# The Misuse of BLUP in Ecology and Evolution

Jarrod D. Hadfield,[1,*] Alastair J. Wilson,[1] Dany Garant,[2] Ben C. Sheldon,[3] and Loeske E. B. Kruuk[1]

1. Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom; 2. Département de Biologie, Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1, Canada; 3. Edward Grey Institute, Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

*Submitted June 16, 2009; Accepted August 5, 2009; Electronically published November 18, 2009*

*Online enhancement:* appendix.

abstract: Best linear unbiased prediction (BLUP) is a method for obtaining point estimates of a random effect in a mixed effect model. Over the past decade it has been used extensively in ecology and evolutionary biology to predict individual breeding values and reaction norms. These predictions have been used to infer natural selection, evolutionary change, spatial-genetic patterns, individual reaction norms, and frailties. In this article we show analytically and through simulation and example why BLUP often gives anticonservative and biased estimates of evolutionary and ecological parameters. Although some concerns with BLUP methodology have been voiced before, the scale and breadth of the problems have probably not been widely appreciated. Bias arises because BLUPs are often used to estimate effects that are not explicitly accounted for in the model used to make the predictions. In these cases, predicted breeding values will often say more about phenotypic patterns than the genetic patterns of interest. An additional problem is that BLUPs are point estimates of quantities that are usually known with little certainty. Failure to account for this uncertainty in subsequent tests can lead to both bias and extreme anticonservatism. We demonstrate that restricted maximum likelihood and Bayesian solutions exist for these problems and show how unbiased and powerful tests can be derived that adequately quantify uncertainty. Of particular utility is a new test for detecting evolutionary change that not only accounts for prediction error in breeding values but also accounts for drift. To illustrate the problem, we apply these tests to long-term data on the Soay sheep (*Ovis aries*) and the great tit (*Parus major*) and show that previously reported temporal trends in breeding values are not supported.

*Keywords:* BLUP, breeding value, quantitative genetics, selection, evolution.

## Introduction

> To state that BLUP is unbiased by changing the usual definition of bias seems to be a rather liberal use of the language. Besides, the term “best” is somewhat misleading. (Blasco 2001, p. 2027)

Best linear unbiased prediction (BLUP) was largely developed by Henderson (e.g., Henderson 1950, 1976) in order to predict the expected phenotype of an animal’s offspring using an individual’s breeding value. This is achieved by using phenotypic information collected both on the individual and its relatives. By obtaining these predictions, animal breeders are able to select parents whose offspring are expected to have desirable properties. Not only does this yield a faster response than direct selection on parental phenotype, but it also means that animal breeders can apply selection even when phenotypic data on the parents are unavailable. An extreme case is in sex-limited traits, such as milk yield in dairy cattle. Because the potential number of calves a bull can sire far exceeds that which a cow can bear, selection is more efficient if bulls as opposed to cows are selected as parents. Phenotypic selection in this case is not possible, because no bull has a milk yield, but with BLUP the expected milk yield of a bull’s daughter can be predicted using the milk yield of that bull’s female relatives. Although developed in this context, BLUP is, and is used as, a more general method for predicting random effects in a variety of fields such as geology and actuarial science (Robinson 1991).

The points that we develop in this article are concerns about the serious problems that may arise when BLUP is used to address questions for which it was not originally intended. In evolutionary quantitative genetics, these questions have included inferring natural selection on breeding value, spatial structuring of breeding values, and changes in breeding value over time (see table 1 for a summary of studies in wild and field populations, many involving the authors of this article). It is in the context of these questions that we will mainly develop our critique of BLUP, although it should be understood that analogous procedures, such as exploring individual effects (Martin and Réale 2008; Moyes et al. 2009) and reaction norms (Brommer et al. 2005; Nussey et al. 2005), suffer from the same sort of problems.

[*] Corresponding author; e-mail: j.hadfield@ed.ac.uk.

Am. Nat. 2010. Vol. 175, pp. 116–125. © 2009 by The University of Chicago. 0003-0147/2010/17501-51356$15.00. All rights reserved.
DOI: 10.1086/648604

**Table 1:** Studies of wild or field populations that have used best linear unbiased prediction to answer questions regarding evolutionary change (E) in breeding value, selection (S) on breeding value, genetic differences between groups (G) of individuals, or genetic covariances (C)

| Species | Population | Trait(s) | Test | Reference |
|---|---|---|---|---|
| Collared flycatcher | Gotland, Sweden | Condition | E + S | Merilä et al. 2001*a* |
| Collared flycatcher | Gotland, Sweden | Condition | S | Merilä et al. 2001*b* |
| Collared flycatcher | Gotland, Sweden | Tarsus length | S | Kruuk et al. 2001 |
| Collared flycatcher | Gotland, Sweden | Clutch size + laying date | E + S + G | Sheldon et al. 2003 |
| Collared flycatcher | Gotland, Sweden | Forehead + wing patch size | E + S | Garant et al. 2004*a* |
| Red deer | Rum, United Kingdom | Antler size | E + S | Kruuk et al. 2002 |
| Red deer | Rum, United Kingdom | Sex-specific fitness | C | Foerster et al. 2007 |
| Wild radish | New York | >3 | C | Agrawal et al. 2002 |
| Wild radish | New York | >3 | S + C | Agrawal et al. 2004 |
| Bighorn sheep | Ram Mountain, Canada | Horn length + body weight | E + S + G | Coltman et al. 2003 |
| Red squirrel | Kluane, Canada | Parturition date | E | Réale et al. 2003 |
| Great tit | Wytham Woods, United Kingdom | Body weight | E + S | Garant et al. 2004*b* |
| Great tit | Wytham Woods, United Kingdom | Body weight | E + S | Garant et al. 2005 |
| Great tit | Vieland, Netherlands | Clutch size | G | Postma and van Noordwijk 2005 |
| Great tit | Vieland, Netherlands | Clutch size | E + G | Postma et al. 2007 |
| Great tit | Hoge Veluwe, Netherlands | Laying date | E + S | Gienapp et al. 2006 |
| Blue tit | Corsica/La Rouvière, France | Tarsus length + body weight | E + S | Charmantier et al. 2004 |
| Scarlet gilia | Colorado | >3 | C | Juenger et al. 2005 |
| Mute swan | Abbotsbury, United Kingdom | Laying date + clutch size | E | Charmantier et al. 2006 |
| Soay sheep | St Kilda, United Kingdom | Body size | E | Wilson et al. 2007 |
| Side-blotched lizard | California | Clutch size | S | Sinervo and McAdam 2008 |
| Red-billed gull | Kaikoura, New Zealand | Body size | E | Teplitsky et al. 2008 |
| Common evening primrose | Colorado | >3 | C | Johnson et al. 2009*b* |
| Common evening primrose | Colorado | >3 | S | Johnson et al. 2009*a* |
| *Arabidopsis thaliana* | Rhode Island | >3 | S + C | Stinchcombe et al. 2009 |

Note: Trait names have been omitted for studies that involve more than three traits. Table is updated from Postma (2006).

We identify three main properties of BLUP that have led to inferential problems, and we present examples that best exemplify each issue. Our first criticism is that the desirable properties of BLUP hold only in the context of predicting the mean of a single breeding value and that these properties do not extend to other aspects of an individual's breeding value (such as the squared deviation from the population mean) and do not extend to higher-level statistics summarizing the distribution of breeding values in a population. Our second criticism is that BLUP is an unbiased predictor of breeding value only when the model used to make the predictions is the correct one (see Postma 2006 for a useful discussion). Although the true model is unknowable, it is often hoped that the model used is close enough for robust conclusions to be drawn. However, in evolutionary biology and ecology, BLUPs are often used to test for patterns in breeding values that are deemed too complicated to be captured in the model used for prediction. In these instances, the distributions of BLUP are biased toward a null model. Although it may be argued that such a test is conservative, it should be borne in mind that the statistical null model is often different from what would be considered the biological or scientific null model. In many instances the statistical null model is that patterns at the genetic level follow patterns at the phenotypic level, as would be observed if those patterns were causally determined by phenotype. For example, when differences between groups of individuals are not explicitly modeled, tests of genetic differentiation using differences in predicted breeding value are not biased toward genetic homogeneity but toward finding genetic differences even when only environmental differences exist. Our third and final criticism is that the large amount of prediction error and complicated patterns of dependence in predicted breeding values are usually not accounted for when quantifying uncertainty, and this can lead to extreme anticonservatism. Some of these criticisms have a long history in the animal breeding literature (e.g., Blasco 2001), and the first two criticisms have already been made in the context of evolutionary biology (Postma 2006; Hadfield 2008; O'Hara et al. 2008). However, the weight of these criticisms has not been widely appreciated, and in conjunction with our final point, we discourage future use of BLUP as an inferential tool in the fields of ecology and evolutionary biology.

We illustrate our three criticisms with both toy and real examples. We define a consistent model to be one in which the BLUPs are predicted using a model that captures the process that is to be explored, and we hold an inconsistent model to be one in which the BLUPs are used to show a pattern that is not explicitly formulated in the model. In general, a consistent model can usually be constructed and the hypothesis formulated directly from the estimated (co)variance components. For example, the strength of selection acting on breeding value (the genetic selection gradient) is defined as the genetic covariance between the trait and fitness divided by the genetic variance for the trait. Using generalized linear mixed models, consistent estimators of these (co)variances, and hence the genetic selection gradient, can be obtained without the need to use BLUPs (Hadfield 2008). One exception where it is necessary to work with individual breeding values rather than some higher-level model parameter (such as a variance) is when trying to detect change in breeding values over the course of a study. However, even in these cases we do not advocate the use of BLUP and strongly recommend using the complete posterior distribution of breeding values from a Bayesian analysis (see Walsh and Lynch 2009 for a review). We illustrate why, using data from our previous studies of the Soay sheep (*Ovis aries*) and the great tit (*Parus major*), in which BLUP methodology had suggested highly significant evolutionary change (Garant et al. 2004*b*; Wilson et al. 2007). We show that the significance of these trends had been overestimated by several orders of magnitude.

## A Consistent Model Resulting in Bias

### *Estimating Additive Genetic Variance*

A well-known example (Henderson 1975), which has been discussed in the context of evolutionary biology (Postma 2006; O'Hara et al. 2008), is the difference between the variance in breeding value BLUPs and variance in true breeding values (the additive genetic variance) that is caused by prediction error. Although nobody would consider using the variance in BLUPs as a measure of additive genetic variance, the example does highlight one of the properties of BLUPs that can cause problems when they are misused. For example, when we fit a quantitative genetic model using restricted maximum likelihood (REML), we obtain asymptotically (with increasing sample size) unbiased estimates of the additive genetic variance in the base population (assuming the model is correct). This estimate, by its derivation, is an asymptotically unbiased estimate of the variance in true breeding values in the base population. However, it is well known that variance in BLUP breeding values is consistently less than the variance in true breeding values, even though they are predicted from a model that is consistent. The reason for this is simply prediction error around the true values. In figure

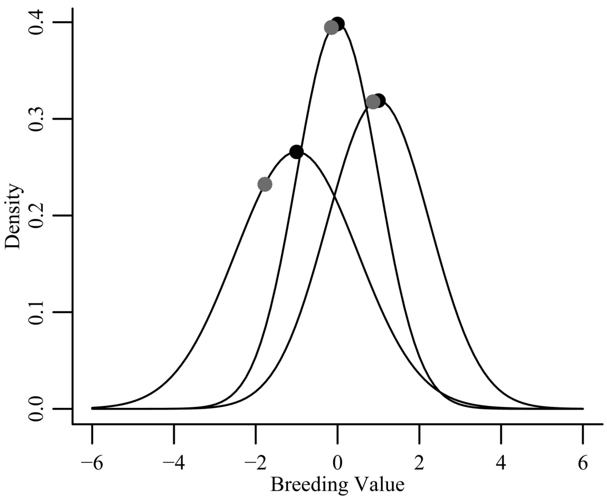

**Figure 1:** The normal densities represent the distribution of three breeding values known with uncertainty. The black dots represent the modes of these distributions that are equal to the best linear unbiased prediction breeding values, and the width of the distributions represents the degree of error in their estimation. The gray dots represent three random draws from the three normal densities and represent a single realization from the distribution of true breeding values.

1, three normal densities have been plotted. Each curve represents our state of knowledge about three individuals' breeding value. The peaks of each curve are the most likely value for each breeding value (*black dots*) and represent the BLUPs. The widths of the curves are related to how certain we are in each breeding value. The gray dots are random draws from these distributions and represent one possible configuration of true breeding values. Because of the uncertainty around each peak, the peaks of the curves tend to be more clustered (have less variance) than possible configurations of true breeding values, and this is why the variance in BLUPs is always less than the additive genetic variance.

The problem becomes even more apparent if we consider an extreme example where the three individuals had no relatives in the sample and had missing phenotype records. In this instance all three curves would be identical, with a mean of 0 and a variance equal to the additive genetic variance. As before, random possible configurations of true breeding values from these distributions would have a variance equal to the additive genetic variance, but the variance in BLUPs would be 0. In summary, although the breeding values have been predicted using a consistent model, the variance in BLUPs are a downward-biased estimator of the variance in true breeding values.

## An Inconsistent Model Resulting in Bias

### *Estimating Selection on Genotype*

One use of BLUPs is to estimate selection on breeding values (Rausher 1992). The genetic selection gradient ($\beta_G$) is defined as the genetic covariance between the trait and fitness divided by the additive genetic variance in the trait

$$\beta_G = \frac{\sigma_{a,w}}{\sigma_a^2}, \quad (1)$$

where $a$ is breeding value and $w$ is fitness. Several studies have estimated selection by using BLUPs ($\hat{a}$) in place of $a$,

$$\hat{\beta}_G = \frac{\sigma_{\hat{a},w}}{\sigma_{\hat{a}}^2}, \quad (2)$$

by regressing individual fitnesses on the BLUPs. Postma (2006) pointed out that because the variance in BLUPs ($\sigma_{\hat{a}}^2$) is a downward-biased estimator of the additive genetic variance ($\sigma_a^2$) for reasons discussed above, $\hat{\beta}_G$ is an upward-biased estimator of $\beta_G$. Hadfield (2008) pointed out that $\sigma_{\hat{a},w}$ is also strongly biased but in a way that depends on the phenotypic selection gradient. This occurs because the BLUPs are predicted using a model that does not explicitly account for any genetic covariance between the trait and fitness. In extreme cases the genetic selection gradient estimated in this way can even have the wrong sign if the environmental covariance between the trait and fitness is large and opposite in sign to the genetic covariance. The solution is to fit a bivariate model of fitness and the trait and estimate the selection gradient directly from the estimated (co)variances (i.e., $\hat{\beta}_G = \hat{\sigma}_{a,w}/\hat{\sigma}_a^2$) without ever touching the BLUPs. In the appendix in the online edition of the *American Naturalist* we provide example code to show how such an analysis could be performed.

This type of model has been used by Etterson and Shaw (2001) to estimate genetic selection differentials for an annual legume, and a few studies have reported genetic correlations between a trait and fitness, also estimated using bivariate models (Kruuk et al. 2002; Sinervo and McAdam 2008). Closer scrutiny of these latter two results are indicative of the problem. For example, using BLUP, Kruuk et al. (2002) estimated the genetic selection differential on antler size to be 0.158, and yet the genetic correlation between antler size and lifetime reproductive success was estimated to be −0.254. Since the genetic correlation is defined as the genetic selection differential multiplied by $(\sigma_w^2/\sigma_a^2)^{1/2}$, the two quantities should have the same sign given the variances have to be positive. The different signs arise because the BLUP-based estimate is biased toward the phenotypic selection differential of 0.449.

### *Estimating Genetic-Spatial Structuring*

In a similar fashion, breeding values predicted using spatially naive models have been used as a test for spatial variation in true breeding values. To show how biases arise that are analogous to the biases caused when using BLUP to obtain genetic selection gradients, we use simple simulations in which a population is subdivided into two interbreeding subpopulations. For ease, we use the Soay sheep pedigree from Wilson et al. (2007), which was collected over a 25-year period (1980–2004) on a feral population living on the islands of St Kilda, northwest Scotland (Clutton-Brock and Pemberton 2004). For each cohort in the pedigree we randomly assigned half of the individuals to either subpopulation A or B. In the first set of simulations, we simulated the scenario in which the mean phenotype of the two subpopulations differed but only because of environmental differences. This was achieved by simulating breeding values down the pedigree assuming an additive genetic variance of 1 but no genetic differences between the groups. Environmental deviations were then added to the breeding values, and these had a variance of 1 for each group, but the mean for group A was −0.5, and the mean for group B was 0.5. A simple animal model was fitted using the program

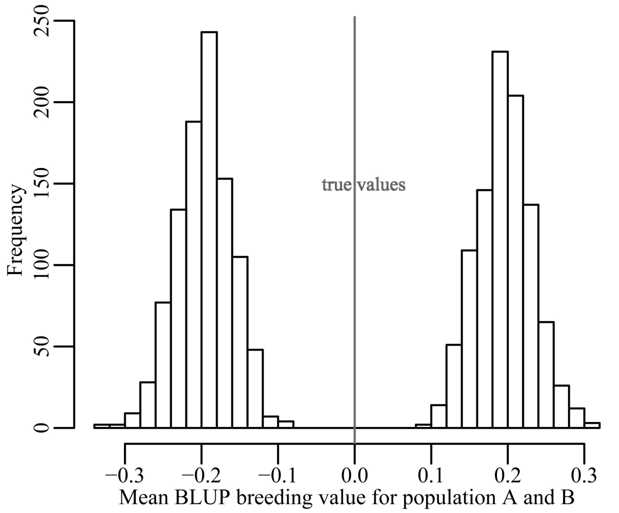

**Figure 2:** Mean breeding value best linear unbiased predictions (BLUPs) for two populations calculated for 1,000 simulated data sets. The true underlying mean breeding values did not differ and had an expectation of 0 (*vertical line*).

ASReml-R (Gilmour et al. 2002) with an intercept, an animal effect (breeding value), and a residual term. This was repeated for 1,000 simulated data sets, and the mean BLUP for the two subpopulations was calculated. Although the two subpopulations are genetically identical, the BLUPs suggest that there are strong genetic differences between the two subpopulations, with an average difference in mean BLUP (A − B) of −0.392 ± 0.001 (see fig. 2). A simple solution is to fit population as a fixed effect. This captures the environmental differences between the two groups, and the difference in mean BLUP between the two subpopulations for this model is close to being unbiased (0.0008 ± 0.0004).

If there were genetic differences between the subpopulations, however, then fitting such a model is still inappropriate because the BLUPs are predicted under the assumption that no genetic differences exist. For example, we could imagine a situation where genetic and environmental differences between the two populations cancel out to give the same mean phenotype—a phenomenon known as countergradient variation (Conover and Schultz 1995). To capture this scenario, we simulated data in the same way as before except the breeding values of animals in the base population were sampled from a normal distribution with unit variance and means of 0.5 (group A) or −0.5 (group B), rather than a common mean of 0. Two models were fitted to the data: one in which population is fitted as a fixed effect and one where population is fitted as both a fixed effect and a genetic group. Genetic groups do not seem to have been used outside of animal breeding but provide a way of modeling genetic structure in the base population (Robinson 1986; Westell et al. 1988). When genetic groups are fitted, all base individuals are assigned to a subpopulation that may differ in mean breeding value. Normal patterns of additive genetic inheritance are still assumed, such that the expected breeding value of an individual with both parents from population A would be 0.5 but an individual with one parent from population A and one from population B would have an expected breeding value of 0. As with fixed effects, the mean breeding value for each subpopulation cannot be uniquely estimated, but differences between them can be estimated when pedigree links exist between them (Quaas 1988). In this example the difference in breeding values between the two populations (A − B) has an expectation of 1, and we obtain estimates of this quantity by (*a*) comparing the mean BLUP of individuals from the two populations predicted using a model without genetic groups and (*b*) directly from the genetic group estimates themselves.

In figure 3 the results are shown for the two models. The histogram on the left is the difference between the mean BLUPs in the standard animal model (with group as a fixed effect) for the 1,000s simulations. These are significantly different from 0, indicating that the BLUPs are on average predicted to be different (−0.067 ± 0.0004). However, the difference has the wrong sign—the data were simulated so that group A's breeding values were a unit higher, although this spatially naive model suggests that group A's breeding

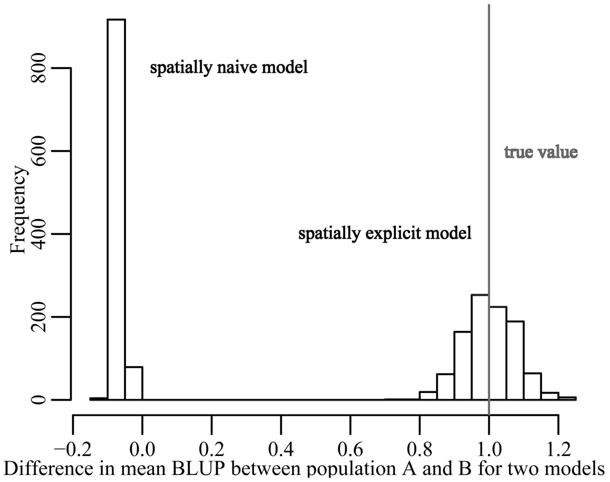

**Figure 3:** Difference between mean breeding value best linear unbiased predictions (BLUPs) for the two populations where genetic structuring of the base population was not explicitly modeled (left-hand distribution) and where genetic groups were fitted (right-hand distribution). The true underlying difference between the mean breeding values was 1 (*vertical line*).

values are actually smaller. The histogram on the right is the difference between the predictions for the two genetic groups, which gives unbiased estimates of the difference (1.002 ± 0.002).

## A Consistent Model without Bias but Strongly Anticonservative

### *Estimating Evolutionary Change*

In the previous sections we demonstrate that BLUP can be misleading because it gives biased parameter estimates. However, additional problems also arise with hypothesis testing and confidence interval estimation because BLUPs are often treated as independent. For instance, in the previous spatial example, where differences between subpopulations were entirely environmental, BLUPs from a model that had subpopulation as a fixed effect had very little bias. However, in this example, significance testing is actually conservative, with only 1 out of the 1,000 simulations being significant at the 5% level if each BLUP is treated as independent. However, in these simulations, individuals were assigned to a subpopulation randomly, and under more realistic scenarios where related individuals are more likely to be from the same subpopulation, these tests will generally be anticonservative. For example, we can rerun the above simulations but give offspring an 83% chance (as opposed to 50%) of belonging to the same population as their parents when both parents come from the same population. Here, 93 of the 1,000 simulations were significant at the 5% level, and presumably this anticonservatism would become worse under assortative mating.

Below, we show how this anticonservatism can become extreme when testing for evolutionary change, either as a response to artificial selection or as a response to natural selection. In order to distinguish between a genetic trend and a phenotypic trend caused by some concurrent environmental change, it is necessary to ask whether breeding values have changed over the course of the study. This is usually achieved by taking the mean BLUP breeding value for each cohort and seeing whether these means increase or decrease over time. When the effects of selection on the data are ignorable (sensu Rubin 1976), then the cohort mean BLUPs have the same expectation as the cohort means of the true breeding values, and the test is unbiased (see Im et al. 1989 for technical details). However, cohort mean BLUPs have less variance than the cohort means of the true breeding values because of prediction error, as discussed above. Moreover, because the prediction errors of relatives are usually positively correlated, there is usually more positive temporal autocorrelation in BLUPs than in true breeding values.

The expected variance and temporal autocorrelation in true breeding values depends on whether we are trying to say something about change in the actual sample means or to generalize from the sample means to what would happen under conceptual repetitions of the same "experiment." The distinction between these two levels of inference is illustrated in the two questions: Has the mean breeding value in the population changed? And has the mean breeding value in the population changed more than we expect by chance (drift; Hill 1971)? For the second and more interesting question, the variance of the cohort mean breeding values would vary due to changes in population size and will show temporal autocorrelation when individuals have relatives in cohorts other than their own. However, this question is often addressed by regressing the cohort mean BLUPs on some measure of time and testing for a significant slope under the assumption that the residuals are independent and identically distributed. This test is very anticonservative and will often reject the null hypothesis (no change) even when there is little evidence that mean breeding values have changed at all, significantly or not.

To understand the problem completely we need to work with three subtly different quantities: cohort means of the BLUPs ($\bar{\hat{\mathbf{a}}}$), cohort means of the actual unobserved breeding values ($\bar{\mathbf{a}}$), and cohort means of breeding values generated under hypothetical repeat sampling of the "experiment" under drift ($\bar{\bar{\mathbf{a}}}$).

If we assume that all variance components are known without error, then the actual cohort mean breeding values have the same expectation as the BLUP estimates, with some variation due to sampling error ($\mathbf{M}$):

$$\bar{\mathbf{a}} \sim N(\bar{\hat{\mathbf{a}}}, \mathbf{M}), \tag{3}$$

and the distribution of hypothetical breeding values have an expectation of 0 (in the absence of selection) but some variance due to drift ($\overline{\mathbf{G}}$):

$$\bar{\bar{\mathbf{a}}} \sim N(\mathbf{0}, \overline{\mathbf{G}}). \tag{4}$$

The key question is whether the change in cohort means of true breeding values ($\bar{\mathbf{a}}$) has been more extreme than what we would observe under drift (i.e., from random fluctuations in $\bar{\bar{\mathbf{a}}}$). This is usually tested by fitting a simple linear regression under the assumption that cohort mean BLUPs ($\bar{\hat{\mathbf{a}}}$) are identical and independently distributed after taking into account the time trend:

$$\bar{\hat{\mathbf{a}}} \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}, \sigma_{\mathrm{r}}^{2}\mathbf{I}), \tag{5}$$

where $\mathbf{X}$ is a design matrix with 1's in the first column and some continuous measure of time in the second column. The associated parameter vector $\hat{\boldsymbol{\beta}}$ is the intercept

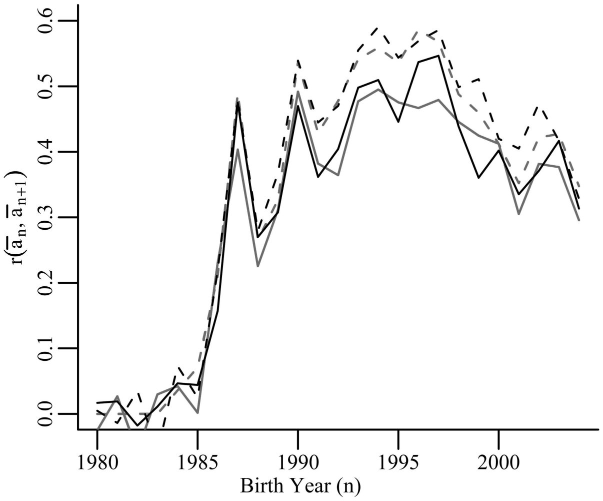

**Figure 4:** Sampling correlation between successive cohort mean breeding values due to correlated prediction error (*solid lines*) and finite sampling/drift (*dashed lines*) for body weight in the Soay sheep (Wilson 2007). The black lines are the posterior prediction error correlations between cohort means due to all model parameter uncertainty. The gray lines are the posterior prediction error correlations between cohort means when the variances are fixed at the posterior mode.

and slope of this regression; $\sigma_r^2$ is the residual variance around the regression, and the identity matrix **I** implies that the residuals are expected to be independent and have equal variance. In part, the validity of this test relies on both prediction error and drift causing independent fluctuations in (predicted) breeding values between generations, such that both $\overline{\mathbf{G}}$ and **M** are close to identity matrices. The term **M** can be derived analytically using well-known results for the prediction error (co)variances of BLUP (e.g., Mrode 1996), and $\overline{\mathbf{G}}$ can be obtained using results from Sorensen and Kennedy (1983). The full derivation of these matrices is given in the appendix (see also Walsh and Lynch 2009), but the key point is that sampling error induces positive correlations between cohort mean BLUPs and drift induces positive correlations between cohort mean breeding values, and so the assumption of independent residuals is never met.

As an example of these issues, we repeated the quantitative genetic analyses of body weight in Soay sheep (model 1 in Wilson et al. 2007) for which evidence of evolutionary change had been found in the form of a significant trend in predicted breeding values. In figure 4, the solid black line represents the expected correlation in prediction error between cohort mean breeding values in successive years, and the dashed black line represents the expected correlation between cohort mean breeding values in successive years due to drift. (i.e., the subdiagonals of **M** and $\overline{\mathbf{G}}$ rescaled to a correlation matrix). Under independence these correlations should be 0.

In reality, the variance components are never known exactly, especially in studies of wild populations where data sets tend to be quite small. If they are not known exactly, then this can induce further sampling correlations in the predicted breeding values (Sorensen and Kennedy 1984). However, the distribution of cohort mean breeding values in this instance is not in any recognizable form, and **M** cannot be obtained analytically. However, using Markov chain Monte Carlo (MCMC), the full posterior distribution of cohort mean breeding values is easy to obtain (Sorensen et al. 1994). Using MCMCglmm (J. D. Hadfield, unpublished manuscript), we fitted the same model to the sheep data (model 1 in Wilson et al. 2007) and obtained 1,000 samples of the joint posterior distribution of breeding values (see appendix). The correlation between successive cohort mean breeding values across the 1,000 samples is plotted as a solid gray line in figure 4. This represents the degree of prediction error correlation in breeding values including that induced by uncertainty in the variance components.

We can also regress each posterior sample of cohort mean breeding value on year to obtain the distribution of the slope coefficient for the genetic trend. This results in 1,000 samples from the posterior distribution of evolutionary change (see fig. 5). The Bayesian posterior mean of the slope and the standard BLUP analysis give identical answers regarding the rate of evolutionary change (0.0026 kg/year). However, the probability that breeding values were actually decreasing during the course of the study

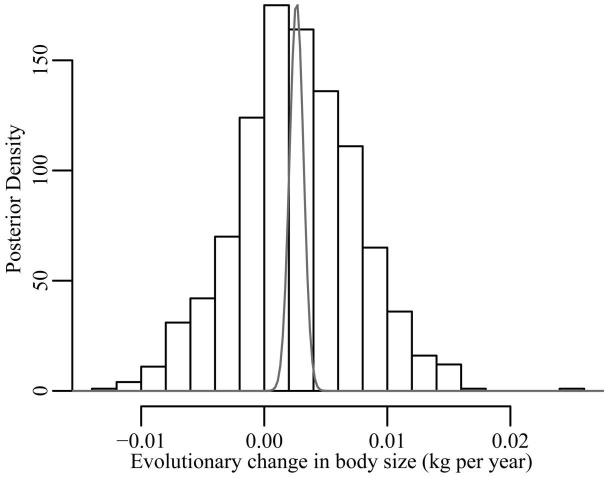

**Figure 5:** Posterior distribution for change in breeding value over time for body weight in Soay sheep (kg/year). The smooth line represents the estimate and sampling error derived from the standard best linear unbiased prediction model where the data were treated as independent.

was 0.283 from the Bayesian analysis. This is in direct contrast to the conclusions that were drawn form the standard BLUP analysis which suggested a significant ($P <$ .0001) increase. To demonstrate that this level of anticonservatism is not specific to the particular data set we used, we also refitted the model used by Garant (2004*a*). In this article, a significant ($P \ll .0001$) increase in breeding value (0.0020 g/year) was reported for fledgling condition in a population of great tits using data collected over a 36-year period (1965–2000) in Wytham Woods, United Kingdom (Perrins 1979). As with the sheep analysis, the Bayesian mean estimate was identical to the published REML estimate, but the significance of the trend was greatly reduced, with the probability of breeding values actually decreasing to 0.045.

These two analyses demonstrate that the $P$ values from regression of mean BLUP are very anticonservative. Furthermore, as formulated above, we are testing only whether the mean breeding value in the population changed after accounting for prediction error. If we want to test for a deterministic response, we need to ask whether this change is more than we expect by chance, by taking into account the variance in breeding values expected under drift ($\overline{\mathbf{G}}$). In the appendix, we provide an analytical test that is valid when the variance components are known without error. However, because this is rarely the case, we propose a test based on posterior predictive simulation. The concept is fairly simple: for each posterior sample we get an estimate of the additive genetic variance that we use to simulate replicated breeding values ($\tilde{\mathbf{a}}$) down the pedigree. Using these replicated breeding values we calculate a regression slope for evolutionary change, which has an expectation of 0 since we have not imposed selection in the simulation but some variation due to drift. We then calculate the proportion of iterations for which the slope calculated from the posterior sample of the actual breeding values exceeds that of the replicate breeding values. This proportion is the probability that the trend could not be due to drift. More formally, we evaluate the posterior predictive test:

$$\int \Pr\,[T(\bar{\tilde{\mathbf{a}}}|\boldsymbol{\theta}, \mathbf{y})] \geq \Pr\,[T(\bar{\hat{\mathbf{a}}}|\boldsymbol{\theta}, \mathbf{y})]d\boldsymbol{\theta}, \tag{6}$$

where the test statistic $T$ is the regression coefficient, and integration is performed over the joint distribution of the remaining model parameters such as the fixed effects and variance components ($\boldsymbol{\theta}$). Not surprisingly, the probability that the reported positive evolutionary change is in fact negative increases when we also take into account drift. For the sheep the probability increased to 0.357, and for the tits the probability increased to 0.127. In reality this should not be too surprising given that the magnitude of the change (as measured in phenotypic standard deviations) was small in each case (0.0008 for adult weight in sheep and 0.002 for fledgling condition in the tits).

## Summary

BLUP is used extensively in the agricultural sciences to predict which individuals should produce the best offspring. It was in this context that BLUP was developed, and for this purpose it performs well under a broad range of circumstances (Mrode 1996). More recently, BLUP has been used to answer a wide range of exciting questions in evolutionary biology and ecology (Postma 2006). We do not intend to diminish the importance of these questions, but we do wish to make the point that BLUP does not give satisfactory answers in these contexts. For many types of problem patterns in breeding values predicted using BLUP have been interpreted as genetic patterns, although the nature of the bias means that many of these patterns may actually be environmental in origin. More importantly, we demonstrate that alternative methods exist for answering most of these questions and that these methods are more powerful and less biased and measure uncertainty more accurately. For most types of analysis, this involves fitting a model that answers the question directly, for example, by specifying genetic groups or by formulating the test in terms of estimated variances and covariances. One consequence of fitting these models correctly will be to reveal the power issues that surround such analyses. Given this, we suggest the real difficulty will be to collect enough relevant data to say something substantive regarding processes that are inherently difficult to measure. If we assume a generation time of about 2 years for both species that we studied, then the rate of change as measured in haldanes was in both cases reasonably close to the median estimates of phenotypic evolutionary change (0.006) compiled by Kinnison and Hendry (2001). However, it should be recognized that on the timescale of most studies this rate of change is negligible. Unfortunately, the use of BLUP methodology has resulted in erroneously high levels of statistical significance, and this has been taken to imply biological significance despite very small effect sizes. We concur with Gienapp et al. (2006) that in most cases the power to reject neutral processes underlying genetic change is limited, especially in long-lived species with small population sizes.

## Acknowledgments

We thank S. Brotherstone, B. Hill, D. Nussey, J. Pemberton, A. Phillimore, J. Stinchcombe, and B. Walsh for discussion and comments on an earlier version of this manuscript. J.D.H. and A.J.W. are funded by the Natural Environment Research Council (United Kingdom) and L.E.B.K. by a Royal Society University Research Fellowship.

## Literature Cited

Agrawal, A. A., J. K. Conner, M. T. J. Johnson, and R. Wallsgrove. 2002. Ecological genetics of an induced plant defense against herbivores: additive genetic variance and costs of phenotypic plasticity. Evolution 56:2206–2213.

Agrawal, A. A., J. K. Conner, and J. R. Stinchcombe. 2004. Evolution of plant resistance and tolerance to frost damage. Ecology Letters 7:1199–1208.

Blasco, A. 2001. The Bayesian controversy in animal breeding. Journal of Animal Science 79:2023–2046.

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.

Charmantier, A., L. E. B. Kruuk, J. Blondel, and M. M. Lambrechts. 2004. Testing for microevolution in body size in three blue tit populations. Journal of Evolutionary Biology 17:732–743.

Charmantier, A., C. Perrins, R. H. McCleery, and B. C. Sheldon. 2006. Evolutionary response to selection on clutch size in a long-term study of the mute swan. American Naturalist 167:453–465.

Clutton-Brock, T. H., and J. M. Pemberton, eds. 2004. Soay sheep: dynamics and selection in an island population. Cambridge University Press, New York.

Coltman, D. W., P. O'Donoghue, J. T. Jorgenson, J. T. Hogg, C. Strobeck, and M. Festa-Bianchet. 2003. Undesirable evolutionary consequences of trophy hunting. Nature 426:655–658.

Conover, D. O., and E. T. Schultz. 1995. Phenotypic similarity and the evolutionary significance of countergradient variation. Trends in Ecology & Evolution 10:248–252.

Etterson, J. R., and R. G. Shaw. 2001. Constraint to adaptive evolution in response to global warming. Science 294:151–154.

Foerster, K., T. Coulson, B. C. Sheldon, J. M. Pemberton, T. H. Clutton-Brock, and L. E. B. Kruuk. 2007. Sexually antagonistic genetic variation for fitness in red deer. Nature 447:1107–1110.

Garant, D., B. C. Sheldon, and L. Gustafsson. 2004*a*. Climatic and temporal effects on the expression of secondary sexual characters: genetic and environmental components. Evolution 58:634–644.

Garant, D., L. E. B. Kruuk, R. H. McCleery, and B. C. Sheldon. 2004*b*. Evolution in a changing environment: a case study with great tit edging mass. 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.

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.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2002. ASReml user guide, release 1.0. http://www.vsn-intl.com.

Hadfield, J. D. 2008. Estimating evolutionary parameters when viability selection is operating. Proceedings of the Royal Society B: Biological Sciences 275:723–734.

Henderson, C. R. 1950. Estimation of genetic parameters. Annals of Mathematical Statistics 21:309–310.

———. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

———. 1976. Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83.

Hill, W. G. 1971. Design and efficiency of selection experiments for estimating genetic parameters. Biometrics 27:293–311.

Im, S., R. L. Fernando, and D. Gianola. 1989. Likelihood inferences in animal breeding under selection: a missing-data theory view point. Genetics Selection Evolution 21:399–414.

Johnson, M. T. J., M. Vellend, and J. R. Stinchcombe. 2009*a*. Evolution in plant populations as a driver of ecological changes in arthropod communities. Philosophical Transactions of the Royal Society B: Biological Sciences 364:1593–1605.

Johnson, M. T. J., A. A. Agrawal, J. L. Maron, and J. P. Salminen. 2009*b*. Heritability, covariation and natural selection on 24 traits of common evening primrose (*Oenothera biennis*) from a field experiment. Journal of Evolutionary Biology 22:1295–1307.

Juenger, T., T. C. Morton, R. E. Miller, and J. Bergelson. 2005. Scarlet gilia resistance to insect herbivory: the effects of early season browsing, plant apparency, and phytochemistry on patterns of seed fly attack. Evolutionary Ecology 19:79–101.

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

Kruuk, L. E. B., J. Merilä, and B. C. Sheldon. 2001. Phenotypic selection on a heritable size trait revisited. American Naturalist 158:557–571.

Kruuk, L. E. B., J. Slate, J. M. Pemberton, S. Brotherstone, F. Guinness, and T. Clutton-Brock. 2002. Antler size in red deer: heritability and selection but no evolution. Evolution 56:1683–1695.

Martin, J. G. A., and D. Réale. 2008. Temperament, risk assessment and habituation to novelty in eastern chipmunks, *Tamias striatus*. Animal Behaviour 75:309–318.

Merilä, J., L. E. B. Kruuk, and B. C. Sheldon. 2001*a*. Cryptic evolution in a wild bird population. Nature 412:76–79.

———. 2001*b*. Natural selection on the genetical component of variance in body condition in a wild bird population. Journal of Evolutionary Biology 14:918–929.

Moyes, K., B. J. T. Morgan, A. Morris, S. J. Morris, T. H. Clutton-Brock, and T. Coulson. 2009. Exploring individual quality in a wild population of red deer. Journal of Animal Ecology 78:406–413.

Mrode, R. A. 1996. Linear models for the prediction of animal breeding values. CAB International, Wallingford.

Nussey, D. H., E. Postma, P. Gienapp, and M. E. Visser. 2005. Selection on heritable phenotypic plasticity in a wild bird population. Science 310:304–306.

O'Hara, R. B., J. M. Cano, O. Ovaskainen, C. Teplitsky, and J. S. Alho. 2008. Bayesian approaches in evolutionary quantitative genetics. Journal of Evolutionary Biology 21:949–957.

Perrins, C. M. 1979. British tits. Collins, Glasgow.

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

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

Quaas, R. L. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science 71:1338–1345.

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., A. G. McAdam, S. Boutin, and D. Berteaux. 2003. Genetic and plastic responses of a northern mammal to climate change. Proceedings of the Royal Society B: Biological Sciences 270:591–596.

Robinson, G. K. 1986. Group effects and computing strategies for models for estimating breeding values. Journal of Dairy Science 69:3106–3111.

———. 1991. That BLUP is a good thing: the estimation of random effects. Statistical Science 6:15–32.

Rubin, D. B. 1976. Inference and missing data. Biometrika 63:581–590.

Sheldon, B. C., L. E. B. Kruuk, and J. Merilä. 2003. Natural selection and inheritance of breeding time and clutch size in the collared flycatcher. Evolution 57:406–420.

Sinervo, B., and A. G. McAdam. 2008. Maturational costs of reproduction due to clutch size and ontogenetic conflict as revealed in the invisible fraction. Proceedings of the Royal Society B: Biological Sciences 275:629–638.

Sorensen, D. A., and B. W. Kennedy. 1983. The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. Theoretical and Applied Genetics 66:217–220.

———. 1984. Estimation of response to selection using least-squares and mixed model methodology. Journal of Animal Science 58: 1097–1106.

Sorensen, D. A., C. S. Wang, J. Jensen, and D. Gianola. 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. Genetics Selection Evolution 26:333–360.

Stinchcombe, J. R., C. Weinig, K. D. Heath, M. T. Brock, and J. Schmitt. 2009. Polymorphic genes of major effect: consequences for variation, selection, and evolution in *Arabidopsis thaliana*. Genetics 182:911–922.

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 USA 105: 13492–13496.

Walsh, B., and M. Lynch. 2009. Evolution and selection of quantitative traits. http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html.

Westell, R. A., R. L. Quaas, and L. D. Vanvleck. 1988. Genetic groups in an animal-model. Journal of Dairy Science 71:1310–1318.

Wilson, A. J., J. M. Pemberton, J. G. Pilkington, T. H. Clutton-Brock, D. W. Coltman, and L. E. B. Kruuk. 2007. Quantitative genetics of growth and cryptic evolution of body size in an island population. Evolutionary Ecology 21:337–356.

Associate Editor and Editor: Ruth G. Shaw

Appearance of the prong-horn antelope in August, the horns being perfect. From The Prong-Horn Antelope by W. J. Hays (*American Naturalist*, 1868, 3:131–133).