 DOI: 10.1086/652469

# Appendix from J. W. McGlothlin et al., "Natural Selection on Testosterone Production in a Wild Songbird Population"

## (Am. Nat., vol. 175, no. 6, p. 687)

## Supplementary Methods and Results

### Sampling for Gonadotropin-Releasing Hormone (GnRH) Challenges

To control for the idiosyncrasies of capture and to obtain robust individual estimates of testosterone production, we attempted to obtain from individual birds four samples each year, collected at four sampling stages across the breeding season (Jawor et al. 2006). We attempted to obtain two samples during early breeding (April 21–May 16) by catching birds at random in baited mist nets and traps. The first GnRH challenge was administered upon each bird's first capture (2003: April 28–May 16, $n = 53$; 2004: April 21–May 11, $n = 46$; combined $n = 99$), and the second was administered after waiting 7–21 days (mean, 10.4 days; 2003: May 6–May 16, $n = 26$; 2004: May 1–May 15, $n = 11$; combined $n = 37$). During early breeding, many birds were beginning to nest, but the exact stage of reproduction was unknown for most of them (dates of first egg were April 26 in 2003 and April 25 in 2004). Some birds were captured and given a GnRH challenge while feeding 6–7-day-old nestlings (2003: May 25–June 29, $n = 14$; 2004: May 20–July 20, $n = 14$; combined $n = 28$). Captures during this stage were made by placing a mist net at the nest. A final set of birds was captured at the end of the breeding season, but before the onset of molt, using baited mist nets (2003: July 15–August 6, $n = 7$; 2004: July 20–August 5, $n = 9$; combined $n = 16$). All sampling periods occurred after the typical early-breeding season testosterone peak (March 26–April 14; Ketterson and Nolan 1992). Overall, five individuals were challenged a total of five times, six were challenged four times, 12 were challenged three times, 28 were challenged two times, and 39 were challenged once. Twenty-three individuals received challenges in both 2003 and 2004, 35 were challenged in 2003 only, and 32 were challenged in 2004 only.

### Analysis of Handling Time

Because our protocol involved taking birds to a central laboratory to perform GnRH challenges, some birds experienced extended handling times, which may have been stressful. Evidence indicates that stressors and the endocrine pathways associated with stress may modulate activity of the hypothalamo-pituitary-gonadal (HPG) axis, but the mechanisms by which this occurs are still unclear. Most work on this question has been conducted in mammals. Studies reviewed by Brann and Mahesh (1991) and Tilbrook et al. (2000) suggest that chronic stress almost always inhibits gonadotropin secretion and reproduction but that the effects of acute stressors are less clear. In female rats, for example, the effect of the adrenocortical stress axis on the HPG axis ranges from inhibitory to stimulatory, depending on the presence of estrogen (Brann and Mahesh 1991). Effects of acute stress or the adrenocortical stress axis on male mammals also vary from inhibitory to stimulatory (Tilbrook et al. 2000).

In our experiment, extended handling times occurred either because of the distance required to transport birds to the laboratory or because many birds were captured at a time. Jawor et al. (2006) reported that increased handling time led to marginally significant decreases in both initial testosterone and GnRH-induced testosterone increase. Because this marginal handling time effect may have implications for assessing natural variation in testosterone in response to GnRH and, hence, selection, here we analyze this effect further.

To determine where the possible effect of handling time arose, we altered the linear mixed model presented in Jawor et al. (2006). Handling times were organized into bins and analyzed as a categorical effect. Handling times were binned into 2–15 min ($n = 14$), 16–30 min ($n = 51$), 31–45 min ($n = 38$), 46–60 min ($n = 27$), 61–91 min ($n = 32$), 91–120 min ($n = 13$), and 121–217 min ($n = 5$). This categorical effect was found to be significant for both initial testosterone ($P = .02$) and postchallenge testosterone (controlling for initial

testosterone, $P = .04$). Post hoc comparisons (Fisher's least significant difference) revealed that for initial testosterone, the only significant differences were between the 16–30-min category, which showed the highest prechallenge levels (fig. A1), and all the other categories except 2–15 min, from which the 16–30-min category differed marginally significantly ($P = .08$). For postchallenge testosterone, the only significant differences were between the 2–15-min category, which showed the highest postchallenge levels (fig. A1), and all the other categories except 15–30 min, which was marginally significant ($P = .08$). Analyzing our data this way did not alter any of the results reported by Jawor et al. (2006), including the significant repeatability of GnRH-induced testosterone increases.

Despite the significant differences among handling time categories, it is clear that GnRH challenges effectively induced significant increases in testosterone, regardless of handling times (fig. A1). GnRH challenge responses were strongest in the 2–15-min category, suggesting that studies wishing to measure the maximum GnRH challenge response should obtain initial samples within 15 min, ideally immediately upon capture. This may explain why, in a previous study, mean GnRH-induced testosterone was somewhat lower than testosterone measured immediately after a simulated territorial intrusion, despite a strong correlation between the two (McGlothlin et al. 2008). In that study, samples after a territorial intrusion were obtained immediately after capture, whereas GnRH challenges required some handling time. However, figure A1 also indicates that handling times up to 3 h and 37 min do not decrease testosterone response to GnRH substantially, at least in this species.

In this study, it would not have been possible to obtain the large sample sizes necessary for measuring selection while minimizing handling time. Therefore, we statistically controlled for handling time when summarizing testosterone measurements (see below). In order to determine whether our results were robust to extreme handling times, we examined a reduced data set consisting of only samples that were obtained in an hour or less. (This analysis could not be performed for samples collected within 15 min because of the very small sample size.) When using the same models reported in Jawor et al. (2006), the effect of handling time on testosterone was drastically reduced in the reduced data set (initial testosterone, $P = .46$; postchallenge testosterone, $P = .41$). GnRH-induced testosterone increases were still repeatable in this model, although the magnitude and significance level was decreased ($r = 0.28$, $P = .06$) compared with the results of Jawor et al. (2006). This change was likely due to the reduced number of repeated measurements available in this data set and thus the reduced power to detect repeatability. Least squares means (see below) were highly correlated between the full and reduced data sets (initial testosterone, $r_{90} = 0.95$; GnRH-induced testosterone increase, $r_{90} = 0.96$). We tested our largest fitness data set (survival) with this reduced data set and found that the selection gradients did not differ qualitatively from those reported in table 1.

## Testosterone Assays

For the analysis of samples, approximately 2,000 cpm of tritiated testosterone was added to allow calculation of recoveries after two extractions with diethyl ether. Extracts were resuspended in 50 $\mu$L of ethanol and diluted to 350 $\mu$L with assay buffer from the kit. From each reconstituted sample, 100 $\mu$L was used to determine recoveries, and duplicate 100-$\mu$L quantities were used in the enzyme immunoassay. Testosterone concentrations were determined with a four-parameter logistic curve-fitting program (Microplate Manager; BioRad Laboratories) and corrected for incomplete recoveries. Intraplate coefficients of variation ranged from 1%–19% (mean, 9%), and interplate variation was 20%. We corrected for interplate variation by multiplying each measurement by the grand mean of assay standards across all plates within the data set and dividing by the plate mean of these standards. More assay details, as well as mean testosterone levels, are reported by Jawor et al. (2006).

To summarize repeated testosterone measurements into a single measurement for each individual, as well as to correct for other variables that may have affected our measurements (Jawor et al. 2006), we fitted general linear models that included individual and sampling stage as categorical predictors and natural-log handling time and mass as continuous predictors. From these models, we estimated the least squares mean for each individual. Mean initial testosterone (ng $mL^{-1}$; natural-log transformed) and GnRH-induced testosterone increase (natural log of postchallenge testosterone minus natural log of initial testosterone) were estimated in separate models; these means were calculated separately for each year (2003, $n = 58$; 2004, $n = 55$). We chose to use GnRH-induced increase rather than absolute postchallenge testosterone in order to reduce collinearity of variables in the selection analyses; postchallenge and initial testosterone levels are positively correlated (Jawor et al. 2006). Selection analyses using postchallenge testosterone (not shown) generated results that were qualitatively similar.

## Paternity Analysis

Total genomic DNA was extracted using standard phenol-chloroform protocols. All individuals were genotyped at five highly polymorphic dinucleotide–repeat microsatellite loci that were assumed to be neutral in juncos: GF01b and GF05 (Petren 1998), designed for medium ground finches (*Geospiza fortis*); Dpu01 and Dpu16 (Dawson et al. 1997), designed for yellow warblers (*Dendroica petechia*); and Cu$\mu$28 (Gibbs et al. 1999), designed for Swainson's thrushes (*Catharus ustulatus*). Combined, these loci had a probability of paternal exclusion of 99.8%. Loci were amplified with fluorescently labeled primers (Operon; Applied Biosystems) in multiplexed polymerase chain reactions using Qiagen multiplex kits and manufacturer-supplied protocols in 10-$\mu$L reactions. The resulting product was then diluted (1 : 20) and mixed with a molecular size standard (GeneScan-500 LIZ, Applied Biosystems), and fragment size was measured with the Applied Biosystems 3730 XL DNA analyzer and GeneMapper 4.0 software. We attempted to genotype each individual at least twice to confirm genotype assignment.

Paternity was assigned using Cervus 3.0 (Marshall et al. 1998; Kalinowski et al. 2007). Candidate sires for each offspring were chosen on the basis of the proximity of the male's capture or territory location to the nest (mean 9.3 males per nestling). The female observed at the nest was assumed to be the genetic dam (intraspecific brood parasitism is extremely rare to absent in our population; Ketterson et al. 1997). The preliminary simulation parameters were based on our data (nine candidate fathers, 99.7% fathers sampled, 94.3% loci typed) and assumed a low mistyping error (1% loci mistyped). A nestling was assigned to a male if Cervus identified him as the sire with 95% confidence (211 nestlings). In addition, the putative (social) father was assigned if Cervus identified him as the sire with lower confidence but still greater confidence than any of the alternative sires under consideration (23 nestlings). For the 26 nestlings for which a male other than the social father was identified as the sire, but with low confidence, we treated them as extrapair young but did not assign them to a male. Six nestlings could not be assigned to a father at all because they or their putative sire were genotyped at fewer than two loci. In all, 95 (32.2%) of 259 nestlings were classified as extrapair young. A more conservative estimate of the rate of extrapair fertilizations that ignores the 26 offspring classified as extrapair young with low confidence is 69 extrapair young (29.6%) of 233 total nestlings. Both estimates are slightly above our population average (28%; N. M. Gerlach and E. D. Ketterson, unpublished data).

## Fitness Components

Estimates of annual survival were based on recapturing or sighting a male in a breeding season (April–August) directly following the one during which we measured response to GnRH. Census methods remained consistent from year to year (Reed et al. 2006). If a male was caught or observed at any time in 2004, he was assigned a survival of 1 for 2003; otherwise, he was assigned a survival of 0. The population was censused in 2005 to estimate 2004 survival, and again in 2006 to check these estimates (two males were observed in 2005 but not in 2004, and two males were observed in 2006 but not in 2005). Recapture probability is high in our population ($0.88 \pm 0.03$; W. Reed and M. Clark, personal communication based on Reed et al. 2006), and males are highly philopatric between breeding seasons (Nolan et al. 2002), suggesting that our estimates of survival closely approximated actual survival. One male given a GnRH challenge was removed from the survival data set because it was killed by a predator while it was caught in a mist net. We obtained 112 survival estimates from 89 unique males (2003, $n = 57$; 2004, $n = 55$; 23 were measured in both years).

Annual offspring production was based on counts of 6-day-old nestlings assigned via paternity analysis (2003, $n = 28$ males; 2004, $n = 20$ males; eight in both years). To explore different pathways leading to total reproductive success, we partitioned total annual offspring production in two ways. First, we separated number of mates (2003, $n = 28$ males; 2004, $n = 28$ males; eight in both years) from offspring per mate (2003, $n = 25$; 2004, $n = 20$; six in both years). These two components of fitness have the advantage of being multiplicative, allowing addition of selection gradients (Arnold and Wade 1984; Wade and Kalisz 1989). Second, we separated within-pair offspring from extrapair offspring (for both: 2003, $n = 28$; 2004, $n = 20$; eight in both years). This partitioning does not generate additive selection gradients, but it does allow us to examine a potential trade-off between the two fitness components (Webster et al. 1995). Sample sizes differ for fitness components because males were excluded from certain analyses due to incomplete information. Specifically, males were excluded from the offspring number data set for a given year if they were the social father of a nest for which paternity could not be assigned due to missing genotypes of either the male or the nestlings (three males) or if they were

the social father of an additional nest known to produce 6-day-old nestlings from which blood samples were not collected (nine males). However, males from the latter group were not excluded from the analysis of mate number if they were found to sire at least one of their social partner's offspring (eight males). The fecundity data set consisted of all males in the offspring data set minus four males that sired no offspring at all. These four males were counted as having zero mating success and unknown offspring per mate. Therefore, the analysis of fecundity did not include males that sired no offspring.

To explore the relationships among different fitness components, we calculated Pearson correlation coefficients for each pair of fitness components and tested the significance of the relationship using generalized linear mixed models due to nonnormality of fitness components and repeated measures. These models controlled for year and age. Because total and extrapair offspring production showed the largest differences between years, we preferentially used them as the dependent variables in the models. Models using offspring or mate counts used a Poisson error structure, and offspring per mate took a normal error structure.

Annual survival was positively correlated with offspring production (table A1). Correlations between annual survival and other reproductive components were not statistically significant but were uniformly positive. Comparing fitness components from the two methods of partitioning reproduction shows that, unsurprisingly, within-pair success was strongly positively correlated with number of offspring per mate, whereas extrapair success was strongly positively correlated with number of mates. Within-pair success was also positively correlated with number of mates, but it showed no significant relationship to extrapair success. Total offspring production appeared to be affected approximately equally by mating success and offspring per mate and more strongly by within-pair than extrapair success.

## Estimation of Total Annual Selection

For directional selection gradients, we used the method of Wade and Kalisz (1989), which corrects for changes in the phenotypic (co)variance matrix, **P**. Total directional selection gradients were calculated using the formula

$$\boldsymbol{\beta}_{\text{total}} = \boldsymbol{\beta}_{\text{surv}} + \mathbf{P}_{\text{surv}}^{-1}\mathbf{P}_{\text{repr}}\boldsymbol{\beta}_{\text{repr}},$$

where $\boldsymbol{\beta}$ represents column vectors of selection gradients, and $\mathbf{P}_{\text{surv}}$ and $\mathbf{P}_{\text{repr}}$ are square matrices with diagonal components equal to 1 and off-diagonal components equal to the covariance between initial testosterone and GnRH-induced testosterone increase ($-0.176$ for survival, $-0.100$ for reproduction). Nonlinear selection gradients cannot be summed in the same way because variance in directional selection contributes to the overall strength of nonlinear selection (McGlothlin 2010). To add nonlinear selection gradients, we used the formula

$$\boldsymbol{\gamma}_{\text{total}} = \boldsymbol{\gamma}_{\text{surv}} + \mathbf{P}_{\text{surv}}^{-1}\mathbf{P}_{\text{repr}}\boldsymbol{\gamma}_{\text{repr}}\mathbf{P}_{\text{repr}}\mathbf{P}_{\text{surv}}^{-1} + \boldsymbol{\beta}_{\text{surv}}\boldsymbol{\beta}_{\text{repr}}^{\text{T}}\mathbf{P}_{\text{repr}}\mathbf{P}_{\text{surv}}^{-1} + \mathbf{P}_{\text{surv}}^{-1}\mathbf{P}_{\text{repr}}\boldsymbol{\beta}_{\text{repr}}\boldsymbol{\beta}_{\text{surv}}^{\text{T}},$$

where $\boldsymbol{\gamma}$ is the square matrix of nonlinear selection gradients and "T" denotes matrix transposition (McGlothlin 2010).

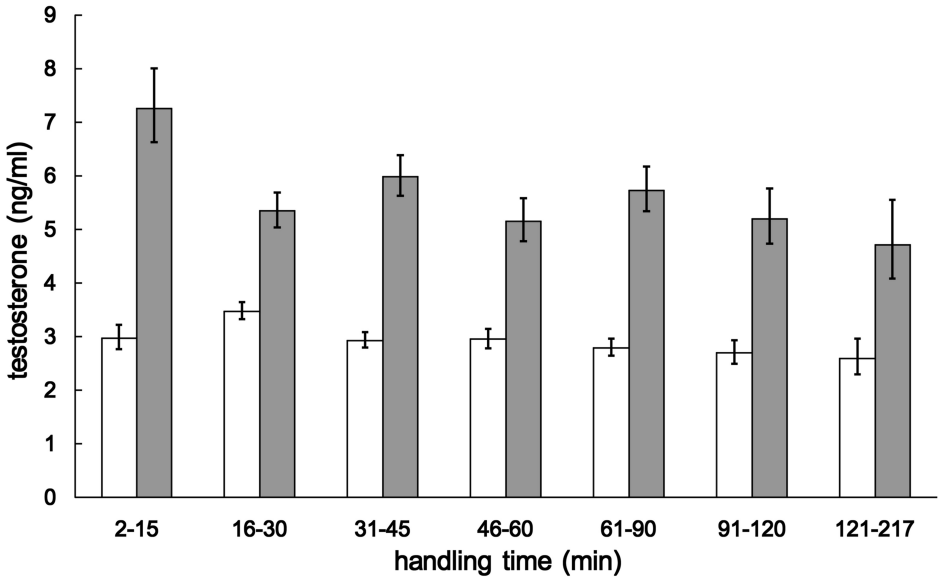


**Figure A1:** Effects of handling time on initial (*white bars*) and post-GnRH-challenge (*shaded bars*) testosterone. Plotted values are predicted means for each category based on a linear mixed model including stage, year, a stage × year interaction, age (years), and mass (g; Jawor et al. 2006). Handling time is treated as a categorical effect. Means and standard errors are back transformed (Jawor et al. 2006).

**Table A1**
Correlations between pairs of fitness components

| Component 1 | Component 2 | *r* | *F* | df | *P* |
|---|---|---|---|---|---|
| **No. offspring** | **Survival** | **.257** | **4.04** | **1, 44** | **.05** |
| **No. offspring** | **No. mates** | **.609** | **19.7** | **1, 24.2** | **<.001** |
| **No. offspring** | **Offspring per mate** | **.746** | **38.7** | **1, 39** | **<.001** |
| **No. offspring** | **Within-pair offspring** | **.885** | **60.3** | **1, 42** | **<.001** |
| **No. offspring** | **Extrapair offspring** | **.454** | **9.41** | **1, 35.6** | **.004** |
| Within-pair offspring | Survival | .214 | 2.16 | 1, 44 | .15 |
| **Within-pair offspring** | **No. mates** | **.307** | **3.94** | **1, 38** | **.05** |
| **Within-pair offspring** | **Offspring per mate** | **.873** | **50.8** | **1, 39** | **<.001** |
| Within-pair offspring | Extrapair offspring | −.014 | .03 | 1, 44 | .86 |
| Extrapair offspring | Survival | .143 | 1.61 | 1, 44 | .21 |
| **Extrapair offspring** | **No. mates** | **.718** | **22.6** | **1, 15.4** | **<.001** |
| Extrapair offspring | Offspring per mate | −.090 | .29 | 1, 41 | .60 |
| No. mates | Survival | .202 | 1.26 | 1, 50 | .22 |
| Offspring per mate | Survival | .148 | .82 | 1, 41 | .37 |

**Note:** Significance testing was performed using generalized linear mixed models with normal (offspring per mate) or Poisson (all other fitness components) error structure. Component 1 was the dependent variable; component 2 was a covariate. Models also included a fixed effect of year and age (in years) as a covariate (not shown). Effects with $P \leq .05$ are shown in bolded type.

## Literature Cited Only in the Appendix

Brann, D. W., and V. B. Mahesh. 1991. Role of corticosteroids in female reproduction. FASEB Journal 5:2691–2698.

Dawson, R. J. G., H. L. Gibbs, K. A. Hobson, and S. M. Yezerinac. 1997. Isolation of microsatellite DNA markers from a passerine bird, *Dendroica petechia* (the yellow warbler), and their use in population studies. Heredity 79:506–514.

Gibbs, H. L., L. M. Tabak, and K. Hobson. 1999. Characterization of microsatellite DNA loci for a Neotropical migrant songbird, the Swainson's thrush (*Catharus ustulatus*). Molecular Ecology 8:1551–1552.

Petren, K. 1998. Microsatellite primers from *Geospiza fortis* and cross-species amplification in Darwin's finches. Molecular Ecology 7:1782–1784.

Tilbrook, A. J., A. I. Turner, and I. J. Clarke. 2000. Effects of stress on reproduction in non-rodent mammals: the role of glucocorticoids and sex differences. Reviews of Reproduction 5:105–113.