010, 2011) reported I×E in phenology in a Dutch population of great tits, but this result could not be replicated by Ramakers et al. (Chapter 10), most likely attributable to the residual structure of the RR(A)Ms.

In this study, we focus on the third problem, i.e. modelling residual variance. We refer to residual variance as the amount of within-individual phenotypic variance left unexplained by the statistical model. This variance is sometimes regarded as 'nuisance' hampering biological predictions, but it has been argued that the residual component in fact contains biologically relevant information (Cleasby and Nakagawa 2011; Nicolaus et al. 2013; Westneat et al. 2015) that, nevertheless, may cause erroneous inferences of variation in plasticity if not appropriately modelled. Nicolaus et al. (2013) found that out of 26 studies of I×E in behavioural and life-history traits, only 5 allowed for heterogeneity in the residual variances (all but one in life-history traits) and concluded for their own study (clutch size in great tits in response to population density) that a RRM with heterogeneous residual variances outperformed a model with homogeneous residual variance. Similarly, Ljungström et al. (2015) found that variation in plasticity in laying date in response to temperature in the sand lizard (*Lacerta agilis*) disappeared when residuals were allowed to vary with the environment. Although sample size in this study might have played a role in the apparent lack of I×E, Ljungström et al. (2015) fitted a residual variance for each environment (year), which may have led to severe overfitting of the model. In contrast, Husby et al. (2010) let residual variances only differ between three year groups in a RRM estimating variation in plasticity in laying date in great tits. The rationale

was that because phenotypic variance in laying date increased with temperature, and temperature increased over time due to climate change, fitting decade-specific residual variances would capture the heteroscedasticity in the RRM, an assumption that was later found to be false (see discussion in Chapter 10). Not all evolutionary ecologists may be aware of this issue or, if they are, may be uncertain as to how to proceed if they suspect their data to have heteroscedasticity.

The 'problem' of heteroscedasticity has long been recognized in the field of animal breeding (Hill 1984). Although the biological importance of the residual variance is increasingly appreciated in the field of ecology and evolution (Nicolaus et al. 2013; Westneat et al. 2015), there appears to be no clear consensus for evolutionary ecologists for assessing whether and how heteroscedasticity may affect estimates of variation in plasticity (I×E) and how it should be dealt with within the context of random regression models (but see Cleasby and Nakagawa 2011). If one is interested in the evolutionary potential of the reaction norm in wild populations (see e.g. Gienapp and Brommer 2014), the main goal is usually to get unbiased estimates of I×E and G×E. To achieve this, behavioural and evolutionary ecologists can make use of advocated mixed-modelling tools (Nussey et al. 2007; Dingemanse and Dochtermann 2013) and use 'basic' random regression models in such a way that it effectively accounts for heterogeneity in residual variances. Alternatively, more sophisticated (and complex) tools are available, e.g. 'double hierarchical generalized linear models' (DHGLM; Lee and Nelder 2006; Rönnegård et al. 2010), which have permeated the ecology and evolution literature to some degree (Westneat et al. 2012; Mulder et al. 2016b) and may be preferred for certain research questions. Here, however, we focus on accurate estimation of variation in reaction norms; fitting heterogeneous residual variances in RRMs should be effective at achieving this but, as pointed out above, a consensus on how to use RRMs in this context and an understanding of how and when heterogeneity in residual variance may affect the accuracy of estimates of I×E (or G×E) is lacking.

In this study, we use a (non-exhaustive) simulation approach to investigate how estimates of I×E, and the statistical power to detect it, are affected by heterogeneity in residual variance that is not appropriately accounted for in the model, and how this effect is mediated by several different characterisations of the data, including the number of environments, the number of observations per environment, the variability in the environment, the amount of variation in reaction norm slopes, and the strength of the association between residual variance and the environment. The aim here was to illustrate in which contexts heteroscedasticity is likely to be problematic in the estimation of I×E and how different residual structures in the random-regression model deal with this heteroskedasticity. Next, we show by simulation how a conventional tool for model selection (Akaike's Information Criterion) performs in detecting heteroskedasticity in the absence of I×E and vice versa. Previous simulation studies have demonstrated how sampling design and size (Martin et al. 2011; Van de Pol 2012) and the choice of the environmental covariate (Gienapp 2018) affect the statistical power and predictive accuracy in detecting I×E, and we therefore do not fully explore the details of these aspects here. Finally, we tested how the methodology applied in the simulations perform in the analysis of the phenology (egg-laying dates) of two long-term study populations of the

great tit (*Parus major*). We use the results of our simulations and empirical analysis to extend existing guidelines for students of behavioural and life-history phenotypic plasticity using random-regression models by shifting the focus on heterogeneity in residual variances.

Methods

Random regression models

A univariate mixed model describing the relationship between trait z and environment x can be written as

$$z_{ij} = a + a_i + bx_{ij} + e_{ij}, \quad (6.1)$$

where z_{ij} is the j^{th} phenotype of the i^{th} individual and the linear function of z_{ij} on environment x_{ij} is characterised by the population-mean intercept a plus the individual deviation a_i, the population-mean slope b and the error term e_{ij}. The random intercept a_i and the error term e_{ij} are both drawn from a normal distribution, where $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. In random regression models (RRMs), each individual is allowed to not only have a different intercept, but also a different slope b_i, so that eqn. (6.1) is rewritten as

$$z_{ij} = a + a_i + (b + b_i)x_{ij} + e_{ij}, \quad (6.2a)$$

where a_i and b_i are assumed to be drawn from (multivariate) normal distributions such that

$$\begin{bmatrix} a \\ b \end{bmatrix}_i \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{a,b} \\ \sigma_{a,b} & \sigma_b^2 \end{bmatrix}_i \right).$$

The error term in eqn. (6.2a) can be drawn from a univariate normal distribution as above, but may sometimes itself be described by some function of the environment such that

$$z_{ijk} = a + a_i + (b + b_i)x_{ijk} + e_{ijk}, \quad (6.2b)$$

where k denotes a group categorizing similar environments (e.g. groups of years with low, intermediate and high temperatures). The error term is then assumed to be drawn from independent, multivariate normal distributions such that

$$e_{ijk} \sim N\left(\begin{bmatrix}0\\ \vdots \\ 0\end{bmatrix}_{ij}, \begin{bmatrix}\sigma^2_{e,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{e,k}\end{bmatrix}_{ij}\right).$$

Note that in reality, e is more likely to vary with x in a more continuous and gradual fashion (whether linearly or not). When error variance (σ^2_e) varies with x in a directional fashion (e.g. a linear increase of decrease), the model of eqn. (6.2a) will likely fail to estimate variation in reaction norm slopes (σ^2_b) accurately (i.e. the estimate will be inflated because the RRM may 'force' reaction norms to merge at one end of the range of x and expand at the other). The model of eqn. (6.2b) should in this case be more appropriate. In empirical datasets, however, we can measure the association between phenotypic variation (σ^2_z) and the covariate of interest (x) but it will be unclear from the surface whether this association is attributable to heterogeneity in σ^2_e, σ^2_z or both.

Simulation objective 1: effect of residual variance structure on estimates and detection rates of I×E

We tested with simulated data whether the estimation of variance in reaction norm slopes, as well as the statistical power to detect it, differed between models with a homogeneous and heterogeneous residual structure. Specifically, we tested whether this difference was mediated by the following factors (see Table 6.1): (1) the mean number of observations per individual (N_o), (2) the total number of different environments (N_x), (3) the variability in the environment (σ^2_x), (4) the variation in slopes (σ^2_b), and (5) the stepwise deviations in environment-specific residual variances ($\Delta\sigma^2_e$) from the mean σ^2_e across environments, used to create a correlation between phenotypic variance (σ^2_z) and the environment (x) due to heterogeneity in σ^2_e (see below). Every combination of parameters (Table 1) was simulated 2000 times.

Environments (X) were randomly drawn from a normal distribution, $x_j \sim N(0, \sigma^2_x)$. Each environment was grouped into one of five equal-interval classes k and received an associated expected σ^2_e using equal-interval deviations from 10 ($\Delta\sigma^2_e$; Table 6.1) to create an association between phenotypic variance and the environment, such that

$$e_{ijk} \sim \begin{cases} N(0, (10 - 2\Delta\sigma^2_e)) \text{ if } x_{ijk} \in [\min(X), N_x/5) \\ N(0, (10 - \Delta\sigma^2_e)) \text{ if } x_{ijk} \in [N_x/5, N_x/5 \cdot 2) \\ N(0, (10)) \text{ if } x_{ijk} \in [N_x/5 \cdot 2, N_x/5 \cdot 3) \\ N(0, (10 + \Delta\sigma^2_e)) \text{ if } x_{ijk} \in [N_x/5 \cdot 3, N_x/5 \cdot 4) \\ N(0, (10 + 2\Delta\sigma^2_e)) \text{ if } x_{ijk} \in [N_x/5 \cdot 4, \max(X)] \end{cases},$$

where X is the vector containing all environments. For example, $\Delta\sigma^2_e = 0.1$ yields expected $\sigma^2_{e,j}$ of 9.8, 9.9, 10.0, 10.1 and 10.2 for the five environmental groups (i.e. low heterogeneity in σ^2_e), whereas $\Delta\sigma^2_e = 2$ yields expected $\sigma^2_{e,j}$ of 6, 8, 10, 12 and 14 (i.e. high heterogeneity

***Table 6.1**. Parameter input in the simulation testing the effect of the residual variance structure in the RRMs to detect variation in reaction norm slopes.*

Parameter	Description	Tested values
1. N_o	Number of observations per individual	2, 5
2. N_x	Number of different environments (years)	20, 40
3. σ_x^2	Variance in the environment	1, 2, 3
4. σ_b^2	Variation in reaction norm slopes	0.003, 0.3, 1.0
5. $\Delta\sigma_e^2$	Equal-interval deviation of four residual variances (σ_e^2) from the average σ_e^2 across environments (here 10), used to create a correlation between phenotypic variance (σ_z^2) and the environment (X). For example, if $\Delta\sigma_e^2 = 0.1$, expected σ_e^2 in five (k; eqn. 6.2b) groups of environments is 9.8, 9.9, 10.0, 10.1, 10.2, i.e. low heterogeneity in realised σ_e^2. (The tested values for $\Delta\sigma_e^2$ led to mean [95% CI] realised correlations between σ_z^2 and X of 0.04 [–0.53, 0.56], 0.18 [–0.42, 0.65], 0.35 [–0.273, 0.74] and 0.56 [0.00, 0.85], respectively, calculated across scenarios.)	0.1, 0.5, 1, 2

in σ_e^2). We opted for this method because the alternative, i.e. drawing $\sigma_{e,j}^2$ based on a given correlation with x_j (e.g. Gienapp 2018), did not strongly drive the correlation between σ_z^2 and x, which is what we were ultimately interested in (see Table 6.1 for the mean realised correlations between σ_z^2 and x). Each individual (N = 500) with N_o observations was randomly assigned to a breeding cohort within the range of X. Individuals randomly received a value for the intercept (a_i) and slope (b_i) (population mean = 0 for both) and their phenotypes in environment x_j were determined following eqn. (6.2b), with e_{ijk} drawn from the k^{th} environmental group as described above. We varied σ_b^2 (Table 6.1) but fixed σ_a^2 to 3; the covariance between a and b was assumed to be zero. The three scenarios for σ_b^2 were chosen based on the estimates gained from studies listed in Table 3 in Nicolaus et al. (2013), which we used to derive the slope variance in proportion to the intercept variance. That is, for all studies that fitted a model on data on the original (non-standardized) scale and reported estimates of $\hat{\sigma}_a^2$ and $\hat{\sigma}_b^2$ (20 pairs of estimates from 6 studies) we divided the $\hat{\sigma}_b^2$ by $\hat{\sigma}_a^2$ and deduced from that 0.001, 0.1 and 0.33 as small, intermediate and large proportions of slope variance in relation to intercept variance (resulting in input σ_b^2 values of 0.001·3, 0.1·3 and 0.33·3, respectively; Table 6.1).

With simulated environments and phenotypes in place, we fitted RRMs with five different variance structures, using the package 'nlme' (Pinheiro et al. 2017). Model 1 had homogeneous residual variance (eqn. 6.2a); the residual structure in the next four models were variations of eqn. (6.2b). For Model 2 and 3, environments were categorized into N_x/5 or N_x/10 equal-interval groups of similar environments, respectively, and estimated residual variance $\hat{\sigma}_e^2$ was partitioned accordingly to capture environmental trends. For Model 4 and 5, environments were again categorized into N_x/5 or N_x/10 groups, but this grouping was done based on consecutive environments, rather than similar environments

(i.e. ignoring the association between σ_e^2 and x); since X was drawn from a random distribution, the grouping thus occurred randomly. Models 4 and 5 served as 'controls' to test whether a heterogeneous residual structure per se affects model performance (note that the number of degrees of freedom, i.e. the difference in the number of parameters, increases with each additional residual variance).

From each model we extracted the estimated variance in reaction norm slopes ($\hat{\sigma}_b^2$). To test the significance of the variance in slopes, we compared each model to a random-intercept model (but keeping the same residual variance structure) with a likelihood-ratio test with 1 degree of freedom. We extracted the proportion of tests with $p < 0.05$ from the 2000 simulation runs.

Simulation objective 2: distinguishing heterogeneous residual variance from I×E

When environmental heterogeneity in phenotypic variance (σ_z^2) is present in the data, the question is whether RRMs can be used to disentangle whether this is caused by heterogeneity in σ_e^2, I×E (fanning reaction norms), or both. In the second simulation, we repeated the analysis of above but focused specifically on relative model performance. We fixed N_0 to 5, N_x to 40 and σ_x^2 to 2. We simulated six scenarios, i.e. all combinations of $\sigma_b^2 =$ 0.01 or 0.7 and $\Delta\sigma_e^2 = 0.1$, 1 or 2 (i.e. low to high correlation between x and σ_z^2; Table 6.1), and assessed relative model performance using Akaike's Information Criterion (AIC; Burnham and Anderson (2002)). The rationale was that if, for example, heterogeneity in σ_e^2 was present but I×E was not, a RRM with a homogeneous residual structure (eqn. 6.2a) may perform better (i.e. have a higher penalized likelihood) than a random-intercept model that incorporated a heterogeneous residual structure. In such a scenario, one would erroneously conclude that there was I×E while in reality there was not. Note that the reverse could equally be true.

We fitted Models 1 to 3 as well as their random-intercept counterparts as described above. For simplicity, we regarded the best fitting model as the most parsimonious one (i.e. with the fewest degrees of freedom) within 2 AIC units from the model with the lowest AIC value (but see caveats in Richards 2005, 2008; Burnham et al. 2011).

Applying RRMs with different residual structures to real data

As a last step we aimed to illustrate how different treatments of the residual variance in RRMs affected estimates of I×E in real data, and how model selection criteria in this context can provide misleading conclusions as to the presence of I×E. We used data of egg-laying dates of first clutches in two of our long-term study populations of the great tit (*P. major*): that of the Hoge Veluwe (HV; 52°01'57"N 5°52'05"E) and the Dutch island of Vlieland (VL; 53°18'N, 5°03'E). In these populations, great tit breeding has been monitored using ~400 and ~500 nest boxes, respectively, since 1955. Briefly, every season (April–June) boxes are checked at least weekly to monitor laying dates, clutch sizes, and number of fledged chicks. Each chick is equipped with a leg ring with a unique identifier, as are the parents, which are captured at the nest box during chick feeding. Most birds (about half) breed only once in their lifetime, although many breed twice or more in subsequent years. Each

year, temperatures are measured in nearby whether stations of the Royal Dutch Meteorological Institute (KNMI; http://projects.knmi.nl/klimatologie/daggegevens/). For HV, this was Deelen station (52°03'N 5°52'12.0"E) throughout the study period. For VL, however, the Vlieland station (53°13'48"N 4°55'12"E) has only been in operation since 1995, so we complemented the data with the nearby Terschelling station (53°22'48"N, 5°21"E), whose measured temperatures correlate strongly with those from the Vlieland station (r^2 = 0.99; Bailey et al., unsubmitted manuscript).

Since the HV study area was reorganised in 1972, we used data from 1973 to 2016; for VL, we used data from 1965 to 2016 (only few breeding records are available from before 1965), with the omission of the years 1981–1985 because of a lack of temperature data. In both areas, brood manipulations were carried out in some years (Both et al. 2000; Postma et al. 2007) but we included these broods in the analyses because they took place during or after clutch completion. In total, we had data on 4890 broods of 3028 females in 44 years in HV and 5250 broods of 3131 females in 47 years in VL. We used spring temperatures (i.e. the mean of daily averages over a specified time window) as the environmental cue for laying date (Gienapp et al. 2005; Visser et al. 2009a; Schaper et al. 2012). We determined the relevant window using a sliding window analysis on population-average laying dates using the 'climwin' package (Bailey and Van de Pol 2017); the best predictive window was from March 11 to April 20 for HV (r^2 = 0.74) and from March 8 to April 21 in VL (r^2 = 0.65; Bailey et al., unpublished manuscript).

With the data in place, we first defined the 'basic' linear mixed-effects model for laying date in our populations in package 'lme4' (Bates et al. 2018). The j^{th} laying date of the i^{th} female in the l^{th} nest box and the h^{th} year is described as

$$z_{ijlh} = a + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij} + \text{nb}_l + \text{yr}_h + e_{ijlh} \quad (6.3a)$$

in the HV population and as

$$z_{ijlhm} = a + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij} + \text{village}_m + \text{nb}_l + \text{yr}_h + e_{ijlhm} \quad (6.3b)$$

in the VL population, where a is the population intercept, a_i is the individual deviation from the population intercept (i.e. a random effect of female identity), b the average slope of the phenotype against the average temperature encountered by individual i ($\bar{T}_i$) and against the individual-centred temperature $\left(T_{ij} - \bar{T}_i\right)$, age_{ij} the female's age (first-year breeder or older) at the time of breeding, nb_l and yr_h the nest box and year, respectively (as random effects), and $e_{ijlh(m)}$ the residual term. In VL (eqn. 6.3b), an additional fixed effect of 'village' (m) was added to denote whether the observation was done within the village on VL or outside of it (birds within the village breed ~5 days earlier). The models of eqns. (6.3a) and (6.3b) (called Model 1) were compared to five different variations on them (Table 6.2): Model 2: partitioning the residual variance as in eqn. (6.2b), dividing the

***Table 6.2**. Model specifications for great tit laying date (z) in the Hoge Veluwe and Vlieland populations.*

Model	Equation	k
1	$z_{ijlh(m)} = a + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijlh(m)}$	1
2	$z_{ijklh(m)} = a + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijklh(m)}$	9
3	$z_{ijklh(m)} = a + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijklh(m)}$	4 / 5
4	$z_{ijlh(m)} = a + a_i + (b + b_i)\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijlh(m)}$	1
5	$z_{ijklh(m)} = a + a_i + (b + b_i)\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijklh(m)}$	9
6	$z_{ijklh(m)} = a + a_i + (b + b_i)\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + \text{age}_{ij}\ (+\text{village}_m) + \text{nb}_l + \text{yr}_h + e_{ijklh(m)}$	4 / 5

Note. k is the number of residual variances estimated, obtained by dividing the number of years by 1 (homogeneous variance), 5 (resulting in 9 groups) or 10 (resulting in 4 or 5 groups in HV and VL, respectively).

number of environments by 5; Model 3: same, but dividing by 10; Model 4: adding an interaction between female identity and individual-centred temperature; Model 5: same as Model 4 but with residual variance partitioned as in Model 2; and Model 6: same as Model 4 but residual variance partitioned as in Model 3 (Table 6.2).

Models were specified in the package 'MCMCglmm' (Hadfield 2010). We opted for this package because the 'nlme' package we used for the simulations does not allow for the inclusion of crossed random effects, and the 'lme4' package does not allow for partitioning residual variances. We used default normal priors for fixed effects, inverse-Wishart priors for the residual variance (V = diag(m) and nu = 0.002, m being the dimension of the matrix, which in this case is k in eqn. (6.2b)) and parameter-expanded priors for the random effects (V = diag(m), nu = m, alpha.mu = 0, alpha.V = diag(m)*625, following Hadfield (2018)). Models were run for a total of $10.1 \cdot 10^6$ simulations, with a burn-in period of 10^5 samples and a thinning interval of 10^4. We report the posterior estimates of slope variance from models 4–6 as well as the Deviance Information Criterion (DIC) for each model as a measure of relative model performance (Spiegelhalter et al. 2002), since there is no Bayesian 'test of significance' like the likelihood-ratio test in a frequentist framework. Hadfield (2018) recommends DIC as the only information criterion for model selection in MCMCglmm, but issues have been raised about using DIC for model comparison in certain contexts (Spiegelhalter et al. 2002; Millar 2009; Hadfield 2018). We therefore used a conservative but reasonable cut-off point of 6 DIC units from the most parsimonious model (ΔDIC). By analogy to frequentist models using AIC, this cut-off point seems to be effective at distinguishing plausible models from those with considerably less support (Richards 2005; Burnham et al. 2011; see also discussion in Spiegelhalter et al. 2002).

Results

Effect of residual variance structure on estimates and detection rates of I×E

As expected, data structure and sample size mediated the effect of the residual variance structure on both the estimates of I×E and the probability of (falsely) detecting it using likelihood-ratio tests. For brevity, we describe here only the scenarios where N_o = 2 and

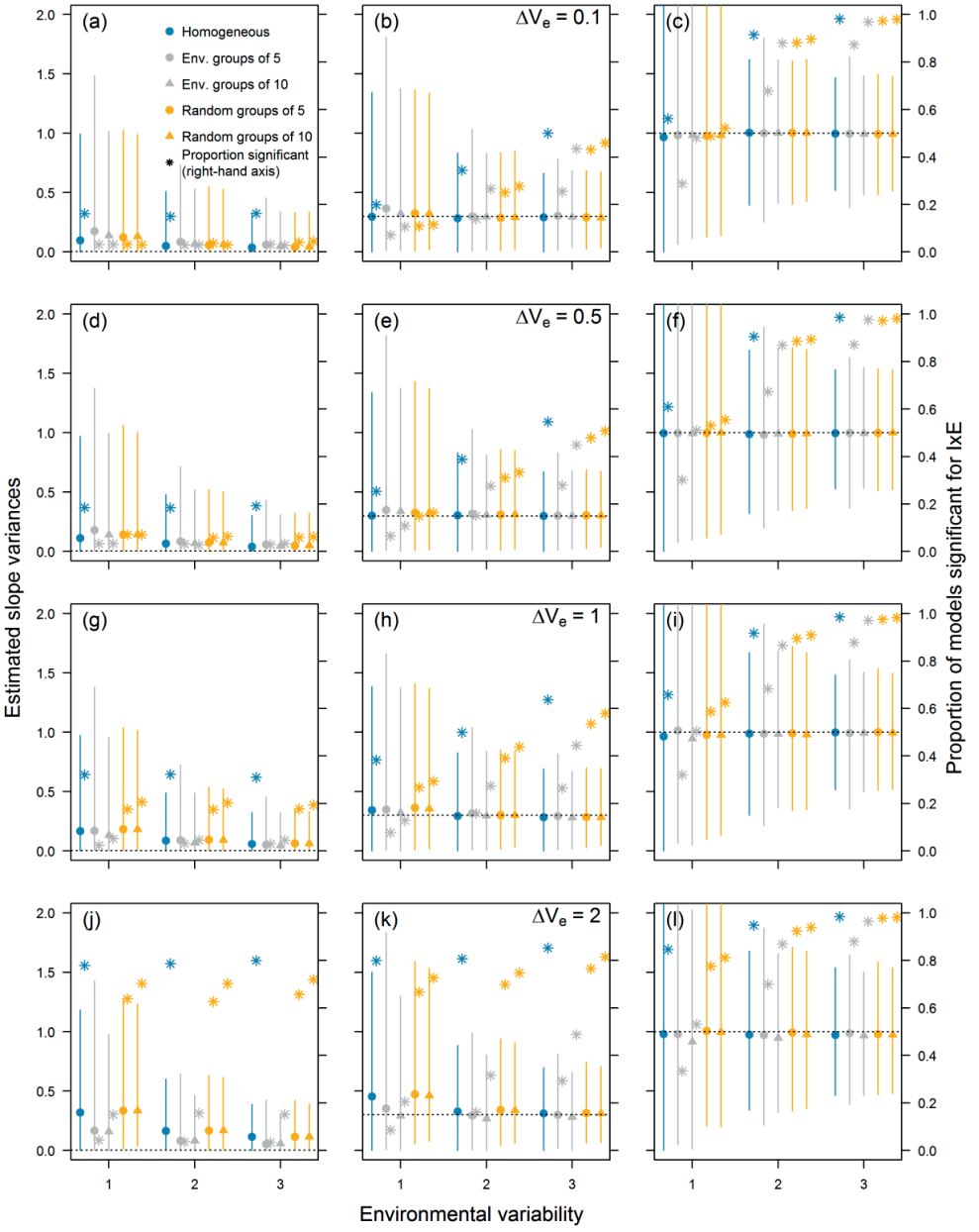


Figure 6.1. *Estimated slope variances (median + 95% CI; left-hand axis) and proportion of significant ($p < 0.05$) models (asterisks, right-hand axis) from different random-regression analyses on different simulated scenarios ($N_o = 2$ and $N_x = 20$ in all panels; see Table 6.1). From top to bottom: the strength of the correlation between phenotypic variance (σ_z^2) and the environment (x) increases through changes in $\Delta\sigma_e^2$ (a–c: 0.1; d–f: 0.5; g–i: 1.0; j–l: 2.0); from left to right: simulated slope variance (σ_b^2) increases (a,d,g,j: 0.003; b,e,h,k: 0.3; c,f,j,l: 1.0), denoted with horizontal dotted lines. The horizontal axis displays the environmental variability (σ_x^2); different colours and symbols display the estimates from models with different residual structures (blue: homogeneous residual structure; grey and yellow: heterogeneous residual structure based on similar environments and through random grouping, respectively, using groups of 5 (circles) or 10 (triangles) environments).*

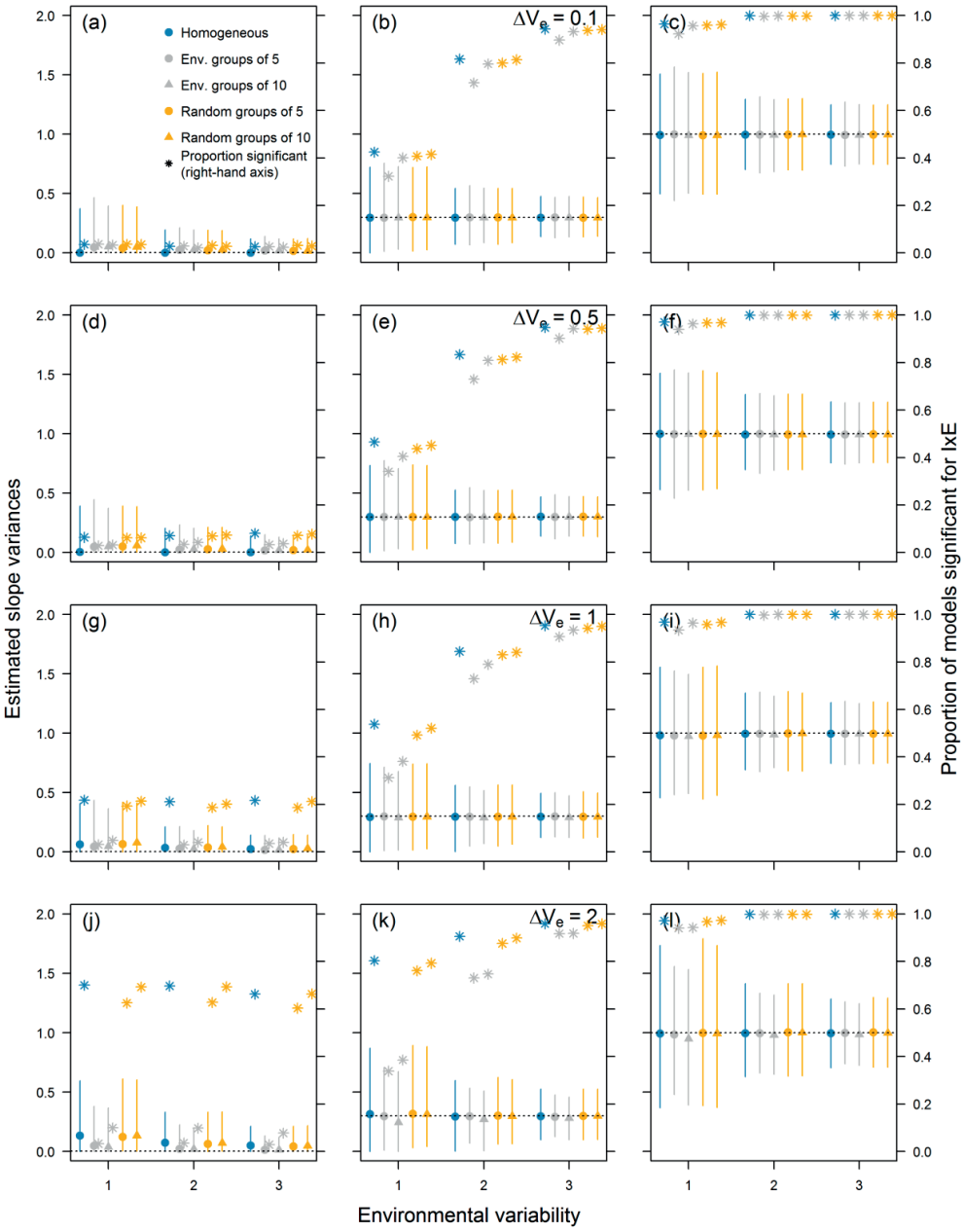


***Figure 6.2**. Estimated slope variances (median + 95% CI; left-hand axis) and statistical power (right-hand axis) from different random-regression models on different simulated scenarios (N_o = 5 and N_x = 20 in all panels; see Table 6.1). See Fig. 6.1 for a description of each panel and the different symbols.*

N_{env} = 20 (Fig. 6.1) and N_o = 5 and N_{env} = 20 (Fig. 6.2). This is because precision and bias in estimates is most affected when sample size is comparatively low (see Supplementary Figs. S6.1 and S6.2 for scenarios where N_{env} = 40). When true slope variance is 0.003, RRMs

consistently overestimate I×E, regardless of the RRM structure deployed (Fig. 1a,d,g,j); this bias decreases across contexts as the environment becomes more variable. As the true heterogeneity in residual variance increases (down the panels in Fig. 6.1), fitting a heterogeneous residual variance structure based on grouped environments reduces the bias in the estimates when the number of groups is low (here two groups of ten environments); that is, the median values move closer to the input value. Fitting more variances (here four groups of five environments) in fact increases the imprecision of the estimates. The same patterns, but less pronounced, can be observed when the input slope variance is of intermediate magnitude (i.e. 0.3; Fig. 6.1b,e,h,k). When input slope variance is substantial (i.e. 1; Fig. 6.1c,f,i,l), median slope estimates almost invariably match the input values reasonably well, regardless of levels of heteroscedasticity and the fitted model, but precision improves substantially as variability in the environment increases. Thus, with a moderate number of environments and 2 observations per individual (Fig. 6.1), the precision of I×E estimates greatly depends on the variability in the environment and when real slope variance is small, failure to fit the proper residual structure may strongly over- or underestimate I×E. An increase in the number of observations per individual (from 2 to 5) can remedy these issues substantially (Fig. 6.2), as can an increase in the number of environments (Figs. S6.1 and S6.2).

Fitting a heterogeneous residual variance structure based on similar environments systematically leads to a reduction in the proportion of models testing significant for I×E (P) when true slope variance is 0.003 ($P \ll 0.2$; left columns in Figs. 6.1 and 6.2). We would therefore rightfully conclude that I×E was absent. Conversely, fitting homogeneous residual variance, or heterogeneous residual variance based on random grouping of environments, increases this proportion as true heterogeneity in residual variance increases, leading to the erroneous conclusion that there is a statistically significant I×E effect. When real slope variance is substantial (i.e. 1), the proportion of significant models is high (> 0.8) in highly variable environments (Fig. 6.1c,f,i,l) and as the number of observations per individual increases, the influence of environmental variability is further reduced (Fig. 6.2c,f,i,l). An exception is when the residual variance is partitioned into environmental blocks of 5: even with high input slope variance, under a low number of observations per individual (Fig. 6.1), 'power' to detect slope variance typically falls below 0.8 when the residual variance is partitioned too excessively. Again, this issue disappears when we have more observations per individual (Fig. 6.2). Concluding, when true slope variance is small and heterogeneity in residual variance is large, fitting the right (heterogeneous) residual structure is crucial to correctly infer statistical evidence for I×E. Moreover, increasing the precision in estimates of I×E and statistical power to detect it when it is there is achieved more easily by increasing the number of observations than by increasing the number of different environments encountered by the same number of individuals (see Figs. S6.1 and S6.2).

Distinguishing heterogeneous residual variance from I×E

Our simulations, for a limited number of scenarios (see Methods), show that whenever there is an association between σ_z^2 and the environment X, simple model comparison using

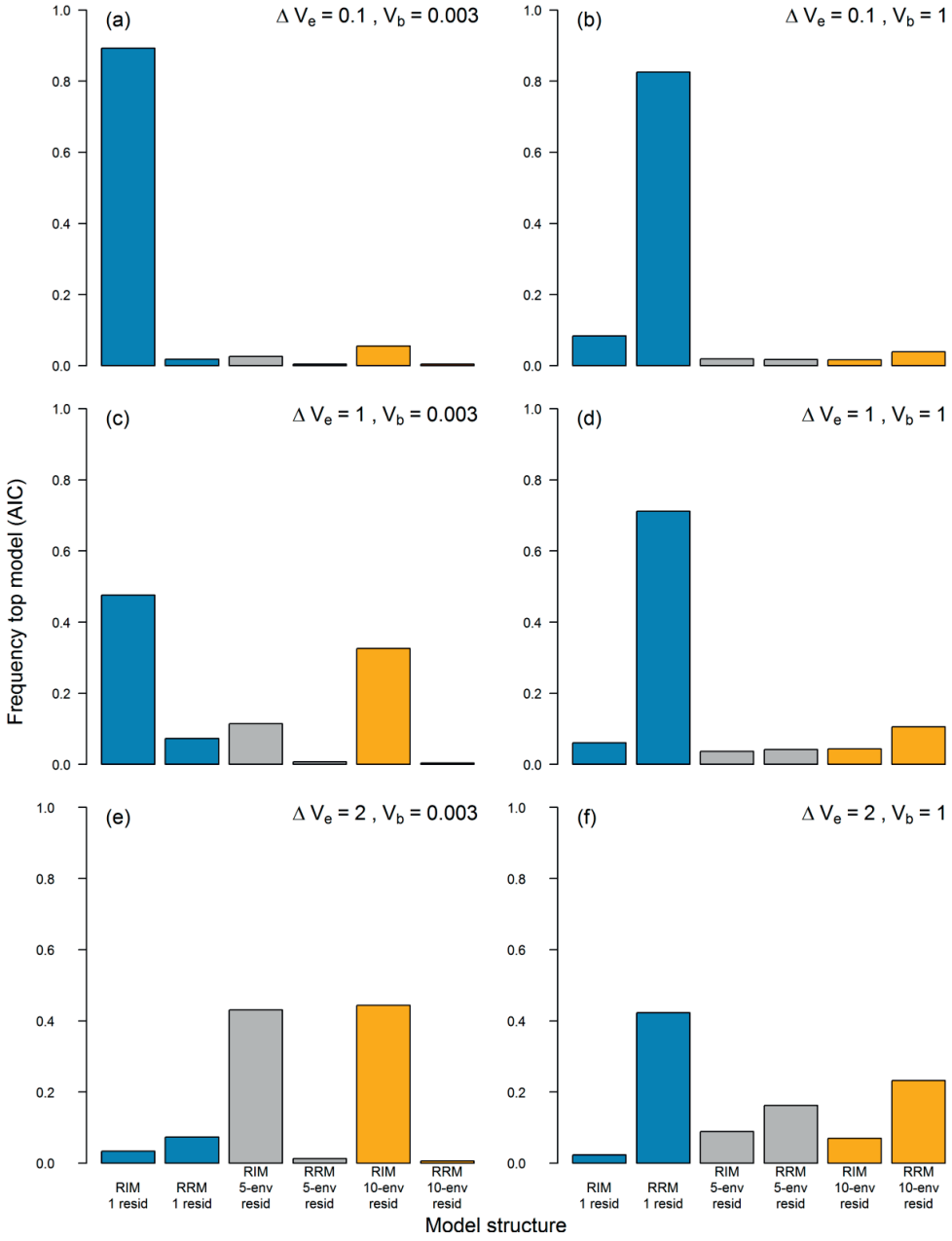


Figure 6.3. *Frequency with which each model is chosen as the top model (based on ΔAIC < 2 and parsimony determined by the total degrees of freedom) under different scenarios (all N_x = 20, N_o = 2 and σ_x^2 = 2), with simulated heterogeneity in residuals ($\Delta\sigma_e^2$) increasing from top to bottom and simulated slope variance (σ_b^2) increasing from left to right. Fitted models (horizontal axes) were random-intercept models (RIM) or random-regression models (RRM) with a homogeneous residual variance structure ('1 resid'; blue bars), heterogeneous partitioned into groups of 5 ('5-env'; grey bars) or groups of 10 environments ('10-env'; orange bars). Note that the meaning of the colours in this figure differs from that in Figs. 6.1 and 6.2.*

AIC is effective at arriving at the qualitative conclusion of whether or not there is statistical evidence for I×E. That is, a combined proportion of > 0.8 of models that appeared as the best model in the selection processes were either RRMs when simulated σ_b^2 was large or random-intercept models (RIMs) when simulated σ_b^2 was small (see Fig. 6.3 for $N_o = 2$ and S6.3 for $N_o = 5$). However, particularly with few observations per individual (Fig. 6.3), selection of the 'correct' model in terms of residual variance structure—which is the one with a structure that matches the simulated data—was achieved at a rate $\ll 0.8$. For example, with an intermediate heterogeneity in residual variance ($\Delta\sigma_e^2 = 1$), models with a homogeneous residual structure were chosen most often (Fig. 6.3c,d). In the most extreme scenario with large slope variance and heterogeneity in residual variance (Fig. 6.3f), both models with and without a heterogeneous residual structure (with 5-env. or 10-env. groups) were selected at competing rates (note that the second blue bar is more or less equally high as the second grey and second orange bar combined). Thus, although qualitatively with respect to the presence or absence of I×E the vast majority of 'best' models was correctly defined (RRMs vs RIMs), this was not the case with respect to the residual structure, which may have implications for (the precision of) estimates of I×E (see discussion).

As expected, increasing the number of observations per individual (Fig. S6.3) improves model selection. At moderate levels of heterogeneity in residual variance, the proportion of selected models having a homogeneous residual variance decreases at $N_o = 5$ compared to $N_o = 2$ (note the strong drop in the blue bars in Fig. S6.3c,d). At high heterogeneity in residual variance (Fig. S6.3e,f), the vast majority of selected models (≥ 0.97) were correctly defined as either RIM or RRMs, respectively, and additionally had a heterogeneous residual structure.

Modelling I×E in great tit laying dates

The two great tit populations differ in the degree of plasticity in laying date with respect to spring temperature (Table 6.3). In HV, the best model arising from DIC model selection was the random-intercept model with a heterogeneous residual structure (Model 2 in Table 6.2). In this population, raw phenotypic (annual) variance in laying dates correlates linearly and positively with mean spring temperature (coefficient + bootstrapped 95% CI: 2.39 [0.702, 4.502]). As the estimate and 95% HPDI for $\hat{\sigma}_b^2$ for Model 5 show, I×E is limited in this population, so the association between σ_z^2 and temperature is not caused by individually differing reaction norms but to other, unmeasured (residual) factors. If we compare RIMs and RRMs while fitting a homogeneous residual structure (Model 1 vs. 4), this conclusion changes radically: now the DIC values suggest a strong preference for Model 4 over Model 1 (ΔDIC = 41.9) with $\hat{\sigma}_b^2$ 4.4 to 4.9 times the size of that of Model 5 or 6.

In VL, the best supported model is Model 4, a RRM with a homogeneous residual structure (Table 6.3). In this population, there is clear evidence for individual reaction norms differing in temperature sensitivity and this evidence is picked up by the RRMs regardless of its residual structure (see $\hat{\sigma}_b^2$ and 95% HPDIs for Models 4–6), concurring

***Table 6.3**. Results of the RRMs on great tit laying dates from the Hoge Veluwe and Vlieland populations.*

Model	Random effects	σ_e^2	Envs. grouped by	No. of residual groups	DIC	ΔDIC	$\hat{\sigma}_b^2$ (95% HPDI)
Hoge Veluwe							
1	Y + NB + I	Ho	44	1	28871.7	159.0	-
2	**Y + NB + I**	**He**	**5**	**9**	**28715.0**	**2.3**	-
3	Y + NB + I	He	10	4	28801.1	88.4	-
4	Y + NB + IxE	Ho	44	1	28829.9	117.1	0.168 (0.018, 0.336)
5	Y + NB + IxE	He	5	9	28712.7	0	0.034 (0.000, 0.123)
6	Y + NB + IxE	He	10	4	28798.2	85.5	0.039 (0.000, 0.135)
Vlieland							
1	Y + NB + I	Ho	47	1	30733.4	867.4	-
2	Y + NB + I	He	5	9	30102.0	236.0	-
3	Y + NB + I	He	10	5	30258.3	392.3	-
4	**Y + NB + IxE**	**Ho**	**47**	**1**	**29866.0**	**0**	**1.893 (1.428, 2.322)**
5	Y + NB + IxE	He	5	9	29885.8	19.8	0.963 (0.428, 1.545)
6	Y + NB + IxE	He	10	5	29905.4	39.4	1.511 (1.032, 2.068)

Note. Y = year, NB = nest box, I = individual, I×E = individual-by-environment interaction, Ho = homogeneous residual variance, He = heterogeneous residual variance. The best models (based on DIC and parsimony) are marked in bold.

with our simulation results (e.g. Figs. 6.1 and 6.2). Importantly, however, the effect size critically depends on the residual structure. Unlike the HV population, raw phenotypic (annual) variances in laying date in VL do not correlate with temperature (–2.932 [–13.880, 1.752] using all years; 0.961 [–1.258, 3.562] when excluding the year 2013 because of an extremely large variance that year). The lack of this association suggests that σ_z^2 covaries nonlinearly with temperature and that this is due to crossing reaction norms and not due to heterogeneity in residual variance, which indeed appears to be the case (see Fig. S6.4).

Discussion

Random regression models are powerful tools to quantify differences in environmentally driven, within-individual phenotypic variation (I×E) across a variety of behavioural and life-history traits and study systems (Nussey et al. 2007; Martin et al. 2011; Van de Pol 2012; Dingemanse and Dochtermann 2013). We have shown by simulation that the precision with which IxE can be estimated is strongly dependent on the level of heterogeneity in residual variance in the data and the way this heterogeneity is subsequently modelled. Importantly, substantial variability in the environment is a prerequisite for reliably estimating—and detecting—variance in reaction norm slopes, although this effect wanes when individuals have observations in many (> 2) environments (cf. Van de Pol 2012).

When these conditions are not met, failure to deal with heteroscedasticity in residuals in an adequate way may strongly impair precision of estimates and the ability of statistical tests to correctly reject or maintain the null hypothesis. We therefore encourage due caution before proceeding to estimate I×E in observational studies (cf. Nicolaus et al. 2013) and suggest an information-theoretic approach to compare the fit of models with different residual structures.

The call for attention to residual variance in (random) regression models is not novel per se in the ecology and evolution literature. Several studies have alluded to both the biological and statistical importance of heteroscedasticity (e.g. Cleasby and Nakagawa 2011; Nicolaus et al. 2013; Westneat et al. 2015). However, in the classic mixed-model 'how-to' paper by Dingemanse and Dochtermann (2013), the implications of heteroscedasticity on model performance and the correct application of alternative methods are not discussed. The same is true for Nussey et al.'s (2007) classic guideline paper for the use of random regression models in studies of phenotypic plasticity. Previous simulation studies that tested the effect of sampling design, sample size, and the choice of the environmental covariate on the performance of random regression models (Martin et al. 2011; Van de Pol 2012; Gienapp 2018) simulated data under the assumption of constant residual variance. Our study adds to previous work by studying heteroscedasticity in a random-regression framework with simulated (and empirical) data with the specific aim to illustrate its effect on model estimates and inference from hypothesis testing.

Cleasby and Nakagawa (2011) perhaps give the most complete practical guidance for ecologists on how to identify and correctly model heteroscedasticity in a standard linear-model framework. They suggested (1) using heteroscedasticity-consistent standard error estimations or (2) or fitting a generalised least-squares model. In their example analysis on experimental data (tarsus length as a function of feeding treatment and sex in house sparrows *Passer domesticus*), the latter was achieved by fitting a residual variance for each treatment–sex combination. This is precisely what we did in our RRMs, with the important difference that the covariate (the environment) is continuous and grouping therefore has to be done 'experimentally' by varying the groups and selecting the most plausible model. Nicolaus et al. (2013) did this by comparing two heterogeneous residual structures when testing variation in plasticity of clutch size with respect to population density and found that partitioning residual variance by environment (i.e. year, as opposed to two groups of environments) yielded the most plausible model. Our simulation results suggest that fitting a heterogeneous residual structure with many groups will be problematic when sample sizes are small (see e.g. the five-environment grouping in Fig. 6.1), potentially due to overfitting of the model. This may also have been the case, for example, in a study on phenology in sand lizards, in which the residual variance in the RRM was estimated for each year (Ljungström et al. 2015). Fitting a homogeneous residual variance in that study led to an estimated slope variance of 10.4 (± 3.4 S.E.) compared to a variance of 4.6 (± 2.4) in intercepts, whereas it decreased to 0 when fitting heterogeneous residual variance. Although the log-likelihood of the model improved considerably compared to a model with a homogeneous residual structure, the best model may actually have been a compromise between the two. Fitting too few groups, on the other hand, may not adequately deal with heteroscedasticity and lead to overestimation of slope variance. We did not explore annual (i.e. for each environment) residual variances in our simulations

because the models were not able to fit them under certain conditions. We therefore strongly suggest that a 'sensitivity analysis' be conducted by changing the number of residual variances stepwise, and judge relative model performance using information criteria. Caution is, however, always warranted when the sample size is low, and it may be reasonable to assume that fitting a residual variance for each environment will result in severe overfitting and potentially erroneous conclusions.

Fitting residual variance for different 'environmental blocks' is an effective way of dealing with heteroscedasticity, but obtaining reliable estimates of I×E naturally starts with the identification of the best 'null' model describing the trait of interest, including the fixed effects on which the variance components are conditioned. Typical reproductive traits such as laying date and clutch size, for example, vary with age. If the phenotypic response to the environment changes with age (A×E; e.g. Van de Pol et al. 2012), individual variation in reaction norm slopes may in fact reflect (unobserved) A×E and not I×E (see discussion in Van de Pol 2012); failing to fit the appropriate age structure in the model may lead to heteroscedasticity and, in turn, to the erroneous conclusion of I×E. Cleasby and Nakagawa (2011) give a comprehensive account of ecological factors generating changes in residual variances across environmental gradients. Their main point, and that of others (e.g. Westneat et al. 2015), is that heteroscedasticity is a perfectly natural biological component of the data that, rather than being just statistical 'nuisance' (Erceg-Hurn and Mirosevich 2008), should inspire researchers to formulate new hypotheses and build their models accordingly.

Recommendations for evolutionary and behavioural ecologists

The results of our simulations (and our empirical data analyses) can be used to draw up a set of guidelines for behavioural and evolutionary ecologists interested in phenotypic plasticity. We acknowledge, again, that we are not the first ones to make recommendations on this topic, an many important recommendations revolving around random regression models and heteroscedasticity more generally have been made by others (Nussey et al. 2007; e.g. Cleasby and Nakagawa 2011; Martin et al. 2011; Van de Pol 2012; Dingemanse and Dochtermann 2013; Nicolaus et al. 2013; Gienapp 2018). It is also important to point out that random regression techniques were originally developed mainly for the field of animal breeding (Henderson 1982; Schaeffer 2004) and developments of tools mainly takes place within this field. There are sophisticated statistical tools available for modelling heteroscedasticity (e.g. so-called 'double hierarchical generalized linear models'; Lee and Nelder 2006; Rönnegård et al. 2010) that may be preferred in some contexts on statistical grounds. We are, however, aware that ecologists may not be sufficiently trained nor have the time our resources to keep up to date with all the latest developments in this complex statistical field, and we would like to present guidelines that can be used within the R environment in software packages and methods that many ecologists will be familiar with (e.g. 'nlme' (Pinheiro et al. 2017), 'MCMCglmm' (Hadfield 2010) and 'ASReml-R' (Butler et al. 2009; Gilmour et al. 2009)).

We assume here that researchers practicing in random regression techniques have at least a basic understanding of linear mixed-modelling procedures in general (see e.g. Dingemanse and Dochtermann 2013) and have a thorough knowledge of the study system so as to incorporate all relevant fixed and random effects. When it comes to random

regression models to estimate I×E (and/or G×E), we suggest the following steps be given sufficient thought:

1. *Plot raw phenotypic variance against the environmental covariate*. Although this may appear trivial, plotting the data prior to analysis can sometimes be quite revealing, because it may give us an idea of whether and how we can expect variances to change with the environment, directionally (e.g. linearly) or not. This may be helpful in deciding if and by how many groups residual variance in the RRM may need to be partitioned. Furthermore, as a 'reality check', we can compare the plot to a plot of individual reaction norms drawn from RRMs (e.g. using 'best linear unbiased predictors' or its equivalents) and visually check if the patterns make sense. In the great tit example, I×E was absent in the HV population despite an association between raw phenotypic variance and temperature, whereas the opposite was true for the VL population.
2. *Compare RRMs with several different residual structures using information criteria.* To our knowledge, there is no clear guideline as to how many residual variances are reasonable, but our simulations suggest that especially when sample size is an issue, more is not necessarily better. In combination with plots of raw phenotypic variance against the environment, the researcher can use informed judgement. A simple approach would be to take the total number of environments (N_x) and divide it by a predetermined number, e.g. by 10, 7, 5, 3, or 1 (i.e. heterogeneous), and by N_x (homogeneous). It should be borne in mind that the more residual groups, the more degrees of freedom are used and the risk of overfitting increases.
3. *Replace the environmental covariate in the RRM with environment-specific mean phenotypes.* When the trait in question does not respond strongly to the environment, estimates of I×E and the power to detect it may be downwardly biased (Gienapp 2018). There may, however, still be undetected I×E and even G×E in the population, which may have implications for the ability of the population to genetically respond to selection. The mean phenotype in a given environment can be used in certain contexts as a substitute for the 'real' environmental driver and in that way serve as a 'yardstick' for testing whether I×E and/or G×E exists in the population (Gienapp 2018; Chapter 9; but see caveats discussed in General Discussion, Chapter 11).
4. *Do a power analysis by simulation.* Whenever the RRM fails to pick up statistical evidence for I×E, the inevitable question arises whether this is due to a true lack of I×E or the lack of statistical power. Simulations can shed light on this. One can simulate a population with differing N_x, N_o, and σ^2_b and play around with parameter values to infer how likely one was to detect I×E in the real data in the first place.

Although not all of the steps may be necessary in every situation, we believe that at least steps 1, 2, and 4 should be carefully considered. Importantly, the chosen residual structure should always be an informed one, and the reader should be informed as to why that particular residual structure was chosen.

Concluding remarks

Random regression models are a powerful statistical tool to estimate phenotypic (and additive genetic) variance across an environmental gradient. Despite its wide use in ecology and evolution, no consensus seems to have reached the research community as to how to treat residual variance. We provide a simulation-informed set of guidelines that students of behavioural or life-history plasticity may adopt to successfully estimate environment-specific individual variances (I×E) and/or genetic variances (G×E). When samples sizes are reasonably large, a simple information-theoretic approach to selecting the best model should help one arrive at the best model explaining the data. We note, however, that when sample sizes are too small, even the most efficient model will not be able to estimate I×E reliably. Defining what is a decent sample size is beyond the scope of this study and has been elegantly demonstrated in previous studies (Martin et al. 2011; Van de Pol 2012). Nevertheless, we encourage researchers to always thoroughly document all statistical procedures (e.g. though R scripts) and report sample sizes, effect sizes and the precision of their estimates, which in the long run will serve the scientific field by enabling biological synthesis across study systems, e.g. in the form of meta-analysis.

Supplementary material S6

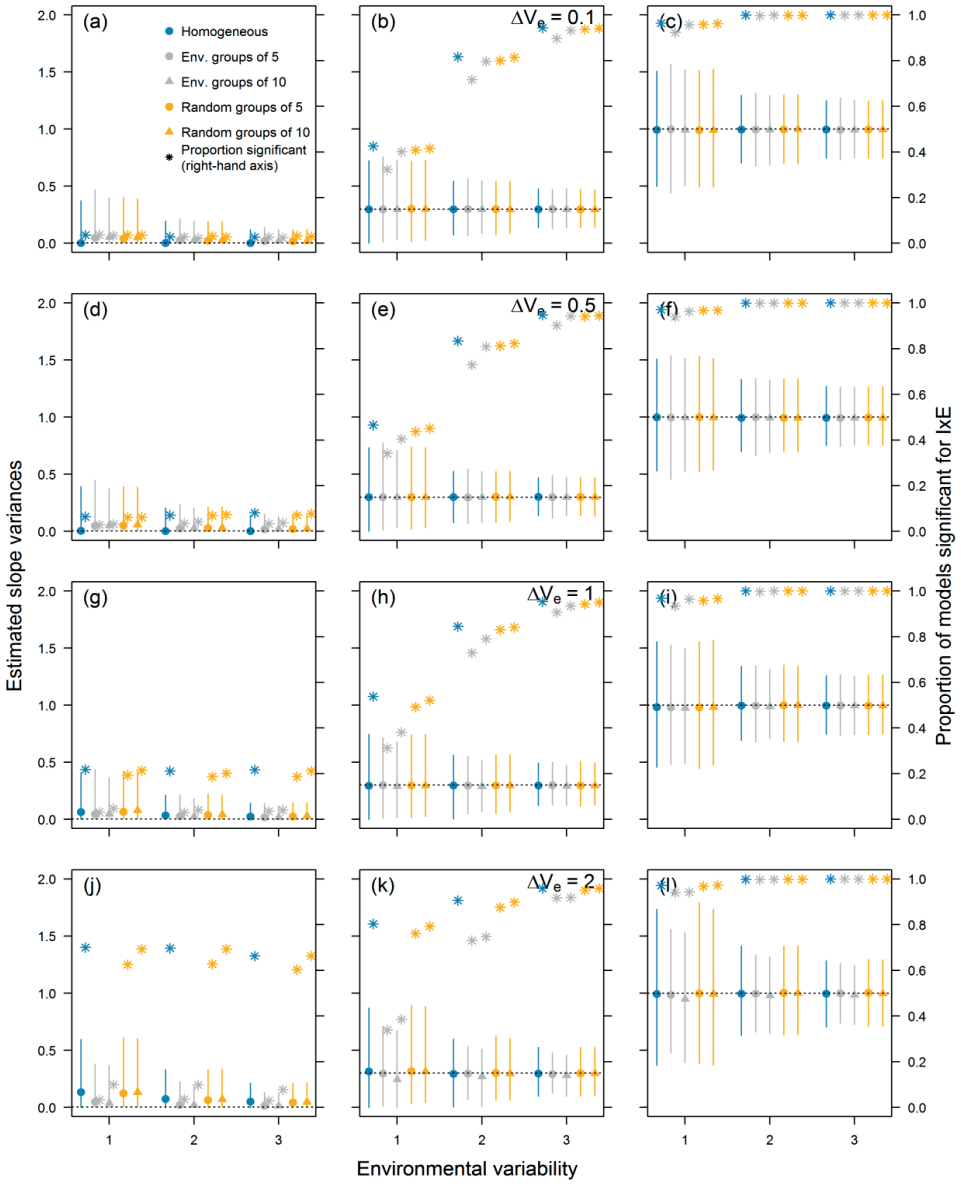


Figure S6.1. *Estimated slope variances (median + 95% CI; left-hand axis) and proportion of significant ($p < 0.05$) models (asterisks, right-hand axis) from different random-regression analyses on different simulated scenarios ($N_o = 2$ and $N_x = 40$ in all panels; see Table 6.1 main text). From top to bottom: the strength of the correlation between phenotypic variance (σ_z^2) and the environment (x) increases through changes in $\Delta\sigma_e^2$ (a–c: 0.1; d–f: 0.5; g–i: 1.0; j–l: 2.0); from left to right: simulated slope variance (σ_b^2) increases (a,d,g,j: 0.003; b,e,h,k: 0.3; c,f,j,l: 1.0), denoted with horizontal dotted lines. The horizontal axis displays the environmental variability (σ_x^2); different colours and symbols display the estimates from models with different residual structures (blue: homogeneous residual structure; grey and yellow: heterogeneous residual structure based on similar environments and through random grouping, respectively, using groups of 5 (circles) or 10 (triangles) environments).*

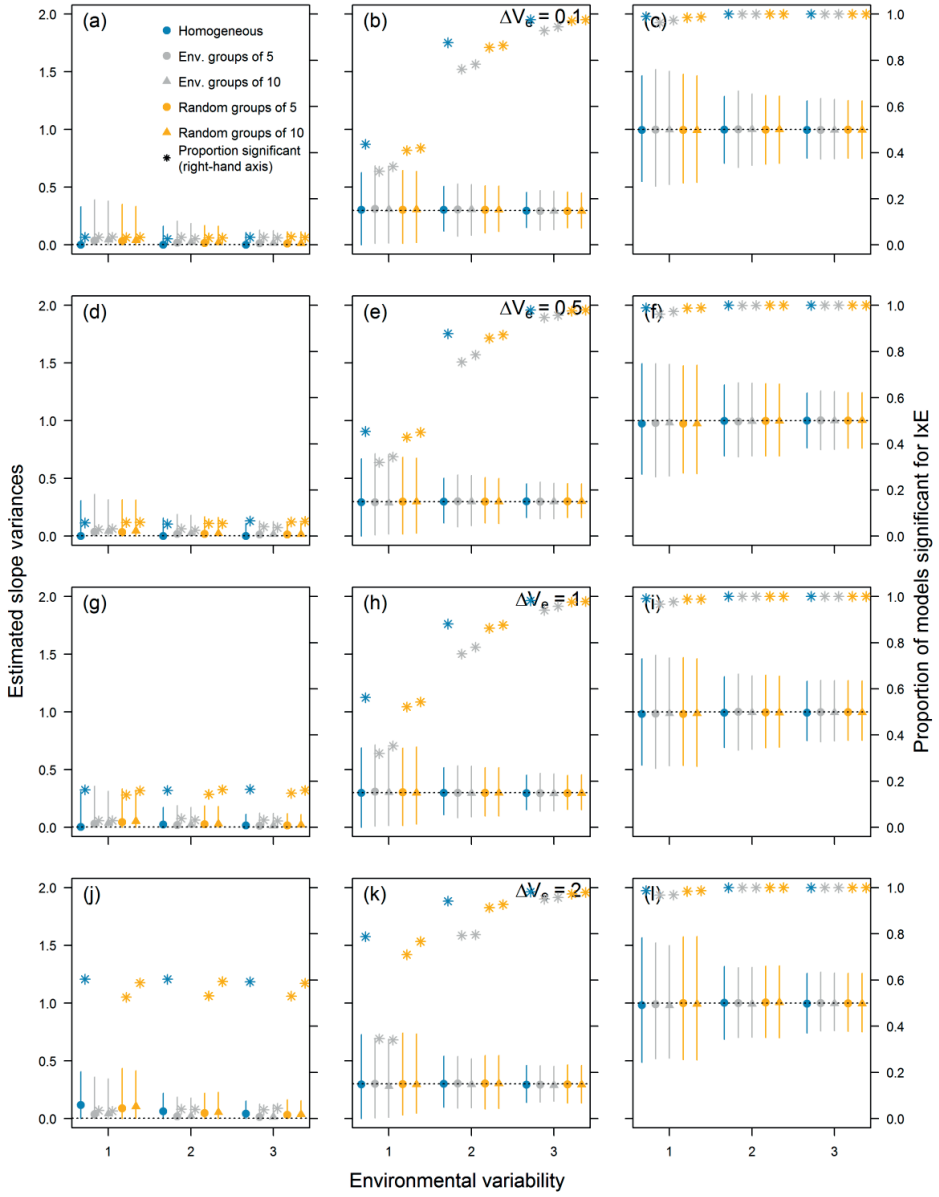


Figure S6.2. *Estimated slope variances (median + 95% CI; left-hand axis) and proportion of significant ($p < 0.05$) models (asterisks, right-hand axis) from different random-regression analyses on different simulated scenarios ($N_o = 5$ and $N_x = 40$ in all panels; see Table 6.1, main text). From top to bottom: the strength of the correlation between phenotypic variance (σ_z^2) and the environment (x) increases through changes in $\Delta\sigma_e^2$ (a–c: 0.1; d–f: 0.5; g–i: 1.0; j–l: 2.0); from left to right: simulated slope variance (σ_b^2) increases (a,d,g,j: 0.003; b,e,h,k: 0.3; c,f,j,l: 1.0), denoted with horizontal dotted lines. The horizontal axis displays the environmental variability (σ_x^2); different colours and symbols display the estimates from models with different residual structures (blue: homogeneous residual structure; grey and yellow: heterogeneous residual structure based on similar environments and through random grouping, respectively, using groups of 5 (circles) or 10 (triangles) environments).*

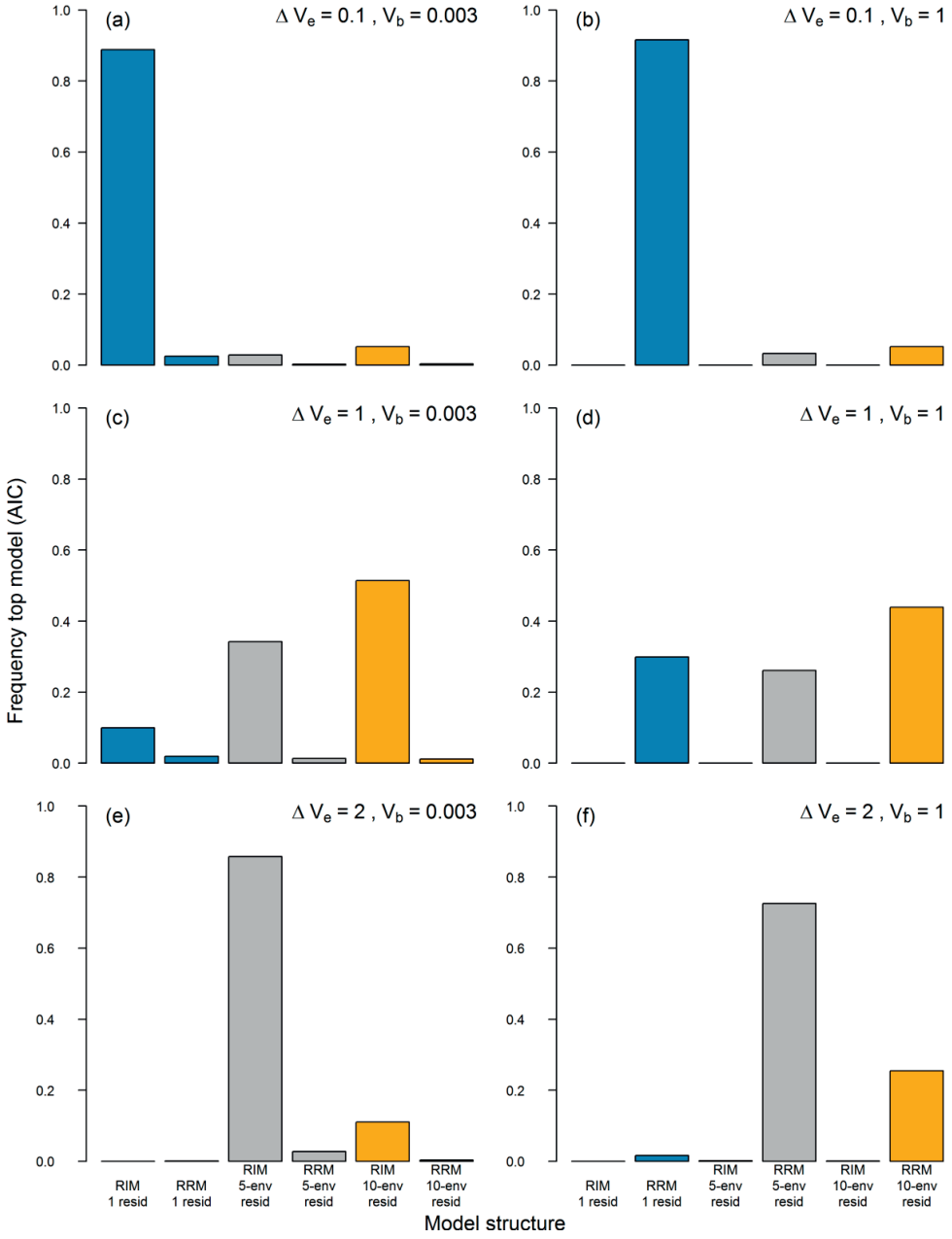


Figure S6.3. *Frequency with which each model is chosen as the top model (based on ΔAIC < 2 and parsimony determined by the total degrees of freedom) under different scenarios (all N_x = 20, N_o = 5 and σ_x^2 = 2), with simulated heterogeneity in residuals ($\Delta\sigma_e^2$) increasing from top to bottom and simulated slope variance (σ_b^2) increasing from left to right. Fitted models (horizontal axes) were random-intercept models (RIM) or random-regression models (RRM) with a homogeneous residual variance structure ('1 resid'; blue bars), heterogeneous partitioned into groups of 5 ('5-env'; grey bars) or groups of 10 ('10-env'; orange bars) environments.*

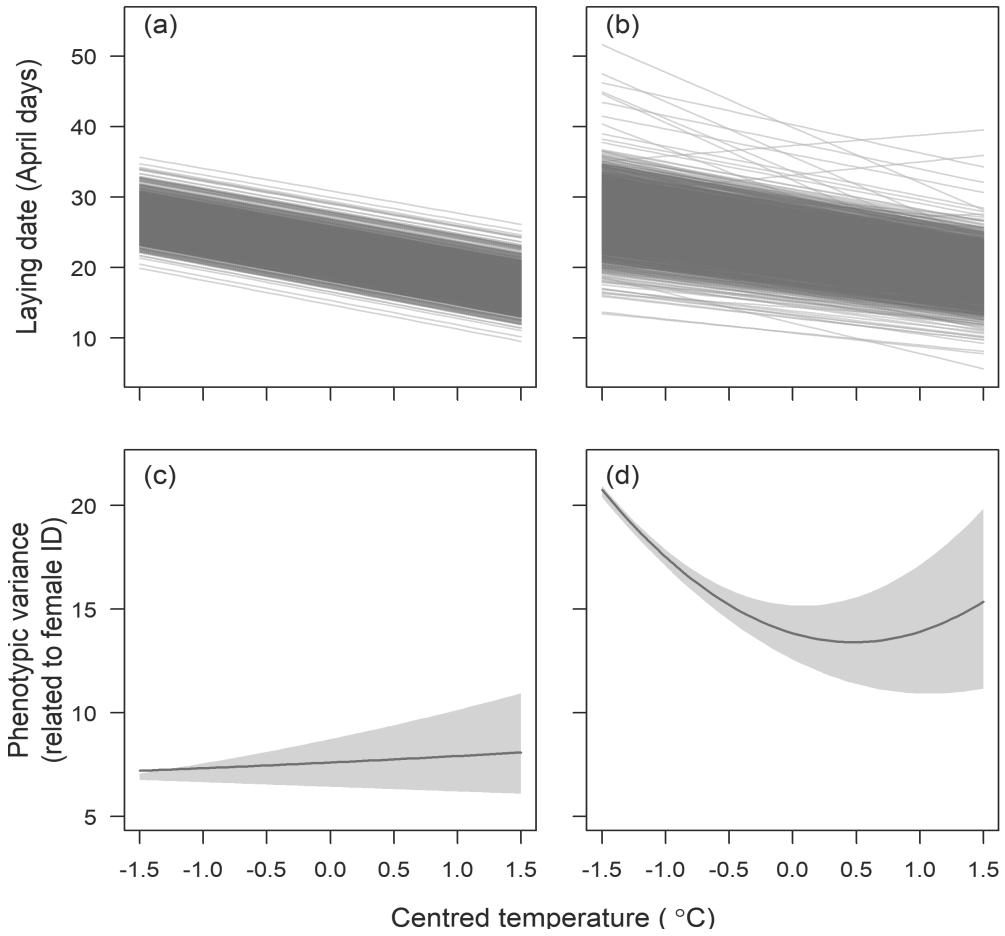


Figure S6.4. *(a,b) Estimated reaction norms (posterior median estimates for each female) and (c,d) estimated phenotypic variance (combined permanent-environment and additive genetic effect; posterior medians and 95% HPDI) for great tit laying dates against individual-centred temperatures in HV (a,c) and VL (b,d). Estimates were obtained from Models 5 (HV) and 4 (VL) in Table 6.3 (main text). Temperature-specific variance estimates (i.e. in the jth environment) were estimated as* $\hat{\sigma}^2_{z,j} = \hat{\sigma}^2_a + 2\hat{\sigma}_{a,b}T_j + \hat{\sigma}^2_b T_j^2$, *with a and b representing intercepts and slopes, respectively, and T representing the (centred) temperature.*

Chapter 7

How to do meta-analysis of open datasets

Antica Culina, Thomas W. Crowther, Jip J.C. Ramakers, Phillip Gienapp &
Marcel E. Visser

Preface

The amount of open data in ecology and evolution is increasing rapidly, yet this resource remains underused. Here, we introduce a new framework and case study for conducting meta-analyses of open datasets, and discuss its benefits and current limitations.

Nature Ecology & Evolution (2018) **2**, *1053–1056*

Introduction

In recent decades, the meta-analysis approach has emerged as the most valuable avenue for scientific progress, along with empirical studies and theoretical models (Husby et al. 2011b; Cadotte et al. 2012; Gurevitch et al. 2018). Traditional meta-analysis combines results from a number of studies (ideally all) conducted on the same research question, to statistically summarize findings, evaluate discrepancies and detect generalizable effects (Gurevitch et al. 2018). The ability to detect overarching patterns makes meta-analyses extremely relevant to evolutionary ecology, which is characterised by highly complex systems, heterogeneous environments and variable methodologies (Jennions et al. 2012; Stewart and Schmid 2015).

Systematic advances in the meta-analysis approach over the last decade have been intended to improve the transparency, replicability, reliability, and impact of data synthesis efforts (Bayliss and Beyer 2015; Lortie et al. 2015; Parker et al. 2016; Gurevitch et al. 2018). However, despite these advances, the major outstanding limitation of any synthesis remains the challenge of accessing a comprehensive range of available data on the topic (Parker et al. 2016). Conventionally, meta-analyses are conducted using effect sizes (i.e. measure of the strength and direction of effects) extracted from the values reported in published studies. These meta-analyses are often limited to studies that focus specifically on the topic of interest (we term these 'target studies'). However, a wealth of useful data is often available in various 'non-target studies' that have attained relevant information to address different research questions. Additional data from 'non-target studies' can enhance the statistical power of meta-analyses (a fact that has been widely accepted and embraced in medical research; Simmonds et al. 2005), as well as considerably reduce current issues with biased effect sizes. These data can be used either on their own, or in a combination with data from targeted studies. Until now, the complex and variable research landscape in ecology and evolution has restricted such data ingestion from non-target studies. However, the increase in data made openly accessible, as now required by many journals, is transforming our capacity to access, evaluate and use raw data from both target and non-target studies. Hence, our potential to survey the data-landscape and gain a comprehensive understanding of the available information has never been greater (Roche et al. 2015). Yet, unlike other scientific fields, this resource remains relatively unexploited in the field of ecology and evolution (Wallis et al. 2013; Evans 2016).

Retrieval of primary data for meta-analysis

Here, we describe how to transparently retrieve and select data, where the information retrieval starts from published (open) datasets, rather than from studies. Our standard is based on existing guidelines for the information retrieval in ecological/evolutionary meta-analysis (Koricheva et al. 2013; Bayliss and Beyer 2015; Lortie et al. 2015; Nakagawa et al. 2017), but adapted specifically for open data. The retrieval and selection process should

be highly transparent - we provide a checklist of the information that needs to be recorded (Table 7.1). This information should ideally supported by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Moher et al. 2009) diagram (Fig. S7.1).

In the **first step** of the approach (Step 1), researchers need to identify the type of data needed to answer the meta-analysis question (or test hypothesis), set appropriate exclusion/inclusion criteria, and choose the search terms (used in a search for the relevant data). This is followed by the data search. In evolutionary ecology datasets are usually scattered across various repositories (e.g. Dryad, Figshare, Zenodo) or published in the supplementary materials associated with a paper. Thus, an effective search should be conducted using data-harvesting platforms that crawl through many different research data repositories that host research data (like Web of Science crawls through journals in a search for articles); some also explore supplementary materials of published papers for additional information. A complete overview of how to navigate the data-landscape by using data-search platforms can be found in Culina et al. (2018). We suggest using DataCite, Base search engine, and DataOne (see below). The original search terms usually need to be adjusted according to the output of the initial search (e.g. when some obviously irrelevant terms appear in the search results). After the initial search, duplicates can be eliminated.

The next steps (second to fourth) describe the screening of the obtained datasets. This starts with screening according to the meta-data (data that describe the dataset) provided by the search platform (**Step two**); these will vary between different search platforms (usually keywords, dataset title, dataset description and/or subject area). Thus, it is important to record and report on which meta-data the screening was based. This step is equivalent to the initial screening of the title, abstract, and keywords in the 'traditional' meta-analysis that starts from published studies. The main difference is that the standards to describe data-sets are less well established than the standards to describe articles (title, abstract, keywords, subject areas). Thus, this screening might be more time-consuming, and lead to the retention of more irrelevant datasets. Next (**Step three)**, each potentially relevant dataset should be opened and screened to identify whether it corresponds to meta-analysis requirements.

The remaining datasets are relevant according to the dataset type, but some will be excluded (**Step four**) as they do not match the specific inclusion criteria or are not fit for use because information crucial to run the desired analysis (to obtain the effect sizes) is missing (equivalent to under-reported effects in the study-centric approach). At this stage, researchers might decide to contact the dataset owner(s).

The final list of datasets is then used to calculate the effect sizes (**Step five**). Ideally, all effect sizes are calculated in the same, standardized way. This process can take several sub-steps. In line with the good scientific practice, and to address the issues of data misinterpretation (Mills et al. 2015), owners of the datasets should be contacted, at the latest, when analysing their data, and asked whether they agree with the way the data was processed (**Step six**). Some data owners specifically ask (in the meta-data files) to be

***Table 7.1**. The checklist of the main steps in conducting meta-analysis that starts from datasets. Following these steps will ensure a transparently conducted meta-analysis, which complies with the current scientific standards.*

Step	What to record (report)
Step 1: What type of data are needed and where/how to obtain them?	• research question/questions • the exact exclusion/inclusion criteria • platform(s) used in search • search terms and syntax (for every platform; whether and how search terms were adjusted)
Step 2: Screening the results according to the metadata provided (keywords, dataset title, the description of the dataset, and/or subject area)	• what meta-data screening was based on • number of excluded results • reasons for exclusion (optional)
Step 3: Open and screen remaining datasets	• number of excluded results • reasons for exclusion (optional)
Step 4: Detailed examination of the datasets. Contacting the authors of the dataset about missing/not clear information	• number of excluded results • reasons for exclusion • whether the authors were contacted and with what outcome
Step 5: Calculate the effect sizes	• statistical procedures to calculate effect sizes
Step 6: Contact the authors to check if they agree with the approach	• contact letter, author responses, dates of contact • datasets excluded based on authors feedback and why
Step 7: Conduct the statistical part of meta-analysis	• the dataset used in meta-analysis • exact models/formulas

contacted directly if there are plans to use the data. Some datasets might be still excluded after this step. Statistical analyses can then be conducted using these effect sizes (**Step seven**) following the existing guidelines (choose an appropriate mode, explore the sources of heterogeneity, account for non-independencies, and, if considered necessary, test for publication bias (Koricheva et al. 2013; Nakagawa et al. 2017)). Statistical analysis can also be conducted using effect sizes calculated from raw data, with those calculated using values reported in published papers (when possible). In this case, information retrieval

protocol should be recorded separately for data and article selection process (Koricheva et al. 2013; Bayliss and Beyer 2015; Nakagawa et al. 2017), and we would further advise on controlling for the source of effect size (data or article) when conducting the statistical analysis.

To demonstrate the information retrieval framework, in Box 7.1 we outline the search for pedigree datasets for the meta-analysis that aimed at evaluating the strength of the evidence for the environmental coupling of heritability and selection (Chapter 9).

***Box 7.1**. Application of the framework: environment, heritability and selection*

While the environment has been shown to influence both selection and heritability of traits (see review in Wood and Brodie III 2016), the number of studies exploring both in the same systems is limited (Wilson et al. 2006; Husby et al. 2011b). As such, there were not enough published studies on this topic to synthesize and generalize the relationship between selection and heritability (Wood and Brodie III 2016). To address this question, Ramakers et al. (Chapter 9) needed data on a) pedigrees or additive genetic relatedness of these individuals, b) individual phenotypes, c) individual fitness, and d) the environment (which we defined by averaging the phenotypes across the population in a given year, and variance-standardizing it across years). We expected that pedigree data represent the limiting source of data. Therefore, we started our search for this type of data deposited in online databases. The details on the data search and data screening process are provided in Chapter 9 and its supplementary material. Here, we provide a summary of the search and data collection process.

After searching through 12 different aggregators of research data repositories (Europe PMC, DataCite, BASE, OpenAIRE, ScienceSerach, DataOne Mercury search, Web of Science Data Citation Index, Scielo, Research Data Australia, DLI Service, Dryad Digital Repository, Data MED), and screening through the original search results (steps 1 to 3), we located 103 animal pedigreed datasets. Different aggregators we consulted identified different parts of the overall sample of datasets (Fig. B7.1.1).

However, after a careful examination of these original 103 pedigree datasets, we were forced to discard 88 either because of a) embargoed data, corrupted or 'encrypted' files, b) insufficient number of individuals with pedigree or c) phenotypic information, d) lack of natural environmental variation in the phenotype (this excluded all laboratory populations), e) too few years included in the data set (at least six years), f) other issues (e.g. non-matching IDs of animals in pedigree and phenotype file) (see Chapter 9, Table S9.1). This left us with 15 datasets for analysis. After analysis, we contacted the original data owners to check whether we had misinterpreted their data, as this was one of the main concerns about the use of open data (Mills et al. 2015). Based on the authors' feedback, we excluded another four datasets for various reasons related to non-random exclusion of individuals from the dataset (potentially leading to biased fitness measurements and quantifications of the environment). The reduction in the overall sample size, from 103 pedigree results obtained, to 11 that we could use in the analysis, drastically reduced a number of taxa and populations represented in the dataset (Amphibians, Fish, Insect, Mammal, Mollusks; Fig. B7.1.2).

We conducted an additional literature search to identify studies that potentially also contained pedigree data (see Chapter 9), identifying three additional datasets yielding data on 49 traits in 15 populations of 9 species in total. Overall, we emailed owners of 18 datasets to check if they agree with the way we analyzed their data. The majority (16) of them offered advice on the analysis (also leading to the exclusion of 4 datasets, see above) and supportive towards our efforts, while two were negative towards the use of their data. After conducting the analysis (including …

***Box 7.1** (continued)*

… heterogeneity analysis), we found that any effects of the environmentally caused coupling between heritability and selection on expected evolutionary response were small.

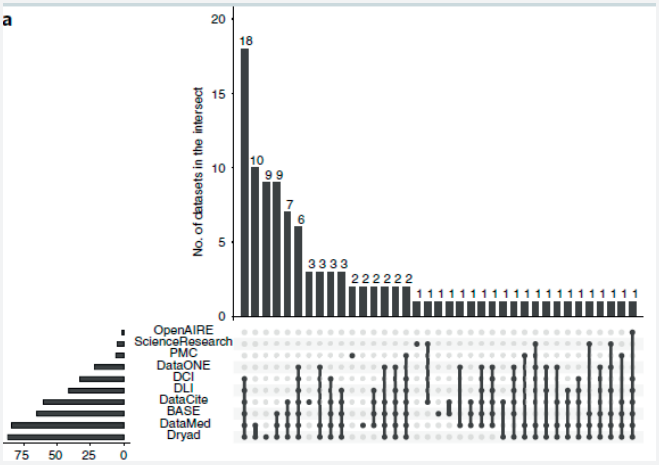


***Figure B7.1.1.** Diagram representing the number of pedigree datasets found by each aggregator (left hand side frequency diagram), and the number of datasets in intersections among aggregators (the main frequency diagram). For example, same 18 datasets were obtained by all: DCI, DLI, DataCite, BASE, DataMed and Dryad search (first horizontal bar), while Dryad search resulted in 9 unique datasets.*

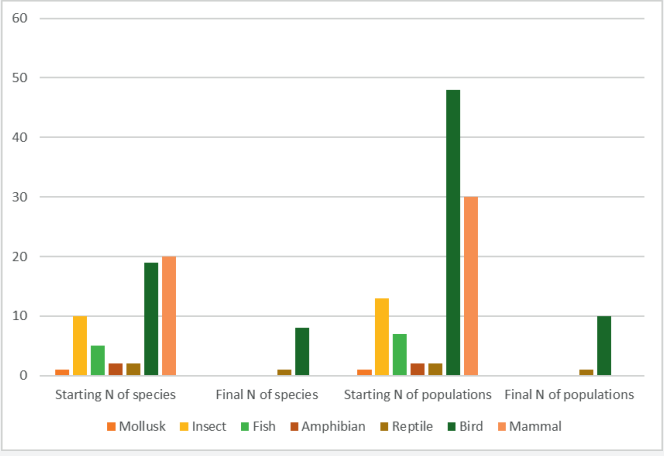


***Figure B7.1.2.** A diagram representing loss of species and populations of different taxonomical groups (Molluscs, insects, Fish, Amphibians, Reptiles, Birds, Mammals) from the original 103 pedigree datasets obtained by search for open pedigree data, to final 11 datasets used in the analysis.*

Benefits of open data to evolutionary ecology meta-analysis

Our case study (Box 7.1) demonstrates an obvious benefit of the information retrieval which starts from published data (rather than published studies): the considerable increase in the data available to conduct meta-analysis (and thus in the amount of research questions that can be addressed; Mengersen et al. 2013). These data can be used on their own to calculate effect sizes for the meta-analysis, or used alongside effect sizes extracted from published studies. In our example a traditional meta-analysis was impossible (only two published studies on the research question, see Box 7.1). Use of open data from studies that themselves addressed another question enabled us to collect enough evidence for meta-analysis. Given that the number of published datasets is greatly increasing across evolutionary and ecological fields (Roche et al. 2015; Culina et al. 2018), the scope of evolutionary ecology meta-analysis can be extended, and not limited only to target studies in the published literature.

An additional benefit of open data is the reduced publication bias that stems from the selective reporting of 'significant' or 'interesting' results (Parker et al. 2016). The under-reporting of weak, negative or unwanted effects (or ambiguous results) is common across scientific disciplines: two reviews showed that basic information (sample size and variance) was missing from generally half of otherwise relevant primary studies collected for meta-analysis in conservation ecology (Côté and Reynolds 2012) and evolutionary ecology (Cassey et al. 2004). Even more worrying is that these under-reported results appear to be a biased sample of all results (Cassey et al. 2004). However, datasets, and effect sizes calculated using published datasets, are less likely to suffer from this kind of issues. Datasets that support published studies can be also used to verify or supplement the results of the study, increasing the number of effect sizes that can be calculated (missing or contradictory reported results).

Finally, meta-analyses conducted using the values reported in studies have to combine effect sizes calculated in a different way (as primary studies analyze their data, and report the results differently). Effect sizes can be calculated in a consistent manner if the original data are used (such as in our case study, Box 7.1), thus leading to directly compatible effect sizes (Mengersen et al. 2013).

Limitations and drawbacks of Open Data in meta-analysis

Despite the apparent benefits, our meta-analysis conducted using non-target research data suffers several limitations. These should not discourage data-driven meta-analysis, but rather be acknowledged, and if possible, adequately resolved.

First, as our case study demonstrates, the description of datasets is often insufficient to enable a sensitive and targeted search. This means that data searches may retrieve a substantial number of irrelevant datasets, whilst also missing some relevant ones. However, this has always been a limitation of meta-analyses, and we believe this will only

improve as the scientific community continues to embrace the advised data standards (e.g. Wilkinson et al. 2016), supported by improvements in the data curation by research institutions, and scientific repositories. The second and related issue is the quality of the retrieved datasets, where a number of datasets might need to be excluded due to the lack of sufficient information. In our case, this reduced the number of species for the analysis, and led to loss of a number of taxonomic groups (Fig. B7.1.2). The third issue is the misinterpretation of data used in meta-analysis (Mills et al. 2015), especially when using non-target studies that addressed different questions from the proposed study. Contacting data owners is probably the best approach to address this issue (e.g. we excluded 4 out of 18 datasets based on owner comments) and should thus be standard in open data meta-analysis. The outlined issues might make meta-analysis based on data more time-consuming compared to traditional meta-analysis, but based on our experience, this will depend from case to case.

Conclusion

The meta-analysis approach has become increasingly important across ecological and evolutionary research fields, having a strong impact on future research, interventions, and policies. Here, we introduce a new standard on how to conduct a data-driven meta-analysis that, in contrast to the conventional meta-analysis, uses research data rather than published studies. This new standard is now possible given that the amount of open research data has been steadily increasing across evolutionary and ecological fields. We show that new questions can be addressed with the use of this ever-growing data-landscape, broadening the scope of meta-analysis in evolutionary ecology. In addition, by embracing open data, evolutionary ecology has the potential to benefit from a spectrum of higher standards and reporting practices brought in the new era of Open Science.

Supplementary Information S7

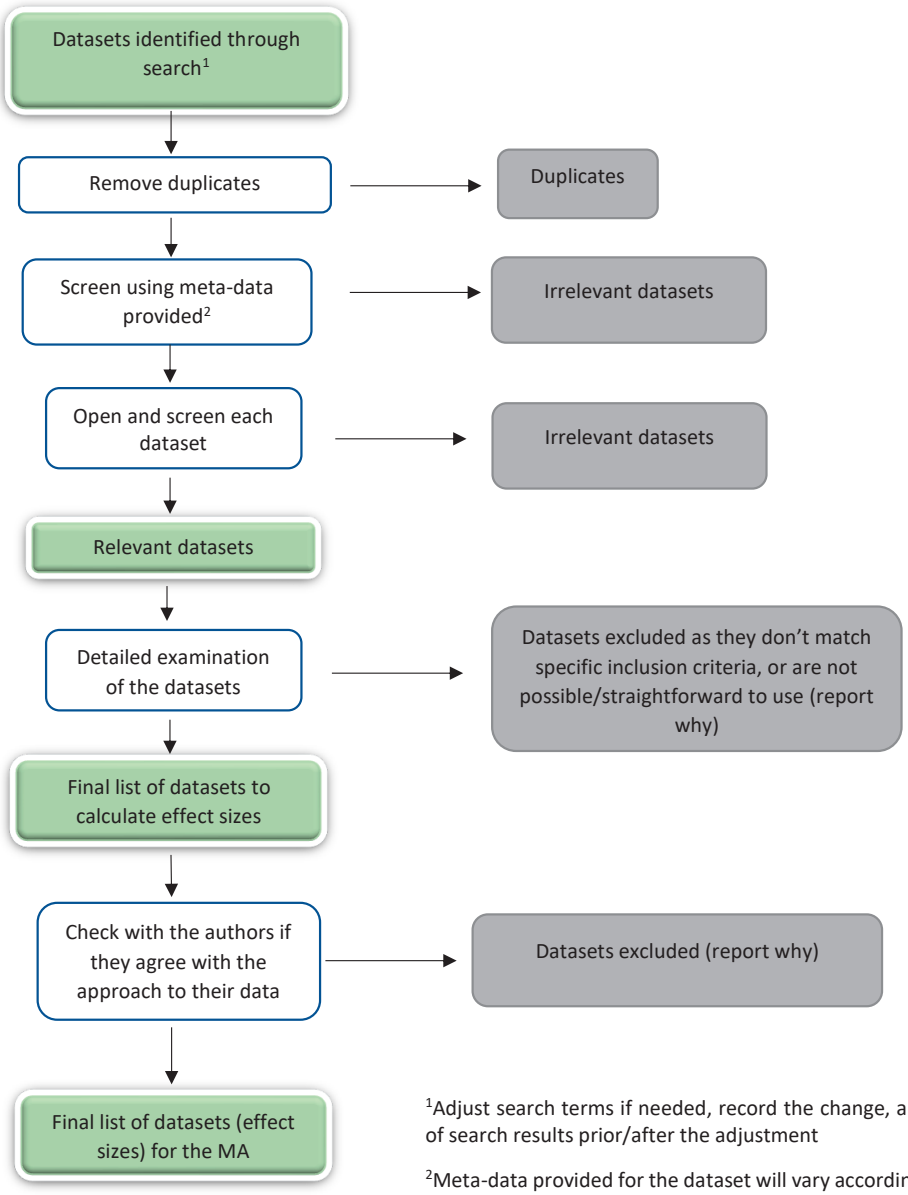


Figure S7.1. *PRISMA diagram for the reporting of the information retrieval in the data-centric approach to meta-analyses. Different sets of data (green boxes) are obtained/kept after the initial (and adjusted) search. The exclusion of irrelevant datasets can happen at different stages, based on the pre-established inclusion/exclusion criteria, and/or due to issues with data. Calculating each effect size can require several sub-steps, comparable for all the datasets (e.g. when the raw data are first used to calculate summary statistics, which is then used to calculate the effect size). This diagram may be adjusted for each case, but should always ensure that the process of data acquisition and screening is transparently documented.*

PART III

Exploring constraints in adaptation: a quantitative genetic approach

Mixed clutch of great tit and blue tit eggs

Chapter 8

Maternal effects in a wild songbird are environmentally plastic but only marginally alter the rate of adaptation

Jip J.C. Ramakers, Marleen M.P. Cobben, Piter Bijma, Thomas E. Reed, Marcel E. Visser & Philip Gienapp

Abstract

Despite ample evidence for the presence of maternal effects (MEs) in a variety of traits, and strong theoretical indications for their evolutionary consequences, empirical evidence to what extent MEs can influence evolutionary responses to selection remains ambiguous. We tested the degree to which MEs can alter the rate of adaptation of a key life-history trait, clutch size, using an individual-based model approach parameterised with experimental data from a long-term study of great tits (*Parus major*). We modelled two types of MEs: (i) an environmentally plastic ME, in which the relationship between maternal and offspring clutch size depended on the maternal environment via offspring condition, and (ii) a 'fixed' ME, in which this relationship was constant. Although both types of ME affected the rate of adaptation following an abrupt environmental shift, the overall effects were small. We conclude that evolutionary consequences of MEs are modest at best in our study system, at least for the trait and the particular type of ME we considered here. A closer link between theoretical and empirical work on MEs would hence be useful to obtain accurate predictions about the evolutionary consequences of MEs more generally.

The American Naturalist (2018) ***191****(5), E144–158*

Introduction

There is increasing recognition among evolutionary biologists of nongenetic (Mameli 2004; Danchin et al. 2011) or indirect genetic (Wolf et al. 1998) mechanisms of inheritance that affect the dynamics of phenotypic adaptation in populations. One such example is that of maternal effects (Mousseau and Fox 1998). In its most general sense, a maternal effect is the degree to which an offspring's phenotype is shaped by properties of the mother other than shared-genes effects, although these maternal properties may themselves have a genetic basis (Willham 1963; Mousseau and Fox 1998; Wolf et al. 1998; Bijma 2011). This can include effects of the maternal trait on the same trait in the offspring, such as litter size in mice (Falconer 1965) or age at maturity in springtails (Janssen et al. 1988), as well as the effect of maternal trait(s) on a different trait in the offspring, such as the effects of host-plant choice of the mother on offspring morphology, or of egg or propagule size on offspring growth rate (Mousseau and Fox 1998; Räsänen and Kruuk 2007). This study is concerned with the former type of maternal effects. Although maternal effects are sometimes thought of as 'nuisance parameters' hampering the prediction of evolutionary trajectories (Räsänen and Kruuk 2007; Danchin et al. 2011), theoretical models and empirical studies show that the presence of such effects can have profound impacts on rates of adaptation (Kirkpatrick and Lande 1989; Bijma 2011; Hoyle and Ezard 2012; McGlothlin and Galloway 2013).

Falconer (1965) described maternal effects, m, as a (partial) linear regression coefficient for offspring trait value on the same maternal trait value. An individual's phenotype z is then the sum of its breeding value A, its environment ε, and the partial maternal-effects regression coefficient times the mother's phenotype z_m ($z = A + \varepsilon + mz_m$). The narrow-sense heritability for the trait, i.e. the proportion of total phenotypic variance attributable to additive genetic effects, may now no longer adequately capture the potential for evolution because the maternal genotype has direct (via additive genetic inheritance) and indirect (via the maternal effect) effects on offspring phenotype. This concept was used by Kirkpatrick and Lande (1989) to devise a model that predicts evolutionary change across generations with the incorporation of phenotypic change due to maternal effects in current and previous generations. When maternal effects are absent, this model reduces to a standard model of additive inheritance. However, when positive maternal effects are present (i.e. $m > 0$, so that a larger maternal trait value results in a larger offspring trait value), the covariance between an individual's breeding value and its trait value exceeds the additive genetic variance for that trait, which facilitates a more rapid change in the mean trait value under directional selection. Negative maternal effects (i.e. $m < 0$, so that a larger maternal trait value results in a smaller offspring trait value) can reduce the response, and possibly even revert it. For example, growth rate (offspring trait) can be impaired by the amount of maternal care (maternal trait); this may have implications for offspring survival and hence the distribution of phenotypes in the next generation, causing an evolutionary time lag (Kirkpatrick and Lande 1989; Wolf et al. 1998).

Theoretical studies of maternal inheritance effects on fitness and rates of adaptation are ample (Kirkpatrick and Lande 1989; Bijma 2011; Hoyle and Ezard 2012; Prizak et al. 2014;

Kuijper and Hoyle 2015). Empirical work mainly comes from short-term studies testing the effect of experimentally manipulated maternal trait values on offspring performance (Schluter and Gustafsson 1993; but see e.g. Dey et al. 2016 and Plaistow and Benton 2009 for studies with more generations; e.g. Beckerman et al. 2006; Rechavi et al. 2011), and some have identified a role for epigenetic effects as an important driver of phenotypic variation in offspring (e.g. Cubas et al. 1999; Champagne 2008). Such short-term studies are, however, insufficient to inform us about the magnitude of maternal effects at longer (micro)evolutionary time scales (i.e. at least tens of generations) in natural populations, for which evidence to date remains scarce (Räsänen and Kruuk 2007; McAdam et al. 2014). Quantitative genetic modelling in long-term observational studies of natural populations can provide insights into maternal sources of phenotypic variation, but require high-quality data that are not always available (Merilä et al. 2001a; Kruuk and Hadfield 2007). Furthermore, if the maternal effect does not reflect a fixed maternal property but varies among breeding events, it is difficult to disentangle maternal from genetic effects. The maternal effect component mz_m on phenotype z (*sensu* Kirkpatrick and Lande 1989), as well as the slope and sign of m, is therefore difficult to estimate in most natural study systems without a highly informative pedigree (but see McAdam and Boutin 2004).

Typically, the role of maternal effects in evolution has been regarded as fixed, i.e. assuming a constant value for m (Kirkpatrick and Lande 1989; Bijma 2011). Different scenarios can then be explored, varying m and predicting its role in adaptation and fitness in combination with other adaptive mechanisms such as phenotypic plasticity and grand-maternal effects (Hoyle and Ezard 2012; Ezard et al. 2014; Prizak et al. 2014). In reality, however, maternal effects may not be fixed but plastic in response to environmental conditions and hence may change from season to season. For example, inbred Seychelles warbler (*Acrocephalus sechellensis*) mothers produce low-quality offspring, which in turn affects offspring survival but only in poor breeding seasons (Richardson et al. 2004). As offspring were cross-fostered in that study and common-environment effects could thus be ruled out, this suggests an environmentally plastic maternal effect mediated through the egg. Similarly, if offspring traits are condition dependent and offspring condition is in turn influenced by a maternal effect at a different rate in different environments, the net maternal effect will then be plastic, i.e. the coefficient m will vary with environments. The ability of m to vary with the environment means that there can be differential selection on the maternal component of the phenotype in different environments and, if the maternal trait is under genetic control, this may hence considerably alter evolutionary trajectories (Kuijper and Hoyle 2015).

Avian clutch size, a major life-history trait, is highly variable in some species and this variability has a genetic basis (Postma and van Noordwijk 2005b). Stabilising selection on clutch size is likely to be strong as deviations from the optimal clutch size compromise offspring viability and recruitment and therefore maternal fitness (Pettifor et al. 1988, 2001; Krementz et al. 1989; Both et al. 1999, 2000; Rodríguez et al. 2016). The maternal effect of the mother's clutch size on her daughters' clutch size is a likely candidate for an environmentally plastic maternal effect as maternal clutch size affects offspring body condition, depending on the environment, and offspring condition likely affects offspring clutch size. Females that lay clutches larger than their individual optima produce offspring

of relatively poor condition (Sanz 1997; Both et al. 2000; Pettifor et al. 2001). If this poor condition persists through to breeding age, then these offspring in turn will lay smaller clutches than predicted by the genes inherited from their parents, because the number of eggs a bird can produce is condition dependent (e.g. Schluter and Gustafsson 1993; but see Merilä et al. 2001a). Their offspring (i.e. the grand-offspring of the original females), now born in "too small" clutches (i.e. a clutch size smaller than the number of young that could be successfully raised), will be relatively heavy because food is shared among fewer nestlings, and may in turn go on to lay (too) large clutches as adults (Haywood and Perrins 1992; Tilgar et al. 2010; but see Haywood 2013). This effect may perpetuate through the generations, although it should wane quickly in stable environments as the phenotype is pulled toward the optimum (cf. Kirkpatrick and Lande 1989; Bijma 2011).

The environmentally plastic nature of *m* becomes apparent when considering the environment-dependent relationship between maternal clutch size and resulting fledgling weight: as there is a trade-off between offspring quantity and quality, offspring fledgling weight will decrease more strongly with maternal clutch size under adverse than favourable conditions (e.g. when food is abundant; Both et al. 2000). The maternal effect will therefore vary with the environment and this has the potential to change the rate of adaptation of a population when it is under directional selection. Adaptation dynamics could be affected in two ways: offspring that fledge in poor condition may survive less well, and those that do survive may produce smaller clutches as adults. Either or both would therefore affect the total strength of selection on maternal clutch size. To predict this in a model, one would therefore need to estimate four important parameters: (i) the narrow-sense heritability of clutch size, (ii) the environment dependency of the clutch size–offspring condition relationship, (iii) survival based on offspring condition as a selection factor on maternal clutch size (note that for the sake of simplicity, this disregards viability selection operating on adults), and (iv) the effect of offspring condition on offspring clutch size (see Fig. 8.1).

In this paper, we addressed the question of to what extent environmentally plastic maternal effects can speed up or slow down the rate of adaptation of clutch size in a wild population using empirically estimated parameter values. We estimated the parameters for an exceptionally well-studied passerine bird, the great tit (*Parus major*), to calculate both environmentally plastic and fixed maternal effects. These were then used in an individual-based model to predict the rate of adaptation of clutch size following an environmental shift. We estimated maternal effects by regressing offspring phenotype on maternally induced offspring condition or maternal phenotype, using a combination of long-term field observations and multi-year experimental manipulations (Both et al. 2000). In keeping with theoretical findings concerning environment-dependent maternal effects (Hoyle and Ezard 2012; Kuijper and Hoyle 2015), we explored two alternative, but related, routes toward quantifying maternal effects: (i) via fledgling weight, which itself is a result of maternal clutch size and the environment (making *m* environmentally plastic); and (ii) via a fixed maternal effect (i.e. using the conventional definition of *m*), where *m* is not environment dependent (Fig. 8.1). Although conceptually simplified, the latter effect may arise, for example, as a result of brood size-mediated androgen levels that may negatively affect offspring fecundity (Naguib et al. 2004; Rutkowska et al. 2005) or of

transgenerational epigenetic inheritance induced by maternal malnutrition (Champagne 2008; Jablonka and Raz 2009). We explicitly used parameters from a wild population to explore realistic evolutionary responses in a key life-history trait under reasonably strong directional selection.

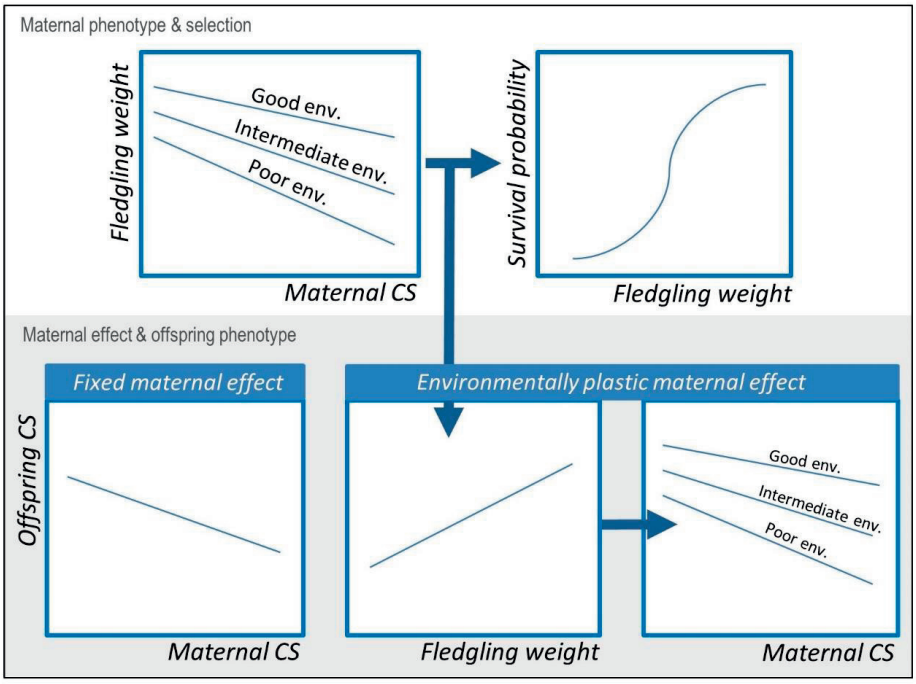


Figure 8.1. *Two alternative approaches to estimating maternal effects underlying avian clutch size. Maternal clutch size (CS) affects offspring condition (weight) depending on the quality of the environment, and this condition in turn drives selection through offspring survival (top row). In one scenario, offspring condition also affects their phenotype (clutch size), which, when regressed against the maternal CS, results in an environmentally plastic maternal effect (right-hand panels, bottom row). Alternatively, the maternal CS directly influences offspring CS independent of the environment, resulting in a 'fixed' maternal effect (left-hand panel, bottom row).*

Methods

Study system

We estimated our model parameters from data from a long-term (1955–present) population study of great tits (*Parus major*), a hole-breeding passerine, at the Hoge Veluwe National Park in the Netherlands (52°02′07″ N 5°51′32″ E). The 171-ha study area, comprising a mixture of deciduous and coniferous forest stands, has ~400 nest boxes that are checked weekly from April–July to score life-history traits including egg-laying/-

hatching date and clutch size. When nestlings are 7–8 days old, the parents are captured in their nest boxes (using spring-door traps), banded and blood-sampled; nestlings are banded, blood-sampled, and weighed at day 15, which is close to the age of fledging. The banding of birds allows for carefully monitoring immigration and offspring recruitment and establishing pedigrees. The study area is surrounded by a matrix of potentially suitable breeding habitat, which facilitates dispersal from and into the focal area. The population has been studied continuously since 1955 and has been subjected to various experiments aimed at manipulating life-history traits such as egg-laying date and clutch size.

The individual-based model

We used an individual-based model to estimate the impact of (environmentally plastic) maternal effects on adaptation. Population size N was roughly 500 in every model generation, assuming no overlapping generations (i.e. the whole adult population is replaced by recruits every year). In each generation, a sex (ratio 1:1) was randomly assigned to individuals and both sexes were paired up randomly for mating. We simulated a total of 1000 generations, i.e. 500 burn-in generations to reach equilibrium conditions, followed by an environmental shift and 500 additional generations, and repeated the process 1000 times. To avoid confusion, we refer to clutch size and fledgling weight as z_{CS} or z_{FW}, respectively, throughout. Parameters requiring estimation from data are summarised in Table 8.1. An example script of the model for the R environment has been uploaded as supplementary material.

(i) Generating a population, genotypes and phenotypes

The clutch size of a given individual (i) in the initial population was defined as

$$z_{CS_i} = \mu_{CS} + A_{CS_i} + M_{CS_i} + \varepsilon_i \quad (8.1)$$

where μ_{CS} is a constant (here 8.0), A_{CS_i} the individual's genotype (breeding value), M_{CS_i} its maternal component (i.e. mz_m as in Falconer 1965) and ε_i the residual component, all initially randomly drawn from a univariate normal distribution with mean 0 and standard deviations $\sqrt{V_{A_{CS}}}$, $\sqrt{V_{M_{CS}}}$, and $\sqrt{V_{z_{CS}} - V_{A_{CS}} - V_{M_{CS}}}$, respectively. For all following generations, z_{CS_i} was calculated as in eqn. (8.1), but A_{CS_i} and M_{CS_i} were no longer randomly drawn but calculated from parameter values in the current generation. Offspring genotype was defined as

$$A_{CS_i} = \frac{A_{CS_{mother_i}} + A_{CS_{father_i}}}{2} + y_i, \quad (8.2)$$

where y_i is the Mendelian segregation error, drawn from a univariate normal distribution with mean 0 and standard deviation $\sqrt{0.5V_{A_{CS}}}$ (Lynch and Walsh 1998).

***Table 8.1**. Summary of input parameters for the individual-based model, estimated from the great tit population at the Hoge Veluwe.*

Parameter		Mixed model component	Estimate	SE	Notation	Details
(i)	Heritability CS[a]	–	0.24	0.04	h^2_{CS}	Tbl S1
(ii)	Fledgling weight as a function of CS and the environment (year)[abc]	1988 (poor env.)	0.53	0.06	α_{FW_j}	Tbl S2; Eqn. 5
		1984 (med. env.)	0.49	0.05	α_{FW_j}	
		1986 (good env.)	0.81	0.05	α_{FW_j}	
		1988:CS	–0.30	0.06	β_{FW_j}	
		1984:CS	–0.08	0.01	β_{FW_j}	
		1986:CS	–0.06	0.01	β_{FW_j}	
(iii)	Offspring survival[b]	Intercept	–19.06	2.92	α_ϕ	Tbl S3; Eqn. 6
		Fledgling weight	1.62	0.33	β_ϕ	
		[Fledgling weight]^2	–0.04	0.01	γ_ϕ	
(iv.i)	Maternal effect via fledgling weight[cd]	Intercept	–0.25	0.06	α_p	Tbl S4; Eqn. 3
		Fledgling weight	0.13	0.05	β_p	
		Coefficient for M_{CS_i} against $z_{CS_{m_i}}$	–0.13		m_p (poor)	
			–0.04		m_p (med.)	
			–0.03		m_p (good)	
(iv.ii)	Maternal effect via maternal CS[ac]	Intercept	–0.25	0.06	α_f	Tbl S5; Eqn. 4
		Maternal CS	–0.21	0.03	m_f	

Additional model input parameters for the initial population

Description	Notation	Estimate	Details
Phenotypic variance[e]	$V_{z_{CS}}$	3.91	
Additive gen. variance	$V_{A_{CS}}$ $(= h^2_{CS} \times V_{z_{CS}})$	0.94	
Relative maternal effect variance[f]			
For m_p model[g]:	M^2_{CS}	0.006	Tbl S4
For m_f model:	M^2_{CS}	0.027	Tbl S5
Maternal effect variance			
For m_p model[h]:	$V_{M_{CS}}$ $(= M^2_{CS} \times V_{z_{CS}})$[h]	0.02	
For m_f model:	$V_{M_{CS}}$ $(= M^2_{CS} \times V_{z_{CS}})$	0.11	

Note: Shown are estimates of intercepts and slopes (and their standard errors) from mixed-effects models detailed in the supplementary tables.

[a] CS = clutch size

[b] Estimates are on a logit scale

[c] Continuous predictor variables were centred around their annual mean before analysis and decentred in the individual-based model

[d] Implicit maternal-effects coefficient m_p is derived by regressing the maternal component M_{CS_i} (eqn. 3) on environment-specific maternal clutch size z_{CS_m}; β_p = 0.10 (0.04 SE) in the model combining fixed and plastic maternal effects

[e] Estimated from the Hoge Veluwe population

[f] Calculated using Nakagawa and Schielzeth's (2013) marginal R^2 for mixed-effects models

[g] In the model combining fixed and plastic maternal effects, this value was 0.003 (Table S5)

[h] In the model combining fixed and plastic maternal effects, this value was 0.01

(ii) Maternal effect

Here, we assumed two types of maternal effects: (i) an environmentally plastic effect via fledgling weight, determined by maternal clutch size in interaction with the environment, and (ii) a fixed effect that only depends on maternal clutch size. In the case of the environmentally plastic maternal effect, M_{CS_i} is calculated as follows:

$$M_{CS_i} = \alpha_p + \beta_p z_{FW_i}, \tag{8.3}$$

where α_p and β_p, the subscript p referring to a plastic maternal effect, are the intercept and slope (i.e. partial regression coefficient), respectively, from a regression of offspring clutch size on offspring fledgling weight z_{FW_i} (estimated whilst controlling for additive genetic effects; see 'Estimating model parameters from data').

The 'fixed' (f) maternal effect is calculated from the maternal clutch size $z_{CS_{m_i}}$ as follows:

$$M_{CS_i} = \alpha_f + m_f z_{CS_{m_i}}, \tag{8.4}$$

where α_f and m_f are the intercept and slope (i.e. a partial regression coefficient) from a regression of offspring clutch size against maternal clutch size (estimated whilst controlling for both fledgling weight and additive genetic effects; see 'Estimating model parameters from data'). Both types of maternal effects were run in separate models (i.e. containing either only the plastic or only the fixed type), where their effect on the rate of adaptation was determined by keeping M_{CS_i} in or removing it from eqn. (8.1).

As eqn. (8.3) models M_{CS_i} as a function of fledgling weight, β_p has to be positive. To intuitively compare both types of maternal effect, we regressed M_{CS_i} resulting from eqn. (8.3) on maternal clutch size $z_{CS_{m_i}}$ for each environment to obtain a negative (partial) regression coefficient m_p associated with each environment. Note that m_p was merely estimated for illustrative purposes; in the model, m_p was incorporated implicitly via the effect of fledgling weight as in eqn. (8.3).

The combined effect of both maternal effects, both represented by partial regression coefficients (see 'Estimating model parameters from data'), was tested in a third model that was defined by extending the maternal-inheritance component in eqn. (8.1) as M_{CS_i} (eqn. 8.3) + M_{CS_i} (eqn. 8.4) (i.e. combining both the plastic and the fixed type in a single model).

(iii) Fledgling weight, survival, and fitness

Offspring fledgling weight z_{FW_i} is a function of maternal clutch size $z_{CS_{m_i}}$. Since fledgling weight is in nature bounded between a minimum and a maximum, it was modelled as a linear function of maternal clutch size assuming a logit scale; this allowed for back-

transformation to get a naturally sigmoidal, asymptotic relationship (see 'Estimating model parameters from data'). Fledgling weight before back-transformation was defined as

$$z'_{FW_i} = \alpha_{FW_j} + \beta_{FW_j} z_{CS_{m_i}}, \quad (8.5)$$

where α_{FW_j} and β_{FW_j} are the intercept and slope related to the j^{th} environment. Fledgling weight z_{FW_i} (calculated as $e^{z'_{FW_i}}/[1 + e^{z'_{FW_i}}] \times [max - min] + min$, with max and min indicating predefined boundaries) affects offspring survival (recruitment) probability, ϕ_i, according to the logistic function

$$\phi_i = \frac{1}{1+e^{-(\alpha_\phi + \beta_\phi z_{FW_i} + \gamma_\phi z^2_{FW})}}, \quad (8.6)$$

where α_ϕ and β_ϕ are the fledgling weight-related intercept and slope; γ_ϕ is the negative slope associated with the quadratic term, as survival was expected to level off and eventually decrease at extremely high fledgling weights (Mulder et al. 2016b). A mother's fitness, W_i, is the product of her clutch size and offspring survival probability, yielding

$$W_i = z_{CS_i}\, \phi_i. \quad (8.7)$$

Note that the index *i* for offspring survival probability ϕ_i is still useful here as all offspring from the same brood are expected to have the same value for ϕ_i. W_i is then scaled up to match the number of recruits that need to be produced to reach *N*:

$$W'_i = W_i \frac{\bar{n}_{E_i}}{\bar{W}}, \quad (8.8)$$

where $\bar{n}_{E_i}$ is the expected mean number of recruits produced per brood pair and $\bar{W}$ is the average fitness over all broods. The actual number of recruits produced by each brood, n_i, is then determined by randomly drawing from a Poisson distribution with $\lambda_i = W'_i$. To quantify the strength of selection, the standardised selection differential (*s*) for a given year (*j*) was calculated following Lande and Arnold (1983):

$$s_j = \frac{Cov\left[\frac{n_{ij}}{\bar{n}_j}, z_{CS_{ij}}\right]}{\sigma_{z_{CS_j}}}, \quad (8.9)$$

where $\bar{n}_j$ is the average number of recruits per brood and $\sigma_{z_{CS_j}}$ the standard deviation of z_{CS_j}.

(iv) Environmental change

To allow model parameters to equilibrate before the stepwise change in the environment we ran the model in an intermediate environment for 500 generations. After this 'burn-in' period, the environment switched to either a poor or a good environment by either increasing (good environment) or decreasing (poor environment) α_{FW_j} and β_{FW_j} in eqn. (8.5) (note that because β_{FW_j} is negative, a higher value means a shallower slope). This means that the population moved to different fitness optima, as the relationship between maternal clutch size and offspring fledgling weight (eqn. 8.5) differed among different environments. The environmental shift was abrupt (i.e. during one generation) and was kept constant for another 500 generations.

Estimating model parameters from data

Four analyses were performed to estimate four parameters necessary for our individual-based model: (i) heritability of clutch size, (ii) the environment-dependent effect of maternal clutch size on fledgling weight, (iii) the effect of fledgling weight on offspring recruitment, and (iv) the maternal effect, i.e. the effect of fledgling weight and/or maternal clutch size on offspring clutch size. All relevant parameters are summarised in Table 8.1; data have been deposited in Dryad Digital Repository (Ramakers et al. 2017).

(i) Heritability of clutch size

To estimate heritability, we used all unmanipulated, first clutches from all birds from 1956–2013 with known identity (n = 5394 observations from 3328 females). We modelled clutch size in an 'animal model' (Henderson 1988; Kruuk 2004) with Gaussian errors, based on restricted maximum likelihood estimation using ASReml-R v. 3 (Gilmour et al. 2009). Fixed effects were age (first-time breeder or older), egg-laying date (centred on the mean value for that year) and year of breeding (as a factor); random effects were female identity ('permanent environment'), maternal identity, nest box identity, and the additive genetic component based on the pedigree. Males do not express a clutch size phenotype but it was assumed here that they carry the genes for clutch size, and hence paternal links were included in the analysis. In the construction of the pedigree, the female's social partner was assumed the genetic father. Molecular analysis in a nearby great tit population has revealed that the proportion of extra-pair young ranges from 6.5 to 12.5% (Van Oers et al. 2008). Such rates are common for tit species (Brommer et al. 2010), but have been found to only marginally affect heritability estimations when sample sizes are sufficiently large (i.e. >100; Charmantier and Réale 2005).

All fixed and random variables contributed significantly to variation in clutch size, with the exception of maternal identity (Table S8.1). Narrow-sense heritability (h^2_{CS}) was estimated at 0.24 (± 0.04 SE).

(ii) Fledgling weight vs. maternal clutch size

To estimate the environment-dependent effect of clutch size on fledgling weight, we cannot rely on observational data since different females likely have different, individually optimised clutch sizes (Pettifor et al. 1988, 2001). We therefore made use of eight years (1983–1990) of brood size manipulations at our study site (Both et al. 2000) to estimate α_{FW_j} and β_{FW_j} in eqn. (8.5). Briefly, each year triplets were formed of nests with the same clutch size and hatching date, within which broods were randomly chosen to be either enlarged or reduced by approximately a half or to remain the same size when chicks were 1–3 days old. The year 1988 differed somewhat in that three broods of different sizes were manipulated to one common brood size. Our aim was to find year-dependent trade-offs between a female's clutch size and her offspring's body condition. Different years are here assumed to represent different environments, i.e. in terms of food availability or breeding-pair densities, with poor years exhibiting the steepest negative slope of fledgling weight versus clutch size and a comparatively low average body condition. We therefore modelled fledgling weight (n = 2145 nestlings) as a function of manipulated brood size,

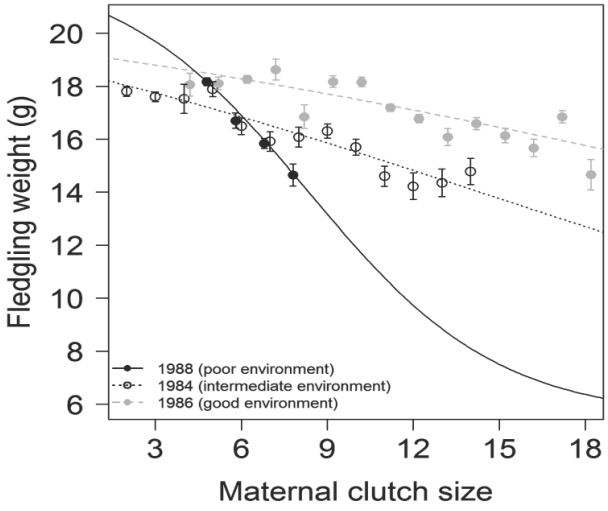


***Figure 8.2**. Great tit fledgling mass at the Hoge Veluwe as a function of maternal clutch size for three representative (good, intermediate and poor) years. Dots are means ± SE and random horizontal spacing was added between years, done for visual purposes only. Trend lines are back-transformed estimates from a linear mixed-effects model with fledgling weight on a logit scale (Tables 1 and S2). Brood sizes are centred values plus the mean for the respective years. Note that data come from experimental brood-size manipulations; 1988 had few manipulated brood sizes compared to other years because in this year the manipulation procedure was slightly different (Both et al. 2000).*

year, and the interaction between the two, as well as original clutch size and hatching date, in linear mixed-effects model using the R package *lme4* (Bates et al. 2018); brood identity (n = 309) nested within female identity (n = 251) served as a random effect. All continuous predictor variables were centred around their mean value for that year; fledgling weight was transformed to a logit scale before analysis ($z'_{FW_i} = p/[1-p]$, where $p = [z_{FW_i} - 5.5]/[22.5 - 5.5]$) to allow for realistic asymptotes at both extremes of the weight spectrum (i.e. 5.5 g < z_{FW} < 22.5 g) after back-transformation. Manipulated brood size, year, and their interaction were highly significant (Table S8.2): years differed in both elevation and slope of the clutch size–fledgling weight relationship (Fig. S8.1). We chose three particular years to represent a 'good', 'intermediate', and 'poor' environment (Fig. 8.2) based on the values for α_{FW_j} (the weight in the average environment) and β_{FW_j} (the steepness of the curve, with the shallower slopes indicating better environments). We chose 1988 to represent the poor environment; note that although the experimental procedure in this year differed somewhat from other years and 1988 might thus be an oddity, its steep slope renders it a suitable 'extreme' scenario.

(iii) Offspring survival probability

We modelled offspring survival probability ϕ based on recruitment probability (which approximates survival) as a function of fledgling weight and the square of fledgling weight. We thus ran a generalised linear mixed-effects model with a logit link to estimate α_ϕ, β_ϕ, and γ_ϕ (eqn. 8.6). Fledglings could either return or not return to the breeding

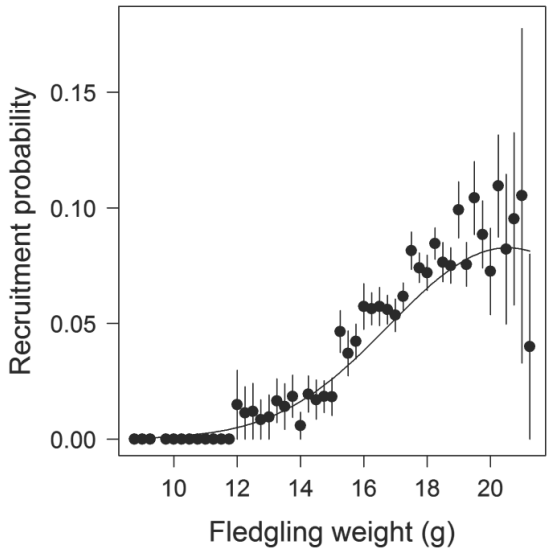


***Figure 8.3**. Recruitment probability of great tits at the Hoge Veluwe as a function of their fledgling weight (see Tables 1 and S3). Means and their standard errors are given; note that grouping was done for visual purposes only and extreme fledgling weights (at both ends of the spectrum) were disregarded due to too few observations.*

population (1/0 response). Brood identity nested within year of breeding was added as a random effect. We used data from 1973–2013 because of few observations in earlier years (n = 24320 nestlings from 3600 broods). Recruitment probability showed a highly significant, quadratic response to fledgling weight (Table S8.3) and was approximately constrained below 0.1 (Fig. 8.3). We also tested whether fledgling weight interacted with year (i.e. whether β_ϕ varied among years) but found no statistical evidence for this (results not shown).

(iv) Maternal effect on offspring clutch size

To estimate the maternal effect on offspring clutch size we explored two routes: one that acts through offspring fledgling weight and where it is environmentally plastic (eqn. 8.3), and one that acts through maternal clutch size and is not environmentally plastic (eqn. 8.4). From the 1973–2013 data set, we extracted all first-year breeding attempts of females with known fledgling weights and known mothers (n = 510). As clutch size is partly under genetic control, we needed to account for this effect when estimating effects of fledgling weight or maternal clutch size on offspring clutch size. To do this, we used the predicted breeding values (PBVs) extracted from the previous animal model (i.e. BLUPs) based on the complete data set. We are fully aware that PBVs can come with substantial and potentially non-random prediction errors (Hadfield et al. 2010), but the data set used to estimate the maternal effect was too small to estimate additive genetic variance reliably, and we are convinced that this approach gives more reliable results. Note that, consequently, our estimates of negative maternal effects are likely somewhat conservative (i.e. show an upward bias) given the positive association between maternal and offspring phenotypes expected from genetic inheritance.

In the first model, maternal clutch size affected offspring clutch size via fledgling weight, which makes it an environmentally plastic effect since the effect of maternal clutch size on fledgling weight varies with the environment (Fig. 8.2 and Fig. 8.4b). Fledgling weight and PBV significantly contributed to variation in offspring clutch size, the former explaining 0.6% of variation and having an estimated slope (β_p) of 0.13 eggs g^{-1} (Table 8.1 and S8.4; Fig. 8.4a). To obtain m_p, which is only implicitly modelled, we subsequently regressed M_{CS_i} against maternal clutch size, which led to three different, environment-dependent values for m_p approximately corresponding to –0.13, –0.04, and –0.03 for poor, intermediate, and good environments, respectively (see Fig. 8.4b; note that these curves are nonlinear).

In the second model, the maternal effect was estimated as the partial regression coefficient for offspring against maternal clutch size (m_f) and was not environmentally plastic (eqn. 8.4). A similar mixed-effects model was run, but maternal clutch size z_{CS_m} (centred around the annual mean value) was added, as well as its interaction with mean population fledgling weight in the current breeding year t ($\bar{z}_{FW_t}$), as a measure of environmental quality, to assess whether the effect of maternal clutch size on offspring clutch size depended on the environment. Besides PBV and individual fledgling weight (z_{FW_i}), maternal clutch size was highly significant and explained 2.7% of variation (Table

8.1 and S8.5; Fig. 8.4b); the maternal-effects coefficient was estimated at $m_f = -0.21$ (note that in this model the effect of fledgling weight (β_p) was reduced to 0.10 eggs g^{-1}, explaining 0.3% of variation). There was no significant effect of $\bar{z}_{FW_t}$, nor was there an interaction between z_{CS_m} and $\bar{z}_{FW_t}$, reinforcing the view that m_f is not environmentally plastic.

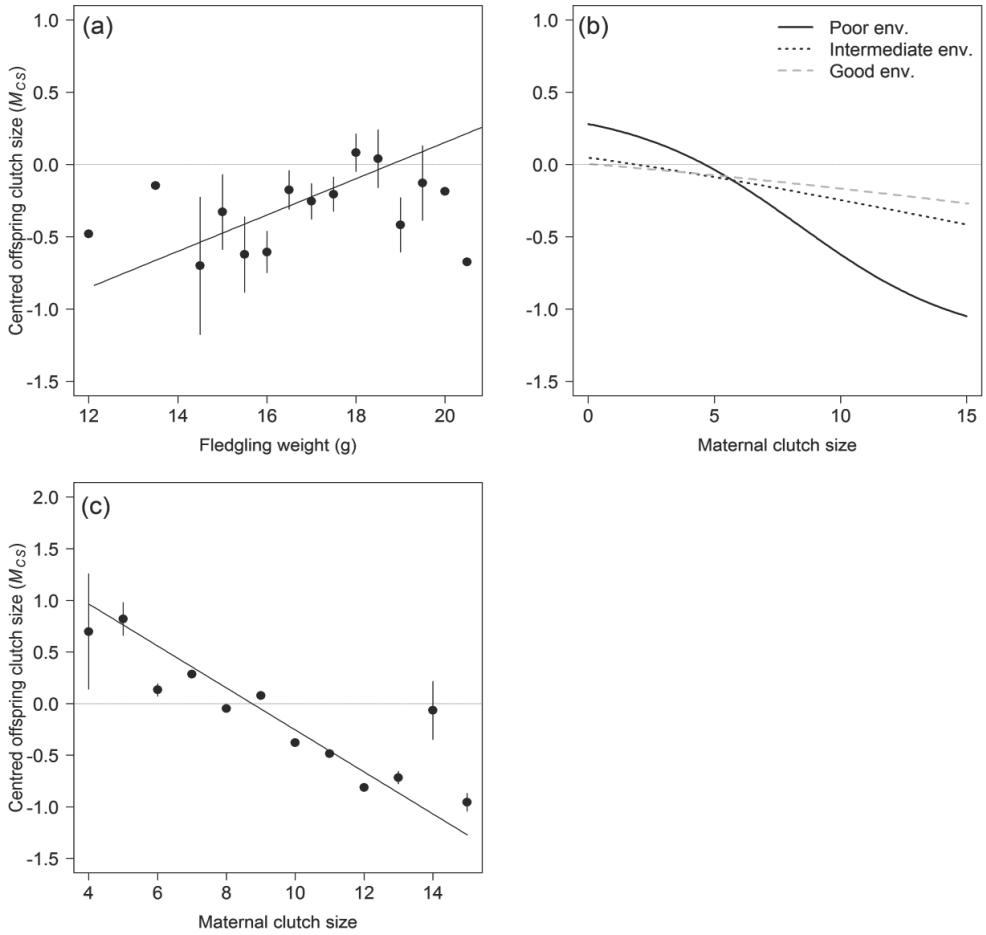


***Figure 8.4**. Maternal effect on offspring clutch size in great tits at the Hoge Veluwe. The environmentally plastic maternal effect operates through fledgling weight (a); plotting the resulting centred maternal clutch size (i.e. the maternal-effects component M_{CS}) against the maternal phenotype in the mother's environment leads to three environment-specific maternal effects m_p (b). The fixed maternal effect m_f is the effect of maternal clutch size on offspring clutch size independently of fledgling weight and is not environmentally plastic (c). Points are means and their standard errors, corrected for PBVs, which was done for visual purposes only. Lines are estimates from linear mixed-effects models, keeping PBVs constant at their mean (Tables S8.4 and S8.5).*

Results

Environmentally plastic maternal effects (m_p)

Environmentally plastic maternal effects only marginally affected the rate of adaptation following an environmental shift to new phenotypic optima, relative to the situation where maternal effects were absent (Fig. 8.5a,b; see Fig. S8.3 for changes in mean maternal effects component $\bar{M}_{CS}$, phenotypic variance $\overline{V_{z_{CS}}}$, fledgling weight $\bar{z}_{FW}$, and selection differential $\bar{s}$).

Under selection for a larger clutch size, offspring survival probability increased after the environmental shift—regardless of maternal clutch size. This is because fledgling weight was little compromised when clutches were large in the new, good environment (Fig. 8.2; see also $\bar{z}_{FW}$ in Fig. S8.3). Surviving offspring in the first generation following the environmental shift, many of them in relatively good condition, laid relatively large clutches that did not result in a reduction of offspring weight. Therefore, a negative m_p coefficient slightly favoured adaptation in the first 100 years following the burn-in. As the new optimum trait value was approached, selection decreased (Fig S8.3), hence diminishing the response in $\bar{z}_{CS}$ compared to the scenario without a maternal effect from ~70 years onward.

Under selection for smaller clutch size in the poor environment, which was much stronger than the selection for larger clutch size because of the narrower fitness peak (Fig. S8.2, S8.3, and S8.6), the initially enhancing effect of a negative m_p coefficient was more pronounced but lasted much shorter. After the environmental shift selecting for smaller clutch size, individuals laid too large clutches, resulting in a low average fledgling weight in generation t and, consequently, a drop in $\bar{z}_{CS}$ in year t + 1. The fixed weight–survival curve (Fig. 8.3) ensured that only the heaviest offspring survived (Fig. S8.3), which in turn would lay relatively large clutches—hence the slight upward tilt following generation t + 2. Again, the negative m_p coefficient pushed $\bar{z}_{CS}$ in the wrong direction, resulting in a lagged response compared to the scenario without a maternal effect from generation 7–8 onward (insets Fig. 8.5a).

The overall effect of the environmentally plastic maternal effect, however, remained small at <0.1 eggs under selection for both larger and smaller clutches compared to the situation without the maternal effect. To illustrate the nature of the environmentally plastic maternal effect more clearly, we ran another set of models where we set the regression coefficient for offspring clutch size against fledgling weight (β_p) to a less realistic 0.5 (resulting in an implicit regression coefficient m_p of –0.52, –0.16, or –0.09 in poor, intermediate, and good environments, respectively; see Methods), whilst keeping the intercept α_p the same. These parameter settings clearly show the potential capacity of the environmentally plastic maternal effect to drive $\bar{z}_{CS}$ and $\bar{A}_{CS}$ (Fig. S8.5a,c).

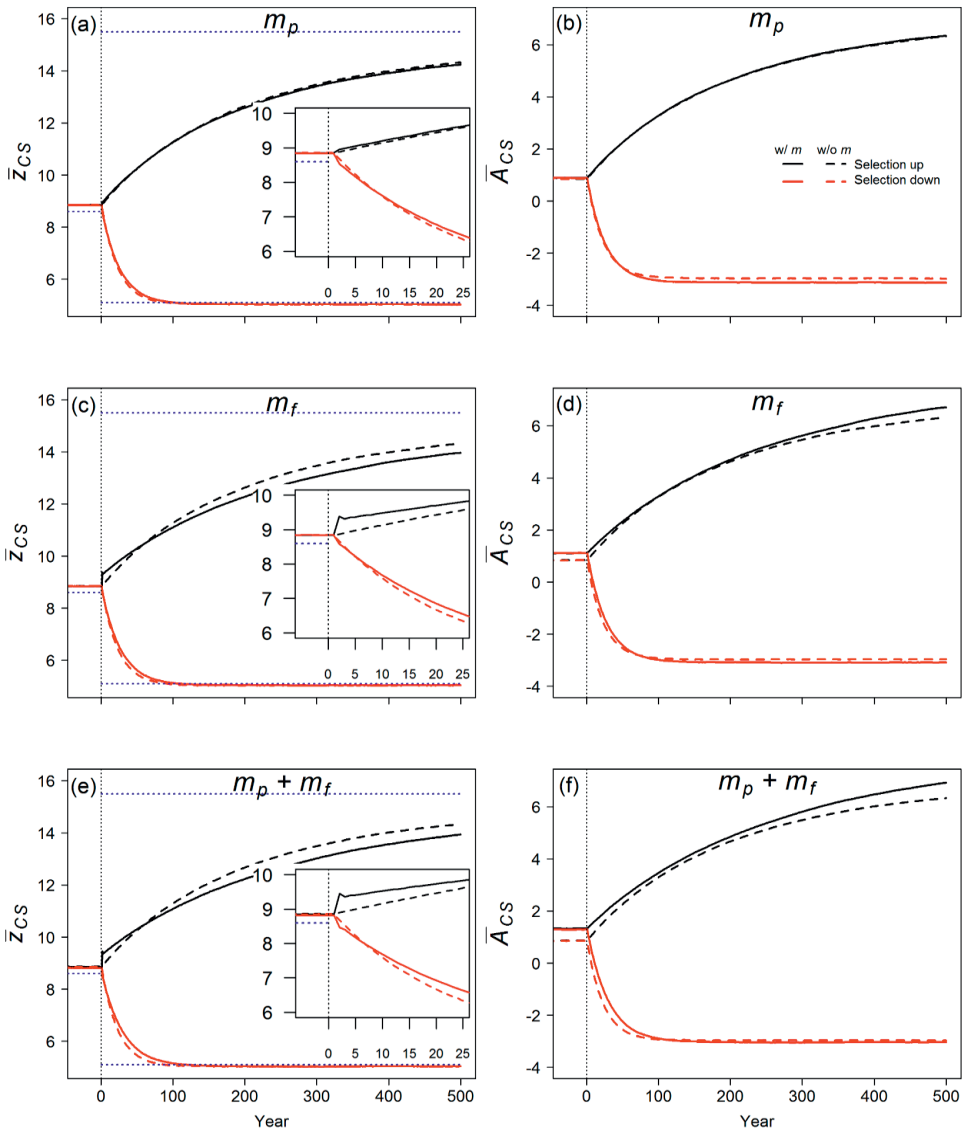


Figure 8.5. *Predicted mean phenotypic (a, c, e) and genetic (b, d, f) change in avian clutch size (modelled after the great tit at the Hoge Veluwe) in response to selection when considering a realistic environmentally plastic (i.e. via fledgling weight; a, b), fixed (i.e. via maternal clutch size; c, d) or combined (e, f) maternal effect (solid lines), or no maternal effect at all (dashed lines). Phenotypic responses in the first 25 years are magnified in the insets in panels a, c, and e. The vertical line denotes the pre-burn-in period, after which selection moves from an intermediate clutch size to either a large (good environment) or small (poor environment) clutch; the blue dotted lines in panels a, c and e denote the 'optimal' phenotype, i.e. z_{CS} at W_{max}. Lines are the means of population averages over 1000 simulation runs. Input parameters are $V_{z_{CS}}$ = 3.91, h^2_{CS} = 0.24, M^2_{CS} = 0.006 (panel a and b) or 0.003 (panel e and f) for the model with the plastic maternal effect, M^2_{CS} = 0.027 for the model with the fixed maternal effect; $m_p \approx$ –0.13, –0.04 and –0.03 (depending on the environment; panel a and b) or –0.10, –0.03 and –0.02 (panel e and f), m_f = –0.21 (see text and Table 8.1 for details).*

Fixed maternal effects (m_f)

In the models in which the maternal effect was fixed, offspring clutch size was independent of fledgling weight, yet the effect of the fixed maternal effect was stronger than that of the environmentally plastic maternal effect (Table 8.1 and S8.5). In the first generation following the environmental shift that selected for larger clutch size, the negative impact of the fixed maternal effect was alleviated as there was little cost to an intermediate clutch size; this ensured that the next generation (t + 1) could lay large clutches that were immediately penalised in the subsequent generation (t + 2; Fig. 8.5c; see also Fig. S8.4). Note that this pattern is reminiscent of an effect of offspring condition on phenotype as in the scenario of the environmentally plastic maternal effect, yet the fixed maternal effect acted independently of the effect of fledgling weight (Table S8.5). Under selection for smaller clutches the negative maternal effect led to a decreased response (i.e. too large clutch sizes), resulting in a lag effect from the first or second generation onward. Note that the initial, adaptive effect of the environmentally plastic maternal effect was less pronounced here, as the strong weight-dependent selection did not affect the phenotype.

The immediate effect of the fixed maternal effect compared to the model without the maternal effect was around 0.6 eggs in the first generation under selection for larger clutches but this effect waned after a few generations and never exceeded 0.6 eggs in subsequent generations. Like in the environmentally plastic maternal effect model, therefore, also the overall effect of a fixed maternal effect remained small. Again, an exaggerated decrease of the coefficient m_f (eqn. 8.4) from –0.21 to –0.5 (but keeping the intercept α_f the same) in an additional set of model runs led to a more distinct effect on adaptation (i.e. adaptive under selection for larger clutches in the short run and maladaptive under both selection scenarios in the long run) and magnified the oscillations observed in the first few generations under selection for larger clutches (Fig. S8.5b,d).

Combining environmentally plastic and fixed maternal effects ($m_p + m_f$)

As the most likely scenario in our great tit study population, the third model that we considered used m_p and m_f as two separate, additive maternal effects, with parameters for both effects taken from Table S8.5 (implicit $m_p \approx$ –0.10, –0.03, or –0.02 for poor, intermediate, and good environments, respectively; m_f = –0.21). This model combined the relatively strong, initially enhancing effect of m_f under selection for larger clutches and the relatively strong, initially enhancing effect of m_p under selection for smaller clutches. Combined, the overall effect of m on $\bar{z}_{CS}$ and $\bar{A}_{CS}$ under selection for smaller clutches was slightly increased (Fig. 8.5e,f) compared to the model with the fixed maternal effect only, but the likely effect in our study population would remain small, making <0.5 eggs difference in the average phenotype between models with and without maternal effects in any generation.

Discussion

Using an individual-based model, parameterised with experimental data from a long-term population study of great tits, we investigated how a specific type of maternal effects—a maternal trait affecting the same trait in the offspring—could affect the rate of adaptation in a population experiencing an environmental shift. We found that the presence of environmentally plastic or fixed (negative) maternal effects in avian clutch size can speed up phenotypic adaptation in the short run and slow it down in the long run, but their effects in real populations are likely very small. This is because the real maternal-effects coefficients—and hence explained variation—were small (Tables 8.1, S8.4, and S8.5). Indeed, the use of higher values for the strength of the maternal effect showed that the model we used resulted in the familiar oscillating pattern in $\bar{z}_{CS}$ over time (Fig. S8.5), as predicted from earlier models that incorporated negative maternal effects (Kirkpatrick and Lande 1989; Bijma 2011). Had we included a realistic adult survival rate (for great tits circa 0.5) in the model, the effect of the maternal effect on the evolutionary response would have been even more reduced due to increased generation time, indicating even more strongly that the evolutionary consequences of the maternal effect on clutch size in our population are negligible. Indeed, had we used extreme parameter values used in theoretical model exercises (e.g. Ezard et al. 2014; Hoyle and Ezard 2012; Prizak et al. 2014), the effects would have been more profound (Fig. S8.5).

A key parameter in our stochastic model was the experimentally derived relationship between clutch size and fledgling weight, as (i) this determined the environmentally plastic maternal effect and (ii) selection on clutch size was largely driven by this relationship. Predicting these environmental scenarios would not have been possible with observational data, as individual optimisation of clutch size (Pettifor et al. 1988, 2001) will render the among-individual relationship of fledgling weight against clutch size flat or even positive. The different relationships depicted in Fig. 8.2 are likely the direct result of population density-dependent food availability in the respective years (Both et al. 2000). By its nature, therefore, the negative slope of the relationship is steepest in poor environments, resulting in strong directional selection for smaller broods as the environment shifts from intermediate to poor (Fig. S8.3 and S8.4); in the good environment, the relationship is much shallower and selection is much weaker (see Fig. S8.6). This imbalance in the strength of selection ensures that, in our model, adaptation is always faster toward smaller *vs* larger clutches. An initially increased response under selection for smaller clutches in the presence of the environmentally plastic maternal effects (Fig. 8.5a,e and Fig. S8.5a) is then merely a result of selection acting against heavy individuals laying too large clutches, which, indeed, is rapidly counteracted in subsequent generations.

The best (empirical) model included both environmentally plastic (m_p) and fixed (m_f) maternal effects (Table S8.5), the latter being the more important source of variation in clutch size (0.3% *vs* 2.7%). Whereas m_p is linked to offspring condition, we have no clear hypothesis as to which mechanism underlies m_f in our population. Non-genetic maternal inheritance has been linked to transgenerational epigenetic effects in several contexts,

including parental care and nutritional stress (Champagne 2008; Jablonka and Raz 2009). In mammals, for example, maternal post-conception protein restriction and prenatal famine induce DNA methylation corresponding to impaired offspring development, with potential consequences for metabolic phenotypes later in life (Tobi et al. 2014; Holland et al. 2016). Rats receiving little grooming as pups show increased stress response and methylation patterns of genes associated with glucocorticoid stress response, setting the stage for their own maternal grooming behaviour as adults (Weaver et al. 2004; Szyf et al. 2005). Such epigenetic mechanisms are likely to reset in every generation (Feng et al. 2010). While this mechanism could theoretically underlie both m_p and m_f in our case, they would be a more likely candidate for m_f as they can be reset in every generation, but more empirical work is needed to elucidate the evolutionary importance of epigenetic inheritance in natural populations (Verhoeven et al. 2016).

We had no indication from our long-term data set that m_f was in any way dependent on the environment, despite a considerable year-to-year variation in clutch size (Table S8.1). Kuijper and Hoyle (2015) have argued that maternal effects are in reality not likely fixed, but have the ability to evolve positive or negative signs depending on the stability of the environment. Interestingly, our empirical estimate of m_f (–0.21) is congruent with Hoyle and Ezard's (2012) derived value for m (–0.2) at which mean population fitness is predicted to be maximised given a moderate degree of autocorrelation ($\rho = 0.25$) between the environment of development and selection. Using an intuitive measure of the quality of the environment, i.e. population-average fledgling weight (see Table S8.2), we find a significantly positive lag-1 autocorrelation of $\rho = 0.36$ ($p < 0.05$). Thus, m_f, in our population, is close to what we would expect to evolve in an environment that, although varying from year to year, exhibits a reasonable degree of predictability. Such a negative maternal effect, whatever the underlying mechanism, is expected to evolve as it tends to reduce phenotypic variance and enhance mean fitness in the population (Hoyle and Ezard 2012; Kuijper and Hoyle 2015).

The trait variation explained by the maternal effect found here as well as in previous studies seems to be small to modest (Räsänen and Kruuk 2007; McAdam et al. 2014). This has obvious implications for their potential consequences for evolutionary change but also raises the question as to why maternal effects seem to be generally weak. If we viewed a maternal effect as an adpative plastic effect to 'prime' offspring optimally for expected environmental conditions, then low predictability of the expected environmental conditions would lead to a reduced or absent maternal effect (Uller 2008), analoguous to non-transgenerational plasticity (Gienapp et al. 2014). We may also expect small maternal effects if adjusting them to varying environmental conditions is costly, analoguous to the costs of phenotypic plasticity (e.g. DeWitt et al. 1998), but our current understanding of costs of plasticity is still limited (Auld et al. 2010). So, maternal effects may be constrained in the same way as other plastic traits and this may explain their small to modest sizes. Furthermore, for maternal effects to evolve to their optimal values, genetic variation in them is required, but our understanding of the (quantitative) genetics of maternal effects in wild populations is even more limited, partly because the necessary data are scarce (McAdam et al. 2014).

In our model, the effects of m_f and m_p on adaptation were projected over a few hundred generations. Realistically, given the transient nature of our empirical approximations, this is the maximum predictive window across which we can endeavour to make projections. Theoretical models that operate at evolutionary time scales predict that environmental shifts are followed by evolution of the maternal effect itself (Kuijper and Hoyle 2015). To complicate matters further, novel environments may release 'cryptic' genetic variation (Lédon-Rettig et al. 2014) as well as increase residual variances (Rowiński and Rogell 2017), affecting the speed with which adaptation can take place (Wilson et al. 2006; cf. Wood and Brodie III 2016; Husby et al. 2011b). These issues, among others, make predicting adaptation at evolutionary time scales (i.e. beyond hundreds of generations) a senseless exercise when the goal is to use 'real' parameters, as these very parameters originate from a mere snapshot of the environment.

This brings us to the question of whether we can quantify real evolutionary responses resulting from maternal effects in wild populations. Indeed, several papers have shown the potential evolutionary importance of maternal effects in wild populations (McAdam and Boutin 2004; e.g. Badyaev 2005; Wilson et al. 2005; McFarlane et al. 2015), corroborated by laboratory studies (e.g. Yanagi and Tuda 2010; McGlothlin and Galloway 2013; Munday et al. 2017). Note that the maternal effects addressed in these studies are the type that in some way represent a female quality or investment (identified as variance components; but see McAdam and Boutin 2004) and therefore differ from our estimated m_p or m_f. The studies cited, making use of past or present selection regimes, showed that the population's capacity to evolve at least partly bears on the presence of maternal effects, but none of the studies has endeavoured to make predictions about future evolutionary trajectories. A way to overcome this would be to make use of estimates originating from populations undergoing substantial directional selection (Kuijper and Hoyle 2015), preferably in combination with long-term cross-fostering experiments (e.g. Postma et al. 2007), which, to date, are rare (Merilä et al. 2001b; Kruuk and Hadfield 2007). The outcome of such long-term studies could serve as input for state-of-the-art models to predict—or hindcast—how a population might evolve in the presence of maternal effects. Combined, these methods may be of use in answering this outstanding question in ecology and evolution.

Our world is changing rapidly, with climate change posing an important threat to populations' persistence (McLaughlin et al. 2002; Thomas et al. 2004). To forecast the viability of populations in the long run, we need to understand the rate at which species can adapt to these novel selection pressures (Visser 2008). We observe apparent evolutionary stasis in several populations (Merilä et al. 2001b), possibly due to the importance of non-Mendelian inheritance systems such as maternal effects. These inheritance systems may greatly affect evolutionary dynamics (Räsänen and Kruuk 2007; Danchin et al. 2011); yet to quantify this in wild populations we need long-term observations of populations under sustained directional selection (Kuijper and Hoyle 2015; cf. McGlothlin and Galloway 2013). Theoretical models can aid in understanding how such inheritance mechanisms can act at evolutionary timescales (Cobben and van Oers 2016) when they are rooted in reality. Basing ourselves on real data, we show that the potential for environmentally plastic maternal effects to alter the rate of adaptation is

limited even under strong, sustained directional selection. To further our understanding of the adaptive potential of non-genetic inheritance, we therefore strongly encourage a closer link between theoretical and empirical work on maternal effects, e.g. through collaboration between research groups with access to real data, to achieve accurate predictions about the evolutionary consequences of maternal effects.

Acknowledgements

Part of this work was funded by an ERC Advanced Grant (339092 - E-Response) to M.E.V. We are grateful to the board of the Hoge Veluwe National Park for allowing us to do fieldwork in all these years. We thank the countless number of students, assistants and volunteers that collected the data, Christiaan Both for making the experimental data available, and Louis Vernooij for carefully managing the database. We are very grateful to Tom Ezard, Charles Baer and an anonymous reviewer for their constructive comments that helped improve the manuscript.

Supplementary Information S8

***Table S8.1**. Results of the linear mixed-model ('animal model') analysis on clutch size in great tits at the Hoge Veluwe (1956–2013; n = 5394 observations from 3328 females).*

Variable		Estimate	SE	F/χ^2 [a]	df	*p*
Fixed effects						
Age mother	1st-timer	9.24	0.05	40.96	1, 4260.6	< 0.0001
	older	9.53	0.05			
Laying date (centred)		-0.08	0.00	250.9	1, 5278.7	< 0.0001
Year of breeding	1956	9.54	1.19	23.57	57, 4755.4	< 0.0001
	1957	8.85	0.54			
	1958	9.95	0.74			
	1959	10.25	0.37			
	1960	9.98	0.33			
	1961	9.96	0.21			
	1962	10.43	0.44			
	1963	10.30	0.36			
	1964	11.07	0.32			
	1965	10.04	0.24			
	1966	10.37	0.22			
	1967	9.37	0.18			
	1968	9.89	0.16			
	1969	9.97	0.16			
	1970	10.09	0.17			
	1971	8.78	0.15			
	1972	9.36	0.15			
	1973	9.35	0.17			
	1974	8.91	0.17			
	1975	7.94	0.14			
	1976	8.35	0.15			
	1977	9.17	0.13			
	1978	9.99	0.17			
	1979	10.00	0.19			
	1980	10.15	0.16			
	1981	8.41	0.16			
	1982	8.26	0.16			
	1983	8.73	0.16			
	1984	7.98	0.15			
	1985	9.61	0.23			

Continued

Table S8.1 *(continued)*

Variable		Estimate	SE	F/χ^2 [a]	df	*p*
	1986	10.61	0.18			
	1987	9.97	0.16			
	1988	7.40	0.14			
	1989	9.22	0.17			
	1990	9.47	0.17			
	1991	7.97	0.15			
	1992	10.01	0.16			
	1993	8.49	0.16			
	1994	9.06	0.16			
	1995	9.58	0.18			
	1996	9.20	0.15			
	1997	10.28	0.18			
	1998	10.33	0.18			
	1999	9.22	0.17			
	2000	9.57	0.19			
	2001	8.12	0.15			
	2002	8.89	0.18			
	2003	9.16	0.14			
	2004	9.30	0.19			
	2005	8.84	0.15			
	2006	9.21	0.21			
	2007	8.69	0.16			
	2008	8.94	0.15			
	2009	10.52	0.18			
	2010	10.19	0.17			
	2011	10.64	0.16			
	2012	8.41	0.14			
	2013	8.10	0.19			
Random effects[b]						
Female identity (permanent environment)		0.50	0.14	13.29	1	0.0003
Additive genetic effect		0.76	0.14	35.54	1	<0.0001
Nest box identity		0.05	0.02	10.22	1	0.0014
Mother identity (maternal effect)		0.16	0.11	2.19	1	0.14
Residual		1.80	0.06			

[a]Conditional Wald *F* tests were used to test significance of fixed effects; likelihood-ratio tests were used for random effects.

[b]Likelihood-ratio tests and parameter estimates based on models excluding non-significant term.

***Table S8.2**. Results of the linear mixed-effects model analysis on the effect of experimentally manipulated brood size on fledgling weight in great tits at the Hoge Veluwe (1983–1990; n = 2145 observations from 309 broods). Marginal and conditional R^2 of the final model (Nakagawa and Schielzeth 2013) were 0.25 and 0.59, respectively.*

Variable		Estimate[b]	SE	*F*	df[c]	*p*
Fixed effects[a]						
Brood size		–0.10	0.02	102.79	1, 287.7	<0.0001
Year of breeding	1983	0.35	0.06	7.45	7, 233.6	<0.0001
	1984	0.49	0.05			
	1985	0.67	0.07			
	1986	0.81	0.05			
	1987	0.79	0.06			
	1988	0.53	0.06			
	1989	0.85	0.09			
	1990	0.62	0.07			
Brood size × year	BS:1983	–0.09	0.02	3.31	7, 260.7	0.0021
	BS:1984	–0.08	0.01			
	BS:1985	–0.06	0.02			
	BS:1986	–0.06	0.01			
	BS:1987	–0.05	0.01			
	BS:1988	–0.30	0.06			
	BS:1989	–0.07	0.03			
	BS:1990	–0.05	0.02			
Orginal clutch size				3.32	1, 288.6	0.07
Hatching date				0.42	1, 284.9	0.52
Random effects		Variance	SD			
Brood ID: mother ID		0.07	0.26			
Mother identity		0.04	0.21			
Residual		0.13	0.36			

[a]Fixed terms expressed in boldface appeared in the final model; continuous variables are centred around their mean for each year.

[b]Parameter estimates are on a logit scale and given only for significant terms.

[c]Denominator degrees of freedom estimated using Kenward–Roger approximation.

***Table S8.3**. Results of the mixed-effects logistic regression analysis on offspring recruitment probability in great tits at the Hoge Veluwe (1973–2013; n = 24320 observations from 3600 broods).*

Variable	Estimate	SE	χ^2	df	p[a]
Fixed effects					
Intercept	–19.06	2.92			
Fledgling weight	1.62	0.33	171.33	1	0.0006
[Fledgling weight]^2	–0.04	0.01	20.74	1	0.0005
Random effects	Variance	SD			
Year of breeding	0.43	0.66			
Brood identity: year of breeding	0.32	0.57			

Note: area under ROC curve: 0.80 (95% CI: 0.79–0.81)

[a]*p*-values of the likelihood-ratio test were simulated using parametric bootstrapping with 2000 simulations

***Table S8.4**. Result of the linear mixed-effects analysis on centred clutch size (M_{CS_i}) estimating the environmentally plastic maternal effect in first-time breeding great tits at the Hoge Veluwe (1973–2013; n = 510).*

Variable	Estimate	SE	F	df[b]	p	R^2 [c]
Fixed effects[a]						
Intercept	–0.25	0.06				
Fledgling weight	0.13	0.05	8.16	1, 498.6	0.0045	0.006
Breeding value	2.95	0.09	1007.39	1, 504.4	<0.0001	0.640
Random effects[d]	Variance	SD				
Year of breeding	0.01	0.12				
Residual	1.27	1.13				

[a]Fixed terms expressed in boldface appeared in the final model; fledgling weight was centred around its mean for each year.

[b]Denominator degrees of freedom estimated using the Kenward–Roger approximation.

[c]Marginal R^2 for fixed effects based on Nakagawa and Schielzeth (2013), termed M^2_{CS} for the maternal effect in Table 8.1 (main text); conditional R^2 of final model: 0.646.

[d]Brood of origin was left out to allow for comparison of relative clutch size among members of the whole population.

***Table S8.5**. Result of the linear mixed-effects analysis on centred clutch size (M_{CS_i}) estimating the fixed maternal effect in addition to the plastic maternal effect in first-time breeding great tits at the Hoge Veluwe (1973–2013; n = 510).*

Variable	Estimate	SE	*F*	df[c]	*p*	R^2 [d]
Fixed effects[a]						
Intercept	–0.25	0.06				
Breeding value	3.17	0.10	1101.40	1, 504.6	<0.0001	0.692
Maternal clutch size (z_{CS_m})[b]	–0.21	0.03	42.83	1, 488.5	<0.0001	0.027
Fledgling weight (z_{FW_i})[b]	0.10	0.04	5.21	1, 497.7	0.0228	0.003
Mean fledgling weight ($\bar{z}_{FW_t}$)			2.41	1, 23.8	0.13	
$z_{CS_m} \times \bar{z}_{FW_t}$			0.02	1, 422.6	0.90	
Random effects[e]	Variance	SD				
Year of breeding	0.02	0.14				
Residual	1.17	1.08				

[a]Fixed terms expressed in boldface appeared in the final model.

[b]Values were centred around the mean for each year.

[c]Denominator degrees of freedom estimated using the Kenward–Roger approximation.

[d]Marginal R^2 for fixed effects based on Nakagawa and Schielzeth (2013), termed M^2_{CS} for the maternal effect in Table 8.1 (main text); conditional R^2 of final model: 0.726.

[e]Brood of origin was left out to allow for comparison of relative clutch size among members of the whole population.

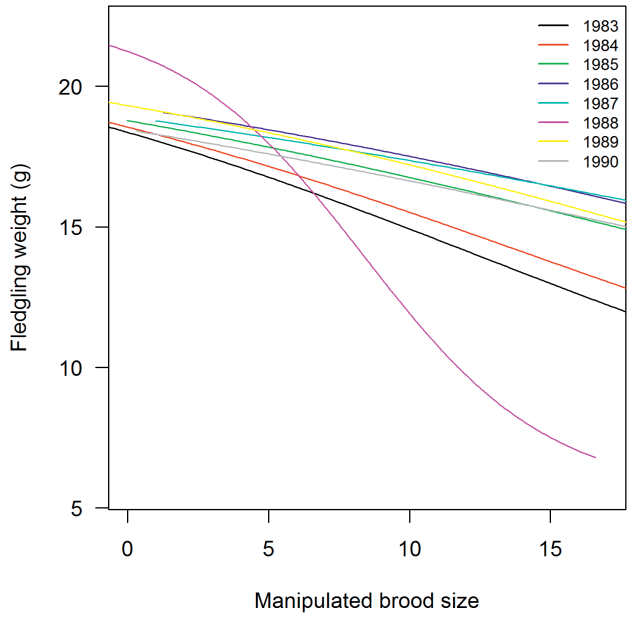


Figure S8.1. *Fledgling weight as a function of maternal clutch size, resulting from eight years of experimental brood size manipulations. Lines are back-transformed regression estimates from a linear mixed-effects model with a logit-transformed response variable.*

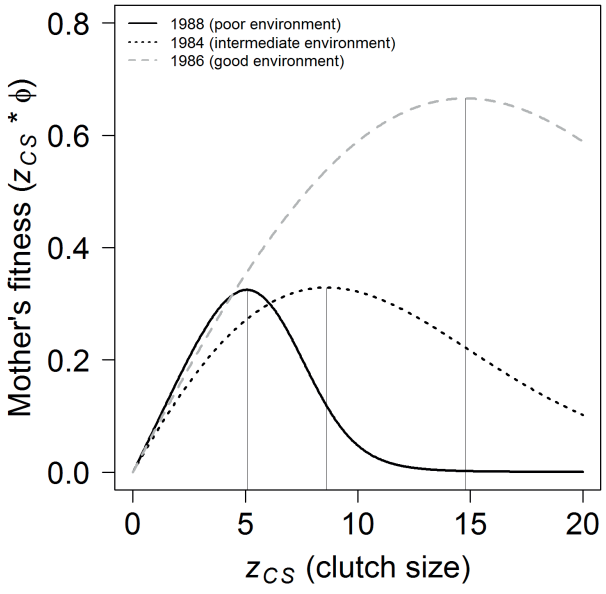


Figure S8.2. *Fitness curves associated with clutch size given different environments (solid line: poor environment; dashed line: intermediate environment, dotted line: good environment). Vertical lines denote the optimum phenotype for each environment.*

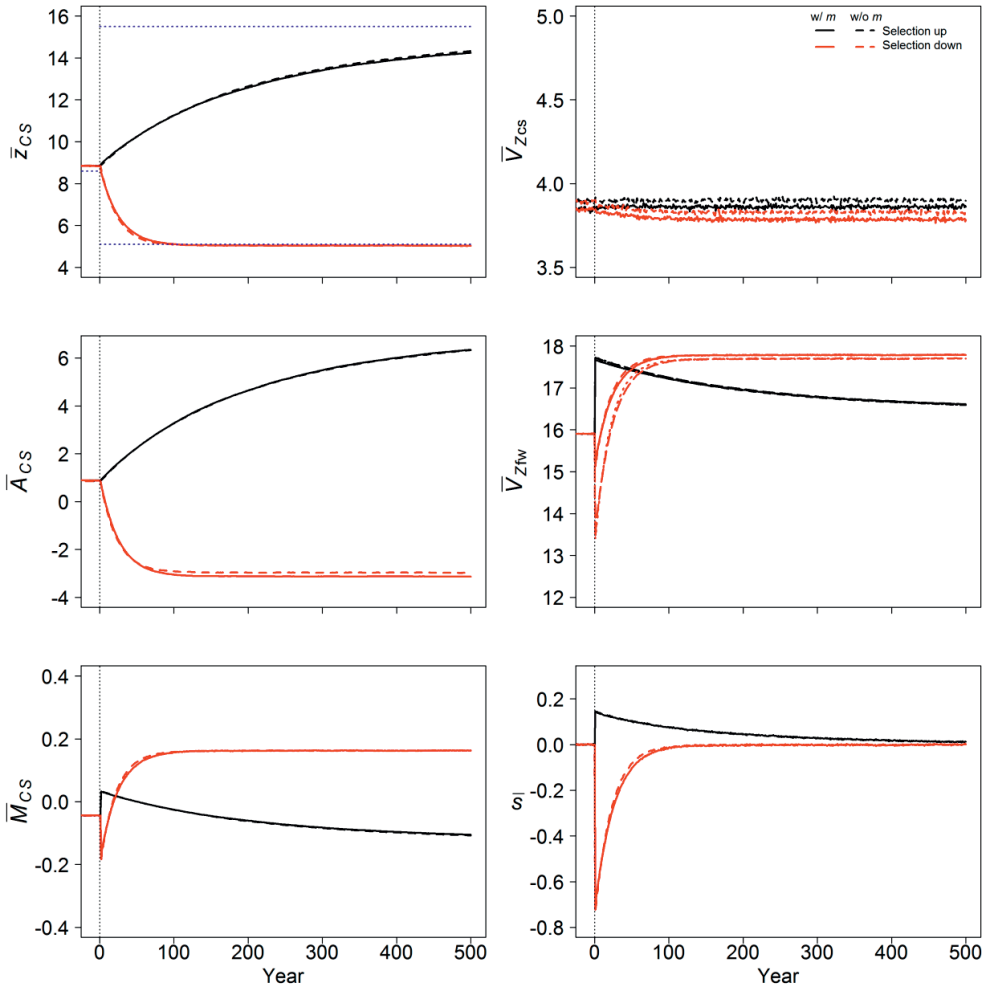


Figure S8.3 *(continued on next page). Environmentally plastic maternal effect-dependent (solid lines) or -independent (dashed lines) changes in mean phenotype $\bar{z}_{CS}$ (avian clutch size), genotype $\bar{A}_{CS}$, additive maternal component $\bar{M}_{CS}$, phenotypic variance $\bar{V}_{z_{CS}}$, fledgling mass $\bar{z}_{FW}$, and standardised selection differentials $\bar{s}$ over a time span of 500 years in an avian population modelled after the biology of the great tit at the Hoge Veluwe. The maternal effect is the effect of maternal clutch size on offspring clutch size via fledgling weight. The vertical line denotes the pre-burn-in period, after which selection moves from an intermediate clutch size to either a large (good environment) or small (poor environment) clutch; the blue dotted line in the $\bar{z}_{CS}$ panel denotes the 'optimal' phenotype, i.e. z at W_{max}. In the $\bar{z}_{FW}$ panel, wide-dashed lines indicate fledgling weight before selection in presence of maternal effects (with solid lines indicating weight after selection), whereas dotted lines indicate weight before selection in absence of maternal effects (with narrow-dashed lines indicating weight after selection). Lines are the means of population averages over 1000 simulation runs. Input parameters are $\bar{V}_{z_{CS}}$ = 3.91, h^2_{CS} = 0.24, M^2_{CS} = 0.006, $m_p \approx$ –0.13, –0.04 or –0.03 (see main text and Table 8.1 for details).*

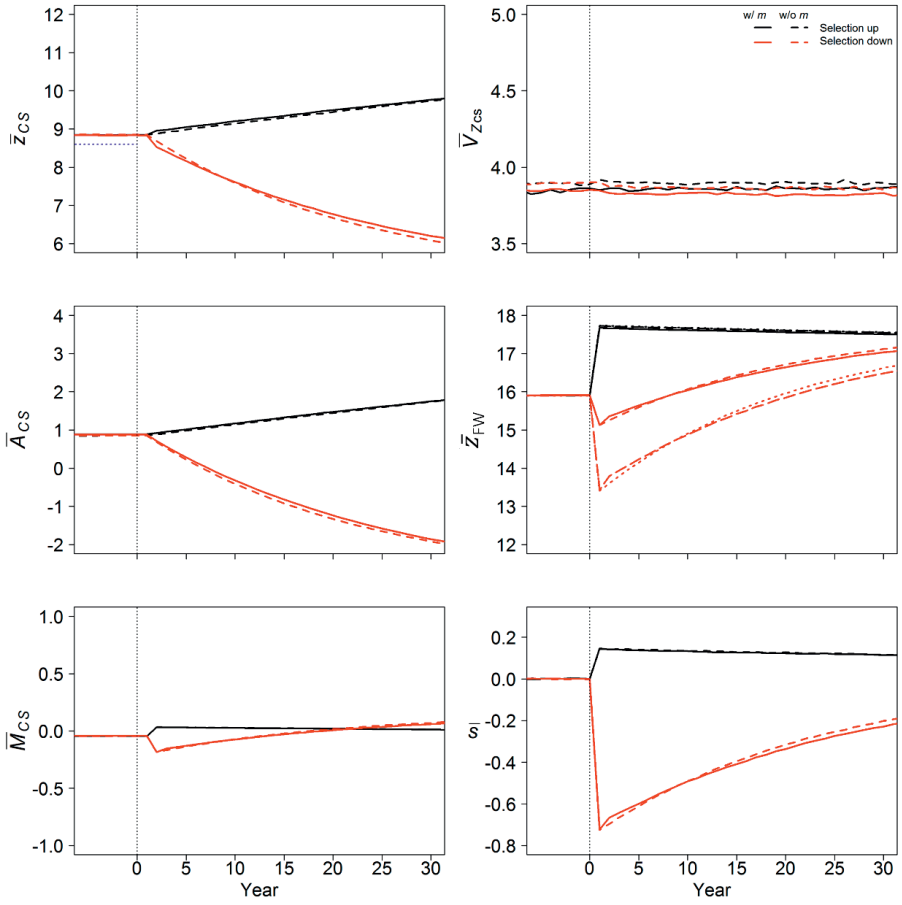


Figure S8.3 *(continued). Snapshot of the model simulation covering the first 30 years following the environmental shift.*

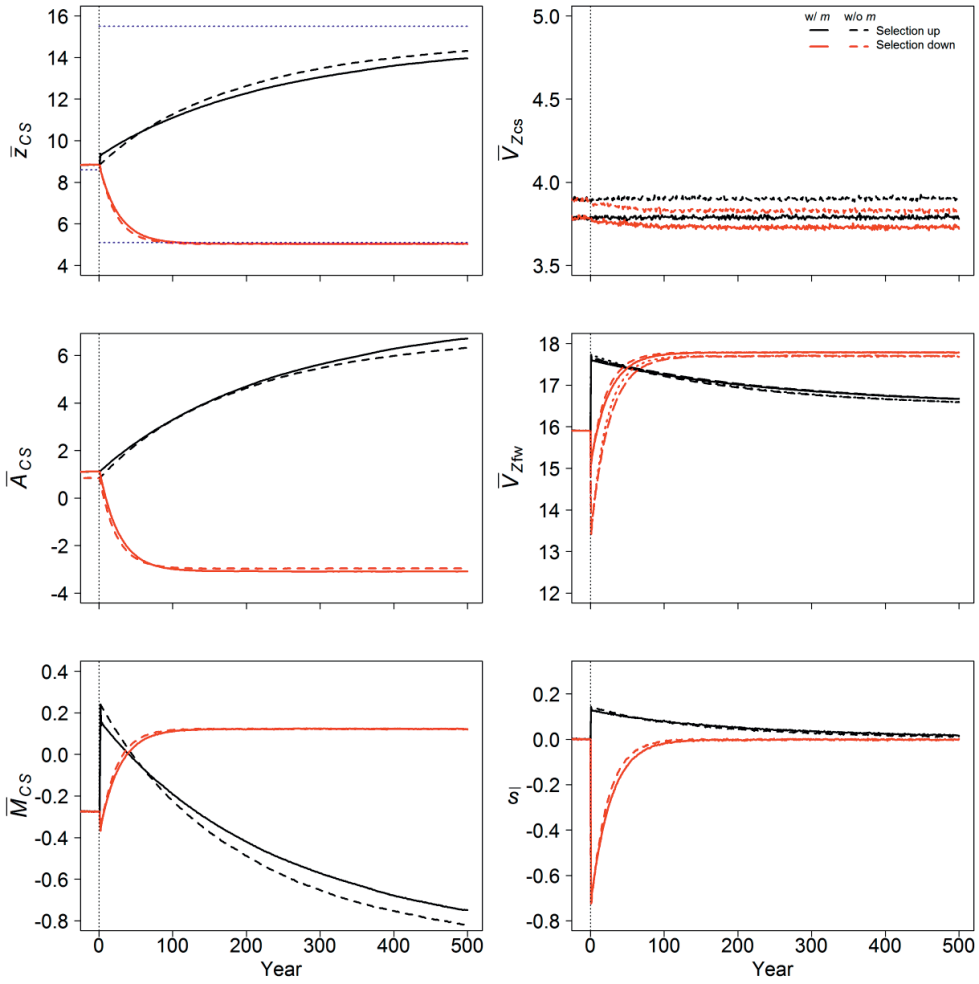


Figure S8.4 *(continued on next page). Fixed maternal effect-dependent (solid lines) or -independent (dashed lines) changes in mean phenotype $\bar{z}_{CS}$ (avian clutch size), genotype $\bar{A}_{CS}$, additive maternal component $\bar{M}_{CS}$, phenotypic variance $\bar{V}_{z_{CS}}$, fledgling mass $\bar{z}_{FW}$, and standardised selection differentials $\bar{s}$ over time span of 500 years in an avian population modelled after the biology of the great tit at the Hoge Veluwe. The maternal effect is the direct effect of maternal on offspring clutch size. The vertical line denotes the pre-burn-in period, after which selection moves from an intermediate clutch size to either a large (good environment) or small (poor environment) clutch; the blue dotted line in the $\bar{z}_{CS}$ panel denotes the 'optimal' phenotype, i.e. z at W_{max}. In the $\bar{z}_{FW}$ panel, wide-dashed lines indicate fledgling weight before selection in the presence of maternal effects (with solid lines indicating weight after selection), whereas dotted lines indicate weight before selection in absence of maternal effects (with narrow-dashed lines indicating weight after selection). Lines are the means of population averages over 1000 simulation runs. Input parameters are $V_{z_{CS}}$ = 3.91, h^2_{CS} = 0.24, M^2_{CS} = 0.027, m_f = –0.21 (see main text and Table 8.1 for details).*

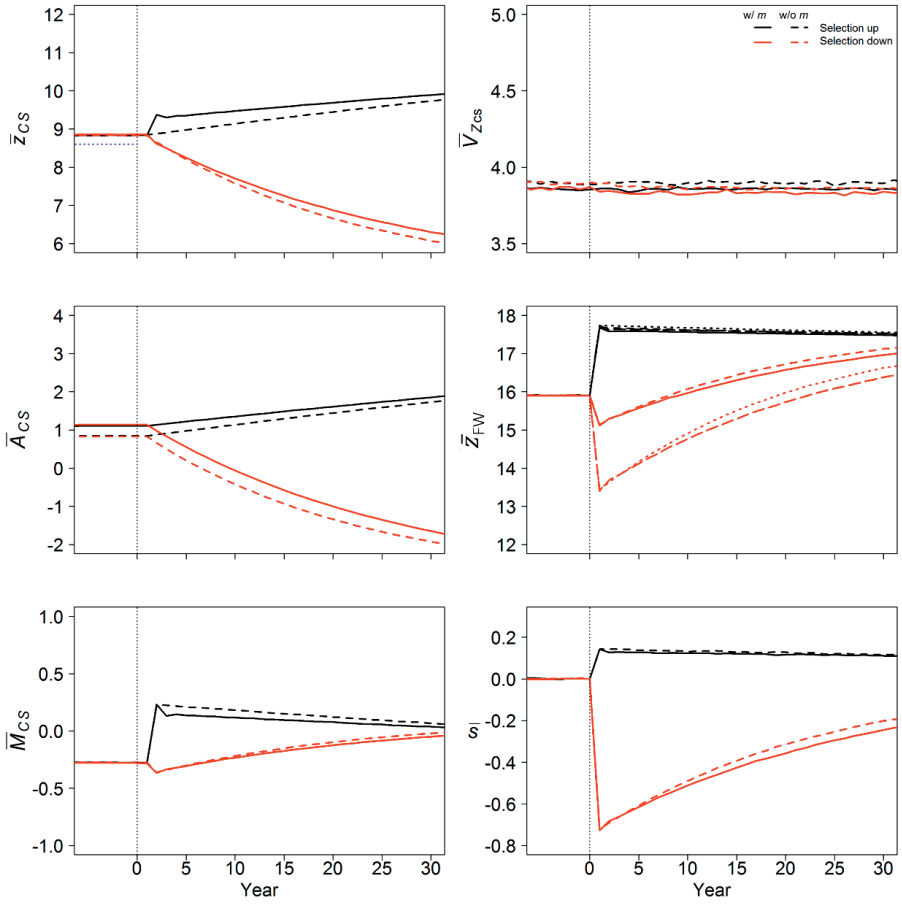


Figure S8.4 *(continued). Snapshot of the model simulation covering the first 30 years following the environmental shift.*

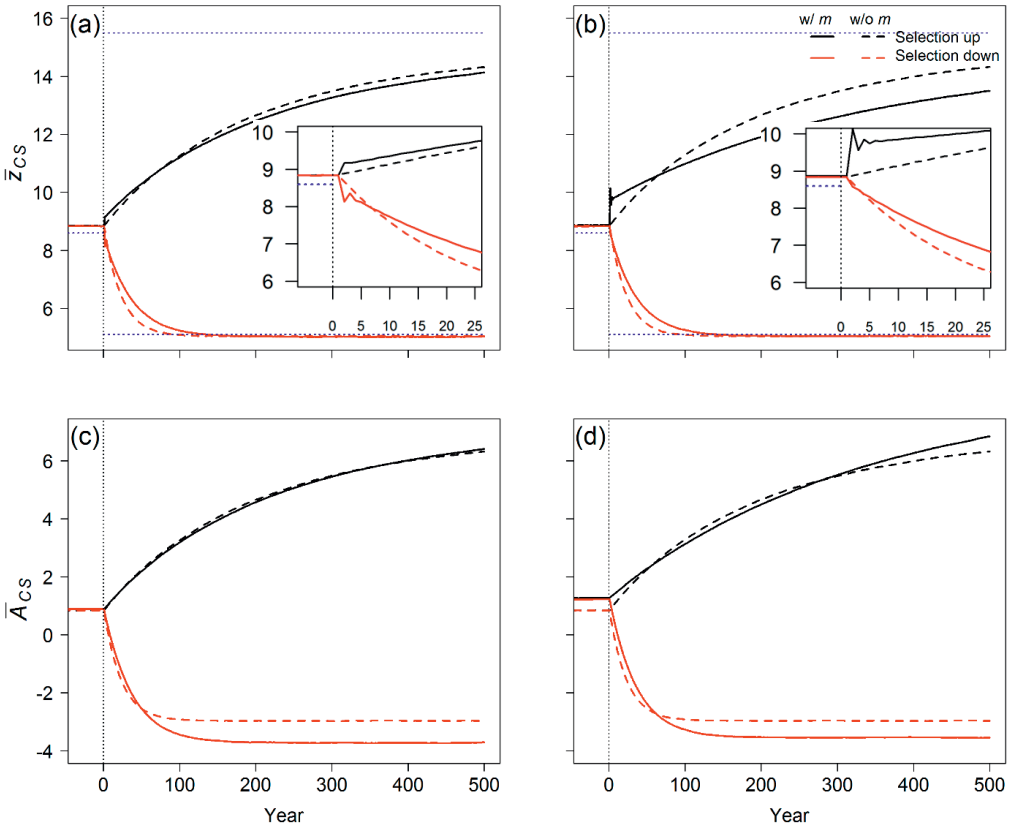


***Figure S8.5**. Exaggerated maternal effects: predicted mean phenotypic (a, b) and genetic (c, d) change in avian clutch size (modelled after the great tit at the Hoge Veluwe) in response to selection when considering an exaggerated, environmentally plastic (i.e. via fledgling weight; a, c) or fixed (i.e. via maternal clutch size; b, d) maternal effect (solid lines), or no maternal effect at all (dashed lines). Phenotypic responses in the first 25 years are magnified in the insets in panels a and b. The vertical line denotes the pre-burn-in period, after which selection moves from an intermediate clutch size to either a large (good environment) or small (poor environment) clutch; the blue dotted lines in panels a and b denote the 'optimal' phenotype, i.e. z at W_{max}. Lines are the means of population averages over 1000 simulation runs. Input parameters are $V_{z_{CS}}$ = 3.91, h^2_{CS} = 0.24, M^2_{CS}= 0.006 (panel a and c) and 0.027 (panel b and d); $m_p \approx -0.52, -0.16$ or -0.09, $m_f = -0.5$ (see main text and Table 8.1 for details).*

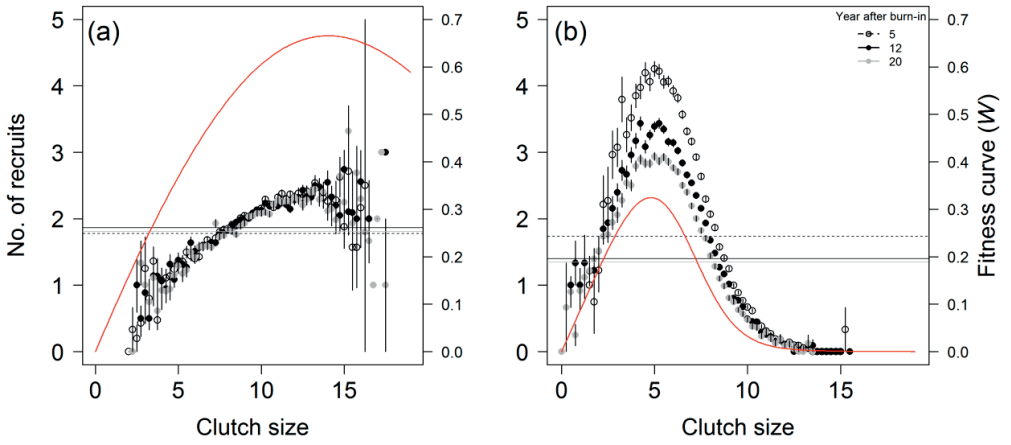


***Figure S8.6**. The effect of maternal clutch size on fitness in a good (a) and poor (b) environment. Shown are the number of recruits (mean ± SE from 100 simulations; primary y axis) in three different generations following the environmental shift (year 5, 12, and 20), with the horizontal lines denoting the mean number of recruits associated with each year. Note that the heights of the 'curves' have no inherent meaning, as the total number of recruits was more or less equal in any given year. The red line denotes the fitness curves for the optimal trait value (eqn.(8.7) in main text; secondary y axis).*

Chapter 9

Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations

Jip J.C. Ramakers, Antica Culina, Marcel E. Visser & Phillip Gienapp

ABSTRACT

Predicting the rate of adaptation to environmental change in wild populations is important for understanding evolutionary change. However, predictions may be unreliable if the two key variables affecting the rate of evolutionary change, heritability and selection, are both affected by the same environmental variable. To determine how general such an environmentally induced coupling of heritability and selection is, and how this may influence the rate of adaptation, we made use of freely accessible, open data on pedigreed wild populations to answer this question at the broadest possible scale. Using 16 populations from 10 vertebrate species, which provided data on 50 traits (body mass, morphology, physiology, behaviour and life history), we found evidence for an environmentally induced relationship between heritability and selection in only 6 cases, with weak evidence that this resulted in an increase or decrease in expected selection response. We conclude that such a coupling of heritability and selection is unlikely to strongly affect evolutionary change even though both heritability and selection are commonly postulated to be environment dependent.

Nature Ecology & Evolution (2018) **2**, *1093–1103*

Introduction

In the face of global environmental change, it is imperative to understand whether and how fast populations can adapt to novel conditions to be 'rescued' by evolution (Carlson et al. 2014). Despite evidence of genetic variance and selection in many wild populations, genetic response to selection (adaptive micro-evolution or rate of adaptation) in natural populations is rarely observed (Gienapp et al. 2008; Merilä and Hendry 2014). An apparent lack of a response to selection may have a variety of biological and/or methodological causes (Merilä et al. 2001b; Kruuk et al. 2003). One potential reason is an environmentally induced coupling between selection and additive genetic variation, which can mask the true evolutionary potential of a population if not recognised. By its definition, selection is mediated by the environment (Darwin 1859; Wade and Kalisz 1990) and has been shown to vary from season to season and between geographical regions, depending on resource availability (including mating opportunity) and predation pressure (e.g. Hairston and Dillon 1990; Grant and Grant 2002; Gosden and Svensson 2008a; Siepielski et al. 2009; Weese et al. 2010; but see Morissey and Hadfield 2012). Genetic variation is, however, also known to vary with the environment (known as genotype-by-environment interaction), being sometimes increased and sometimes reduced under benign conditions (e.g. when mean fitness in the population is high; Hoffman and Merilä 1999; Lédon-Rettig et al. 2014; Wood and Brodie III 2016), although the ecological drivers of changes in the genetic variance–covariance matrix remain largely unknown (Wood and Brodie III 2015).

Although the environmental dependency of both selection and genetic variation has been thoroughly documented, our knowledge on how they may interact to result in evolutionary change in natural populations is very limited. In their recent review, Wood and Brodie (2016) identified 23 studies that measured environmental effects on selection and 28 studies that measured environmental effects on additive genetic variation. Overall, reviewing a great variety of taxa, environments and traits, they found that environmental effects on selection and genetic variance were broad and inconsistent. Importantly, most studies on environment-dependent genetic variation were done in laboratory settings (and those on selection mostly in wild populations) and extrapolating laboratory findings to natural conditions is not necessarily straightforward. To date, only two studies of natural populations have measured how both genetic variation and selection within the same trait covaried across environments. A study on Soay sheep (*Ovis aries*) demonstrated increased selection for a higher birth weight in harsh environments, whereas total genetic variance was highest in benign environments (Wilson et al. 2006). The opposite was found in the great tit (*Parus major*), where warmer springs, which are associated with increased mismatch between offspring energetic demands and food availability, were associated with stronger selection for early egg-laying as well as high additive genetic variance for that trait (Husby et al. 2011b). Thus, in the former example, selection and genetic variance covaried with the environment in opposite directions, whereas in the latter example they did so in the same direction. The negative covariance between selection and genetic variation in Soay sheep led to an 21% decrease in expected response to selection as opposed to a situation where genetic variance was assumed to not vary with the

environment (Wilson et al. 2006). In the great tit, the positive association between additive genetic variance and selection gradients resulted in a 20% increase in predicted response to selection as compared to a situation where heterogeneity in both selection and genetic variance was ignored (Husby et al. 2011b). A more recent study investigated the environmental dependency of genetic variance and selection in several morphological traits in the Soay sheep population, but did not explicitly address the relationship between them, presumably since environment-dependent genetic variance was found to be absent (Hayward et al. 2018). The environmental coupling of selection and genetic variance (or heritability) may therefore provide an important explanation for the discrepancy between observed and expected responses to selection in some natural populations, but the prevalence of this mechanism—and how it may alter the expected response to selection—in wild populations remains largely unknown.

We investigated the prevalence and strength of an environmentally induced correlation between heritability and selection—and its expected evolutionary consequence—in a variety of wild populations. We searched for multiannual, pedigreed datasets on wild populations freely accessible from online data repositories and used these data to quantify environment-dependent additive genetic variation (using random regression animal models) and standardised selection gradients for a suite of life-history, morphological, behavioural, physiological and body mass traits. We then regressed heritability against selection for 50 traits from 16 populations and compared expected selection responses with and without considering environmental heterogeneity in heritability. We had no specific expectation as to the prevalence of a correlation between heritability and selection but, if anything, expected it to be more common in life-history and morphological traits, since selection in these traits tends to be strong and variable (Kingsolver et al. 2001). Our approach using open data (Culina et al. 2018) speaks to recent recommendations to use available data to address novel, outstanding questions in ecology and evolution that transcend a single study system (Whitlock et al. 2010; Hampton et al. 2013).

Results

Data acquisition and author response

We performed a search in online data repositories (see Methods) for multiannual (≥ 6 years) datasets containing pedigrees of wild populations accompanied by phenotypic measures on individually marked animals. From 106 acquired pedigreed datasets (Supplementary Table 9.1), we used 14 that were suitable for our analysis (see Methods). We added one unpublished dataset from our own database (pied flycatcher, *Ficedula hypoleuca*). These 15 datasets comprised 16 different populations, spanning ten species, eight of which were avian species, one a lizard, and one a mammal (Table 9.1).

Authors were generally supportive of the use of their data. We contacted 14 authors (of 18 datasets) about our use of their data and found that 4 datasets were not usable. This was mainly related to a bias in our approximation of the environment, i.e. the population-

Table 9.1. Overview of studies used in the gene-by-environment and selection analyses. *The studies and traits listed are the ones that met inclusion criteria and could be successfully analysed (see Table S9.1 for an untrimmed overview of datasets).*

Species	Refs.	Locality	Aim of original study	Trait	N_{obs}	N_{ind}	N_{year}	h^2 (s.e.)	Select-ion?
Cinclus cinclus	1,2	ZU	Estimate biases in inbreeding depression	Wing length (mm)	1132	672	18	**0.464 (0.066)**	F
Chyonomis nivalis	3	CW	Predict genetic changes in body mass	Body mass (g)	3382	931	9	a: **0.188 (0.064)**	F
								j: 0.069 (0.056)	V
				Body length (mm)*	2761	791	8	a: **0.139 (0.062)**	–
								j: 0.093 (0.070)	V
				Tail length (mm)*	3382	931	9	a: **0.281 (0.079)**	–
								j: 0.101 (0.054)	V
Cyanistes caeruleus	4,5	TA	Estimate the developmental stability of behavioural syndromes	Nestling handling aggression	6149	6149	8	**0.235 (0.043)**	V
				Adult handling aggression	1633	1103	8	**0.283 (0.057)**	V
				Nestling breath rate (breaths/s)	5863	5863	7	**0.266 (0.037)**	V
				Adult breath rate (breaths/s)	1526	1031	7	**0.194 (0.063)**	V
	6,7	DR	Investigate spatial variation in G-matrix in populations with contrasting population history and selective environment	Incubation duration (d)	1104	740	24	**0.195 (0.074)**	V
				Laying date	1104	740	24	**0.214 (0.094)**	V
				Clutch size	1104	740	24	**0.345 (0.104)**	V
				Wing length (mm)	2916	1597	24	**0.374 (0.040)**	V
				Body mass (g)	2916	1597	24	**0.347 (0.038)**	V
		EP	As above	Incubation duration (d)	997	637	35	0.011 (0.102)	V
				Laying date	997	637	35	0.043 (0.148)	V
				Clutch size	997	637	35	0.108 (0.145)	V
				Wing length (mm)	2260	1187	26	**0.287 (0.065)**	V
				Body mass (g)	2260	1187	26	**0.332 (0.065)**	V
	8,9	EB	Quantify selection on parental care to explain stasis in evolution of offspring body size	Wing length (mm)	1677	847	8	**0.448 (0.151)**	V
				Body mass (g)	1677	847	8	**0.262 (0.117)**	V
Falco tinnunculus	10,11	SP	Quantify multivariate heredity of colouration, mass and immunity	Tail-band width (mm)	688	444	17	**0.699 (0.086)**	F
Ficedula hypoleuca	12	HV	None (published for the purpose of this paper)	Laying date	3044	2211	39	**0.149 (0.072)**	F
Hirundo rustica	13,14	SP	Estimate genetic correlation between arrival date and life-history traits	Spring arrival date	2337	1407	17	0.131 (0.076)	F
Lacerta agilis	15,16	AS	Test for trade-off between offspring size and number	Clutch size	472	288	10	0.294 (0.066)	F
				Laying date	370	236	9	0	F
				Mean offspring mass (g)†	452	279	10	**0.384 (0.064)**	F

Continued

Table 9.1. Overview of studies used in the gene-by-environment and selection analyses *(continued)*

Species	Refs.	Locality	Aim of original study	Trait	N_{obs}	N_{ind}	N_{year}	h^2 (s.e.)	Select-ion?
Parus major	17,18	WW	Estimate genetic variance in colour expression across the visual spectrum	Plumage reflectance at 349 nm‡	2904	1618	6	0.015 (0.035)	V
				Plumage reflectance at 449 nm‡	2901	1616	6	**0.194 (0.051)**	V
				Plumage reflectance at 549 nm‡	2901	1616	6	**0.098 (0.043)**	V
				SWS ratio (plumage reflectance)	2901	1616	6	**0.336 (0.051)**	V
				Double cone (plumage reflectance)	2901	1616	6	**0.092 (0.044)**	V
				Wing length (mm)	2892	1614	6	**0.478 (0.054)**	V
				Body mass (g)	2878	1613	6	**0.357 (0.052)**	V
	19,20	WW	Investigate the genetic architect of a suite of parameters in two populations	Adult body mass	2919	1358	12	0.004 (0.034)	V
				Offspring fledgling weight (g)†	3162	328	13	0.022 (0.107)	V
				Wing length	3206	1408	12	0.055 (0.042)	V
		HV/WH	As above	Adult body mass (g)	1543	477	16	**0.472 (0.027)**	F§
				Clutch size	1585	943	17	0.058 (0.181)	F§
				Offspring fledgling weight (g)†	8569	744	17	0	F§
				Wing length	1908	1275	17	**0.158 (0.140)**	F§
	21,22	HV	Test for bias in selection on life-history traits	Clutch size	4054	2861	57	**0.318 (0.055)**	F
				Laying date	4054	2861	57	**0.157 (0.052)**	F
		VL	As above	Clutch size	3700	2368	52	**0.306 (0.044)**	F
				Laying date	3700	2368	52	**0.277 (0.048)**	F
	23,24	HV	Estimate heritability of within-family variance in fledgling weight	Fledgling weight (g)	17535	17535	36	**0.235 (0.027)**	V
				Clutch size	2175	1598	36	**0.282 (0.082)**	F
Perisoreus infaustus	25,26	OB	Disentangle plastic and genetic changes in body mass	Body mass (g)	1619	1025	30	**0.408 (0.058)**	F
				Wing length (mm)	1453	1016	28	**0.516 (0.056)**	F
Passerculus sandwichensis	27,28	KI	Investigate the relationship between heritability/evolvability and selection	Day-8 to Yr-1 wing length (mm)	2839	2469	20	a: **0.353 (0.072)**	F
								j: **0.430 (0.063)**	F
				Day-8 to Yr-1 tarsus length (mm)*	1913	1615	20	a: **0.398 (0.070)**	–
								j: **0.292 (0.080)**	V
				Day-8 to Yr-1 body mass (g)	2469	2362	20	a: **0.064 (0.018)**	F
								j: **0.330 (0.059)**	V

Note. **Locality**: AS = Asketunnan, Sweden; CW = Churwalden, Switzerland; DR = D-Rouvière, France; EB = Edinburgh, UK; EP = E-Pirio, France; HV = Hoge Veluwe, NL; KI = Kent Island, Canada; OB = Ostrobothnia, Finland; SP = Spain; TA = Tammisaari, Finland; VL = Vlieland, the Netherlands; WH = Westerheide, NL; WW = Wytham Woods / Bagley Woods, UK; ZU = Zürich, Switzerland. **References**: 1,2 (Becker et al. 2016b, a); 3 (Bonnet et al. 2017); 4,5 (Class and Brommer 2015b, a); 6,7 (Delahaie et al. 2017a, b); 8,9 (Thomson et al. 2017b, a); 10,11 (Kim et al. 2013b, a); 12 (Ramakers et al. 2018); 13,14 (Teplitsky et al. 2011b, a); 15,16 (Ljungström et al. 2016b, a) ; 17,18 (Evans and Sheldon 2015a, b); 19,20 (Santure et al. 2015b, a); 21,22 (Reed et al. 2016b, a); 23,24 (Mulder et al. 2016b, a); 25,26 (Gienapp and Merilä 2014b, a);

27,28 (Wheelwright et al. 2014b, a). **Selection**: F = fecundity; V = viability; '–' = disregarded due to fixed nature of trait in adults. $\mathbf{N_{obs}/N_{ind}/N_{year}}$: number of observations/individuals/years (environments); $\boldsymbol{h^2}$: narrow-sense trait heritability (a: adult; j: juvenile; significant values in boldface).

*'Fixed' trait: may change from juvenile to adult stage but are assumed to be relatively constant within adult lifespan.

†Trait considered maternal.

‡Trait consitutes one out of a range of 198 2-nm bands; three bands equally spaced apart and spanning most of the gradient were chosen for analysis.

§Fecundity based on number of fledglings, not recruits.

mean trait value (see Methods) and selection in a given year when a non-random portion of the population was not represented in the dataset. Only in two cases authors were initially reluctant to cooperate, but all authors eventually informed us about the appropriateness of our analyses of their data (see Chapter 7 for a full account on author correspondence associated with this article).

Estimating environment, heritability and selection

From the included datasets, we extracted a total of 50 morphological, behavioural, physiological, life-history and body mass traits. We used these traits first to estimate a standardised measure of the environment, the standardised annual population-mean trait value (Yates and Cochran 1938; Finlay and Wilkinson 1963; Lynch and Walsh 1998; James 2009). We estimated the heritability (h^2, the relative additive genetic variation) of the traits and found that the majority showed significant heritable variation within the population (Table 9.1). We then fitted random regression animal models (RRAMs) with an interaction between the additive genetic effect and the standardised measure of the environment. We extracted environment-dependent heritability estimates resulting from these RRAMs (as heritability determines the short-term evolutionary change) and regressed them against annual standardised selection gradients (Lande and Arnold 1983; Hereford et al. 2004; Morrissey and Sakrejda 2013) (β'; Fig. S9.2), while accounting for uncertainty in both predictor and response. In 6 out of 50 cases (all in bird species), this led to a statistically significant relationship between selection and heritability (Fig. 9.1; Table S9.2). None of the 14 life-history traits exhibited such a relationship, despite considerable variation in both selection and heritability. We found a positive, significant relationship in nestling body mass in *Passerculus sandwichensis*, based on viability selection on nestlings (slope [95% bootstrapped confidence interval] = 0.102 [0.045, 0.191], r^2 = 0.369 [0.089, 0.596]). As the only morphological example, nestling tarsus length in *P. sandwichensis* showed a significantly negative correlation based on viability selection (slope = –0.057 [–0.118, –0.023], r^2 = 0.148 [0.038, 0.340]). Finally, four avian physiological and behavioural traits exhibited a significant association between heritability and selection, all based on viability selection: *P. major* plumage reflectance at 349 nm (slope = 0.018 [0.009, 0.037], r^2 = 0.284 [0.052, 0.507]), 549 nm (slope = –0.190 [–0.440, –0.040], r^2 = 0.467 [0.054, 0.949]) and spectral sensitivity (double cone; slope = –0.055 [–0.173, –0.010], r^2 = 0.248 [0.027, 0.751]), and *Cyanistes caeruleus* adult handling aggression (slope = 0.001 [0.0004, 0.003], r^2 = 0.009 [0.002, 0.021]).

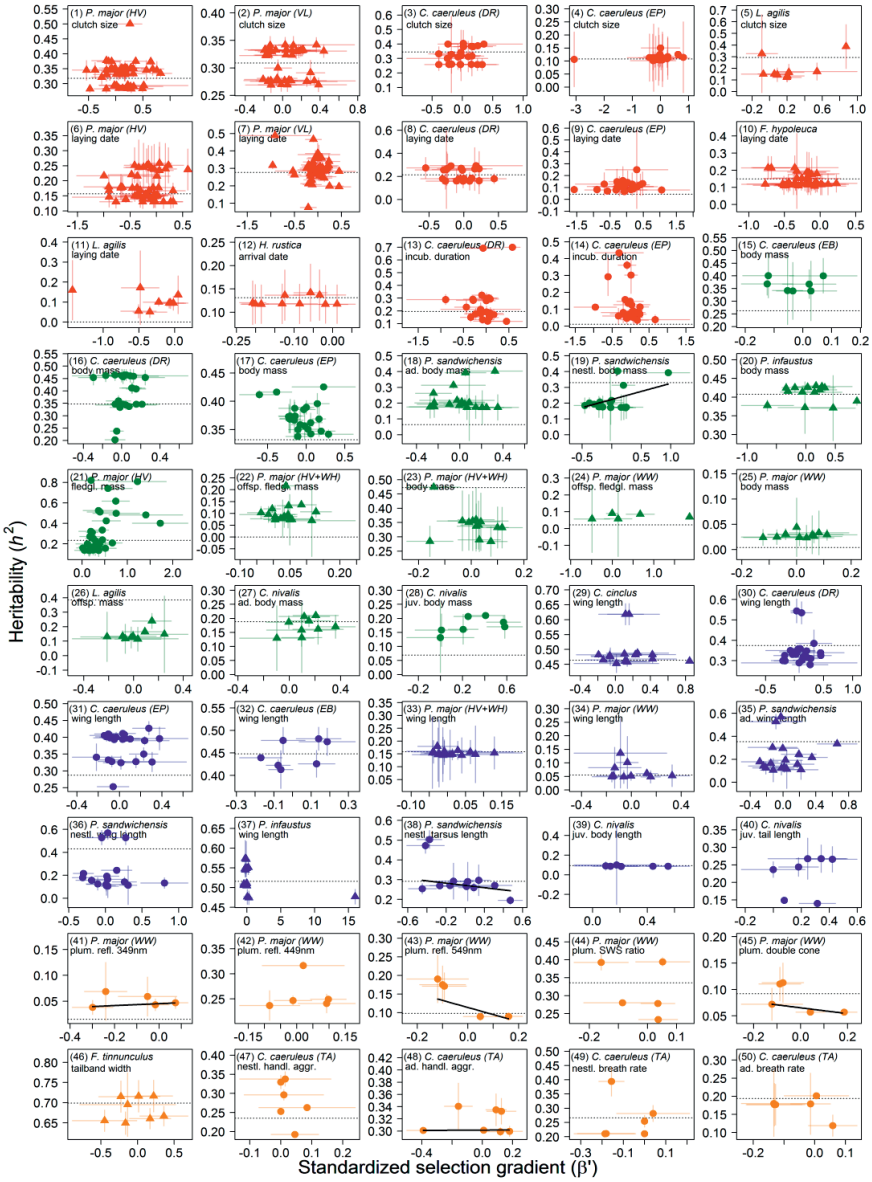


Figure 9.1. Heritability as a function of the standardized selection gradient. *Standard errors (SEs) are omitted when $SE_{h^2} > 0.5$ and $SE_{\beta'} > 1$ for visual aid. Regression lines result from weighted least-squares regression models (weights: $1/[(SE_{h^2})^2]$), with bootstrapping to account for uncertainty in β', shown only when the 95% CI did not include zero. Colours denote different trait classes (red: life history; green: body mass; blue: morphology; orange: miscellaneous), whereas shapes indicate selection based on survival (circles) or based on number of fledglings or recruits (triangles). Dotted horizontal lines denote the constant heritability as estimated from a standard animal model. Duplicate traits (from same population but different dataset) are not shown. Data sources by panel:* ***1,2,6,7*** *(Reed et al. 2016a);* ***5,11,26*** *(Ljungström et al. 2016a);* ***3,4,8,9,13–15,17,30,31*** *(Delahaie et al. 2017b);* ***10*** *(Ramakers et al. 2018);* ***12*** *(Teplitsky et al. 2011a);* ***15,32*** *(Thomson et al. 2017a);* ***18,19,35,36,38*** *(Wheelwright et al. 2014a);* ***20,37*** *(Gienapp and Merilä 2014a);* ***21*** *(Mulder et al. 2016a);* ***22–25,33,34*** *(Santure et al. 2015a);* ***27,28,39,40*** *(Bonnet et al. 2017);* ***29*** *ref. (Becker et al. 2016a);* ***41–45*** *(Evans and Sheldon 2015b);* ***46*** *(Kim et al. 2013a);* ***47–50*** *(Class and Brommer 2015a).*

A formal meta-analysis on the correlation coefficient r from each heritability–selection regression, correcting for independence of traits within studies (weighted linear mixed-effects model with random effect 'study'), reaffirmed that the overall correlation was weak and not dependent on the class of trait (Fig. 9.2). We found similar results when we disregarded non-avian traits.

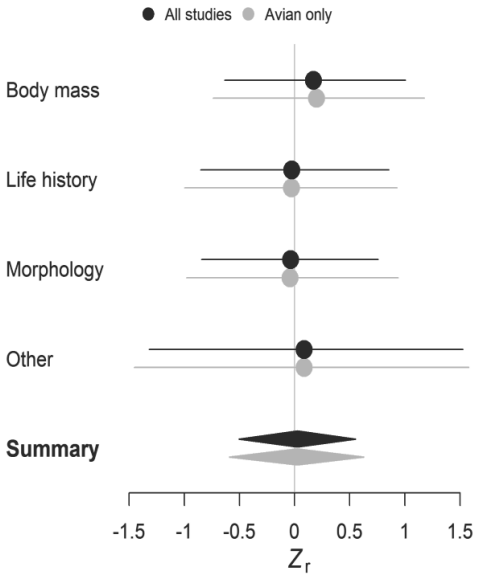


Figure 9.2. Meta-analysis on the heritability–selection correlation coefficients. *Coefficients r were standardised using Fisher's Z transformation prior to analysis. Estimates and bootstrapped 95% CIs are shown, predicted from a linear mixed-effects model and unconditioned on the random term 'study area'. The summary statistic results from a model that included only the intercept as a fixed term. Estimates from an analysis excluding non-avian traits are shown for comparison.*

Comparing expected responses to selection

Environmental coupling of (additive) genetic variance and selection can affect the predicted response to selection (Wilson et al. 2006; Husby et al. 2011b). We therefore predicted for the six datasets identified above the standardised selection response for each environment j (R'_j), assuming either constant or environment-dependent heritability ($R'_j = h^2\beta'_j$ or $h^2_j\beta'_j$). When we calculated the mean difference in response across environments between the two approaches (accounting for uncertainty in estimates), we found that environmental variation in heritability significantly affected the mean expected response in all six cases, but this effect was not in a consistent direction (i.e. either reduced in case of a negative association or increased in case of a positive association; Table 9.2). Finally, we modelled the directional difference in expected response as a function of the correlation coefficient between h^2 and β' for all datasets (cf. Wood and Brodie III 2016), and found that the difference in expected response was not affected by this correlation coefficient (slope = 0.002 [–0.001, 0.004]; Fig. 9.3).

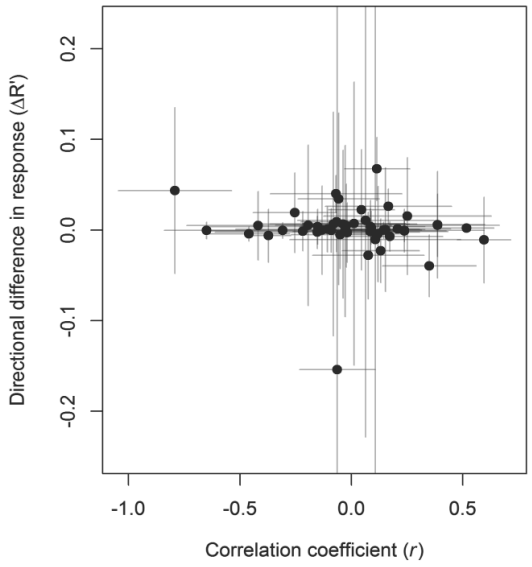


Figure 9.3. No effect of a correlation between heritability and selection on differences in selection response. *Correlation coefficients (r ± standard errors) result from WLS regressions of heritability against standardised selection gradients; $\Delta R'$ (± standard errors) is the mean, directional difference between expected responses to selection assuming varying vs. constant heritability. Each data point represents a single trait–species–population combination.*

Table 9.2. Predicted selection response assuming constant vs. environment-dependent heritability. *Predicted response (R') differed in absolute terms from year to year under the two approaches for all six cases where a correlation between heritability and selection was found; in none of these cases the difference was in a consistent direction.*

Species	Trait	$\Delta R'_{absolute}$ [95% CI]	$\Delta R'_{directional}$ [95% CI]
Cyanistes caeruleus	Adult handling aggression	0.031 [0.015, 0.049]	0.005 [–0.026, 0.031]
Parus major	Plumage refl. (at 349 nm)	0.007 [0.003, 0.012]	0.005 [–0.001, 0.011]
	Plumage refl. (at 549 nm)	0.012 [0.005, 0.021]	0.005 [–0.007, 0.016]
	Double cone plumage refl.	0.008 [0.003, 0.014]	–0.001 [–0.010, 0.007]
Passerculus sandwichensis	Nestling tarsus length	0.059 [0.038, 0.085]	0.005 [–0.034, 0.046]
	Nestling body mass	0.072 [0.046, 0.101]	–0.019 [–0.062, 0.024]

Note. R' is measured in phenotypic standard deviations. Estimates of differences were calculated using bootstrapping procedures.

Discussion

Little evidence for environmental coupling of heritability and selection

We investigated the prevalence of an environmentally induced relationship between heritability and selection across traits and study systems by using open data available in data repositories. Our study extends the limited evidence for this phenomenon (Wilson et al. 2006; Husby et al. 2011b) to 50 traits from 10 species in 16 populations. Relying on

robust statistical methods to (i) quantify the relationship between heritability and selection, (ii) synthesise results of different studies using meta-analysis, and (iii) infer expected evolutionary response, we conclude that, despite being a current topic in ecology and evolution (Hoffman and Merilä 1999; Garant et al. 2004b; Garant et al. 2005; Lédon-Rettig et al. 2014; Wood and Brodie III 2016), its evolutionary importance in natural populations is small—at least for the range of species for which we have data.

So far, only two studies have investigated this relationship within the same trait and population (Wilson et al. 2006; Husby et al. 2011b). Reanalysis of egg-laying date in the Hoge Veluwe great tit population (Husby et al. 2011b) yielded different results, potentially linked to the different approximation of the environment (see below), although the environment in that population, i.e. mean spring temperature, explains much of the variation in the trait ($r^2 = 0.66$) (Visser et al. 2006). The correlations between selection and heritability or additive genetic variance found by Husby et al. (2011b), however, were marginally or non-significant, respectively, and were not subjected to rigorous correction for uncertainty like our bootstrapping methods. Thus, even in a population with (i) a strong link between the environment (temperature) and a life-history trait (laying date) and (ii) demonstrated increases in selection and additive genetic variance under increased temperatures, evidence for an environmental link between heritability and selection was modest at best. Heritability of life-history traits is generally found to be low (Houle 1992; Lynch and Walsh 1998; Charmantier et al. 2014), potentially due to high environmental variance (Price and Schluter 1991) or genetic canalisation (Stearns and Kawecki 1994), but life-history traits are inherently likely to exhibit gene-by-environment interactions whenever selection pressures vary with the environment, because of their close association with fitness (Price and Schluter 1991). It is, then, remarkable that heritability was not related to selection in any of the life-history traits investigated here (Fig. 9.1), even though substantial variation existed in the strength of selection (Fig. S9.2).

Finding a significant relationship between heritability and selection requires sufficient statistical power. Although the number of years and individuals varied considerably between study systems (Table 9.1), significant relationships were not exclusively found in the largest datasets (Fig. 9.1). A visual inspection of the components that make up this relationship, as well as the relationship between selection and the environment, suggested that statistical significance was neither influenced by the variance in the predictor and response variables nor by the number of years or the total number of observations available (Fig. S9.3). Given the larger statistical power associated with larger datasets, the lack of the sought correlation in our largest datasets suggests that the effect size is likely too small to be biologically meaningful.

Using an analytical model informed by data from a literature review, Wood and Brodie III (2016) predicted that the strength of the relationship between selection and genetic variance would impact the mean and, to a greater degree, variance in responses to selection across hypothetical populations. Yet even in the few cases in which we demonstrated a reasonably strong relationship between heritability and selection (Fig. 9.1; cf. Wilson et al. 2006), this was not sufficient to fuel a change in the rate of expected response to selection (Table 9.2). This is partly because both components of the

relationship came with prediction error that needed to be accommodated in the estimation of the response. From the studies investigated here, we therefore conclude that even when we find environmental coupling between heritability and selection, its net effect on the predicted evolutionary change is small and is hence an unlikely explanation for potential discrepancies between observed and expected responses to selection in natural populations (Merilä et al. 2001b).

Methodological considerations

An important aspect in analysis of genotype-by-environment interactions, i.e. testing whether genetic variance and heritability differ among environments, is the choice of the environmental variable. However, in most of the analysed datasets no such environmental variable was included. Instead of obtaining such data from other sources and testing whether the chosen variable was predictive for the trait in question, we used environment-specific, population-mean trait values as the environmental variable (covariate) in our analyses, an accepted practice in animal and plant breeding (Lynch and Walsh 1998; James 2009). This approach has three major advantages. First, the daunting task of searching environmental data relevant to each trait becomes unnecessary. Second, it enables the inclusion of traits for which it is difficult to conceive and collect environmental data (compare, for example, breeding time in great tits, which is strongly temperature dependent (Visser et al. 2009a), with a physiological trait like handling aggression in blue tits *C. caeruleus*, for which no clear environmental component has been identified, despite substantial year-to-year and residual variation (Class and Brommer 2015b)). Third, because the population-mean phenotype encompasses all unmeasured or unobserved components of the environment, it will generally be an accurate representation of the environment for the trait of interest (Lynch and Walsh 1998), circumventing the problem of misidentifying the relevant environmental component and, consequently, erroneously inferring the presence or absence of variation in reaction norm slopes. For example, in a population of collared flycatchers (*Ficedula albicollis*), Brommer et al. (2005) found significant between-individual variation in breeding-time reaction norms in response to average temperatures in spring, but not to rainfall or North Atlantic Oscillation, even though these variables correlated well with breeding time. Similarly, Husby et al. (2010) could show between-individual variation in reaction norms for breeding time in great tits while Charmantier et al. (2008) did not find this in the same population when using a different environmental variable. Indeed, simulations have shown that random regression models with 'mean trait' as the environment yielded similar variation in reaction norm slopes to models with a 'real' environmental driver of the trait (Gienapp 2018). In contrast, using other environmental measures that did not drive the trait but correlated with the 'real' environment to a decreasing degree (r = 0.9 to 0.1) yielded increasingly downwardly biased estimates of both the slope and the variance in the reaction norm. This is an important finding because it shows that environment-specific mean phenotypes can serve as a 'yardstick' when testing for gene-by-environment interactions (Gienapp 2018).

Ideally, heritability should be estimated at the same level as where selection operates, because the correlation at this level is what ultimately matters. Since selection is generally

estimated at an annual level (where each year captures all components of the environment), heritability should be estimated at this level too. This would, however, require an enormous number of individuals in each year to estimate the annual genetic variances reliably—and hence generally not be feasible. Using (continuous) environmental covariates instead to estimate genetic variance along an environmental gradient (Schaeffer 2004; Nussey et al. 2007) is the next best option, and the best way to do this is to choose a metric that captures most features of the environment in a given year (which annual mean phenotypes do). This alleviates the need to establish a link between an environmental covariate and selection, which will not necessarily be informative when investigating the correlation between heritability and selection—in particular when statistical power is limited.

A concern when estimating selection in natural populations is to identify the real target of selection (Lande and Arnold 1983; Hereford et al. 2004; Hadfield 2008). The use of the Breeders' Equation to predict evolutionary change in natural populations has therefore been advised against, and the Robertson-Price identity has been suggested as an appropriate alternative (Hadfield 2008; Morrissey et al. 2010). However, estimating the genetic covariance between a trait and fitness at an annual basis to estimate variation in selection is rarely, if ever, possible, due to the large datasets required to reliably estimate genetic covariances. Furthermore, Reed et al. (2016b) showed that in a wild population of great tits, environmental bias in phenotypic selection estimates for egg-laying date and clutch size is small at best. A similar conclusion was reached by Morrissey and Ferguson (2011), who showed for brook charr (*Salvelinus fontinalis*) that estimates of phenotypic selection on body size are highly congruent with estimates of genetic selection.

Benefits and limitations of open data

One important development in ecology and evolution in recent years has been the requirement to make the data used to produce the results of studies (usually published studies) publicly available (Whitlock et al. 2010; Mills et al. 2015), leading to a surge in data output onto online data repositories. The potential advantages of open data archiving in revolutionising the natural sciences are now increasingly recognised (Hampton et al. 2013; Culina et al. 2018). Yet Evans (2016) showed that data from long-term population studies archived in Dryad Digital Repository are never used by third parties. Our multi-study approach makes extensive use of such long-term data to address an outstanding question in evolutionary ecology. Indeed, the use of open data comes with important logistical and ethical issues (Mills et al. 2015, 2016; Whitlock et al. 2016) that need to be addressed before biological conclusions can be safely drawn. Our study, however, shows that it can be done successfully (see also Chapter 7).

From the 106 initially considered datasets in our example, we could eventually use only 14 (plus the previously unpublished pied flycatcher dataset), due to various reasons such as small and/or biased sample sizes, a lack of appropriate fitness data, unusable pedigrees (e.g. relatedness matrices, which we were unable to use after data manipulation because they required a specific ordering of the individuals in the phenotype file), and a low number of years. Moreover, the data were heavily biased towards birds and mammals (50

and 31 datasets, respectively). We therefore need to make the cautionary note that we cannot necessarily extrapolate the evolutionary importance of an environmental correlation between selection and genetic variance across a wider range of taxa. The general taxon bias in quantitative genetic studies of wild populations toward birds and mammals can be explained by the fact that linking individual offspring to their parents, necessary to construct a pedigree, is comparably straightforward (Clutton-Brock and Sheldon 2010). Relatedness matrices based on genomic markers may make pairwise relatedness estimates a less stringent requirement in evolutionary studies in the future and in that way greatly augment the taxonomic scale at which important evolutionary questions can be addressed (Gienapp et al. 2017). Time will resolve issues like samples sizes and number of years, but whether or not a dataset is suitable will ultimately depend on the type of analysis and the type of data required. In the era of Open Science that encourages publication of datasets while increasing their quality, it is but a matter of time before taxon biases in multi-annual meta-studies similar to ours may dissipate. Such long-term datasets of individually marked animals are invaluable tools in ecology and evolution and will inevitably serve to elucidate the ecological and evolutionary consequences of environmental change (Visser 2008; Clutton-Brock and Sheldon 2010).

Methods

Data acquisition

In May and July 2016 we conducted a search for datasets that contained pedigree information on a wild species through twelve different aggregators of research data repositories (Europe PMC, DataCite, BASE, OpenAIRE, Science Research, DataOne Mercury search, Web of Science Data Citation Index, Scielo, Research Data Australia, DLI Service, Dryad Digital Repository, DataMED). These aggregators collect information on datasets (e.g. title, keywords, abstract and description) that have been deposited in different data repositories, and allow for search through multiple data sources in one search interface. Datasets were tracked using fixed search terms (see Supplementary Methods SM9.1); search results were screened based on title, abstract, dataset description, and/or keywords, if available. Remaining datasets were further checked for relevance by opening the data files and/or reading the related publication if necessary, leaving only datasets containing pedigree information for a wild or captive animal population. Recording of datasets was done according to PRISMA guidelines (Moher et al. 2009; Chapter 7).

Next, we screened and filtered this data subset (103 datasets) to keep those where: (i) the pedigree file could potentially be used (i.e. when the file was not embargoed, corrupted or otherwise unsuitable for our particular analysis, e.g. relatedness matrices lacking the specific links between parents and offspring); (ii) the pedigree contained a sufficient number of individuals (final datasets had, on average, >40 observations/individuals per year); (iii) individuals in the pedigree also had information on a phenotype on which selection could act; (iv) there was natural environmental

variation in the phenotype (this excluded all laboratory populations); (v) the associated phenotype file contained at least six years of data; and (vi) there were no additional issues (e.g. non-matching IDs of animals in pedigree and phenotype file). In addition to these 103 datasets, we did an additional search in Web of Science (on 9 September 2017; see Supplementary Methods SM9.1) and from the resulting 396 studies, we discovered three additional suitable datasets overlooked by the initial search, using the inclusion criteria above. Lastly, we added our own, previously unpublished data from the long-term study of pied flycatchers (*Ficedula hypoleuca* (Ramakers et al. 2018); see Visser et al. (2015) for more information on that population), totalling 107 retrieved datasets (Table S9.1).

The total number of datasets included in the analysis amounted to 15, covering 10 species from 16 populations and a variety of life-history, morphological, physiological, behavioural and body mass traits (Table 9.1). This excludes datasets that initially appeared suitable to us but whose suitability for our analysis was refuted by the original authors (see 'Enquiring with original authors'; Table 9.1).

Quantifying the environment

None but two of the final datasets provided information about the environment. Therefore, we used a standardised protocol to quantify the environment. For each year, we calculated the population-mean trait value ($\bar{x}$) as a measure of the general environment and mean- and variance-standardized it across seasons/sites:

$$E'_j = \frac{\bar{x}_j - \mu_{\bar{x}}}{\sigma_{\bar{x}}},$$

where j denotes the j^{th} season, and μ and σ the grand mean and standard deviation, respectively. Note that this measure does not identify any specific environmental parameter but captures the environment as a whole in a specific season. The method is commonly used in animal and plant breeding studies in a process called 'joint-regression analysis', where genotype-specific interactions are partitioned into a component explained by mean population performance and a residual component (Lynch and Walsh 1998, pp. 672–678). It was first proposed by Yates and Cochran (1938) and later brought into prominence in a barley yield experiment by Finlay and Wilkinson (1963), and has now become widely accepted in the plant- and animal-breeding literature (Lynch and Walsh 1998; James 2009). It has the advantage that all of the complex (and potentially unobserved) features of the environment are integrated into a single measure, allowing for the ranking of seasons in terms of overall environmental quality. Note that this method disqualifies traits that do not vary at the annual level (i.e. fixed adult traits were not used in our analyses).

One complication with our measure of the environment is that such a measure is potentially biased when a non-random portion of the population in a given season is removed from the dataset (e.g. because certain individuals are never sampled), or when changes in the demographic structure of the population strongly affect the mean trait

value. When this was the case (see 'Enquiring with original authors'; Table S9.1), the dataset was dropped from further analysis.

Standard trait heritability

For each of the traits in our full data (Table 1), we tested for evidence of additive genetic variance following a standardized protocol. First, we constructed 'minimum adequate' mixed-effects models (MAMs) with the trait of interest as response variable (all with Gaussian errors) using restricted maximum likelihood (REML) estimation in ASReml-R (Butler et al. 2009; Gilmour et al. 2009). This method provides a fast and efficient away of estimating variance components and allows for the inclusion of additive genetic effects. Note, however, that we used a Bayesian approach to estimate environment-dependent heritability estimates, as this allows for estimation of posterior confidence regions, which we needed to reliably account for uncertainty in our environment-dependent heritability estimates in subsequent analysis (see 'Genotype-by-environment analysis'). Fixed effects were the environment (E'), as continuous variable, and additional effects provided in the dataset, based on mixed-effects models in the associated original paper. Significance of these effects, as well as that of interactions between effects, was tested with conditional Wald F tests, removing non-significant ($p > 0.05$) terms in a backward stepwise manner (but always keeping E'). Random effects were those identified in the original papers (always containing a 'permanent environment' effect, i.e. individual ID, when there were multiple observations of the same individual), but sometimes we constructed our own additional effects when deemed biologically appropriate (e.g. in nestling traits, 'nest-box ID' and 'year' were combined to identify common-environment effects within a single brood). Significance of random effects was tested using likelihood-ratio tests ($D = 2[\log(L_{m1}) - \log(L_{m0})]$, where D is asymptotically χ^2 distributed with one degree of freedom). The MAM was extended to an 'animal model' (Henderson 1988; Kruuk 2004) (AM) by adding a random additive genetic effect based on the pedigree with maternal and paternal links (see references in Table 1 for how pedigrees were constructed). Thus, the AMs took the form

$$\boldsymbol{y} = \boldsymbol{X_1\beta}_{E'} + \cdots + \boldsymbol{X_n\beta} + \boldsymbol{Z_1pe} + \boldsymbol{Z_2a} + \cdots + \boldsymbol{Z_n}u + \varepsilon,$$

where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{X_{1\ldots n}}$ and $\boldsymbol{Z_{1\ldots n}}$ are the design matrices relating the fixed ($\boldsymbol{\beta}$) and random effects ($\boldsymbol{pe}$, permanent environment; $\boldsymbol{a}$, additive genetic; u, other) to $\boldsymbol{y}$ and ε is the error term. The narrow-sense heritability was calculated as $h^2 = \sigma^2_{\boldsymbol{a}}/\sigma^2_P$, where σ^2_P represents the total phenotypic variance comprising all variance components, conditioned on the fixed effects (Table 9.1).

Genotype-by-environment analysis

To model the interaction between additive genetic variance and the environment (G×E), we extended the AM to a random regression animal model (RRAM) using the

'MCMCglmm' package (Hadfield 2010; Hadfield 2018) (ignoring years with <8 observations). In the RRAMs, we allowed the environment to interact with both the permanent environment (if present) and the additive genetic effect:

$$\boldsymbol{y} = \boldsymbol{X_1\beta}_{E'} + \cdots + \boldsymbol{X_n\beta} + \boldsymbol{Z_1}(\boldsymbol{pe}, E', n_1) + \boldsymbol{Z_2}(\boldsymbol{a}, E', n_1) + \cdots + \boldsymbol{Z_n}u + \varepsilon,$$

where n_1 is the first-order polynomial of the regression function. Fixed and random terms were those identified from the (M)AMs; note that because E' explains most of the variation related to seasonal effects, it replaced the random effect of year in most analyses. We constructed two 2×2 unstructured variance–covariance matrices for the intercept and the slope of the permanent environment and the additive genetic effect:

$$\mathbf{P} = \begin{bmatrix} \sigma^2_{\boldsymbol{pe}_{int}} & \sigma_{\boldsymbol{pe}_{int},\boldsymbol{pe}_{sl}} \\ \sigma_{\boldsymbol{pe}_{sl},\boldsymbol{pe}_{int}} & \sigma^2_{\boldsymbol{pe}_{sl}} \end{bmatrix} \text{ and } \mathbf{G} = \begin{bmatrix} \sigma^2_{\boldsymbol{a}_{int}} & \sigma_{\boldsymbol{a}_{int},\boldsymbol{a}_{sl}} \\ \sigma_{\boldsymbol{a}_{sl},\boldsymbol{a}_{int}} & \sigma^2_{\boldsymbol{a}_{sl}} \end{bmatrix}.$$

In cases where there was no permanent-environment effect but only a maternal or common-environment effect (in juvenile-only traits), only the **G** matrix was fitted. To avoid artificial inflation of slope variance estimates in the **P** and **G** matrices due to heterogeneity in residual variance across the environmental gradient, we partitioned the residual component ε into 'environmental blocks'(Lillehammer et al. 2009), following categorisation of environments into n equal-interval groups. Thus, we fitted the residual matrix as an $n \times n$ matrix with independent variances,

$$\mathbf{R} = \begin{bmatrix} \sigma^2_{\boldsymbol{\varepsilon}_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{\boldsymbol{\varepsilon}_n} \end{bmatrix},$$

where n was the number of environments divided by 5, but was always ≥3 (e.g. in a dataset with 20 environments $n = 4$, but with 10 environments $n = 3$).

To illustrate that our environmental metric (E') was valid in this context, Gienapp (2018) ran random regression models on simulated data using several different quantifications of the environment related to a 'true' environmental driver of the phenotype, as well as the annual trait mean. He found no evidence that variance estimates of reaction norm intercepts and slopes were biased by the annual trait mean (relative to the 'true' environmental driver) and showed that this metric outperformed environmental correlates. Although we concur that additive genetic variance may not only be affected by current environmental conditions but also be the outcome of past selection processes, it is evident from many quantitative genetic studies of wild populations that year-to-year variation in phenotypes is mostly attributable to phenotypic plasticity and that the share of genetic change from year to year is generally very small and undetectable (Gienapp et

al. 2008; Merilä and Hendry 2014). Consequently, we believe that using environment-specific mean trait values will lead to more reliable results than an environmental variable that correlates too weakly with the real driver of plasticity, thereby underestimating variation in (genetic) reaction norm slopes.

To obtain independent samples in the MCMC sampling process, we used a thinning interval of 20,000 in all models, with a burn-in period of 200,000 samples and a total effective sample size of 250 (i.e. 5,200,000 samples). In exploratory stages of the analysis, we found that a larger effective sample size (1000) did not affect the posterior estimates, but these models take substantially longer to complete. Effective sample size in all models included never fell substantially below 250 and autocorrelation between samples was almost always <0.1 but never exceeded 0.2 for any variance component; models that did not meet these criteria were discarded (not listed in Table 1). For the residual term, we specified Inverse-Wishart (IW) priors (V = diag(x) and nu = 1.002, where x is the dimension of the matrix). For the random terms we explored two alternative priors: the IW prior (specifications as above) and parameter-expanded (PE) priors (V = diag(x), nu = x, alpha.mu = 0, alpha.V = diag(x)*500). Although both priors yielded similar results in most cases, the posterior variances tended to be smaller when real variance was close to zero under the PE compared to the IW prior. This is in agreement with previously voiced concerns that the IW prior may behave poorly when true variance is close to zero (Gelman 2006; Schuurman et al. 2016; Hadfield 2018). We therefore only present posterior estimates from the models based on PE priors. We refrained from 'significance' testing of the G×E interaction, because—issues concerning model-selection criteria such as DIC aside (Spiegelhalter et al. 2002; Millar 2009; Hadfield 2018)—of potentially limited power in the smaller datasets and our main interest in testing the covariance between h^2 and selection. We instead opted for a pragmatic approach and used the highest posterior density intervals (HPDIs) to account for uncertainty in all subsequent analyses. The rationale behind this was that if we had excluded all 'non-significant' G×E interactions, of which some may have been false negatives, we may have overlooked a potentially strong covariance between h^2 and selection (see below). By accounting for the uncertainties in environment-specific h^2 estimates, the true negatives in G×E will not lead to a spurious covariance between h^2 and selection.

The posterior mean variance for each variance component in environment j was derived from the estimated **G** and **P** matrix as (De Jong 1990)

$$\sigma_j^2 = \sigma_{int}^2 + 2\sigma_{int,sl}\boldsymbol{\beta}_{E'_j} + \sigma_{sl}^2\boldsymbol{\beta}_{E'_j}^2.$$

The 95% HPDIs were likewise derived from the upper and lower HPDI matrices. Environment-dependent heritability was defined as the mean of posterior variance estimates, $h_j^2 = \sigma_{\boldsymbol{a}_j}^2/(\sigma_{\boldsymbol{a}_j}^2 + \sigma_{\boldsymbol{pe}_j}^2 + \cdots + \sigma_{\boldsymbol{\varepsilon}_j}^2)$, with 95% HPDIs estimated from the lower and upper HPD limits of each variance component. Standard errors of h_j^2 were then calculated as half the 95% HPDI divided by 1.96.

Selection

To quantify selection on the trait in a given environment, we made use of provided reproductive fitness data (number of offspring or recruits) or survival data (Table 9.1). When such data were not provided, we inferred (annual) reproductive success by linking animals to sires and dams in the pedigree using their birth year (when available). If we could not infer annual recruits from the pedigree, we determined survival from one year to the next by identifying reappearance of individuals in the dataset in subsequent years, assuming the last year of appearance was the last year the individual was alive. As with quantifying the environment (see above), inferring fitness is problematic if a non-random portion of the population appears in the dataset (aside from the non-random disappearance due to selection; see also Hadfield 2008). When this was likely to be problematic (see 'Enquiring with original authors'; Table S9.1) the dataset in question forewent inclusion in the analysis.

To estimate annual, standardised selection gradients (β'), we constructed general(ised) additive models (GAMs, package 'mgcv' (Wood 2017)), where the fitness component was the response variable following either a Gaussian, Poisson or negative binomial distribution for fecundity measures (number of offspring produced or recruits), depending on the distribution of the data, or a binomial distribution for survival (1/0 response). As fixed effects, we initially included an interaction between year and the trait of interest and used it as a null model to identify additional significant fixed effects (using F or χ^2 tests) that influenced the fitness measure (e.g. age or sex and additional quantitative traits). Based on these findings, we ran annual GAMs (without 'year') and calculated annual β' using the 'gam.gradients' function from the 'gsg' package (Morrissey and Sakrejda 2015). This procedure estimates β's as

$$\beta' = \frac{Cov(w,z)}{\sigma_z},$$

where the numerator is the covariance between the trait and relative fitness, i.e. the partial regression coefficient after taking into account the effect of traits potentially simultaneously under selection, and the denominator is the standard deviation of the trait, following Lande and Arnold (1983; Morrissey and Sakrejda 2013). Standard errors of β' were estimated through parametric bootstrapping (1000 iterations).

Covariance between selection and heritability: a meta-analysis

As we were interested in studying the effect of environmental variation in selection and genetic variance on selection response, we examined the (linear) relationship between heritability and selection. We refrained from making this analysis conditional on the presence of an underlying correlation between β' and E', because in cases where statistical power may be an issue, such a two-step approach would decrease the likelihood of detecting a real relationship between h^2 and β' if datasets were omitted based on this criterion. A similar reasoning applied to testing for an underlying relationship between h^2

and E' (see above). For each dataset, we regressed h^2 against β' in linear weighted least-squares (WLS) regressions, weighting data points by 1/[(standard error of h^2)^2]. To account for uncertainty in the predictor, β', we substituted each of its values (j) with a randomly drawn value from a random normal distribution (n = 1000, $\mu = \beta'_j$ and σ = standard error of β'_j) and iterated the entire process 1000 times. We obtained the mean and the 0.025 and 0.975 quantiles (i.e. the 95% bootstrapped confidence interval CI) of the model estimates (intercepts and slopes) resulting from these iterations; estimates were considered statistically significant if the 95% CI did not include 0. Note that in reality, estimates of β' are not entirely independent because some individuals are included in multiple estimates, potentially affecting the estimates from (W)LS regression models. We believe, however, that this issue was sufficiently accounted for by our pragmatic bootstrapping approach.

When estimating the covariance between selection and heritability we took the sign of the estimated selection gradients into account, i.e. we did not correlate heritability with the *absolute* strength of selection (cf. Wood and Brodie III 2016). The rationale was that (1) it is biologically relevant whether there is selection for larger or smaller trait values and (2) using absolute or signed selection gradients has different implications for evolutionary change. If a correlation between absolute strength of selection and heritability exists, the overall selection response will not be altered because episodes of strong selection in either direction are always coupled with either high or low heritability. This is, however, not the case when signed selection estimates are used, because in this case strong selection in one direction is coupled with low heritability, whereas strong selection in the other direction is coupled with high heritability.

To examine the overall correlation coefficient across studies and trait types, we performed a meta-analysis using the mean correlation coefficients (r) and their standard errors (SE_r, i.e. half the 95% CI divided by 1.96) resulting from each bootstrapped regression model. Following Nakagawa and Cuthill (2007), we transformed coefficients prior to meta-analysis to Fisher's Z,

$$Z_r = 0.5 \times \ln\left(\frac{1+r}{1-r}\right).$$

Variance in Z_r was calculated as (Niemelä and Dingemanse 2018)

$$\sigma^2_{Z_r} = {SE_r}^2 \times \left(\frac{1}{[1+r]\times[1-r]}\right)^2.$$

We estimated the (weighted) mean correlation coefficient (n = 50) in a linear mixed-effects model (REML, package 'lme4' (Bates et al. 2018)) with trait type (life history, body mass, morphology, or other) as a fixed effect, study area (i.e. by species; n = 16) as a random effect, and $1/\sigma^2_{Z_r}$ as weights. We initially included a random effect of species, which explained 0 variance and was therefore removed from the model (note that the bias toward passerine birds in the acquired datasets precluded phylogenetic analysis). Mean Z_r and

95% CI, predictions unconditioned on the random term, were calculated for each trait type and from a null model excluding the fixed term (i.e. intercept only) through bootstrapping with 1000 iterations. The procedure was repeated on a subset of the data that excluded non-avian traits (n = 43 coefficients, 14 studies). To quantify the consistency among studies, we estimated for both sets of analysis (all data or avian-only) the heterogeneity (I^2, the proportion of variance that cannot be explained by chance) in the random-effects components for the random-only models (see Nakagawa and Santos 2012 for details). Residual variance in Z_r was estimated at 1.55 and 1.76, respectively, whereas 'study' variance was 0.002 in both cases. Error variance (σ_m^2 in Nakagawa and Santos (2012)) was small (0.016 and 0.014, respectively) and I^2 was estimated at 0.99 in both cases.

Expected response to selection

To quantify the consequence of a covariance between h^2 and β' on the response to selection, we predicted the absolute response to selection under the assumption of constant *vs* varying heritability following the Breeder's Equation (Lande and Arnold 1983; Falconer and Mackay 1996), i.e. $R'_j = R_j{\sigma_z}^{-1} = h^2\beta'_j$ *vs.* $h_j^2\beta'_j$. Note that the expected response is in units standard deviation (Lande and Arnold 1983; Hereford et al. 2004) (hence σ_z), indicated by the apostrophe. The standard error for R'_j was derived by adding up the relative standard errors of h^2 (or h_j^2) and β'_j. We then calculated the mean absolute (1) and directional (2) difference in response between the two approximations ($\Delta R'$), with the assumption that non-constant heritability does affect the response from any one year to the next (1) and that this difference is directional (2), i.e. positive when the correlation between h^2 and β' is positive and vice versa (Wilson et al. 2006; Husby et al. 2011b). We estimated mean $\Delta R'$ across seasons in a linear model without an intercept and with a fixed effect of 'study'. As a response variable, $\Delta R'$ in each environment (j) was determined as the difference between two randomly drawn (absolute) values for R'_j from two random normal distributions (n = 1000, $\mu = R'_j$ and σ = standard error of R'_j). Mean $\Delta R'$ was derived as the mean, study-specific intercept from 1000 iterations, along with the 0.025 and 0.975 quantiles (i.e. 95% CI).

Loosely based on Wood and Brodie III (2016), we estimated whether the strength of the relationship between heritability and selection affected expected (difference in) selection response. We repeated the procedure above for all the datasets (except those for which $h^2 = 0$) and calculated the expected, mean directional difference (± standard error) in expected response to selection assuming varying *vs.* constant heritability ($\Delta R'$). We also extracted the correlation coefficients, r, along with their 95% CIs, from each WLS regression model described in the previous section and calculated standard errors of r as half the 95% CI divided by 1.96. We ran a WLS regression model with $\Delta R'$ as a response variable and 1/[(standard error of $\Delta R'$)^2] as weights. The correlation coefficient r was the predictor, randomly drawn from a random normal distribution (n = 1000, $\mu = r$ and σ = standard error of r); the procedure was iterated 1000 times and mean estimates and the 0.025 and 0.975 quantiles (95% CI) were extracted. We also tested this relationship with 'study area' as a random effect in a linear mixed-effects model, but found that this factor explained 0 variance.

Enquiring with original authors

A potential danger of using open data is that the investigator may not be familiar with the study system and therefore make false assumptions about the data (Mills et al. 2015, 2016; Whitlock et al. 2016). Hence, for every dataset potentially suitable for analysis, we wrote a letter to the leading author and/or principal investigator of the associated paper, informing them about the general project aim, as well as a description with specifics regarding the use of their dataset (see Supplementary Methods SM9.2). The description contained information about which data files we used, what our study aim was using their datasets, how we went about preparing the data for analysis (e.g. combining multiple files, (re)construction of the pedigree, calculation of the environment based on the population-mean trait value, identification of reproductive performance or survival), how we analysed the data (including which variables we included in the (M)AMs and RRAMs) and a brief overview of tentative findings. We were specifically interested in the authors' verdict on our quantification of the environment and fitness. All analyses presented here are based on datasets that were deemed 'appropriately used' by the original authors. A common concern with discarded datasets was that reproductive success or survival could not be reliably inferred, for example because a non-random portion of recruits disperse away from the study area, or because surviving individuals were not included in the dataset because they had no phenotype. Similarly, non-random dropping of individuals was likely to affect the estimation of the environment (E'), in which case the dataset forewent inclusion in the analysis. We refer the reader to Table S9.1 for a full list of considered datasets and the reason for their exclusion. We report on the author correspondence in more detail in Chapter 7.

Data availability and code availability

Raw data used in the analyses can be found in the references listed in Table 9.1; DOIs for each dataset can be found in Table S9.1 (online). Data used for the weighted regression analysis, estimating predicted response to selection, and meta-analysis can be found in Tables S9.2 and S9.3 (online). R code examples for each analysis are available as a supplementary text file (online).

Acknowledgements

We are grateful to all the original data owners who found time in their tight schedules to reply to our e-mails, or who otherwise contributed to the discussion (these include people whose data we did not end up using for our analyses). Specifically, we thank—in no particular order—Ben Sheldon, Arild Husby, Erik Postma, Brigitte Schloegl, Katharina Foerster, Juan Fargallo, Jon Brommer, Barbara Class, Anders Møller, Gabriella Ljüngstrom, Simon Blanchet, Nat Wheelwright, Marion Nicolaus, Niels Dingemanse, Scott Sakaluk, Alastair Wilson, Jarrod Hadfield, Joey Hubbard, Jane Reid, Peter Arcese, Elise

Huchard, Tim Clutton-Brock, Josephine Pemberton, Susan Johnston, Timothée Bonnet, and Boris Delahaie. Shinichi Nakagawa kindly advised us on the meta-analysis. Thomas Reed, Arild Husby and three anonymous reviewers provided useful comments on our manuscript. This work was funded in part by NIOO Strategic funds (SM1521) and an ERC Advanced Grant (339092 - E-Response) to M.E.V.

Supplementary Information S9

Contents:

Supplementary Tables (not presented here)

These tables can be found in the online version of this paper at:

https://doi.org/10.1038/s41559-018-0577-4

- Table S9.1.1. This table contains all considered datasets that met initial screening criteria. It lists 107 datasets: 103 found during the initial search using aggregators of research data repositories; 3 found during an additional search on Web of Science; and 1 (previously unpublished) dataset from our own database.
- Table S9.1.2. Data repositories associated with each dataset in Table S1.1.
- Table S9.2. Data necessary to reproduce Figure 1 in the main text (heritability–selection regressions) and to calculate expected response to selection.
- Table S9.3. Data necessary to replicate the meta-analysis on the correlation coefficient of the relationship between heritability and selection.

Supplementary Methods (SM) and Figures

This file describes the methods used to acquire the datasets used in the analyses (1) and the correspondence with the original owners of the data to ensure validity of our analyses (2), as well as the supplementary figures referenced in the main text (3).

1. Acquiring data
2. Correspondence with the authors of the datasets
3. Figures

Supplementary Codes (not presented here)

The code can be found in the online version of this paper at:

https://doi.org/10.1038/s41559-018-0577-4

- This text file contains all codes necessary to replicate the analyses in R. Note that Supplementary Code (SC) 1–4 provide examples for a specific trait from a specific dataset; SC5–7 can be used to run all analyses using the data in Table S9.2 and S9.3.

SM9.1: Acquiring data

General methods, search criteria, and dataset selection process

We conducted a search for datasets that contained pedigree information on a wild species (either wild or captive population) through 12 different aggregators of research data repositories and Dryad Digital Repository (http://datadryad.org/) (between May and July 2016): Europe PMC (http://europepmc.org/); DataCite (https://search.datacite.org/); BASE (https://www.base-search.net/); OpenAIRE (https://www.openaire.eu/search/find/); Science Research (http://scienceresearch.com/scienceresearch/); DataOne Mercury search (https://cn.dataone.org/onemercury/); Web of Science Data Citation Index (http://apps.webofknowledge.com); Scielo (http://www.scielo.org/php/index.php); Research Data Australia (https://researchdata.ands.org.au/); DLI Service (https://dliservice.research-infrastructures.eu/index.html#/); and Data MED (https://datamed.org/). These aggregators collect information on datasets (e.g. title, keywords, abstract, description) that are deposited in different data repositories, and allow for search through multiple data sources in a single search interface.

Our inclusion terms were ["pedigree" OR "relatedness matrix"]. Based on the results of this initial search, we decided to use several exclusion terms in a refined search: "dog food", cultivar*, "family tree", "family pedigree", "middle age", "middle aged", nation, child*, medicine, medical, adolescent, adolescence, autism, diagnosis, "family health", "risk factor*", patient, pancer, wheat, schizophrenia, poplar, maize, "mental health", soya bean, soya beans (all connected with OR). These were the most common terms that appeared in the description of datasets that relate to plants and to humans (e.g. mental health, autism, diagnosis, maize, soya beans). Because the functionality of search differs between different aggregators, we adjusted search terms and search syntax accordingly (see below). If the aggregator allowed for results filtering (e.g. based on the scientific area) we used this option to filter out the irrelevant datasets from the initial list of datasets. Datasets were then screened based on Title, Abstract, Dataset description, and/or Keywords, if available (different aggregators provide all or some of these). Remaining datasets were then further checked for the relevance by opening the data files and/or reading the related publication if necessary, leaving only datasets containing pedigree information for an animal population (wild or captive).

Next, we screened and filtered this data subset (106 datasets, Supplementary Table 1) to keep those where:

(i) the pedigree file could potentially be used (i.e. when the file was not embargoed, corrupted or 'encrypted');

(ii) the pedigree contained a sufficient number of individuals (final datasets had, on average, >40 observations/individuals per year);

(iii) individuals in the pedigree also had information on a phenotype on which selection could act;

(iv) there was potential, natural environmental variation in the phenotype (this excluded all laboratory populations);

(v) the associated phenotype file contained at least six years of data;

(vi) there were no additional issues (e.g. non-matching IDs of animals in pedigree and phenotype file).

In addition to these datasets, we did an additional search in Web of Science (on 9 September 2017) using the following key words: "("animal model" OR "quantitative genetic*" OR pedigree*) AND ("natural population*" OR "wild population*") NOT (plant OR experiment)", disregarding studies published before 2010 (because data publication was not a standard journal policy before that time, and none of the previously retrieved datasets were from before that year). From the resulting 396 studies, we discovered three additional suitable datasets overlooked by the initial search, using the inclusion criteria above.

Overall, we located 106 datasets containing animal pedigree/phenotype data (Table S9.1, excluding Pied flycatcher *Ficedula hypoleuca*, which makes 107 datasets). After we applied the above screening, we ended up with the 18 datasets for the final analysis (excluding the Pied flycatcher dataset; see main text), 4 of which were omitted after correspondence with the original authors. The reasons for the exclusion of the pedigreed dataset from the final analysis are given in Table S9.1.

Details of the search (exact search terms, date of search), and search result screening process for each data aggregator

1. **Europe PMC**

http://europepmc.org/

Europe PMC is primary aggregator of scientific journals and their publications. However, it provides links to datasets related to these publications. Thus we used our main search syntax to search the full text of publications, and added that only articles with appendix, supplement, or tables that contained the words 'pedigree' or 'relatedness matrix' be included.

Date searched: 24-05-2016

Used the 'Advanced search' option (with synonyms on)

Search syntax:

All Bibliographic Fields ("pedigree*" OR "relatedness matrix" NOT "Cultivar* " NOT "family tree" NOT "family pedigree" NOT "dog food" NOT "Nation" NOT "Child" NOT "Medicine" NOT "medical" NOT "Adolescent" NOT "Autism" NOT "family health" NOT "risk factors" NOT "patient" NOT "diagnosis" NOT "cancer" NOT "clinical" NOT "wheat" NOT "schizophrenia" NOT "poplar"

NOT "maize") Article Sections: AND (APPENDIX:"pedigree*" OR SUPPL:"pedigree*" OR TABLE:"pedigree*" OR APPENDIX:"relatedness matrix" OR SUPPL:"relatedness matrix" OR TABLE:"relatedness matrix")

Initial results: **304**

Screened by title: leaves **64**

Screened by abstract: leaves **34**

Further check for relevance: leaves **6** pedigree datasets

Used in the final analysis: **1**

2. **DataCite**

https://search.datacite.org/

Date searched: 24-05-2016

Used the 'Advanced search' option

Search syntax:

Search in all fields: (pedigree*" OR "relatedness matrix") -Cultivar* -"family tree" -"family pedigree" -"dog food" -Nation -Child -Medicine -medical -Adolescent -Autism -"family health" -"risk factors" -patient -diagnosis -cancer -clinical –wheat –schizophrenia –poplar -maize

Initial results: **409**

Screened by title: leaves **148**

Screened by abstract: leaves **101**

- Some of these belonged to the same data package, and we combined these to obtain final **64** datasets

Further check for relevance: leaves **59** pedigree datasets

Used in the final analysis: **7**

3. **BASE**

https://www.base-search.net/

Date searched: 7-06-2016

Search in all fields the search syntax:

(pedigree "relatedness matrix") -(cultivar* nation child medicine medical adolescent autism patient diagnosis cancer clinical wheat schizophrenia poplar maize "family tree" "family pedigree" "dog food" "risk factors") with document type 'Primary data'

Initial results: **355**
Remove duplicated results: leaves **239**
Screened by title: leaves **173**
Screened by abstract: leaves **144**

- Some of these belonged to the same data package, and we combined these to obtain final **91** datasets

Further check for relevance: leaves **65** pedigree datasets
Used in the final analysis: **8**

4. **OpenAire**

https://www.openaire.eu/search/find/
Date searched: 25-05-2016

Search for 'Research Data' with search terms: pedigree, pedigrees (as it does not allow use of "", thus cannot search for "relatedness matrix")
Gives **3** results, **1** of which is relevant, but excluded after contacting the authors

5. **Science Research**

http://scienceresearch.com/scienceresearch/
Date searched: 16-6-2016

Search syntax:
(Pedigree* OR "relatedness matrix") NOT cultivar* NOT nation NOT child NOT medicine NOT medical NOT adolescent NOT adolescence NOT autism NOT patient* NOT diagnosis NOT cancer NOT clinical NOT wheat NOT schizophrenia NOT poplar NOT maize NOT "family tree" NOT "family pedigree" NOT "risk factors" NOT "risk factor" NOT "dog food" NOT "middle aged" NOT "middle age" NOT "family health" NOT "mental health" NOT "soya bean" NOT "soya beans"

Include the additional results (this option appears after the search starts)

Initial results: **426**
Result filtering: leaves **123**

- Select publications in Topics:

Genetics (59)

Populations -> tick all in inbreeding (12), Wild (10), Natural population (6), genetic variability (2), small populations (2),

Pedigree analysis -> tick all in Study (5), pedigree data (28),

Large pedigrees -> tick all in Study (3), Genetic Pedigree (2), Animal (11), inbreeding depression (10)

Structure -> tick all in population structure (5), Seleciton (2),quantitative trait (9),potential (7)

- Topics excluded: Molecular (11), Markers (10), Method (10), Genomic (8), Cattle (6), Check library (40), Estimation (40), breeding (26), Complex (9), Analysese (8), Human Pedigree (7), chemistry (5), region (5), E coli (4), Regulation (4), University (4), role (4)

Screened by abstract and title: leaves **28**
Further check for relevance: leaves **5** pedigree datasets
Used in the final analysis: **0**

6. **DataOne, Mercury search**

https://cn.dataone.org/onemercury/

Date searched: 15-06-2016

Search syntax:

Pedigree* or ”relatedness matrix”

cultivar* nation child* medicine medical adolescent adolescence autism patient* diagnosis cancer clinical wheat schizophrenia poplar maize "family tree" "family pedigree" "risk factors" "risk factor" "dog food" "middle age" "middle aged" "family health" "mental health" "soya bean" "soya beans"

- untick the ’Direct access data available’

Initial results: **69**

Screened by title and abstract: leaves **33**

- Some of these belonged to the same data package, and we combined these to obtain final **29** datasets

Further check for relevance: leaves **22** pedigree datasets

Used in the final analysis: **4**

7. **Web of Science Data Citation Index (DCI)**

http://apps.webofknowledge.com/

Date searched: 19-05-2016

Two searches:

a) by TITLE:

(pedigree* OR "relatedness matrix") NOT Cultivar* NOT “family tree” NOT “family pedigree” NOT “dog food” NOT Nation NOT Child NOT Medicine NOT medical NOT Adolescent NOT Autism NOT “family health” NOT “risk factors” NOT patient NOT diagnosis NOT patient NOT diagnosis NOT cancer NOT clinical NOT wheat NOT schizophrenia NOT poplar NOT maize

Initial results: **145**

Result filtering:

a) By Data Type: leaves **134**

- Exclude: NUCLEOTIDE SEQUENCING INFORMATION (9); EXPRESSION PROFILING BY ARRAY (2); METHYLATION PROFILING BY HIGH THROUGHPUT SEQUENCING (1)

b) By Source Title: leaves **128**

- Exclude: ARRAYEXPRESS ARCHIVE (4), SCHOLARSARCHIVE OSU (1), EUROPEAN NUCLEOTIDE ARCHIVE (1)

Screened by title: leaves **86**

Screened by abstract: leaves **63**

b) by TOPIC:

(pedigree* OR "relatedness matrix") NOT Cultivar* NOT “family tree” NOT “family pedigree” NOT “dog food” NOT Nation NOT Child NOT Medicine NOT medical NOT Adolescent NOT Autism NOT “family health” NOT “risk factors” NOT patient NOT diagnosis NOT patient NOT diagnosis NOT cancer NOT clinical NOT wheat NOT schizophrenia NOT poplar NOT maize

Initial results: **782**

Result filtering:

a) By WoS Category: leaves **663**

- Excluded BIOCHEMISTRY MOLECULAR BIOLOGY (119)

b) By Source Title: leaves **555**

- Excluded: EUROPEAN NUCLEOTIDE ARCHIVE (65); ARRAYEXPRESS ARCHIVE (27); GWAS CENTRAL (9); DATABASE OF GENOTYPES AND PHENOTYPES DBGAP (7)

c) By Data Type: leaves **469**

- Excluded: APPLICATION OCTET STREAM (11); TEXT PLAIN (8); VIDEO X MSVIDEO (2); APPLICATION X BZIP2 (1); IMAGE TIFF (71)

Screened by title: leaves **276**

Screened by Abstract : leaves **99**

Further check for relevance of both searches**:** leaves **33** pedigree datasets

Used in the final analysis: **4**

8. **SCIELO**

http://www.scielo.org/php/index.php

Date Searched: 16-06-2016

(pedigree*) AND NOT (child*) AND NOT (cultivar*) AND NOT (nation) AND NOT (medicine) AND NOT (medical) AND NOT (adolescent) AND NOT (adolescence) AND NOT (autism) AND NOT (patient*) AND NOT (diagnosis) AND NOT (cancer) AND NOT (clinical) AND NOT (wheat) AND NOT (schizophrenia) AND NOT (poplar) AND NOT (maize) AND NOT ("family tree") AND NOT ("family pedigree") AND NOT ("risk factors") AND NOT ("risk factor") AND NOT ("dog food") AND NOT ("middle age") AND NOT ("middle aged") AND NOT ("family health") AND NOT ("mental health") AND NOT ("soya bean") AND NOT ("soya beans") AND NOT soybeans

Initial results: **154**

Filtered by ' Scielo Subject Areas' - results in Biological Science: leaves **57**

Screened by title/abstract: leaves **0**

9. Research Data Australia

https://researchdata.ands.org.au/

Date Searched: 16-06-2016

Search for: pedigree* OR "relatedness matrix"

Initial results: **11**

Screened by title: leaves **0** pedigree datasets

10. DLI Service

https://dliservice.research-infrastructures.eu/index.html#/

Date Searched: 14-09-2016

Search for:

pedigree – 99 results

pedigrees – 4 results

Screened by Title/Abstract: leaves **48**

Further check for relevance: leaves **41** pedigree datasets

Used in the final analysis: **5**

11. Data MED

https://datamed.org/

Date Searched: 15-09-2016

Search Syntax:

Pedigree OR Pedigrees OR "Relatedness Matrix" OR "Kinship Matrix" NOT Cultivar NOT Cultivars NOT "Family Tree" NOT "Family Pedigree" NOT "Dog Food" NOT Nation NOT Child NOT Children NOT Medicine NOT Medical NOT Adolescent NOT Autism NOT "Family Health" NOT "Risk Factors" NOT Patient NOT Diagnosis NOT Cancer NOT Clinical NOT Wheat NOT Schizophrenia NOT Poplar NOT Maize

Initial results: **346**

Result filtering (by Source): leaves **303**

- Included: Dryad (300); PeerJ (2); Zenodo (1)
- Excluded: BioProject (91); OmicsDI (41); ArrayExpress (28); dbGaP (17); GEO (2); GEMMA (1)

Screened by Title/Description/Keywords: leaves **184**

- depending on the record, some record contain information on all three elements, and some miss one or two elements
- some of these belonged to the same data package, and we combined these to obtain final **126** datasets

Further check for relevance: leaves **84** pedigree datasets

Used in the final analysis: **9**

Dryad Digital Repository

http://datadryad.org/

Date Searched: 15-09-2016

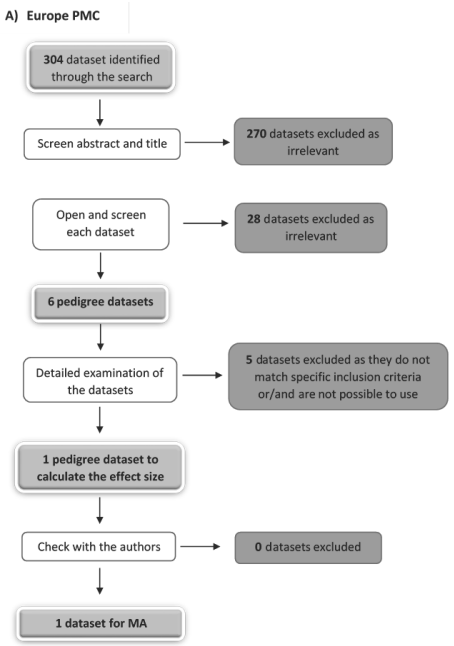
Search Syntax: (Pedigree OR Pedigrees OR "Relatedness Matrix" OR "Kinship Matrix") NOT Cultivar NOT Cultivars NOT "Family Tree" NOT "Family Pedigree" NOT "Dog Food" NOT Nation NOT Child NOT Children NOT Medicine NOT Medical NOT Adolescent NOT Autism NOT "Family Health" NOT "Risk Factors" NOT Patient NOT Diagnosis NOT Cancer NOT Clinical NOT Wheat NOT Schizophrenia NOT Poplar NOT Maize

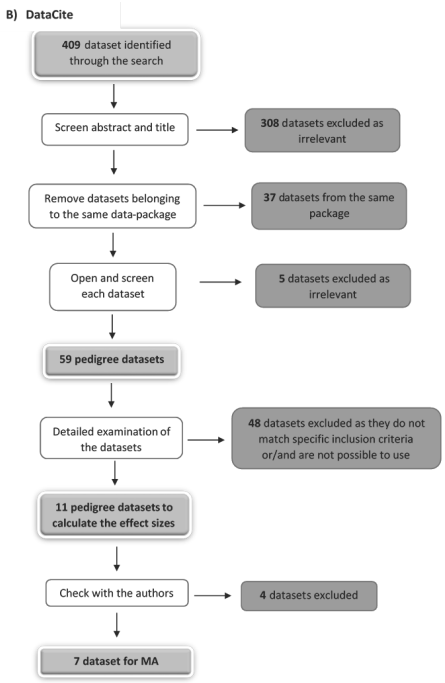
Initial results: **185**

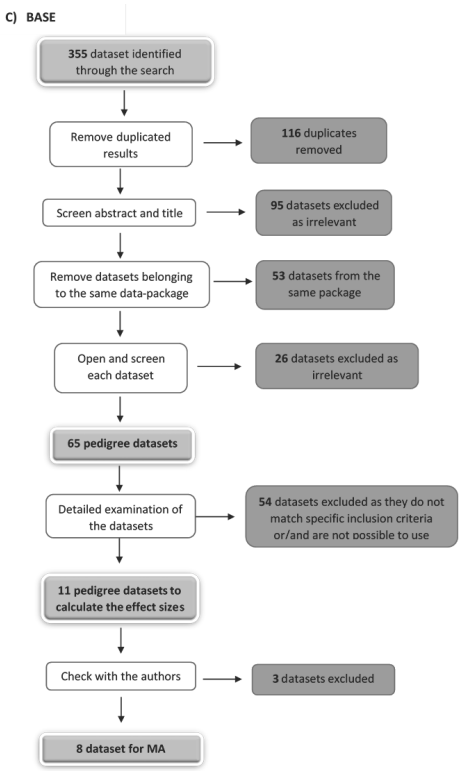
Screened by Title/Abstract: leaves **134**

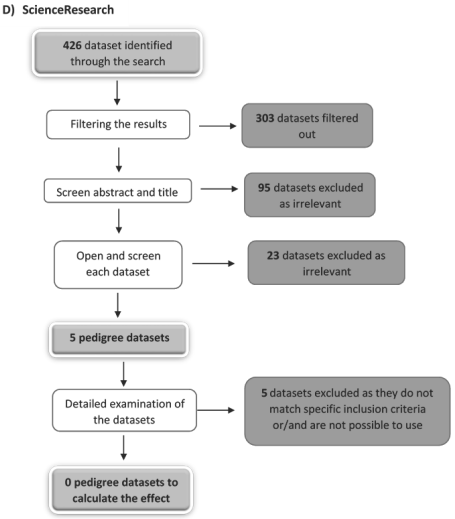
Further check for relevance: leaves **87** pedigree datasets

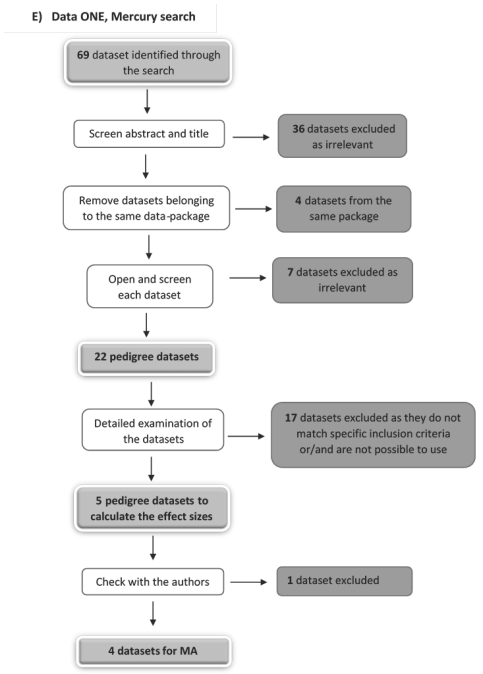
Used in the final analysis: **11**

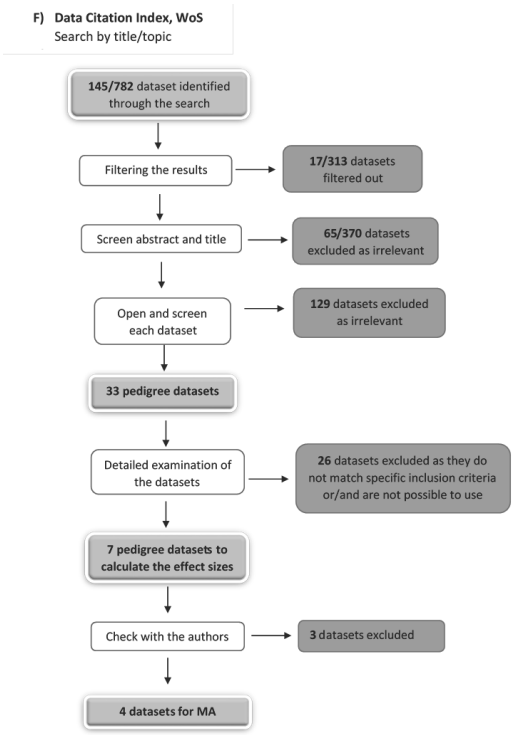

A) Europe PMC
304 dataset identified through the search
Screen abstract and title
270 datasets excluded as irrelevant
Open and screen each dataset
28 datasets excluded as irrelevant
6 pedigree datasets
Detailed examination of the datasets
5 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
1 pedigree dataset to calculate the effect size
Check with the authors
0 datasets excluded
1 dataset for MA

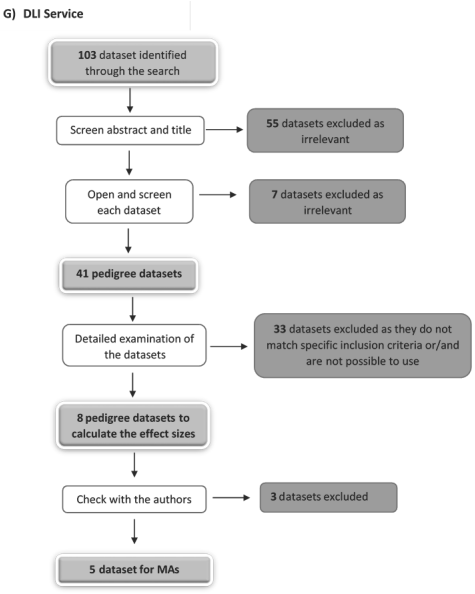

B) DataCite
409 dataset identified through the search
Screen abstract and title
308 datasets excluded as irrelevant
Remove datasets belonging to the same data-package
37 datasets from the same package
Open and screen each dataset
5 datasets excluded as irrelevant
59 pedigree datasets
Detailed examination of the datasets
48 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
11 pedigree datasets to calculate the effect sizes
Check with the authors
4 datasets excluded
7 dataset for MA

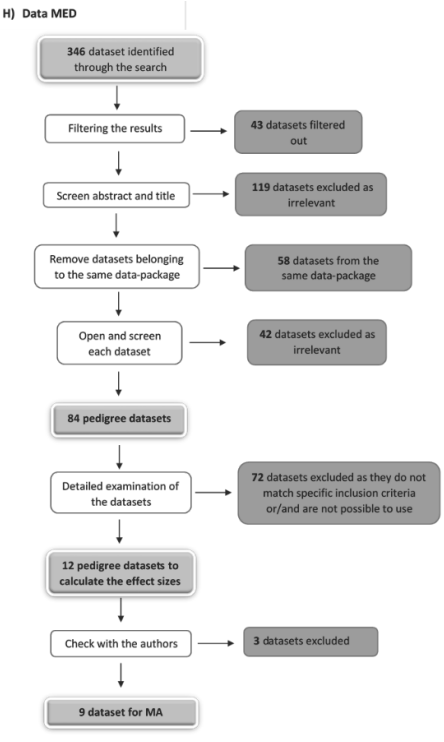

C) BASE
355 dataset identified through the search
Remove duplicated results
116 duplicates removed
Screen abstract and title
95 datasets excluded as irrelevant
Remove datasets belonging to the same data-package
53 datasets from the same package
Open and screen each dataset
26 datasets excluded as irrelevant
65 pedigree datasets
Detailed examination of the datasets
54 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
11 pedigree datasets to calculate the effect sizes
Check with the authors
3 datasets excluded
8 dataset for MA


D) ScienceResearch
426 dataset identified through the search
Filtering the results
303 datasets filtered out
Screen abstract and title
95 datasets excluded as irrelevant
Open and screen each dataset
23 datasets excluded as irrelevant
5 pedigree datasets
Detailed examination of the datasets
5 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
0 pedigree datasets to calculate the effect


E) Data ONE, Mercury search
69 dataset identified through the search
Screen abstract and title
36 datasets excluded as irrelevant
Remove datasets belonging to the same data-package
4 datasets from the same package
Open and screen each dataset
7 datasets excluded as irrelevant
22 pedigree datasets
Detailed examination of the datasets
17 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
5 pedigree datasets to calculate the effect sizes
Check with the authors
1 dataset excluded
4 datasets for MA


F) Data Citation Index, WoS
Search by title/topic
145/782 dataset identified through the search
Filtering the results
17/313 datasets filtered out
Screen abstract and title
65/370 datasets excluded as irrelevant
Open and screen each dataset
129 datasets excluded as irrelevant
33 pedigree datasets
Detailed examination of the datasets
26 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
7 pedigree datasets to calculate the effect sizes
Check with the authors
3 datasets excluded
4 datasets for MA


G) DLI Service
103 dataset identified through the search
Screen abstract and title
55 datasets excluded as irrelevant
Open and screen each dataset
7 datasets excluded as irrelevant
41 pedigree datasets
Detailed examination of the datasets
33 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
8 pedigree datasets to calculate the effect sizes
Check with the authors
3 datasets excluded
5 dataset for MAs


H) Data MED
346 dataset identified through the search
Filtering the results
43 datasets filtered out
Screen abstract and title
119 datasets excluded as irrelevant
Remove datasets belonging to the same data-package
58 datasets from the same data-package
Open and screen each dataset
42 datasets excluded as irrelevant
84 pedigree datasets
Detailed examination of the datasets
72 datasets excluded as they do not match specific inclusion criteria or/and are not possible to use
12 pedigree datasets to calculate the effect sizes
Check with the authors
3 datasets excluded
9 dataset for MA

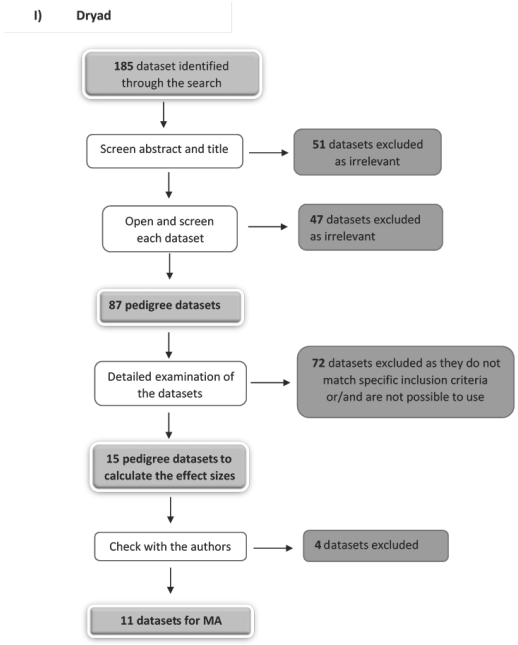


Figure SAD9.1. *PRISMA diagrams for search for pedigreed datasets, and data screening process for different research data aggregators consulted. The reasons for the exclusion of pedigree datasets from the final analysis can be found in Supplementary Table 9.1.1.*

SM9.2: Correspondence with the authors of the datasets

A) General letter used to approach authors:

Dear author(s),

At the Netherlands Institute of Ecology (NIOO-KNAW) we are running a number of long-term studies on individually marked birds, including several population studies on hole-breeding passerines initiated in 1955. We now regularly deposit part of our data in Open Data repositories when journals require us to do so. This has led to the question as to what extent such data can be properly used for answering different questions. To study this we have started a large-scale study on Open Access Data (OAD) on individually marked animals stored in online data depositories by authors upon publication of their scientific papers. The overarching aim of this project, led by Prof. Marcel E. Visser and run by Dr. Antica Culina, is to investigate the usefulness of OAD for answering biological questions, for example by combining data to estimate overall effect sizes of biological phenomena (as opposed to traditional meta-analysis) or by simply testing new hypotheses. Within this project, several investigators are attempting to address certain specific hypotheses (please see the project

website, https://nioo.knaw.nl/en/meta-analysis-meets-open-science, for more information on this subject).

As a participant in this project, I am specifically interested in the prevalence of gene-by-environment (GxE) and selection-by-environment (SxE) interactions across populations and species. For this purpose, we have acquired a number of multi-annual data sets of wild populations containing morphological, physiological, behavioural or life-history trait measurements and pedigree information to tackle this problem. Note that although most data sets lack information about the environment (because the data are not intended for answering these questions), we refrained from enquiring with the authors as this exercise particularly dictates the use of what has already been made available by them. Your data set [*reference to dataset and paper*] is one of the selected data sets. To answer our research questions, we have had to make certain assumptions about your data and, where necessary, reformatted them to make them fit for our analysis.

One concern about the use of OAD for testing new hypotheses is that the data may be misinterpreted, therefore leading to wrong conclusions (see Mills et al. 2016: *Solutions for Archiving Data in Long-Term Studies: A Reply to Whitlock et al.*, *Trends in Ecology & Evolution*, *31*(2), 85-7. DOI: 10.1016/j.tree.2015.12.004). We are therefore writing to all authors (from whom we used OAD) to check whether this is the case; we will report on the frequency of misinterpretation in our manuscript to inform future investigators wishing to embark on similar projects and stress the importance of getting into contact with the original owners of the data. I would like to ask you if you could check whether what we have done makes sense to you—that is, have we made any false assumptions about your data or adapted them in a way that does not concur with your knowledge of the system? For this purpose, I kindly refer to the attached document, which gives a concise description as to how we went about preparing and analyzing your data.

I would very much appreciate your input in this to make sure we can draw sensible conclusions. I would like to point out that your original work will be duly acknowledged in our manuscript. The resulting paper will contain a table with all the data sets used, we shall provide a description of the aim of each original study and both your data and associated paper will be properly cited. I would also like to stress that we are not interested in replicating your original results or verifying the rigorousness of your original analyses. Lastly, I would like to point out that we shall be discreet with respect to your response and report any issues in the manuscript anonymously.

Thank you very much for your cooperation.

Kind regards, on behalf of the whole team,

Jip Ramakers

PhD student
Netherlands Institute of Ecology (NIOO-KNAW)
Droevendaalsesteeg 10
6708 PB Wageningen
The Netherlands
J.Ramakers@nioo.knaw.nl / Jip.Ramakers@gmail.com

B) Specific letter addressing the use of the data

Addressee(s)

Dr. X

Their.names@emailaddress.com

Data

Paper: [paper associated with the dataset]

Data source: [link to repository from which data were retrieved]

Datasets used: [file names]

Aim of our study

To test for evidence of heritability × selection interaction (via GxE and environment-dependent selection) on trait(s) *X* in species *Y*.

Data preparation

As environmental data were not available, we used our own measure of environment to keep analysis across datasets and studies comparable. As the environment, we considered the population-average trait value in a given year ('av.z'); these values were standardized across seasons: $av.z_{std} = (av.z_i - mean(av.z))/sd(av.z)$, where *i* denotes a given year. Note that our chosen measurement of the environment does not identify any specific environmental parameter; rather, it captures the overall environmental 'quality' by looking at the average *X* of individuals in that year (see addendum for rationale behind this approach).

GxE analysis

We first determined the minimum adequate mixed models (Gaussian errors) for each trait in ASReml-R. Using conditional Wald F tests, we identified as fixed effects [*relevant fixed effects*], plus our newly derived measure for the environment ($av.z_{std}$). Likelihood-ratio tests were used to identify important random effects: [*statistically important random effects*]. GxE was then estimated using random regression analysis in MCMCglmm (i.e. by modelling an interaction between the additive genetic effect and $av.z_{std}$, as well as an interaction between individual ID and $av.z_{std}$).

Selection analysis

[*Here we explained how we derived annual recruits and/or survival from the dataset and what assumptions we used to determine selection*].

Tentative results

[*Here we provided tentative results of the analysis, if any*]

Addendum: Rationale behind using mean trait value as a measure of the environment

Given the lack of a suited variable most datasets describing the environment affecting each trait of interest, the mean performance of individuals with respect to the trait of interest in a given environment (year) was used instead. The method is commonly used in animal and plant breeding studies in a process called 'joint-regression analysis', where genotype-specific interactions are partitioned into a component explained by mean population performance and a residual component (Lynch and Walsh 1998, pp. 672–678). The method was first proposed by Yates and Cochran (1938) and later brought into prominence by Finlay and Wilkinson (1963), who applied it to yields of varieties of Barley, where the site mean yield was a predictor of individual variety yields. This approach became widely accepted in the plant- and animal-breeding literature (Lynch and Walsh 1998, James 2009). The method has the advantage that all of the complex (and potentially unobserved) features of the environment are integrated into a single measure. This has two important implications for our analyses: (1) the daunting quest for potentially important environmental variables becomes unnecessary, and (2) the environmental fixed component in our statistical models, now substituting a random effect of 'year' or 'season' to explain environmental variance, can now be used to rank years in terms of overall environmental quality. This approach obviously cannot reveal any detailed links between phenotypes and components of the environment. Note, however, that this is not the purpose of our study. We therefore feel that our use of the population-average trait value in a given year adequately deals with the environmental conditions faced by members of the population, provided that this average is based on the entire population or at least a random portion of it.

References

Finlay, K. W., and G. N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research **14**:742–754.

James, J. W. 2009. Genotype by envionment interaction in farm animals.*in* J. van der Werf, H.-U. Graser, R. Frankham, and C. Gondro, editors. Adaptation and fitness in animal populations: Evolutionary and breeding perspectives on genetic resource management. Springer, Dordrecht, The Netherlands.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sutherland, MA.

Yates, F., and W. G. Cochran. 1938. The analysis of groups of experiments. Journal of Agricultural Science **28**:556–580.

SM9.3: Supplementary Figures

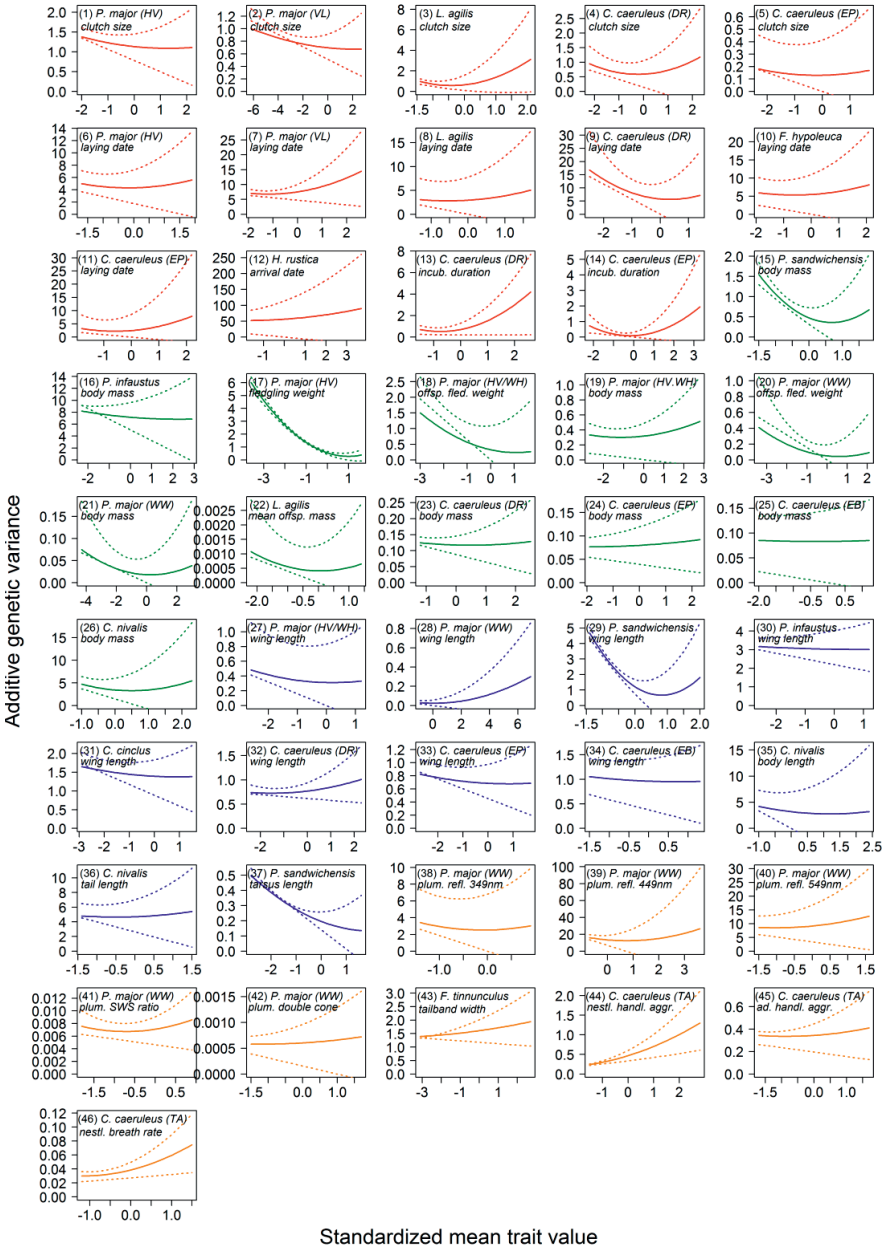


Figure S9.1. *Additive genetic variance (posterior mean and 95% HPDI) as a function of the environment (standardised mean trait value). Colours represent different trait types (red: life history; green: body mass; blue: morphology; orange: miscellaneous). Note that the increase in the width of the HPDIs at higher environmental values is due to uncertainty in the covariance (off-diagonal) estimate in the G matrix. Also note that this figure has fewer panels than Figure S2 and Figure 1 (main text), because G×E was estimated using the whole population (conditioned on the age of individuals), whereas adults and juveniles were separated in subsequent analyses. See main text for details.*

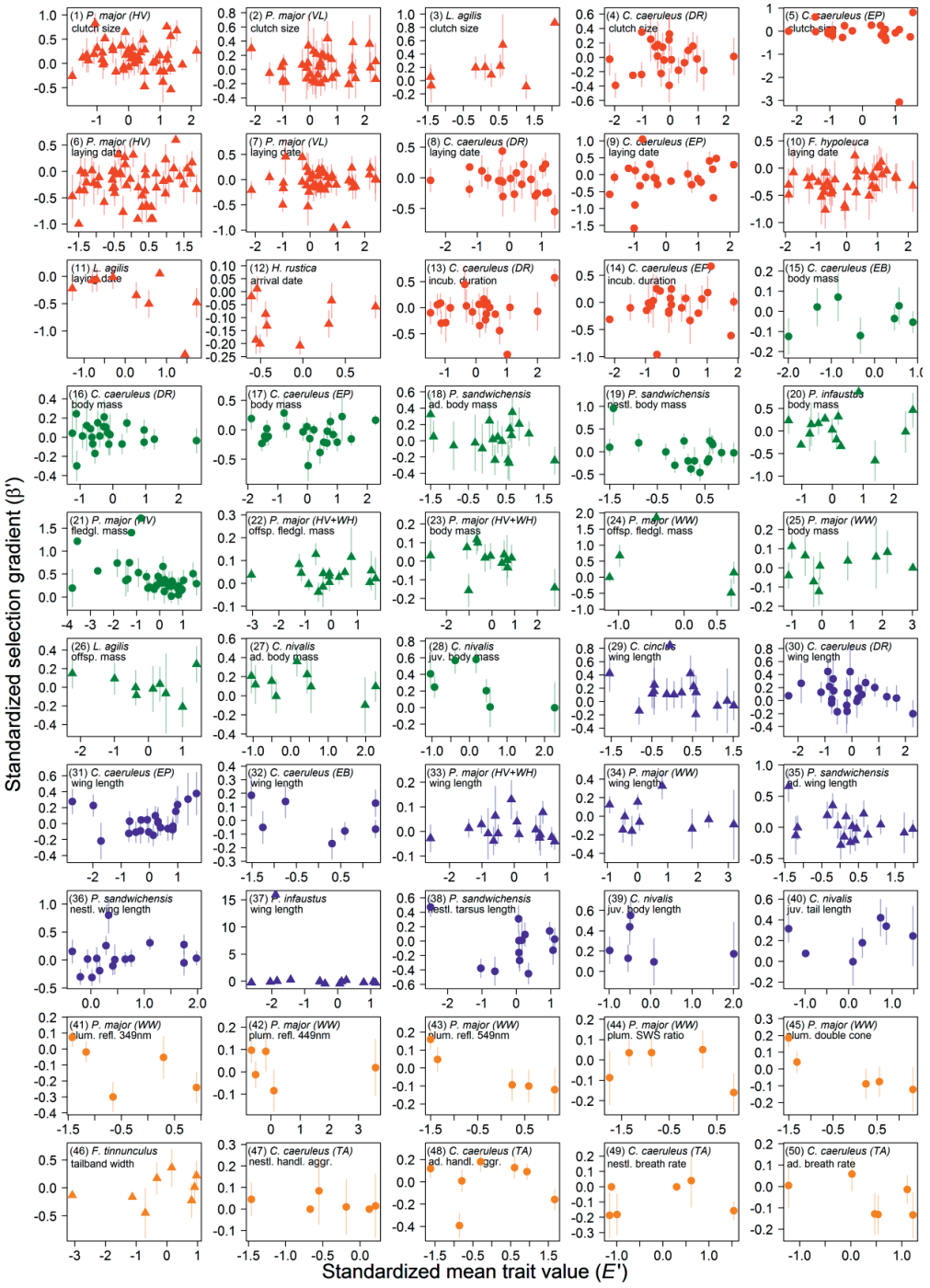


Figure S9.2. *Standardised selection gradients plotted against the environment (i.e. the standardised mean trait value). Colours represent different trait types (red: life history; green: body mass; blue: morphology; orange: miscellaneous), whereas shapes indicate selection based on survival (circles) or based on number of fledglings or recruits (triangles). Standard errors (SEs) are given for the selection gradient (shown only when SE < 0.5).*

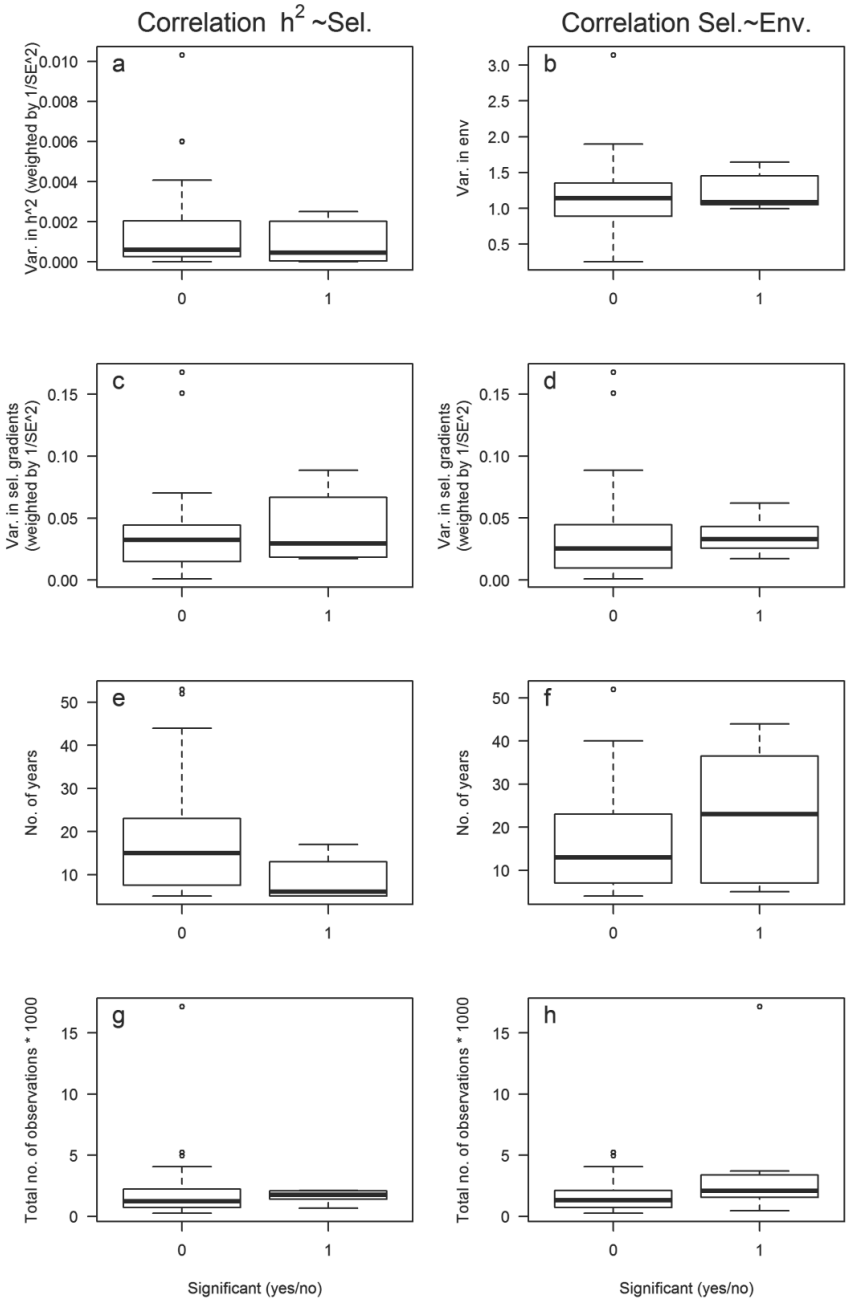


Figure S9.3. *Evaluating statistical power for the heritability–selection relationship (**a**, **c**, **g**, **e**) and for the selection–environment relationship (**b**, **d**, **f**, **h**). **a** Weighted variation in heritability estimates; **b** weighted variation in the environment; **c**, **d** weighted variation in selection gradients; **e**, **f** the number of years available; **g**, **h** the total number of observations for each analysis. The bottom axes denote whether the sought correlation was significant (1) or not (0; n = 6 vs. 44 [**a**, **c**, **g**, **e**] and n = 7 vs. 43 [**b**, **d**, **f**, **h**]). Note that although **g** and **h** were added for completeness, they are not wholly informative for statistical power, which really is a balance between total sample size, number of years (the statistical units in the regression models) and the quality of the pedigree (to estimate heritability reliably).*

Chapter 10

Phenological mismatch drives selection on elevation, but not on slope, of breeding time plasticity in a wild songbird

Jip J.C. Ramakers, Phillip Gienapp & Marcel E. Visser

Abstract

Phenotypic plasticity is an important mechanism for populations to respond to fluctuating environments, yet may be insufficient to adapt to a directionally changing environment. To study whether plasticity can evolve under current climate change, we quantified selection and genetic variation in both the elevation (RN_E) and slope (RN_S) of the breeding time reaction norm in a long-term (1973–2016) study population of great tits (*Parus major*). The optimal RN_E (the caterpillar biomass peak date regressed against the temperature used as cue by great tits) changed over time, whereas the optimal RN_S did not. Concordantly, we found strong directional selection on RN_E, but not RN_S, of egg-laying date in the second third of the study period; this selection subsequently waned, potentially due to increased between-year variability in optimal laying dates. We found individual and additive genetic variation in RN_E but, contrary to previous studies on our population, not in RN_S. The predicted and observed evolutionary change in RN_E were, however, marginal, due to low heritability and the sex limitation of laying date. We conclude that adaptation to climate change can only occur via micro-evolution of RN_E, but this will necessarily be slow and potentially hampered by increased variability in phenotypic optima.

Evolution (2018) **73**, *175–187*

Introduction

Phenotypic plasticity is an important mechanism by which an individual can adapt its phenotype in response to fluctuating environmental conditions (Schlichting and Pigliucci 1998; Pigliucci 2001). For example, life-history traits such as phenology (e.g. timing of breeding or migration) or litter size are often phenotypically plastic, and this plasticity is adaptive as it tracks the environmental variability in the optimal phenotype (Scheiner 1993). Timing of avian seasonal reproduction (or laying date) is an illustrative example in this respect; since laying date is an important determinant of reproductive success and its optimal timing varies from year to year (Visser and Both 2005; Verhulst and Nilsson 2008), it is often phenotypically plastic to environmental conditions, usually spring temperatures (Brommer et al. 2005; Nussey et al. 2005b; Charmantier et al. 2008; Avilés et al. 2014). Similarly, avian clutch size can be phenotypically plastic with respect to population density, such that birds maximize the number of successfully raised, high-quality offspring under varying levels of food availability and competition (Ricklefs 1980; Both et al. 2000; Sæther et al. 2016). More generally, many forms of animal behaviour are highly context dependent in a wide range of taxa (Dingemanse et al. 2010), making phenotypic plasticity ubiquitous in nature (Pigliucci 2001).

Phenotypic plasticity can be described by a reaction norm (Woltereck 1909; Scheiner 1993); that is, the (often assumed to be linear) function of the phenotype against the environment, characterised by the intercept or elevation (i.e. the phenotype in the average environment) and slope (i.e. the sensitivity of the phenotype to the environment). The degree of plasticity may vary among individuals (individual-by-environment interaction or I×E) and this variance may have a partly genetic basis (genotype-by-environment interaction or G×E), making phenotypic plasticity itself an evolvable trait (Scheiner 1993; Van Tienderen and Koelewijn 1994; Via et al. 1995). In a directionally changing environment, evolution of the reaction norm may be necessary because the environmental driver of the trait no longer accurately predicts future environmental conditions, rendering plasticity alone insufficient to respond to environmental change (Visser 2008). Quantifying variation in reaction norms is therefore imperative for understanding evolutionary processes because it can elucidate whether populations are capable of responding to such directional selection. Predicting such responses may be difficult when G×E leads to nonlinear changes in genetic variation across environments (Tomkins et al. 2004; Turelli and Barton 2004; Kokko and Heubel 2008), or when genetic variation and selection are negatively correlated with one another (Wood and Brodie III 2016; but see Chapter 9). Ultimately, the extent to which phenotypic plasticity modulates evolutionary processes will be highly context dependent (Hoffman and Merilä 1999).

A largely unexplored aspect in the light of directional environmental change is how selection on consumer phenology translates to selection on the reaction norm (Visser 2008). If the optimal reaction norm—i.e. the relationship between resource phenology and the environment driving phenology of the consumer (the 'cue')—does not change over time, selection on neither the elevation nor the slope of the consumer reaction norm should occur. If, on the other hand, the sensitivity of the optimal reaction norm to the

environmental cues changes over time, this may lead to two scenarios: selection for increased phenotypic plasticity to track the plasticity of the resource (Nussey et al. 2007; Lande 2009; Gienapp et al. 2014), or selection for reduced plasticity when the cue environment changes to such an extent that the consumer can no longer accurately predict the resource phenology (De Jong 1999; Reed et al. 2010). Alternatively still, if the elevation of the optimal reaction norm changes over time, selection on the elevation, rather than the slope, of the consumer reaction norm should occur. Global climate warming has led to the disruption of phenological synchrony between trophic levels (Visser and Holleman 2001; Both et al. 2006; Durant et al. 2007; Both et al. 2009a; Schultz et al. 2009; Thackeray et al. 2010; Thackeray et al. 2016; Visser and Gienapp in press) and in some cases to directional selection on consumer phenology (Van Noordwijk et al. 1995; Visser et al. 1998). It is unclear whether selection on phenology reflects selection on the elevation, the slope or both components of the reaction norm. Most of what we know about the evolutionary dynamics of the reaction norms stems from theoretical work or laboratory experiments (see Scheiner 1993; De Jong 1999; Van Asch et al. 2007; Lande 2009) or from phylogenetic and population comparisons of reaction norms (Murren et al. 2014), with very few empirical examples of how selection as well as (additive) genetic variation led to an evolutionary change in the reaction norm in wild populations (Van Asch et al. 2013; cf. Carter et al. 2017).

Several long-term, vertebrate study populations have been shown to exhibit phenotypic plasticity in phenology (e.g. Réale et al. 2003b; Brommer et al. 2005; Nussey et al. 2005a; Nussey et al. 2005b; Charmantier et al. 2008). Evidence for I×E and G×E, however, is overall mixed. For example, phenotypic plasticity of laying date against spring temperature in Dutch great tits (*Parus major*) was shown to vary between individuals (I×E), with part of this variation being heritable (G×E; Nussey et al. 2005b). In a UK great tit population, on the other hand, there was no I×E (Charmantier et al. 2008). In a re-analysis for both populations, phenotypic, but not genetic, variation was found to be present in the elevation as well as the slope of the reaction norm in both populations (Husby et al. 2010).

In general, heritable variation in phenological traits is widespread (e.g. Van Noordwijk et al. 1981; Blondel et al. 1990; Réale et al. 2003b; Sheldon et al. 2003), which in the absence of variation in plasticity slopes should reflect variation in the elevation of the reaction norm. Evidence for G×E in wild populations is, however, rare (Wood and Brodie III 2016; Hayward et al. 2018; but see Chapter 9), and estimates of selection on reaction norms are even scarcer. This is because it requires estimating selection on both the elevation and the slope, which is a statistically challenging procedure (see discussion in Weis and Gorman (1990) and in Brommer et al. (2012)). Several studies used random regression techniques to get individual estimates (best linear unbiased predictions, or BLUPs) for reaction norm elevation and slope and performed a separate analysis on these components to quantify selection on them (e.g. Brommer et al. 2005; Nussey et al. 2005b). Such a two-step approach is now considered inappropriate (Hadfield et al. 2010; Morrissey and Liefting 2016) and an alternative method has been suggested to use random regression models to estimate variation in reaction norms as well as selection thereon in a single analysis (Brommer et al. 2012). Application of this method to estimate selection on reaction norms in general has been rare (Brommer et al. 2012; Hayward et al. 2014). Thus, quantitative estimates of both

(genetic) variation in and selection on reaction norms in wild populations are rare or ambiguous, and there is a clear need to empirically quantify the evolutionary dynamics of reaction norms in the light of environmental change (Visser 2008).

To address this gap, we quantified selection on and predicted evolution of the reaction norm of timing of breeding (laying date) in response to temperature (an important environmental cue; Visser et al. 2009a; Schaper et al. 2012) in a Dutch long-term (1973–2016) study population of the great tit (*Parus major*) at the Hoge Veluwe (HV). Laying date is a labile trait that can be expressed several times by an individual; this means that each female has her own reaction norm. Timing of breeding is under increased directional selection (for earlier dates) in this population due to a climate change-driven mismatch with the caterpillar food peak (Visser et al. 1998; Reed et al. 2016b). We tested (i) whether the optimal (linear) reaction norm—determined by regressing the caterpillar biomass peak date against the temperature used by great tits to time laying date—changed over time, both in its elevation and slope; (ii) whether a change in the optimal reaction norm over time led to selection on the phenotypic reaction norm of the consumer (great tit); and (iii) whether there was genetic variation in great tit reaction norms. Based on the results, we used a quantitative genetic model to predict quantitatively the amount of expected change in the population reaction norm (elevation and slope) due to years of selection, and verified the outcome by comparing the predicted reaction norms with those observed in the wild. Combined, these results should elucidate whether the breeding-time reaction norm can evolve given sustained directional selection.

Methods

Data collection and preparation

Data were collected from a long-term great tit (*Parus major*) population at the Hoge Veluwe National Park (HV; 52°02′07″ N, 5°51′32″ E, central Netherlands). The HV population is situated within a matrix of natural habitat, facilitating dispersal from and into the study area, and has been monitored continuously since 1955. Nest boxes are provided in excess (~400) in suitable habitat. Each breeding season from April to July, boxes are checked at least once a week to monitor the breeding activity of hole-breeding passerines. Clutch size is noted and laying date (i.e. the date when a female's first egg of her first clutch in that season is laid) is calculated based on the number of eggs in the nest, assuming that one egg is laid per day. During the chick-feeding phase, parents and chicks are captured at the nest box and ringed, allowing for establishing a 'social' pedigree. Extra-pair paternity in the neighbouring population of Westerheide ranges from 6.5% to 12.5% of all chicks (Van Oers et al. 2008), a common rate for tit species (Brommer et al. 2010) that has been found to only marginally affect the accuracy of heritability estimates when sample sizes are sufficiently large (Charmantier and Réale 2005; Firth et al. 2015).

Temperature data were retrieved as daily averages from a nearby weather station of the Royal Dutch Meteorological Institute (KNMI; Deelen station: 52°05′N, 5°87′E; http://projects.knmi.nl/klimatologie/daggegevens/). Temperature data were averaged

over the period from March 11 to April 20, which is the time window yielding the strongest correlation between annual mean daily average temperature and annual mean laying date (r^2 = 0.74). This was done using a sliding-window approach in the 'climwin' R package (Bailey and Van de Pol 2017); details of the analyses are given elsewhere (Bailey et al. in prep.). We used these mean daily average temperatures as proxies for the environmental cues great tits use to time their reproduction, which in reality is an intricate interplay between day length and changes in temperature over time (Gienapp et al. 2005; Schaper et al. 2012).

The peak date of food availability has been estimated since 1985, using frass samples (see Visser et al. 2006 for details). The most common caterpillars are of the winter moth (*Operophtera brumata*) and the oak leaf roller (*Tortrix viridana*), although other species are present. Caterpillar peak date correlates very well with the mean temperature from March 22 to May 16 (r^2 = 0.81, resulting from a sliding window analysis); we used this relationship to hindcast the caterpillar peak date from 1973 to 1984 (all temperatures fell within the analysed range).

Here, we consider brood data from 1973 to 2016, as the study area was reorganised following a major storm in 1972, and the latest data on recruitment of fledged offspring (our fitness measure) was available for 2017. During the study period, a number of broods were manipulated in brood size. None of the broods that we considered were affected in their laying date, but brood size manipulations could affect the reproductive success of a brood. We therefore included broods whose size was manipulated in the laying date analyses, but removed them from annual selection analysis (but not for analysis of selection on the reaction norm; see below). Sample sizes per analysis are given in Table 10.1.

***Table 10.1**. Sample sizes used in all analyses.*

Analysis	Year span	N_{years}	$N_{females}$	N_{broods}
1. Optimal laying date vs. environment	1973–1987	15	-	-
	1988–2001	14	-	-
	2002–2016	15	-	-
2. Selection on plasticity[a]	1969–1987	19	456	1272
	1988–2001	14	365	953
	2002–2012	11	280	691
3. Quantifying ***G***-matrix	1973–2016	44	3028	4890
4. Selection (β_z) on laying date[b]	1973–2016	44	>2347	3662
5. Observed reaction norms	1973–1987	15	1026	1650
	1988–2001	14	993	1551
	2002–2016	15	1126	1689

[a]Year span here indicates cohort span; birds with incomplete lifetime reproductive success (LRS) at the end of the dataset were omitted, whereas LRS of birds breeding in 1973 were complemented with brood data from previous years, hence making the cohort span 1969–2012 (see text). Only birds with ≥ 2 breeding events were included here.

[b]Reduced dataset without manipulated broods; exact $N_{females}$ is unknown because this analysis includes broods whose mother could not be identified.

Change in the relationship between optimal phenotype and the environment

If the optimal egg-laying date (ELD_θ, i.e. the laying date at which fitness is maximised) advanced at a faster rate over time than the observed laying date, this could lead to selection on (1) the reaction norm elevation, if the sensitivity of the optimal reaction norm to temperature (i.e. plasticity) remained constant over time, or (2) the slope, if the sensitivity of ELD_θ to temperatures changed over time. We therefore determined ELD_θ in each year and regressed it against the temperature cues used by the birds (i.e. the mean temperature across March 11 – April 20). ELD_θ was defined as the peak date of caterpillar biomass minus 33 (see Chevin et al. 2015). The rationale behind this was that it takes approximately 33 days from the laying date to the moment when chicks' food requirements are highest, assuming a clutch size of 9–10, followed by 12–13 days of incubation, and a peak in food demands at chick age 9–11 d (e.g. Keller and Van Noordwijk 1994; Mols et al. 2005).

To test whether the optimal reaction norm changed over time, we divided the datasets into three equal-interval time periods (spanning 14–15 years: 1973–1987, 1988–2001, 2002–2016; Table 10.1). We then fitted a linear model with LD_θ as a function of temperature and time period. We assessed statistical significance of model terms by bootstrapping (1000 iterations, bias-corrected and accelerated confidence intervals) estimates for each period, as well as the temperature slope. We then ran a model that contained an interaction between temperature and period and bootstrapped the change in the slope of the optimal reaction norm between periods to determine whether the slope changed over time. The year 1991 was excluded from this analysis as an exceptionally late frost spell in that year damaged all fresh oak leaves and hence the food peak was exceptionally late (see also Visser et al. 2002; 2006).

Selection on the slope and elevation of the great tit reaction norm

To estimate period-specific (fecundity) selection on plasticity, we assigned each female with at least two breeding attempts to a breeding cohort (based on the first year a bird was observed breeding) and split the phenotypic dataset into three time periods as above. A female's lifetime reproductive success (LRS) was the sum of all recruits she produced over her lifetime. To avoid truncation of LRS of birds breeding in 1973, we added broods from earlier years (before 1973) to complete each individual's LRS. For the same reason, we removed all observations of birds from the 2013 breeding cohort or later, as some birds from these cohorts were still known to be breeding in 2017 and therefore had incomplete LRS. Hence, the periods in this analysis were 1969–1987, 1988–2001 and 2002–2012 (Table 10.1).

As pointed out above, some broods had been manipulated, likely affecting the female's fitness. Since we were interested in the lifetime fitness consequences of being more or less phenotypically plastic with respect to temperature, as well as having a higher or lower mean response (i.e. the elevation of the reaction norm), we opted to include manipulated broods in this analysis. The reason for this was that because the probability of having any of her broods manipulated increases with a female's age, discarding data from

manipulated females would lead to a biased subset of shorter-lived individuals. Consequently, the inclusion of manipulated broods in our analysis may create some noise in the selection estimates, but we nevertheless believe it is a superior approach to removing a non-random subset of individuals from the data.

For each of the three periods, we fitted a bivariate random regression model (RRM) in 'MCMCglmm' (Hadfield 2010; Hadfield 2018) with laying date as a Gaussian and LRS as a (overdispersed) Poisson trait. Breaking down this bivariate model in scalar notation, the laying date (z) of the i^{th} individual in the j^{th} year in the k^{th} nest box within the l^{th} 'environmental block' was modelled as

$$z_{ijkl} = \alpha_z + a_i + b_i\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + age_{ij} + nb_k + yr_j + e_{z,ijl}, \quad (10.1)$$

where α_z is the overall mean laying date (intercept), a_i and b_i are the individual intercept and slope, respectively, related to temperature (T) and b is the population-level slope, age_{ij} is the female's age (first-year breeder, older, or unknown), nb_k and yr_j are the k^{th} nest box and j^{h} year, respectively (treated as random effects with estimated variance $nb_k \sim N(0, \sigma^2_{nb})$ and $yr_j \sim N(0, \sigma^2_{yr})$) and $e_{z,ijl}$ is the residual term. This residual term was estimated as $e_{z,ijl} \sim N(0, \sigma^2_{e,l})$, where l is one of two equal-interval groups of years with similar temperatures, which was done to accommodate changes in residual variance along the temperature gradient (Lillehammer et al. 2009; Nicolaus et al. 2013). Temperature was divided into a within-individual ($b_i\left(T_{ij} - \bar{T}_i\right)$) and a between-individual ($b\bar{T}_i$) component, following Van de Pol and Wright (2009). This method disentangles any effect of temperature on laying date caused by having observed certain individuals only under certain temperatures and others under different temperatures (a between-subject effect of T on z) from a real, within-subject effect (but note that for the purpose of estimating variance in intercepts, centring may not be desired in all context; see Kreft et al. (1995)).

A female's LRS (W) is described as

$$\log(E[W]_i) = \alpha_W + c_i + e_{W,i}, \quad (10.2)$$

where α_W is the overall mean LRS, c_i is the individual intercept and $W_i \sim Poisson(E[W]_i)$. We set the residual term, $e_{W,i}$, to a fixed small variance (0.01) since LRS is a repeated measures trait and laying date is not; ideally, we would constrain variance to 0 (Brommer et al. 2012; Morrissey et al. 2012a), but since the MCMC chain does not mix well under these conditions a minimum non-zero value needs to be assigned.

Variance in intercepts (a) and slopes (b) of the reaction norm (eqn. 10.1), as well as in intercepts of LRS (c; eqn. 10.2) were jointly fitted as the random effect 'female identity', such that individual values were estimated as

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix}_i \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{a,b} & \sigma_{a,c} \\ \sigma_{a,b} & \sigma_b^2 & \sigma_{b,c} \\ \sigma_{a,c} & \sigma_{b,c} & \sigma_c^2 \end{bmatrix}_i \right),$$

i.e. drawn from a 3×3 unstructured, phenotypic covariance matrix with the variance components on the diagonals and the covariances between them on the off-diagonals.

To obtain independent samples from the MCMC sampling process, we used a thinning interval of 10^4, a burn-in period of 10^5 and a total sample size of 10.1×10^6 (i.e. an effective sample size of 1000). We used default normal priors for the fixed effects and parameter-expanded priors for the random terms (V = diag(n), nu = n, alpha.mu = 0, alpha.V = diag(n)*25^2, n being the dimension of the matrix), following Hadfield (2018). For the residual variance in eqn. (10.1), we used inverse-Wishart priors with V = diag(2) and nu = 0.002.

Predicting changes in the reaction norm

We used a multivariate version of the breeder's equation to predict changes in the reaction norm as a result of selection across the three time periods (Table 10.1). The evolutionary response to selection from one year to the next is defined as the trait heritability times the selection differential ($R = h^2S$, i.e. the breeder's equation; Falconer and Mackay 1996). Even in its multivariate form (Lande 1979), however, this would require estimating the heritability of both components of the reaction norm (elevation and slope), which is not straightforward (Hadfield et al. 2010). We therefore used a derivation of the breeder's equation for reaction norms (Van Tienderen and Koelewijn 1994). In their Appendix 1, Van Tienderen and Koelewijn (1994) show that the selection gradients on reaction norm components, i.e. intercept and slope in our case, can be calculated from environment-specific—here hence annual—selection gradients:

$$\boldsymbol{\beta}_g = \boldsymbol{\Phi}^t \boldsymbol{\beta}_z, \tag{10.3}$$

where $\boldsymbol{\beta}_g$ is the vector of selection gradients on intercept and slope, $\boldsymbol{\Phi}^t$ is the transposed matrix of vectors consisting of the environmental values in environments *i...k* with a leading 1, and $\boldsymbol{\beta}_z$ the vector of selection gradients in environments *i...k*. The (expected) genetic change in the breeding values for intercept and slope is given by the multivariate breeders' equation:

$$\Delta\overline{\boldsymbol{g}} = \boldsymbol{G}\boldsymbol{\beta}_g, \tag{10.4}$$

where $\Delta\overline{\boldsymbol{g}}$ is the change in intercept and slope breeding values, and $\boldsymbol{G}$ the additive genetic variance-covariance matrix for intercept and slope. Substituting eqn. (10.3) and (10.4)

allows the change in reaction norm components ($\Delta\overline{\boldsymbol{g}}$) as a result of selection in year j following

$$\Delta\overline{\boldsymbol{g}} = \boldsymbol{G}[1, x_j]^t \beta_{z_j}, \quad (10.5a)$$

where $[1, x_j]^t$ is a transposed vector characterising the environment in a given year (x_j), and β_{z_j} the selection gradient. We then only need to accommodate generation time by multiplying the response by the proportion of the breeding population in environment j+1 represented by recruits ($p_{recr_{j+1}}$) and further halve the response because laying date is only expressed by females (following Gienapp et al. 2006):

$$\Delta\overline{\boldsymbol{g}} = \boldsymbol{G}[1, x_j]^t \beta_{z_j} \times p_{recr_{j+1}} \times 0.5. \quad (10.5b)$$

To estimate $\boldsymbol{G}$, we fitted a univariate RRM on the entire dataset (i.e. all years and all observations) that included an additive genetic term ('animal model', RRAM; Henderson 1988; Kruuk 2004) in 'MCMCglmm'. Random effects were as in eqn. (10.1), with the addition of an additive genetic term (via the pedigree) to estimate additive genetic variance in laying date. The difference was that we allowed both the permanent environment term (i.e. female identity) and the additive genetic term to vary with grand-mean-centred, as opposed to individual-mean-centred, temperature. Thus, the RRAM took the form:

$$z_{ijkl} = \alpha_z + a_i + A_i + (b + b_i + B_i)\left(T_{ij} - \overline{T}\right) + age_{ij} + nb_k + yr_j + e_{z,ijl}, \quad (10.6)$$

where definitions and indices are as for eqn. (10.1), but where the individual intercepts and slopes of the laying date–temperature reaction norm are now estimated on the grand-mean-centred temperature ($\overline{T}$) and partitioned into a permanent environment (a_i and b_i) and additive genetic component (A_i and B_i; breeding values). Phenotypic and additive genetic variance in these components was estimated using two separate, 2×2 unstructured phenotypic and additive genetic ($\boldsymbol{G}$) variance–covariance matrices containing σ_a^2 and σ_b^2 or σ_A^2 and σ_B^2, respectively, as well as the covariance between each component. The residual term was estimated from a 4×4 matrix to allow for independent and heterogeneous variance as in eqn. (10.1), where 4 is the rounded number of years divided by 10, grouped based on temperature. We extracted posterior medians from $\boldsymbol{G}$, along with 95% HPDIs (note that because variance estimates are constrained to be positive, they have a skewed posterior sample distribution when close to 0, hence making the median a more appropriate point summary than the mean). We applied the same prior structure and sampling procedures as for eqn. (10.1). Estimates resulting from the RRAM were robust to excluding individuals with only one or only two breeding records.

To estimate the selection gradient, β_{z_j}, we ran annual Generalized Additive Models (GAMs; package 'mgcv'; Wood 2017) on unmanipulated broods, with annual reproductive success (ARS; number of recruited offspring) as response variable with a negative binomial distribution and laying date as the predictor variable. We then used the 'gam.gradients' function from the 'gsg' package (Morrissey and Sakrejda 2015) to calculate β_{z_j} (see Lande and Arnold 1983) along with the standard error through parametric bootstrapping (1000 iterations). Females with unknown identity were included in these analyses as they comprise a potentially biased subset of individuals that laid their eggs too early or too late, in some cases leading to brood desertion before the nestling stage when adults could be identified.

With $\boldsymbol{G}$ and β_z in place, we could predict the evolution of the reaction norm throughout the study period. Since the reaction norm has to be estimated across years (because each year has only one temperature average and one breeding event), we could not predict change from the first year onward. Instead, we estimated the observed population reaction norm across years in period 1 (see Table 10.1) using a slightly modified version of eqn. (10.1) ($z_{ijkl} = \alpha_z + a_i + b\left(T_{ij} - \bar{T}_i\right) + b\bar{T}_i + age_{ij} + nb_k + yr_j + e_{z,ijl}$, a random-intercept model) and predicted the evolutionary outcome of selection in all subsequent years (i.e. excluding years in period 1) using eqn. (10.5b). Note that the predicted reaction norm after selection is then a property of a single year and will hence never accurately match the observed reaction norm, which is necessarily estimated across multiple years. We therefore view the predicted change as the upper limit of possible change within the studied period.

Since both $\boldsymbol{G}$ and $\boldsymbol{\beta_z}$ came with estimation errors, we accommodated uncertainty in predicted cumulative change in the following way. We used the upper and lower 95% HPDI of $\boldsymbol{G}$ to calculate the variance (V) for $\boldsymbol{G}\left[1, x_j\right]^t$ in each year by squaring its standard error (half the 95% HPD range divided by 1.96); similarly, we squared the standard error of β_{z_j} to get its variance. The standard error of the full product was then calculated following simple error-propagation rules (Taylor 1997):

$$SE_{\Delta\bar{g}} = \sqrt{\left(\boldsymbol{G}\left[1, x_j\right]^t\right)^2 V_{\beta_{z_j}} + {\beta_{z_j}}^2 V_{\boldsymbol{G}\left[1, x_j\right]^t}}, \tag{10.7}$$

which, in the case of eqn. 10.5b, was in turn multiplied by $p_{recr_{j+1}} \times 0.5$. We calculated the cumulative difference across years ($\Delta\bar{g}_{cum.}$) by summing $\Delta\bar{\boldsymbol{g}}$ of each year; the 95% CI was calculated as $\Delta\bar{g}_{cum.} \pm 1.96 \times (SE_{\Delta\bar{g}_1}{}^2 + \cdots + SE_{\Delta\bar{g}_n}{}^2)^{0.5}$, where n is the number of years in period 2 and 3 combined. We need to make the cautionary note that although this way of calculating of errors should work satisfactorily when posterior distributions approximate normality, it may cause upward bias (i.e. be conservative) for variance estimates that are skewed because they are bounded at 0.

We visually compared the predicted reaction norm with the observed reaction norms in each of the three periods, which we estimated in 'MCMCglmm' using a trimmed

version of eqn. (10.6) as described above and using the same priors and MCMC sampling procedures as described for eqn. (10.1). Although we cannot compare the predicted and observed reaction norms quantitatively for the reasons given above, we do present them alongside one another as the latter serves as an empirical validation of the former.

Results

Change in optimal reaction norm over time

We found that the optimal reaction norm for laying date (estimated as the date of the caterpillar biomass peak versus the temperatures in the time window as used by great tits to time their reproduction) changed to earlier dates over three distinct periods (Figure 10.1a); the optimal phenotype (ELD_θ) advanced by 7.72 days (95% CI: [–10.89, –4.79]) from the first to the second period, and by 8.45 days (95% CI: [–12.75, –5.13]) from the first to the third period. There was, however, no interaction between temperature and time period (change in slope coefficient from first to second period: –1.54 [–4.19, 2.15]; from the first to the third period: 0.58 [–2.10, 3.31]), indicating that the response of LD_θ to temperature did not change over the three periods (slope coefficient across all years: –3.15 [–4.36, –1.76] days/°C). Besides the mean, inter-annual variance in optimal laying date changed across the three periods (24.31 [10.82, 50.36], 29.98 [15.09, 64.27] and 63.70 [30.56, 149.37], respectively).

Table 10.2. *Posterior medians (and 95% HPDIs) of the phenotypic (co)variance matrix resulting from an analysis of selection on the reaction norm of egg-laying date (ELD) against temperature (via lifetime reproductive success (LRS); eqns. 10.1 and 10.2) for three distinct cohort periods in the HV great tit population, excluding females with only one observation (see Table 10.1 for sample sizes). Variance estimates are given on the diagonals, whereas covariance estimates are on the off-diagonals. Covariance estimates whose HPDI did not include zero are marked in bold.*

	$ELD_{intercept}$	ELD_{slope}	LRS
Period 1			
$ELD_{intercept}$	6.48 (4.78, 8.09)		
ELD_{slope}	0.29 (–0.26, 1.11)	0.15 (0.00, 0.77)	
LRS	–0.32 (–0.74, 0.11)	–0.02 (–0.25, 0.20)	1.12 (0.89, 1.40)
Period 2			
$ELD_{intercept}$	9.88 (7.67, 12.62)		
ELD_{slope}	1.00 (–0.32, 2.80)	0.43 (0.00, 1.61)	
LRS	**–0.98 (–1.58, –0.32)**	–0.24 (–0.76, 0.15)	1.51 (1.14, 1.97)
Period 3			
$ELD_{intercept}$	7.92 (5.52, 10.70)		
ELD_{slope}	–0.24 (–1.04, 0.55)	0.21 (0.00, 0.69)	
LRS	–0.53 (–1.15, 0.13)	–0.22 (–0.46, 0.05)	1.23 (0.85, 1.62)

Selection on the great tit reaction norm

Along with a change to earlier dates in the elevation of the optimal reaction norm, we found statistical evidence for directional selection on the elevation of the great tit reaction norm in period 2; that is, lifetime reproductive success (LRS) covaried negatively with the elevation, indicating selection for a lower reaction norm (Table 10.2). Despite overwhelmingly negative annual selection gradients (Table S10.1), this selection on the intercept weakened in period 3, potentially related to the high variability in phenotypic optima in that period. We found no evidence for directional selection on slopes, and estimates of slope variation were overall small and zero-bound (Table 10.2). This low slope variance was also found when we used the entire 1973–2016 dataset (Table 10.3; see Discussion).

Table 10.3. *Model estimates resulting from the RRAM (eqn. 10.6) quantifying the* ***G*** *matrix for great tit laying date in the HV population between 1973 and 2016 (see Table 10.1 for sample sizes). The permanent-environment and additive genetic (co)variance components are marked in bold.*

Parameter	Posterior median	95% HPDI	
Fixed effects			
Intercept	21.84	20.82	22.77
Age (old)	-	-	-
Age (unkn.)	1.27	0.42	2.13
Age (young)	1.78	1.51	2.05
Temperature (T_c)	–3.28	–3.92	–2.68
Random effects			
V_{year}	9.65	5.92	14.45
$\boldsymbol{V_{PE}}$	**3.24**	**0.88**	**5.68**
$\boldsymbol{Cov(\beta 0_{PE}, \beta 1_{T_c})}$	**0.06**	**–0.16**	**0.47**
$\boldsymbol{V_{PE \times T_c}}$	**0.02**	**0.00**	**0.15**
$\boldsymbol{V_A}$	**4.38**	**1.98**	**6.70**
$\boldsymbol{Cov(\beta 0_A, \beta 1_{T_c})}$	**0.07**	**–0.20**	**0.43**
$\boldsymbol{V_{A \times T_c}}$	**0.02**	**0.00**	**0.15**
V_{NB}	1.22	0.85	1.66
$V_{R_{3.64-5.16℃}}$	11.09	9.40	13.05
$V_{R_{5.16-6.68℃}}$	13.63	12.20	14.99
$V_{R_{6.68-8.20℃}}$	19.21	17.35	21.17
$V_{R_{8.20-9.72℃}}$	19.27	16.32	21.90

Note. T_c = mean-centred temperature; V_x = variance component associated with each random effect (*PE* = permanent environment, *A* = additive genetic, *NB* = nest box; *R* = residual); $Cov(\beta 0, \beta 1)$ = intercept–slope covariance.

Predicted and observed great tit reaction norms

There was ample additive genetic variation in the elevation, but not the slope, of the great tit reaction norm (Table 10.3). Based on selection patterns in Table 10.2, we may expect the

reaction norm to evolve toward an earlier average over time, but its slope to change only marginally because of the weak selection on it. Using the quantitative genetic model (eqn. 10.5b), we found that both the elevation and the slope of the reaction norm were expected to evolve only marginally over time (Figure 10.1b; cumulative change in elevation across years in period 2 and 3: –2.34 [–4.20, –0.48] days; in slope: –0.04 [–0.09, 0.02]; see Table S10.1 for annual predicted responses). The small predicted response in the elevation—compared to the change in the optimal reaction norm—despite strong selection in period 2 was due to the generation time of about two years and the sex limitation of the trait; when we disregarded these factors (eqn. 10.5a), the predicted response in elevation was substantially stronger (cumulative change in elevation: –10.30 [–18.63, –1.98] days; in slope: –0.16 [–0.40, 0.08]; Table S10.1).

In close agreement with the predicted reaction norms, the observed reaction norm showed no distinct advancement over the three periods (Figure 10.1c), with largely overlapping 95% HPDIs for the intercepts (posterior medians and 95% HPDIs for period 1–3: 41.80 [33.40, 49.29], 41.27 [30.72, 53.50] and 42.73 [35.85, 50.81]).

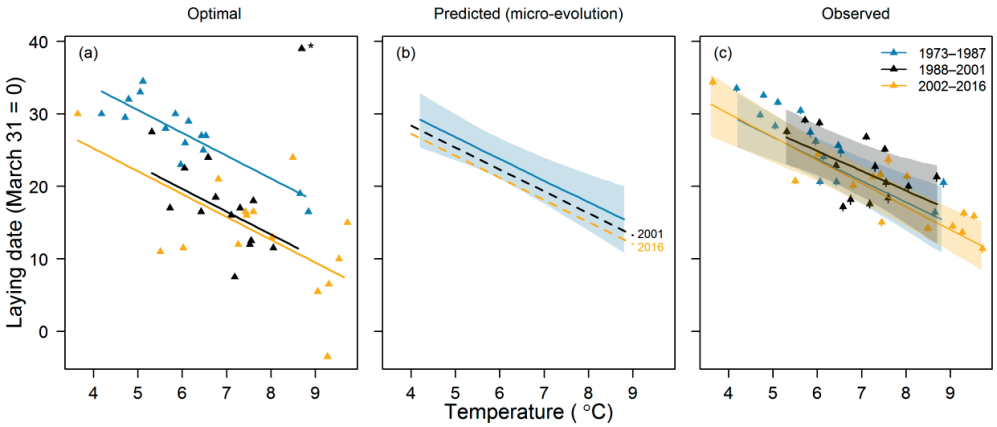


Figure 10.1. *Optimal (**a**), predicted (**b**), and observed (**c**) reaction norms of laying date against spring temperature in three consecutive periods (blue: 1973–1987; black: 1988–2001; orange: 2002–2016) in the HV great tit population. (**a**) The optimal laying date for each year was the caterpillar peak date minus 33 days. Lines are estimates from linear regressions, excluding 1991 (marked by an asterisk) because late frost damaged the oak leaves in that year. (**b**) The black and orange dashed lines are the predicted evolutionary deviation from the observed reaction norm in period 1 (solid blue line) by the end of the second (2001) and third (2016) time period, respectively, based on cumulative change due to annual selection (eqn. (10.5b); see Methods for interpretation). (**c**) Observed laying dates are yearly averages (± 1 s.e.m.); solid lines and shadings are the regression lines and 95% HPDI regions from a univariate mixed-effects model of laying date against temperature (i.e. the mean individual-level slope).*

Discussion

We studied whether the optimal reaction norm of great tit laying date against temperature changed over three consecutive time periods in a long-term study population and whether this resulted in intensified directional selection on—and an evolutionary response in—its two components (elevation and slope, i.e. the sensitivity to the environmental variable). We found that, whereas the optimal laying date (as predicted by spring temperatures and determined by the timing of the caterpillar peak) across the temperature range (the elevation of the optimal reaction norm) advanced over the past decades, the sensitivity of the optimal laying date to temperature (the slope of the optimal reaction norm) did not change. In agreement with this, there was selection on the great tit reaction norm elevation, but not on the slope, in the second third of our time series, but this selection waned in the third period, potentially due to the increased variability in phenotypic optima (Fig. 10.1a). Despite directional selection, we predicted quantitatively that the elevation of the reaction norm would shift only marginally over time (a maximum of a few days), simply because of the low heritability, the generation time of about two years (Garant et al. 2004a; Kvist et al. 2007) and the sex limitation of laying date (Caro et al. 2009). Indeed, the actual (observed) reaction norm did not change over time, neither in slope nor elevation. Our results suggest that, despite the apparent lack of an evolutionary response in our population, adaptation of timing of breeding to climate change can only occur through evolution of the elevation of the reaction norm, as evolution of increased (or decreased) phenotypic plasticity is not possible.

Selection on plasticity in our study population was limited because the slope of the optimal reaction norm did not change over time. However, even if there had been substantial directional selection on the slope, the response to selection would have been weak as there was very little variation in individual reaction norm slopes, both phenotypically and additive genetically (Tables 10.2 and 10.3). This qualitative statement can be quantified by comparing the variances in environment-specific expected values of laying date attributable separately to variation in reaction norm elevations (i.e. the estimates from Table 10.3) and slopes (i.e. estimated variance multiplied by the variance in centred temperatures; see Appendix 7 in Morrissey and Liefting 2016). This was 3.24 [0.88, 5.68] and 4.38 [1.98, 6.70] for the phenotypic and additive genetic intercepts, respectively, whereas it was only 0.05 [0.00, 0.34] and 0.04 [0.00, 0.34] for the respective slopes. Thus, the contribution of variance in slopes to variance in expected laying dates in a given environment was negligible.

The lack of variation in (phenotypic) plasticity slopes (I×E) contradicts previous work on our study population (Nussey et al. 2005b; Husby et al. 2010, 2011). The source of this discrepancy lies in the residual variance structure of the random regression models. Besides the problems inherent to using BLUPs in this context, Nussey et al. (2005b) fitted homogeneous residual variance, thereby forcing any heterogeneity in residual variance caused by the environment (i.e. temperature) to be estimated by the phenotypic covariance matrix and hence inflating estimates of I×E. Husby *et al.* (2010; 2011b), on the other hand, fitted a heterogeneous residual structure but grouped years by decade, assuming that

variance in laying date increased over the years because temperature increased over time. This assumption does not hold up, however, as we found that the variance in laying date in HV correlated very weakly with year (slope = 0.001 [–0.22, 0.22]; r^2 = 0.00 [0.00, 0.001]), whereas it correlated reasonably well with temperature (2.40 [0.70, 4.49]; r^2 = 0.13 [0.01, 0.36]). In general, it would make more sense to partition residual variance based on years with similar environments rather than based on decades. Indeed, when we fitted the model of eqn. (10.6) (but dropping the G×E term for efficiency) on the same data subset (1973–2006) as in Husby *et al.* (2010, 2011) and partitioned residual variance into three groups by decade, we found substantial individual variation in slopes (posterior median slope variance and 95% HPDI: 0.91 [0.36, 1.53]). When we refitted the model on the same data subset (1973–2006) with residual variance partitioned into three groups based on temperature, however, the slope variance decreased substantially (0.25 [0.00, 0.87]). These results were confirmed by simulations; when we simulated a population with a small slope variance, specifying the incorrect residual structure in the random regression model led to upward bias in the variance estimate and a high false-positive rate (Supplementary Methods SM2, Fig. S10.1). Thus, we conclude that variation in plasticity in HV really is limited, reinforcing the notion that, despite previous studies (Nussey et al. 2005b), there can be only limited response to selection on plasticity in this population (cf. e.g. Brommer et al. 2012; Hayward et al. 2014).

Although we can only speculate about the reasons for an apparent lack of I×E and G×E in our population, previous simulation studies suggest it is unlikely that we had limited statistical power (or precision in our case) to detect either (Martin et al. 2011; Van de Pol 2012). Illustrative in this respect are the narrow 95% HPDIs for slope variances (Table 10.3), which seem to indicate that the RRAM was able to estimate I×E (strictly speaking PE×E and G×E) with a fair amount of precision. To put this in perspective, when we applied a version of the model of eqn. 10.6 (that included individual-mean-centred temperatures) to another of our long-term great tit study populations on the Dutch island of Vlieland (VL), we found substantial evidence for I×E in laying date (posterior median slope variance and 95% HPDI: 1.32 [0.53, 1.89]), but not G×E (0.10 [0.00, 0.67]; unpublished results), both estimated with substantially lower precision (the zero constraint of the estimates notwithstanding). The HV and VL datasets are quite similar in terms of the number of individuals and observations and we should therefore have been able to detect I×E in HV if it was really there. Since the pedigree in VL is more informative ('deeper') than the one in HV because of limited immigration from the mainland (Postma and Van Noordwijk 2005a; Gienapp et al. 2013b), the VL data may be better suited to test for G×E, but even there G×E appeared to be limited. (No caterpillar biomass data are available for that population, making it an unsuitable dataset for the purpose of this paper.) This apparent absence of G×E is consistent with the notion that G×E in general is hard to detect in wild populations when not in an experimental setting (Gienapp and Brommer 2014; Wood and Brodie III 2016; Hayward et al. 2018; Chapter 9). Nevertheless, since I×E is the upper limit of G×E (Gienapp and Brommer 2014), the absence of I×E in our study system in HV suggests that G×E is absent. The reason for this absence may be that early-spring temperatures have historically been highly predictive of the food peak and selection for being well matched with this peak has been strong (Visser et al. 2006; Reed et al. 2013b),

possibly eroding (genetic) variation in plasticity in laying date over time (cf. Tomkins et al. 2004; Turelli and Barton 2004).

Despite selection on, and ample additive genetic variation in, the elevation of the reaction norm (Tables 10.2 and S10.1), the observed, period-specific reaction norms changed only marginally over the course of time, as was also predicted from the quantitative genetic model (eqn. 10.5b). A close look at Figure 10.1c reveals that the population is 'sliding' up and down the same reaction norm; that is, the observed temperature range has become wider and individuals are still using these temperatures to time the onset of reproduction (see also Visser et al. 2006). Clearly, given the increase in mismatch with the caterpillar peak in this population, this is not an adaptive strategy (Thomas et al. 2001; Reed et al. 2013b); the population needs to evolve a lower elevation to become better matched with the caterpillar peak. The apparent lack of such a shift is most likely explained by the low heritability of elevation, the generation time and the sex limitation of laying date, as shown by our quantitative genetic model (eqn. 10.5b). Additionally, increased variability in phenotypic optima in the last third of the study period may further hamper adaptation as early-breeding genotypes no longer consistently have a reproductive advantage.

In the quantitative genetic model, we used annual phenotypic selection gradients to predict the change in the reaction norm from one year to the next. One concern with using phenotypic (annual) selection gradients is that estimates of selection in reality reflect an environmental covariance between the trait and fitness or selection on a correlated trait (Lande and Arnold 1983; Price et al. 1988; Hadfield 2008). In our population, however, we know that such environmental bias in selection is limited and that estimates of selection at the phenotypic and genetic level are very similar (Gienapp et al. 2006; Reed et al. 2016b). Thus, the necessary ingredients for genetic adaptation—genetic variation and selection—are real and present in HV, but evolutionary change is simply too small to be detected due to the large environmental variation in laying date among years (see also discussion in Gienapp et al. 2006). Such small rates of adaptation can put a strain on population persistence in the longer run as adaptation continues to be outpaced by climate change (Gienapp et al. 2013a; Carlson et al. 2014; Visser and Gienapp in press; but see Reed et al. 2013a,b).

Many populations have thus far responded to the increased mismatch with the phenology of their food through phenotypic plasticity (Charmantier and Gienapp 2014; Merilä and Hendry 2014), but it has been predicted that this cannot be sufficient in the long run as climate change continues to disrupt synchrony between trophic levels (Visser et al. 2004a; Visser 2008; Thackeray et al. 2010; Carlson et al. 2014; Thackeray et al. 2016; Visser and Gienapp in press). Consequently, we need to know whether an absence of G×E in phenology as reported here is the general case or an exception. As pointed out above, such studies in natural populations are rare (but see Chapter 9), which is likely due to the logistical challenges of obtaining the necessary pedigree, so far constraining these studies to mainly birds and mammals. Replacing the observational, social pedigree by relatedness estimated from genetic markers has earlier been suggested (Ritland 1996; Moore and Kukuk 2002), but applications of this approach remained unsuccessful (Coltman 2005;

Garant and Kruuk 2005; Csilléry et al. 2006). However, with the advent of high-throughput, high-density genotyping to 'non-model' species, 'genomic' relatedness estimates should have become sufficiently reliable to replace pedigrees (Gienapp et al. 2017) and thereby allow us to study genetic variation in and selection on phenotypic plasticity both in labile and fixed traits in a broader range of taxa.

To date, evidence for successful evolutionary rescue through a genetic shift in the reaction norm remains rare (Merilä and Hendry 2014). One textbook example of successful evolutionary rescue is that of the great tit's most important food source, caterpillars of the winter moth (*O. brumata*); three Dutch populations have now restored the match of their hatching date with oak (*Quercus robur*) bud burst through a genetic shift in the elevation of the reaction norm (Van Asch et al. 2013). It remains unclear whether such a change will be also observable in vertebrate populations, as the methods thus far deployed have been largely insufficient to infer evolutionary change (Merilä and Hendry 2014). We provide an important step to this discussion by using rigorous statistical tools to reveal the evolutionary potential of a key life-history trait in a reaction norm context.

Acknowledgements

We are grateful to the board of the Hoge Veluwe National Park for allowing us to do fieldwork all these years, to the countless number of students, assistants, and volunteers who collected the data, and to Jan Visser and Louis Vernooij for carefully managing the database. We are grateful to Thomas E. Reed for the inspiring discussions that eventually led to this manuscript. Sam Scheiner, Michael Morrissey, and an anonymous referee provided useful comments on a previous version of the manuscript. We thank Jarrod Hadfield for his help with the model specifications. Funding for this project was provided by an ERC advanced grant (339092-E-Response) to M.E.V.

Supplementary Information S10

Content

SM1 – Predicting evolution of the reaction norm

Table S10.1. *Annual selection gradients (β_z), absolute (T) and mean-centred (T_c) mean spring temperature, and the predicted annual evolutionary changes (Δ) in the reaction norm elevation (intercept, Int) and slope (Slp), along with standard errors (SE), as a result of annual selection on great tit laying date from the second time period onward (i.e. excluding 1973–1987 because the 'initial' reaction norm was estimated over these first 15 years). Predicted changes were estimated while accounting for generation time and the sex limitation of laying date (**a**) and when not accounting for these factors (**b**). See main text and eqn. (10.5a) and (10.5b) for details.*

Year	β_z	SE_{β_z}	T (°C)	T_c	(a) Generation time and sex limitation included				(b) Generation time and sex limitation not included			
					ΔInt	SE_{Int}	ΔSlp	SE_{Slp}	ΔInt	SE_{Int}	ΔSlp	SE_{Slp}
1973	0.007	0.049	5.05	–1.84	-	-	-	-	-	-	-	-
1974	–0.010	0.028	8.65	1.76	-	-	-	-	-	-	-	-
1975	0.025	0.036	4.18	–2.71	-	-	-	-	-	-	-	-
1976	–0.018	0.021	5.96	–0.93	-	-	-	-	-	-	-	-
1977	–0.045	0.035	5.11	–1.78	-	-	-	-	-	-	-	-
1978	0.010	0.048	5.62	–1.27	-	-	-	-	-	-	-	-
1979	–0.006	0.022	5.84	–1.05	-	-	-	-	-	-	-	-
1980	–0.067	0.041	6.06	–0.83	-	-	-	-	-	-	-	-
1981	–0.028	0.040	8.84	1.95	-	-	-	-	-	-	-	-
1982	–0.074	0.034	6.43	–0.46	-	-	-	-	-	-	-	-
1983	0.000	0.073	6.47	–0.42	-	-	-	-	-	-	-	-
1984	0.037	0.744	4.71	–2.18	-	-	-	-	-	-	-	-
1985	–0.204	0.084	6.53	–0.36	-	-	-	-	-	-	-	-
1986	–0.036	0.109	4.79	–2.10	-	-	-	-	-	-	-	-
1987	–0.165	0.060	6.14	–0.75	-	-	-	-	-	-	-	-
1988	–0.191	0.132	7.10	0.21	–0.070	0.130	–0.001	0.003	–0.792	1.470	–0.013	0.038
1989	–0.032	0.043	7.30	0.41	–0.024	0.030	0.000	0.001	–0.153	0.192	–0.003	0.007
1990	–0.079	0.031	7.60	0.71	–0.068	0.033	–0.001	0.003	–0.358	0.176	–0.006	0.016
1991	0.006	0.043	8.69	1.80	0.007	0.023	0.000	0.001	0.056	0.174	0.001	0.005
1992	–0.059	0.074	6.58	–0.31	–0.041	0.200	–0.001	0.003	–0.205	0.992	–0.003	0.015
1993	–0.165	0.081	7.52	0.63	–0.131	0.078	–0.002	0.006	–0.693	0.413	–0.012	0.030
1994	–0.200	0.349	6.42	–0.47	–0.228	0.551	–0.003	0.010	–0.878	2.122	–0.011	0.040
1995	–0.078	0.073	6.05	–0.84	–0.048	0.062	–0.001	0.002	–0.239	0.309	–0.003	0.008
1996	–0.118	0.089	5.31	–1.58	–0.146	0.121	–0.001	0.004	–0.524	0.435	–0.004	0.013
1997	–0.075	0.088	6.75	–0.14	–0.086	0.277	–0.001	0.005	–0.309	0.994	–0.004	0.018
1998	–0.188	0.075	7.55	0.66	–0.203	0.274	–0.004	0.010	–0.827	1.119	–0.015	0.040
1999	–0.052	0.087	8.05	1.16	–0.050	0.550	–0.001	0.011	–0.235	2.578	–0.005	0.052
2000	–0.092	0.026	7.18	0.29	–0.112	0.048	–0.002	0.004	–0.364	0.156	–0.006	0.014
2001	–0.114	0.037	5.72	–1.17	–0.096	0.038	–0.001	0.003	–0.466	0.185	–0.005	0.013
2002	–0.008	0.027	7.42	0.53	–0.004	0.034	0.000	0.001	–0.013	0.110	0.000	0.002
2003	–0.151	0.063	7.26	0.37	–0.178	0.088	–0.003	0.007	–0.698	0.347	–0.012	0.028
2004	–0.023	0.030	8.02	1.13	–0.025	0.042	0.000	0.001	–0.078	0.133	–0.002	0.004
2005	–0.055	0.047	9.53	2.64	–0.059	0.063	–0.002	0.004	–0.223	0.240	–0.006	0.014

Continued

Table S10.1 *(continued)*

Year	β_z	SE_{β_z}	T (°C)	T_c	(a) Generation time and sex limitation included				(b) Generation time and sex limitation not included			
					ΔInt	SE_{Int}	ΔSlp	SE_{Slp}	ΔInt	SE_{Int}	ΔSlp	SE_{Slp}
2006	–0.111	0.060	6.02	–0.87	–0.125	0.070	–0.001	0.004	–0.470	0.265	–0.005	0.014
2007	–0.205	0.069	9.27	2.38	–0.206	0.101	–0.005	0.012	–0.918	0.450	–0.023	0.052
2008	–0.119	0.050	5.50	–1.39	–0.127	0.067	–0.001	0.003	–0.461	0.243	–0.004	0.012
2009	–0.044	0.052	9.30	2.41	–0.050	0.059	–0.001	0.003	–0.195	0.231	–0.005	0.012
2010	–0.003	0.039	8.49	1.60	–0.001	0.050	0.000	0.001	–0.002	0.181	0.000	0.004
2011	–0.082	0.032	9.05	2.16	–0.092	0.050	–0.002	0.005	–0.325	0.177	–0.008	0.018
2012	0.011	0.045	7.45	0.56	0.012	0.039	0.000	0.001	0.062	0.203	0.001	0.004
2013	–0.095	0.060	3.64	–3.25	–0.078	0.052	0.000	0.001	–0.392	0.260	0.000	0.003
2014	–0.092	0.046	9.72	2.83	–0.067	0.047	–0.002	0.004	–0.359	0.253	–0.010	0.022
2015	–0.004	0.045	6.81	–0.08	0.007	0.037	0.000	0.001	0.038	0.190	0.001	0.003
2016	–0.057	0.032	7.60	0.71	–0.053	0.031	–0.001	0.002	–0.281	0.164	–0.005	0.012
Total	**-**	**-**	**-**		–2.341	0.948	–0.037	0.027	–10.304	4.248	–0.164	0.122

S2 – Simulation: testing the effect of the residual variance structure on detecting I×E

The lack of between–individual variation in reaction slopes (I×E) in the Hoge Veluwe (HV) population contradicts previous studies (Nussey et al. 2005b; Husby et al. 2010; Husby et al. 2011b). Here, we show by simulation that specifying the residual variance structure of the random regression model incorrectly will lead to wrong inferences about the presence of I×E (for simplicity, we disregard G×E here, but the same reasoning applies).

The model

We partly used the posterior estimates from the random regression animal model presented in Table 10.3 (main text) as basis for the simulation. We simulated a population of 1000 individuals with each roughly 2.6 observations (which is the mean no. observations per individual in HV when disregarding females that bred only once), distributed over 44 environments (years). We tested the effect of different residual structure on three scenarios, i.e. with small (0.1), intermediate (0.5) and large (1) variation in reaction norm slopes.

First, we drew 44 random temperatures from a normal distribution based on real temperature data (mean = 6.9, sd = 1.5). We then randomly drew a number of observations for each individual from a Poisson distribution, such that the mean per individual approximated 2.6. We then randomly assigned each individual to a cohort (and hence the temperatures they were exposed to). We randomly assigned a reaction norm to each individual by drawing an intercept (a_i) and a linear slope (b_i) from a random, normal distribution with mean = 0 and σ^2 = 3.5 and either 0.1, 0.5 or 1, respectively. Temperature values were individually mean–centred (T_i), and phenotypes (laying date or ELD) were derived as $ELD_{ij} = a_i + b_i T_{ij} + bT_{av,i} + e_{ij}$, where $T_{av,i}$ is the average temperature experienced by the individual and e_{ij} is the error term. The error term was randomly drawn from a normal distribution with mean = 0 and σ^2 = 10.8, 13.4, 19.2 or 19.9, depending on the temperature in that environment; these values were taken from Table 10.3 (main text) and were used to make variance in ELD dependent on temperature, as is the case in our great tit population.

For each slope variance scenario (small, intermediate, or large), we fitted three mixed–effects models in ASReml–R (Butler et al. 2009; Gilmour et al. 2009), each a variation on *LD ~ individual–centred temperature + individual–mean temperature, random = Individual×individual–centred temperature* (i.e. 9 scenarios): (i) with residual variance partitioned into four temperature blocks, i.e. years equally divided based on temperature; (ii) residual variance partitioned into four 'decadal' blocks, i.e. based on consecutive years; and (iii) homogeneous residual structure. Each model was tested against a simpler model (*LD ~ individual–centred temperature + individual–mean temperature, random = Individual*) to test for significance of I×E using likelihood–ratio tests with 1 degree of freedom. Starting values in ASReml–R were set such that they matched the input values.

The whole procedure was iterated 1000 times. The R script has been uploaded as a separate text file.

Results and discussion

As expected, slope variance estimates matched the input values nicely when we used heterogeneous residual variance based on temperature blocks in each of the three slope variance input scenarios (Figure S10.1a, c and e). When slope variance was small (Fig S10.1a), specifying the wrong residual structure inflated the estimates; with larger input values, however, this bias largely disappeared (panels c and e). Again as expected, power to detect I×E at a low slope variance was limited when using the appropriate residual structure, but strikingly, specifying the wrong residual structure led to large false positive rates (Fig S10.1b). Again, this discrepancy between models disappeared as true slope variance increased (panels d and f).

We conclude that specifying the right residual structure is essential for making correct inference of the presence of I×E (or G×E). This in itself is not a new insight (e.g. Gienapp and Brommer 2014), but it stresses the importance of carefully assessing which parameter drives variation in a phenotype. In the great tit example, this is clearly temperature and not year as a proxy for temperature. When true variance in slopes is small, therefore, an incorrectly specified residual structure will lead to both quantitatively and qualitatively different results (i.e. whether or not there is I×E). When variance is substantial, however, the chance of making a qualitatively (and perhaps quantitatively) wrong inference may in fact be reasonably small.

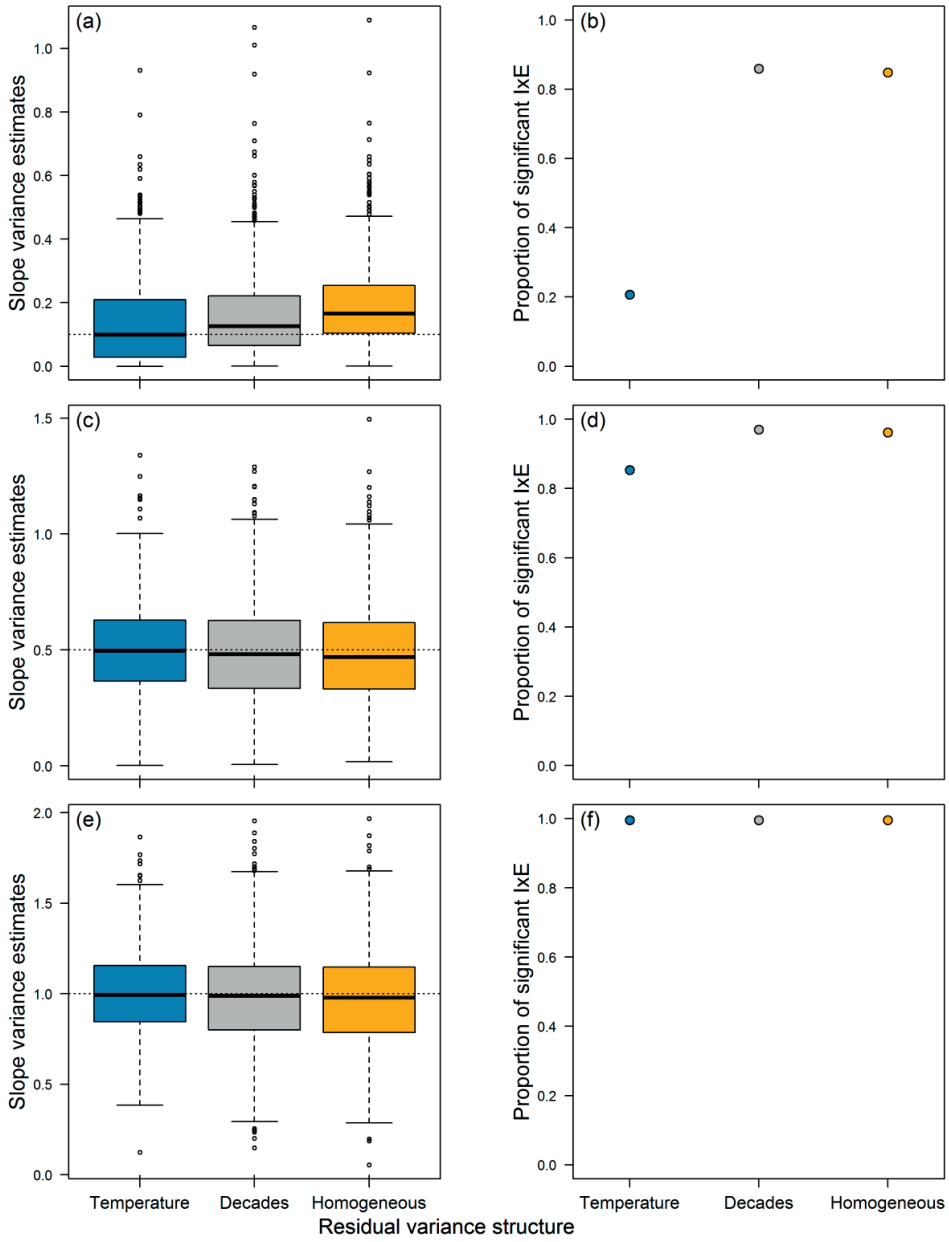


Figure S10.1. *Results of the simulation testing the effect of the residual structure of the random regression model on the estimate of variance in slopes (a, c, e) and the respective statistical power to detect this variance (b, d, f). Input values for slope variance are denoted by the horizontal, dotted lines.*

Chapter 11

General discussion

Many traits of interest to evolutionary ecologists, ranging from life history and behaviour to more 'tangible' traits related to morphology and physiology, show heritable genetic variation (Postma 2014). This allows us to do quantitative predictions of evolutionary trajectories, provided that selection is acting on this variation—and that we are able to correctly estimate it. The step from quantifying heritable variation and selection to adequate predictions of evolution is, however, not straightforward, and many attempts to demonstrate that selection on genetic variation has led to evolutionary change have led to erroneous conclusions (Merilä et al. 2001b). Quantitative methods are, however, improving, with an increasing number of studies that aim to provide tangible guidelines for evolutionary ecologists (e.g. Postma 2006; Hadfield et al. 2010; Morrissey et al. 2010, 2012; Morrissey and Liefting 2016). But to understand evolutionary processes in the wild, we first and foremost need a thorough understanding of the system we are working on and identify the ecological factors that are the drivers of selection.

In this thesis, I address how populations are coping with environmental change and which ecological processes potentially affect the rate of adaptation. To answer this broad question, I broke down my thesis into several parts. I first set out to explain the observed consequences of global climate change on bird biology and the possible impacts in the future. In the second part of my thesis, I used a combination of experiments and long-term data to study constraints in adaptation to climate change in a population of great tits. In the final part of the thesis, I used advanced quantitative genetic tools and simulations to study outstanding questions revolving around the evolutionary potential of wild vertebrate populations. Combined, these studies should improve our general understanding of the phenotypic and genetic variation in populations, the selective forces acting on this variation and how ecological factors may (or may not) modulate plastic or genetic responses. Related to this part, in the Intermezzo I explored existing and novel methodologies and provided guidelines for students of ecology and evolution to improve our scientific approaches to study adaptation in the wild.

In this chapter I aim to bring all these results together and to discuss them in the light of what we know about the ecological and evolutionary processes not only in great tits, but in wild populations in general. I will also reflect on the methods that I have used and provide my ideas about how I think the field of ecology and evolution in long-term population studies should move forward from hereon.

Coping with a warming world

This thesis, as many of its predecessors in the field of evolutionary ecology, was sparked by the observation that the environment is changing in many aspects under human influence (Rockström et al. 2009; Steffen et al. 2015; Scheffers et al. 2016). Climate change is one of the most pervasive forces driving selection in natural populations on a global scale (Sala et al. 2000). In **Chapter 2** of this thesis, I explored the particular case of birds. With an estimated 18,000 species on this planet (Barrowclough et al. 2016) it is impossible

to estimate the biological, ecological and evolutionary effects of climate change on every species, but it will be clear that the effects—in the great variety in which they come—will depend strongly on geography (e.g. the degree of seasonality) and altitude, the extent of warming or climatic variability, ecological niche, morphology and physiology, the degree of (geographical) isolation, migration status, and many other factors. Importantly, non-migratory species that cannot evade the selective pressures of the environment in which they reside need to cope with changing conditions by either plastically changing their phenotype or by adapting genetically toward the new optimum (Parmesan 2006). From **Chapter 2** it becomes clear that what we know about the ecological and evolutionary consequences of climate change is still limited. We know, for example that birds can shift their geographical distributions as (a)biotic conditions change (Thomas and Lennon 1999; Hitch and Leberg 2007; Huntley et al. 2008), but the majority of data are from sightings reported from e.g. citizen science projects or constant-effort monitoring sites that lack individual-level fitness data. We still lack a clear understanding of which factors drive range expansion or shift, if this expansion or shift is adaptive, and how it affects community compositions and species interactions. When species shift 'successfully' they may be faced with a plethora of novel challenges, e.g. because temperature warming trends vary between seasons (Easterling et al. 1997) or their food cannot shift at a similar rate (Van der Putten et al. 2010). Transplant experiments, where some species are introduced into novel habitat, combined with long-term monitoring, may provide useful insights into the adaptive value of range expansion or shifts under climate change (Hargreaves et al. 2013).

A general pattern in the study of the ecological and evolutionary effects of climate change in natural populations is that there is (1) a strong (biased) focus on phenology (the seasonal timing of biological events) and (2) a strong bias toward certain taxa. The focus on phenology is understandable for obvious reasons, such as that it is the first aspect of change we typically see in long-term monitoring projects: the first bird eggs found increasingly early in spring (Crick et al. 1997), advanced spawning dates in anurans (Beebee 1995), or advanced avian migration (Gienapp et al. 2007). It is also a very important subject of study since phenology, like other life-history traits, is closely related to fitness, as survival and reproductive success in seasonal organisms is often tightly linked to the phenology of other organisms (Cushing 1969; Durant et al. 2007). Reported changes in phenology due to climate warming are ubiquitous across taxa (Parmesan 2006; Thackeray et al. 2010; Thackeray et al. 2016; Kharouba et al. 2018). A remaining challenge, however, is to get a broader understanding of whether these phenological changes are adaptive (Merilä and Hendry 2014) and whether they are plastic or genetic changes (Gienapp et al. 2008; Charmantier and Gienapp 2014). Phenological mismatch due to differential shifts among trophic layers are often reported (e.g. Visser and Both 2005; Kharouba et al. 2018) but only rarely have they been directly linked to selection on phenology (Visser and Gienapp in press). The case study of the great tit at the Hoge Veluwe (Visser et al. 1998) is a classic one and is complemented by a few others (e.g. Plard et al. 2014; Bowers et al. 2016; Marrot et al. 2018), although some of these used mere proxies (e.g. spring temperature) for mismatch. One obvious reason for the lack of demonstrated associations between phenological mismatch and selection is the logistical challenge

associated with either getting decent estimates of selection (survival or recruitment of offspring needs be monitored) or quantifying phenological (mis)match. While the latter issue is a subject of discussion (Miller-Rushing et al. 2010; Lindén 2018), I aimed to provide a guideline in **Chapter 5** for ecologists as to how measure phenological match in highly seasonal environments (see next section).

The second pattern observed in studies of phenology is that there is a severe taxon bias. To reliably study the evolutionary consequences of climate change, we need to be able to separate plastic from genetic responses, and this can only be done when we have phenotypic data for individually marked individuals that can be tracked across multiple phenotypic events (Clutton-Brock and Sheldon 2010). This naturally confines long-term studies to mainly birds and mammals. As we have seen in **Chapter 9**, from the 106 datasets retrieved from open data repositories there was a severe bias towards birds (50) and mammals (31). Much of the evolutionary work on phenological shifts due to climate change stems from study populations that happened to have been studied for various purposes before the effects of climate change became apparent in the first place. These populations are a valuable source of information, but it is clear that if we are to truly comprehend the effects of phenological shifts on selection and, ultimately, demography and population dynamics, we need to expand our research efforts beyond the typical 'evolution model species' such as great tits and include different taxonomic groups (cf. e.g. Phillimore et al. 2012). Such efforts, however, will be costly and take time to build up before they are useful for evolutionarily relevant questions (Clutton-Brock and Sheldon 2010).

Fitness consequences of avian reproductive timing

Identifying constraints in adaptation

The proximate and ultimate causes of seasonal reproductive timing in birds have long intrigued evolutionary ecologists. It started back in the previous century when David Lack (1950) postulated that seasonal timing in Europe's birds is an adaptation to the seasonal availability of food for their offspring. Laying date varies, however, strongly between individuals and years within populations. Perrins' (1970) seminal paper argues that the timing of breeding for each individual female is a result of the food conditions she experiences in the pre-laying period, and differences between females may reflect differences in 'quality' (e.g. body condition or physiological state) that enable high-quality birds to attain the best breeding territories (see discussion in Verhulst and Nilsson 2008). Birds can therefore not breed in early-spring harsh conditions when sufficient resources cannot be attained, likely not because of a 'physiological boundary' (Perrins 1970), but because of severe fitness (e.g. survival) costs of breeding under adverse conditions (Monaghan et al. 1998; Visser and Lessells 2001). As the climate changes and food availability peaks earlier (Visser et al. 1998), we need to know whether birds can advance with it. This was the subject of **Chapters 3** and **4**.

In **Chapter 3,** I specifically explored the *constraints* hypothesis of reproduction posited by Visser et al. (2012). In the Hoge Veluwe great tit population we see an increased mismatch between timing of the caterpillar biomass peak date and the timing of the food demands of the chicks (Visser et al. 1998; Visser et al. 2006; **Chapter 10**). Under the constraints hypothesis, birds cannot advance their laying date sufficiently to restore the mismatch, possibly because temperatures during the egg-laying period have not advanced as much as the temperatures relevant for the caterpillar phenology later in the season (Visser et al. 2006). It has therefore been suggested that it is adaptive to be slightly mismatched with the food peak if that entails relaxed survival costs of breeding early (Lof et al. 2012; Visser et al. 2012). This is, however, a difficult hypothesis to test. Numerous experiments have attempted to advance laying date to test the effect of female quality vs. timing per se on the seasonal decline of reproductive success (Verhulst and Nilsson 2008). An inevitable result of many methods to advance, such as supplemental feeding (to induce early laying) and cross-fostering clutches (to advance hatch dates), is that the condition of the female is likely affected. I will therefore argue in this section that feeding experiments alone cannot provide us with a clear-cut test of the hypothesis, nor can single years of any experiment.

The idea of the experiment in **Chapter 3** was to advance laying through supplemental feeding and cease feeding upon the start of egg-laying in a subgroup; if the treatment was successful and birds were laying under adverse conditions, cessation of feeding would confer a fitness cost, because these birds would have to complete costly laying (and incubation) (Visser and Lessells 2001) without sufficient resources. On the other hand, birds that continued to be fed would enjoy fitness benefits of being better matched with the food peak because supplemental feeding continued during egg-laying. There are a few assumptions that need to be met for experiments like this one to work: (1) supplemental feeding should be effective at advancing laying date substantially compared to control (unfed) birds, e.g. by roughly a week (since the food peak is generally a few weeks 'wide' (see **Chapter 5**) and therefore shifts need to be sufficient to have measurable fitness effects); (2) advanced birds should meet genuinely adverse conditions; and (3) there should be no substantial carry-over effects of the supplemental feeding into the laying/incubation phase. In **Chapter 3**, I suggested that conditions were poor regardless of whether birds laid early or late and showed that supplemental feeding was not successful at advancing birds. In fact, any supplemental feeding conferred a fitness advantage, likely because of a carry-over effect to the chick-feeding phase. In a way, this is another way of testing the hypothesis: now all birds are 'constrained' (the average number of fledglings in 2015 was 2.7 ± 0.3 SE, whereas the average over 2010–2017 was 5.0 ± 0.1 SE) and supplemental feeding lifts this constraint. This constraint is, however, not directly related to constraints on reproductive timing. I conclude from this that multiple years of experiments are necessary to (1) test whether supplemental feeding—ceased upon the start of egg-laying—genuinely confers a fitness disadvantage under poor (pre-)laying conditions, and (2) to properly judge the carry-over effects of feeding through to the chick-rearing phase in a range of environmental conditions.

Feeding experiments like the one in **Chapter 3** are likely not sufficient to test whether birds are constrained to breed earlier if they fail to provide evidence for a lack of carry-

over effects. In **Chapter 4**, I described the first results of a large selection experiment that aims to 'cleanly' manipulate laying date. Eggs from birds genomically selected for their laying date (early and late) in aviaries were fostered in the wild and the phenotype (laying date) of the resulting recruits in the following year were recorded. Such an experiment is necessary because the experimental advancement in laying date is not contingent on changing the condition of the female. We only have little data for these selection-line birds in the wild, but tentative results suggest that the early and late selection line birds differ in their realised laying dates in the expected direction. Sample size notwithstanding, and given that the fostering of selection-line eggs will continue for several more breeding seasons, the results are quite promising, since they suggest that the genomic selection experiment was effective at diverging the selection lines. Genomic selection, i.e. on genomically estimated breeding values (GEBV; Gienapp et al. 2019) typically results in faster phenotypic responses across generations than conventional selection based on the phenotype (Meuwissen et al. 2016). The expected phenotypic response to genomic selection (using the breeder's equation) was quite substantial given the low heritability of laying date (typically ~0.2), leading to an expected divergence of ~2.5 days and a realised divergence of ~6 days in laying dates in aviaries between the selection lines over just three generations (Verhagen et al. in review).

The hypothesis in **Chapter 4** is that birds that breed earlier than the natural population have reduced fitness (e.g. through reduced survival) if there are in fact constraints to early reproduction, but is difficult to predict (1) by how much birds need to advance for there to be fitness costs and (2) whether we can reasonably expect birds to achieve that advance. Gienapp and Visser (2006) manipulated food availability during breeding in the previous year in an attempt to alter the decision of the females in the current year, and found that manipulated females advanced their laying date by about 6 days. Unfortunately, sample sizes were too small to assess the fitness consequences of this advance. Interestingly, the reported change of 6 days by Gienapp and Visser (2006) coincides with the reported difference in the aviaries discussed above (Verhagen et al. in review), as does the observed difference between the recruited selection-line birds in the wild (**Chapter 4**). Moreover, Visser et al. (2009a) showed that laying dates from females measured in the wild and in aviaries correlate well. This may suggest that the divergence of six days found in the aviaries is what we could reasonably expect to see as we continue to bring selection-line birds into the wild. Given that six days is roughly one standard deviation of the within-year variance in laying date at the Hoge Veluwe, such a divergence could have strong fitness implications, but such effects would likely only become apparent in years with poor food conditions (e.g. Nager et al. 1997).

I conclude this section with the realisation that (1) feeding experiments may only be helpful in testing the constraints hypothesis of reproduction when the manipulation of laying date is successful, carry-over effects are known, and the effect of multiple environmental conditions (years) is tested, and that (2) genomic selection for earlier (and later) breeding can provide a promising test provided the experiment is carried out for a sufficient number of years (Vaughn and Young 2010). Provisioning rates (**Chapter 3**) and measures of daily energy expenditure during chick feeding (Te Marvelde et al. 2011) will further help us identify the proximate causes of reproductive success. As of yet, we have

no clear answer to the question of whether great tits are in fact constrained to breed earlier because of the associated fitness costs.

Quantifying phenological mismatch as a driver of selection: is there room for improvement?

In the previous section I discussed constraints on adaptation to a warming climate, and experiments to reveal these constraints. In **Chapter 5**, I studied the ultimate driver of selection for earlier breeding, which is the advance of (peak) food availability over time. This principle, that consumer phenology should be matched with that of the resource, is formally described as the *match/mismatch* hypothesis (Cushing 1969; Durant et al. 2007; see **Chapter 2**). In many ecological studies of phenological mismatch, mismatch has been described as the difference in mean phenology or another convenient, temporal parameter such as the date of first occurrence (e.g. Kharouba et al. 2018), including our own work on the great tit (Visser et al. 1998; Visser et al. 2006; Reed et al. 2013b; Chapter 10). Durant et al. (2007) noted for marine study systems, however, that the temporal overlap is just as important because it determines whether organisms really are mistimed. Miller-Rushing et al. (2010) and Lindén (2018) illustrated this point nicely from a theoretical point of view. **Chapter 5** aimed to answer the question of whether our quantification of mismatch is in fact appropriate. My conclusion was that, with the methods at hand, the 'classic' measure of mismatch (match of peak dates in phenology) performs better (or at least just as well) than a measure of temporal overlap between phenological distributions in models explaining variation in offspring recruitment and selection differentials for laying date.

The conclusions from **Chapter 5** deserve a mention here because they will have implications for how other long-term studies of predator–prey interactions may or should be designed. Visser and Gienapp (in press) identified only a handful of studies that have linked selection on phenology directly to phenological mismatch, likely because only few researchers have access to such data. When such data are available, they are generally not detailed enough to allow for a careful description of the temporal distribution of food resources *sensu* Durant et al. (2007). For the great tits in the Hoge Veluwe, we have fairly accurate description of season-wide food availability, as do a few other (avian) research groups (e.g. Cresswell and McCleery 2003; Vatka et al. 2011). As I pointed out in **Chapter 5**, these data can never be sufficient to accurately estimate the phenological overlap between consumer needs and food availability because it requires making nontrivial assumptions about the interaction between consumer and prey. First, to assess whether the consumer-needs barrier is exceeded, we would need to know exactly how much food is available in space and time—which prey-density estimates such as mass per m^2 cannot provide because it assumes that they can simply be multiplied by the total available area, an assumption that is unlikely to be true. More importantly, the consumer (e.g. a food-provisioning great tit) may find and keep returning to a single tree that is teeming with prey, making prey availability in the remainder of the forest irrelevant. Lastly, and perhaps most importantly, an experiment in great tits showed that prey-encounter rates were not proportionate to prey densities and that these rates increased by 72% (and not 100%) with a doubling in prey densities (Mols et al. 2004). Mols et al. (2004) showed, moreover, that encounter rate decreased when a caterpillar-laden tree had been previously

exploited by other great tits, either because the caterpillars responded to the previous encounter by hiding or because the most conspicuous caterpillars had already been removed. This shows that the relationship between prey availability and prey-encounter rates is far from straightforward, and therefore even the most accurate estimates of food availability in the wild will likely not help us quantify the temporal overlap between the consumer and the effectively available prey.

Naturally, I underline the notion that the complete (temporal) interaction between consumer and resource is what really determines the effective phenological (mis)match (Miller-Rushing et al. 2010; Lindén 2018), but I caution that quantifying this interaction will in most cases be logistically very challenging at the very least. With funding already often being the limiting factor in keeping long-term population studies going (Clutton-Brock and Sheldon 2010), I would argue that budgets and manpower may best be invested elsewhere, e.g. sampling different food sources at different parts of the season (e.g. in the pre-egg-laying period) or by studying the phenology of multiple trophic layers to gain a full understanding of the ecosystem that we are studying.

Predicting evolution in the wild

The backbone of this thesis is the quantifying of evolutionary processes in wild populations (Part III). The study of evolutionary parameters in wild populations has long intrigued evolutionary biologists since the pioneering work on heritability of avian life history (Perrins and Jones 1974; Van Noordwijk et al. 1980) and morphology (Boag and Grant 1978), and since animals models were brought into prominence by evolutionary ecologists in the early 2000s (Kruuk 2004), the number of heritability estimates in the wild have soared (Postma 2014). At the same time, evolutionary questions have become more complex and have gone beyond simply estimating additive genetic variation in traits to questions such as "How does genetic variation vary with the environment?" and "Are populations genetically adapting in response to global warming?". Both questions appear non-trivial and the methods applied in an attempt to answer these questions have long been insufficient to warrant any biologically meaningful conclusions in many studies (e.g. Gienapp et al. 2008; Hadfield et al. 2010; Merilä and Hendry 2014). My primary aim for Part III of this thesis, as well as the preceding chapters in the Intermezzo, was to make predictions of evolutionary trajectories, not for the sake of the study species at hand, but with a focus on how this was achieved. In a way, these chapters therefore balance on the interface of evolutionary biology and statistical methodology, for the sake of making evolutionary predictions more reliable.

Improving our predictive methods in ecology and evolution.

A recurring problem in evolution studies of wild populations is that predictions do not match observations (Merilä et al. 2001b). In practice this means that the application of the breeder's equation (Lande and Arnold 1983; Falconer and Mackay 1996), which was

originally designed in the field of animal breeding, does not yield reliable estimates of evolutionary response in wild populations (Morrissey et al. 2010). For the breeder's equation to work, we need reliable estimates of genetic variation (or heritability) and selection. This may be problematic, for example, if we fail to account for the fact that genetic variation (Hoffman and Merilä 1999) or genetic covariances (Wood and Brodie III 2015) change with the environment, selection varies with the environment (Hairston and Dillon 1990; Gosden and Svensson 2008b; but see Morrissey and Hadfield 2012) or estimates thereof are misleading (Morrissey et al. 2010; Bonnet et al. 2017), or a combination of varying selection and genetic variation (Wood and Brodie III 2016).

With the advent of animal models in evolutionary biology came also the realisation that long-term population studies can be used to quantify variation in reaction norms (e.g. Postma and van Noordwijk 2005b), and in some cases whether this variation has a genetic basis (G×E). Pioneering studies on avian and mammalian populations (Brommer et al. 2005; Nussey et al. 2005a,b) used mixed-modelling tools to obtain individual estimates of plasticity but did this using methods (analysis of Best Linear Unbiased Predictors obtained *a posteriori* from the models) that are now considered inappropriate (Hadfield et al. 2010). The idea behind it was, however, appealing: quantify components of individual reaction norms, assess (phenotypic) selection on these components, and use animal models to estimate genetic variation therein and assess whether this genetic variation changes over time or an environmental gradient. We have come a long way from these studies and alternative methods for estimating I×E/G×E and selection on reaction norm components have been advocated (e.g. Nussey et al. 2007; Brommer et al. 2012).

There are, however, two major observations that I make from the evolutionary ecology literature (more specifically related to I×E/G×E). First, although much attention has been devoted on the use of mixed-modelling approaches to quantify variation in behaviour and life-history in the wild (e.g. Nussey et al. 2007; Van de Pol and Wright 2009; Van de Pol 2012; Dingemanse and Dochtermann 2013), few studies actually address the effect that the residual variance structure can have on the accuracy of the prediction of variation in reaction norms. That residual variance contains important biological information is well recognised (e.g. Westneat et al. 2015), but surprisingly few studies of phenotypic plasticity have incorporated this in their analysis (Nicolaus et al. 2013). In **Chapter 6**, I ran simulation to show that, particularly when sample sizes are limited, heteroscedasticity can lead to erroneous conclusions about the presence an extent of I×E (and hence also G×E). This analysis revealed the obvious, but surprisingly enough, not many ecologists have been aware of the magnitude of the problem (Nicolaus et al. 2013), nor are there any clear guidelines as to how to tackle this problem in an intuitive way (but see e.g. Cleasby and Nakagawa 2011). Likely, many (if not most) ecologists dealing with this type of evolutionary questions are not statisticians by training and may not naturally be exposed to the dense, equation-ridden animal-breeding literature from which much of the statistical methodology is derived. Practical guidelines will therefore be useful because evidently the way heteroscedasticity is treated can radically change our conclusions about the natural world (see **Chapter 10**).

The second observation is that some particular questions have been only very rarely studied, even though they are commonly postulated to be of potential evolutionary importance. A clear example is that of the environmental coupling between genetic variation and selection, which may impair or enhance adaptation depending on the direction and the strength of the association between them (Wood and Brodie III 2016). Meta-analyses allow us to synthesise the results of multiple studies in a single analysis, but this only works when effect sizes (along with standard errors and sample sizes) are available, which in this particular case was true for only two studies (Wilson et al. 2006; Husby et al. 2011b). Ironically, the results of these two studies probably reflected statistical artefacts more than any biologically meaningful phenomenon (see discussion in **Chapter 9**). In **Chapter 9**, I specifically assessed the presence and the evolutionary implications of a coupling between heritability and selection in a variety of traits from 10 different species in 16 populations and verified indeed that, for the case of the great tit (Husby et al. 2011b), such coupling likely does not exist (I could not access the data for the other study, which involved Soay sheep *Ovis aries* (Wilson et al. 2006), but the results of a recent re-analysis on that study population suggests indeed that such a coupling does not exist in Soay sheep either (Hayward et al. 2018)). **Chapters 6**, **7** and **9** are therefore complementary to one another in that they, respectively, (1) provide guidelines to correctly identify I×E/G×E, (2) explain how outstanding questions can be answered at a (potentially very) broad taxonomic scale by using data that are openly available, and (3) provide a worked example of the use of such open data to show that the perceived evolutionary significance of a coupling between selection and genetic variation is likely overestimated.

Towards better prediction: combining data sources and the importance of understanding the ecology

A common feature of the chapters in Part III of this thesis is the attempt to integrate as many sources of information as possible to (1) achieve accurate evolutionary predictions and (2) gain a full understanding of the study system at hand. This is particularly useful in light of the observation that estimates of (phenotypic) selection may be misleading (e.g. Morrissey et al. 2010; Bonnet et al. 2017) and the more general finding that most longitudinal studies that report a phenotypic change cannot unequivocally prove that this change is (partly) due to micro-evolution (e.g. Charmantier and Gienapp 2014; Merilä and Hendry 2014). As Merilä and Hendry (2014) pointed out, we need a combination of methods and use common sense to be able to make reliable evolutionary predictions—and to understand them. Paraphrasing this, it means that we need to truly understand the ecology of the species and integrate as many relevant sources of information as possible.

In **Chapter 8**, I ventured away from phenology and used clutch size as a trait to study whether a relatively understudied ecological force (Räsänen and Kruuk 2007), maternal effects experienced in the rearing environment (*sensu* Falconer 1965), could affect the rate of adaptation to novel environments. This chapter differed from all other chapters in that I used individual-based models to predict evolutionary trajectories forward at an ecological timescale. To achieve this, I parameterised the model with data from our long-term study population of great tits at the Hoge Veluwe. I found that maternal clutch size affected offspring fledging weight, and that offspring from large clutches in turn produced

smaller clutches. Importantly, estimating the pathway between maternal clutch size and fledgling weight was possible only because we were able to link observational with experimental data; because each female is believed to optimise her clutch size, the decline in fledgling weight as maternal clutch size increases can only be demonstrated with experimental data (Tinbergen and Daan 1990; Pettifor et al. 2001; but see Tinbergen and Both 1999). **Chapter 8** illustrates that in order to keep evolutionary predictions as realistic as possible, in certain cases a combination of correlational and experimental data, as well as predictive modelling exercises, is needed that can help us test and develop new hypotheses that are relevant to evolutionary ecologists—and not only to theoretical biologists (e.g. Hoyle and Ezard 2012; Ezard et al. 2014; Kuijper and Hoyle 2015).

In **Chapter 10**, I quantified phenotypic selection on and additive genetic variation in great tit laying-date plasticity to better understand the evolutionary dynamics of reaction norms. I found that, although there was phenotypic and additive genetic variation in the elevation of the reaction norms, this was not the case for the slope. I also found that there has been selection on this elevation due to a shifting caterpillar peak. However, I predicted quantitatively that there was (or could be) no substantial response to selection on the elevation or slope of the reaction norm; in the case of the elevation this was because of large environmental variation in laying dates and in the case of slope this was because of lack of additive genetic variation and selection. It was possible to study selection on and the evolution of reaction norms largely because (the evolutionary parameters of) the phenologies of the great tit (e.g. Lack 1950; Van Balen 1973; Van Noordwijk et al. 1981; Visser et al. 1998; Cresswell and McCleery 2003; Nussey et al. 2005b; Gienapp et al. 2006; Visser et al. 2006; Charmantier et al. 2008; Reed et al. 2013b; Reed et al. 2016b) and that of other important components of the food chain (winter moth *Operopthera brumata* and its main host plant, oak *Quercus robur*) (Buse et al. 1999; Visser and Holleman 2001; Van Asch et al. 2013; Salis et al. 2018) are exceptionally well known. We know, for example, that estimates of phenotypic selection on laying date in the Hoge Veluwe are largely in agreement with those estimated at the genetic level (Gienapp et al. 2006; Reed et al. 2016b). We also know that the caterpillar peak date is an important driver of selection (Visser et al. 1998; Visser et al. 2006; Reed et al. 2013b; Chapter 5). This allowed us to get reliable estimates of selection on the reaction norm and at the same time, as a reality check, compare it to our expectations based on the association between the temperature cue and the peak date of caterpillars. Previous studies on this population suggested there was additive genetic variation in and selection on plasticity slopes (Nussey et al. 2005b; Husby et al. 2011b). We now know this is not the case and we can explain why it does not need to be the case (**Chapter 10**).

A forward look

Some thoughts and recommendations for longitudinal evolutionary studies

Long-term population studies are a vital asset in the study of the evolutionary consequences of environmental change (Visser 2008; Clutton-Brock and Sheldon 2010).

The importance of these longitudinal studies in the field of quantitative genetics and a general outlook of where the field is headed is reported in great detail elsewhere (e.g. Charmantier et al. 2014). Specific recommendations about the quantification of evolutionary parameters, such as that of selection (Morrissey et al. 2010, 2012), or the use of predicted breeding values in evolutionary ecology (Postma 2006; Hadfield et al. 2010), have also been given in great detail elsewhere and are beyond the scope of my thesis. I have made some important observations in my thesis that are in my view helpful in improving our understanding of evolutionary processes and our predictive methods. Some of these particularly pertain to the study of (genetic variation in) reaction norms, which was an important component of much of the quantitative genetic analysis in this thesis.

Studying variation in plasticity. First, in the study of variation in plasticity, I noticed that the treatment of residual ('unmeasured') variation is generally underappreciated, or at least guidelines as to how to effectively deal with heteroscedasticity are lacking (**Chapter 6**). When evolutionary ecologists use any type of regression technique, more care should be given to the potential effect of heteroscedasticity, and in **Chapter 6** I have given guidelines as to how to approach this. Briefly, a promising route would be to first test for an association between phenotypic variance and the environment in question and then fit multiple random regression models with different residuals structures and use information criteria to select the best model. Statistical power (i.e. sample sizes) will often be an issue and it is therefore recommendable to always complement empirical analysis with simulations (e.g. Johnson et al. 2015). Moreover, multiple environments may need to be identified and tested before dismissing the presence of I×E or G×E in the population. This could include the population-mean phenotype as a proxy for the combined effect of all (unknown) environmental variables (Gienapp 2018; **Chapter 9**). For this method, however, due caution is warranted since changes in mean phenotypes between environments may be underlain by factors not directly related to the current environment, for example, a genetic change in response to selection or a cohort of juveniles growing to an adult body size from one season to the next. This reiterates the point that we need to understand our study system well before we can truly understand evolutionary processes in wild populations.

More generally, phenotypic (and additive genetic) variation—and selection thereon—should be studied in a reaction norm context whenever we have access to multiple observations per individual (Postma and van Noordwijk 2005b). Quantifying the sensitivity of traits to the environment can give important insights into why, for example, traits exhibit more or less genetic variation (a trait that is extremely responsive to the environment will likely have a low heritability because the environment is responsible for much of the variation in the phenotype) or whether selection on a given phenotype in a given environment acts on the mean value or the slope of the reaction norm (see **Chapter 10**). Moreover, the slopes of individual reaction norms relative to the population-level reaction norm may tell us whether shifts in phenotypes over time are a plastic response to

a changing environment or whether there is—perhaps—evolution at work (Gienapp and Brommer 2014).

Open data as a tool to move the science forward at high speed. The use of Open Data should and will greatly enhance the science of ecology and evolution (Whitlock et al. 2010; Hampton et al. 2013). Many important evolutionary questions, some of which may not have been posed yet, can be addressed at a broad scale beyond a single study system if these data are appropriately used (**Chapter 7** and **9**). They key would be to first formulate an outstanding question, determine which type of data are needed to answer this question, find the data (Culina et al. 2018) and perform the analysis. Sometimes it may help to know what types of data are available and let that inspire research questions. As a note of caution, however, the use of data always needs to be communicated with the original author to avoid misinterpretation and erroneous conclusions. Finally, results of the data-driven meta-study should be synthesised using formal meta-analysis to obtain robust conclusions.

Combining multiple sources of data. To make reliable (inference of) evolutionary predictions, it is desirable to combine as many sources of information as possible. This could be, for example, a combination of observational (correlational) and experimental data for the association between phenotypes and measures of fitness. Also, additional data about the environment (e.g. food availability) should help us understand selective pressures operating on the traits of interest without having to resort to (potentially less reliable) environmental proxies (e.g. Arlt and Pärt 2017; Marrot et al. 2018). Lastly, simulations can help us understand future evolutionary trajectories when parameterised with realistic values obtained from observational and experimental data.

Final note: a view on using complex quantitative genetics in observational studies

I shall complete this outlook with a short view on some aspects of the field of quantitative genetics in wild populations. In this thesis, I have used quantitative genetics to (1) calculate breeding values for laying date in great tits (**Chapter 3**), (2) calculate heritability of clutch size and make evolutionary predictions based on individual-based model (**Chapter 8**), and (3) quantify the genetic (co)variance matrix for reaction norms to facilitate the prediction of evolutionary trajectories (**Chapters 9** and **10**). When pedigrees are sufficiently informative and contain a large number of families (e.g. ≥ 100), animal models have proven to be very powerful in estimating heritabilities (e.g. Charmantier and Réale 2005) and long-term population studies continue to produce better estimates over time (Postma 2014).

Evolutionary ecologists are increasingly interested in complex questions that involve multi-dimensional (additive) genetic variance-covariance matrices, for example to estimate environmental variation in genetic variation (Hoffman and Merilä 1999) or the additive genetic covariance between two traits (e.g. Sheldon et al. 2003; Class and Brommer 2015b) or between a trait and fitness to measure selection at the genetic level (e.g. Morrissey et al. 2012a; Reed et al. 2016b). Particularly when one is interested in the estimation of (additive) genetic covariances, such analyses may not be helpful in many

situations because they typically require large samples to reach sufficient statistical power (Steppan et al. 2002). In **Chapter 9**, we regressed environment-specific heritability estimates against environment-specific selection. In an ideal situation, we would have constructed multivariate animal models that would allow us to directly estimate the additive genetic covariance between fitness and trait value, which would be too computationally heavy for most (or all) of the datasets, especially since this would have to be done for each environment (year), decreasing sample sizes even more. In a study on great tits, estimating the additive genetic covariance between fitness and two life-history traits (i.e. selection at the genetic level) was only possible when the entire dataset (60 years) was divided into three groups of years of strong, weak, and medium phenotypic selection (Reed et al. 2016b), because estimating genetic selection annually would require a very large annual sample size. The estimation of the additive genetic covariance between female fitness and her liability to produce extra-pair offspring in song sparrows (*Melospiza melodia*) led to weak, statistically indiscernible effects, like partly due to the lack of power (Reid 2012). Studies that show statistically discernible (additive) genetic covariances usually have a large sample size, i.e. (repeated) observations of at least a few thousand individuals (e.g. Morrissey et al. 2012b; Class and Brommer 2015b).

The above might indicate that in many practical (non-experimental) situations we will not be able to test reliably whether genetic (co)variance matrices change over time (by splitting the dataset in groups of years) or whether annual selection indeed acts at the genetic level (by estimating matrices at the annual level). This is not to suggest that evolutionary biologists should not aspire to ask these intriguing questions, but they may have to realise that (1) observational studies may not always be suitable to ask complex questions (such as testing genetic maternal effects, which will be only possible when they can be separated from common-environment effects, which is typically only achieved in cross-foster experiments (Kruuk and Hadfield 2007)), and (2) more powerful methods to study the genetic relationships between individuals may be needed to be able to answer such questions. For example, relatedness matrices based on genomic markers (as opposed to a pedigree) are less sparse and may therefore be more powerful in detecting genetic variation and covariation in traits (Gienapp et al. 2017). Ultimately, the power to estimate such estimates will largely depend on the number of phenotypes measured, as illustrated by an example of the additive genetic covariance between *Tau* (the free-running daily period length under constant conditions) and laying date in great tits (Box 11.1). Despite an impressive number of wild birds genotyped (for the selection-line experiment; see **Chapter 4**) and a powerful 'genomic' pedigree, I was not able to find a genetic covariance between *Tau* measured in birds in captivity and laying dates of (different) birds in the wild. Although this may have been because of a genuine lack of a genetic covariance (cf. Verhagen et al. in review), the lack of a covariance between wild and captive laying dates (the same trait) suggests that statistical power was likely an additional issue (Box 11.1).

Evolutionary biologists specifically aspiring to answer these exciting questions in future projects should therefore carefully consider whether available observational data are suitable or whether experimental evolution in the wild (which typically increases power to detect the sought effects) (e.g. Postma et al. 2007) may be more appropriate, as well as whether genomic tools may be a fruitful avenue for them.

Box 11.1. *Quantifying the genetic covariance between laying date in the wild and* Tau*.*

Introduction

Circadian rhythms constitute an essential part of the organisation of every-day life in many organisms. Many internal processes, e.g. cell division, gene expression and hormone secretion, as well as rest–activity cycles, are regulated through these rhythms (Takahashi et al. 2001). In birds in laboratory conditions, daily activity patterns have been shown to entrain to daily light–dark cycles, and constant darkness or dim-light conditions lead to a shift in activity onset (Kumar et al. 2000; De Jong et al. 2015). This shift occurs because the internal free-running period, called *Tau*, is not exactly 24h and varies between individuals. In great tits, *Tau* is moderately to strongly heritable (h^2 = 0.48–0.86) (Helm and Visser 2010; Laine et al. in review). Variation in *Tau* in general is believed to be coupled with variation in fitness (Vaze and Sharma 2013), such that there is selection for being well entrained to the natural light–dark cycle (Reppert and Weaver 2002).

Seasonal timing of reproduction in birds is directly mediated by temperature (e.g. Visser et al. 2009a), but the priming of the reproductive system follows a circannual clock set by photoperiod (Helm et al. 2013). Since photoperiod involves the measuring of day length, it has been suggested that the circadian clock is also involved in seasonal timing of reproduction (e.g. the *clock* gene). Associations between *clock* gene variation and breeding phenology have indeed been reported in birds (Liedvogel et al. 2009; Caprioli et al. 2012), suggesting the involvement of this gene in adaptation to novel environmental conditions. Hence, variation in *Tau* may be genetically correlated with phenology in great tits (but see Verhagen et al. in review). This is difficult to determine for wild animals because they need to be taken indoors for a longer period of time. Powerful quantitative genetic methods may help here when laboratory-retrieved phenotypes (*Tau*) can be linked to 'wild' phenotypes (egg-laying date; ELD) of relatives via the pedigree. Here I therefore aimed to **estimate the additive genetic covariance** (cov_A) between ELD and *Tau* in great tits.

Phenotyping and constructing a genomic relatedness matrix

A total of 154 birds (76 males, 78 females) were held in captivity in a large-scale genomic selection experiment for phenology (Verhagen et al. in review); these birds were more or less equally divided in numbers over three selection generations. Selection on ELD was done based on genomic breeding values for ELD, calculated based on ~2000 wild birds genotyped on a 650k SNP chip (see Gienapp et al. 2019 for details). For each bird, the period length (*Tau*) was determined using the methods described in Spoelstra et al. (2018). The mean period length (± SD) was 23.70 (± 0.16). For the wild population (Hoge Veluwe), we had 3527 ELDs from 2092 genotyped females (ELD variance: 24.8).

From the genotyped birds, a 'genomic' relatedness matrix (GRM) was constructed, based on the average allelic correlation across loci between two genotypes (Gienapp et al. 2017; see Laine et al. in review for more details). The advantage of this GRM over a pedigree-based relatedness matrix is that the relatedness between any individual can be calculated—and not just direct relatives. The GRM is hence less sparse and preserves more power and accuracy in estimating genetic parameters (Gienapp et al. 2017). This is necessary since the 154 lab birds, which had been bred in captivity for a couple of generations, had no direct wild relatives.

Estimating the additive genetic covariance between Tau *and ELD*

Both *Tau* and ELD were mean- and variance-standardized prior to analysis, to get both traits on the same scale. We first fitted univariate 'minimum adequate models' (MAM) in ASReml-R (Butler et al. 2009; Gilmour et al. 2009) for both *Tau* and ELD. Through backward elimination of non-significant fixed effects using conditional Wald *F* tests, we determined that the age of the breeding female (first-year breeder or older) affected ELD, and that sex affected *Tau.* Likelihood-ratio tests confirmed an effect of female identity for repeated measures on ELD ($\hat{\sigma}^2$ = 0.17 ± 0.05), and …

Box 11.1 *(continued)*

… mother identity for *Tau* ($\hat{\sigma}^2$ = 0.27 ± 0.13). Next, we fitted univariate animal models, adding the inverse GRM to estimate the additive genetic effect in both traits. There was significant heritable variation in both traits (Table B11.1.2). When we verified there was significant additive genetic variation, we built bivariate models for ELD and *Tau* that included the fixed and random effects identified in the MAMs, with independent normal residual variance (i). We then added the GRM, constraining cov_A to 0 (ii). The final model included the GRM with an unconstrained covariance (iii).

No additive genetic covariance between Tau *and ELD*

The bivariate animal models showed no significant additive genetic correlation between *Tau* and ELD (Tables B11.1.1a and B11.1.2). The cov_A estimate was small with a large standard error (–0.056 ± 0.162). This was likely the result of a combination of small effects and low power, given the large standard error. To test this, we ran another set of bivariate models, in which *Tau* was replaced by the ELDs of these same birds in captivity (ELD_{lab}) (Verhagen et al. in review). The expectation was that the model would be able to pick up the cov_A since the same genes are responsible for the same trait. The estimated cov_A was comparatively large (–0.188 ± 0.156), but also here the standard error was substantial (Tables B11.1.1b and B11.1.2). Statistical power is therefore likely an issue here.

To be able to estimate cov_A between wild ELD and ELD_{lab} or indeed *Tau* reliably, we probably need more than 154 individuals, as estimating cov_A is usually a data-hungry exercise (Steppan et al. 2002). Since relatedness matrices based on genomic markers contain information about the whole sampled population at once, the power to detect genetic correlations is most likely limited by the number of phenotypes available, rather than the number of 'families' (Gienapp et al. 2017). Unfortunately, although the genotyping of ~2000 wild birds and the phenotyping of the 154 birds in captivity is unique and impressive in its scale and the costs of genotyping are likely to decrease over time, it still remains an expensive exercise. The required budgets, manpower and other resources (e.g. bird housing and facilities) may prevent other studies on vertebrate organisms to phenotype as many or even more individuals in the near future.

Table B11.1.1 *Results of likelihood-ratio tests comparing bivariate (i) minimum adequate models with models with (ii) an additive genetic term fitted for both traits separately and (iii) the additive genetic covariance between them.*

Random-effects structure	LogLik	χ^2	*df*	*p*
(a) Wild egg-laying date (ELD_W) vs. Tau				
i. MAM	–1680.4			
ii. MAM + inverse GRM (constrained covariance)	–1670.1	20.65	2	< 0.0001
iii. MAM + inverse GRM (unconstrained covariance)	–1670.0	0.14	1	0.35
(b) ELD_w vs ELD_{lab}				
i. MAM	–1687.9			
ii. MAM + inverse GRM (constrained covariance)	–1674.1	27.52	2	< 0.0001
iii. MAM + inverse GRM (unconstrained covariance)	–1673.7	0.88	1	0.17

MAM = minimum adequate model (in (1): MotherID for *Tau* and FemaleID for ELD_W; in (2): FemaleID for ELD_W).

Table B11.1.2*. Estimates (SE) of additive genetic variation (diagonals) and additive genetic covariance (off-diagonals).*

	ELD_W	*Tau*	ELD_{lab}
ELD_W	**0.171 (0.050)**		
Tau	–0.056 (0.162)	**0.523 (0.222)**	
ELD_{lab}	–0.188 (0.156)		**0.398 (0.146)**

Significant effects marked in bold. Variance estimates are relative to 1.

Conclusion

In this thesis, I explored how populations are coping with environmental change and which ecological processes potentially affect the rate of adaption. I did this using a combination of three approaches. In **Part I** of my thesis, I explored how avian populations are affected by climate change in general, and I concluded in this chapter that there is a lot more that we need to learn about the adaptive value of (phenotypic) changes induced by climate change. In **Part II**, I aimed to identify the ecological constraints to adaptation in avian breeding time using a combination of experiments and long-term observational data. I concluded that it is too early as of yet to conclude whether great tits are indeed constrained to breed earlier, which is necessary to restore the phenological match with their main food source, the caterpillars. I also concluded that our current definition of phenological match, i.e. the match in peak dates consumer and resource phenology, probably adequately describes observed demographic processes. In **Part III**, I explored adaptation and evolutionary dynamics in long-term study populations using quantitative genetic tools. I concluded that in order for us to make reliable predictions of evolutionary processes and to assess the generality of our findings, we need to (1) have a thorough understanding of the ecology of the species under study and integrate multiple sources of information, (2) use quantitative genetic tools in a reaction norm context, and (3) explore the availability of Open Data as an opportunity to answer novel evolutionary questions at a broad (taxonomic or geographic) scale.

Much is yet to be discovered about how populatons are coping with environmental change in general, in particular beyond the few (very) well-studied model species. The central question of my thesis, posed in **Chapter 1** and at the beginning of this section, hence cannot—and should not—receive an unequivocal answer. I have, however, shown that major questions posed by evolutionary ecologists are best solved using a 'bottom-up' approach, where we first identify and carefully break down the ecological factors driving (selection on) phenotypic varation and use that knowledge as a basis for quantifying evolutionary dynamics. Careful consideration of the ecological and evolutionary forces at play have enabled me to reveal that what we initially thought might constrain adaptation in wild populations, may not necessarily be true. Nevertheless, I *have* shown that the great tits of the Hoge Veluwe are under selective pressure to breed earlier and suggested that micro-evolution might not be sufficient to evade extinction in the long run. As evolutionary ecologists, we need to keep inspired to ask outstanding novel questions but, as our methods improve and our data expand, also be willing to revisit the old ones. This will enable us to once more refine our methods and our basic understanding of our study systems, and eventually improve our predictions of evolution in the wild.

References

Ahola, M., T. Laaksonen, K. Sippola, T. Eeva, K. Rainio, and A. Lehikoinen. 2004. Variation in climate warming along the migration route uncouples arrival and breeding dates. Global Change Biology 10:1610–1617.

Alerstam, T. 2011. Optimal bird migration revisited. Journal of Ornithology 152:5–23.

Archibald, J. D. 2011. Extinction and radiation: how the fall of dinosaurs led to the rise of mammals. John Hopkins University Press.

Arlt, D. and T. Pärt. 2017. Marked reduction in demographic rates and reduced fitness advantage for early breeding is not linked to reduced thermal matching of breeding time. Ecology and Evolution 7:10782-10796.

Arnaud, C. M., P. H. Becker, F. S. Dobson, and A. Charmantier. 2013. Canalization of phenology in common terns: genetic and phenotypic variations in spring arrival date. Behavioral Ecology 24:683–690.

Arnold, T. W. 1993. Factors affecting egg viability and incubation time in prairie dabbling ducks. Canadian Journal of Zoology 71:1146-1152.

Auld, J. R., A. A. Agrawal, and R. A. Relyea. 2010. Re-evaluating the costs and limits of adaptive phenotypic plasticity. Proceedings of the Royal Society B-Biological Sciences 277:503–511.

Avilés, J. M., M. Molina-Morales, and J. Gabriel Martínez. 2014. Climatic effects and phenological mismatch in cuckoo-host interactions: a role for host phenotypic plasticity in laying date? Oikos 123:993-1002.

Badyaev, A. V. 2005. Maternal inheritance and rapid evolution of sexual size dimorphism: Passive effects or active strategies? American Naturalist 166:S17–S30.

Bailey, L. D. and M. Van de Pol. 2017. Package 'climwin': Climate Window Analysis. https://cran.r-project.org/web/packages/climwin/index.html.

Barbraud, C. and H. Weimerskirch. 2003. Cimate and density shape population dynamics of a marine top predator. Proceedings of the Royal Society B-Biological Sciences 270:2111–2116.

Barrowclough, G. F., J. Cracraft, J. Klicka, and R. M. Zink. 2016. How Many Kinds of Birds Are There and Why Does It Matter? PLoS One 11:e0166307.

Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. Bojesen, H. Singmann, B. Dai, F. Scheipl, G. Grothendieck, P. Green, and J. Fox. 2018. Package 'lme4': Linear Mixed-Effects Models using 'Eigen' and S4. CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Bayliss, H. R. and F. R. Beyer. 2015. Information retrieval for ecological syntheses. Research Synthesis Methods 6:136-148.

Becker, P. J. J., J. Hegelbach, L. F. Keller, and E. Postma. 2016a. Data from: Phenotype-associated inbreeding biases estimates of inbreeding depression in a wild bird population. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.53g68.

Becker, P. J. J., J. Hegelbach, L. F. Keller, and E. Postma. 2016b. Phenotype-associated inbreeding biases estimates of inbreeding depression in a wild bird population. Journal of Evolutionary Biology 29:35–46.

Beckerman, A. P., T. G. Benton, C. T. Lapsley, and N. Koesters. 2006. How effective are maternal effects at having effects? Proceedings of the Royal Society B-Biological Sciences 273:485–493.

Beebee, T. J. C. 1995. Amphibian breeding and climate. Nature 374.

Bell, G. P. 1990. Birds and mammals on an insect diet: a primer on diet composition analysis in relation to ecological energetics. Studies in Avian Biology 13.

Beniston, M., D. B. Stephenson, O. B. Christensen, C. A. Ferro, C. Frei, S. Goyette, K. Halsnaes, T. Holt, K. Jylhä, and B. Koffi. 2007. Future extreme events in European climate: an exploration of regional climate model projections. Climatic Change 81:71-95.

Betts, M. M. 1955. The food of titmice in oak woodland. Journal of Animal Ecology 24:282–323.

Bêty, J., G. Gauthier, and J.-F. Giroux. 2003. Body condition, migration, and timing of reproduction in snow geese: a test of the condition-dependent model of optimal clutch size. American Naturalist 162:110–121.

Bêty, J., J.-F. Giroux, and G. Gauthier. 2004. Individual variation in timing of migration: causes and reproductive consequences in greater snow gees (*Anser caerulescens atlanticus*). Behavioral Ecology and Sociobiology 57:1–8.

Bijma, P. 2011. A general definition of the heritable variation that determines the potential of a population to respond to selection. Genetics 189:1347–1359.

Blondel, J., P. Perret, and M. Maistre. 1990. On the genetical basis of the laying-date in an island population of blue tits. Journal of Evolutionary Biology 3:469-475.

Boag, P. T. and P. R. Grant. 1978. Heritability of external morphology in Darwin's finches. Nature 274:793.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127-135.

Bonnet, T., P. Wandeler, G. Camenisch, and E. Postma. 2017. Bigger is fitter? Quantitative genetic composition of selection reveals an adaptive evolutionary decline of body mass in a wild rodent population. PLoS Biology 15:e1002592.

Both, C. 2010a. Flexibility of timing of avian migration to climate change masked by environmental constraints en route. Current Biology 20:243–248.

Both, C. 2010b. Food availability, mistiming, and climatic change. Pp. 129–147 *in* A. P. Møller, H. P. Fiedler, and P. Berthold, eds. Effects of climate change on birds. Oxford University Press, Oxford, UK.

Both, C., A. V. Artemyev, B. Blaauw, R. J. Cowie, A. J. Dekhuijzen, T. Eeva, A. Enemar, L. Gustafsson, E. V. Ivankina, and A. Järvinen. 2004a. Large–scale geographical variation confirms that climate change causes birds to lay earlier. Proceedings of the Royal Society of London B: Biological Sciences 271:1657-1662.

Both, C., A. V. Artemyev, B. Blaauw, R. J. Cowie, A. J. Dekhuijzen, T. Eeva, A. Enemar, L. Gustafsson, E. V. Ivankina, A. Järvinen, N. B. Metcalfe, N. E. I. Nyholm, J. Potti, P.-A. Ravussin, J. J. Sanz, B. Silverin, F. M. Slater, L. V. Sokolov, J. Török, W. Winkel, J. Wright, H. Zang, and M. E. Visser. 2004b. Large-scale geographical variation confirms that climate change causes birds to lay earlier. Proceedings of the Royal Society B-Biological Sciences 271:1657–1662.

Both, C., S. Bouwhuis, C. M. Lessells, and M. E. Visser. 2006. Climate change and population declines in a long-distance migratory bird. Nature 441:81–83.

Both, C., J. M. Tinbergen, and M. E. Visser. 2000. Adaptive density dependence of avian clutch size. Ecology 81:3391–3403.

Both, C., M. van Asch, R. G. Bijlsma, A. B. van den Burg, and M. E. Visser. 2009a. Climate change and unequal phenological changes across four trophic levels: constraints or adaptations? Journal of Animal Ecology 78:73–83.

Both, C., M. van Asch, R. G. Bijlsma, A. B. van den Burg, and M. E. Visser. 2009b. Climate change and unequal phenological changes across four trophic levels: constraints or adaptations? Journal of Animal Ecology 78:73-83.

Both, C. and M. E. Visser. 2001. Adjustment to climate change is constrained by arrival date in a long-distance migrant bird. Nature 411:296–298.

Both, C., M. E. Visser, and N. Verboven. 1999. Density-dependent recruitment rates in great tits: the importance of being heavier. Proceedings of the Royal Society B-Biological Sciences 266:465–469.

Bowers, E. K., J. L. Grindstaff, S. S. Soukup, N. E. Drilling, K. P. Eckerle, S. K. Sakaluk, and C. F. Thompson. 2016. Spring temperatures influence selection on breeding date and the potential for phenological mismatch in a migratory bird. Ecology 97:2880-2891.

Bradshaw, C. J. A., B. W. Brook, S. Delean, D. A. Fordham, S. Herrando-Pérez, P. Cassey, R. Early, C. H. Sekercioglu, and M. B. Araújo. 2014. Predictors of contraction and expansion of area of occupancy for British birds. Proceedings of the Royal Society B-Biological Sciences 281:20140744.

Bridge, E. S., K. Thorup, M. S. Bowlin, P. B. Chilson, R. H. Diehl, R. W. Fléron, P. Hartl, R. Kays, J. F. Kelly, W. D. Robinson, and M. Wikelski. 2011. Technology on themove: recent and forthcoming innovations for tracking migratory birds. BioScience 61:689–698.

Brommer, J. E., J. S. Alho, C. Biard, J. R. Chapman, A. Charmantier, A. Dreiss, I. R. Hartley, M. B. Hjernquist, B. Kempenaers, J. Komdeur, T. Laaksonen, P. K. Lehtonen, T. Lubjuhn, S. C. Patrick, B. Rosivall, J. M. Tinbergen, M. van der Velde, K. van Oers, T. Wilk, and W. Winkel. 2010. Passerine extrapair mating dynamics: a Bayesian modeling approach comparing four species. American Naturalist 176:178–187.

Brommer, J. E., P. Kontiainen, and H. Pietiäinen. 2012. Selection on plasticity of seasonal life-history traits using random regression mixed model analysis. Ecology and Evolution 2:695–704.

Brommer, J. E., J. Merilä, B. C. Sheldon, and L. Gustafsson. 2005. Natural selection and genetic variation for reproductive reaction norms in a wild bird population. Evolution 59:1362–1371.

Brommer, J. E., K. Rattiste, and A. J. Wilson. 2008. Exploring plasticity in the wild: laying date–temperature reaction norms in the common gull *Larus canus*. Proceedings of the Royal Society B-Biological Sciences 275:687–693.

Brown, C. R. and M. B. Brown. 2000. Weather-mediated natural selection on arrival time in cliff swallows (*Petrochelidon pyrrhonota*). Behavioral Ecology and Sociobiology 47:339–345.

Bürger, R. and M. Lynch. 1994. Evolution and extinction in a changing environment: a quantitative-genetic analysis. Evolution 49:151–163.

Burnham, K. P. and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65:23-35.

Buse, A., S. J. Dury, R. J. W. Woodburn, C. M. Perrins, and J. E. G. Good. 1999. Effects of elevated temperature on multi-species interactions: the case of Pedunculate Oak, Winter Moth and Tits. Functional Ecology 13:74–82.

Butler, D., B. R. Cullis, A. R. Gilmour, and D. J. Gogel. 2009. ASReml-R Reference Manual, Release 3.0. Department of Primary Industries and Fisheries, Brisbane, Qld, Australia.

Cadotte, M. W., L. R. Mehrkens, and D. N. Menge. 2012. Gauging the impact of meta-analysis on ecology. Evolutionary Ecology 26:1153-1167.

Caprioli, M., R. Ambrosini, G. Boncoraglio, E. Gatti, A. Romano, M. Romano, D. Rubolini, L. Gianfranceschi, and N. Saino. 2012. Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. PLoS One 7:e35140.

Carlson, S. M., C. J. Cunningham, and P. A. H. Westley. 2014. Evolutionary rescue in a changing world. Trends in Ecology & Evolution 29:521–530.

Caro, S. P., A. Charmantier, M. M. Lambrechts, J. Blondel, J. Balthazart, and T. D. Williams. 2009. Local adaptation of timing of reproduction: females are in the driver's seat. Functional Ecology 23:172-179.

Carter, M. J., M. I. Lind, S. R. Dennis, W. Hentley, and A. P. Beckerman. 2017. Evolution of a predator-induced, nonlinear reaction norm. Proc. R. Soc. B 284:20170859.

Cassey, P., J. G. Ewen, T. M. Blackburn, and A. P. Møller. 2004. A survey of publication bias within evolutionary ecology. Proceedings of the Royal Society of London B-Biological Sciences 271:S451-S454.

Champagne, F. A. 2008. Epigenetic mechanisms and the transgenerational effects of maternal care. Frontiers in Neuroendocrinology 29:386–397.

Charmantier, A., D. Garant, and L. E. B. Kruuk. 2014. Quantitative genetics in the wild. Oxford University Press, Oxford, UK.

Charmantier, A. and P. Gienapp. 2014. Climate change and timing of avian breeding and migration: evolutionary versus plastic changes. Evolutionary Applications 7:15–28.

Charmantier, A., R. H. McCleery, L. R. Cole, C. M. Perrins, L. E. B. Kruuk, and B. C. Sheldon. 2008. Adaptive phenotypic plasticity in response to climate change in a wild bird population. Science 320:800–803.

Charmantier, A. and D. Réale. 2005. How do misassigned paternities affect the estimation of heritability in the wild? Molecular Ecology 14:2839–2850.

Chevin, L. M., M. E. Visser, and J. Tufto. 2015. Estimating the variation, autocorrelation, and environmental sensitivity of phenotypic selection. Evolution 69:2319-2332.

Christians, J. K., M. Evanson, and J. J. Aiken. 2001. Seasonal decline in clutch size in European starlings: a novel randomization test to distinguish between the timing and quality hypothesis. Journal of Animal Ecology 70:1080–1087.

Class, B. and J. E. Brommer. 2015a. Data from: A strong genetic correlation underlying a behavioural syndrome disappears during development because of genotype-age interactions. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.443g2.

Class, B. and J. E. Brommer. 2015b. A strong genetic correlation underlying a behavioural syndrome disappears during development because of genotype-age interactions. Proceedings of the Royal Society B-Biological Sciences 282:20142777.

Clausen, K. K. and P. Clausen. 2013. Earlier Arctic springs cause phenological mismatch in long-distance migrants. Oecologia 173:1101-1112.

Cleasby, I. R. and S. Nakagawa. 2011. Neglected biological patterns in the residuals. Behavioral Ecology and Sociobiology 65:2361-2372.

Close, R. A., M. Friedman, G. T. Lloyd, and R. B. Benson. 2015. Evidence for a mid-Jurassic adaptive radiation in mammals. Current Biology 25:2137-2142.

Clutton-Brock, T. and B. C. Sheldon. 2010. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. Trends in Ecology & Evolution 25:562–573.

Cobben, M. M. P. and K. van Oers. 2016. Bolder takes all and the role of epigenetics. A comment on Canestrelli et al. Trends in Ecology & Evolution 31:498–499.

Coltman, D. W. 2005. Testing marker-based estimates of heritability in the wild. Molecular Ecology 14:2593-2599.

Côté, I. M. and J. D. Reynolds. 2012. Meta-analysis at the intersection of evolutionary ecology and conservation. Evolutionary Ecology 26:1237-1252.

Cresswell, W. and R. McCleery. 2003. How great tits maintain synchronization of their hatch date with food supply in response to long-term variability in temperature. Journal of Animal Ecology 72:356–366.

Crick, H. Q. P., C. Dudley, D. E. Glue, and D. L. Thomson. 1997. UK birds are laying eggs earlier. Nature 388:526.
Csilléry, K., T. Johnson, D. Beraldi, T. Clutton-Brock, D. Coltman, B. Hansson, G. Spong, and J. M. Pemberton. 2006. Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. Genetics 173:2091-2101.
Cubas, P., C. Vincent, and E. Coen. 1999. An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401:157–161.
Culina, A., M. Baglioni, T. W. Crowther, M. E. Visser, S. Woutersen, and P. Manghi. 2018. Navigating the unfolding open data landscape in ecology and evolution. Nature Ecology & Evolution 2:420–426.
Cushing, D. 1969. The regularity of the spawning season of some fishes. Journal du Conseil 33:81-92.
Cushing, D. 1990. Plankton production and year-class strength in fish populations: an update of the match/mismatch hypothesis. Pp. 249-293. Advances in Marine Biology.

Dalhaug, L., I. M. Tombre, and K. E. Erikstad. 1996. Seasonal decline in clutch size of the barnacle goose in Svalbard. Condor 98:42–47.
Dall, S. R., A. M. Bell, D. I. Bolnick, and F. L. Ratnieks. 2012. An evolutionary ecology of individual differences. Ecology Letters 15:1189-1198.
Danchin, É., A. Charmantier, F. A. Champagne, A. Mesoudi, B. Pujol, and S. Blanchet. 2011. Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. Nature Reviews Genetics 12:475–486.
Danks, H. V. 1987. Insect dormancy: an ecological perspective. Biological Survey of Canada (Terrestrial Artropods), Ottawa, ON.
Darwin, C. 1859. On the Origin of Species by Means of Natural Selection. John Murray, London.
De Jong, G. 1990. Quantitative genetics of reaction norms. Journal of Evolutionary Biology 3:447–468.
De Jong, G. 1999. Unpredictable selection in a structured population leads to local genetic differentiation in evolved reaction norms. Journal of Evolutionary Biology 12:839-851.
De Jong, M., L. Jeninga, J. Q. Ouyang, K. van Oers, K. Spoelstra, and M. E. Visser. 2015. Dose-dependent responses of avian daily rhythms to artificial light at night. Physiology & Behavior 155:172–179.
Delahaie, B., A. Charmantier, S. Chantepie, D. Garant, M. Porlier, and C. Teplitsky. 2017a. Conserved G-matrices of morphological and life-history traits among continental and island blue tit populations. Heredity 119:76–87.
Delahaie, B., A. Charmantier, S. Chantepie, D. Garant, M. Porlier, and C. Teplitsky. 2017b. Data from: Conserved G-matrices of morphological and life-history traits among continental and island blue tit populations. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.03mn0.
DeWitt, T. J., A. Sih, and D. S. Wilson. 1998. Costs and limits of phenotypic plasticity. Trends in Ecology & Evolution 13:77–81.
Dey, S., S. R. Proulx, and H. Teotónio. 2016. Adaptation to temporally flutcuating environments by the evolution of maternal effects. PLoS Biology 14:e1002388.
DiCiccio, T. J. and B. Efron. 1996. Bootstrap condidence intervals. Statistical Science 11:189–228.
Dingemanse, N. J. and N. A. Dochtermann. 2013. Quantifying individual variation in behaviour: mixed-effect modelling approaches. Journal of Animal Ecology 82:39–54.
Dingemanse, N. J., A. J. Kazem, D. Réale, and J. Wright. 2010. Behavioural reaction norms: animal personality meets individual plasticity. Trends in Ecology & Evolution 25:81-89.
Dunn, P. O. and A. P. Moller. 2014. Changes in breeding phenology and population size of birds. Journal of Animal Ecology 83:729-739.
Dunn, P. O. and D. W. Winkler. 1999. Climate change has affected the breeding date of tree swallows throughout North America. Proceedings of the Royal Society B-Biological Sciences 266:2487–2490.
Dunn, P. O., D. W. Winkler, L. A. Whittingham, S. J. Hannon, and R. J. Robertson. 2011. A test of the mismatch hypothesis: How is timing of reproduction related to food abundance in an aerial insectivore? Ecology 92:450-461.
Durant, J. M., D. O. Hjermann, T. Anker-Nilssen, G. Beaugrand, A. Mysterud, N. Pettorelli, and N. C. Stenseth. 2005. Timing and abundance as key mechanisms affecting trophic interactions in variable environments. Ecology Letters 8:952-958.
Durant, J. M., D. Ø. Hjermann, G. Ottersen, and N. C. Stenseth. 2007. Climate and the match or mismatch between predator requirements and resource availability. Climate Research 33:271–283.

Easterling, D. R., B. Horton, P. D. Jones, T. C. Peterson, T. R. Karl, D. E. Parker, M. J. Salinger, V. Razuvayev, N. Plummer, P. Jamason, and C. K. Folland. 1997. Maximum and minimum temperature trends for the globe. Science 277:364–367.

Embree, D. G. 1970. The diurnal and seasonal pattern of hatching of winter moth eggs, *Operophtera brumata* (Geometridae: Lepidoptera). Canadian Entomologist 102:759–768.

Erceg-Hurn, D. M. and V. M. Mirosevich. 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. American Psychologist 63:591.

Evans, S. R. 2016. Gauging the purported costs of public data archiving for long-term population studies. PLoS Biology 14:1–9.

Evans, S. R. and B. C. Sheldon. 2015a. Colour in a new light: a spectral perspective on the quantitative genetics of carotenoid coloration. Functional Ecology 29:96–103.

Evans, S. R. and B. C. Sheldon. 2015b. Data from: Colour in a new light: a spectral perspective on the quantitative genetics of carotenoid coloration. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.1cn02.

Ezard, T. H. G., R. Prizak, and R. B. Hoyle. 2014. The fitness costs of adaptation via phenotypic plasticity and maternal effects. Functional Ecology 28:693–701.

Falconer, D. S. 1965. Maternal effects and selection response *in* S. J. Geerts, ed. Genetics today: Proceedings of 11th International Congress of Genetics. Pergamon Press, Oxford, UK.

Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Longmans Green, Harlow, Essex, UK.

Feng, S., S. E. Jacobsen, and W. Reik. 2010. Epigenetic reprogramming in plant and animal development. Science 330:622–627.

Field, C. B., V. R. Barros, D. J. Dokken, K. J. Mach, M. D. Mastrandrea, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea, and L. L. White. 2014. IPCC, 2014: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Cambridge University Press, Cambridge, UK, and New York, USA.

Finke, M. 2002. Complete nutrient composition of commercially raised invertebrates used as food for insectivores. Zoo Biology 21:269–285.

Finlay, K. W. and G. N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research 14:742–754.

Firth, J. A., J. D. Hadfield, A. W. Santure, J. Slate, and B. C. Sheldon. 2015. The influence of nonrandom extra-pair paternity on heritability estimates derived from wild pedigrees. Evolution 69:1336–1344.

Gamelon, M., J. Tufto, A. L. K. Nilsson, K. Jerstad, O. W. Røstad, N. C. Stenseth, and B.-E. Sæther. 2018. Environmental drivers of varying selective optima in a small passerine: A multivariate, multiepisodic approach. Evolution 72:2325-2342.

Garant, D. and L. E. Kruuk. 2005. How to use molecular marker data to measure evolutionary parameters in wild populations. Molecular Ecology 14:1843-1859.

Garant, D., L. E. Kruuk, R. H. McCleery, and B. C. Sheldon. 2004a. Evolution in a changing environment: a case study with great tit fledging mass. The American Naturalist 164:E115-E129.

Garant, D., L. E. B. Kruuk, R. H. McCleery, and B. C. Sheldon. 2004b. Evolution in a changing environment: a case study with great tit fledging mass. The American Naturalist 164:E115–E129.

Garant, D., L. E. B. Kruuk, T. A. Wilkin, R. H. McCleery, and B. C. Sheldon. 2005. Evolution driven by differential dispersal within a wild bird population. Nature 433:60–65.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1:515–534.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Gelman, A., Y.-S. Su, J. Hill, M. G. Pittau, J. Kerman, T. Zheng, and V. Dorie. 2016. Package 'arm': Functions to accompany A. German and J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, 2007. CRAN.

Ghalambor, C. K., J. K. McKay, S. P. Carroll, and D. N. Reznick. 2007. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. Functional Ecology 21:394-407.

Gibb, J. and M. M. Betts. 1963. Food and food supply of nestling tits (Paridae) in Breckland pine. The Journal of Animal Ecology:489-533.

Gienapp, P. 2018. The choice of the environmental covariate affects the power to detect individual variation in reaction norm slopes. BioRxiv: https://doi.org/10.1101/311217

Gienapp, P. and T. Bregnballe. 2012. Consequences of timing of migration and breeding in cormorants. PLoS One 7:e46165.

Gienapp, P. and J. E. Brommer. 2014. Evolutionary dynamics in response to climate change *in* A. Charmantier, D. Garant, and L. E. B. Kruuk, eds. Quantitative genetics in the wild. Oxford University Press, Oxford, UK.

Gienapp, P., M. P. L. Calus, V. N. Laine, and M. E. Visser. 2019. Genomic selection on breeding time in a wild bird population. Evolution Letters Advance online publication: https://doi.org/10.1002/evl3.103.

Gienapp, P., S. Fior, F. Guillaume, J. R. Lasky, V. L. Sork, and K. Csilléry. 2017. Genomic quantitative genetics to study evolution in the wild. Trends in Ecology & Evolution 32:897–908.

Gienapp, P., L. Hemerik, and M. E. Visser. 2005. A new statistical tool to predict phenology under climate change scenarios. Global Change Biology 11:600–606.

Gienapp, P., R. Leimu, and J. Merilä. 2007. Responses to climate change in avian migration time—Microevolution versus phenotypic plasticity. Climate Research 35:25–35.

Gienapp, P., M. Lof, T. E. Reed, J. McNamara, S. Verhulst, and M. E. Visser. 2013a. Predicting demographically sustainable rates of adaptation: can great tit breeding time keep pace with climate change? Philosophical Transactions of the Royal Society B 368:20120289.

Gienapp, P. and J. Merilä. 2014a. Data from: Disentangling plastic and genetic changes in body mass of Siberian jays. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.4v514.

Gienapp, P. and J. Merilä. 2014b. Disentangling plastic and genetic changes in body mass of Siberian jays. Journal of Evolutionary Biology 27:1849–1858.

Gienapp, P., E. Postma, and M. E. Visser. 2006. Why breeding time has not responded to selection for earlier breeding in a songbird population. Evolution 60:2381–2388.

Gienapp, P., T. E. Reed, and M. E. Visser. 2014. Why climate change will invariably alter selection pressures on phenology. Proceedings of the Royal Society B-Biological Sciences 281:20141611.

Gienapp, P., C. Teplitsky, J. S. Alho, J. A. Mills, and J. Merilä. 2008. Climate change and evolution: disentangling environmental and genetic responses. Molecular Ecology 17:167–178.

Gienapp, P., A. J. van Noordwijk, and M. E. Visser. 2013b. Genetic background, and not ontogenetic effects, affects avian seasonal timing of reproduction. Journal of Evolutionary Biology 26:2147–2153.

Gienapp, P. and M. E. Visser. 2006. Possible fitness consequences of experimentally advanced laying dates in Great Tits: differences between populations in different habitats. Functional Ecology 20:180-185.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide. Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Giorgi, F. and P. Lionello. 2008. Climate change projections for the Mediterranean region. Global and Planetary Change 63:90–104.

Gosden, T. P. and E. I. Svensson. 2008a. Spatial and temporal dynamics in a sexual selection mosaic. Evolution 62:845–856.

Gosden, T. P. and E. I. Svensson. 2008b. Spatial and temporal dynamics in a sexual selection mosaic. Evolution 62:845-856.

Gosler, A. 1993. The Great Tit. Hamlyn Limited, London.

Grant, P. R. and B. R. Grant. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. Science 296:707–711.

Grieco, F. 2002. Time constraint on food choice in provisioning blue tits, *Parus caeruleus*: the relationship between feeding rate and prey size. Animal Behaviour 64:517-526.

Grieco, F., A. J. van Noordwijk, and M. E. Visser. 2002. Evidence for the effect of learning on timing of reproduction in blue tits. Science 296:136-138.

Griggio, M. and A. Pilastro. 2007. Sexual conflict over parental care in a species with female and male brood desertion. Animal Behaviour 74:779–785.

Gunnarsson, T. G. and G. Tomasson. 2011. Flexibility in spring arrival of migratory birds at northern latitudes under rapid temperature changes. Bird Study 58:1–12.

Gurevitch, J., J. Koricheva, S. Nakagawa, and G. Stewart. 2018. Meta-analysis and the science of research synthesis. Nature 555:175.

Gwinner, E. 1996. Circannual clocks in avian reproduction and migration. Ibis 138:47–63.

Hadfield, J. 2008. Estimating evolutionary parameters when viability selection is operating. Proceedings of the Royal Society B-Biological Sciences 275:723–734.

Hadfield, J. 2018. MCMCglmm Course Notes. https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 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. American Naturalist 175:116–125.

Hairston, N. G. and T. A. Dillon. 1990. Fluctuating selection and response in a population of freshwater copepods. Evolution 44:1796–1805.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11:156–162.

Hansen, T. F. and D. Houle. 2004. Evolvability, stabilizing selection, and the problem of stasis. Pp. 130-150 *in* M. Pigliucci, and K. Preston, eds. Phenotypic integration: Studying the ecology and evolution of complex phenotypes. Oxford University Press, Oxford.

Hargreaves, A. L., K. E. Samis, and C. G. Eckert. 2013. Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. The American Naturalist 183:157-173.

Harrison, T. J. E., J. A. Smith, G. R. Martin, D. E. Chamberlain, S. Bearhop, G. N. Robb, and S. J. Reynolds. 2010. Does food supplementation really enhance productivity of breeding birds? Oecologia 164:311–320.

Hayward, A. D., D. H. Nussey, A. J. Wilson, C. Berenos, J. G. Pilkington, K. A. Watt, J. M. Pemberton, and A. L. Graham. 2014. Natural Selection on Individual Variation in Tolerance of Gastrointestinal Nematode Infection. PLoS Biology 12:e1001917.

Hayward, A. D., J. M. Pemberton, C. Berenos, A. J. Wilson, J. G. Pilkington, and L. E. Kruuk. 2018. Evidence for Selection-by-Environment but Not Genotype-by-Environment Interactions for Fitness-Related Traits in a Wild Mammal Population. Genetics 208:349-364.

Haywood, S. 2013. Origin of evolutionary change in avian clutch size. Biological Reviews 88:895–911.

Haywood, S. and C. M. Perrins. 1992. Is clutch size in birds affected by environmental conditions during growth? Proceedings of the Royal Society B-Biological Sciences 249:195–197.

Helm, B., R. Ben-Shlomo, J. Sheriff Michael, A. Hut Roelof, R. Foster, M. Barnes Brian, and D. Dominoni. 2013. Annual rhythms that underlie phenology: biological time-keeping meets environmental change. Proceedings of the Royal Society B: Biological Sciences 280:20130016.

Helm, B. and M. E. Visser. 2010. Heritable circadian period length in a wild bird population. Proceedings of the Royal Society B: Biological Sciences 277:3335-3342.

Henderson, C. R. 1982. Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. Biometrics 38:623–640.

Henderson, C. R. 1988. Theoretical basis and computational methods for a number of different animal models. Journal of Dairy Science 71:1–16.

Hendry, A. P. 2017. Eco-evolutionary Dynamics. Princeton University Press, Oxfordshire.

Hendry, A. P. and M. T. Kinnison. 1999. Perspective: 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.

Herrel, A., K. Huyghe, B. Vanhooydonck, T. Backeljau, K. Breugelmans, I. Grbac, R. Van Damme, and D. J. Irschick. 2008. Rapid large-scale evolutionary divergence in morphology and performance associated with exploitation of a different dietary resource. Proceedings of the National Academy of Sciences of the United States of America 105:4792–4795.

Hill, W. 1984. On selection among groups with heterogeneous variance. Animal Science 39:473-477.

Hitch, A. T. and P. L. Leberg. 2007. Breeding distributions of North American bird species moving north as a result of climate change. Conservation Biology 21:534–539.

Hoffman, A. A. and J. Merilä. 1999. Heritable variation and evolution under favourable and unfavourable conditions. Trends in Ecology & Evolution 14:96–101.

Høgda, K. A., S. R. Karlsen, and I. Solheim. 2001. Climate change impact on growing season in Fennoscandia studied by a time series of NOAA AVHRR NDVI data. IGARSS 2001, Sydney.

Holland, M. L., R. Lowe, P. W. Caton, C. Gemma, G. Carbajosa, A. F. Danson, A. A. M. Carpenter, E. Loche, S. E. Ozanne, and V. K. Rakyan. 2016. Early-life nutrition modulates the epigenetic state of specific rDNA genetic variants in mice. Science 353:495–498.

Holt, R. D. 1990. The microevolutionary consequences of climate change. Trends in Ecology & Evolution 5:311–315.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics 130:195–204.

Howard, C., P. A. Stephens, J. W. Pearce-Higgins, R. D. Gregory, and S. G. Willis. 2015. The drivers of avian abundance: patterns in the relative importance of climate and land use. Global Ecology and Biogeography 24:1249–1260.

Hoyle, R. B. and T. H. G. Ezard. 2012. The benefits of maternal effects in novel and in stable environments. Journal of the Royal Society Interface 9:2403–2413.

Huntley, B., Y. C. Collingham, S. G. Willis, and R. E. Green. 2008. Potential impacts of climatic change on European breeding birds. PloS One 3:e1439.

Husby, A., S. M. Hille, and M. E. Visser. 2011a. Testing mechanisms of Bergmann's rule: phenotypic decline but no genetic change in body size in three passerine bird populations. American Naturalist 178:202–213.

Husby, A., D. H. Nussey, M. E. Visser, A. J. Wilson, B. C. Sheldon, and L. E. B. Kruuk. 2010. Contrasting patterns of phenotypic plasticity in reproductive traits in two great tit (*Parus major*) populations. Evolution 64:2221–2237.

Husby, A., M. E. Visser, and L. E. B. Kruuk. 2011b. Speeding up microevolution: the effects of increasing temperature on selection and genetic variance in a wild bird population. PLoS Biology 9:e1000585.

Jablonka, E. and G. Raz. 2009. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. The Quarterly Review of Biology 84:131–176.

James, J. W. 2009. Genotype by environment interaction in farm animals *in* J. van der Werf, H.-U. Graser, R. Frankham, and C. Gondro, eds. Adaptation and fitness in animal populations: Evolutionary and breeding perspectives on genetic resource management. Springer, Dordrecht, The Netherlands.

Janssen, G. M., G. de Jong, E. N. G. Joosse, and W. Scharloo. 1988. A negative maternal effect in springtails. Evolution 42:828–834.

Jenni, L. and M. Kery. 2003. Timing of autumn bird migration under climate change: advances in long-distance migrants, delays in short-distance migrants. Proceedings of the Royal Society B-Biological Sciences 270:1467–1471.

Jennions, M. D., A. T. Kahn, C. D. Kelly, and H. Kokko. 2012. Meta-analysis and sexual selection: past studies and future possibilities. Evolutionary Ecology 26:1119-1151.

Johnson, P. C., S. J. Barry, H. M. Ferguson, and P. Müller. 2015. Power analysis for generalized linear mixed models in ecology and evolution. Methods in ecology and evolution 6:133-142.

Jonzen, N., A. Hedenstrom, and P. Lundberg. 2007. Climate change and the optimal arrival of migratory birds. Proceedings of the Royal Society B-Biological Sciences 274:269-274.

Keller, L. F. and A. J. Van Noordwijk. 1994. Effects of local environmental conditions on nestling growth in the great tit (*Parus major* L.). Ardea 82:349–362.

Kharouba, H. M., J. Ehrlén, A. Gelman, K. Bolmgren, J. M. Allen, S. E. Travers, and E. M. Wolkovich. 2018. Global shifts in the phenological synchrony of species interactions over recent decades. Proceedings of the National Academy of Sciences of the United States of America 115:5211–5216.

Kim, S., J. A. Fargallo, P. Vergara, and J. Martínez-Padilla. 2013a. Data from: Multivariate heredity of melanin-based coloration, body mass and immunity. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.03mn0.

Kim, S., J. A. Fargallo, P. Vergara, and J. Martínez-Padilla. 2013b. Multivariate heredity of melanin-based coloration, body mass and immunity. Heredity 111:139–146.

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. The American Naturalist 157:245–261.

Kirkpatrick, M. and R. Lande. 1989. The evolution of maternal characters. Evolution 43:485–503.

Kokko, H. 1999. Competition for early arrival in migratory birds. Journal of Animal Ecology 68:940-950.

Kokko, H. and K. Heubel. 2008. Condition-dependence, genotype-by-environment interactions and the lek paradox. Genetica 134:55–62.

Koricheva, J., J. Gurevitch, and K. Mengersen. 2013. Handbook of meta-analysis in ecology and evolution. Princeton University Press.

Kreft, I. G., J. De Leeuw, and L. S. Aiken. 1995. The effect of different forms of centering in hierarchical linear models. Multivariate Behavioral Research 30:1-21.

Krementz, D. G., J. D. Nichols, and J. E. Hines. 1989. Postfledging survival of European starlings. Ecology 70:646–655.

Kristensen, N. P., J. Johansson, J. Ripa, and N. Jonzén. 2015. Phenology of two interdependent traits in migratory birds in response to climate change. Proceedings of the Royal Society B-Biological Sciences 282:20150288.

Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the 'animal model'. Philosophical Transactions of the Royal Society B 359:873–890.

Kruuk, L. E. B. and J. Hadfield. 2007. How to separate genetic and environmental causes of similarity between relatives. Journal of Evolutionary Biology 20:1890–1903.

Kruuk, L. E. B., J. Merilä, and B. C. Sheldon. 2003. When environmental varation short-circuits natural selection. Trends in Ecology & Evolution 18:207–209.

Kuijper, B. and R. B. Hoyle. 2015. When to rely on maternal effects and when on phenotypic plasticity? Evolution 69:950–968.

Kumar, V., E. Gwinner, and T. Van't Hof. 2000. Circadian rhythms of melatonin in European starlings exposed to different lighting conditions: relationship with locomotor and feeding rhythms. Journal of Comparative Physiology A 186:205-215.

Kvist, L., T. Arbabi, M. Paeckert, M. Orell, and J. Martens. 2007. Population differentiation in the marginal populations of the great tit (Paridae: *Parus major*). Biological Journal of the Linnean Society 90:201-210.

Lack, D. 1950. The breeding seasons of European birds. Ibis 92:288-316.

Laine, V. N., E. Atema, P. Vlaming, I. Verhagen, A. C. Mateman, J. J. C. Ramakers, K. van Oers, and M. E. Visser. in review. The genomics of circadian timing in a wild bird, the great tit (*Parus major*).

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

Lande, R. 2009. Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. Journal of Evolutionary Biology 22:1435–1446.

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

Lédon-Rettig, C. C., D. W. Pfennig, A. J. Chunco, and I. Dworkin. 2014. Cryptic genetic variation in natural populations: a predictive framework. Integrative & Comparative Biology 54:783–793.

Lee, Y. and J. A. Nelder. 2006. Double hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society: Series C (Applied Statistics) 55:139-185.

Lehikoinen, A. and R. Virkkala. 2016. North by north-west: climate change and directions of density shifts in birds. Global Change Biology 22:1121–1129.

Leinonen, T., R. S. McCairns, R. B. O'hara, and J. Merilä. 2013. Q_{ST}–F_{ST} comparisons: evolutionary and ecological insights from genomic heterogeneity. Nature Reviews Genetics 14:179.

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. The American Naturalist 138:1315-1341.

Lepage, D., G. Gauthier, and A. Reed. 1998. Seasonal variation in growth of greater snow goose goslings: the role of food supply. Oecologia 114:226-235.

Lessells, C. M. and P. T. Boag. 1987. Unrepeatable repeatabilities: a common mistake. The Auk 104:116–121.

Liedvogel, M., M. Szulkin, S. C. L. Knowles, M. J. Wood, and B. C. Sheldon. 2009. Phenotypic correlates of *Clock* gene variation in a wild blue tit population: evidence for a role in seasonal timing of reproduction. Molecular Ecology 18:2444–2456.

Lillehammer, M., J. Ødegård, and T. H. E. Meuwissen. 2009. Reducing the bias of estimates of genotype by environment interactions in random regression sire models. Genetics Selection Evolution 41:30.

Lindén, A. 2018. Adaptive and nonadaptive changes in phenological synchrony. Proceedings of the National Academy of Sciences of the United States of America 115:5057–5059.

Ljungström, G., M. Stjernstedt, E. Wapstra, and M. Olsson. 2016a. Data from: Selection and constraints on offspring size-number trade-offs in sand lizards (*Lacerta agilis*). Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.700jq.

Ljungström, G., M. Stjernstedt, E. Wapstra, and M. Olsson. 2016b. Selection and constraints on offspring size-number trade-offs in sand lizards (*Lacerta agilis*). Journal of Evolutionary Biology 29:979–990.

Ljungström, G., E. Wapstra, and M. Olsson. 2015. Sand lizard (*Lacerta agilis*) phenology in a warming world. BMC Evolutionary Biology 15:206.

Lof, M. E., T. E. Reed, J. M. McNamara, and M. E. Visser. 2012. Timing in a fluctuating environment: environmental variability and asymmetric fitness curves can lead to adaptively mismatched avian reproduction. Proceedings of the Royal Society B-Biological Sciences 279:3161-3169.

Lortie, C. J., G. Stewart, H. Rothstein, and J. Lau. 2015. How to critically read ecological meta-analyses. Research synthesis methods 6:124-133.

Lynch, M. and R. Lande. 1998. The critical effective size for a genetically secure population. Animal Conservation 1:70–72.

Lynch, M. and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sutherland, MA.

Mameli, M. 2004. Nongenetic selection and nongenetic inheritance. The British Journal for the Philosophy of Science 55:35–71.

Marra, P. P., C. M. Francis, R. S. Mulvihill, and F. R. Moore. 2005. The influence of climate on the timing and rate of spring bird migration. Oecologia 142:307–315.

Marrot, P., A. Charmantier, J. Blondel, and D. Garant. 2018. Current Spring Warming as a Driver of Selection on Reproductive Timing in a Wild Passerine. Journal of Animal Ecology 87:754–764.

Martin, J. G., D. H. Nussey, A. J. Wilson, and D. Réale. 2011. Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. Methods in Ecology and Evolution 2:362-374.

McAdam, A. G. and S. Boutin. 2004. Maternal effects and the response to selection in red squirrels. Proceedings of the Royal Society B-Biological Sciences 271:75–79.

McAdam, A. G., D. Garant, and A. J. Wilson. 2014. The effects of others' genes: maternal and other indirect genetic effects *in* A. Charmantier, D. Garant, and L. E. B. Kruuk, eds. Quantitative Genetics in the Wild. Oxford University Press, Oxford, UK.

McFarlane, S. E., J. C. Gorell, D. W. Coltman, M. M. Humphries, S. Boutin, and A. G. McAdam. 2015. The nature of nurture in a wild mammal's fitness. Proceedings of the Royal Society B-Biological Sciences 282:20142422.

McGlothlin, J. W. and L. F. Galloway. 2013. The contribution of maternal effects to selection response: an empirical test of competing models. Evolution 68:549–558.

McKinnon, E. A., K. C. Fraaser, and B. J. M. Stutchbury. 2013. New discoveris in landbird migration using geolocators, and a flight plan for the future. The Auk 130:211–222.

McKinnon, L., M. Picotin, E. Bolduc, C. Juillet, and J. Bety. 2012. Timing of breeding, peak food availability, and effects of mismatch on chick growth in birds nesting in the High Arctic. Canadian Journal of Zoology-Revue Canadienne De Zoologie 90:961-971.

McLaughlin, J. F., J. J. Hellmann, C. L. Boggs, and P. R. Ehrlich. 2002. Climate change hastens population extinctions. Proceedings of the National Academy of Sciences of the United States of America 99:6070–6074.

Mengersen, K., J. Gurevitch, and M. D. Schmid. 2013. Meta-Analysis of Primary Data. Pp. 300–313 *in* J. Koricheva, J. Gurevitch, and K. Mengersen, eds. Handbook of Meta-analysis in Ecology and Evolution. Princeton University Press, Princeton.

Merilä, J. 2012. Evolution in response to climate change: In pursuit of the missing evidence. BioEssays 34:811–818.

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

Merilä, J., L. E. B. Kruuk, and B. C. Sheldon. 2001a. Natural selection on the genetical component of variance in body condition in a wild bird population. Journal of Evolutionary Biology 14:918–929.

Merilä, J., B. C. Sheldon, and L. E. B. Kruuk. 2001b. Explaining stasis: microevolutionary studies in natural populations. Genetica 112–113:199–222.

Mertens, J. 1977. Thermal conditions for successful breeding in Great Tits (Parus major L.). Oecologia 28:1-29.

Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. Animal frontiers 6:6-14.

Millar, R. B. 2009. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' Factors. Biometrics 65:962–969.

Miller-Rushing, A. J., T. T. Høye, D. W. Inouye, and E. Post. 2010. The effects of phenological mismatches on demography. Philosophical Transactions of the Royal Society B-Biological Sciences 365:3177-3186.

Mills, J. A., C. Teplitsky, B. Arroyo, A. Charmantier, P. H. Becker, T. R. Birkhead, P. Bize, D. T. Blumstein, C. Bonenfant, S. Boutin, A. Bushuev, E. Cam, A. Cockburn, S. D. Cote, J. C. Coulson, F. Daunt, N. J. Dingemanse, B. Doligez, H. Drummond, R. H. Espie, M. Festa-Bianchet, F. Frentiu, J. W. Fitzpatrick, R. W. Furness, D. Garant, G. Gauthier, P. R. Grant, M. Griesser, L. Gustafsson, B. Hansson, M. P. Harris, F. Jiguet, P. Kjellander, E. Korpimaki, C. J. Krebs, L. Lens, J. D. Linnell, M. Low, A. McAdam, A. Margalida, J. Merila, A. P. Moller, S. Nakagawa, J. A. Nilsson, I. C. Nisbet, A. J. van Noordwijk, D. Oro, T. Part, F. Pelletier, J. Potti, B. Pujol, D. Reale, R. F. Rockwell, Y. Ropert-Coudert, A. Roulin, J. S. Sedinger, J. E. Swenson, C. Thebaud, M. E. Visser, S. Wanless, D. F. Westneat, A. J. Wilson, and A. Zedrosser. 2015. Archiving Primary Data: Solutions for Long-Term Studies. Trends in Ecology & Evolution 30:581-589.

Mills, J. A., C. Teplitsky, B. Arroyo, A. Charmantier, P. H. Becker, T. R. Birkhead, P. Bize, D. T. Blumstein, C. Bonenfant, S. Boutin, A. Bushuev, E. Cam, A. Cockburn, S. D. Cote, J. C. Coulson, F. Daunt, N. J. Dingemanse, B. Doligez, H. Drummond, R. H. Espie, M. Festa-Bianchet, F. Frentiu, J. W. Fitzpatrick, R. W. Furness, D. Garant, G. Gauthier, P. R. Grant, M. Griesser, L. Gustafsson, B. Hansson, M. P. Harris, F. Jiguet, P. Kjellander, E. Korpimaki, C. J. Krebs, L. Lens, J. D. Linnell, M. Low, A. McAdam, A. Margalida, J. Merila, A. P. Moller, S. Nakagawa, J. A. Nilsson, I. C. Nisbet, A. J. van Noordwijk, D. Oro, T. Part, F. Pelletier, J. Potti, B. Pujol, D. Reale, R. F. Rockwell, Y. Ropert-Coudert, A. Roulin, J. S. Sedinger, J. E. Swenson, C. Thebaud, M. E. Visser, S. Wanless, D. F. Westneat, A. J. Wilson, and A. Zedrosser. 2016. Solutions for archiving data in long-term studies: A reply to Whitlock et al. Trends in Ecology & Evolution 31:85–87.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Medicine 6:e1000097.

Mols, C. M., K. van Oers, L. M. Witjes, C. M. Lessells, P. J. Drent, and M. E. Visser. 2004. Central assumptions of predator–prey models fail in a semi–natural experimental system. Proceedings of the Royal Society of London B-Biological Sciences 271:S85-S87.

Mols, C. M. M., A. J. Van Noordwijk, and M. E. Visser. 2005. Assessing the reduction of caterpillar numbers by Great Tits *Parus major* breeding in apple orchads. Ardea 93:259–269.

Monaghan, P., R. G. Nager, and D. C. Houston. 1998. The price of eggs: increased investment in egg production reduces the offspring rearing capacity of parents. Proceedings of the Royal Society B-Biological Sciences 265:1731–1735.

Moore, A. J. and P. F. Kukuk. 2002. Quantitative genetic analysis of natural populations. Nature Reviews Genetics 3:971.

Morrissey, M. B. and M. M. Ferguson. 2011. A test for the genetic basis of natural selection: an individual-based longitudinal study in a stream-dwelling fish. Evolution 65:1037–1047.

Morrissey, M. B. and J. D. Hadfield. 2012. Directional selection in temporally replicated studies is remarkably consistent. Evolution 66:435-442.

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. Journal of Evolutionary Biology 23:2277–2288.

Morrissey, M. B. and M. Liefting. 2016. Variation in reaction norms: statistical considerations and biological interpretation. Evolution 70:1944-1959.

Morrissey, M. B., D. J. Parker, P. Korsten, J. M. Pemberton, L. E. B. Kruuk, and A. J. Wilson. 2012a. The prediction of adaptive evolution: empirical application of the Secondary Theorem of Selection and comparison to the Breeder's Equation. Evolution 66:2399–2410.

Morrissey, M. B. and K. Sakrejda. 2013. Unification of regression-based methods for the analysis of natural selection. Evolution 67:2094–2100.

Morrissey, M. B. and K. Sakrejda. 2015. Package 'gsg': Calculation of selection coefficients. https://cran.r-project.org/web/packages/gsg/gsg.pdf.

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

Mousseau, T. A. and C. W. Fox. 1998. The adaptive significance of maternal effects. Trends in Ecology & Evolution 13:403–407.

Mulder, H. A., P. Gienapp, and M. E. Visser. 2016a. Data from: Genetic variation in variability: phenotypic variability of fledging weight and its evolution in a songbird population. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.2qv8n.

Mulder, H. A., P. Gienapp, and M. E. Visser. 2016b. Phenotypic variability of fledging weight and its evolution in a songbird: do families differ genetically? Evolution 70:2004–2016.

Munday, P. L., J. M. Donelson, and J. A. Domingos. 2017. Potential for adaptation to climate change in a coral reef fish. Global Change Biology 23:307–317.

Murren, C. J., H. J. Maclean, S. E. Diamond, U. K. Steiner, M. A. Heskel, C. A. Handelsman, C. K. Ghalambor, J. R. Auld, H. S. Callahan, and D. W. Pfennig. 2014. Evolutionary change in continuous reaction norms. The American Naturalist 183:453-467.

Nager, R. G., C. Ruegger, and A. J. van Noordwijk. 1997. Nutrient or energy limitation on egg formation: a feeding experiment in great tits. Journal of Animal Ecology 66:495–507.

Naguib, M., K. Riebel, A. Marzal, and D. Gil. 2004. Nestling immunocompetence and testosterone covary with brood size in a songbird. Proceedings of the Royal Society B-Biological Sciences 271:833–838.

Nakagawa, S. and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591–605.

Nakagawa, S., D. W. Noble, A. M. Senior, and M. Lagisz. 2017. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. BMC Biology 15:18.

Nakagawa, S. and E. S. A. Santos. 2012. Methodological issues and advances in biological meta-analysis. Evolutionary Ecology 26:1253–1274.

Nakagawa, S. and Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4:133–142.

Nicolaus, M., J. E. Brommer, R. Ubels, J. M. Tinbergen, and N. J. Dingemanse. 2013. Exploring patterns of variation in clutch size–density reaction norms in a wild passerine bird. Journal of Evolutionary Biology 26:2031-2043.

Nielsen, J. T. and A. P. Møller. 2006. Effects of food abundance, density and climate change on reproduction in the sparrowhawk *Accipiter nisus*. Oecologia 149:505–518.

Niemelä, P. T. and N. J. Dingemanse. 2018. Meta-analysis reveals weak associations between intrinsic state and personality. Proceedings of the Royal Society B: Biological Sciences 285:20172823.
Nilsson, J.-Å. 1994. Energetic bottle-necks during breeding and the reproductive cost of being too early. Journal of Animal Ecology 63:200–208.
Nur, N. 1984. The consequences of brood size for breeding blue tits II. Nestling weight, offspring survival and optimal brood size. The Journal of Animal Ecology:497-517.
Nussey, D. H., T. H. Clutton-Brock, D. A. Elston, S. D. Albon, and L. E. Kruuk. 2005a. Phenotypic plasticity in a maternal trait in red deer. Journal of Animal Ecology 74:387-396.
Nussey, D. H., E. Postma, P. Gienapp, and M. E. Visser. 2005b. Selection on heritable phenotypic plasticity in a wild bird population. Science 310:304–306.
Nussey, D. H., E. Postma, P. Gienapp, and M. E. Visser. 2005c. Selection on heritable phenotypic plasticity in a wild bird population. Science 310:304-306.
Nussey, D. H., A. J. Wilson, and J. E. Brommer. 2007. The evolutionary ecology of individual phenotypic plasticity in wild populations. Journal of Evolutionary Biology 20:831–844.

O'Connor, R. J. 1975. The influence of brood size upon metabolic rate and body temperature in nestling Blue Tits Parus caeruleus and House Sparrows Passer domesticus. Journal of Zoology 175:391–403.

Pace, N. R. 2009. Mapping the tree of life: progress and prospects. Microbiology and Molecular Biology Reviews 73:565-576.
Palumbi, S. R. 2001. Humans as the world's greatest evolutionary force. Science 293:1786-1790.
Parker, T. H., W. Forstmeier, J. Koricheva, F. Fidler, J. D. Hadfield, Y. E. Chee, C. D. Kelly, J. Gurevitch, and S. Nakagawa. 2016. Transparency in ecology and evolution: real problems, real solutions. Trends in Ecology & Evolution 31:711-719.
Parmesan, C. 2006. Ecological and evolutionary responses to recent climate change. Annual Review of Ecology and Systematics 37:637–669.
Parmesan, C. and G. Yohe. 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421:37–42.
Perrins, C. and P. Jones. 1974. The inheritance of clutch size in the Great Tit (*Parus major* L.). The Condor 76:225-229.
Perrins, C. M. 1970. The timing of birds' breeding seasons. Ibis 112:242–255.
Pettifor, R. A., C. M. Perrins, and R. H. McCleery. 1988. Individual optimization of clutch size in great tits. Nature 336:160–162.
Pettifor, R. A., C. M. Perrins, and R. H. McCleery. 2001. The individual optimization of fitness: variation in reproductive output, including clutch size, mean nestling mass and offspring recruitment, in manipulated broods of great tits *Parus major*. Journal of Animal Ecology 70:62–79.
Phillimore, A. B., J. D. Hadfield, O. R. Jones, and R. J. Smithers. 2010. Differences in spawning date between populations of common frog reveal local adaptation. Proceedings of the National Academy of Sciences of the United States of America 107:8292–8297.
Phillimore, A. B., S. Stålhandske, R. J. Smithers, and R. Bernard. 2012. Dissecting the contributions of plasticity and local adaptation to the phenology of a butterfly and its host plants. The American Naturalist 180:655-670.
Piersma, T. and J. Drent. 2003. Phenotypic flexibility and the evolution of organismal design. Trends in Ecology & Evolution 18:228-233.
Pigliucci, M. 2001. Phenotypic plasticity; beyond nature and nurture. Synthesis in ecology and evolution. John Hopkins University Press, Baltimore, MD.
Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R. C. Team. 2017. Package 'nlme': Linear and Nonlinear Mixed Effects Models. R package version 3.1–131.
Plaistow, S. J. and T. G. Benton. 2009. The influence of context-dependent maternal effects on population dynamics: an experimental test. Philosophical Transactions of the Royal Society B 364:1049–1058.
Plard, F., J.-M. Gaillard, T. Coulson, A. J. M. Hewison, D. Delorme, C. Warnant, and C. Bonenfant. 2014. Mismatch Between Birth Date and Vegetation Phenology Slows the Demography of Roe Deer. PLoS Biology 12:e1001828.
Porlier, M., A. Charmantier, P. Bourgault, P. Perret, J. Blondel, and D. Garant. 2012. Variation in phenotypic plasticity and selection patterns in blue tit breeding time: between-and within-population comparisons. Journal of Animal Ecology 81:1041-1051.
Postma, E. 2005. Evolutionary genetics of life-history traits in a structured environment: understanding variation in clutch size and laying date in great tits (*Parus major*). PhD thesis Utrecht University. Ponsen & Looijen B.V., Wageningen.

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

Postma, E. 2014. Four decades of estimating heritabilities in wild vertebrate populations: improved methods, more data, better estimates? *in* A. Charmantier, D. Garant, and L. E. B. Kruuk, eds. Quantitative genetics in the wild. Oxford University Press, Oxford.

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

Postma, E. and A. J. van Noordwijk. 2005b. Genetic variation for clutch size in natural populations of birds from a reaction norm perspective. Ecology 86:2344–2357.

Postma, E., J. Visser, and A. J. van Noordwijk. 2007. Strong artificial selection in the wild results in the predicted small evolutionary change. Journal of Evolutionary Biology 20:1823–1832.

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

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

Price, T. and D. Schluter. 1991. On the low heritability of life-history traits. Evolution 45:853–861.

Prizak, R., T. H. G. Ezard, and R. B. Hoyle. 2014. Fitness consequences of maternal and grandmaternal effects. Ecology and Evolution 4:3139–3145.

Pyke, G. H., H. R. Pulliam, and E. L. Charnov. 1977. Optimal foraging: a selective review of theory and tests. The Quarterly Review of Biology 52:137-154.

Ramakers, J. J. C., M. M. P. Cobben, P. Bijma, T. E. Reed, M. E. Visser, and P. Gienapp. 2017. Data from: Maternal effects in a wild songbird are environmentally plastic but only marginally alter the rate of adaptation. Dryad Digital Repository: https://doi.org/10.5061/dryad.4fv71.

Ramakers, J. J. C., A. Culina, M. E. Visser, and P. Gienapp. 2018. Data from: Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations. Dryad Digital Repository. https://doi.org/10.5061/dryad.3mk7f2q.

Ramsay, S. L. and D. Houston. 1997. Nutritional constraints on egg production in the blue tit: a supplementary feeding study. Journal of Animal Ecology 66:649–657.

Räsänen, K. and L. E. B. Kruuk. 2007. Maternal effects and evolution at ecological time-scales. Functional Ecology 21:408–421.

Rausher, M. D. 1992. The measurement of selection on quantitative traits: biases due to environmental covariances between traits and fitness. Evolution 46:616–626.

Réale, D., D. Berteaux, A. McAdam, and S. Boutin. 2003a. Lifetime selection on heritable life-history traits in a natural population of red squirrels. Evolution 57:2416-2423.

Réale, D. and N. J. Dingemanse. 2010. Personality and individual social specialisation. Pp. 417–441 *in* T. Székely, J. Moore, and J. Komdeur, eds. Social Behaviour: Genes, Ecology and Evolution. Cambridge University Press, Cambridge.

Réale, D., A. G. McAdam, S. Boutin, and D. Berteaux. 2003b. Genetic and plastic responses of a northern mammal to climate change. Proceedings of the Royal Society of London B-Biological Sciences 270:591–596.

Réale, D., S. M. Reader, D. Sol, P. T. McDougall, and N. J. Dingemanse. 2007. Integrating animal temperament within ecology and evolution. Biological reviews 82:291-318.

Rechavi, O., G. Minevich, and O. Hobert. 2011. Transgenerational inheritance of an acquired small RNA-based antiviral response in *C. elegans*. Cell 147:1248–1256.

Reed, T. E., P. Gienapp, and M. E. Visser. 2015. Density dependence and microevolution interactively determine effects of phenology mismatch on population dynamics. Oikos 124:81-91.

Reed, T. E., P. Gienapp, and M. E. Visser. 2016a. Data from: Testing for biases in selection on avian reproductive traits and partitioning direct and indirect selection using quantitative genetic models. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.1255v.

Reed, T. E., P. Gienapp, and M. E. Visser. 2016b. Testing for biases in selection on avian reproductive traits and partitioning direct and indirect selection using quantitative genetic models. Evolution 70:2211–2225.

Reed, T. E., V. Grotan, S. Jenouvrier, B.-E. Saether, and M. E. Visser. 2013a. Population Growth in a Wild Bird Is Buffered Against Phenological Mismatch. Science 340:488-491.

Reed, T. E., S. Jenouvrier, and M. E. Visser. 2013b. Phenological mismatch strongly affects individual fitness but not population demography in a woodland passerine. Journal of Animal Ecology 82:131–144.

Reed, T. E., R. S. Waples, D. E. Schindler, J. J. Hard, and M. T. Kinnison. 2010. Phenotypic plasticity and population viability: the importance of environmental predictability. Proceedings of the Royal Society B-Biological Sciences 277:3391–3400.

Regular, P. M., A. Hedd, W. A. Montevecchi, G. J. Robertson, A. E. Storey, and C. J. Walsh. 2014. Why timing is everything: Energetic costs and reproductive consequences of resource mismatch for a chick-rearing seabird. Ecosphere 5.

Reid, J. M. 2012. Predicting evolutionary responses to selection on polyandry in the wild: additive genetic covariances with female extra-pair reproduction. Proc. R. Soc. B:rspb20121835.

Reid, J. M., J. M. Travis, F. Daunt, S. J. Burthe, S. Wanless, and C. Dytham. 2018. Population and evolutionary dynamics in spatially structured seasonally varying environments. Biological Reviews.

Reppert, S. M. and D. R. Weaver. 2002. Coordination of circadian timing in mammals. Nature 418:935.

Reusch, T. B. and T. E. Wood. 2007. Molecular ecology of global change. Molecular Ecology 16:3973-3992.

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology 86:2805-2814.

Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology 45:218-227.

Richardson, D. S., J. Komdeur, and T. Burke. 2004. Inbreeding in the Seychelles warbler: environment-dependent maternal effects. Evolution 58:2037–2048.

Ricklefs, R. E. 1980. Geographical variation in clutch size among passerine birds: Ashmole's hypothesis. The Auk 97:38–49.

Ritland, K. 1996. A marker-based method for inferences about quantitative inheritance in natural populations. Evolution 50:1062-1073.

Rizopoulus, D. 2009. Package 'bootStepAIC': Model selection by bootstrapping the step AIC() procedure. CRAN.

Robertson, A. 1966. A mathematical model of the culling process in dairy cattle. Animal Production 8:95–108.

Roche, D. G., L. E. Kruuk, R. Lanfear, and S. A. Binning. 2015. Public data archiving in ecology and evolution: how well are we doing? PLoS Biology 13:e1002295.

Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, and H. J. Schellnhuber. 2009. A safe operating space for humanity. Nature 461:472.

Rodríguez, S., A. J. van Noordwijk, E. Álvarez, and E. Barba. 2016. A recipe for postfledging survival in great tits *Parus major*: be large and be early (but not too much). Ecology and Evolution 6:4458–4467.

Rönnegård, L., M. Felleki, F. Fikse, H. A. Mulder, and E. Strandberg. 2010. Genetic heterogeneity of residual variance-estimation of variance components using double hierarchical generalized linear models. Genetics Selection Evolution 42:8.

Root, T. L., J. T. Price, K. Hall, R,, S. H. Schneider, C. Rosenzweig, and J. A. Pounds. 2003. Fingerprints of global warming on wild animals and plants. Nature 421:57–60.

Rowe, L., D. Ludwig, and D. Schluter. 1994. Time, condition, and the seasonal decline of avian clutch size. American Naturalist 143:698–722.

Rowiński, P. K. and B. Rogell. 2017. Environmental stress correlates with increases in both genetic and residual variances: A meta-analysis of animal studies. Evolution 71:1339–1351.

Royama, T. 1966. Factors governing feeding rate, food requirement and brood size of nestling great tits Parus major. Ibis 108:313-347.

Royama, T. 1970. Factors governing the hunting behaviour and selection of food by the great tit (Parus major L.). The Journal of Animal Ecology:619-668.

Rubolini, D., N. Saino, and A. P. Møller. 2010. Migratory behaviour constrains the phenological response of birds to climate change. Climate Research 42:45–55.

Ruffino, L., P. Salo, E. Koivisto, P. B. Banks, and E. Korpimäki. 2014. Reproductive responses of birds to experimental food supplementation: a meta-analysis. Frontiers in Zoology 11:80.

Rutkowska, J., M. Cichoń, M. Puerta, and D. Gil. 2005. Negative effects of elevated testosterone on female fecundity in zebra finches. Hormones and Behavior 47:585–591.

Sæther, B.-E., M. E. Visser, V. Grøtan, and S. Engen 2016. Demonstrating r- and K-selection in a wild bird population. Proceedings of the Royal Society B-Biological Sciences 283:20152411.

Sala, O. E., F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, and A. Kinzig. 2000. Global biodiversity scenarios for the year 2100. science 287:1770-1774.

Saladin, V., D. Bonfils, T. Binz, and H. Richner. 2003. Isolation and characterization of 16 microsatellite loci in the Great Tit *Parus major*. Molecular Ecology Notes 3:520-522.

Salewski, V., W. M. Hochachka, and W. Fiedler. 2010. Global warming and Bergmann's rule: do central European passerines adjust their body size to rising temperatures? Oecologia 162:247–260.

Salis, L., E. van den Hoorn, D. G. Beersma, R. A. Hut, and M. E. Visser. 2018. Photoperiodic cues regulate phenological carry-over effects in an herbivorous insect. Functional Ecology 32:171-180.

Santure, A. W., I. De Cauwer, M. R. Robinson, J. Poissant, B. C. Sheldon, and J. Slate. 2013. Genomic dissection of variation in clutch size and egg mass in a wild great tit (Parus major) population. Molecular Ecology 22:3949-3962.

Santure, A. W., J. Poissant, I. De Cauwer, K. van Oers, M. R. Robinson, J. L. Quinn, M. A. M. Groenen, M. E. Visser, B. C. Sheldon, and J. Slate. 2015a. Data from: Replicated analysis of the genetic architecture of quantitative traits in two wild great tit populations. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.5t32v.

Santure, A. W., J. Poissant, I. De Cauwer, K. van Oers, M. R. Robinson, J. L. Quinn, M. A. M. Groenen, M. E. Visser, B. C. Sheldon, and J. Slate. 2015b. Replicated analysis of the genetic architecture of quantitative traits in two wild great tit populations. Molecular Ecology 24:6148–6162.

Sanz, J. J. 1997. Clutch size manipulation in the Pied Flycather: effects on nestling growth, parental care and moult. Journal of Avian Biology 28:157–162.

Sanz, J. J., S. Kranenbarg, and J. M. Tinbergen. 2000. Differential response by males and females to manipulation of partner contribution in the great tit (Parus major). Journal of Animal Ecology 69:74-84.

Sasvari, L. 1986. Reproductive effort of widowed birds. The Journal of Animal Ecology:553-564.

Schaefer, H.-K., W. Jetz, and K. Böhning-Gaese. 2008. Impact of climate change on migratory birds: community reassembly versus adaptation. Global Ecology and Biogeography 17:38–49.

Schaeffer, L. R. 2004. Application of random regression models in animal breeding. Livestock Production Science 86:35–45.

Schaper, S. V., A. Dawson, P. J. Sharp, P. Gienapp, S. P. Caro, and M. E. Visser. 2012. Increasing Temperature, Not Mean Temperature, Is a Cue for Avian Timing of Reproduction. American Naturalist 179:E55-E69.

Scheffers, B. R., L. De Meester, T. C. Bridge, A. A. Hoffmann, J. M. Pandolfi, R. T. Corlett, S. H. Butchart, P. Pearce-Kelly, K. M. Kovacs, and D. Dudgeon. 2016. The broad footprint of climate change from genes to biomes to people. Science 354:aaf7671.

Scheiner, S. M. 1993. Genetics and evolution of phenotypic plasticity. Annual Review of Ecology and Systematics 24:35-68.

Schlichting, C. D. and M. Pigliucci. 1998. Phenotypic evolution: a reaction norm perspective. Sinauer Associates, Sunderland, MA.

Schluter, D. and L. Gustafsson. 1993. Maternal inheritance of condition and clutch size in the collared flycatcher. Evolution 47:658–667.

Schultz, M. T., J. F. Piatt, A. M. A. Harding, A. B. Kettle, and T. I. Van Pelt. 2009. Timing of breeding and reproductive performance in mirres and kittiwakes reflect mismatched seasonal prey dynamics. Marine Ecology Progress Series 393:247–258.

Schuurman, N. K., R. P. P. P. Grasman, and E. L. Hamaker. 2016. A comparison of Inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. Multivariate Behavioral Research 51:185–206.

Sekercioglu, C. H., S. H. Schneider, J. P. Fay, and S. R. Loarie. 2008. Climate change, elevational range shifts, and bird extinctions. Conservation Biology 22:140–150.

Sheldon, B. C., L. E. B. Kruuk, and J. Merilä. 2003. Natural selection and inheritance of breeding time and clutch size in the collared flycatchers. Evolution 57:406–420.

Sheridan, J. A. and D. Bickford. 2011. Shrinking body size as an ecological response to climate change. Nature Climate Change 1:401–406.

Siepielski, A. M., D. DiBattista, and S. M. Carlson. 2009. It's about time: the temporal dynamics of phenotypic selection in the wild. Ecology Letters 12:1261–1276.

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. Ecology letters 16:1382-1392.

Siikamäki, P. 1998. Limitation of reproductive success by food availability and breeding time in pied flycatchers. Ecology 79:1789-1796.

Simmonds, M. C., J. P. Higginsa, L. A. Stewartb, J. F. Tierneyb, M. J. Clarke, and S. G. Thompson. 2005. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clinical Trials 2:209-217.

Singer, M. C. and C. Parmesan. 2010. Phenological asynchrony between herbivorous insects and their hosts: signal of climate change or pre-existing adaptive strategy? Philosophical Transactions of the Royal Society B-Biological Sciences 365:3161-3176.

Small-Lorenz, S. L., L. A. Culp, T. B. Ryder, T. C. Will, and P. P. Marra. 2013. A blind spot in climate change vulnerability assessments. Nature Climate Change 3:91–93.

Smith, J. J., S. T. Hasiotis, M. J. Kraus, and D. T. Woody. 2009. Transient dwarfism of soil fauna during the Paleocene–Eocene Thermal Maximum. Proceedings of the National Academy of Sciences of the United States of America 106:17655–17660.

Smith, R. J. and F. R. Moore. 2005. Arrival timing and seasonal reproductive performance in a long-distance migratory landbird. Behavioral Ecology and Sociobiology 57:231-239.

Soo, R. M., J. Hemp, D. H. Parks, W. W. Fischer, and P. Hugenholtz. 2017. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. Science 355:1436-1440.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society B 64:583–639.

Spoelstra, K., I. Verhagen, D. Meijer, and M. E. Visser. 2018. Artificial light at night shifts daily activity patterns but not the internal clock in the great tit (Parus major). Proc. R. Soc. B 285:20172751.

Stearns, S. C. and T. J. Kawecki. 1994. Fitness sensitivity and the canalization of life-history traits. Evolution 48:1438–1450.

Steffen, W., W. Broadgate, L. Deutsch, O. Gaffney, and C. Ludwig. 2015. The trajectory of the Anthropocene: the great acceleration. The Anthropocene Review 2:81-98.

Stenseth, N. C. and A. Mysterud. 2002a. Climate, changing phenology, and other life history and traits: Nonlinearity and match-mismatch to the environment. Proceedings of the National Academy of Sciences of the United States of America 99:13379-13381.

Stenseth, N. C. and A. Mysterud. 2002b. Climate, changing phenology, and other life history traits: nonlinearity and match–mismatch to the environment. Proc Natl Acad Sci USA 99:13379–13381.

Steppan, S. J., P. C. Phillips, and D. Houle. 2002. Comparative quantitative genetics: evolution of the G matrix. Trends in Ecology & Evolution 17:320-327.

Stevenson, I. R. and D. M. Bryant. 2000. Avian phenology: Climate change and constraints on breeding. Nature 406:366–367.

Stewart, G. B. and C. H. Schmid. 2015. Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies. Research synthesis methods 6:109-110.

Stinchcombe, J. R., M. T. Rutter, D. S. Burdick, P. Tiffin, M. D. Rausher, and R. Mauricio. 2002. Testing for environmentally induced bias in phenotypic estimates of natural selection: theory and practice. The American Naturalist 160:511-523.

Stutchbury, B. J. M., S. A. Tarof, T. Done, E. Gow, P. M. Kramer, J. Tautin, J. W. Fox, and V. Afanasyev. 2009. Tracking long-distance songbird migration by using geolocators. Science 323:896.

Szyf, M., I. C. G. Weaver, and F. A. Champagne. 2005. Maternal programming of steroid receptor expression and phenotype through DNA methylation. Frontiers in Neuroendocrinology 26:139–162.

Takahashi, J. S., F. W. Turek, and R. Y. Moore. 2001. Circadian clocks. Handbook of Behavioral Neurobiology. Springer US, Boston, MA.

Taylor, J. 1997. Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements. University Science Books, New York.

Te Marvelde, L., S. V. Schaper, and M. E. Visser. 2012a. A Single Long Day Triggers Follicle Growth in Captive Female Great Tits (Parus major) in Winter but Does Not Affect Laying Dates in the Wild in Spring. PLoS One 7.

Te Marvelde, L. and M. E. Visser. 2012. Manipulation of Life-History Decisions Using Leptin in a Wild Passerine. PLoS One 7:e34090.

Te Marvelde, L., S. L. Webber, H. A. J. Meijer, and M. E. Visser. 2011. Mismatched reproduction is energetically costly for chick feeding female great tits. Functional Ecology 25:1302–1308.

te Marvelde, L., S. L. Webber, H. A. J. Meijer, and M. E. Visser. 2012b. Energy expenditure during egg laying is equal for early and late breeding free-living female great tits. Oecologia 168:631-638.

Teplitsky, C. and V. Millien. 2014. Climate warming and Bergmann's rule through time: is there any evidence? Evolutionary Applications 7:156–168.

Teplitsky, C., J. A. Mills, J. S. Alho, J. W. Yarrall, and J. Merilä. 2008. Bergmann's rule and climate change revisited: Disentangling environmental and genetic responses in a wild bird population. Proceedings of the National Academy of Sciences of the United States of America 105:13492–13496.

Teplitsky, C., N. G. Mouawad, J. Balbontín, F. de Lope, and A. P. Møller. 2011a. Data from: Quantitative genetics of migration syndromes: a study of two barn swallow populations. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.41np7.

Teplitsky, C., N. G. Mouawad, J. Balbontín, F. de Lope, and A. P. Møller. 2011b. Quantitative genetics of migration syndromes: a study of two barn swallow populations. Journal of Evolutionary Biology 24:2025–2039.

Thackeray, S. J., P. A. Henrys, D. Hemming, J. R. Bell, M. S. Botham, S. Burthe, P. Helaouet, D. G. Johns, I. D. Jones, and D. I. Leech. 2016. Phenological sensitivity to climate across taxa and trophic levels. Nature 535:241.

Thackeray, S. J., T. H. Sparks, M. Frederiksen, S. Burthe, P. J. Bacon, J. R. Bell, M. S. Botham, T. M. Brereton, P. W. Bright, L. Carvalho, T. Clutton-Brock, A. Dawson, M. Edwards, J. M. Elliott, R. Harrington, D. Johns, I. D. Jones, J. T. Jones, D. I. Leech, D. B. Roy, W. A. Scott, M. Smith, R. J. Smithers, I. J. Winfield, and S. Wanless. 2010. Trophic level asynchrony in rates of phenological change for marine, freshwater and terrestrial environments. Global Change Biology 16:3304-3313.

Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. Ferreira de Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. van Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. T. Peterson, O. L. Phillips, and S. E. Williams. 2004. Extinction risk from climate change. Nature 427:145–148.

Thomas, C. D. and J. J. Lennon. 1999. Birds extend their ranges northwards. Nature 399:213.

Thomas, D. W., J. Blondel, P. Perret, M. M. Lambrechts, and J. R. Speakman. 2001. Energetic and fitness costs of mismatching resource supply and demand in seasonally breeding birds. Science 291:2598–2600.

Thomson, C. E., F. Bayer, N. Crouch, S. Farrell, E. Heap, E. Mittell, M. Zurita-Cassinello, and J. Hadfield. 2017a. Data from: Selection on parental performance opposes selection for larger body mass in a wild population of blue tits. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.m2c10.

Thomson, C. E., F. Bayer, N. Crouch, S. Farrell, E. Heap, E. Mittell, M. Zurita-Cassinello, and J. Hadfield. 2017b. Selection on parental performance opposes selection for larger body mass in a wild population of blue tits. Evolution 71:716–732.

Tilgar, V., R. Mänd, P. Kilgas, and M. Mägi. 2010. Long-term consequences of early ontogeny in free-living Great Tits *Parus major*. Journal of Ornithology 151:61–68.

Tinbergen, J. and M. Dietz. 1994. Parental energy expenditure during brood rearing in the great tit (*Parus major*) in relation to body mass, temperature, food availability and clutch size. Functional Ecology:563-572.

Tinbergen, J. M. and C. Both. 1999. Is clutch size individually optimized? Behavioral Ecology 10:504–509.

Tinbergen, J. M. and S. Daan. 1990. Family planning in the great tit (*Parus major*): optimal clutch size as integration of parent and offspring fitness. Behaviour 114:161-190.

Tobi, E. W., J. J. Goeman, R. Monajemi, H. Gu, H. Putter, Y. Zhang, R. C. Slieker, A. P. Stok, P. E. Thijssen, F. Müller, E. W. van Zwet, C. Bock, A. Meissner, L. H. Lumey, P. Eline Slagboom, and B. T. Heijmans. 2014. DNA methylation signatures link prenatal famine exposure to growth and metabolism. Nature Communications 5:5592.

Tomkins, J. L., J. Radwan, J. S. Kotiaho, and T. Tregenza. 2004. Genic capture and resolving the lek paradox. Trends in Ecology & Evolution 19:323-328.

Tomotani, B. M., F. T. Muijres, J. Koelman, S. Casagrande, and M. E. Visser. 2017. Simulated moult reduces flight performance but overlap with breeding does not affect breeding success in a long-distance migrant. Functional Ecology:1–13.

Torti, V. M. and P. O. Dunn. 2005. Variable effects of climate change on six species of North American birds. Oecologia 145:486–495.

Toupoint, N., L. Gilmore-Solomon, F. Bourque, B. Myrand, F. Pernet, F. Olivier, and R. Tremblay. 2012. Match/mismatch between the Mytilus edulis larval supply and seston quality: effect on recruitment. Ecology 93:1922-1934.

Trivelpiece, W. Z., J. T. Hinke, A. K. Miller, C. S. Reiss, S. G. Trivelpiece, and G. M. Watters. 2011. Variability in krill biomass links harvesting and climate warming to penguin population changes in Antarctica. Proceedings of the National Academy of Sciences of the United States of America 108:7625–7628.

Turelli, M. and N. H. Barton. 2004. Polygenic Variation Maintained by Balancing Selection: Pleiotropy, Sex-Dependent Allelic Effects and G x E Interactions. Genetics 166:1053–1079.

Tylianakis, J. M., R. K. Didham, J. Bascompte, and D. A. Wardle. 2008. Global change and species interactions in terrestrial ecosystems. Ecology Letters 11:1351-1363.

Uller, T. 2008. Developmental plasticity and the evolution of parental effects. Trends in Ecology & Evolution 23:432–438.

Van Asch, M., L. Salis, L. J. M. Holleman, B. van Lith, and M. E. Visser. 2013. Evolutionary response of the egg hatching date of a herbivorous insect under climate change. Nature Climate Change 3:244–248.

Van Asch, M., P. H. Van Tienderen, L. J. M. Holleman, and M. E. Visser. 2007. Predicting adaptation of phenology in response to climate change, an insect herbivore example. Global Change Biology 13:1596-1604.

Van Asch, M. and M. E. Visser. 2007. Phenology of forest caterpillars and their host trees: The importance of synchrony. Annual Review of Entomology 52:37–55.

Van Balen, J. H. 1973. A Comparative Sudy of the Breeding Ecology of the Great Tit *Parus major* in Different Habitats. Ardea 61:1–93.

Van de Pol, M. 2012. Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. Methods in Ecology and Evolution 3:268-280.

Van de Pol, M., L. D. Bailey, N. McLean, L. Rijsdijk, C. R. Lawson, and L. Brouwer. 2016. Identifying the best climatic predictors in ecology and evolution. Methods in Ecology and Evolution 7:1246-1257.

Van de Pol, M., H. L. Osmond, and A. Cockburn. 2012. Fluctuations in population composition dampen the impact of phenotypic plasticity on trait dynamics in superb fairy-wrens. Journal of Animal Ecology 81:411-422.

Van de Pol, M. and J. Wright. 2009. A simple method for distinguishing within-versus between-subject effects using mixed models. Animal behaviour 77:753-758.

Van der Graaf, S., J. Stahl, A. Klimkowska, J. P. Bakker, and R. H. Drent. 2006. Surfing on a green wave – how plant growth drives migration in the Barnacle Goose *Branta leucopsis*. Ardea 94:567–577.

Van der Jeugd, H. P., G. Eichhorn, K. E. Litvin, J. Stahl, K. Larsson, A. J. van der Graaf, and R. H. Drent. 2009. Keeping up with early springs: rapid range expansion in an avian herbivore incurs a mismatch between reproductive timing and food supply. Global Change Biology 15:1057–1071.

Van der Jeugd, H. P. and R. McCleery. 2002. Effects of spatial autocorrelation, natal philopatry and phenotypic plasticity on the heritability of laying date. Journal of Evolutionary Biology 15:380–387.

Van der Putten, W. H., M. Macel, and M. E. Visser. 2010. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. Philosophical Transactions of the Royal Society B 365:2025–2034.

Van Noordwijk, A. J., R. H. McCleery, and C. M. Perrins. 1995. Selection for the timing of great tit breeding in relation to caterpillar growth and temperature. Journal of Animal Ecology 64:451–458.

Van Noordwijk, A. J., J. H. van Balen, and W. Scharloo. 1980. Heritability of ecologically important traits in the great tit. Ardea 68:193–203.

Van Noordwijk, A. J., J. H. van Balen, and W. Scharloo. 1981. Genetic variation in the timing of reproduction in the great tit. Oecologia 49:158–166.

Van Oers, K., P. J. Drent, N. J. Dingemanse, and B. Kempenaers. 2008. Personality is associated with extrapair paternity in great tits, *Parus major*. Animal Behaviour 76:555–563.

Van Tienderen, P. H. and H. P. Koelewijn. 1994. Selection on reaction norms, genetic correlations and constraints. Genetics Research 64:115-125.

Vatka, E., M. Orell, and S. RytkÖnen. 2011. Warming climate advances breeding and improves synchrony of food demand and food availability in a boreal passerine. Global Change Biology 17:3002-3009.

Vaughn, K. J. and T. P. Young. 2010. Contingent Conclusions: Year of Initiation Influences Ecological Field Experiments, but Temporal Replication is Rare. Restoration Ecology 18:59-64.

Vaze, K. M. and V. K. Sharma. 2013. On the adaptive significance of circadian clocks for their owners. Chronobiology International 30:413-433.

Veiga, J. P. 1992. Hatching asynchrony in the house sparrow: a test of the egg-viability hypothesis. The American Naturalist 139:669-675.

Verboven, N. and M. E. Visser. 1998. Seasonal variation in local recruitment of great tits: the importance of being early. Oikos:511-524.

Verhagen, I., P. Gienapp, V. N. Laine, A. C. Mateman, K. Van Oers, and M. E. Visser. in review. Genetic and phenotypic responses to genomic selection for timing of breeding in a wild songbird. Functional Ecology.

Verhoeven, K. J. F., B. M. vonHoldt, and V. L. Sork. 2016. Epigenetics in ecology and evolution: what we know and what we need to know. Molecular Ecology 25:1631–1638.

Verhulst, S. and J.-Å. Nilsson. 2008. The timing of birds' breeding seasons: a review of experiments that manipulated timing of breeding. Philosophical Transactions of the Royal Society B 363:399–410.

Via, S., R. Gomulkiewicz, G. De Jong, S. M. Scheiner, C. D. Schlichting, and P. H. Van Tienderen. 1995. Adaptive phenotypic plasticity: consensus and controversy. Trends in Ecology & Evolution 10:212–217.

Visser, M. E. 2008. Keeping up with a warming world; assessing the rate of adaptation to climate change. Proceedings of the Royal Society B-Biological Sciences 275:649–659.

Visser, M. E. and C. Both. 2005. Shifts in phenology due to global climate change: the need for a yardstick. Proceedings of the Royal Society B-Biological Sciences 272:2561-2569.

Visser, M. E., C. Both, and M. M. Lambrechts. 2004a. Global climate change leads to mistimed avian reproduction. Advances in Ecological Research 35:89–110.

Visser, M. E., C. Both, and M. M. Lambrechts. 2004b. Global climate change leads to mistimed avian reproduction. Pp. 89-110 *in* A. P. Moller, W. Fielder, and P. Berthold, eds. Birds and Climate Change.

Visser, M. E., S. P. Caro, K. van Oers, S. Schaper, and B. Helm. 2010. Phenology, seasonal timing and circannual rhythms: towards a unified framework. Phil Trans R Soc B 365:3113–3127.

Visser, M. E. and P. Gienapp. in press. Seasonal timing in a warming world: a review of the phenological mismatch hypothesis. Nature Ecology & Evolution.

Visser, M. E., P. Gienapp, A. Husby, M. B. Morrissey, I. de la Hera, F. Pulido, and C. Both. 2015. Effects of spring temperatures on the strength of selection on timing of reproduction in a long-distance migratory bird. PLoS Biology 13:e1002120.

Visser, M. E. and L. J. M. Holleman. 2001. Warmer springs disrupt the synchrony of oak and winter moth phenology. Proceedings of the Royal Society B-Biological Sciences 268:289–294.

Visser, M. E., L. J. M. Holleman, and S. P. Caro. 2009a. Temperature has a causal effect on avian timing of reproduction. Proceedings of the Royal Society B-Biological Sciences 276:2323–2331.

Visser, M. E., L. J. M. Holleman, and P. Gienapp. 2006. Shifts in caterpillar biomass phenology due to climate change and its impact on the breeding biology of an insectivorous bird. Oecologia 147:164–172.

Visser, M. E. and C. M. Lessells. 2001. The cost of egg production and incubation in great tits (*Parus major*). Proceedings of the Royal Society of London B-Biological Sciences 268:1271–1277.

Visser, M. E., A. C. Perdeck, J. H. van Balen, and C. Both. 2009b. Climate change leads to decreasing bird migration distances. Global Change Biology 15:1859–1865.

Visser, M. E., B. Silverin, M. M. Lambrechts, and J. M. Tinbergen. 2002. No evidence for tree phenology as a cue for the timing of reproduction in tits *Parus* spp. Avian Science 2:77–86.

Visser, M. E., L. te Marvelde, and M. E. Lof. 2012. Adaptive phenological mismatches of birds and their food in a warming world. Journal of Ornithology 153 (Suppl 1):S75–S84.

Visser, M. E., A. J. van Noordwijk, J. M. Tinbergen, and C. M. Lessells. 1998. Warmer springs lead to mistimed reproduction in great tits (*Parus major*). Proceedings of the Royal Society B-Biological Sciences 265:1867–1870.

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

Wallis, J. C., E. Rolando, and C. L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS One 8:e67332.

Walther, G.-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Bairlein. 2002. Ecological responses to recent climate change. Nature 416:389–395.

Weaver, I. C. G., N. Cervoni, and F. A. Champagne. 2004. Epigenetic programming by maternal behavior. Nature Neuroscience 7:847–854.

Weese, D. J., S. P. Gordon, A. P. Hendry, and M. T. Kinnison. 2010. Spatiotemporal variation in linear natural selection on body color in wild guppies (*Poecilia reticulata*). Evolution 64:1802–1815.

Weigensberg, I. and D. A. Roff. 1996. Natural heritabilities: can they be reliably estimated in the laboratory? Evolution 50:2149-2157.

Weis, A. E. and W. L. Gorman. 1990. Measuring selection on reaction norms: an exploration of the Eurosta-Solidago system. Evolution 44:820-831.

Westneat, D. F., M. Schofield, and J. Wright. 2012. Parental behavior exhibits among-individual variance, plasticity, and heterogeneous residual variance. Behavioral Ecology 24:598–604.

Westneat, D. F., J. Wright, and N. J. Dingemanse. 2015. The biology hidden inside residual within-individual phenotypic variation. Biological Reviews 90:729-743.

Wheelwright, N. T., L. F. Keller, and E. Postma. 2014a. Data from: The effect of trait type and strength of selection on heritability and evolvability in an island bird population. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.dv0qt.2.

Wheelwright, N. T., L. F. Keller, and E. Postma. 2014b. The effect of trait type and strength of selection on heritability and evolvability in an island bird population. Evolution 68:3325–3336.

Whitlock, M. C., J. L. Bronstein, E. M. Bruna, A. M. Elison, C. W. Fox, M. A. McPeek, A. J. Moore, M. A. F. Noor, M. D. Rausher, L. Rieseberg, M. G. Ritchie, and R. G. Shaw. 2016. A balanced data archiving policy for long-term studies. Trends in Ecology & Evolution 31:84–85.

Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data archiving. American Naturalist 175:145–146.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, and P. E. Bourne. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018.

Willham, R. L. 1963. The covariance between relatives for characters composed of components contributed by related individuals. Biometrics 19:8–27.

Williams, T. D. 2012. Physiological adaptations for breeding in birds. Princeton University Press, Princeton, New Jersey.

Wilson, A. J., D. W. Coltman, J. M. Pemberton, A. D. J. Overall, K. A. Byrne, and L. E. B. Kruuk. 2005. Maternal genetic effects set the potential for evolution in a free-living vertebrate population. Journal of Evolutionary Biology 18:405–414.

Wilson, A. J., J. M. Pemberton, J. G. Pilkington, D. W. Coltman, D. V. Mifsud, T. H. Clutton-Brock, and L. E. B. Kruuk. 2006. Environmental coupling of selection and heritability limits evolution. PLoS Biology 4:e216.

Wilson, A. J., D. Réale, M. N. Clements, M. M. Morrissey, E. Postma, C. A. Walling, L. E. B. Kruuk, and D. H. Nussey. 2010. An ecologist's guide to the animal model. Journal of Animal Ecology 79:13–26.

Winkler, D. W. and P. E. Allen. 1996. The seasonal decline in tree swallow clutch size: physiological constraint or strategic adjustment? Ecology 77:922–932.

Wolf, J. B., E. D. Brodie III, J. M. Cheverud, A. J. Moore, and M. J. Wade. 1998. Evolutionary consequences of indirect genetic effects. Trends in Ecology & Evolution 13:64–69.

Woltereck, R. 1909. Weitere experimentelle Untersuchungen über Artveränderung, speziel über das Wesen quantitativer Artunterschiede bei Daphniden. Verhandlungen der Deutschen Zoologischen Gesellschaft 19:110–173.

Wood, C. W. and E. D. Brodie III. 2015. Environmental effects on the structure of the G-matrix. Evolution 69:2927-2940.

Wood, C. W. and E. D. Brodie III. 2016. Evolutionary response when selection and genetic variation covary across environments. Ecology Letters 19:1189–1200.

Wood, S. 2017. Package 'mgcv': Mixed GAM Computation Vehicle with Automatic Smoothness Automation. https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

Yanagi, S.-i. and M. Tuda. 2010. Interaction effect among maternal environment, maternal investment and progeny genotype on life history traits in *Callosobruchus chinensis*. Functional Ecology 24:383–391.

Yates, F. and W. G. Cochran. 1938. The analysis of groups of experiments. Journal of Agricultural Science 28:556–580.

Yee, T. W. 2010. The VGAM package for categorical data analysis. Journal of Statistical Software 32:1-34.

Summary

The environment is changing and this is exerting selection pressures on wild populations. For example, the timing of phenological events such as reproduction and migration are driven by temperatures and climate change is leading to a differential shift in timing of phenology among trophic levels, in some cases leading to selection on consumer phenology. Individuals are often phenotypically plastic, meaning that they can change their phenotype (e.g. breeding time) in response to environmental conditions. This allows them to track the changing environment to some degree but ultimately a genetic change is necessary to safeguard populations from extinction in the long run. Many wild populations so far, however, could not be shown to be undergoing any genetic adaptation (e.g. a shift in their phenology) over time. Quantitative genetics, i.e. the study of the genetics of quantitative (polygenic) traits, is commonly used to identify the evolutionary parameters (genetic variation and selection) in wild populations to predict evolution, but for such predictions to be successful we need to understand the ecological factors underlying (constraints to) adaptation. In this thesis, I aimed to get an understanding of how populations are coping with environmental change and which ecological processes affect the rate of adaptation. I did this using a combined approach of field experiments to identify ecological constraints and long-term observations to make evolutionary predictions in the great tit (*Parus major*) and other vertebrate species.

In the first part of my thesis (**Chapter 2**), I provided a broad overview of what is known about the effects of climate change on the general biology of birds. Birds are affected by climate change in several ways: they may change (1) their geographical distribution due to a shifting 'bioclimatic envelope', (2) advance their timing of phenological events such as breeding and migration, (3) undergo morphological changes (e.g. in body size) as an energetic adaptation, and (4) undergo demographic changes as a direct result of changes in reproductive success or survival. There has been a strong bias in the literature on phenology and therefore we still have a lot to learn about the ecological and evolutionary consequences related to these other aspects of phenotypic change. Whether observed changes in phenology are due to plasticity or due to genetic change remains an open question.

Dutch great tits have been under increased selection for earlier laying due to increased mismatch with the caterpillar peak (the main food for their nestlings), but we see little (phenotypic) response. The lack of a response may be caused by an energetic constraint associated with breeding too early under harsh conditions, such that birds that do breed earlier may pay fitness costs. In the second part of my thesis I aimed to test whether birds were constrained to breed earlier. In **Chapter 3**, I used experimental food supplementation food prior to and during egg-laying to test whether females that were tricked into laying early would pay fitness costs (due to brood desertion or reduced chick-provisioning efforts) once food supplementation ceased upon the start of laying. Food supplementation was not effective at advancing laying, and any food increased, rather than decreased, fitness. Because food supplementation alone is not sufficient to test the

constraints hypothesis (e.g. because it cannot distinguish whether birds are constrained by food or are just missing the essential cues to advance their breeding) we need an additional, clean manipulation of egg-laying date that does not affect the body condition of a female. In **Chapter 4**, I described the first results of a large-scale experiment in which great tits are genomically selected to breed early or late. Eggs produced by females from these selection lines were brought to the wild and raised by foster parents. I showed that selection lines (late vs early) did not differ in any aspect of early-life fitness (fledging success, nestling weight at fledging), but that the fitness parameters differed slightly between selection-line birds and their wild counterparts. Since only 11 birds from these fostered birds survived until breeding in 2018 (including 2 females from the early and 3 from the late selection line), we could not test whether these birds indeed bred respectively earlier or later, or whether earlier laying indeed led to higher fitness costs. I concluded that multiple years (with different environmental conditions and an increased sample size) would be needed to conclude whether birds are indeed constrained to breed earlier.

Ultimately, breeding success in great tits is largely determined by the match of the offspring needs with the caterpillar abundance. In **Chapter 5**, I explored the notion that to clearly understand phenological mismatch—and to determine whether birds really are mismatched—we need a thorough, temporal description of offspring needs and food availability to quantify the amount of temporal overlap between these distributions. I found that the classical way of defining mismatch, i.e. the difference in peak dates between great tit and caterpillar phenology, outperformed a more comprehensive measure that described the temporal overlap in a model explaining variation in offspring survival and selection for laying date. I concluded that a simple measure of mismatch in highly seasonal study systems is likely to be best for describing demographic processes, and that more complex measures are likely infeasible in most practical situations.

In the third part of my thesis, I deployed state-of-the-art quantitative genetic modelling approaches to unravel patterns of selection, genetic variation, and evolutionary response to selection in a reaction-norm context using long-term, pedigreed datasets of wild populations. Some methods to achieve this were explored in the preceding Intermezzo (**Chapter 6** and **7**). In **Chapter 8**, I investigated whether maternal effects as a form of transgenerational plasticity could affect the rate of adaptation in great tits. Using experimental and long-term data, I was able to show that the clutch size of a great tit is partly dependent on her body weight at fledgling and that it is negatively associated with the clutch size of her own mother. Such a negative maternal effect could constrain adaptation to a novel environment with selection for a larger of smaller clutch. We showed by simulation, however, that this negative maternal effect would likely have little impact on the rate of adaptation.

Phenological changes over time do not always match evolutionary predictions; one potential reason for this discrepancy is an unrecognised environmentally induced coupling between selection and the heritability of the trait. In **Chapter 9**, I investigated how general such a coupling is in wild vertebrate populations, and whether such a coupling affects the expected rate of adaptation. The expectation was that if heritability and selection are negatively associated, this constrains adaptation because little genetic

variation is present under strong selection and vice versa. Making use of openly available datasets (see **Chapter 7**), we managed to estimate environment-specific heritability and selection in 50 traits from 16 populations of 10 species. We found that heritability and selection are only rarely associated and that this association is an unlikely explanation for apparent evolutionary stasis observed in wild populations.

Great tits respond strongly to temperature through phenotypic plasticity; this plasticity is described by a reaction norm, the linear function consisting of an elevation (the laying date in the average environment) and a slope (the sensitivity to the environment). Since different individuals have different reaction norms, selection on laying date may result in an evolutionary shift in the reaction norm. In **Chapter 10**, I found that individual great tits differ genetically in the elevation of the reaction norm, but not in its slope, and this reaction norm is under selection due to the advance in the caterpillar peak over time. I predicted quantitatively, however, that such evolution has been—and will be—too slow to be detected due to the high environmental variability in laying dates.

To conclude, I investigated the evolutionary potential of populations and aimed to identify ecological constraints in adaptation. I found that there is still a lot we need to learn about the ecological and evolutionary consequences of climate change beyond the few well-known study systems, including effects on demography and population viability. Experiments aimed at unravelling the fitness costs of breeding too early are inconclusive and warrant further investigation (with more samples and multiple environments). Powerful quantitative genetic tools are available to evolutionary ecologists to quantify evolutionary trajectories but these models must be based on reality to obtain reliable predictions. I have suggested in this thesis that realistic predictions could be benefited by the integration of multiple data sources (i.e. long-term observational and experimental data) and simulations. The use of open data can aid in achieving this through the answering of novel research questions at a broad taxonomic or geographic scale. Most importantly, we need a thorough understanding of the most important components of the ecosystem of our study species. Only then can we make sense of our evolutionary predictions.

Samenvatting

Onze omgeving is aan het veranderen en dit leidt tot veranderingen in de selectiedruk in wilde populaties. De timing van fenologische (ofwel seizoensgebonden) processen als voortplanting en migratie wordt bijvoorbeeld in belangrijke mate gedreven door temperatuur; klimaatverandering zorgt voor een differentiële verschuiving in de timing van fenologie tussen trofische lagen (bijv. tussen predator en prooi) en dit leidt in sommige gevallen tot een versterkte selectiedruk voor een vroegere timing bij de predator. Individuen vertonen vaak 'fenotypische plasticiteit': ze zijn in staat hun fenotype aan te passen aan de omgeving. Dit stelt hen in staat om tot op zekere hoogte met veranderingen in de omgeving om te gaan, maar deze plasticiteit zal niet altijd voldoende zijn om zich aan te kunnen passen aan een sterk veranderende omgeving, hetgeen noopt tot genetische aanpassing (micro-evolutie) om het lokaal uitsterven van de populatie te voorkomen. Empirisch bewijs dat populaties zich over de tijd inderdaad genetisch aanpassen (bijv. d.m.v. een verschuiving in hun fenologie) is echter schaars. Kwantitatieve genetica, de studie die zich bezigt met de genetica van kwantitatieve (polygene) eigenschappen, wordt vaak gebruikt in wilde populaties om de evolutionaire parameters (genetische variatie en selectie) te kwantificeren en voorspellingen van evolutie te doen. Echter, voor betrouwbare voorspellingen moeten we de ecologische factoren kennen die ten grondslag liggen aan adaptatie—of het belemmeren ervan. Het doel van mijn proefschrift was om te begrijpen hoe populaties omgaan met een veranderende omgeving en welke ecologische processen de snelheid van adaptatie beïnvloeden. Ik maakte gebruik van een combinatie van zowel veldexperimenten om deze processen te identificeren als langetermijndata om evolutionaire voorspellingen te doen voor de koolmees (*Parus major*) en andere gewervelde soorten.

In het eerste deel van mijn proefschrift (**Hoofdstuk 2**) gaf ik een breed overzicht van wat er bekend is over de effecten van klimaatverandering op de algemene biologie van vogels. Vogels worden op verschillende manieren door klimaatverandering beïnvloed: ze kunnen (1) veranderingen tonen in hun geografische verspreiding omdat hun 'klimaatniche' verschuift, (2) hun fenologische processen zoals voortplanting en migratie vervroegen, (3) morfologische veranderingen (bijv. lichaamsgrootte) ondergaan als een vorm van energetische adaptatie, en (4) demografische veranderingen ondergaan als een direct gevolg van veranderingen in hun reproductief succes en overleving. Veel van wat we weten over de effecten van klimaatverandering slaat op fenologie; er valt dus nog veel te leren over de ecologische en evolutionaire gevolgen van de andere genoemde processen. Of de waargenomen veranderingen in fenologie enkel een plastische of ook een genetische grondslag hebben blijft een open vraag.

Nederlandse koolmezen staan onder toenemende selectiedruk voor een vroegere eilegdatum door een toenemend tijdsverschil (*mismatch*) met de rupsenvoorraad (een belangrijk bestanddeel van het voedsel voor hun kuikens). We zien echter een zeer kleine (fenotypische) respons op deze selectiedruk. Het uitblijven van een respons kan worden veroorzaakt door een energetische beperking die samengaat met het leggen en bebroeden

van eieren in het vroege voorjaar onder barre omstandigheden; vogels die wel (te) vroeg broeden betalen mogelijk een fitnessprijs. In het tweede deel van mijn proefschrift testte ik of vogels inderdaad belemmerd zijn in het vervroegen van de voortplanting. In **Hoofdstuk 3** voerde ik vogels experimenteel bij tijdens de (voor)eilegfase om te testen of vrouwtjes die eerder begonnen met leggen als reactie op het voeraanbod een prijs zouden betalen (d.m.v. vroege nestverlating of gereduceerde voerfrequenties later in het seizoen) wanneer dit bijvoeren gestopt werd op het moment van aanvang van de eileg. Het bijvoeren leidde niet tot een vervroeging van de leg maar leidde tot een verhoogde, en niet een verlaagde, fitness. Omdat bijvoeren *alleen* geen uitsluitsel kan geven of vogels energetisch belemmerd worden om vroeger te leggen (omdat het moeilijk is te achterhalen of bijvoeren leidt tot opheffing van een energietekort of dat het simpelweg fungeert als een aanwijzing om te gaan leggen), is er tevens een experiment nodig dat de legdatum manipuleert zonder daarbij de fysieke gesteldheid van het vrouwtje te beïnvloeden. In **Hoofdstuk 4** beschreef ik de eerste resultaten van een grootschalig experiment waarbij koolmezen op het genoom werden geselecteerd op een vroege of late legdatum. De eieren die werden geproduceerd door broedparen van de selectielijnen in volières werden naar het wild gebracht en de kuikens grootgebracht door pleegouders. De twee selectielijnen (vroeg vs. laat) verschilden op geen enkele manier in hun fitness (uitvliegsucces, gewicht pullen), maar de fitness parameters verschilden licht tussen de selectielijndieren en die van de natuurlijke populatie. Omdat slechts 11 vogels van de selectielijndieren overleefden tot het volgende broedjaar (2018; twee vogels van de vroege en drie van de late lijn) konden we niet statistisch testen of deze vogels inderdaad verschilden in hun legdatum en of het vroeger leggen inderdaad leidde tot verlies in fitness. Ik concludeerde dat meerdere jaren (met verschillende omstandigheden en een grotere steekproef) nodig zijn om te concluderen of vogels inderdaad belemmerd worden om hun eileg te vervroegen.

Broedsucces wordt bij koolmezen uiteindelijk bepaald door de mate van synchronisatie (*match*) tussen de voedselbehoefte van de kuikens en de rupsenbeschikbaarheid. In **Hoofdstuk 5** onderzocht ik de stelling dat als we fenologische (*mis*)*match* als fenomeen willen begrijpen—en willen kunnen bepalen of vogels echt uit de pas lopen met het voedsel—we een grondige, seizoensoverbruggende beschrijving van zowel voedselbehoefte als -beschikbaarheid moeten hebben om zo de mate van temporale overlap tussen deze twee componenten te kunnen bepalen. Ik ontdekte dat de klassieke manier om *mismatch* te beschrijven—d.w.z. het verschil in piekdatums tussen de fenologie van de koolmezen en de rupsen—een betere statistische verklaring geeft voor variatie in zowel de overlevingskansen voor de jongen als voor selectie op eilegdatum dan de temporale overlap tussen voedselbehoefte en -beschikbaarheid. Ik concludeerde dat een eenvoudige maat voor *mismatch* in een seizoensgebonden ecosysteem waarschijnlijk de beste verklaring geeft voor demografische processen, en dat complexere maten in de praktijk waarschijnlijk zelden betrouwbaar zijn.

In het derde deel van mijn proefschrift paste ik vernieuwende kwantitatief genetische modellen toe om patronen in selectie, genetische variatie, en evolutionaire responsen bloot te leggen in wilde populaties waar langetermijnwaarnemingen en stamboomgegevens beschikbaar voor waren. Hierbij maakte ik gebruik van 'reactienormen' (lineaire functies

die fenotypische plasticiteit beschrijven). Sommige van de methodieken die ik hiervoor gebruikte worden in detail uitgelegd in het voorgaande Intermezzo (**Hoofstuk 6** en **7**). In **Hoofdstuk 8** onderzocht ik of maternale effecten (als een vorm van plasticiteit tussen generaties) de snelheid van evolutionaire adaptatie kunnen beïnvloeden. Met behulp van experimentele en langetermijndata kon ik laten zien dat de legselgrootte (het aantal eieren in een nest) van een moederkoolmees deels afhankelijk is van haar eigen lichaamsgewicht als pul (kuiken) en dat deze legselgrootte negatief correleert met dat van haar moeder. Een dergelijk 'negatief maternaal effect' zou adaptatie aan een nieuwe omgeving (met een andere optimale legselgrootte) kunnen belemmeren. Met behulp van simulaties konden we echter laten zien dat dit negatieve maternale effect waarschijnlijk heel weinig invloed heeft op de snelheid van adaptatie.

Fenologische veranderingen over de tijd komen niet altijd overeen met evolutionaire voorspellingen. Eén potentiële reden hiervoor is een onzichtbare onderliggende correlatie tussen omgevingsafhankelijke selectie en erfelijkheid van de bestudeerde eigenschap. In **Hoofdstuk 9** onderzocht ik hoe algemeen een dergelijke correlatie tussen selectie en erfelijkheid is in wilde populaties van gewervelde soorten, en of deze correlatie de snelheid van adaptie beïnvloedt. De verwachting was dat als erfelijkheid en selectie negatief correleren (dus als in een bepaalde omgeving erfelijkheid laag is en selectiedruk hoog), adaptatie belemmerd wordt en vice versa. We maakten op een unieke manier gebruik van *Open Data* (vrij beschikbare data; zie **Hoofdstuk 7**) om omgevingsafhankelijke erfelijkheid en selectie te schatten voor 50 eigenschappen in 16 populaties van 10 soorten. Erfelijkheid en selectie correleerden slechts zelden, en deze correlatie is dus zeer waarschijnlijk niet debet aan de discrepantie tussen evolutionaire voorspellingen en waargenomen fenologische veranderingen in het wild.

Koolmezen reageren in hun voortplantingsgedrag sterk op temperatuur d.m.v. fenotypische plasticiteit. Deze plasticiteit wordt beschreven d.m.v. een reactienorm, de lineaire functie die bestaat uit een intercept (de hoogte van de regressielijn, ofwel de eilegdatum bij een gemiddelde temperatuur) en de helling (de responsgevoeligheid van eilegdatum t.o.v. temperatuur); ieder vrouwtje heeft een eigen reactienorm. In **Hoofdstuk 10** ontdekte ik dat vrouwtjes genetisch verschillen in de intercept maar niet in de helling van de reactienorm, en dat de reactienorm onder verhoogde selectiedruk staat door de vervroeging van de rupsenpiek over de tijd. Ik kon echter kwantificeren dat de evolutionaire respons ondanks deze selectiedruk te traag is geweest om deze waar te kunnen nemen, en voorspelde dat dit zo zal blijven, doordat variatie in eilegdata in grote mate door (niet-genetische) omgevingsfactoren wordt beïnvloed.

In conclusie onderzocht ik het evolutionair potentieel van populaties en had ik als doel de ecologische factoren in kaart te brengen die adaptatie belemmeren. Ik ontdekte dat we nog veel te weten moeten komen over de ecologische en evolutionaire gevolgen van klimaatverandering—waaronder de effecten op demografie en de levensvatbaarheid van populaties—buiten de enkele zeer goed bestudeerde soorten. Experimenten gericht op het blootleggen van de fitnesskosten die gepaard gaan met een te vroege voortplanting zijn niet eenduidig en behoeven verdere studie (met grotere steekproeven en diverse ecologische omstandigheden). Krachtige kwantitatief-genetische methodieken staan klaar

om gebruikt te worden door evolutionair ecologen om evolutionaire processen te kwantificeren, maar deze modellen moeten op de werkelijkheid gestoeld zijn om betrouwbare voorspellingen te waarborgen. Ik heb in mijn proefschrift geopperd dat realistische voorspellingen verwezenlijkt kunnen worden d.m.v. het integreren van verschillende databronnen (langetermijnobservaties en experimenten) en simulaties. Het gebruik van *Open Data* kan helpen dit te bereiken door het beantwoorden van nieuwe onderzoeksvragen op een zo breed mogelijke taxonomische of geografische schaal. Op de eerste plaats komt echter de noodzaak dat we grip hebben op de belangrijkste componenten van het ecosysteem van de soorten die we bestuderen. Alleen dan kunnen we betrouwbare—en begrijpelijke—evolutionaire voorspellingen doen.

Curriculum vitae

Jip Ramakers was born on April 14, 1986 in Baarlo, the Netherlands, and grew up on the countryside where he enjoyed spending time with friends in fields and woodlands and being in nature. Following secondary school he went on to study Forest and Nature Conservation at Van Hall Larenstein University of Applied Sciences in Velp, and after obtaining his bachelor's degree in 2009 he continued his education at Van Hall Larenstein in Leeuwarden, where he studied International Wildlife Management and received a second degree in January 2012. In 2012 he started a voluntary assistantship at the Netherlands Institute of Ecology (NIOO-KNAW) in Wageningen, working with Dr. Kamiel Spoelstra on the effects of artificial light at night on wildlife. Captivated by the science at NIOO and motivated to obtain a PhD position, he enrolled in the Environmental Biology master's programme at Utrecht University in February 2013. During this master's he continued working with Kamiel Spoelstra on the effects of light at night on foraging and commuting behaviour in bats. In a second project, he collaborated with Dr. Teague O'Mara and Dr. Dina Dechmann (Max Planck Institute of Ornithology, Radolfzell, Germany) and Dr. Rachel Page (Smithsonian Tropical Research Institute, Panamá) to study social-information transfer in group-living fruit bats in Gamboa, Panamá. He obtained his Master's degree in February 2015, with the distinction *cum laude*.

In January 2015 he started his PhD on the evolutionary ecology and quantitative genetics of seasonal timing and other reproductive traits with Prof. Dr. Marcel E. Visser and Dr. Phillip Gienapp at NIOO-KNAW, the results of which are presented in this thesis. In March 2019 he started as a postdoctoral researcher statistical genetics with Prof. Dr. Fred A. van Eeuwijk at Biometris, part of Wageningen University & Research in Wageningen.

List of publications

Ramakers, J.J.C., Gienapp, P., & Visser, M.E. (2018). Phenological mismatch drives selection on elevation, but not on slope, of breeding time plasticity in a wild songbird. *Evolution 73*, 175–187.

Ramakers, J.J.C., Culina, A., Visser, M.E., & Gienapp, P. (2018). Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations. *Nature Ecology & Evolution 2*, 1093–1103.

Culina, A., Crowther, T.W., **Ramakers, J.J.C.**, Gienapp, P., & Visser, M.E. (2018). How to do meta-analysis of open datasets: comment. *Nature Ecology & Evolution 2*, 1053–1056.

Ramakers, J.J.C., Cobben, M.M.P., Bijma, P., Reed, T.E., Visser, M.E., & Gienapp, P. (2018). Maternal effects in a wild songbird are environmentally plastic but only marginally alter the rate of adaptation. *The American Naturalist 191*(5), E144–E158.

Spoelstra, K., **Ramakers, J.J.C.**, van Dis, N.E., & Visser, M.E. (2018). No effect of artificial light of different colors on commuting Daubenton's bats (*Myotis daubentonii*) in a choice experiment. *Journal of Experimental Zoology Part A: Ecological and Integrative Physiology* 1–5.

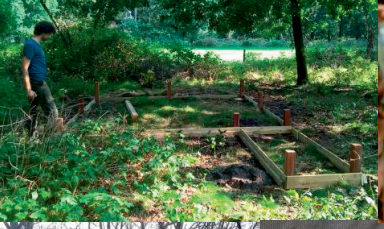
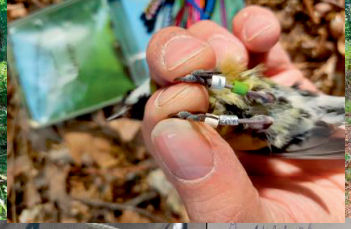
Spoelstra, K., van Grunsven, R.H.A., **Ramakers, J.J.C.**, Ferguson, K.B., Raap, T., Donners, M., Veenendaal, E.M., & Visser, M.E. (2017). Response of bats to light with different spectra: light-shy and agile bat presence is affected by white and green, but not red light. *Proceedings of the Royal Society B-Biological Sciences 284*(1855), 20170075.

Tomotani, B.M., **Ramakers, J.J.C.**, & Gienapp, P. (2016). Climate change impact: Birds. In: eLS. John Wiley & Sons, Ltd: Chichester. doi:10.1002/9780470015902.a0020484.pub2

Ramakers, J.J.C., Dechmann, D.K.N., Page, R.A., & O'Mara, M.T. (2016). Frugivorous bats prefer information from novel social partners. *Animal Behaviour 116*, 83–87.

Ramakers, J.J.C., Dorenbosch, M., & Foppen, R.P.B. (2014). Surviving on the edge: a conservation-oriented habitat analysis and forest edge manipulation for the hazel dormouse in the Netherlands. *European Journal of Wildlife Research 60*, 927-931.

Smith, P., Pheasey, H., Atkinson, K., **Ramakers, J.**, & Sarvary, J. (2012). The Didelphimorphia (Didelphidae) of Reserva Natural Laguna Blanca, Departamento San Pedro, Paraguay. *Acta zoológica lilloana* 56: 141–153.



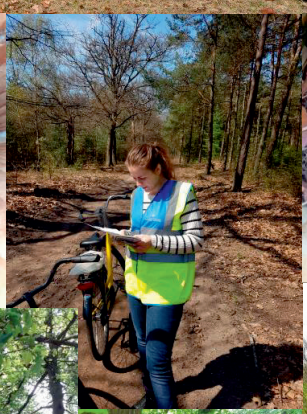


GEDEELTE HOGE VELUWE
2011












136







Acknowledgements

This thesis would be non-existent if it weren't for the numerous people who, each in their own way, contributed to the work presented here.

First and foremost, I would like to express my deepest gratitude to my promotor, **Marcel**, who not only made this project possible financially, but who has also been a true inspirer, motivator, and mentor throughout my journey. You continuously inspired me when embarking on new projects and, while never involved with the very nitty-gritty analytical details, always managed to keep the bigger picture and the biology of the system in sight and prevented me from wandering astray. Most certainly my (future) career as a researcher will be owed largely to the doors that you opened for me.

Phillip, as my co-promotor you have been the true engine of my PhD. Not only have your analytical skills been instrumental in my development as a quantitative geneticist, you were also always there to have a nice chat during the highly needed lunch or coffee breaks. Thank you for all the things that you have taught me, for the interesting discussions, and for keeping patient when being asked ignorant questions over and over again. I wish you the best of luck furthering your career in Bergenhusen.

Before I continue to thank the people that have been important to me scientifically, I need to highlight two people in particular: my paranymphs **Kamiel** and **Irene**. At a point when I had no idea about what to do with my life, **Kamiel** took me under his wings. He invested time in me by teaching me everything he knew about bats, helped me believe in myself and develop myself as a researcher, introduced me to some very nice people (including his family!), helped me get enrolled in the master's programme by putting in a good word and eventually tipped me about the ad for this PhD. Kamiel, this is one of the rare occasions where it is totally justified to state that without you this thesis would not have happened *in the very literal sense*! **Irene**, my 'PhD sibling': let me start by thanking you for your invaluable help during my project at a time when your schedule was occasionally even more hectic than mine! You and I started on the same project. Although our ways parted soon scientifically, I have always felt a natural connection between the two of us. Sharing office, our long conversations about life as a PhD (tough at times), and the nice dinners with you and **Renske** have created long-lasting memories that I will continue to cherish. Thank you both for being my paranymphs!

Scientific ideas never emerge out of the blue. They come from attending conferences, reading interesting papers, and of course having inspiring discussions with many equally inspiring people. In my case, the list of people is endless but I would like to list at least the entire Animal Ecology department; the seniors **Bart, Kees, Kate**, **Martijn**, and **Arie**, as well as the large flock of awesome PhD students and postdocs: **Barbara**, **Lies**, **Lysanne**, **Lucia**, **Maaike**, **Thomas**, **Nina B.**, **Rascha**, **Kees**, **Els**, **Nelleke**, **Magali**, **Henk Jan**, **Krista**, **Bernice**, **Melanie**, **Davide B.**, **Chiel**, **Natalie**, **Götz**, **Antica**, **Davide D.**, **Gretchen**, **Filipe**, **Liam**, **Nina M.**, **Callum**, **Jenny**, **Lyanne**, **Cas**, **Judith**, **Monique**, **Veronika**, and **Marleen**. Some of you may have been more heavily involved than others in scientific discussions, but in

all cases you have been wonderful colleagues whom I enjoyed spending time with. People like you make NIOO a place worth returning to for a cup of coffee—or perhaps future collaborations? Furthermore, I want to thank **Tom** and **Piter** for their invaluable expert input that resulted in at least one paper.

A big word of thanks to the supporting staff and students, without whom the scientific engine would come to a halt. The animal caretakers (**Ruben**, **Marylou**, **Anouk**, **Franca**, and **Coretta**), the gentlemen from the technical department (**Wim**, **Gilles**, **Jeroen**, and of course **Eke**), **Christa**, **Agata**, **Cynthia**, **Manon**, **Martijn**, **Bart**, and **Piet**: thank you for all your help and support and for being nice colleagues during all these years. **Peter** and **Louis**, '*Buurman & Buurman'*, you deserve a separate mention here because it is you two who I have bothered the most in these past four years; building the cages at the Hoge Veluwe, though virtually fruitless in the end, was particularly fun and memorable. **Gerda** and the ladies at the reception, you have made my stay at NIOO very pleasant, as have my chats at the coffee machine with the people from the AqE department. **Henri**, thanks for showing me (a glimpse of) what it is like to be a field ornithologist! Working with you in the field was as useful and informative (because you kept my students in check) as it was pleasant. My wonderful students prevented my projects from becoming a logistical nightmare. **Emma**, **Imre**, **Mees**, **Mireille**, **Sylvana**, **Martijn** and **Maartje**: I wish you all the best starting your careers as ecologists/researchers.

Graag wijd ik een woord van dank aan mijn familie en schoonfamilie die altijd voor me klaar staan en mij zowel moreel als logistiek (lees: financieel) gesteund hebben bij het bereiken van mijn doelen. Bedankt, oom **Rick,** voor de mooie foto's die in mijn boekje prijken en dank, zusje **Maja**, voor je mooie illustraties die de omslag van mijn proefschrift sieren.

Ik ben veel verschuldigd aan mensen die me direct of indirect, al dan niet inhoudelijk gesteund hebben—en ongetwijfeld ben ik er veel vergeten. Maar de grootste steun die ik in de afgelopen jaren heb mogen ontvangen komt van mijn persoonlijke advocate, mijn maatje door dik en dun, mijn verloofde, maar bovenal mijn grote liefde **Hanna**. In veel opzichten ben jij de sleutelfactor achter zowel het behalen van mijn PhD als mijn groei als man. Zo heb je mijn vaak drukke werkschema moeten accepteren en heb je mij altijd—al ver voor de aanvang van mijn PhD—gestimuleerd om het beste uit mezelf te halen. Als de nood hoog was tijdens het veldseizoen, was je ook niet te beroerd om mij te helpen, in het veld of elders. Als ik in het buitenland was wachtte jij geduldig op me. Als ik het even niet zag zitten was jij er om me op te beuren. Ik noem dat liefde. Ik ben vereerd jou vanaf 5 juli 2019 mijn vrouw te mogen noemen.

De tied vluuëg, maar wat waor ut sjoeën! Haije!

WIAS Training and Education Statement

With the training and education activities listed below, the PhD candidate has complied with the requirements set by the graduate school for the Wageningen Institute of Animal Sciences, which comprises a minimum total of 30 ECTS (= 20 weeks of activities).

Basic training activities (1.8 ECTS)

WIAS Introduction Day (2015)
Course *Philosophy of Science and/or Ethics (2015)*

Disciplinary competences (14.2 ECTS)

Research proposal PhD (2015)
Course *Getting started with ASREML*, ABG, WUR (2015)
Course *Animal Experimentation (Art. 9),* KNAW (2015)
Course *Life History Theory*, RUG (2015)
Course *Genotype by environment interaction, uniformity and stability*, WIAS, WUR (2015)
Course *Survival Analysis*, PE&RC, WUR (2016)
Course *Bayesian Statistics*, PE&RC, WUR (2017)

Professional competences (4.7 ECTS)

Course *Essential skills*; WIAS, WUR (2016)
Course *Career orientation*; EPS, WUR (2017)
Journal clubs (literature review) at NIOO-KNAW (2015–2018)

Presentation skills (4 ECTS)

WIAS Science Day (poster), Wageningen, NL (2016)
Netherlands Annual Ecology Meeting (oral), NERN, Lunteren, NL (2016)
European Society of Evolution (ESEB) congress (poster), Groningen, NL (2017)
Joint Evolution congress of the Amer. Evol. Soc. and ESEB (oral), Portland, Oregon (2017)
Netherlands Society of Evolutionary Biology (NLSEB) congress (poster), Ede, NL (2018)
27th International Ornithological Congress (oral), Vancouver (2018)

Teaching competences (6 ECTS)

Supervising students' thesis (two bachelor's (BAS) and two master's (MSc)) (2017-2018)
Lecturing master course *Environmental Processes,* Leiden University (2016–2018)
Examination of master students course *Ecology of Life Histories*, WUR (2018)

Colophon

The research presented in this thesis was conducted at the Department of Animal Ecology of the Netherlands Institute of Ecology (NIOO-KNAW), Wageningen.

The research described in this thesis was financially supported by the European Research Council.

This thesis is NIOO Thesis 169

Photos of free-ranging great and blue tits by courtesy of Rick Sanders

Cover art by Maja Ramakers

Printed by ProefschriftMaken.nl, Wageningen, the Netherlands