Calibration of Laboratory Models in Population Genetics

Robert A. Skipper, Jr.
Department of Philosophy
University of Cincinnati

ML 0374
Cincinnati, OH 45221-0374

Email: skippera@email.uc.edu

Forthcoming in *Perspectives on Science*
Running header "Calibration of Laboratory Models in Population Genetics"

Abstract count: 97
Word count: 11,112 (including abstract, notes, references, and figures)

Abstract

This paper explores the calibration of laboratory models in population genetics as an experimental strategy for justifying experimental results and claims based upon them following Franklin (1986, 1990) and Rudge (1996, 1998). The analysis provided undermines Coyne et al.'s (1997) critique of Wade and Goodnight's (1991) experimental study of Wright's (1931, 1932) *Shifting Balance Theory*. The essay concludes by further demonstrating how this analysis bears on Diamond's (1986) claims regarding the weakness of laboratory experiments as evidence, and further how the calibration strategy fits within Lloyd's (1987, 1998) account of the confirmation of ecological and evolutionary models.

Acknowledgements

Thanks to Lindley Darden, Mike Dietrich, Dave Rudge, and an anonymous reviewer for *Perspectives on Science* for comments that helped improve the paper. Remaining infelicities are my own. Much of the research for this paper was supported by National Science Foundation Grant SBE-9818095; any opinions, findings and conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect those of the National Science Foundation. Thanks to Lisa Lloyd and Mike Wade who helped make the grant research possible.

1. Introduction

This paper explores calibration of laboratory models in population genetics as an experimental strategy for justifying experimental results and claims based upon them. Calibration as such an experimental strategy has been discussed by Franklin (1986, 1990, 1997) and Rudge (1996, 1998). In those works, the intuition is that scientists have greater reason to believe in the validity of the results of an experimental apparatus when the apparatus can be shown to reproduce known phenomena. This paper explicates a type of calibration that neither Franklin nor Rudge consider in any detail, but which fits within their frameworks. I argue that the apparatuses of laboratory models in population genetics are calibrated to parameters describing natural populations in order to provide scientists greater reason to believe that artificial selection experiments done in the laboratory correspond to natural selection processes in natural populations. I further demonstrate how my analysis of calibration bears on claims about the weakness of laboratory experimental evidence made by Diamond (1986) and, further, how the calibration strategy fits within Lloyd's (1987, 1988) account of the confirmation of ecological and evolutionary models. Ultimately, my view is that calibration of laboratory models is a key experimental strategy (following Darden's 1991 conception of a strategy) for justifying claims about the relationship between laboratory experiments and natural processes.[1]

The paper is driven by a case study in contemporary population genetics, somewhat familiar to philosophers of biology, i.e., Wade and Goodnight's (1991) experimental study of Wright's (1931, 1932) *Shifting Balance Theory*. The calibration strategy under scrutiny in this paper is manifest particularly in defending Wade and Goodnight's (1991) results from recent

[1] Darden (1991, p. 20) defines "strategy" as a hypothetical or advisory method, procedure, practice, or principle (used in cycles of discovery, evaluation, and revision).

criticism by Coyne, Barton, and Turelli (1997). So, the case study, in a narrow sense, is played out as a defense I explore of Wade and Goodnight against Coyne, Barton, and Turelli's criticism, and in a broad sense played out as my articulation of a key epistemological feature of the calibration of laboratory models in population genetics. I begin in section (2) with a brief review of Franklin's and Rudge's explications of calibration. In section (3), I present in some detail Wade and Goodnight's experimental study of Wright's *Shifting Balance Theory*. In section (4), I present and respond to Coyne, Barton, and Turelli's critique of Wade and Goodnight, highlighting calibration as key to the response. In section (5), I discuss my analysis of calibration in a more philosophical context. Section (6) is reserved for concluding remarks.

2. Franklin and Rudge on Calibration

Via historical case studies from high energy physics, Franklin (1986, 1990) introduced a set of experimental strategies he packaged as an epistemology of experiment. Franklin's broad aim was to demonstrate that claims based on experimental results are warranted, *pace* the then key social constructivists of science, such as Collins (1985). Franklin offered nine experimental strategies in his earlier work, of which one was the strategy of experimental checks and calibration.[2] To demonstrate the reasonableness of the strategies, Franklin justifies each using the Bayesian epistemology. Franklin (2002) has continued and expanded his work on epistemology of experiment. Here, I constrain my discussion to calibration.

More important for my purposes is Rudge's (1996, 1998) application of Franklin's epistemology of experiment into the domain of evolutionary biology. Only a small handful of philosophers of biology have critically examined the nature and epistemology of experiment in

[2] Franklin's (1986, 1990) strategies include the following: experimental checks and calibration, reproducing artifacts known to be present, intervention, independent confirmation using different experiments, elimination of error, using the results to argue for their validity, using an independently well-corroborated theory of the phenomena to explain the results, using an apparatus based on a well-corroborated theory, using statistical arguments.

evolutionary biology. Notable work has been done by Brandon (1994), Diamond (1986), Dietrich (1996), and Griesemer and Wade (1988). Rudge's work is probably the most systematic, because he aims to provide, as Franklin did for high energy physics, a description of the character of experiment in evolutionary biology as well as its epistemology. Rudge (1996, 1998) takes Franklin's epistemology and critically examines it against a handful of detailed historical case studies, most notably Kettlewell's studies of industrial melanism in the peppered moth, *Biston betularia* (Rudge 1998). Rudge found that Franklin's experimental strategies were used in the Kettlewell work and, further, that additional, Franklin-like strategies could be drawn from that work.[3] Further, like Franklin, Rudge "epistemologizes" his experimental strategies using the Bayesian epistemology.

Recently, Rudge (2001) has compared his Bayesian analysis of the Kettlewell work with an analysis of that work using Mayo's (1996) error statistical epistemology of experiment. Rudge now thinks that Mayo's epistemology of experiment better captures the epistemology of the Kettlewell experimental work. But the issue exercising Rudge here is one over the Bayesian and error statistical approach to understanding the growth of knowledge, and not whether the experimental strategies identified in Rudge (1996, 1998) apply. At any rate, I want to avoid the debate over the relative merits of Bayesian and error statistical justifications of experimental strategies, but nevertheless attach calibration as an experimental strategy to Lloyd's (1987, 1988) work on confirmation of ecological and evolutionary models.

According to Rudge (1998, pp. 346-347), the basic intuition behind assessing experimental checks and calibration is that biologists have greater reason to believe in the validity of the results of an experimental apparatus when the apparatus can be shown to reproduce known phenomena. In ecological genetics, e.g., field studies (i.e., observational or

[3] Rudge (1998, p. 352) mentions the use of controls as an additional strategy.

experimental studies done in naturally occurring populations) often make use of certain kinds of traps. The traps are used to capture specific organisms for marking, release, and recapture. The data is used for such things as estimates of gene frequency fluctuation. Rudge says that assuring the traps are in proper working order is part and parcel of experimental checks and calibration (Rudge 1998, p. 348). Further checks and calibrations, related to the ecological genetics example, might include a review of literature to determine whether a specific kind of trap is appropriate for the kind of study one is doing (Rudge 1998, p. 348). Rudge is echoing Franklin: If an apparatus reproduces known phenomena, then a scientist is justified in her belief that the apparatus is reliable. And if the apparatus is reliable, then the results produced by it during an experiment are warranted.

Rudge stretches Franklin's conception of "apparatus" (following Kohler 1994) to make his point, above. For Rudge, the apparatus includes microscopes, vials, dyes, nets, incubators, and even organisms, and populations of organisms (cf. Griesemer and Wade 1988, p. 71). The calibration of each of these parts of the apparatus are, as Rudge argues, important for assessing whether the experimental apparatus is working properly. After all, in many experiments in evolutionary biology, the evolutionary processes causing change in gene frequencies are under scrutiny. And the apparatuses used in those studies include traps, cages, pipettes, vials, incubators, and populations of organisms. But notice that Rudge's point above is made by emphasizing the specifically instrumental parts of the apparatus. He only briefly mentions calibration of the actual *populations of organisms* that are part of the apparatus (Rudge 1996, p. 266, 2000).

Consider laboratory models, where a laboratory model is an implementation of the materials, apparatus, and maintenance procedures for the apparatus structured by an

experimental design (e.g., Griesemer and Wade 1988, Lloyd 1987, 1988). Calibration of the relevant populations is important for assessing whether the experimental apparatus is working properly as Rudge would argue. But, as Rudge does not argue, and which is equally if not more important, calibration of the relevant populations is crucial to justifying claims based on experimental results that the processes at work in laboratory populations could work in nature (cf. Griesemer and Wade 1988, pp. 71, 81, 84). The idea is this: Claims based on artificial selection experiments done in the laboratory can be justifiably extrapolated to claims about natural selection operating in natural populations if the laboratory model, particularly features of the apparatus as articulated here –the populations of organisms– are matched, or tuned, or *calibrated*, to relevant features of natural populations. The next two sections articulate this point by contextualizing it within a scientific case.

3. An Experimental Study of *The Shifting Balance Theory*

In 1991, Wade and Goodnight published, in *Science*, the results of what they took to be an experimental (laboratory) test of Wright's *Shifting Balance Theory* (SBT) (Wade and Goodnight 1991). At the time, and it seems to me even now, Wade and Goodnight's study represents the most important contemporary experimental study of the SBT in existence. Indeed, it is the only contemporary *comprehensive* laboratory study of the SBT.

Wright developed his SBT in his classic (1931) and (1932) papers and it is fully articulated in his four volume *magnum opus*, *Evolution and the Genetics of Populations* (1968, 1969, 1977, 1978). What Wright was looking for in 1931 and 1932 were the ideal conditions for evolution to occur given specific assumptions about the relationship between Mendelian heredity and the adaptive value of gene complexes (Wright 1931, p. 158, 1932, p. 163). Wright in particular emphasized what he thought of as the universality of gene interaction, or epistasis, and

its role in the evolutionary process. The consequences of Wright's emphasis on epistasis led him to the view that the "adaptive landscape" of the possible field of gene combinations in a population of organisms was "hilly" (Wright 1932). That is, for Wright, the evolutionary consequences of epistasis meant that genes adaptive in one combination would be maladaptive in another. As such, the surface of the adaptive landscape would be made up of adaptive "peaks" and "valleys." Wright's problem, i.e., the problem of "peak shifts," was to articulate the evolutionary process by which a population could make its way to the highest adaptive peak in the fastest time.

Wright argued that the ideal conditions to solve his problem meant that evolution occurred in structured, smallish populations with specific mutation and migration rates. Wright's SBT captures these ideal conditions in what he called the shifting balance process. A rough summary of the SBT, sufficient for my purposes, goes as follows: Evolution proceeds via a shifting balance process through three phases: **Phase I** – *Random genetic drift causes subpopulations semi-isolated within the global population, i.e., demes, to lose fitness*; **Phase II** – *Mass selection on complex genetic interaction systems raises the fitness of those subpopulations*; **Phase III** – *Interdemic selection then raises the fitness of the large or global population*. [4]

By 1991, biologists did not know whether the conditions of Wright's phase III (in his theoretical model) were met by populations in nature (Wade and Goodnight 1991, p. 1015). Indeed, only mathematical investigations, and importantly that by Crow et al. (1990), had been done. The key issue concerning phase III of the SBT is how population structured selection raises the fitness of the global population. More specifically: Population structured selection is

[4] For more detailed, historical presentations of the development of Wright's SBT, see Gayon (1998), Hodge (1992), Provine (1971, 1986), Skipper (2000, 2002).

selection driven by migration. Members of local populations that are especially fit migrate to another local population. There, they mate with members of the new-found local population. The traits of the migrants are then spread through the new-found local population. This causes the fitness of the global population to be raised via interdemic selection through differential dispersion (Wade and Goodnight 1991, p. 1015).

Figure 1 below is Wade and Goodnight's graphical depiction of their experimental design for their experimental test, as they call it, of the third phase of Wright's SBT. The idea behind the experiment is as follows. Wade and Goodnight argue that if they can show that interdemic fitness can increase in response to selection driven by differential dispersion of migrants, then they have shown that the third phase of Wright's SBT can occur. If features of the laboratory models can further be calibrated so that there is some identifiable correspondence to natural populations, then Wade and Goodnight believe that there is evidence that Wright's third phase is likely to occur in nature. Further, because the third phase of Wright's SBT explicitly involves features of Wright's shifting balance process that are endemic to its other two phases (e.g., random genetic drift) and more generally its underlying assumption of the role of gene interaction in evolution, then Wade and Goodnight believe that their experiment is a comprehensive test of Wright's SBT.

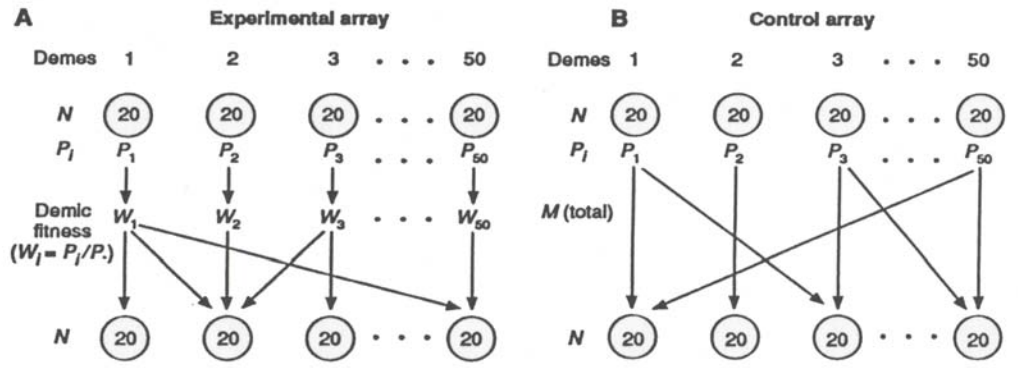

Fig. 1. (**A**) Schematic diagram of the laboratory model of Wright's shifting balance process: N is the number of breeding adults used to establish each deme at the start of each generation and P_i, productivity, is the number of adult progeny produced by N parents in 60 days; demic fitness, W_i, is equal to P_i divided by the average productivity, $P_.$. (**B**) Schematic diagram of a control treatment: the number of migrants, M *(total)*, in this array is determined by the number of migrants in the corresponding experimental array.

Fig. 1: Reproduction of Wade and Goodnight's figure depicting their experimental design (Wade and Goodnight 1991, p. 1016). Figure 1A depicts the experimental lines in which selection is imposed. Figure 1B depicts the control lines in which only migration, with no selection, occurs. Note especially the computation of relative fitness of demes.

Wade and Goodnight's experiment consisted of three experimental and three control treatments using laboratory populations of the flour beetle, *Tribolium castaneum*. Each treatment followed either an experimental array or a control array. Each array consisted of 50 demes of 20 breeding adults (N) each. The founders were selected at random from laboratory stock. Fitness of the demes (demic fitness, W) was measured as the census number of the adult offspring produced by the 20 breeding adults in a 60-day period (P, for productivity).

After 60 days, census data were used to assign relative interdemic fitnesses. Demes with larger populations had higher relative fitness than demes with smaller ones. Based on the measures, migrants in the experimental treatments would be dispersed to different demes (this was the imposition of Wright's third phase). In the control group, migration was random, but the

number of total migrants (*M*) was determined by the number of migrants in the corresponding experimental array. Phase III was imposed in the following way: In the experimental array (A), migrants from demes that are high in productivity are placed in demes that are low in productivity. So, there is (artificial) migration from demes of high fitness to demes of low fitness (following Crow et al. 1990). In the control array (B), there is the same amount of migration, but the dispersion is not from demes of high productivity to demes of low productivity. Rather, it is non-patterned.

Figure 2 below illustrates one way of imposing population structured selection driven by differential dispersion. In the experimental array (A), P_1 has the highest relative fitness. So, migrants are taken from it and put into the migrant pool *M*. P_3 has a relative fitness of 1.00, so no migrants are taken. Because the relative fitness of P_2 is less than 1.00, P_2 receives migrants from the migrant pool *M* but contributes none. All demes now have 20 members. Notice that migrants were taken from a deme of high fitness to a deme of low fitness. This is the essence of Wright's third phase. In the control array (B), the migration pattern from $P_1 \ldots P_n$ to the migrant pool *M* and back into $P_1 \ldots P_n$ is random, but the same numbers migrate (which is a function of the experimental array's census).

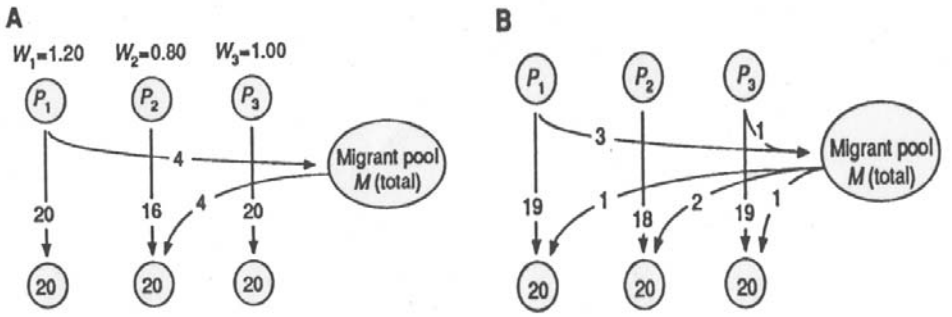

Fig. 2. (**A**) An example of the method used to impose phase (iii) of Wright's process on the experimental array of demes. Because deme 1 has a demic fitness in excess of 1.00, it contributes migrants to the pool. Because deme 2 has a fitness of less than 1.00, it receives migrants from the pool. Deme 3 has a demic fitness exactly equal to 1.00, so it neither contributes nor receives migrants. (**B**) An example of the method used to move migrants at random among demes in the control array. Deme 1 contributes three migrants to the migrant pool but receives only one migrant. Deme 2 receives two migrants although it contributes none to the pool. Deme 3 contributes and receives the same number of migrants. Total migrants in the control migrant pool is four because it is determined by the total number of migrants in (A), the experimental array.

Fig. 2: Reproduction of an example of Wade and Goodnight's (1991, p. 1017) imposition of selection driven by differential dispersion. Fig. 2A is the experimental array; Fig. 2B is the control array.

The experimental (E) treatments were imposed three times and each had their own control (C) treatments. The pairs of experimental (E) and control (C) treatments differed with regard to the frequency with which artificial selection was imposed. In the first (E1-C1) treatment, experimental and control protocols were imposed every generation. In the E2-C2 pair, the protocols were imposed every two generations: In alternating, non-selection generations, 20 adults were selected at random from each deme in the experimental array (E2) of 50 demes and each group of 20 was used to reestablish a deme; the same was done in the C2 control (there was no migration of any kind in the odd E2-C2 generations). In the E3-C3 pair, the protocols were imposed every third generation analogous to the E2-C2 treatment.

Wade and Goodnight varied the imposition of differential and random migration in the three trials in order to investigate the interaction between random genetic drift and (directional)

interdemic selection. Because Wade and Goodnight varied the treatments in the way that they did, they were able to further investigate Wright's shifting balance process as a whole rather than simply the third phase. The results, summarized graphically, are reproduced in Figure 3 below.

The interdemic selection differential in the three experimental treatments is the difference between the mean productivity of the selected parent demes ($\sum_i w_i P_i / 50$) and the mean productivity of the unselected demes (*P.*), divided by the standard deviation of the unselected array (see Figure 1B). The standardized response to interdemic selection is the difference between the means of the experimental and control treatment pairs divided by the standard deviation of the control. The regression of the standardized response on the cumulative selection differential is the realized interdemic heritability.

An additive population genetics model was used to predict results. On the additive model, the expected results are that the three treatments will show that the responses to selection, i.e., the differences in mean phenotype between the progeny generation and the previous generation, are homogeneous (Wade and Goodnight 1991, p. 1017). That is, the slopes of the lines in the E1-C1, E2-C2, E3-C3 ought to all be homogeneous. Such results are expected because on an additive selection model, according to which random genetic drift and directional selection oppose each other, the response to interdemic selection should be proportional to the selection differential, i.e., the difference in mean phenotype between the selected parents and the entire parental population. After all, on the additive assumption random genetic drift acts to increase genetic variation among demes whereas directional interdemic selection should, given the assumption of additivity of genetic variance, act to decrease variation among demes.[5]

[5] The assumption of additivity of genetic variance means effects of alleles on phenotypes are averaged over the variety of genetic combinations in which they occur; an allele is said to act additively if its effect is independent of the other alleles with which it is combined. Notice that this assumption contrasts with Wright's emphasis on the non-additive component of genetic variance, i.e., the statistical effects of the interaction of an allele with the other alleles

Indeed, using an additive model, the expected F_{ST} values (inbreeding coefficients measuring population structure, or the among-deme genetic variation, see Wright 1951a, 1969) can be computed (Wade and Goodnight 1991, p. 1017). Given a ratio of effective population size (N_e) –the number of individuals in a theoretical ideal population having the same magnitude of genetic drift as the actual population– to observed deme size (N) (where $N_e < N$) of 0.90, the expected F_{ST} increases almost linearly by generation: 0.0278 after one generation, 0.0548 after two generations, and 0.0811 after three generations (with no migration). If there are only additive genetic effects on fitness (and if the environment is held constant), then the interdemic selection differential is proportional to F_{ST} . This means that the interdemic selection differential in the E2 treatment should be twice that of the E1 treatment and the E3 treatment three times that of the E1 treatment. The frequency of selection in Wade and Goodnight's experiments is the inverse of the above. So, the smaller selection differential applied every generation in E1 should produce the same response as the selection differential in E2, or three times that in E3.

During the experiment, Wade and Goodnight measured effective deme size and random genetic drift using a single locus (Wade and Goodnight 1991, p. 1018). Observed F_{ST} (in C2 at generation 13) was 0.0288. In the next generation, the F_{ST} was 0.0470. There appears to be an almost linear increase in F_{ST} (as the additive model predicted). Wade and Goodnight further found that the ratio of effective population size to observed deme size was 0.90, agreeing with their estimates from the additive model discussed above. In spite of the agreement between expected and observed F_{ST} and (N_e/N) ratio, Wade and Goodnight did not get the expected results: As is manifest, the slopes of the lines in Figure 3 are heterogeneous, not homogeneous.

---

with which it is combined, such as epistasis and dominance. The role of the additive model in contrast to Wright's emphasis on non-additive genetic variance will become clearer as the discussion proceeds.

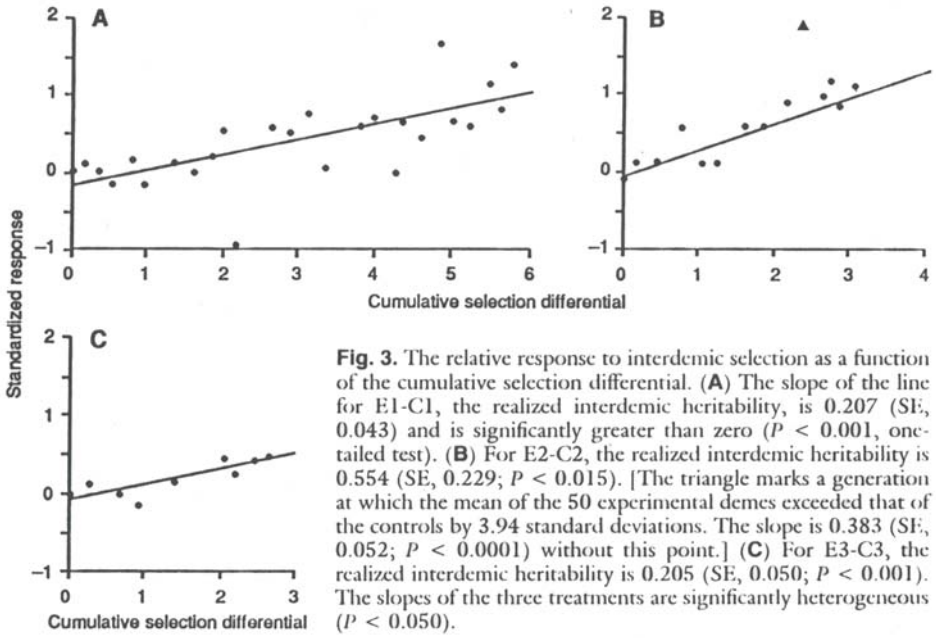

Fig. 3. The relative response to interdemic selection as a function of the cumulative selection differential. (**A**) The slope of the line for E1-C1, the realized interdemic heritability, is 0.207 (SE, 0.043) and is significantly greater than zero ($P < 0.001$, one-tailed test). (**B**) For E2-C2, the realized interdemic heritability is 0.554 (SE, 0.229; $P < 0.015$). [The triangle marks a generation at which the mean of the 50 experimental demes exceeded that of the controls by 3.94 standard deviations. The slope is 0.383 (SE, 0.052; $P < 0.0001$) without this point.] (**C**) For E3-C3, the realized interdemic heritability is 0.205 (SE, 0.050; $P < 0.001$). The slopes of the three treatments are significantly heterogeneous ($P < 0.050$).

Fig. 3: Reproduction of Wade and Goodnight's (1991, p. 1017) table of experimental results. See Wade and Goodnight's caption in the figure.

In Figure 3, the regression of the responses to selection on the cumulative selection differential for all treatments (E1-C1, E2-C2, E3-C3) are shown. This is the realized interdemic heritability. All of the slopes are greater than zero indicating a statistically significant response to the imposed interdemic selection in all treatments. The heterogeneity of the slopes are contrary to the additive model's prediction that because selection is standardized across the X-axis in all treatments the responses should all be equal (Wade and Goodnight 1991, p. 1018): The realized interdemic heritability was two times greater in the E2 (0.383) than in the E1 treatment (0.207) (in spite of the reduced selection frequency). Additionally, the realized interdemic heritability in E1 (0.207) was equivalent to that of E3 (0.205) (again in spite of the reduced frequency of selection in the E3). Finally, even though E2 and E3 were nearly equal in total heritability (*S*)

(2.826 vs. 2.565) and average population size given migration (N_m) (0.99 vs. 0.86) the response to selection was two times larger in the E2.

According to Wade and Goodnight, if there are non-additive gene effects on fitness of the sort Wright emphasized, then random genetic drift can operate to more rapidly produce significant heritable variation among demes in gene combinations. However, Wade and Goodnight did not know the distribution of adaptive peaks owing to co-dependent gene complexes. Moreover, Wade and Goodnight did not know the extent to which interdemic selection by differential dispersion could export the gene combinations to other demes of lower fitness. Wade and Goodnight do cite Crow et al. (1990), who suggest that the rate could be small. Nevertheless, Wade and Goodnight conclude that their experimental design is a laboratory model of Wright's shifting balance process (based on the realized response to their imposition of third phase selection).

Based on their experimental study, Wade and Goodnight (1991, p. 1018) conclude that phase III of Wright's shifting balance process can cause an evolutionary change in mean fitness compared to control populations (non-shifting balance populations) with identical amounts of island model migration. Wade and Goodnight further claim that because their statistical measures of genetic differentiation among demes (i.e., population structure) in their observations match, are tuned to, or are calibrated to a variety of natural populations of animals and plants (e.g., Loveless and Hamrick's 1984 review of population structure field data, Wright's 1978 review of numerous field observations of population structure of various species), Wright's SBT is likely to be an important description of evolution in nature. Wade and Goodnight use the *b* allele in the C2 treatment to measure among-deme genetic variation and found F_{ST} values

between 0.03 and 0.05. Wright (1931, 1932, 1977, 1978) argued that the shifting balance process described by the SBT should affect evolution with this degree of genetic subdivision.

Loveless and Hamrick (1984) and Wright (1978) contain comprehensive reviews of data concerning population size and structure among plants and animals. Loveless and Hamrick pay special attention to the ecological determinants of the genetic structure of populations of plants. They review isozyme data from 163 studies representing 124 plant taxa to examine the associations between ecological factors and the distribution of allozyme variation. Each taxon was classified across 10 ecological traits, breeding system, floral morphology, mode of reproduction, pollination mechanism, seed dispersal, phenology, life cycle, timing of reproduction, successional stage and geographic range. Measures of the genetic structure of demes were reported for 19 species of plants across 29 studies (Loveless and Hamrick 1984, pp. 86-89). Of the 19 species, 13 had G_{ST} values (a generalized measure of F_{ST}) between 0.03 and 0.05, consonant with Wade and Goodnight's laboratory model.

Wright (1978, especially chapters 2, 3, 6, 9) provides a comprehensive review of population structure data from a staggering number of studies. Wright found F_{ST} values between 0.03 and 0.05 in various species of trees, perennials such as *Phlox pilosa*, the desert annual *Linanthus parryae*, as well as varieties of *Drosophila*, the land snail *Cepaea nemoralis*, the rusty lizard, *Sceloporus olivaceu*s, the prairie mouse, *Peromyscus polionotus*, and human groups based on blood group data. As in the Loveless and Hamrick (1984) review, Wright's statistical analyses cut across numerous ecological factors. The aim of Wright's review, the last volume of his four volume *magnum opus*, is a sweeping analysis of variability within and among natural populations. The work is not without its critics (e.g., Coyne et al. 1997), and itself critically

reanalyzes historical work Wright, with others (such as Dobzhansky), had done during the period of the evolutionary synthesis (1930s-1950s).

Ultimately, because the population structure of the population in Wade and Goodnight's laboratory model is calibrated to population structure of a variety of natural populations, and that population structure is ideal for Wright's shifting balance processes to occur, there is reason to think that the shifting balance process as observed in the laboratory corresponds to evolutionary processes as they occur in nature (see also Crow 1991 for a similar, brief, interpretation of Wade and Goodnight 1991). More abstractly, Wade and Goodnight think that because their laboratory system is calibrated to natural systems, the claims made about the laboratory system can be extrapolated to natural systems. In the next section of the paper, the calibration strategy is applied in the context of a recent critique of Wade and Goodnight's (1991) experimental study by Coyne et al. (1997).

4. A Response to Coyne, Barton and Turelli's Critique of Wade and Goodnight

In 1997, Coyne, Barton, and Turelli published what is now a well-known critique of Wright's SBT (Coyne et al. 1997). In that paper, Coyne et al. argue that Wright's SBT describes a relatively insignificant evolutionary process, opting for the mass selection alternative originally described by R. A. Fisher (1930, 1958, 1999). Briefly, Fisher's *Large Population Size Theory* (LPST) embraces assumptions at odds with Wright's SBT. In particular, Fisher emphasized that the ultimate fate of a gene is determined by its average effect with respect to fitness, i.e., that alleles act additively on phenotypes, within large, randomly mating or panmictic populations so that co-dependent gene complexes arise as a by-product of the effects of selection acting on single genes (Wade 1992, p. 38). During the period 1929-1962, Fisher and Wright engaged in

numerous and sometimes acrimonious debates over the evolutionary implications of the differences between their theories (Provine 1986).

Reminiscent of the exchanges between Fisher and Wright, Coyne et al.'s (1997) paper spurred an exchange with Wade and Goodnight over the relative significance of the SBT versus the LPST (Wade and Goodnight 1998, Goodnight and Wade 2000, Coyne et al. 2000). In the evolutionary biology community today, Fisher's LPST probably has the most support, even if Wright's contributions to evolutionary biology are widely hailed. However, Skipper (2002) analyzed the general logic of the exchange between Coyne et al. and Wade and Goodnight and argued in defense of Wade and Goodnight and, therefore, Wright. The critique of Wade and Goodnight's (1991) study raised here is but a highly focused portion of Coyne et al.'s (1997) paper and is discussed here in more detail than in Skipper (2002).

Coyne et al. (1997, p. 663) have argued that Wade and Goodnight's (1991) conclusion is unwarranted because their results are ambiguous. That is, Coyne et al. claim that Wade and Goodnight do not adequately impeach the alternative (Fisherian) hypothesis to Wright's shifting balance hypothesis that mass selection on additive genetic effects accounts for the experimental results of the 1991 study.[6] According to Coyne et al., the root causes of the ambiguity of the results are problems with experimental control and experimental replication of treatments in the same experiment. In what follows, I discuss both problems and then construct a response to them. The crux of my response is that Wade and Goodnight's calibration of a specific feature of

[6] There are two ways to take Coyne et al.'s (1997) critique of Wade and Goodnight (1991). The first way is this: Coyne et al. claim that Wade and Goodnight's experiment does not demonstrate that Wright's SBT produces a faster response to selection than Fisher's LPST. Taking Coyne et al.'s critique this way makes it a straw man, since nowhere do Wade and Goodnight claim that their 1991 results show that Wright's SBT produces a faster response to selection than Fisher's LPST. The second, and more charitable way to take Coyne et al.'s critique, in fact the way I take the critique in my discussion, is that Coyne et al. claim that Wade and Goodnight's experiment does not show that Wright's SBT and not Fisher's LPST accounts for the reported response to selection.

laboratory populations of *Tribolium castaneum* to a variety of natural populations shows that their results are not ambiguous as Coyne et al. (1997) charge.

Coyne et al.'s first complaint, the complaint regarding experimental controls, goes as follows. Wade and Goodnight's results (in Figure 3) are claimed to support Wright's SBT and specifically that the interaction between epistasis (non-additive genetic effects on fitness), population structure, and interdemic selection account for the response to selection reported. However, according to Coyne et al., without a mass selection control group to compare to the shifting balance experimental group, Wade and Goodnight's experiment does not impeach the alternative that mass selection on additive genetic effects accounts for the experimental data.

On Coyne et al.'s view, interdemic selection can occur on any mode of gene action, including the simple case of selection for an additive trait governed by a single locus. Wade and Goodnight's imposition of selection by differential dispersion is based on an experimental design that assumes that demes contribute to the next generation in proportion to the deme's mean phenotypic value. This method is based on the quantitative genetics concept of family-selection, or the view that only the family-groups (demes) with the highest mean fitness can contribute to the next generation (Falconer 1989). Coyne et al. (1997, p. 663) claim that the family-selection design allows that selection can modify almost any trait (of the beetle or any other organism), including traits that are commonly understood as modified by mass selection (like bristle number in *Drosophila*).

In order for Wade and Goodnight to show that the SBT unambiguously accounts for their data, they must, say Coyne et al., compare an experimental shifting balance treatment against a mass selection control treatment. The direct comparison would provide Wade and Goodnight with a way to distinguish between traits affected by interdemic selection and traits affected by

mass selection. But Wade and Goodnight do no such thing –the control arrays are unselected (see above, especially Figure 1). According to Coyne et al., because Wade and Goodnight do not control for mass selection, there is no justifiable way for them to make the claim that they have provided empirical support for Wright's SBT. Only a precise comparison between interdemic selection on the shifting balance process and mass selection will resolve the problem.

Coyne et al.'s second objection to Wade and Goodnight's experimental study of the SBT concerns the ambiguity of experimental results due to lack of replication of the experimental treatments (E1-E3). Wright (1931, 1932) claims that non-additive genetic effects on fitness are foundational to understanding the evolutionary process. Indeed, as we saw at the beginning of section (3) above, Wright includes such effects as core elements of evolution in his SBT. Wade and Goodnight (1991) claim to have supported Wright's inclusion via the results of their shifting balance process laboratory experiment represented in Figure 3. Recall that Wade and Goodnight claimed that because the slopes of the lines in the E1-E3 treatments were heterogeneous that non-additive effects on fitness were present (see especially E2, selection every other generation). The conclusion is that the heterogeneity of the slopes shows that non-additive genetic effects are involved in the observed response to selection. Coyne et al. (1997, p. 663) object that because Wade and Goodnight only used one experimental line for each of the three experimental treatments they have not appropriately removed doubt that their data is not affected by artifacts of the experimental design.

According to Coyne et al., the results in the E2 experimental line could be due to, e.g., some sort of sampling artifact in that line. Had Wade and Goodnight replicated all three lines, there may be reason to claim that the results of the E2 treatments could be used to support a claim that non-additive genetic effects on fitness were an evolutionary factor. However, Coyne et

al. doubt even that proposition: “Even if their [Wade and Goodnight’s 1991] experimental results were seen in multiple lines, however, it would not prove the existence of epistasis: It is also the case that in models of group selection acting on a single locus, the largest response can occur with intermittent selection” (Coyne et al. 1997, p. 663; see the supporting study by Wilson and Colwell 1981). Coyne et al. conclude that Wade and Goodnight’s experimental results are ambiguous: The results do not impeach the alternative mass selection hypothesis.

Coyne et al.’s charge of ambiguity of Wade and Goodnight’s experimental results is weak. Indeed, both arguments that Coyne et al. mount, the first regarding lack of a specific type of experimental control and the second regarding lack of experimental replication, can be shown to be dubious with a single line of reasoning: Coyne et al. focus on two methods for justifying observational and experimental claims, viz., control and replication of treatments in the same experiment. However, Wade and Goodnight’s experimental design does not depend on the experimental methods that Coyne et al. cite and in the way that Coyne et al. argue. Indeed, specific to Coyne et al.’s two complaints, Wade and Goodnight’s experimental design depends upon the following key method, viz., calibration of features of the laboratory model to data from a variety of natural populations in the justification of their experimental claim. Because Coyne et al. fail to consider adequately Wade and Goodnight’s experimental design, their critique is weak.

Consider, first, Coyne et al.’s charge that there is a problem with Wade and Goodnight’s experimental controls. Here, Coyne et al. focus on the lack of a mass selection control group in Wade and Goodnight’s E1-E3 treatments. Indeed, Wade and Goodnight use, as a control, a set of *unselected* treatments as a way to demonstrate a response to interdemic selection. However, as a further and alternative way of justifying their claim that Wright’s third phase is operating, Wade and Goodnight do the following. They argue that a feature of their laboratory apparatus,

specifically F_{ST} measures of population structure, is calibrated to a variety of plant and animal systems in nature (viz., the above cited reviews by Loveless and Hamrick 1984 and Wright 1978). Recall that for the shifting balance process to work, the right kind of population structure is key. Moreover, the precise population structure reported by Wade and Goodnight is indicative of non-additive genetic variance (Wade 1991, 1992, Wade and Goodnight 1998, Wright 1978). For these two reasons, it is crucial for Wade and Goodnight to emphasize the calibration of their F_{ST} measures.

In broad strokes, Wade and Goodnight use the calibration strategy to ground their claim that there is a non-additive basis to their response to interdemic selection and impeach the alternative hypothesis that the genetic basis of the response is additive. More specifically, Wade and Goodnight use the calibration strategy to bolster their claim that there is a non-additive genetic basis to their reported response to interdemic selection in two ways. The first way corresponds to Franklin's (1986, 1990) and Rudge's (1996, 1998) analysis of calibration discussed in section (2) above. That is, Wade and Goodnight tune F_{ST} measures of their laboratory population to natural populations to demonstrate that their apparatus is working properly, thereby warranting their experimental results. The second way that Wade and Goodnight use the calibration strategy is to warrant their broader claim that their experimental results provide reason to think that Wright's shifting balance process is likely to occur in nature. Analysis of how the latter use of calibration works, a central aim of this paper, is spelled out in some detail in section (5) below.

Coyne et al. miss Wade and Goodnight's use of calibration as a strategy for disambiguating between allowable interpretations of their experimental results. Yet, the two ways in which Wade and Goodnight use the calibration strategy plausibly trump Coyne et al.'s

demand for a mass selection control group. After all, via calibration Wade and Goodnight provide independent reasons for thinking that the apparatus is working properly and, as we will see below, they extrapolate their claims about the laboratory model to natural systems.

A general response more illustrative of my line of reasoning against Coyne et al. can be made to their second charge, regarding the lack of replication of experimental lines. The idea is that Coyne et al. are myopically focusing on a specific experimental methodology when what they must do, if they are to accurately depict Wade and Goodnight's experimental design, is criticize the host of experimental methods that Wade and Goodnight in fact use. Differently put, there is a set of experimental methods, in particular calibration, that Wade and Goodnight use as a package to justify their claim (and Wright's) that the shifting balance process is operating. Arguably, the package of methods that Wade and Goodnight use trumps the singular methods on which Coyne et al. focus. Anyway, Coyne et al. do not even consider this point because they do not scrutinize the actual experimental methods that Wade and Goodnight use. It is on these grounds that Coyne et al.'s criticism is weak (if not also unfair).

In the above defense of Wade and Goodnight (1991), I drive a wedge between alternative methods for justifying claims based on experimental results. Coyne et al. complain that Wade and Goodnight fail to exercise proper experimental controls and to replicate their experimental lines. My view is that the calibration strategy is a key part of a package of strategies employed by Wade and Goodnight and that it trumps Coyne et al.'s critique. In the next section of the paper, I show how the calibration strategy provides the justification it does for Wade and Goodnight's claim that their laboratory experimental results show that Wright's shifting balance process is likely to occur in nature. Further, I show how the strategy might fit into an epistemology of experiment reflective of the case study discussed.

5. Calibration as an Experimental Strategy

As we have seen, Franklin and Rudge have discussed calibration as an experimental strategy for warranting claims that an experimental apparatus is working properly, which thereby warrants the results produced by using that apparatus. As I view the strategy, calibration of laboratory models further warrants hypotheses that artificial selection experiments done in the laboratory correspond to the way selection works in natural populations. So, calibration is not merely a strategy for justifying results in the way Franklin and Rudge argue, it is further a strategy that is used to justify results from laboratory experiments by demonstrating that those results can be extrapolated to natural populations.

But how does calibration do what I claim it does? I argue for the following related claims: (a) Calibration of laboratory models increases the justificatory strength of empirical claims based on laboratory experiments; (b) Calibration accomplishes (a) via its connection with specific elements of Lloyd's (1987, 1988) confirmation view. Indeed, making these two points clear articulates the conceptual foundations of the crux of my defense of Wade and Goodnight against Coyne et al.'s critique in section (4) above.

Consider claim (a). A claim of Diamond's (1986) that is of special interest here is that he views the evidence gathered from different sorts of experimental procedures, as he calls them (Diamond 1986, p. 3), as weighed differently in evaluation contexts. Diamond distinguishes between three different kinds of experiments, viz., laboratory experiments, field experiments, and natural experiments. According to Diamond, the evidence gathered from natural experiments is the strongest, and so on down the line to laboratory experiments, which provide the weakest evidence, all else equal.

Diamond claims that laboratory (and field) experiments have an air of artificiality about them that can weaken their justificatory role relative to the "reality" of natural experiments (Diamond 1986, p. 5). The artificiality of laboratory and field experiments consists in, importantly, experimenter intervention in the experimental system. Wade and Goodnight's imposition of phase III of Wright's SBT is an artificial perturbation in their *Tribolium* population. Indeed, such intervention is intentional artificial selection on populations to observe its downstream effects on genetic variation. On Diamond's view, the observation of the same perturbation occurring naturally, i.e., without experimenter interference, would be less artificial, or more realistic (in the sense that phase III would occur *naturally*). As such, the evidence of a natural perturbation would be stronger than evidence from an intentional one, all else equal (Diamond 1986, pp. 5, 9, 13, 19). To be sure, intervention is not the only sort of artificiality in evolutionary biology experiments: Use of, e.g., laboratory populations in highly controlled environments is a further instance. (So, evidence from a laboratory experiment might, e.g., be weaker than evidence from a field experiment given that the field experiment might involve the use of natural populations, but also intervention.)

It is possible to see how the relative strength of evidence given its experimental source (i.e., from the laboratory or the field) arises in the case of Wade and Goodnight's experiment. In Wade and Goodnight's study concern is expressed regarding whether phase III selection is an important evolutionary process in *nature* (Wade and Goodnight 1991, p. 1018). Wade and Goodnight assume that the laboratory imposition of phase III on the flour beetle population is not enough to establish such a conclusion. Indeed, other ways of providing justification of that claim are required in addition to the intervention. Wade and Goodnight used, as we saw, calibration of population structure of their laboratory population of beetles to a variety of *natural* populations.

The idea is to increase the strength of the claim that phase III selection is important in nature as opposed to (merely) achievable in the laboratory by showing that the populations are comparably structured to those in nature. The strength of the empirical claim is increased by increasing what Diamond calls the reality of the laboratory experiment. That is, by tempering the air of artificiality of the laboratory intervention using the calibration strategy, Wade and Goodnight increase the degree of reality Diamond requires for experimental results to gain strength. I consider this idea a bit more fully below.

Given the above, Diamond's claim about the greatest weakness of laboratory experiments is weakened given the calibration strategy. That is, what calibration does is provide a practical articulation of what Griesemer and Wade (1988, p. 77) claim is the character of the inference from causes observed in the laboratory to causes in nature. Their idea is that the scientist finds matches between the causes in the laboratory and causes in nature. My view is that calibration of laboratory populations to natural populations is a key experimental strategy for demonstrating those matches. According to Griesemer and Wade (1988), analogical reasoning drives the matching between laboratory model and natural system. I have no dispute with Griesemer and Wade about the analogical character of the matching inference. My aim has been to specify the practical strategy used for making that inference and how to view that strategy as warranting the claim being made. The ensuing discussion of claim (b) above, that calibration increases the justificatory strength of empirical claims based on laboratory experiments via its connection with specific elements of Lloyd's (1987, 1988) confirmation view, fleshes out my claim about warrant.

Lloyd (1987, 1988) has argued that confirmation of evolutionary and ecological models can be practically specified by the following three strategies, viz., (1) fit between theoretical

model and data; (2) independent support for aspects of a particular theoretical model; and (3) variety of evidence (which includes varieties of instances of fit and independent support).[7] Following Lloyd (1988, pp. 146-52), briefly, fit between model and data is the determination of a matching between the predictions of a theoretical model and experimental results. Demonstrating fit relies on straightforward statistical curve fitting techniques. Independent support for aspects of a particular model concerns independently garnered empirical evidence for specific assumptions built into a particular theoretical model. Variety of evidence amplifies the two strategies above and comes in three forms, viz., variety of instances of fit, variety of independent support for aspects of a model, and variety of types of support.

Lloyd sees herself as describing at least part of the set of techniques evolutionary biologists use to relate theoretical models to natural systems. More specifically, she sees her three confirmation strategies as concretely articulating techniques that are part and parcel of an experimental model (following Suppes 1962). In the context of my discussion of calibration, the laboratory model (defined in section 2 above) corresponds to a specific instantiation (in the laboratory) of Lloyd's conception of an experimental model. And I see my discussion as specifying even more concretely Lloyd's confirmation account as an articulation of an experimental model. That is, I see calibration as a specifically *experimental* strategy that is used to spell out how to understand the relative strength of the claims that there is fit between model and data, independent support for aspects of a model, and variety of evidence.

Variety of evidence is at the core of Lloyd's confirmation account. Variety of evidence as a strategy itself is at bottom iterations and/or combinations of the strategies of fit between theoretical model and data and independent support for aspects of a theoretical model. Lloyd's

[7] Lloyd discusses these strategies as criteria. But I see no cost to substituting my "strategy" for her "criterion," as I do not think there is any specific weight placed on her use of "criterion."

intuition behind variety of evidence is that numerous demonstrations of, e.g., fit, or, e.g., independent support, can increase the strength of the hypothesis that a theoretical model corresponds to the natural system modeled. That is, 10 instances of fit over data for populations of size 1000 places a relevant hypothesis in a different confirmation situation than 10 instances of fit over data for populations ranging in size between 1 and 1,000,000 (Lloyd 1988, p. 148). A similar claim can be made for independent support. The more times, the more different ways, and the more assumptions are tested independently, the greater the empirical support for an hypothesis made about a model that is based on the assumption. Variety of types of support is just a mixture of variety of instances of fit and variety of independent support. So, an hypothesis made about a theoretical model that has a variety of instances of fit along with a variety of instances of independent support for its assumptions is in a different situation regarding its confirmation than a theory with only one instance of fit.

Here is how I see the explicit connection between the strategy of evaluating variety of evidence and calibration in grounding the strength of Wade and Goodnight's claim. Wade and Goodnight's laboratory model included, as a part of its apparatus, laboratory populations of *Tribolium*. Their aim was to use those populations to determine whether Wright's phase III and, more generally, Wright's SBT, could be considered to describe an important evolutionary process. As I discussed in sections (3) and (4) above, the main way Wade and Goodnight argued that Wright's SBT is important in evolution is by calibrating the population structure of the flour beetle population to the same aspect of other natural populations (of plants and animals, reviewed by Loveless and Hamrick 1984, Wright 1978). As we saw, the F_{ST} measure indicating population structure was chosen for calibration because such population structure is crucial to the operation of Wright's shifting balance process. Moreover, such population structure is indicative

of non-additive genetic variance, a foundational assumption of Wright's SBT. Wade and Goodnight's view is, it seems to me, as follows: If such a match between an aspect of the laboratory population and populations in nature exists, and that factor is central to bringing about evolutionary change on Wright's SBT, then the shifting balance process can be (according to Wade and Goodnight) considered as likely to describe an important evolutionary process. The calibration of the laboratory population, a part of the apparatus (which also included, vials, flour, incubators, microscopes, etc.), played a crucial role in Wade and Goodnight's evaluation of their results.

Wade and Goodnight did not merely calibrate their flour beetle population to, e.g., another laboratory or natural population of flour beetles. Rather, they calibrated their data to several, various natural populations for which the relevant population structure data had been previously collected (Loveless and Hamrick 1984, Wright 1978). So, in order to strengthen their empirical claim that Wright's SBT is important in evolution, Wade and Goodnight provided a *variety of evidence*, specifically, a *variety of independent support of aspects of a theoretical model* by calibrating their laboratory model to natural populations. Differently put, a key assumption built into Wright's SBT is that populations are structured in specific ways. This assumption provides for specific conditions for the shifting balance process to be realized, specifically the interaction between epistasis, migration, and interdemic selection. Relevant to Wade and Goodnight's experimental study, F_{ST} values should be in the range 0.03-0.05 to satisfy the assumption. By calibrating their laboratory model to natural populations, Wade and Goodnight provide independent support for an aspect of Wright's SBT (which is a family of theoretical models). As we saw, F_{ST} values for Wade and Goodnight's laboratory population fit F_{ST} values in the 0.03-0.05 range across a variety of populations found in nature, viz., 13

different species of plant reviewed by Loveless and Hamrick (1984) and numerous species of plant and animal reviewed by Wright (1978). Further, by providing numerous and varied instances of independent support for this key assumption of Wright's SBT, Wade and Goodnight increase the confirmatory standing of their hypothesis via a variety of evidence.

The use of calibration by Wade and Goodnight fits the core of Lloyd's confirmation account nicely. One way to increase the confirmatory standing of an hypothesis is to provide independent support for aspects of a theoretical model in a variety of ways. Wade and Goodnight increased the confirmatory standing of Wright's SBT by providing, via the calibration of their laboratory model, a variety of evidence in the form of a variety of independent support of aspects of the SBT.

What I have argued so far is that the key epistemological strategy for understanding the justification of Wade and Goodnight's claim that Wright's SBT is an important evolutionary process in nature is driven by the calibration of the apparatus, i.e., the flour beetle population, of their laboratory model to populations in nature. This calibration strategy allows Wade and Goodnight to claim that the shifting balance process imposed in the laboratory corresponds to instances of evolution in nature. Further, the way in which the calibration strategy was used, i.e., by marshalling a variety of independent support for aspects of the SBT, increases the strength of their claim.

I have said that calibration as an experimental strategy specifies Lloyd's (1987, 1988) confirmation account. Lloyd takes her account to be a practical specification of the evaluation of experimental models. I take myself to have extended that specification by articulating a concrete experimental strategy that is used to generate the very types of justifications that Lloyd (rightly) demands. My working hypothesis is that such could be done for a number of experimental

strategies, such as those delineated by Franklin (1986, 1990, 2002) and Rudge (1996, 1998). However, doing so here is beyond the scope of the essay, because another key aim has been to defend Wade and Goodnight (1991) from charges leveled at them by Coyne, Barton, and Turelli (1997).

At any rate, Franklin and Rudge have articulated calibration, among other strategies, in their own epistemological frameworks. As we saw, Franklin (1986, 1990) and one version of Rudge (1998) used the Bayesian epistemology to run their epistemology of experiment. Rudge (2001) later argued that Mayo's (1996) error statistical approach better depicts his case, i.e., industrial melanism in the peppered moth. I present calibration as part of, and connected to, a further set of theory evaluation strategies (following Darden 1991 on the concept of a strategy). Nevertheless, a Bayesian or error statistical gloss on the strategies presented here could, at least to some extent, be used to more formally "epistemologize" them.

6. Conclusion

I have argued that Wade and Goodnight (1991) used calibration as an experimental strategy for warranting their claim that Wright's SBT is an important evolutionary process. Moreover, I have shown how their use of calibration defends them from Coyne et al.'s (1997) objection that the results are ambiguous. Ultimately, my view is that calibration of laboratory models is a key experimental strategy for warranting claims made via laboratory experiments in population genetics. Calibration allows an experimenter in population genetics to warrant claims about the correspondence between laboratory systems and natural systems.

References:

Brandon, R. 1994. "Theory and Experiment in Evolutionary Biology." *Synthese* 99: 56-73.

Collins, H. 1985. *Changing Order*. London: Sage Publications.

Coyne, J. A., N. H. Barton, and M. Turelli. 1997. "Perspective: A Critique of Sewall Wright's Shifting Balance Theory of Evolution." *Evolution* 51: 643-671.

Coyne, J. A., N. H. Barton, and M. Turelli. 2000. "Is Wright's Shifting Balance Process Important in Evolution?" *Evolution* 54: 306-317.

Crow, J. F. 1991. "Was Wright Right?" *Science* 278: 973.

Crow, J. F., W. R. Engels, and C. Denniston. 1990. "Phase Three of Wright's Shifting Balance Theory." *Evolution* 44: 233-247.

Darden, L. 1991. *Theory Change in Science: Strategies from Mendelian Genetics.* New York: Oxford University Press.

Diamond, J. 1986. "Overview: Laboratory Experiments, Field Experiments, and Natural Experiments." Pp. 3-22 in *Community Ecology*. Edited by Diamond, J. and T. Case. New York: Harper and Row Publishers.

Dietrich, M. 1996. "Monte Carlo Experiments and the Defense of Diffusion Models in Molecular Population Genetics." *Biology and Philosophy* 11: 339-356.

Falconer, D. 1989. *Introduction to Quantitative Genetics*. 3rd edition. New York: Longmans.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*. Oxford, UK: Oxford University Press.

Fisher, R. A. 1958. *The Genetical Theory of Natural Selection.* 2nd Edition. New York: Dover Publications, Inc.

Fisher, R. A. 1999. *The Genetical Theory of Natural Selection. A Complete Variorum Edition.* Edited by Bennett J. New York: Oxford University Press.

Franklin, A. 1986. *The Neglect of Experiment*. Cambridge, UK: Cambridge University Press.

Franklin, A. 1990. *Experiment: Right or Wrong*. Cambridge, UK Cambridge University Press.

Franklin, A. 1997. "Calibration." *Perspectives on Science* 5: 31-80.

Franklin, A. 2002. *Selectivity and Discord: Two Problems of Experiment*. Pittsburgh: University of Pittsburgh Press.

Gayon, J. 1998. *Darwinism's Struggle for Survival: Heredity and the Hypothesis of Natural Selection.* Cambridge, UK: Cambridge University Press.

Goodnight, C. J. and M. J. Wade. 2000. "The Ongoing Synthesis: A Reply to Coyne, Barton, and Turelli." *Evolution* 54: 317-324

Griesemer, J. and M. J. Wade. 1988. "Laboratory Models, Causal Explanations and Group Selection." *Biology and Philosophy* 3: 67-96.

Hodge, M. J. S. 1992. "Biology and Philosophy (Including Ideology): A Study of Fisher and Wright." Pp. 231-293 in *The Founders of Evolutionary Genetics*. Edited by Sarkar, S. Dordrecht: Kluwer.

Kohler, R. E. 1994. *Lords of the Fly:* Drosophila *Genetics and the Experimental Life*. Chicago: University of Chicago Press.

Lloyd, E. A. 1987. "Confirmation of Evolutionary and Ecological Models." *Biology and Philosophy* 2: 277-293.

Lloyd, E. A. 1988. *The Structure and Confirmation of Evolutionary Theory*. Princeton: Princeton University Press.

Loveless, M. D. and J. L. Hamrick. 1984. "Ecological Determinants of Genetic Structure in Plant Populations." *Annual Review of Ecology and Systematics* 15: 65-.95.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: Unversity of Chicago Press.

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics*. Chicago: University of Chicago Press.

Provine, W. B. 1986. *Sewall Wright and Evolutionary Biology*. Chicago: University of Chicago Press.

Rudge, D. W. 1996. *A Philosophical Analysis of the Role of Natural Selection Experiments in Evolutionary Biology*. Ph.D. dissertation. University of Pittsburgh.

Rudge, D. W. 1998. "A Bayesian Analysis of Strategies in Evolutionary Biology." *Perspectives on Science* 6: 341-360.

Rudge, D. W. 2000. "The Complementary Roles of Observation and Experiment: Th. Dobhzansky's Genetics of Natural Populations IX and XII." *History and Philosophy of the Life Sciences* 22: 167-186.

Rudge, D. W. 2001. "Kettlewell from an Error Statistician's Point of View." *Perspectives on Science* 9: 59-77.

Skipper, Jr. R. A. 2000. *The R. A. Fisher-Sewall Wright Controversy in Philosophical Focus: Theory Evaluation in Population Genetics*. Ph.D. dissertation. University of Maryland, College Park.

Skipper, Jr. R. A. 2002. "The Persistence of the R.A. Fisher-Sewall Wright Controversy." *Biology and Philosophy* 17: 341-367.

Suppes, P. 1962. "Models of Data." Pp. 252-261 in *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Edited by Nagel, E., P. Suppes, and A. Tarski. Palo Alto, CA: Stanford University Press.

Wade, M. J. 1991. "Genetic Variances for Rate of Population Increase in Natural Populations of Flour Beetles, *Tribolium* SPP." *Evolution* 45: 1574-1584.

Wade, M. J. 1992. "Sewall Wright: Gene Interaction and the Shifting Balance Theory." Pp. 35-62 in *Oxford Surveys in Evolutionary Biology, Vol. 8*. Edited by Futuyma, D. and J. Antonivics. New York: Oxford University Press.

Wade, M. J. and C. J. Goodnight. 1991. "Wright's Shifting Balance Theory: An Experimental Study." *Science* 253: 1015-1018.

Wade, M. J. and C. J. Goodnight. 1998. "Perspective: The Theories of Fisher and Wright in the Context of Metapopulations: When Nature Does Many Small Experiments." *Evolution* 52: 1537-1548.

Wilson, D. S. and R. K. Colwell. 1981. "Evolution of Sex Ratio in Structured Demes." *Evolution* 35: 882-897.

Wright, S. 1931. "Evolution in Mendelian Populations." *Genetics* 16: 97-159.

Wright, S. 1932. "The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution." *Proceedings of the Sixth Annual Congress of Genetics* 1: 356-366.

Wright, S. 1951. "The Genetical Structure of Populations." *Annals of Eugenics* 15: 323-354.

Wright, S. 1968. *Evolution and the Genetics of Populations, Vol. 1: Genetic and Biometric Foundations*. Chicago: University of Chicago Press.

Wright, S. 1969. *Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies*. Chicago: University of Chicago Press.

Wright, S. 1977. *Evolution and the Genetics of Populations, Vol. 3: Experimental Results and Evolutionary Deductions*. Chicago: University of Chicago Press.

Wright, S. 1978. *Evolution and the Genetics of Populations, Vol. 4: Variability Within and Among Populations*. Chicago: University of Chicago Press.