ORIGINAL ARTICLE

doi:10.1111/evo.13325

On the standardization of fitness and traits in comparative studies of phenotypic selection

Stephen P. De Lisle[1,2] **and Erik I. Svensson**[1]

[1] *Evolutionary Ecology Unit, Department of Biology, Lund University, Sölvegatan 37, 223 62 Lund, Sweden*

[2] *E-mail: stephen.de_lisle@biol.lu.se*

Received May 23, 2017
Accepted August 3, 2017

Comparisons of the strength and form of phenotypic selection among groups provide a powerful approach for testing adaptive hypotheses. A central and largely unaddressed issue is how fitness and phenotypes are standardized in such studies; standardization across or within groups can qualitatively change conclusions whenever mean fitness differs between groups. We briefly reviewed recent relevant literature, and found that selection studies vary widely in their scale of standardization, but few investigators motivated their rationale for chosen standardization approaches. Here, we propose that the scale at which fitness should be relativized should reflect whether selection is likely to be hard or soft; that is, the scale at which populations (or hypothetical populations in the case of a contrived experiment) are regulated. We argue that many comparative studies of selection are implicitly or explicitly focused on soft selection (i.e., frequency and density-dependent selection). In such studies, relative fitness should preferably be calculated using within-group means, although this approach is taken only occasionally. Related difficulties arise for the standardization of phenotypes. The appropriate scale at which standardization should take place depends on whether groups are considered to be fixed or random. We emphasize that the scale of standardization is a critical decision in empirical studies of selection that should always warrant explicit justification.

KEY WORDS: Experimental design, hard selection, phenotypic selection, relative fitness, soft selection.

Evolution is governed by distinct processes of natural selection of phenotypes and the heritable transmission of those phenotypes across generations (Fisher 1930). This separation between natural selection and evolutionary change was first characterized by Darwin, but fully formalized in the 20th century (Darwin 1859, Price 1970). Notably, Lande and Arnold (1983) exploited then-recent theoretical developments (Lush 1945, Price 1970, 1972) to develop statistical tools, based on multiple regressions, that enabled empiricists to quantify the strength and shape of natural selection in extant populations, using data from fitness or fitness components. In the Lande-Arnold special case of Price's equation, we can describe between-generation evolution by characterizing the effects of selection on within-generation phenotypic change:

$$\Delta\, \bar{z}_s = \frac{\mathrm{cov}(W,\ z)}{\bar{W}} = \mathrm{cov}(w,\ z) = \beta \mathrm{P}, \qquad (1)$$

where W is individual absolute fitness, w is individual relative fitness ($W/\bar{W}$), β is the linear regression of relative fitness on phenotype z, P is the phenotypic variance of z, and the subscript s on the left-hand side signifies change due to selection alone. Thus, evolutionary change depends on the relationship between phenotype and absolute fitness *relative* to the population mean absolute fitness. The trait(s) z are typically variance-standardized (Lande and Arnold 1983), although some evolutionary biologists argue for mean-standardization (Hereford et al. 2004). Standardization is done to make estimates of selection comparable across different traits and studies.

The statistical methodology introduced by Lande and Arnold (1983) was an important methodological advance that enabled quantification of how selection operates in wild populations, but on its own provides little insight into the ecological causes of natural selection (Grafen 1987, 1988, Wade and Kalisz 1990). Understanding why and how some phenotypes offer a fitness advantage in a population(s) requires manipulative or natural experiments, via perturbation of the traits under selection (Mitchell-Olds and Shaw 1987), ecological agents of selection (Wade and

Kalisz 1990), or “double-level” experimental manipulations of both the phenotypic traits and the putative ecological selective agents (Svensson & Sinervo 2000). Determining whether phenotypic divergence among populations is a likely outcome of divergent natural or sexual selection requires comparing patterns of variation in selection with the direction of phenotypic divergence (Zeng 1988, Chenoweth et al. 2010). Thus, arguably some of the most interesting and general applications of phenotypic selection analysis entail statistical comparison of the strength or form of selection among different groups, such as among demes, populations, species, or levels of an experiment. Although advocated early after the introduction of the Lande-Arnold approach, such comparative selection studies remained exceedingly rare for decades. However, recently there has been rapid increase in studies employing among-group comparison of phenotypic selection, whether in controlled experiments, correlative studies in the wild, or in large-scale meta-analyses of published selection studies (Kingsolver et al. 2001, Chenoweth et al. 2012, Siepielski et al. 2017).

A large literature exists on the statistical difficulties in estimating selection within a population dealing with various methodological considerations and questions such as: What is the best way to fit a regression (Morrissey 2014)? What if the data are not normally distributed (Mitchell-Olds and Shaw 1987)? How should one accommodate environmental covariances between traits and fitness (Rausher 1992)? What about imperfect detection of marked individuals and the associated effects on the selection estimates (Waller and Svensson 2016)? However, there is limited discussion of the conceptual and statistical issues and biological considerations associated with among-group comparisons of selection (Svensson and Sinervo 2004, Chenoweth et al. 2012, ter Horst et al. 2015, Bolstad 2017, ter Horst et al. 2017).

Here, we address the question of how and why fitness and phenotypes should be standardized in comparative studies of selection of multiple populations or subpopulations of a single species. We begin by reviewing the proximate sources of variation in selection between groups, and how the scale of fitness relativization can influence conclusions under different scenarios. We perform a brief review of the relevant literature and show that authors vary in their approach to relativization and standardization and often provide little justification for these decisions. We then show that the concepts of soft and hard selection, well known to population geneticists but perhaps less known to empiricists interested in phenotypic selection, can be used to rationalize how relative fitness should ideally be calculated in comparative studies. A coherent approach to phenotypic standardization follows from the nature of “treatment” assignment to groups that are compared. Misidentifying the appropriate scale of relativization can lead to qualitatively wrong conclusions about treatment effects. Our main conclusion is that the appropriate scale of fitness relativization is an underappreciated biological, and not only a statistical, issue that depends on the processes that may generate variation in selection and the ecology of the study organism(s) in question. We argue that rigorous empirical studies should provide explicit justification for specific choices of standardization, and these standardizations should ideally be grounded in both evolutionary theory and natural history of the study organism.

Background: Sources of Variation in Phenotypic Selection

The strength or form of natural selection can differ between groups for three nonexclusive reasons, which are all illustrated in Figure 1. First, the two groups could experience absolute fitness functions that differ in shape, as illustrated in Figure 1A. Second, differences in mean absolute fitness between groups can lead to differences in the relationship between relative fitness and phenotype whenever group mean fitness is used to calculate relative fitness. These differences are illustrated in Figure 1B, C, and E. That is, even if the shape of the relationship between absolute fitness and phenotype is constant across groups, a difference in the intercept creates a scenario where the scale of fitness relativization plays a role in determining group differences in selection. Third, groups could experience the same absolute fitness function, but the groups could differ in the location of their phenotypic distribution (Fig. 1D). This is particularly important when the two underlying fitness function are nonlinear (e.g., selection toward an optimum), but will nonetheless be important generally because even if selection is linear groups will then differ in mean fitness (see next). Distinguishing among causes of variation in selection will be difficult or impossible when groups differ in their phenotypic distribution; we address this further below in the Section “Fixed Groups, Random Groups, and the Standardization of Phenotypes.”

We would of course like to underscore that all three sources of variation could contribute to group differences in selection simultaneously. Thus, different groups may differ in mean fitness, may have different phenotypic distributions, and could be evolving on fitness surfaces that differ fundamentally in shape. A key implication of group differences in mean absolute fitness, however, is that the scale of relativization plays a critical role in determining group differences in the strength of selection. This effect is illustrated in Figure 1, where quantitative differences in selection strength are illustrated in the left and the right panels for different forms of relativization. However, more extreme scenarios are also possible, such as when fitness differences lead to reversed treatment effects under global versus local fitness relativization (Fig. 2). Here, two groups (red and blue) differ both in their mean absolute fitness and in the magnitude of the covariance between absolute fitness and phenotype (Fig. 2A). Whether the strength

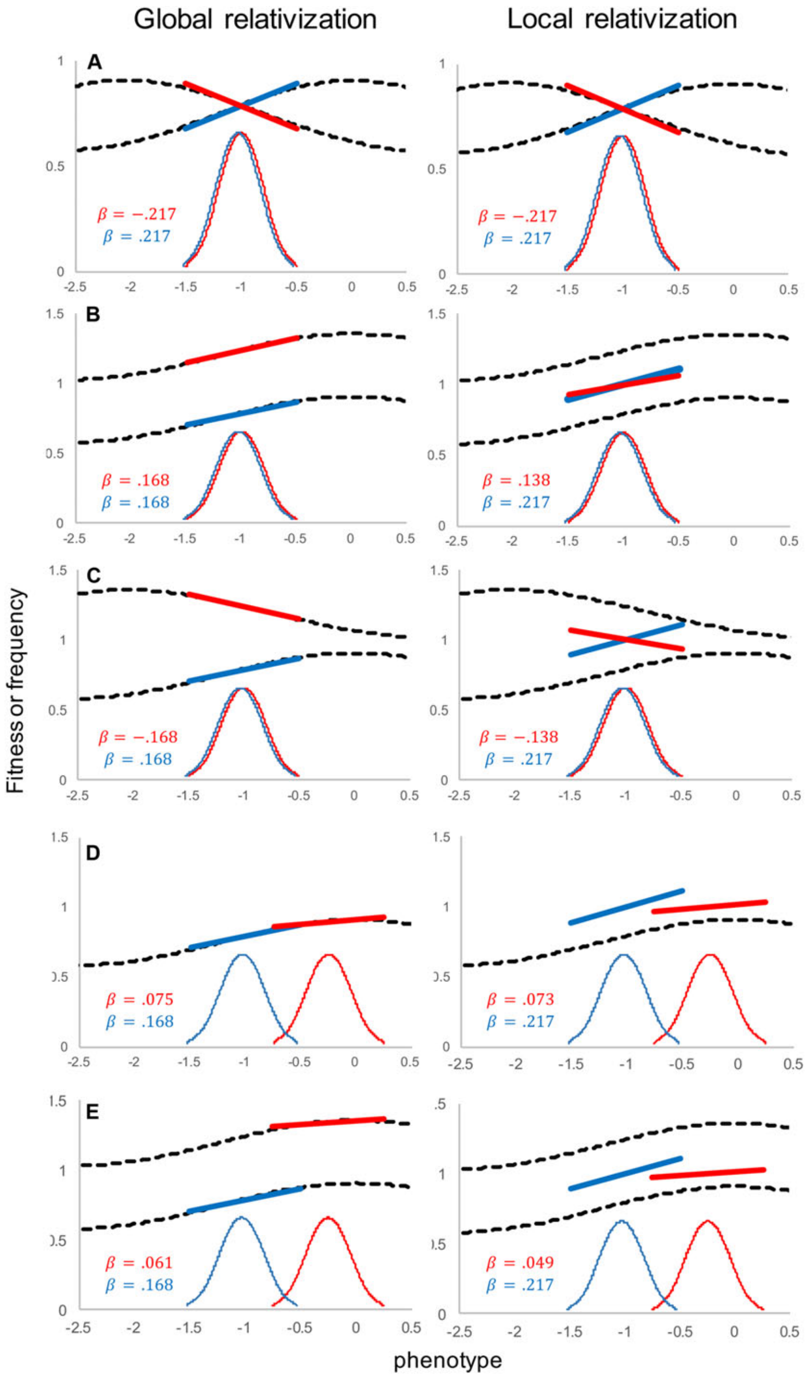

Figure 1. Global and local fitness relativization under different scenarios of variation in selection. Phenotypic distributions of two hypothetical groups are illustrated in red and blue, absolute Gaussian fitness functions are black dashed lines and indicate directional selection toward an optimum (width of the fitness function, ω = 1.1, optimum, θ = 0 or −2 [in B and C], fitness at optimum = 0.91 or 1.36). In panels (A–C), both groups have the same phenotypic distribution, but are subject fitness functions that either differ in location of the optimum (A), mean absolute fitness (B), or both (C). Panels (D) and (E) indicate a scenario where populations differ in their phenotypic distribution (centered at −1 or −0.25) but are evolving on a common fitness surface (D) or differing fitness surfaces (E). In all cases where mean absolute fitness differs between groups (every panel except A), the scale of relativization changes the inferred strength of directional selection and the magnitude of among-group difference in selection. Selection gradients were calculated as the covariance between phenotypic values within 0.5 units of the group mean phenotype and relative fitness (absolute fitness divided by within-group [local] or pooled [global] mean fitness) divided by the phenotypic variance, which was held constant. For the global standardization, regression lines were plotted with an intercept through the group absolute mean fitness for clarity. Note y-axis scale differs for Panel (A).

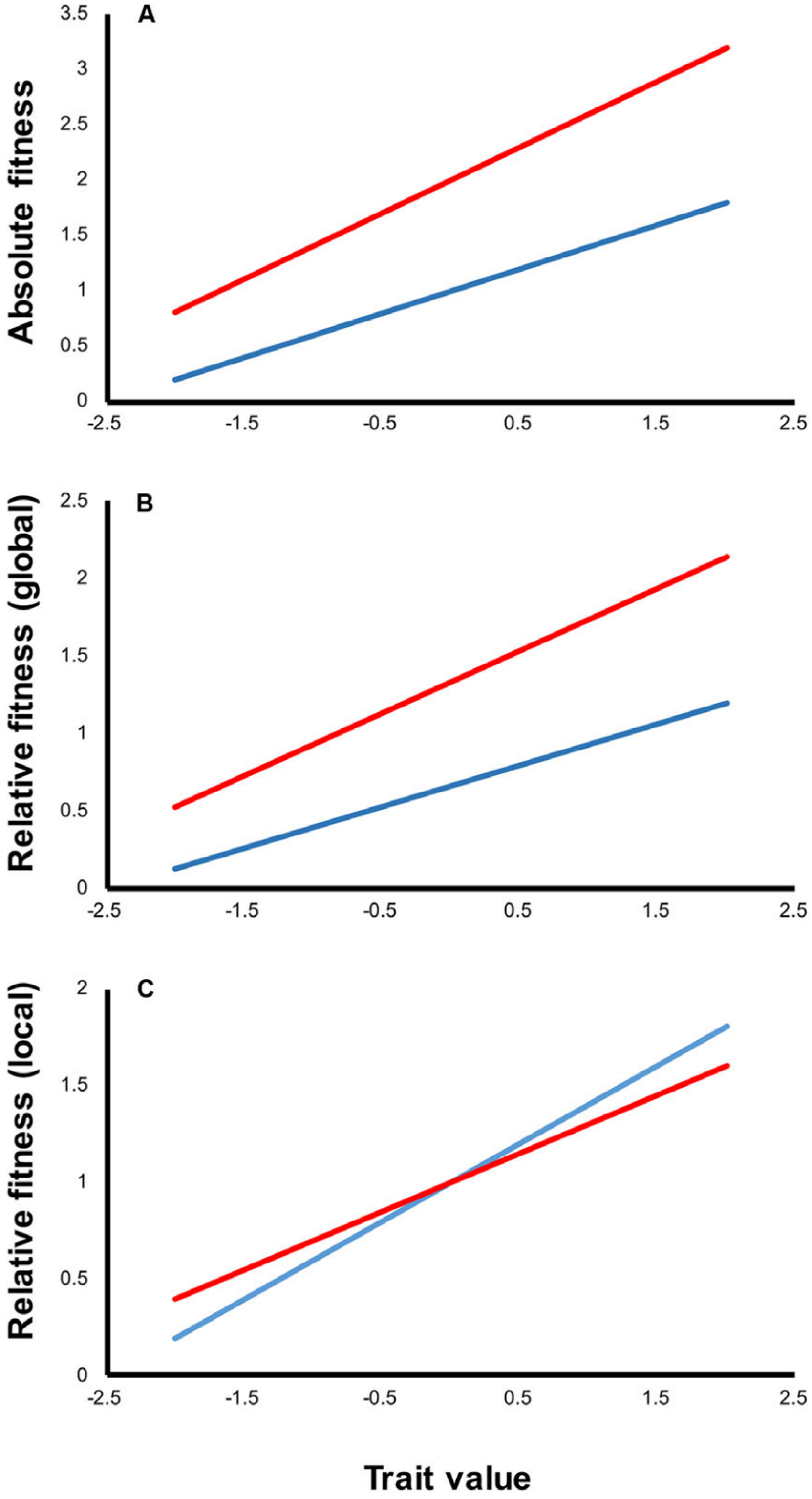

Figure 2. Scale of relativization can reverse conclusions under realistic conditions. Two groups differ in both mean fitness (under the implied assumption of similar phenotypic distributions in the two groups) and the covariance between phenotype and absolute fitness, illustrated in (A). These differences hold in regressions of relative fitness on phenotype in the case where relative fitness is calculated using the global (experiment-wide) mean fitness (B). However, when fitness is relativized locally, using group-specific mean fitness, the group effect on the strength of selection is reversed (C). This occurs because the higher covariance between phenotype and absolute fitness in the red group is more than offset by the group's higher mean fitness. Values used in this figure: β (absolute fitness) 0.4 (blue), 0.6 (red); mean absolute fitness 1 (blue) 2 (red).

of phenotypic selection—the covariance between phenotype and relative fitness—is stronger in a given group depends on the scale at which fitness is relativized. If the global mean is used, the group differences reflect the covariance between phenotype and absolute fitness (Fig. 2B). However, if the group-specific means are used to relativize fitness, the group effect might become reversed (Fig. 2C). In this case, the choice of relativization scale qualitatively changes conclusions on treatment effects. Although the absolute fitness–phenotype covariance is higher in the red group, this is offset by the higher mean fitness when relative fitness is calculated within groups (Fig. 2C).

Although the above discussion applies to any fitness component, it is important to point out some special properties of relative fitness calculated from binary fitness components, such as mating success or survival. Whenever absolute fitness of one category (e.g., "unmated") is treated as zero, relative fitness depends on the frequency of successful versus unsuccessful individuals in the population (Brodie and Janzen 1996), which is closely connected to the mating skew (Kokko and Lindstrom 1997). Thus, the situation of a difference in mean absolute fitness, in the case of a binary fitness component, can be equivalently expressed as a difference in the frequency of successful versus unsuccessful individuals between groups. Although this seems trivially true, it is important because the frequency of successful versus unsuccessful individuals is key to determining the strength of selection when fitness is binary, and thus information on this frequency is an essential piece of information for estimating selection (see also Price et al. 2000); that is, the frequency of successful individuals will influence mean fitness in the population, that is, the denominator when relativizing individual fitness and is thus inversely related to the opportunity for selection (Arnold and Wade 1984a,b).

Literature Review

We briefly reviewed the recent literature for recent empirical studies that compared the strength or form of selection among groups within species or between pairs of closely related species (e.g., congeners). We began our search by examining every paper citing the methodological book chapter by Chenoweth et al. (2012), who outlined a sequential model building approach and other contemporary methods for the statistical comparison of selection among groups. We then expanded this search to include some other relevant and well-cited papers published between 2000 and 2012 as well as other relevant selection papers published after 2012 that do not cite Chenoweth et al. Our search was haphazard, but appropriate for our purpose: to obtain a representative sampling of the literature of comparative studies of phenotypic selection published in respected journals in the field, rather than a complete systematic list of all available papers. We examined each paper to determine whether fitness was relativized across the whole experiment or at a lower level and whether phenotypes were standardized across the whole experiment or at a lower level. Many investigators simply reported that standardized selection gradients were estimated separately for each group; although this does not specify the scale of standardization, we took this to imply local standardization and studies reported in this way are marked with an asterisks in Table 1. We also noted whether explicit justification or elaboration of these standardization choices were provided. By explicit justification or elaboration we mean some statement as to why the particular level of fitness relativization or trait standardization was chosen. We were generous in our assessment of justification; any elaboration for the rationale of standardization scale was considered explicit justification. We also emphasize that our aim is not to call into question the results or reporting of any particular study; rather we seek to establish the realized standard in the field for both the scale of standardization and the degree of justification provided for standardization decisions. The relevant quotations from the methods section of each study, used to ascertain standardization decisions reported in Table 1, are reproduced in the Supporting Information.

A summary of this literature review is provided in Table 1. Noteworthy is that there is much variation in the scale of fitness and phenotype standardization, even across studies originating from the same research group and using similar methodology. More importantly, few studies provide explicit justification for standardization choice, even by our simple standards (see above): only four of 20 studies justify or elaborate on the scale of fitness relativization, whereas five of 20 justify the scale of trait standardization. In many cases it was not made completely clear what level was chosen for standardization. Noteworthy also is that many of the laboratory studies of mating success listed in Table 1 fix mating rates across treatment groups, for example, by selecting the same number of mated and unmated males for phenotypic analysis across groups. Thus, although in this case relativization decisions are inconsequential, because absolute fitness in this case has been standardized across groups, such an experimental design caries an implicit assumption of global fitness relativization, because any treatment effects on male mating rate are eliminated in such a design. Although we believe the authors of many of the studies in Table 1 likely had some rationale for choosing a particular standardization scale, it is not clear from most papers what that particular rationale was. We find this general pattern worrying, in light of the interpretive difficulties suggested by Figures 1 and 2.

Hard Selection, Soft Selection, and the Relativization of Fitness

Under realistic biological scenarios as illustrated in Figures 1 and 2, how should we then relativize fitness? Intuition suggests the answer lies in the scale of comparison. It seems trivially true

Table 1. Summary of recent comparative studies of phenotypic selection.

Study	Scale of fitness relativization	Explicit justification/ elaboration provided	Scale of phenotypic standardization	Explicit justification/ elaboration provided	Fitness component	Selection	Design
Keagy et al. (2016)	Global	Yes	Global[*]	No	Territory establishment	Sexual	Fixed
Curtis et al. (2013)[1]	Global[*]	No	Global	yes	Mating success	Sexual	Random
De Lisle and Rowe 2015	Local	Yes	Global	Yes	Growth rate	Competition	Fixed
Gershman et al. 2014	Global[*]	No	Global[*]	No	Mating success	Sexual	Fixed
White and Rundle 2015	Local[*]	No	Local[*]	No	Mating success/ territory Establishment	Sexual	Fixed
Rundle and Dyer 2015	Local	No	Local	No	Mating success	Sexual	Fixed
Gosden et al. (2014)[2]	Absolute fitness	No	Unspecified	No	Mating success	Sexual	Fixed
Punzalan et al. (2010)[3]	Local	No	Local	No	Mating success	Sexual	Random
Svensson and Sinervo (2004)	Both	Yes	Local and global	Yes	Survival	Competition/ predation	Fixed
Gosden and Svensson (2008)	Local[*]	No	Global	No	Mating success	Sexual	Random
Chenoweth et al. (2010)	Absolute fitness	No	Unstandardized	No	Mating success	Sexual	Random
Rundle et al. (2008)[4]	Absolute fitness	No	Unstandardized	No	Mating success	Sexual	Random
Rundle and Chenoweth (2011)	Local	No	Local	No	Mating success	Sexual	Random

(*Continued*)

Table 1. Continued.

Study	Scale of fitness relativization	Explicit justification/ elaboration provided	Scale of phenotypic standardization	Explicit justification/ elaboration provided	Fitness component	Selection	Design
Steele et al. (2011)	Local	No	Local	No	Mating success	Sexual	Random
Rundle et al. (2009)[5]	Absolute fitness	No	Global	Yes	Mating success	Sexual	Fixed
Calsbeek and Cox (2010)[6]	Unspecified	No	Local	No	Survival	Competition/ predation	Fixed
Calsbeek (2009)[7]	Unspecified	No	Unspecified	No	Survival	Competition/ predation	Both
Losos et al. (2004)	Local*	No	Local*	No	Survival	Predation	Random
Start and Gilbert (2016)	Both	Yes	Global	No	Survival	Predation	Random
Austen and Weis (2015)	Local	No	Both	Yes[8]	Total seed mass	Total seed mass	Fixed

*Implied but not explicit (e.g., a statement that standardized selection gradients were estimated separately by group).

[1] Separate congeners considered, but relevant because selection gradients were compared in a common statistical model.

[2] Compared male and female selection, unclear if or how fitness was relativized for reported gradients.

[3] From the cross-sectional analysis.

[4] CHCs were standardized within individuals to be placed on a relative scale, but not standardized otherwise.

[5] States fitness was relative, but gives binomial absolute fitness in the regression equation.

[6] We assume population = island.

[7] Unclear whether population = island, plot, or the entire study.

[8] Justification provided for global relativization.

that in comparative meta-analyses of selection, estimates from geographically widely separated populations and phylogenetically distant taxa (e.g., Kingsolver et al. 2001; Siepielski et al. 2017), the only appropriate approach would be to relativize fitness within "groups" (e. g., different species). The issue becomes progressively more ambiguous, however, when comparing selection among populations that are connected by varying levels of gene flow or in a manipulative experiment where comparisons are made among treatment groups with different ecological settings. This ambiguity begs for theoretical justification of a general methodological and rigorous approach to standardization. Here, we connect this problem to two classical population genetic models of selection, hard (density- and frequency-independent) and soft (density- and frequency-dependent) selection. These classical population genetic models help to understand evolution in a metapopulation and can provide a theoretical rationale for relativization decisions in field studies of natural and sexual selection. Our main goal here is to strengthen the links between evolutionary quantitative genetics and population genetics, which historically rely on a common theoretical foundation, as outlined by R. A. Fisher and Sewall Wright (Lynch and Walsh 1998). We feel that some of the insights from early population genetic theory about how to maintain genetic polymorphisms in spatially heterogeneous environments (e.g., Levene 1953) were forgotten by researchers in the later-developed quantitative genetic tradition. The reason for this was that the problem of maintaining genetic variation became less of a focus (for right or wrong reasons) with the development of evolutionary quantitative genetics theory (Lynch and Walsh 1998), especially for empiricists, compared to single-locus population genetics, as mutation-selection balance was thought to solve that problem for quantitative traits under some simplifying assumptions (Lande 1976; but see Turelli 1984, Barton and Turelli 1989). Although connections between early population genetic models and quantitative genetic analysis of phenotypic selection have been made in a few cases, most notably in the form of contextual analysis, the utility of these classical models appears limited to specific experimental designs where frequency is manipulated directly (Goodnight et al. 1992; Weis et al. 2015). Further, these models do not directly treat density-dependence, a key component of soft selection.

Wallace (1968, 1975) introduced the concepts of hard and soft selection in an attempt to reconcile the dilemma of the mutation load that would be expected in natural populations given observed levels of variation (Haldane 1937, Reznick 2016). Under the most simplistic assumptions of classical population genetics, selection is hard and the fitnesses of genotypes are density- and frequency-independent. Individual fitness depends only on underlying genotypes, and not on the abundance or local genetic composition of the local deme and environment. Under such hard selection, each local deme in the metapopulation contributes propagules to the next generation in proportion to the mean fitness of that environment. Thus, hard selection acts on individual fitness relative to the mean absolute fitness across demes (groups). Conversely, under soft selection, individual fitness depends on both its genotype and the density and genetic composition of the local deme. Selection is therefore both density- and frequency-dependent, and the contribution of each environment in a metapopulation is independent of average phenotype of individuals in that environment. Under soft selection, selection acts on individual fitness relative to fitness of other individuals in the local deme.

Wallace provided several examples of types of mutations and traits that would be under hard versus soft selection (Wallace 1975). Generally, competition over limited resources is expected to generate conditions of soft selection, because selection then becomes both frequency- and density-dependent, due to a limited local carrying capacity set by resource availability. The contribution of a deme to the next generation will primarily be a function of the amount of resources available. Sexual selection could also often be seen as soft, where male mating success is determined in part by the abundance and phenotype frequency distribution of other males that it interacts with (Wallace 1975), yet the contribution of the deme to a metapopulation will depend on female abundance and not male mating success. Traits that affect cold tolerance, for example, would be expected under hard selection; whether an individual survives a cold spell will largely or entirely depend on its genotype and is likely to be relatively independent of the genotypic distribution surrounding it. Thus, whether selection is hard or soft depends on the biological and ecological mechanisms that generate selection. In some cases, particularly manipulative experiments, authors engaging in comparative studies of phenotypic selection may have a priori hypothesis as to the ecological causes of selection. We suggest that a priori hypotheses of the potential causes of selection should be used as informative mechanisms and should guide the decision of how to relativize fitness; mechanisms likely to generate hard selection warrant global relativization, while under soft selection local relativization is appropriate.

Following previous work (Gomulkiewicz and Kirkpatrick 1992, Kelley et al. 2005), we can define treatment-and experiment-wide selection gradients under soft and hard selection. The average strength of selection in a replicated experiment characterized by soft selection, where selection favors phenotypes that maximize local mean fitness, can be expressed as

$$\bar{\beta}_s = \sum_j f_j \sum_i f_i \operatorname{cov}\left(\frac{W}{\bar{W}_{ji}},\ z\right) \mathrm{P}^{-1}, \qquad (2)$$

where $\bar{W}_{ji}$ is the mean absolute fitness in the ith replicate of the jth treatment group, and f is frequency $\sum_i f_i = 1$. In the case of hard selection, fitness is taken relative to the experiment-wide grand mean, or equivalently, selection estimates are weighted by the ratio of group-specific mean fitness over grand mean fitness:

$$\begin{aligned}\bar{\beta}_h &= \sum_j f_j \sum_i f_i \left(\frac{\bar{W}_{ji}}{\bar{W}}\right) \operatorname{cov}\left(\frac{W}{\bar{W}_{ji}},\, z\right) \mathrm{P}^{-1} \\ &= \frac{1}{\bar{W}} \sum_j f_j \sum_i f_i \operatorname{cov}(W,\, z)\, \mathrm{P}^{-1},\end{aligned} \tag{3}$$

where $\bar{W}$ is the grand mean absolute fitness across all i and j. Statistical comparisons of treatment groups could be accomplished in a mixed model with among-replicate variance in slopes (Chenoweth et al. 2012) or in a linear model with experimental unit-specific estimates of β treated as the response.

Of course, hard and soft selection represents two extremes of what is likely to be a continuum of scales at which selection may act. A more general version of Equations (2) and (3) should ideally accommodate such a continuum. Following the population genetic models of Whitlock (2002) and Agrawal (2010), we can express the continuum between hard and soft selection by defining individual relative fitness w as:

$$w = (1-b)\,\frac{W}{\bar{W}_j} + b\frac{W}{\bar{W}}, \tag{4}$$

where b is a parameter between 0 and 1 that determines the softness of selection. This expression differs from that considered by Whitlock and Agrawal in that they treat individual genetic quality, rather than (an estimate of) absolute fitness. When $b = 0$, selection is completely soft and acts on individual fitness relative only to within-group mean fitness $\bar{W}_j$. In terms of the statistical decision of how to relativize fitness in empirical field studies (Table 1), this would be equivalent to relativizing fitness locally, rather than globally. When $b = 1$, selection is completely hard (fitness is relativized globally), and values in between represent the varying degrees of hardness of selection. This formulation has the appealing property that for the special case when within-group and grand mean fitness would be equal, $\bar{W}_j = \bar{W}$, b and the distinction between hard and soft selection is inconsequential.

Using Equation (4), we can obtain a general expression for the mean strength of selection in a replicated experiment:

$$\bar{\beta} = \sum_j f_j \quad \sum_i f_i \operatorname{cov}\left(\left((1-b)\,\frac{W}{\bar{W}_{ij}} + b\frac{W}{\bar{W}}\right),\; z\right)\mathrm{P}^{-1} \tag{5}$$

that accommodates an arbitrary hardness of selection. Equation (5) clarifies the relationship between hard and soft selection. Although in many cases workers will have an a priori interest in the extremes of completely hard or soft selection (e.g., Kelley et al. 2005), in some cases insight may be gained by exploring if or how treatment effects vary with b.

We provide a brief empirical example to illustrate the potential utility of the description of relative fitness provided in Equation (4). We reanalyzed some previously published data examining phenotypic selection on head shape in a salamander (De Lisle and Rowe 2015). In this study, De Lisle and Rowe (2015) manipulated density of adult newts (*Notophthalmus viridescens*) in replicated artificial ponds. Their goal was to determine if density increases the strength of sexually antagonistic selection on head shape, as expected under the hypothesis that resource competition contributes to the evolution of sexual dimorphism in this multivariate trait. To do this, they estimated both the strength of disruptive selection and its geometric alignment with sexual dimorphism, and tested for a density effect on these values. This density-dependence can be summarized as $\mathbf{s}^{\mathbf{T}}\boldsymbol{\gamma}_{\text{high}}\mathbf{s} - \mathbf{s}^{\mathbf{T}}\boldsymbol{\gamma}_{\text{low}}\mathbf{s}$, where $\mathbf{s}$ is the discriminant function vector defining the sexes and $\boldsymbol{\gamma}$ is the matrix of nonlinear selection gradients from a mixed model with relative fitness as the response. $\boldsymbol{\gamma}$ is composed of quadratic selection gradients (in this case, for four traits: snout-vent length, gape, jaw length, and head depth) on the diagonal and correlational selection gradients on the off-diagonal; separate $\boldsymbol{\gamma}$ were estimated for each density treatment (high and low) following formal rejection of the hypothesis that $\boldsymbol{\gamma}$ was equal across treatments. Positive values of the effect size above indicate that the fitness-advantage for morphologically extreme males and females is greater at high density compared to low density. In their analysis, De Lisle and Rowe treated selection as completely soft and used pond means to relativize fitness. This was justified because the type of selection of interest, that arising from ecological character displacement, is expected to arise locally within each experimental pond due to local competitive interactions with other individuals in the particular replicate (i.e., local tank). Thus, an individual competes with other individuals in its pond, and individual fitness relative to these competitors, rather than fitness relative to the larger population as a whole across all ponds, influences the local selection gradient. In natural ponds selection via adult growth rate would also be expected to be soft, because the contribution of each pond to a metapopulation of newts is likely limited by resource abundance more so than adult mean phenotype.

We reanalyzed De Lisle and Rowe's data, this time treating relative fitness according to Equation (4) and exploring how density-dependence varied across the continuum of models dichotomized by hard and soft selection (dryad data from original study: https://doi.org/10.5061/dryad.md23g). We found that the strength of density-dependent sexually antagonistic selection was highest under a model of completely soft selection, where fitness is relativized within ponds, and gradually decreases under increasingly hard models of selection (Fig. 3). However, the positive effect size under completely hard selection indicates that, in this study, qualitative conclusions on the nature of density-dependent selection remain unchanged regardless of how fitness is relativized. The decreasing effect size with increasing hardness

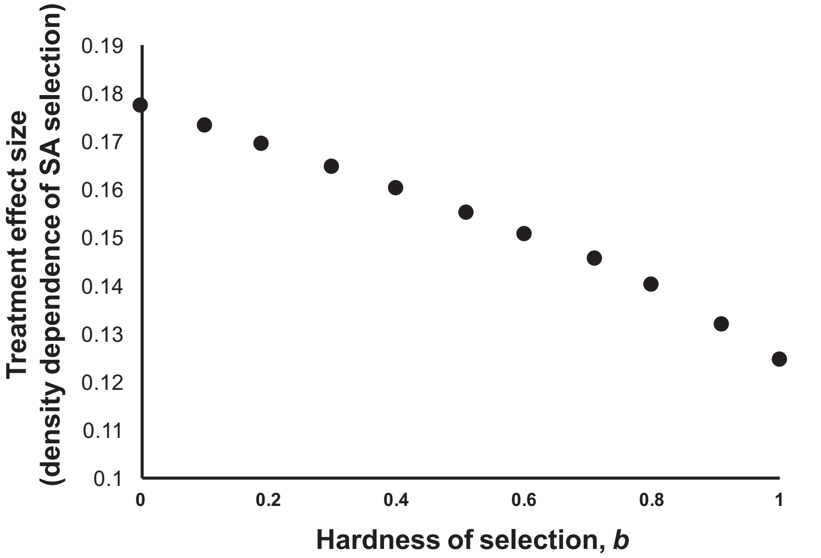

Figure 3. **The strength of density-dependence of sexually antagonistic natural selection declines with hardness of selection, *b*. Data are from a reanalysis of a previously published experiment examining density effects on disruptive selection in red spotted newts, *N. viridescens* (De Lisle and Rowe 2015). Effect size on the y-axis is the difference in curvature of the fitness surface along the axis of morphological sexual dimorphism, calculated as $s^T\gamma_{high}s - s^T\gamma_{low}s$, where s is the discriminant function vector defining the sexes and γ is the matrix of nonlinear selection gradients from a mixed model with relative fitness as the response. Relative fitness was calculated as in Equation (4), with the corresponding values of *b* given on the x-axis. Note that although De Lisle and Rowe treated selection as completely soft, their estimate of effect size differs slightly because they calculated absolute fitness by positivizing negative growth via addition of tank-specific constants, instead of addition of an experiment-wide constant as done here to make absolute fitness comparable across experimental units.**

of selection (Fig. 3) is consistent with the a priori hypothesis and reinforces the early conclusions in this study that sexually antagonistic selection in this system arises causally from local resource competition between the sexes.

FIXED GROUPS, RANDOM GROUPS, AND THE STANDARDIZATION OF PHENOTYPES

The question about the appropriate scale of phenotypic standardization comes down to the issue of whether the phenotypic distribution can be controlled across treatment groups, that is, whether variation in the phenotypic distribution is fixed or random. Here, by variation, we mean any differences between groups in either the phenotypic mean or phenotypic variance. By “random” and “fixed” we refer to the source of variation in a predictor variable (Rencher and Christensen 2012), in this case some aspect (e.g., moment) of the phenotypic distribution. This is related to but slightly different from how treatment effects would be treated in a linear model. Under our proposed definition, population variation in selection may be treated as a random factor in a mixed model, even though variation in P is fixed. For example, Rundle et al. (2009) estimated variation in female preference (sexual selection) among populations of *Drosophila serrata*, treating among-population variation in selection as random. However, by mating females to a common set of males, variation in the phenotypic distribution (or in this case, lack thereof) was fixed (Chenoweth et al. 2012). Had they mated females to males from their source populations, variation in the phenotypic distribution would have been random.

Alternatively, consider an experiment in which case local phenotypic frequencies of phenotypes were manipulated to assess the frequency-dependent aspects of selection. In this case, variation in the phenotypic distribution is fixed. Thus, our use of “random” and “fixed” designs refers to whether variation in the phenotypic distribution among experimental units reflects a sample of uncontrolled variation in the phenotypic distribution (e.g., Gosden and Svensson 2008) or a specific subset of possible the phenotypic distributions (e. g. De Lisle and Rowe 2015).

The case of random distribution of the phenotypic is illustrated in the left panel of Figure 4. In this case, selection differs across categories (A) and (B), but populations in these categories also differ in their phenotypic distributions. Separating effects of the phenotypic distribution and category on selection in this case are difficult. One approach would be standardize phenotypes within populations, and then model selection as a function of category and mean (unstandardized) phenotype. In this scenario, selection estimates can be compared across groups differing in their phenotypic distributions. Differences in selection due to variation in the phenotypic distribution can then be assessed,

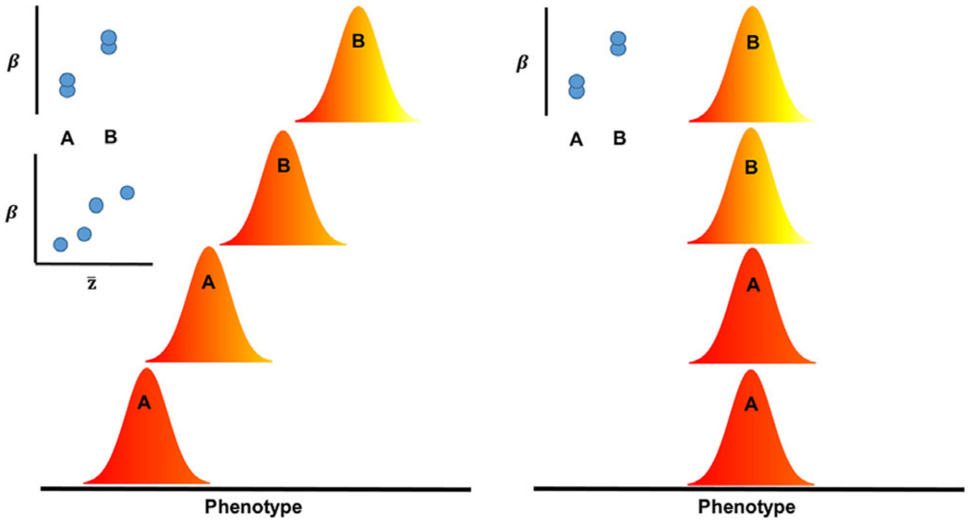

Figure 4. Random versus fixed experimental designs for comparing phenotypic selection. The left panel indicates a random design, where replicate populations (with phenotypic distributions represented by normal probability density functions) corresponding to two types of interest, (A) and (B), are chosen from existing variation. In this case, the design is random because variation in the phenotypic distribution among replicates represents uncontrolled random differences. Selection (illustrated by differential shading under the phenotype distributions) differs across fixed factors (A) and (B), and the populations in the two categories also differ in the location of their phenotypic distribution as a result of the random design. In this case, assessment of a main effect of (A) and (B) independent of among-population variation in $\bar{z}$ would require incorporating $\bar{z}$ as a covariate in a linear model (either a mixed model with relative fitness as a response to estimate β, or in a higher level analysis of the β estimates). In the right panel, the phenotypic distribution is controlled in replicate populations, rather than sampled from random variation, and effects of factors (A) and (B) are not confounded with variation in the phenotypic distribution. In general, in random designs, it will be difficult or impossible to ascertain the scenarios illustrated in Figure 1.

although the hypothesis that groups are also evolving on different selective surfaces can never be completely rejected when groups also differ in the phenotypic distribution because these scenarios are not exclusive. For example, Steele et al. (2011) used a design with random variation in the phenotypic distribution to investigate causes of variation in sexual selection on male body size in the damselfly *Enallagma aspersum*. They found that strength and form of selection was governed by both the location (relative to females) and the variance of the phenotypic distribution. The approach of Steele et al. illustrates how within-group standardization in a random design can be combined with higher level analysis of the importance of the phenotypic distribution in generating variation in selection.

Alternatively, when variation in the phenotypic distribution is fixed it may be appropriate to standardize phenotypes across the entire experiment, because in this case the phenotypic distribution is either constant across groups (e.g., right panel of Fig. 4) or is likely to be part of the manipulation of interest in comparisons among groups. For example, in an artificial pond experiment with *N. viridescens*, De Lisle and Rowe (2015) manipulated the phenotypic distribution to achieve two levels of phenotype frequency. In this case, standardizing within treatments would be inappropriate, because the goal was to examine how selection on a given phenotype changed with frequency; standardizing within groups would have made such comparisons impossible because phenotypes would not be comparable across treatments.

Discussion

Understanding the ecological causes of natural selection is key to understanding phenotypic evolution and the evolution of biological diversity. The statistical comparison of selection among groups provides one powerful approach for such studies. However, among-group comparison of selection is complicated by the fact that selection can differ for multiple reasons: differences in the covariance between phenotype and fitness, different phenotypic distributions, or differences in mean fitness. Whether these proximate sources of variation in selection can be recovered in among-group comparisons will depend on the scale at which fitness is relativized and at what level phenotypic traits are standardized. Thus, inherent in all comparative studies of selection is a critical and implicit, but often underappreciated, choice on the scale at which standardization should be carried out. Here, we have suggested that in general fitness should be relativized within groups whenever selection is soft—that is, whenever groups correspond to the scale at which (phenotype-dependent) population regulation is expected to occur—and relativized across groups whenever selection is hard. Many evolutionary biologists are interested in forms of selection that would be characterized as frequency- and density-dependent, that is, soft selection (Reznick 2016). Under this scenario, fitness should be relativized at the level in an experiment corresponding to the scale where the competitive interactions between individuals take place. That spatial scale is usually the lowest level of replication (e.g., local neighborhoods; Brandon 1990, Svensson and Sinervo 2000; 2004).

One counterpoint is that if researchers are primarily interested in the functional properties of how a phenotype determines fitness, and how that may change across environments, then the relationship between phenotype and *absolute* fitness could be of interest, which circumvents the issues raised here. Such forms of selection driven by, for example, global biomechanical performance of the phenotype that do not necessarily change with density or frequency. This might be similar to hard selection, that is, the relationship between fitness and local phenotype frequency distribution might not matter. Although such scenarios occur and this could be a sensible argument in some cases (and several studies in Table 1 take this approach), it is worthwhile to recall why linear selection is of interest to begin with: because the covariance between relative fitness and phenotype (or equally the variance weighted regression) determines the change in population mean phenotype. Many evolutionary biologists are interested in linear selection because it is the theoretically appropriate predictor of evolutionary change in mean phenotype regardless the shape of the joint distribution of phenotype and fitness (Rice 2004). If we are interested in obtaining a more nuanced understanding of the fitness surface, then linear/polynomial approximations are inappropriate to begin with (Schluter 1988; Schluter and Nychka 1994). A rich literature exists on such nonparametric approximations of the fitness surface, and these approaches are perhaps most appropriate if the primary interest is in uncovering functional relationships between phenotype and absolute fitness.

A recent discussion about the ecological causes and mechanisms of nonadditive selection through indirect ecological effects illustrate some of the conceptual problems we have raised in this article (ter Horst et al. 2015; 2017; Bolstad 2017). Consider the case where one species (e.g., an herbivore) exerts selection on a trait of another species (e.g., plant defense). Does the exerted selection remain constant when another species (e.g., a second herbivore) is present? Or do indirect ecological effects (e.g., interactions between herbivores) lead to nonadditive selection, and thus "diffuse" coevolution (Inouye and Stinchcombe 2001)? Key to testing this hypothesis is comparison of the strength of selection on a focal species when both manipulated species are present to a null model of additive selection constructed using selection estimates obtained from each species alone (ter Horst et al. 2015). ter Horst et al. (2015) argued that in such an experiment, only nonadditive effects that result from changes in the absolute fitness–phenotype relationship should qualify as indirect ecological effects on selection; interactive effects arising from a second-species effect on focal mean fitness are irrelevant, as such effects do not to represent a physical interaction between the two herbivore species (ter Horst et al. 2017). Following this logic, ter Horst et al. (2015) argued that fitness should be relativized across all treatments in experiments on nonadditive selection. However, global relativization in this case raises several difficulties. First, relativizing fitness across the experiment will not estimate a strength of selection that is useful for predicting response to selection (which in this case must be soft) or standardized in a way to make it comparable to other studies. That is, the interaction gradient estimate will be biased by the mean finesses of the other treatment groups in the experiment. Second, real indirect ecological effects on selection could manifest as mean-fitness differences (Bolstad 2017). For instance, the presence of a competitor could reduce consumption rates of an herbivore, leading to an interaction for focal mean fitness, without a change in absolute fitness–phenotype covariance. Third, total selection on a trait is likely to follow multiplicatively from different fitness components or result from several different interactions, rather than selection gradients being additive (see Bolstad 2017 for in-depth discussion and motivation). The controversy over assessment of nonadditive selection (Bolstad 2017, ter Horst et al. 2017) and the associated issue of the appropriate level of fitness relativization underscores the conceptual challenges associated with among-group comparisons of selection. In these and similar controversies it may be useful to explore if or how treatment effects change with the softness of selection, for example, in an approach similar to that taken in Figure 3.

Although we have attempted to suggest guidelines for the standardization choice in comparative selection analyses, in many cases the appropriate approach will depend on nuances of study design, biological details of the study system, and the question of interest. For example, in comparing the strength or form of selection across the sexes, theory justifies the separate relativization of male and female fitness (components) because under stable 50:50 sex ratios male and female mean fitness must be equal (Fisher 1930), although they may not appear to in any particular situation because of chance effects or the of use different components to approximate male and female fitness (e.g., mating success vs. fecundity). That is, selection in males and females is conceptually related to soft selection, because each sex (environment) will contribute an equal (fixed) fraction of gametes to the fertilized zygotes that will comprise the next reproductive generation (Levene 1953, Kidwell et al. 1977). Yet, in an experiment designed to study the evolution of departures from a Fisherian sex ratio, separate relativization of male and female fitness would be inappropriate. In another example, studies of "evolutionary rescue," an effect where adaptation by natural selection increases population mean fitness of threatened populations and so rescues them from extinction, absolute fitness may be a more important metric than relative fitness (Gomulkiewicz and Holt 1995). Thus, appropriate standardization choice may vary depending on the specific questions that are being addressed (see also terHorst et al. 2015; 2017).

Statistical descriptions of natural selection are eminently useful yet plagued with conceptual and statistical difficulties. Such difficulties should be seen as methodological challenges and motivate the formulation of appropriate a priori hypotheses. Here we have highlighted the scale of fitness and phenotype standardization as one issue that we believe has received less direct attention than deserved. Our view that experiment-wide relativization of fitness may rarely be justified might be controversial to some evolutionary biologists, but should be natural to others (e.g., evolutionary ecologists working with how local interactions influence fitness). The need for more explicit justification for standardization decisions in comparative studies of phenotypic selection is therefore a key issue that needs to be addressed in every comparative selection study.

ACKNOWLEDGMENTS

SPD is grateful for past discussions on these issues with D. Punzalan and S. Wadgymar. Comments from two anonymous reviewers improved the presentation of this work. Funding was obtained from the Swedish Research Council (VR: grant no. 621-2012-3768) to EIS and the John Templeton Foundation (grant no. 60501).

LITERATURE CITED

Agrawal, A. F. 2010. Ecological determinants of mutation load and inbreeding depression in subdivided populations. Am. Nat. 176:111–122.

Arnold, S. J., and M. J. Wade. 1984a. On the measurement of natural and sexual selection: theory. Evolution 38:709–719.

———. 1984b. On the measurement of natural and sexual selection: applications. Evolution 38:720–734.

Austen, E. J., and A. E. Weis. 2015. What drives selection on flowering time? An experimental manipulation of the inherent correlation between genotype and environment. Evolution 69:2018–2033.

Barton, N. H., and M. Turelli. 1989. Evolutionary quantitative genetics: how little do we know? Ann. Rev. Genet. 23:337–370.

Bolstad, G. H. 2017. Quantifying nonadditive selection caused by indirect ecological effects: comment. Ecology 98:278–282.

Brandon, R. N. 1990. Adaptation and environment. Princeton Univ. Press, Princeton, NJ.

Brodie, E.D., III, and F. J. Janzen. 1996. On the assignment of fitness values in statistical analyses of selection. Evolution 50:437–442.

Calsbeek, R. 2009. Experimental evidence that competition and habitat use shape the individual fitness surface. J. Evol. Biol. 22:97–108.

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

Chenoweth, S. F., H. D. Rundle, and M. W. Blows. 2010. The contribution of selection and genetic constraints to phenotypic divergence. Am. Nat. 175:186–196.

Chenoweth, S. F., J. Hunt, and H. D. Rundle. 2012. Analysing and comparing the geometry of individual fitness surfaces. Pp. 126–149 *in* E. I. Svensson and R. Calsbeek, eds. The adaptive landscape in evolutionary biology. Oxford Univ. Press, Oxford, U.K.

Curtis, S., J. L. Sztepanacz, B. E. White, K. A. Dyer, H. D. Rundle, and P. Mayer. 2013. Epicuticular compounds of *Drosophila subquinaria* and *D. recens*: identification, quantification, and their role in female mate choice. J. Chem. Ecol. 39:579–590.

Darwin, C. 1859. On the origin of species by means of natural selection. J. Murray, London.

De Lisle, S. P., and L. Rowe. 2015. Ecological character displacement between the sexes. Am. Nat. 186:693–707.

Fisher, R. 1930. The genetical theory of natural selection. Oxford Univ. Press, Oxford, U.K.

Gershman, S., M. Delcourt, and H. D. Rundle. 2014. Sexual selection on *Drosophila serrata* male pheromones does not vary with female age or mating status. J. Evol. Biol. 27:1279–1286.

Gomulkiewicz, R., and R. D. Holt. 1995. When does evolution by natural-selection prevent extinction? Evolution 49(1):201–207.

Gomulkiewicz, R., and M. Kirkpatrick. 1992. Quantitative genetics and the evolution of reaction norms. Evolution 46:390–411.

Goodnight, C. J., J. M. Schwartz, and L. Stevens. 1992. Contextual analysis of models of group selection, soft selection, hard selection, and the evolution of altruism. Am. Nat. 140:743–761.

Gosden, T. P., and E. I. Svensson. 2008. Spatial and temporal dynamics in a sexual selection mosaic. Evolution 62:845–856.

Gosden, T.P., H.D. Rundle, and S.F. Chenoweth. 2014. Testing the correlated response hypothesis for the evolution and maintenance of male mating preferences in *Drosophila serrata*. J. Evol. Biol. 27:2106–2112.

Grafen, A. 1987. Measuring sexual selection: why Bother? Pp. 221–223 *in* J. W. Bradbury and M. Andersson, eds. Sexual selection: testing the alternatives. Jon Wiley & Sons, New York.

———. 1988. On the uses of data on lifetime reproductive success. Pp. 454–471 *in* T. H. Clutton-Brock, ed. Reproductive success. Univ. of Chicago Press, Chicago.

Haldane, J. B. S. 1937. The effect of variation on fitness. Am. Nat. 71:337–349.

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

Inouye, B., and J. R. Stinchcombe. 2001. Relationships between ecological interaction modifications and diffuse coevolution: similarities, differences, and causal links. Oikos 95:323–360.

Keagy, J., L. Lettieri, and J. W. Boughman. 2016. Male competition fitness landscapes predict both forward and reverse speciation. Ecol. Lett. 19:71–81.

Kelley, J. L., J. R. Stinchcombe, C. Weinig, and J. Schmitt. 2005. Soft and hard selection on plant defence traits in *Arabidopsis thaliana*. Evol. Ecol. Res. 7:287–302.

Kidwell, J. F., M. T. Clegg, F. M. Stewart, and T. Prout. 1977. Regions of stable equilibria for models of differential selection in the two sexes under random mating. Genetics 85:171–183.

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

Kokko, H., and J. Lindstrom. 1997. Measuring the mating skew. Am. Nat. 149:794–799.

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

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

Levene, H. 1953. Genetic equilibrium when more than one ecological niche is available. Am. Nat. 87:331–333.

Losos, J. B., T. W. Schoener, and D. A. Spiller. 2004. Predator-induced behaviour shifts and natural selection in field-experimental lizard populations. Nature 432:505–508.

Lush, J. L. 1945. Animal breeding plans. Iowa Univ. Press, Ames, IA.

Lynch, M., and B. Walsh. 1998. Genetics and the analysis of quantitative traits. Sinaur Associates, Inc., Sunderland, MA.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical inference and biological interpretation. Evolution 41:1149–1161.

Morrissey, M. B. 2014. In search of the best methods for multivariate selection analysis. Methods Ecol. Evol. 5:1095–1109.

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

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

Price, T., C. Brown, and M. Bomberger Brown. 2000. Evaluation of selection on cliff swallows. Evolution 54:1824–1827.

Punzalan, D., F. H. Rodd, and L. Rowe. 2010. Temporally variable multivariate sexual selection on sexually dimorphic traits in a wild insect population. Am. Nat. 175:401–414.

Rausher, M. D. 1992. The measurement of selection on quantitative traits: biases due to environmental covariances. Evolution 46:616–626.

Rencher, A. C., and W. F. Christensen. 2012. Methods in multivariate analysis. 3rd ed. John Wiley & Sons, Hoboken, New Jersey.

Reznick, D. 2016. Hard and soft selection revisited: how evolution by natural selection works in the real world. J. Hered. 107:3–14.

Rice, S. H. 2004. Evolutionary theory: mathematical and conceptual foundations. Sinaur Associates, Sunderland, MA.

Rundle, H. D., and S. F. Chenoweth. 2011. Stronger convex (stabilizing)selection on homologous sexual display traits in females than in males: a multipopulation comparison in *Drosophila serrata*. Evolution 65:893–899.

Rundle, H. D., S. F. Chenoweth, and M. W. Blows. 2008. Comparing complex fitness surfaces: among-population variation in mutual sexual selection in *Drosophila serrata*. Am. Nat. 171:443–454.

———. 2009. The diversification of mate preferences by natural and sexual selection. J. Evol. Biol. 22:1608–1615.

Rundle, H. D., and K. A. Dyer. 2015. Reproductive character displacement of female mate preferences for male cuticular hydrocarbons in *Drosophila subquinaria*. Evolution 69:2625–2637.

Schluter, D. 1988. Estimating the form of natural selection on a quantitative trait. Evolution 42:849–861.

Schluter, D., and D. Nychka. 1994. Exploring fitness surfaces. Am. Nat. 143:597–616.

Siepielski, A. M., M. B. Morrissey, M. Buoro, S. M. Carlson, C. M. Caruso, S. M. Clegg, T. Coulson, J. Di Battista, K. M. Gotanda, C.D. Francis et al. 2017. Precipitation drives global variation in natural selection. Science 355:959–962.

Start, D., and B. Gilbert. 2016. Host-parasitoid evolution in a metacommunity. Proc. R. Soc. B. 283:20160477.

Steele, D. B., A. M. Siepielski, and M. A. McPeek. 2011. Sexual selection and temporal phenotypic variation in a damselfly population. J. Evol. Biol. 24:1517–1532.

Svensson, E., and B. Sinervo. 2000. Experimental excursions on adaptive landscapes: density-dependent selection on egg size. Evolution 54: 1396–1403.

Svensson, E. I., and B. Sinervo. 2004. Spatial scale and temporal component of selection in side-blotched lizards. Am. Nat. 165:726–734.

ter Horst, C. P., J. A. Lau, I. A. Cooper, K. R. Keller, R. J. La Rosa, A. M. Royer, E. H. Schultheis, T. Suwa, and J. K. Conner. 2015. Quantifying nonadditive selection caused by indirect ecological effects. Ecology 96:2360–2369.

ter Horst, C. P., J. A. Lau, and J. K. Conner. 2017. Quantifying nonadditive selection caused by indirect ecological effects: reply. Ecology 98:1171–1175.

Turelli, M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Popul. Biol. 25:138–193.

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

Wallace, B. 1968. Polymorphism, population size, and genetic load. Pp. 87–108 *in* R. C. Lewontin, ed. Population biology and evolution. Syracuse University Press, Syracuse, NY.

———. 1975. Hard and soft selection revisited. Evolution 29:465–473.

Waller, J., and E. I. Svensson. 2016. The measurement of selection when detection is imperfect: how good are naïve methods? Methods Ecol. Evol. 7:538–548.

Weis, A. E., K. M. Turner, B. Petro, E. J. Austen, and S. M. Wadgymar. 2015. Hard and soft selection on phenology through seasonal shifts in the general and social environments: a study on plant emergence time. Evolution 69:1361–1374.

White, A. J., and H. D. Rundle. 2015. Territory defense as a condition-dependent component of male reproductive success in *Drosophila serrata*. Evolution 69:407–418.

Whitlock, M. C. 2002. Selection, load, and inbreeding depression in a large metapopulation. Genetics 160:1191–1202.

Zeng, Z.-B. 1988. Long-term correlated response, interpopulation covariation, and interspecific allometry. Evolution 42:363–374.

Associate Editor: D. Roff
Handling Editor: M. Noor

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Supplementary Material