The left hand side of the Fundamental Theorem of Natural Selection

Alan Grafen[a]

[a] *Zoology Department and St John's College, Oxford University, Postal Address: St John's College, Oxford OX1 3JP, UK*

Abstract

The fundamental theorem of natural selection is explained here in very simple terms, suitable for students. The biological significance of the left hand side – the rate of change in mean fitness due to changes in gene frequencies, which is also described as the rate of change due to natural selection – has been regarded since 1972 as problematic, but here a simple graph is used to show that Fisher's poor explanation was of a robust and simple intuition. Simple numerical examples show the theorem at work with fixed genotypic fitness under two different mating systems, with bland density dependence, and also with fitnesses determined by an evolutionary game. The content of the theorem has long been taken for granted by whole-organism evolutionary biologists, though in an imprecise way, even while mathematical population geneticists have been, in sequence, wrongly proving it false, wrongly proving it requires more assumptions than Fisher admitted, and accepting the truth of the theorem as Fisher proved it, but doubting its biological significance. An important emphasis on the instantaneous nature of natural selection, and of its measurement, emerges from the argument. Price's disappointments with the content of the theorem are directly confronted. The new explanation allows us to recognise the central place the theorem already occupies in evolutionary biology, and to begin to incorporate more fully the insights embedded in it.

Keywords: biological fitness, natural selection, partial change in mean fitness, additive genetic component, formal Darwinism

Email address: alan.grafen@sjc.ox.ac.uk (Alan Grafen)

[1] *Competing interests statement:* "Declarations of interest: none".

[2] *Funding statement:* "This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors".

1. Introduction

An inquisitive student in ethology, animal behaviour, or behavioural ecology, interested in the role of natural selection in theory and practice, can find much to be unhappy about. Many undergraduates are taught that natural selection causes organisms to act as if they are maximising their fitness, and also told that we are not quite sure what fitness is. This is a curious position, and furthermore one that is not peripheral, but central to those subjects. This paper leaves fitness-maximisation for future discussion, but is concerned with the more modest ideas that natural selection is an improving process, and that it improves fitness, as well as what fitness is. Improvement is uncontroversial in a very simple situation, with haploid asexual organisms, when only one locus is under consideration, and individual organisms are all the same, thus excluding population structure, for example age structure. A concept of fitness as simply number of descendants (including the organism itself, if it survives) will fulfil most requirements. But what about diploidy, sexual reproduction, multiple loci, age and other structure? Most organisms studied in the field have most, and many have all, of these characteristics. Can the action of natural selection be characterised as 'improving' with these complexities? If so, what is the nature of the 'fitness' that is improved?

These are the questions to which Fisher (1930)'s fundamental theorem of natural selection finds answers. Unfortunately, biologists were distracted by mathematical population geneticists, whose serious misreading of the theorem prevented them understanding even the statement of the theorem until Ewens (1989) brought to their attention the first serious attempt to articulate Fisher's proof, by Price (1972b). By 1989, biologists had been finding their own fragmentary answers to the important questions, and anyway the conclusion of Price and the mathematical population geneticists was that, although the newly comprehended theorem was true, they could find no useful significance in it. Their doubt centred on the meaning of the left hand side of the theorem. The present paper aims to persuade the reader that the left hand side of the theorem has a very important meaning, and that the theorem has very satisfactory answers to the questions of the previous paragraph. Biologists should acknowledge the fundamental theorem as the central result of evolutionary biology, and understand more recent work as partial replacement or, in some cases, partial extension. The methodology of the theorem is ripe for systematic development, to form a new core theory.

An elaboration and formal treatment of Fisher's theorem and proof has been provided previously (Grafen, 2015a). After some scene-setting in Section 2, the present paper introduces in Section 3 a new graph and partition of the total change in mean fitness. This clarifies and justifies Fisher's identification of the change in mean fitness due to changes in gene frequencies with the change due

to natural selection, which resolves the meaning of the left hand side. The right-hand side of the theorem is given some less rigorous attention in Section 4, to indicate its application in more general settings, some included in Fisher's own theorem and some going beyond it, and also connections to other ideas, such as the Price Equation. Then, numerical simulations in Section 5 show the partition and the theorem at work in simple cases, and show more concretely what they mean. The following section deals head-on with the disappointment in the theorem felt by Price (1972b) and most subsequent commentators, arguing that this was due to misunderstanding of the theorem, and of what general truths about natural selection are available. The paper closes with a discussion of the current status of fundamental theorem.

2. Preliminaries and context

The original statement of the theorem is 'The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time' (Fisher, 1930, page 35). Some ambiguities are resolved in the following restatement (Grafen, 2015a):

The part of the rate of change in mean fitness of any population of organisms that is due to natural selection at any time	=	the additive genetic variance in fitness at that time.

Unfortunately, Fisher did not explicitly define the central quantity 'fitness of an individual', and indeed the first two major rederivations, by Price (1972b) and Lessard (1997), differ from each other on this core question. Fisher published the theorem in a book intended for a popular audience, and so his presentation was more verbal and less mathematical than it might have been, which is one cause of the theorem's chequered history.

The second preliminary is that Fisher's theorem was in continuous time, so he discussed the rate of change in mean fitness over time. Grafen (2015a) derives a parallel discrete-time version, and we follow Ewens and Lessard (2015) in employing discrete time. The right hand side of the theorem, which is presented in a simplified version in Section 5.1, has a divisor of the mean number of offspring in the discrete form. Otherwise, the only point to note is that Price (1972b)'s notation of $\partial_{NS}\overline{w}$, which is appropriate for continuous time, will be replaced with the less perplexing $\Delta_{NS}\overline{w}$ when this notation is needed.

Next, a key point is that the theorem aims to distinguish the effect of natural selection from evolutionary change in general. Fisher makes no secret of this ambition – the first sentence of *The Genetical Theory of Natural Selection* is famously 'Natural Selection is not Evolution' – and some authors have taken this point (Price, 1972b; Frank and Slatkin, 1992; Gayon, 1998), though others

regard Fisher's intention as unfulfilled (Ewens, 1989, 1992). We will see that this partly underlies the lack of appreciation by mathematical population geneticists (e.g. Ewens and Lessard, 2015), whose main interest seems to be in total change in the mean fitness, rather than in isolating that part of it due to natural selection.

Finally, the history of the theorem may be followed up in Edwards (1994, 2014). The view from mathematical population genetics has recently been provided by Ewens and Lessard (2015), and the fundamental theorem is developed and linked to other methodologies by Frank (1997, 1998). Without further ado, we introduce the expository graph and partition.

3. A three-way partition of evolutionary change

The purpose of this section is to explain in conceptual terms the meaning of the left hand side of the fundamental theorem. We first establish a three-way partition of the evolutionary change in the mean of a quantitative trait, and will discuss each component in turn, one of which will turn out to be the left hand side of the theorem *in the special case when the trait whose mean is being studied is fitness itself.* The simple algebra that derives it very generally is relegated to Appendix Appendix A. Here, the reader is directed to Figure 1, which has five panels. For each panel, we can calculate a mean value for the trait, using

$$\bar{x} = p_{11}x_{11} + p_{12}x_{12} + p_{22}x_{22},$$

where p is the population proportions of each genotype (A_1A_1, A_1A_2, A_2A_2) and x is the trait values. Note that the trait values for a given genotype may be different at different times for various kinds of reasons. A prime indicates as usual 'in the next year'.

The mean calculated for Panel 1 is the mean trait this year, and the mean for Panel 5 is the mean trait next year. By looking at the differences from one panel to the next, we decompose the difference between this year's and next year's means into an initial four components. Table 1 shows this decomposition. The first part of the change, from the first to second panel, is always zero, as guaranteed by the theory of linear regression. More is said about the regression in Appendix Appendix B. The remaining three components represent our three-way partition of the change in the mean of a trait from this year to next year, which we can represent formally as

$$\overline{x} + \Delta_{AG}\overline{x} + \Delta_{NAG}\overline{x} + \Delta_{PGL}\overline{x} = \overline{x}'.$$

Each component of the partition has a name. The move from the second to third panel is the *additive genetic component* of the change (AG), and from the third to fourth is the *non-additive genetic component* (NAG) These deal only

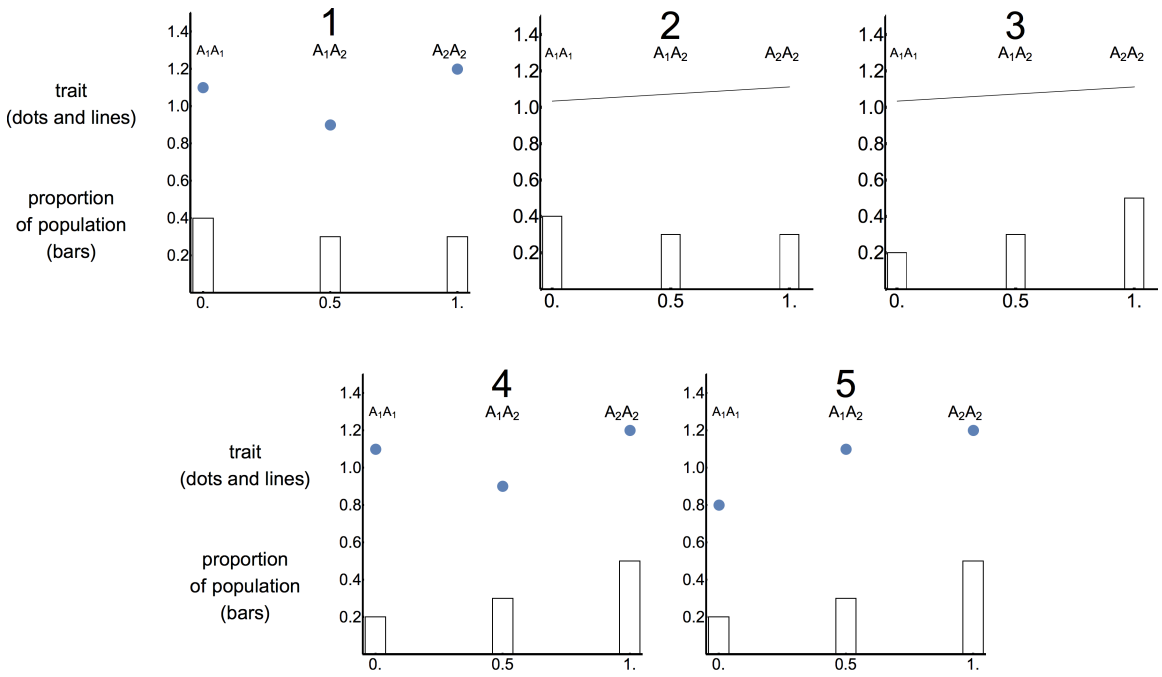

Figure 1: A partition of the total change in the mean of a quantitative trait. Each panel shows on the x-axis the frequency of the allele A_2 in each of the three genotypes (A_1A_1 at 0, A_1A_2 at 0.5, and A_2A_2 at 1), and the y-axis shows both the proportion of the population comprised of those genotypes (for the histogram bars) and the trait value for each genotype (the dots or line). The histograms show the population proportions this year (Panels 1 and 2) or next year (Panels 3, 4 and 5). The dots show either the value of the trait for each genotype this year (Panels 1 and 4) or next year (Panel 5). The straight line (Panels 2 and 3) is the best-fitting straight line for this year's trait-values, by least squares, weighting by this year's population proportions. For each panel, a mean value of the trait can be calculated on the basis of the population proportions and the trait-values (where there are dots) or fitted trait values (where there is a line).

with trait values and fitted trait values based on this year's values. The final change assumes next year's population proportions, and is the change in mean based on the differences between the years in the trait values assigned to each genotype, and we call this the *phenotype-genotype link* component (PGL).

It is important to note that partitions of this kind are a routine tool in population genetics (see, for example Crow and Kimura, 1970), going back at least as far as Fisher (1918). Most similar to the one here are those of Price (1970, 1972a) and Frank (e.g. Frank, 1997, 1998, 2012). So far as I know, this particular partition has not been proposed before.

Table 1: The decomposition of the change in the mean of a trait, using the sequence of population proportions and trait-values in Figure 1.

Panel	Proportions	Trait	Value	Difference
1	this year	this year's values	$\overline{x}$	
				Zero
2	this year	this year's regression		
				$\Delta_{AG}\overline{x}$
3	next year	this year's regression		
				$\Delta_{NAG}\overline{x}$
4	next year	this year's values		
				$\Delta_{PGL}\overline{x}$
5	next year	next year's values	$\overline{x}'$	

The central argument is first that our aim is to find a part of the total change that can usefully be ascribed to natural selection, rather than any other causal force, and second that the additive genetic component represents that part. We proceed by showing that the PGL component should not be regarded as due to natural selection, with an argument that has significant implications that may not be immediately obvious, and then that the NAG component should also not be regarded as due to natural selection, while the AG component should. The additive genetic component, when the trait being studied is fitness itself, turns out to be the left hand side of the fundamental theorem. These conclusions show that natural selection acts and we measure it according to various contingent circumstances at the moment of measurement, as the examples of Section 5 will show: we will refer to the 'instantaneous' nature of natural selection and its measurement. The omission of the PGL term removes dependence on effects between different moments.

The PGL component is settled by an appeal to a central and important principle of Darwinism. The PGL component is about a change that takes place in the future, after natural selection has already acted. Darwin's central point was that natural selection was an automatic process that worked without foresight, and required no intelligence to predict the future. The mechanical 'physics and physiology' aspect of natural selection is key to the role it plays in explaining biological design in a scientific way. Thus, on this matter of high principle, we can exclude the PGL component from the desired measure of the effect of natural selection. This has important consequences that will be returned to, especially in the numerical examples when we look at an evolutionary game.

Important patterns for the AG and NAG components are now shown *in*

the case where there are no fitness differences between genotypes. We will show first, that the AG component is always zero and second, the NAG component can be non-zero. As natural selection must have precisely no effect in that situation, this strongly suggests selecting the AG component as the measure of natural selection. (Frank (1997, pp 1725–1727)) interestingly analyses an example of the Price Equation without selection, but says 'The calculations are relatively easy because there is no selection and thus no partitioning of change into components.', and so does not distinguish additive from non-additive components.) Consider first the AG component, and keep in mind that it is the difference between Panels 2 and 3 of Figure 1. As there are no fitness differences, the Hardy-Weinberg theorem tells us that the population gene frequency does not change. This implies that the average individual *gene* frequency does not change, even though the *genotype* frequencies change. Referring to our graphs in Figures 1 and 2, if the mean on the x-axis does not change, then when calculating the mean of the quantity on the y-axis in terms of the straight line, that mean does not change either. With a non-linear curve, changing the mean on the x-axis *does* change the mean on the y-axis. This is one central property of straight lines and regressions needed for our argument here. It follows that the mean, when calculated from the straight line, does not change either, no matter how much the *genotype* frequencies may change. The genotype frequencies may well change, even in the absence of selection: under random mating, the population relaxes into Hardy-Weinberg proportions; equally, under selfing (automixis), the heterozygote is halving in frequency each year and adding equal quantities of both homozygotes to the existing numbers. As a special case of the general principle, notice that the fraction of the population moving from 0.5 to 0 has an equal and opposite effect on the mean to the fraction moving from 0.5 to 1, because the fitnesses come from a straight line.

Now we consider the NAG component, which is calculated from the difference between Panels 3 and 4 of Figure 1. The selfing case now does change the mean trait whenever the trait values do not lie on a straight line. In the figure, the mean is increased each year, because both homozygote trait values lie above the line. When heterozygotes halve in frequency, the other half appears in equal portions as each of the homozygotes: this swap increases the mean in the diagram (and would decrease the mean of the trait if the homozygote fitnesses lay below the line). This change in the mean clearly results from segregation of the alleles, and in the absence of fitness differences between genotypes cannot be considered as due to natural selection. The important point about these NAG effects is that they apply when the trait is height, and would also apply when the trait is chosen to be fitness itself, and this shows that some part of the change in mean fitness should definitely not be regarded as due to natural selection. The NAG effect on fitness will not be quantitatively exactly equal

to the parallel effect on height, because the change in population proportions, which will inevitably happen when the trait is fitness but may not in the case of height, will interact with the NAG effect. The resulting discrepancies will, in all the cases we see here at least, be proportionally small compared to the NAG effects themselves.

The case of no fitness differences has been formalised by Grafen (2015b) as a 'principle of neutrality' for testing proposals to measure natural selection. Grafen explains how this principle illuminates the two choices of definition for 'fitness' made by Price (1972b) and Lessard (1997). Lessard's choice (which originates with Kimura, 1958) is incompatible with the principle, and the principle also resolves a large element of ambiguity in Price's definition.

These considerations about the components of the partition shown in Table 1 explain and fully justify Fisher's identification of the change 'due to all changes in gene ratio' (Fisher, 1930, page 35), on the one hand, with the change 'produced by Natural Selection', 'determined by natural selection' and 'ascribable to natural selection' (the first two on page 42 of the 1930 edition, and the third on page 57 of Fisher (1941)). The fact that it puzzled Price (1972b) so much may possibly be put down to his failure to appreciate the full significance of Fisher's aim of identifying a component due to natural selection as opposed to evolution as a whole. The puzzlement is surprising in light of Price's own other work at the same time, as the first term in what is now called the Price Equation (Price, 1970, 1972a) is numerically equal to the additive genetic component here, as noted in Section 4.

The basic argument completed, we note that in the numerical examples of Section 5 we will see how these definitions work out in practice. Now we note reasons why this identification of the AG component with 'due to natural selection' should be regarded as suitable and satisfactory in a wider context. First, it is genes, not gene combinations, that are passed to offspring and especially to grandoffspring. Thus, as Fisher (1930, p. 32) puts it when discussing the additive component of stature, 'The expected value [sc. breeding value, additive genetic component] will not necessarily represent the real stature, though it may often be a good approximation to it, but its statistical properties will be more intimately involved in the inheritance of real stature than the properties of that variate itself.' Second, a species could change its genotypic constitution rapidly by switching from random mating to selfing – and then back again – without harming biological design. Species that can reproduce sexually and asexually are the foundation for Williams' famous 'balance argument', and include most aphids, many plants, most fungi, and many single-celled eukaryotes Williams (1975); of these, those whose asexuality takes the form of selfing must be hermaphodites, including many plants. Fisher hypothesises a plant species that outcrosses or inbreeds depending on the presence of a pollinator in each

given year. On the other hand, changes in gene frequency take time and work. This is the argument of Fisher (1941) as elaborated in his much-quoted letter to Oscar Kempthorne (Fisher's letter to O. Kempthorne on 18 February, 1955, reproduced in Bennett, 1983, p229). A third reason is that the differences between humans and chimps are likely the result of changes in gene frequencies. A change in linkage or epistasis has never, so far as I know, been suggested as causing an adaptive change that underlies population differences. For all of these reasons, the distinction made on technical grounds fits in comfortably with other ideas about natural selection.

Thus, Fisher's thinking about the significance of the left hand side of the fundamental theorem turns out to be not so hard to understand. The arguments are almost all in Fisher's writings, with the relevant intentions. It is a matter for historians of science to explain why it has taken so long for biologists to pick up on the relevant clues and expand on Fisher's cryptic explanations.

This section is closed with Table 2, which summarises Fisher's view of the biological significance of the three different components of the change in mean fitness. The fundamental theorem is what allows us to know the sign of the additive genetic component. The non-additive genetic component will be, in a rough sense, equally likely to be favouring an adaptive change as resisting it. For any model in which the non-additive genetic component favours an adaptive change there will be a 'mirror model' in which it resists change to an equal extent, where the mirror model differs only in details of the genetic architecture and not in terms of the interaction of phenotypes with the enviroment. This is what justifies the conclusion that it has no connection to design and adaptive complexity.

Table 2: The three components of the total change in mean fitness, and their biological significance.

Component	Sign	Determined by	Connection to design
Additive Genetic	Non-negative	Additive genetic variance in fitness	The creative force in evolution
Non-additive Genetic	Positive or negative	Genetic details	None
Phenotype-Genotype Link	Positive or (usually) negative	Environmental change *sensu latu*	Presents challenges to natural selection

Thus, the fundamental theorem's result that the change in mean fitness due to natural selection equals a variance, which cannot be negative, has a very significant biological meaning, and can be taken as the mathematical version of

Darwin's argument that natural selection is an improving process. As Ewens and Lessard (2015) point out, the complications of Mendelian genetics make this formulation more complicated than in the very simple situation of haploidy. Fisher shows that the intuition based on haploidy has a very general and simple representation in the case of diploidy.

4. Generalities and connections

The reader keen to see numerical examples, or to find relief from conceptual abstraction, should move first to Section 5, where numerical examples are developed to show the theory at work. Here, we take a moment to consider the right-hand side of the theorem in the light of what we have learned about the left-hand side, pointing out the many generalisations available beyond the simple one-locus genetics with an unstructured population of the previous section, and also connections to the Price Equation. Fisher proved his fundamental theorem for an age-structured population, and with arbitrary genetic architecture, though he failed to give in explicit form the definition of the fitness of an individual. His derivation is expanded and completed by Grafen (2015a).

First, Fisher's genetic architecture allows arbitrary numbers of loci and alleles, with arbitrary dominance relations within each locus, arbitrary epistasis between loci, and also arbitrary linkage and arbitrary linkage disequilibrium. Further, no assumption is made about the mating system, and for structured populations the theorem holds whether demographic equilibrium holds or not. While it may seem very surprising that the theorem can hold in all this generality, Darwin's argument that natural selection is an improving process is very persuasive, and yet makes no assumptions about these matters (of course, Darwin was ignorant of Mendelian genetics and so could not even have considered making these assumptions). Fisher tamed all of this complexity with the same principle of least squares that he wielded to such good effect in much of his statistical work. The extension of the argument of the previous section is only a matter of replacing the simple linear regression with a multiple linear regression in which the frequency of each gene is an explanatory variable and fitness remains as the response variable, as detailed in Appendix Appendix B. All the conclusions then follow in just the same way. There is one restriction on the genetic architecture, which is that alleles all obey Mendelian rules and belong to the same coreplicon (Cosmides and Tooby, 1981): thus, usually we would think of autosomes for mammals, and not include sex-linked or mitochondrial loci.

The second major generalisation is to structured populations, which comes down to the question of how to calculate fitness. The previous section simply takes number of offspring as the measure of fitness, in a model with non-overlapping generations. Fisher's very general theorem comes unsurprisingly

with a very precise way of measuring fitnesss. We have managed with simple parametrisation so far because the examples have been very simple. Once generations overlap, we need Fisher's concept of reproductive value to assign a numerical value to each age class. Sticking with our discrete time model, we can now follow Williams (1966). He employed his understanding of the fundamental theorem to explore survival/fertility tradeoffs, and in this founding paper of life-history theory, he worked out a 'Williams' reproductive value' by adding up the expected reproductive value from the surviving self (the probability of survival times the reproductive value of an individual in the next age class) and the expected reproductive value from offspring (expected number of offspring times the reproductive value of a newborn). This sum was taken by Williams to be what Fisher meant by fitness, and used as the maximand that underlies life history theory. However, if we denote the Williams reproductive value of individual i by W_i, and the reproductive value from Fisher's calculation for an individual in i's age class as v_i, then Grafen (2015a) shows that Fisher's definition of fitness is

$$m_i = \frac{W_i - v_i}{v_i}.$$

This definition has some surprising implications. Like Williams' version, fitness is not a single tombstone measure for an individual, but instead an individual has a fitness at each moment in time. Special properties stemming from Fisher's ingenious construction include that the average fitness of each age class is equal, and furthermore equal to the population's Malthusian parameter, with the notable implication that the average fitness is *zero* in a stationary population. These and other points are explored by Grafen (2015a,b). The importance here is that by making the appropriate definition of fitness, the arguments of the previous section apply to structured populations too.

The value of the theorem can be regarded as lying precisely in the accompanying definition of fitness. The theorem itself can seem somewhat innocuous - yes, natural selection is an improving process. But to find how to define fitness in a complex model, with age structure, is to discover what the quantity is that natural selection improves — an achievement of considerable interest, that resolves one of the most profound and simple questions at the heart of Darwinism. Does selection act on survival, or on fecundity, or on some combination of the two? Presumably the latter, but which combination? Fisher's definition of fitness answers this question, and the fact that the fundamental theorem holds with *that* definition is what guarantees Fisher is right. Similarly, more general models than Fisher's have their definitions of fitness validated by their role in a version of the fundamental theorem (Grafen, 2015b; Batty et al., 2014; Crewe et al., 2017). Gayon (1998, section 7.2.3) presents the fascinating history of the recognition that fertility was relevant to natural selection, and the conceptual struggle it engendered within biology: Fisher resolved it.

We now turn to two connections of the theorem. The history of the view of the theorem from mathematical population genetics is told in different ways by Ewens (1979), Edwards (1994) most comprehensively, Ewens (2004), and Edwards (2014), and will not be repeated here. It is interesting, however, to accept Ewens and Lessard (2015) as the latest word from that subject, in a paper entitled 'On the meaning and significance of the fundamental theorem of natural selection'. They have separately published the most significant work in mathematical population genetics on the theorem (Ewens, 1989, 2004; Lessard, 1997). In their joint paper, Ewens and Lessard are very relucant to accept a measure of the change in mean fitness that omits any evolutionary forces, and repeatedly point to causes that the fundamental theorem does not include in its left hand side. The argument of the current paper is that distinguishing the effects of natural selection from other parts of evolutionary change is the whole point of the theorem, and this logically requires omitting some causes from the measurement: as Ewens and Lessard nowhere even discuss making this distinction, it is not surprising that they come to a negative view of the value of the theorem.

The second connection is with Price's equation. Price (1972b) was the first to tackle Fisher's logic with sufficient energy and understanding to make serious sense of the proof of the fundamental theorem, although there were elements still missing (Grafen, 2015a). Simultaneously, Price (1970, 1972a) was working on what has become known as the Price equation, and it is natural to ask whether there is a connection between the components of the change in the mean of a trait provided above, and the components identified by Price. The answer is subtle. Price (1970)'s formula is for the change in a gene frequency, or an arbitrary weighted sum of gene frequencies. If this formula is applied to the breeding value of a trait (which equals its additive genetic component), then the covariance term (including the division by mean number of successful gametes) for the breeding value equals the additive genetic component for the trait in the partition above. The second of Price's terms represents mutation (not included in the framework of this paper or of Fisher's) and also the variability due to the randomness of meiosis, as well as bias at meiosis, but does not include the non-additive genetic component or the phenotype-genotype link component. Moving to a later work by the same author, Price (1972a)'s equation does track the change in the mean of a trait, and so far as I can make out, under the most straightforward application, the covariance term equals the additive genetic component of our partition, and the second term includes the non-additive genetic and phenotype-genotype link components. This connection between the fundamental theorem and the Price Equation essentially agrees with that proposed by Frank (1997, 1998).

5. Numerical Examples

Numerical examples will be helpful for many biologists, who understand general principles by concrete applications, and those who are happy with conceptual arguments at the start, but now need to see how to apply the ideas in practice. Here, some simple examples of Ewens and Lessard (2015) are repurposed and extended to illustrate some important general points. Four sets of examples will be given. We begin by assuming there is no change in the phenotype-genotype link: the first set of examples adopt a linear connection between genotype and phenotype, while the second set has a non-linear connection. These examples show that the additive genetic component acts in a very suitable way for a measure of natural selection, and show quite dramatically that the non-additive genetic component definitely does not. In both cases, following Ewens and Lessard, we give a random mating simulation and a selfing simulation, and this allows the suitability of the measures to be displayed in a wider range of situations.

The first two sets of examples define fitness as 'number of offspring' and, under this definition, number of offspring can be fitness only if the population is growing indefinitely (if mean fitness is greater than one) or shrinking indefinitely (if less than one). Biologists often assume a 'bland density dependence', under which the relative frequencies of genotypes follow the 'number of offspring' definition, but the total size of the population is kept fixed. The third set of examples repeats the second set, but with a definition of fitness that implements this bland density dependence. Essentially, we move partway towards Fisher's precise definition in this case. This device is usefully employed in the fourth set of examples, in which we assume the population is playing the Hawk-Dove game, with genetically determined strategies. This allows a numerical analysis of how selection acts, and an illustration of the failure that Dawkins (1976) predicts for the 'conspiracy of doves'. This example has been regarded as problematic by Lehmann and Rousset (2014) and others, and it is important that the additive genetic component continues to represent the effect of natural selection in a frequency-dependent model.

There are many further possible examples, employing age structure or other population structure, and multiple alleles and multiple loci, which could usefully show how the theorem applies in further cases, and further exploration is encouraged. We first set out Ewens and Lessard's setting in slightly more detail, to pin down the assumptions of the numerical examples.

5.1. The underlying model

Ewens and Lessard (2015)'s notation and result are simplified here still further, to present a model that matches their numerical simulations. A note for mathematicians should begin by saying that Ewens and Lessard (2015) assume

an infinite diploid sexually reproducing population with non-overlapping generations in discrete time. Mating may be random or automictic selfing. They follow Fisher in considering only autosomal loci, and also in not distinguishing the sexes: we can think of this as a hermaphroditic population. We take advantage of the notation to state the theorem in a one-locus two-allele version of the already very simple setting of Ewens and Lessard (2015), so readers are not kept completely in the dark about its mathematical nature. For proofs, readers are referred to Ewens and Lessard (2015) for this simple case, and to Grafen (2015a) for Fisher's original case.

Our earlier notation was chosen for consistency, so we have a single autosomal locus with two alleles A_1 and A_2. The frequencies of the genotypes A_1A_1, A_1A_2 and A_2A_2 are p_{11}, p_{12} and p_{22}, with $p_{11}+p_{12}+p_{22}=1$; the fitnesses are initially fixed parameters, w_{11}, w_{12} and w_{22}. These assumptions are enough to calculate the gene frequencies next year, but not the genotype frequencies, because we have made no assumption about how individuals pair up to mate: this point is important to bear in mind, because the fundamental theorem is proved with a similar innocence. And if we do not know the genotypes, we cannot calculate the mean fitness next year either.

The mean fitness is simply calculated as

$$\overline{w} = p_{11}w_{11} + p_{12}w_{12} + p_{22}w_{22}.$$

Next, we write down the genetic variance in fitness as (by definition in a model with no environmental effects, simply the expected squared deviation from the mean):

$$\sigma^2 = p_{11}(w_{11} - \overline{w})^2 + p_{12}(w_{12} - \overline{w})^2 + p_{22}(w_{22} - \overline{w})^2.$$

Next is a concept known as the 'additive genetic variance'. The whole variance allows each genotype to have its own fitness. The additive genetic variance allows each allele to have a given, additive effect, which implies that the heterozygote is exactly intermediate between the homozygotes. This means fitting a straight line through the fitness values, appropriately weighted by genotype frequencies. See Figure 2, and for details, Appendix Appendix B. The part of the variance accounted for by the straight line is the additive genetic variance, while the rest is the non-additive genetic variance. Let the fitted values of the genotypes, that is, the fitnesses predicted for each genotype by the straight line, be $\hat{w}_{11}$, $\hat{w}_{12}$ and $\hat{w}_{22}$. The formula for the additive genetic variance is

$$\sigma_A^2 = p_{11}(\hat{w}_{11} - \overline{w})^2 + p_{12}(\hat{w}_{12} - \overline{w})^2 + p_{22}(\hat{w}_{22} - \overline{w})^2. \quad (1)$$

This gives us the second ingredient for the theorem.

The third ingredient is notated by Ewens and Lessard (2015, following Price

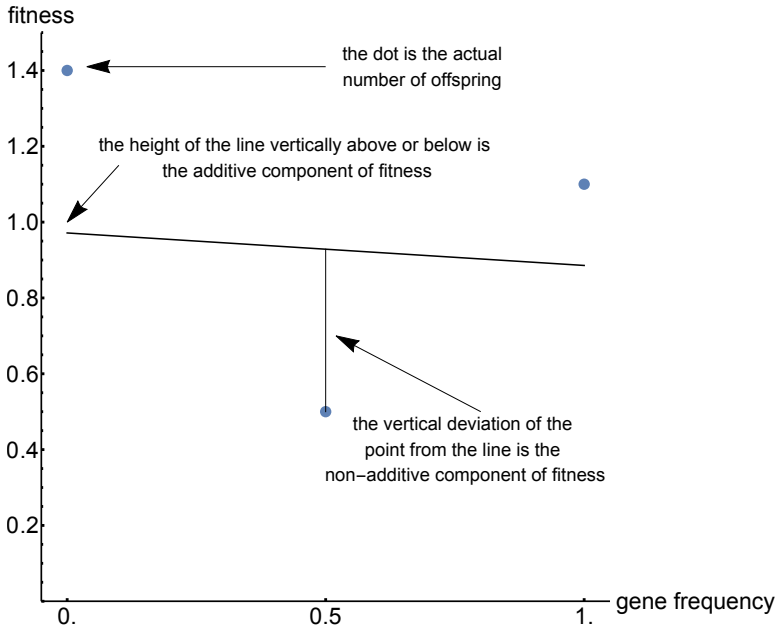

Figure 2: Additive and non-additive genetic effects on fitness The points show illustrative fitness values. The straight line is the best fitting straight line through the points, weighting by the frequencies of the genotypes. The additive fitness values, or the breeding values of fitness, are given by the height of the line for each genotype. The point for genotype g is at height w_g, the line is at height $\hat{w}_g$, and the deviation of point from line is $\tilde{w}_g$. The *average effect* of substituting A_1 for A_2 is height of the line at 0 minus the height at 0.5, which equals the height at 0.5 minus the height at 1: it is thus minus one half times the slope of the line.

(1972b)) as $\partial_{NS}\overline{w}$ (the 'NS' stands for 'Natural Selection'), which corresponds to our notation above of $\Delta_{AG}\overline{w}$. The equivalence between 'due to changes in gene frequencies' and 'due to natural selection' has been much discussed and questioned by the authors of these and other papers, but Section 3 above sets out a strong case for it. The use of ∂ by Price reflects the continuous-time setting of this model, whereas it is less mysterious in the discrete setting employed here to use simply Δ_{NS}.

Our simplified version of Ewens and Lessard (2015)'s version of the theorem

is:

$$\Delta_{NS}\overline{w} = \frac{\sigma_A^2}{\overline{w}}. \tag{2}$$

Ewens and Lessard (2015)'s version looks the same, but the terms are defined more widely than here (allowing many alleles and loci). Fisher's version is much more general still as we saw above, but it looks different: as he uses continuous time, there is no denominator on the right hand side, and he didn't actually notate the left hand side at all. In words in our special case, we can say

$$\text{The increase in population } \overline{w} \text{ due to changes in gene frequencies} = \frac{\text{The additive genetic variance in fitness}}{\text{Mean fitness}}.$$

Ewens (1989) was the first to exhibit the denominator in the discrete-time version. This use of fitness works in our examples so far. If we employed Fisher's sense fully, the denominator would become one plus mean fitness (Grafen, 2015a).

This establishes what the theorem is in our special case, and shows how we can calculate the additive genetic variance. The examples will show simple natural selection at work, cases in which it is strikingly clear that not all the change in mean fitness should count as due to natural selection, that the additive genetic component usefully measures natural selection, and that the measurement of natural selection is made in the circumstances of the moment, and does not look ahead to changes in the population constitution, even when those changes are brought about by natural selection itself.

5.2. Linear versus non-linear traits

The computations in the present paper were all performed in Mathematica 11.1, Wolfram Research, Champaign, IL, USA, and used arbitrary precision numbers, so that the accuracy was tracked of each number produced. The workbook is available as supplementary information. In general, approximately one digit of accuracy was lost for each year. All the numbers reported in the tables are accurate to the precision shown. While checking numerical calculations, a few errors in the tables of Ewens and Lessard (2015) were noted. For example, in the top row of their Table 1, the value for $\sigma_A^2/\overline{w}$ is given as 0.011429, but this is twice the correct value of 0.005714, which moreover does equal σ_A^2 divided by $\overline{w}$ as given in the same row of the table. Thus, a general caution is appropriate in reading their tables, though there seems to be no consequent error in the text of the paper. We now turn to contrast linear with non-linear traits.

Ewens and Lessard (2015)'s examples begin with the simple diploid locus of the previous section, and vary three things. First, the set of initial frequencies of the genotypes, p_{11}, p_{12} and p_{22}. Second, the fitnesses w_{11}, w_{12} and w_{22},

which are shown in the example of Figure 3 to lie in a straight line. Third, they intriguingly follow Fisher (1941) in considering two mating systems: first, the usual random mating, and second, automictic selfing, in which two sexually produced gametes from the same individual fuse to form an individual in the next year. Ewens and Lessard (2015) give four tables of results. Here, we begin with their Example 1, which has initial genotype frequencies of $(0.2, 0.4, 0.4)$ and fitnesses of $(1.1, 1.0, 0.9)$. We begin with the random mating case. First notice that the fitnesses are in a straight line, and so the slope of the best fitting straight line is clearly -0.2. The points lie on the line, so the additive genetic variance is just equal to the total genetic variance.

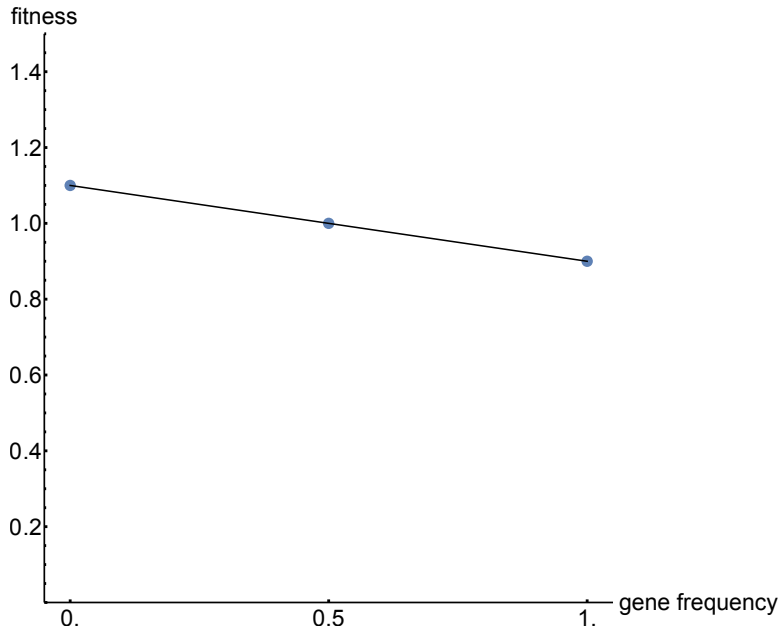

Figure 3: Purely additive genetic effects on fitness in Ewens and Lessard (2015)'s Example 1. The points show the fitness values assumed in Example 1. Here, the straight line passes exactly through the points, and the weights therefore make no difference. The slope is -0.2. The non-additive components of the fitnesses are all zero.

When we run the simulation, the most obvious thing to happen is that allele A_1 increases in frequency: it is more advantageous than A_2. As it does

Table 3: Ewens and Lessard's Example 1 under random mating. The table shows statistics for successive years. The starting frequencies were $(0.2, 0.4, 0.4)$, and the trait-values were $(1.1, 1.0, 0.9)$ for the genotypes (A_1A_1, A_1A_2, A_2A_2). p_1 is the frequency of A_1; $\overline{w}$ and $\Delta\overline{w}$ are the mean fitness and the change in it to the next year; $\Delta_{AG}\overline{w}$, $\Delta_{NAG}\overline{w}$ and $\Delta_{PGL}\overline{w}$ are the additive genetic, non-additive genetic and phenotype-genotype link components of the change in mean fitness; $\Delta\overline{x}$ is the change that would take place if the trait values were for height not fitness, and if the all fitness values were equal.

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	0.9800	0.00571	0.00571	0.	0.	0.
0.429	0.9857	0.00497	0.00497	0.	0.	0.
0.453	0.9907	0.00500	0.00500	0.	0.	0.
0.478	0.9957	0.00501	0.00501	0.	0.	0.
0.503	1.0007	0.00500	0.00500	0.	0.	0.
0.528	1.0057	0.00496	0.00496	0.	0.	0.

so, the mean fitness increases. Clearly the increase in the frequency of A_1 is caused by natural selection, but of course the genotype frequencies are also being dictated by Mendel's rules. However, because of the linearity of the fitted line, we can work out the average fitness of the population by knowing only the gene frequency next year – how it is distributed between the genotypes won't affect the mean fitness. Thus, we can say that *all* of the change in mean fitness is caused by the change in gene frequencies, and that all of the change in mean fitness is brought about by natural selection. Relevant numbers are shown in Table 3, which confirm that the total change equals Fisher's partial change, as we have just identified it should. Note that the change in mean fitness is largest in the first year, and it then falls before growing again. This is because the population is not in Hardy-Weinberg proportions, and the initial frequences of $(0.2, 0.4, 0.4)$ are more spread out than the Hardy-Weinberg frequencies with the same gene frequencies, which are $(0.16, 0.48, 0.36)$. 'More spread out' implies higher variance, and the fundamental theorem tells us that the improvement due to natural selection equals the variance. After the first year, we begin each year in Hardy-Weinberg proportions, and the variance then increases because the gene frequency is approaching 0.5. Once it passes 0.5, the variance reduces each year - the actual peak of fitness change occurs just before 0.5 because $\Delta_{AG}\overline{w}$ equals the variance divided by mean fitness, which is slowly increasing over time.

The second simulation for Example 1 assumes the extreme mating system of selfing. Selfing systematically reduces the frequency of heterozygotes, which throw off homozygotes of both kinds, and so the genotype frequencies vary in

Table 4: Ewens and Lessard's Example 1 under selfing. As Table 3, except there random mating is assumed.

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	0.9800	0.00571	0.00571	0.	0.	0.
0.429	0.9857	0.00787	0.00787	0.	0.	0.
0.468	0.9936	0.00898	0.00898	0.	0.	0.
0.513	1.0026	0.00945	0.00945	0.	0.	0.
0.560	1.0120	0.00948	0.00948	0.	0.	0.
0.607	1.0215	0.00921	0.00921	0.	0.	0.

a systematic way, eventually ending up with the whole population at the single fittest homozygote. Although this is more complicated than random mating, it is still true that the fitted line for fitnesses is linear, and that natural selection causes all of the change in mean fitness, even though the mating system has done something different with the genotype frequencies in the two cases. See Table 4.

The tables show how important quantities of interest vary over the years for both mating systems. The first column shows the frequency of A_1, then the mean fitness is given along with its change from that year to the next, followed by the change's partition into additive genetic, non-additive genetic and phenotype-genotype link components. The final column looks at how the trait would change if it were an ordinary trait like height, and if the fitnesses were all equal. In this example, there is no change in the mean height, and so the values are all identically zero. The linearity of fitnesses (and heights for the final column) imply that the non-additive genetic and phenotype-genotype link components, and the change in mean height, are all exactly zero. All change in mean fitness is due to natural selection.

Table 5: Ewens and Lessard's Example 2 under random mating. As Table 3, except that the fitness of A_1A_1 is 1.2 not 1.1.

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	1.0000	0.00736	0.01143	−0.00407	0.	−0.00400
0.440	1.0074	0.01027	0.01014	0.00012	0.	0.00000
0.475	1.0176	0.01080	0.01067	0.00013	0.	0.00000
0.511	1.0284	0.01123	0.01110	0.00013	0.	0.00000
0.548	1.0397	0.01156	0.01142	0.00014	0.	0.00000
0.585	1.0512	0.01174	0.01160	0.00013	0.	0.00000

Ewens and Lessard (2015)'s Example 2 introduces changes in mean fitness

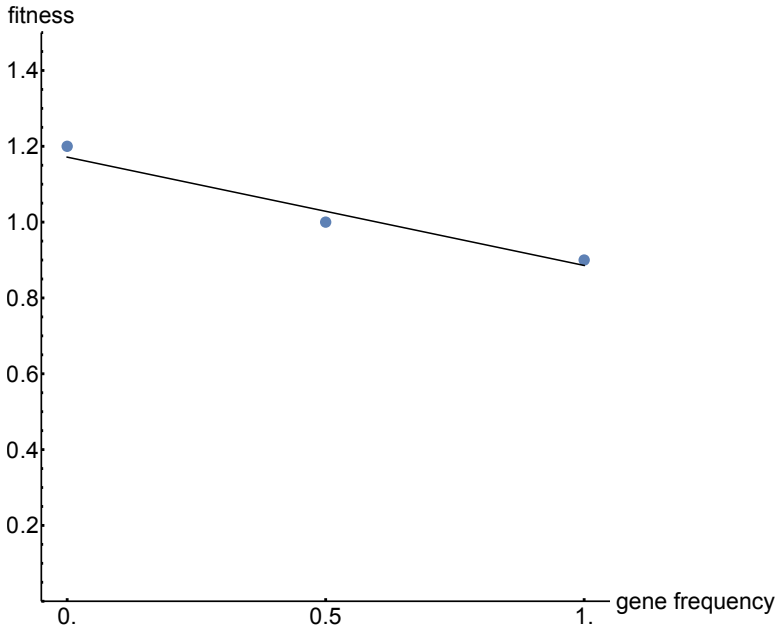

Figure 4: Genetic effects on fitness in Ewens and Lessard (2015)'s Example 2 include non-additive effects. The points show the fitness values assumed in Example 2. The straight line is the best fitting straight line through the points, weighting by the frequencies of the genotypes. The best-fitting slope will always lie between the slope based on the left two points (-0.4) and on the right two points (-0.2).

that are *not* due to natural selection. The only difference from Example 1 is that the fitnesses are $(1.2, 1.0, 0.9)$, so only the fitness of A_1A_1 has been changed, from 1.1 to 1.2. This matters because the fitnesses no longer lie in a straight line, as shown in Figure 4. No longer does the mean gene frequency on the x-axis tell us by itself what the mean fitness on the y-axis will be, because the distribution across the x-values matters and not just the mean, as we saw in Section 3. How does the example work out? Of course, A_1 increases even faster in frequency now, and natural selection is stronger. But there is another force affecting mean fitness.

To understand this extra force, we can examine the $\Delta\overline{x}$ column. This is the change in mean height (assuming the trait values represent height), in the

absence of fitness differences. In the first year, this change is negative, as have just seen, and it equals −0.00400. Thereafter, the population is in Hardy-Weinberg proportions, and so this value reverts to zero. Accordingly, we see that the change in mean fitness has this same reduction, so the effect of natural selection is added together with this non-additive effect to obtain the final total change in mean fitness. The −0.00400 is clearly not an effect of natural selection, and any reasonable way of measuring natural selection would exclude it. This difference is evidently caused by how the alleles are segregated into genotypes, and how mating changes that segregation, and not by natural selection. This is the first example in which two different causal factors contribute to the change in mean fitness.

Table 6: Ewens and Lessard's Example 2 under selfing. As Table 5, except that there random mating is assumed.

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	1.0000	0.02200	0.01143	0.01057	0.	0.01000
0.440	1.0220	0.02223	0.01695	0.00528	0.	0.00500
0.497	1.0442	0.02198	0.01943	0.00255	0.	0.00245
0.562	1.0662	0.02108	0.01987	0.00120	0.	0.00117
0.628	1.0873	0.01952	0.01896	0.00055	0.	0.00055
0.691	1.1068	0.01746	0.01721	0.00025	0.	0.00025

It is illuminating to compare to the selfing case shown in Table 6. As natural selection operates through offspring number, there should be no difference in the first year between the force of natural selection in the two mating systems, and indeed $\Delta_{AG}\overline{w}$ is the same. The total change in mean fitness is much higher under selfing (0.022 vs 0.007), and this is indeed explained by a large positive value of $\Delta\overline{x}$. The increase in homozygosity would have increased the mean value of x by 0.01, and added to the natural selection effect this gives a total of 0.021, very close to the actual change. Note that $\Delta\overline{x}$ continues to be non-zero, because as the gene frequencies change under selection, in each year reproduction with equal fitnesses would result in a change of genotype frequencies, as the heterozygote continues to throw off both homozygotes. However, as the heterozygote becomes very rare, this effect will vanish in later years.

The effect of natural selection and the change in the mean height do not add up to give the exact total change in the simulation. This is because the change in mean height is calculated here at the starting parental genotype frequencies, but we could also have calculated it at the offspring genotype frequenices. Neither is exactly right, but somewhere in between would be. Thus, the small proportional discrepancy should not disturb us.

We round off our discussion of the first and second sets of examples. In the

first, only natural selection is at work, and Fisher's theorem exactly matches the change in $\overline{w}$. In the second example, segregation was clearly an important force, and should certainly not be counted as natural selection. By measuring the change in mean height, we measured the size of the change due to segregation, and this showed that Fisher's 'change due to natural selection' matched very well the total change in $\overline{w}$ with the effect of segregation netted off. Thus, there is a clear biological sense to identifying some of the change in $\overline{w}$ as due to natural selection and, so far, in admittedly limited circumstances, Fisher's calculation has performed well. We now turn to incorporating density-dependence into the calculations.

5.3. The effect of density dependence

In this section we approach a more Fisherian definition of fitness in the simplest case where it matters, namely bland density dependence. The mathematics is simple, as having calculated the mean fitness $\overline{w}$ we simply calculate the fitnesses of the genotypes as relative to $\overline{w}$. As we will continue to use w for fitness, as we make it more Fisherian, we will use n for number of offspring, for Ewens and Lessard (2015)'s parameters. So, their offspring numbers will from now on be denoted n_{11}, n_{12} and n_{22}, with a mean of $\overline{n}$. The fitnesses themselves are defined from now on as $w_{11} = n_{11}/\overline{n}$, etc. Tables 7 and 8 show how this changes the important output statistics for the random mating and selfing cases of Example 2. The mean fitness is now forced to be 1 in every year, and it follows that the change in mean fitness is always zero. However, natural selection obviously operates in just the same way whether we have bland density dependence or not. Thus $\Delta\overline{w}$ is always zero: natural selection makes a positive contribution, which is then balanced out by the phenotype-genotype link component, which has so far been zero. This makes perfect sense, as the fitness changes are due to density dependence, which makes sure the population size does not change, thus the formality in the table exactly matches the reality in the population.

We see in Table 7 that $\Delta_{AG}\overline{w}$ has the same jump due to the relaxation into Hardy-Weinberg proportions as noted above in Table 5. The NAG component begins as negative, as it includes the negative $\Delta\overline{x}$ caused by this relaxation, but then becomes slightly positive. This is because there is a small change in gene frequency caused by selection, and the accompanying changes in genotype frequencies would have changed mean fitness even if the genotype-specific fitnesses had remained the same. $\Delta\overline{x}$ also shows same pattern, reducing to zero in the sexual case, because it is calculated on the basis of no fitness differences and so no change in gene frequencies, but roughly halving in the selfing case (with the NAG component following suit) in Table 8. The PGL component balances all these forces in keeping with the assumption of constant population size. The increase in $\Delta_{AG}\overline{w}$ is caused by the heterozygotes diminishing, which leads to an

Table 7: Ewens and Lessard's Example 2 under random mating, with bland density dependence. As Table 5, except for the density dependence

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	1.0000	0.00000	0.01143	−0.00407	−0.00736	−0.00400
0.440	1.0000	0.00000	0.01007	0.00012	−0.01019	0.00000
0.475	1.0000	0.00000	0.01048	0.00013	−0.01061	0.00000
0.511	1.0000	0.00000	0.01079	0.00013	−0.01092	0.00000
0.548	1.0000	0.00000	0.01098	0.00013	−0.01111	0.00000
0.585	1.0000	0.00000	0.01104	0.00013	−0.01117	0.00000

increase in variance as the population becomes more extreme. This will eventually reduce again, as the population concentrates on the fitter homozygote (A_1A_1) and the variance eventually diminishes to zero for that reason.

Table 8: Ewens and Lessard's Example 2 under selfing, with bland density dependence. As Table 7, except there random mating is assumed.

p_1	$\overline{w}$	$\Delta\overline{w}$	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$
0.400	1.0000	0.00000	0.01143	0.01057	−0.02200	0.01000
0.440	1.0000	0.00000	0.01658	0.00517	−0.02175	0.00489
0.497	1.0000	0.00000	0.01861	0.00245	−0.02105	0.00234
0.562	1.0000	0.00000	0.01864	0.00113	−0.01977	0.00110
0.628	1.0000	0.00000	0.01744	0.00051	−0.01795	0.00051
0.691	1.0000	0.00000	0.01555	0.00023	−0.01577	0.00023

The difference in the behaviour of $\overline{w}$ under density dependence points up a gulf between biologists and mathematical population geneticists. The latter seem fixed on $\overline{w}$ and the change in $\overline{w}$, so this model must look very different to them. However, for biologists the bland density dependence changes nothing about the action of natural selection. Adaptations would arise in same way and at the same speed. The search for *completing* Fisher's partial change to the total change in mean fitness, begun by Kimura (1958), is pointless under bland density dependence, as we know in advance the answer is zero: yet this certainty tells us nothing about the action of natural selection. Fisher's partial change in mean fitness is the right guide when studying natural selection as opposed to evolution as a whole. In the next section, we turn to a more interesting reason for the fitnesses of genotypes to change over the years.

5.4. Less unrealistic fitnesses: the hawk-dove game

The phenotype-genotype link component was non-zero, and generally negative, under bland density dependence. Here we study how the three-way partition behaves when the population is playing an evolutionary game. The hawk-dove game has been implemented alongside diploid genetics since at least Maynard Smith (1982, page 40), and we repeat these calculations, and add only how our three-way partition illuminates the operation of natural selection. Suppose individuals play the hawk-dove game (Maynard Smith and Price, 1973), and have their offspring number determined by a baseline fitness of 5 plus the payoff from the game with $v = 2.5$ and $c = 5$. We suppose homozygotes A_1A_1 play Hawk and A_2A_2 play Dove, while the heterozygotes in the first instance play the two strategies with half a chance each. There is no mystery about what happens in this case (Maynard Smith, 1982). The genes change frequency until the phenotypes are in the equilibrium frequencies that produce equal payoffs to Hawk and Dove. Table 9 shows how the analysis of fitness change proceeds, but first, the information provided in these tables has changed.

Table 9: The hawk-dove game. The game parameters were that baseline offspring number is 5, $v = 2.5$, and $c = 5$. Bland density dependence and random mating are assumed. The starting frequencies were $(0.05, 0.1, 0.85)$ for the genotypes (A_1A_1, A_1A_2, A_2A_2). The chances of playing Hawk were $(1, 0.5, 0)$. As $\overline{w} = 1$ and $\Delta\overline{w} = 0$ at all times, these columns have been omitted. Not included in previous tables are p_H, the frequency of the phenotype Hawk, and α_w and α_H, the average effect of substituting A_1 for A_2 on fitness and on the probability of playing Hawk, respectively.

p_1	p_H	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$	α_w	α_H
0.100	0.100	0.001677	0.000000	-0.001677	0.000000	0.0803	0.5000
0.110	0.110	0.001204	0.000000	-0.001204	0.000000	0.0783	0.5000
0.118	0.118	0.001229	0.000000	-0.001229	0.000000	0.0768	0.5000
0.126	0.126	0.001248	0.000000	-0.001248	0.000000	0.0753	0.5000
0.134	0.134	0.001262	0.000000	-0.001262	0.000000	0.0736	0.5000
0.143	0.143	0.001270	0.000000	-0.001270	0.000000	0.0720	0.5000

In the tables in this section, the columns for $\overline{w}$ and $\Delta\overline{w}$ have been suppressed, as it is now no surprise that with bland density dependence, these variables always equal 1 and 0, respectively. Three new columns are shown. The proportion of the population playing Hawk p_H shows how the phenotypes are changing, and the direction of selection. With our parameters, the equilibrium value is $p_H^* = 1/2$, and selection will favour Hawk and A_1 while p_H is less than a half, and favour Dove and A_2 when p_H is greater than a half. The average effect of A_1 for the trait 'probability of playing Hawk' is denoted α_H, and the average effect of A_1 for fitness is denoted α_w. (Fisher's concept of average effect is explained in the legend to Figure 2 as the effect of substituting one

A_1 for one A_2 on the given variable, according to the best fitting straight line.) α_H is always 1/2 in the first example, with codominance, because switching an A_2 to an A_1 will always increase the bearer's chance of playing Hawk by 1/2. However, dominance in the later tables will cause α_H to vary with gene frequencies, as the appropriate weights over the non-linear graph change the slope of the best-fitting straight line, as illustrated in Figure 2. The average effect for fitness will change with gene frequencies, and be positive (negative) when A_1 is associated statistically with higher (lower) fitness. It will be zero when $p_H = p_H^*$ because then both phenotypes, and therefore all three genotypes, have equal fitness, and the best-fitting straight line has zero slope.

Table 10: The hawk-dove game with dominance. As Table 9, except that the initial frequencies were reversed to $(0.85, 0.1, 0.05)$, and the chances of playing Hawk were $(1, 0, 0)$, corresponding to A_2 and Dove being dominant.

p_1	p_H	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$	α_w	α_H
0.900	0.850	0.004310	0.009444	-0.013753	0.007876	-0.1287	0.6538
0.883	0.780	0.003530	-0.000027	-0.003503	0.000000	-0.1308	0.8833
0.870	0.757	0.003034	-0.000023	-0.003011	0.000000	-0.1157	0.8698
0.857	0.734	0.002562	-0.000019	-0.002544	0.000000	-0.1021	0.8567
0.844	0.713	0.002134	-0.000015	-0.002119	0.000000	-0.0901	0.8441
0.832	0.693	0.001759	-0.000012	-0.001747	0.000000	-0.0794	0.8323

Returning to Table 9, with the results of the first Hawk-Dove simulation, the population starts with many fewer Hawks than the equilibrium value, and so there is comparatively strong selection for A_1 and Hawk. We see p_1 and p_H increasing. There is no NAG component because of codominance, and so the AG component is simply counter-balanced by the PGL component to produce a total of zero. Notice that the PGL component takes account of the changes in the fitness of each genotype that are caused by the change in the frequency of Hawks, but we do not count that as natural selection because of our requirement that natural selection is measured 'instantaneously'. The AG component drops a lot after the first year because of the relaxation into Hardy-Weinberg proportions, and then as in parallel circumstances above with fixed fitnesses, the AG component increases because the genotype frequencies produce increased variance as the frequency of A_1 increases to 1/2. Here, however, the move towards the equilibrium frequency of Hawk is making the fitnesses of Hawks and Doves more equal, and this counteracts the variance effect to some extent. α_H is fixed at 1/2 as noted in the previous paragraph, but α_w decreases, as the increasing similarity of fitnesses of Hawks and Doves causes a substitution of A_1 for A_2 to have a smaller effect on fitness.

We switch to dominance of A_2 for the next example in Table 10, and see that the NAG component has sprung to life. The phenotype-genotype link

Table 11: The hawk-dove game starting just below the equilibrium frequency of the phenotype Hawk. As Table 9, except that the initial genotype frequencies, approximately $(0.067, 0.383, 0.550)$, were chosen to make $p_H = 0.45$ (0.05 below the equilibrium value) and to be in Hardy-Weinberg proportions, and the chances of playing Hawk were $(1, 1, 0)$, corresponding to A_1 and Hawk being dominant. The equilibrium being approached would have all column values equal to zero except for $p_1 = 1 - \sqrt{1/2} \approx 0.293$, $p_H = 1/2$, and $\alpha_H = \sqrt{1/2} \approx 0.707$.

p_1	p_H	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$	α_w	α_H
0.258	0.450	0.000100	0.000000	-0.000100	0.000000	0.0161	0.7416
0.261	0.455	0.000083	0.000000	-0.000082	0.000000	0.0146	0.7385
0.264	0.459	0.000068	0.000000	-0.000068	0.000000	0.0133	0.7357
0.267	0.463	0.000057	0.000000	-0.000056	0.000000	0.0120	0.7331
0.269	0.466	0.000047	0.000000	-0.000047	0.000000	0.0109	0.7308
0.271	0.469	0.000039	0.000000	-0.000039	0.000000	0.0099	0.7286

component, because we continue to assume bland density dependence, exactly compensates for the other two components to ensure a zero change in mean fitness overall. This simulation begins on the other side of the equilibrium, and the frequencies of A_1 and Hawk are coming down. The fitnesses of Hawk and Dove are getting more similar over the years, and the induced variance-reducing effect more than compensates for the variance-increasing effect of moving towards more genetic variability in strategy: the AG component drops throughout, and the relaxation into Hardy-Weinberg proportions slightly reduces the first drop. The value of $\Delta\overline{x}$ shows the effect of relaxation into Hardy-Weinberg proportions, which makes a positive NAG component in the first year, overwhelming the negative value that shows up in later years. Whereas the NAG component after the first year was positive in earlier tables, here it is negative, and whether it assists or impedes the action of natural selection depends on genetic details such as dominance.

The next two Tables 11 and 12 show starting positions chosen to be much closer to the equilibrium value of 0.5 Hawk phenotypes. Both tables have the opposite dominance to the previous case. The changes in the frequencies of A_1 and Hawk are much smaller, as the fitness differences are now much smaller. These simulations were also begun in Hardy-Weinberg proportions, so there is no initial surge in the NAG component, which is what allows it to be zero to the accuracy of table, and therefore omitted.

The average effect on probability of playing Hawk is converging to its equilibrium value, which is not close to zero, as the phenotype differences continue to cover the range between 0 and 1. However, the average effects of fitness become much smaller, and are of opposite sign in the two tables because they are on opposite sides of the equilibrium, and will become zero at the equilibrium.

Table 12: The hawk-dove game starting just above the equilibrium frequency of the phenotype Hawk. As Table 11, except that initial frequencies, approximately $(0.108, 0.442, 0.450)$, were chosen to ensure $p_H = 0.55$ (0.05 above the equilibrium value).

p_1	p_H	$\Delta_{AG}\overline{w}$	$\Delta_{NAG}\overline{w}$	$\Delta_{PGL}\overline{w}$	$\Delta\overline{x}$	α_w	α_H
0.329	0.550	0.000103	0.000000	-0.000103	0.000000	-0.0153	0.6708
0.326	0.545	0.000085	0.000000	-0.000085	0.000000	-0.0139	0.6742
0.323	0.541	0.000070	0.000000	-0.000070	0.000000	-0.0127	0.6772
0.320	0.538	0.000058	0.000000	-0.000058	0.000000	-0.0116	0.6800
0.317	0.534	0.000048	0.000000	-0.000048	0.000000	-0.0105	0.6825
0.315	0.531	0.000040	0.000000	-0.000040	0.000000	-0.0096	0.6848

The third and fourth tables show slower natural selection than the first two, because the additive genetic variance in fitness is lower. Had we gone even closer to the equilibrium, the changes would at some stage have become zero to the accuracy of the tables.

The whole understanding that natural selection operates at the level of individuals, and that mean fitness can go down as a result of natural selection operating in a frequency-dependent environment was fully and articulately explored by Dawkins (1976) some decades ago under the chapter heading 'The Conspiracy of Doves': if all individuals agreed to play Dove, then the mean population fitness would be higher, but natural selection will favour a Hawk mutant, and the frequency of Hawk will rise in a population of Doves. Of course, this sense of 'fitness' is one that applies before bland density dependence is implemented.

We now consider what all these examples tell us.

5.5. Lessons learned from the examples

We recap the important conclusions that the examples have shown in a concrete way, without the numerical details. The total evolutionary change in $\overline{w}$ from one year to the next is shown to be influenced by two forces apart from natural selection. Tables 5 and 6 show that segregation of alleles changes mean fitness, both through relaxation into Hardy-Weinberg proportions and through the automictic production of homozygotes, in ways that should clearly not be regarded as caused by natural selection. Equally, bland density dependence affects mean fitness from Table 7 onwards, by ensuring it is always equal to 1, and this force is also quite distinct from natural selection. If we wish to study natural selection, therefore, we need to find a way of measuring just the part of the change in $\overline{w}$ that is due to natural selection. In all the examples, $\Delta_{AG}\overline{w}$ has fulfilled this role, matching biological intuition about when and how much it acts. The neutrality principle, that with constant fitnesses throughout a population natural selection does not act, helped us identify which causes should count as natural selection.

The measurement of the effect of natural selection takes no account of how future fitnesses are altered as a result of the change in gene frequencies this year. We saw above why, in principle, that was a good idea. In practice, it means we consider only the fitnesses this year and how they change gene frequencies this year. This 'instantaneous' nature of natural selection, and of measuring it, is important. It holds true in structured populations too, when the reproductive values employed for weighting are derived from the schedule of birth and death rates this year, even though one effect of natural selection is likely to be to change those schedules. Interestingly, although Fisher understood that selection changes mean fitness as did other forces, his discussion in the sections following the statement of the fundamental theorem never clearly comes to terms with the fact that the very definition of fitness (at least, the reproductive value weights) are likely to be constantly changing.

The average effects presented in Tables 9 to 12 change in meaningful ways in relation to the state of the population, in particular to the frequency of Hawk, as they will generally do with frequency-dependent selection. They show how the selective pressures are acting on the alleles. The fact that average effects change is sometimes raised as problematic for understanding the fundamental theorem (e.g. Ewens and Lessard, 2015). However, in combination with an understanding that natural selection should be measured instantaneously, these changes in average effects should be regarded as an opportunity for illumination rather than as a difficulty. Average effects are a way to discuss how natural selection operates in general, that can abstract from the genetic architecture.

The AG component of change is always positive (or very occasionally zero). The NAG component is zero when the link between individual gene frequency and fitness is linear, and otherwise can be positive or negative depending on the genetic details. Tables 3, 4 and 9 have a zero NAG component because of linearity. Tables 5 to 8 have a positive value (after the first year, when relaxation into Hardy-Weinberg proportions sometimes makes it negative). Table 10 shows a negative NAG effect after the first year, illustrating that whether NAG assists or opposes natural selection is down to genetic details. The PGL component is zero when the fitnesses of genotypes do not change from year to year, but under bland density dependence, it adjusts so that the sum of all three components equals zero, ensuring that $\overline{w} = 1$ and $\Delta\overline{w} = 0$. Usually, this makes the PGL component negative, as it must oppose the non-negative AG component; though it could be positive if the NAG effect were sufficiently negative to more than balance the positive AG component. The AG component equals the additive genetic variance of fitness divided by the mean fitness, says the discrete-time version of the fundamental theorem. We saw in the examples how a large variance in fitness did lead to a large AG component, and to larger gene frequency changes. These patterns justify the description of the biological meaning of the

partition components in Table 2.

Thus, the observed patterns make sense of the three-way partition and of the fundamental theorem's way of measuring natural selection. Exploring the hawk-dove game itself which would lead in other directions.

6. The disappointments of Price (1972b)

Readers who are unfamiliar with the preceding literature on the fundamental theorem may understandably wish to skip this section, but it is important to state how the clarifications here go beyond what has been previously published. Price (1972b) made the first serious study of Fisher's argument, largely reconstructing it, but expressed his disappointment with the result in the following terms, in which he uses M for mean fitness :

> One defect is the device of treating non-additive gene effects as 'environment'. ... A much graver defect is the matter of the shifting standard of 'fitness' that gives the paradox of M tending always to increase and yet staying generally close to zero. Much more interesting would be a theorem telling of increase in 'fitness' defined in terms of some fixed standard. Thus there is a challenge here to find a deeper definition of this elusive concept 'fitness' and to give a deeper and sharper explanation of why it increases and under what conditions. (pages 139–140 Price, 1972b)

where recall that Fisher's mean fitness is zero in static populations. These two criticisms, along with the final hope, are the main criticisms of the meaning of the fundamental theorem in the subsequent literature (e.g. Ewens, 1989, 1992; Frank and Slatkin, 1992; Frank, 1997, 1998; Ewens and Lessard, 2015). We take them in turn.

The three-way partition of Section 3 shows why the non-additive genetic component should certainly not count towards the effect of natural selection, on the grounds that this component can be non-zero even in the absence of fitness differences. Price complains that it counts as 'environment', but what it is called, once it is agreed that it should not constitute part of natural selection, is of little importance. Fisher's thinking may have been coloured by the statistical frame of mind. The additive genetic component of a trait is the part predicted by a particular linear regression model, and in statistics all the other forces are lumped together into the 'error' or 'residual'. As genuine environmental variation is also included in this error term, it would be convenient and of little matter to include non-additive genetic effects in the same category. Frequency-dependent effects, in which the frequency of a gene affects the fitnesses and average effects, can naturally be understood as part of the 'environment' of a gene. It is true that the non-additive genetic component is harder to interpret

in this way, but the crucial conclusion is not threatened that it is the additive genetic component that represents the effect of natural selection. (Frank, 1997, Appendix A) gives an example with all fitnesses equal and therefore no selection, but does not explicitly distinguish the additive and non-additive components of the change in mean trait.

The second 'shifting standard' criticism is much repeated. When expressed as assuming average effects are fixed, this could give the unfortunate impression that Fisher required an additional unstated assumption for his theorem, beyond the absence of mutation and the absence of meiotic drive. Ewens and Lessard (2015, pages 62–63) discuss changing average effects, and repeatedly assert that the fundamental theorem therefore omits 'an important aspect of the evolutionary process' and similar phrases. They do not discuss the goal of isolating the effect of natural selection from other causes, and it is almost a necessary consequence of doing so that some aspects of the evolutionary process will be omitted – indeed, that could be seen as the whole point. But what is Price looking for when he makes this second criticism? The issue here is that he dreams of something that may be entirely illusory. If you want a result that describes natural selection as an improving process and tells you what is improved, then the fundamental theorem is that result. It employs the 'instantaneous nature' of natural selection and of its measurement. Price regards the changeability of the standard of fitness as a problem, but this is to wish away the biology and seek to replace it with mathematical simplicities. A biologist must surely embrace the changeability of the standard of fitness, and regard the fundamental theorem as appropriately mirroring the biological properties of natural selection, which are irretrievably contingent. An organism will be designed for the biological world in which it lives, and the arrival of a new competitor species, or a spreading mutation that changes an existing species of pathogen, will alter the world, and alter what counts as good design in it.

The final hope in the quotation from Price above is swept away by understanding the messiness and unpredictability of biology, and the impossibility of matching them against a simple mathematical standard, but there is one particular form of the hope that is very relevant to understanding the history of the reception of the theorem. Mathematical population geneticists are usually mathematicians who specialise in dynamical systems, and their commonest *modus operandi* is to define a dynamical system representing genotype frequency changes that they then analyse to discover the resulting dynamical behaviour. For them, the fundamental theorem is a will-o'-the-wisp. At first sight, a quantity that must always increase is of enormous significance, and some time was spent in the 1950s and 1960s pursuing the idea of fitness as a Lyapunov function or gradient function (see Edwards, 1994, pages 469–470). The changeability of fitness makes it wholly unsuitable for this mathematical purpose, but to com-

plain about average effects changing as a result of frequency-dependent selection, or non-additive genetic effects, is to miss the larger point that they will change when the environment changes, which it continually does. The theorem is, therefore, always liable to disappoint students of dynamical systems, whatever biological insights it may have to offer.

7. Discussion

What is the current status of the fundamental theorem of natural selection? It is like the air breathed by whole-organism biologists, including ethologists, animal behaviourists, and behavioural ecologists, in that they assume natural selection is an improving process, and that it improves fitness. Further, they usually admit to not knowing exactly what fitness is, but rightly regard it as similar to lifetime number of offspring, and in complicated cases involving reproductive value. Although this is a matter for historians of science, I suggest that the source of this assumption is Fisher (1930)'s *The Genetic Theory of Natural Selection* – but not directly. Mediated through Hamilton (1964) and Williams (1966), both of whom were directly inspired by Fisher and by his fundamental theorem, biologists have taken for granted that a numerical quantity called fitness is increased by natural selection. Thus, Fisher would presumably look at biology today and be pleased that his ideas have taken hold so firmly, while being puzzled that the conceptual core of his work, the fundamental theorem, is not often cited as the source.

Thus, in a history-of-ideas sense, the fundamental theorem is already fundamental to whole branches of modern biology. On the other hand, in a philosophy-of-ideas sense, the theorem's current status as currently understood in its formal expression, is very low. Mathematical population geneticists, who might be expected to be best positioned to understand it, continue to express uncertainty and doubt about the meaning of the fundamental theorem, most recently and authoritatively Ewens and Lessard (2015). The most obvious issue to emerge is what are the implications of the history vs philosophy discordance, and how should it be resolved in light of the new articulation of the theorem above?

First, it is significant that mathematical population geneticists and whole-organism biologists have different interests and approaches. Ewens and Lessard (2015) make no mention of distinguishing natural selection from the whole effect of evolution, and yet the analysis above makes clear that this was Fisher's main purpose. Thus, the failure of mathematical population geneticists to support what biologists need from the fundamental theorem should not be given much weight. A new kind of analysis, employing average effects and linking them to changing selective pressures, would explore the theorem more effectively from a strictly biological point of view. This understanding of the mathematical population geneticists' position, combined with the clarification of the left hand

side above, frees us to conclude that the fundamental theorem can reasonably be cited for the claim that natural selection is an improving process, and that the quantity it improves is fitness. It is vital to this conclusion that Fisher's argument provides a methodology for defining fitness, as well as defining it in the particular setting of the fundamental theorem itself.

The distinction between natural selection improving fitness, on the one hand, and causing organisms to act as if maximising their fitness, on the other, is very important here. Most whole organism biologists are happy to adopt the stronger maximising position, at least as a working hypothesis, though doubts are often articulated (for a small sample, see e.g. Lehmann and Rousset, 2014; Orzack, 2014). The additional arguments to move from one to the other will not be considered here, but are obviously worth investigating for a fuller reconciliation of current biological practice with Fisher's theorem.

Fisher himself compared his theorem to the Second Law of Thermodynamics, and the central place of fitness in modern whole-organism biology certainly supports this comparison. But a theorem that says 'quantity Q increases', or relevantly in both cases, 'the change in Q brought about by a specific cause is non-negative' has little of interest in the statement itself — the whole work and interest is in constructing the quantity Q, whether that is entropy or fitness. This is obscured in the very simple models of the fundamental theorem that lack age structure or other population structure, as the obvious measure of fitness will work just fine. But with age structure, there is a lot of work to do to derive a concept of fitness that can take part in any theorem about increase. Thus, it is the accompanying definition of fitness that should be admired and recognised as the hallmark of genius in the fundamental theorem. It is surprising that there exists a suitable quantity that can be defined in the maelstrom of multilocus Mendelian genetics and population structure that produces an individually-defined value and represents quality of design. The current paper has worked with the simple models because clarifying the left hand side is, by itself, a substantial work of exegesis. Future explorations of the theorem could well focus on the nature of fitness.

The Nobel prize acceptance speech of Prigogine (1978) for his work on entropy makes fascinating reading. The Second Law of Thermodynamics is also about a partial change over time, and Prigogine describes its application as more a project than a theorem – that is, working with the principle is an aid to interesting enquiry, rather than simply a source of unquestioned truths. It will be interesting to see how the acceptance of the fundamental theorem in biology improves our means of enquiry. One fruitful avenue is defining fitness in more sophisticated models, incorporating environmental uncertainty, different types of population structure, and interactions between individuals who may be related. A start has been made to this project (Batty et al., 2014; Grafen,

2015b; Crewe et al., 2017). Thus, Fisher's analogy between the fundamental theorem and the Second Law is well made, and biologists would do well to take it seriously.

Why did it take so long for biologists to understand the fundamental theorem? This is a topic for historians of science, and may reward study. A first glance suggest that it was a complex idea, necessarily so for a complex topic. Biologists were too simple-minded in their reading of the theorem for the first few decades. Ethologists, animal behaviourists and behavioural ecologists have generally taken for granted the gist of the theorem, though often without knowing there is a theorem involved at all, and took comfort from it when pursuing projects based on fitness-maximisation (a controversial link not pursued in this paper). Mathematical biologists had developed their own approach to studying population genetics before they reached the sophistication required to understand the theorem. And, it must be admitted, following Price (1972b), that Fisher did explain it very badly indeed. One saving thought for Fisher is that he was so far ahead of his contemporaries that, no matter how good his explanation, it is doubtful anyone would have understood the theorem for forty or more years anyway. Fisher had simply bitten off vastly more than anyone else could, at the time, chew.

Some other puzzles that are not discussed in this paper may also concern the inquisitive student who figured in the introduction. For example, despite the claim that natural selection improves fitness, one obvious limit to this power occurs in the case of overdominance, of which the most prominent example is sickle-cell anaemia (Allison, 1954). The working of the fundamental theorem would be illuminated by analysing that example, and others, within its formalism. The most obvious gap in the theory to date is incorporating inclusive fitness ideas into the fundamental theorem framework, and reconciling these two major areas of theory is an obvious target for future work.

The first sentence of Fisher's Preface, quoted already above, is 'Natural Selection is not Evolution', and he goes on to set out his intention to create a theory of natural selection as opposed to evolution as a whole. The other main point in the Preface is the distinction between mathematical and biological imaginations, and the need for a substantial body of mathematical work in biology to match that of physics. One test of achieving that body of work is an articulated understanding of the fundamental theorem, and its acceptance as the central principle of evolutionary biology.

Acknowledgments

I am grateful to Prof. Philippe Huneman, for inviting me to present this work at a conference at l'Institut d'histoire et de philosophie des sciences et

des techniques in Paris in November 2017, and to the audience there for helpful comments.

Appendix A. The derivation of the three-way partition of $\Delta\overline{x}$

In broad outline, we follow Price (1972a). Frank (1998) made many partitions of this general kind, but not this one. The partition will be applied to an arbitrary quantitative trait x, of which fitness w will turn out to be merely a specially interesting case. The trait of genotype g is x_g, and $\hat{x}_g$ is the predicted value from the straight line. The deviations of the trait values for each genotype from additivity are $\tilde{x}_g = x_g - \hat{x}_g$. The regression is specified in Appendix Appendix B. We follow Ewens and Lessard (2015) in using the subscript g to represent a genotype – in the examples, g might equal 11, 12 or 22, but the formula applies much more widely than a single biallelic diploid locus. We add a prime to indicate 'next year', and agree to use Δ to mean the change from this year to the next. The partition is derived as follows:

$$\begin{aligned}
\overline{x} &= \sum p_g x_g \\
\overline{x}' &= \sum p'_g x'_g \\
\Delta\overline{x} = \overline{x}' - \overline{x} &= \sum p'_g x'_g - \sum p_g x_g \\
&= \sum (p_g + \Delta p_g) \cdot (x_g + \Delta x_g) - \sum p_g x_g \\
&= \sum x_g \Delta p_g + \sum (p + \Delta p_g) \Delta x_g \\
&= \sum \hat{x}_g \Delta p_g + \sum \tilde{x}_g \Delta p_g + \sum (p + \Delta p_g) \Delta x_g \\
&= \Delta_{AG}\overline{x} + \Delta_{NAG}\overline{x} + \Delta_{PGL}\overline{x}
\end{aligned}$$

The three components will be called the 'additive genetic component', the 'non-additive genetic component', and the 'phenotype-genotype link component'. This is an exact partition of $\Delta\overline{x}$, which applies with multiple loci and multiple alleles at each locus, and arbitrary dominance and epistasis. One significance is that, when the trait is fitness, the first component of the partition is the left hand side of the fundamental theorem. This partition is illustrated for the simple biallelic single locus case in Figure 1.

Appendix B. The linear regression defining additive genetic value

This appendix describes how the best fitting straight line is found predicting mean trait value from gene frequencies, and more generally the best fitting regression plane when there are more than two alleles. We begin with a set of genotypes indexed by g and a mean trait value for each genotype, say $\overline{x}_g$,

where the mean is taken over the population. Suppose we have alleles in a set T indexed by t, and that the gene frequency of allele t in genotype g is z_{tg}. z_{tg} sums to one over the subset of alleles at one locus.

The least-squares linear multiple regression has one datapoint for each genotype. The response variable is $\overline{x}_g$, and there is one regressor for each allele t, whose numerical values are given by z_{tg} – it is convenient not to include a constant in the model. The regression is weighted by p_g, the population proportions of the genotypes. The regressors are collinear, with one degree of collinearity for each locus. There may be further collinearity if certain allele combinations do not occur. The regression provides predicted values, which are unique despite the collinearity, which we call $\hat{x}_g$ to reflect the notation used in Figure 2. On the same basis, we call the residuals $\tilde{x}_g = \overline{x}_g - \hat{x}_g$.

The predicted values are the breeding values, which are also called the additive genetic component of the trait. The additive genetic variance is the variance of the predicted values.

The regression would become more complicated when there is population structure, and the interested reader is referred for details to Grafen (2015a,b), including for a reconstruction of Fisher's argument in this case. The explanation of the meaning of the left hand side of the theorem remains unaltered by this considerable sophistication of the concept of fitness. Note that an equivalent setup for finite populations has each individual as a datapoint in the regression, with reproductive value as weights in structured populations.

Finally, we look at why the change in mean from Panel 1 to Panel 2 in Figure 1 is zero. The means are $\sum p_g x_g$ and $\sum p_g \hat{x}_g$, so the difference is $\sum p_g \tilde{x}_g$. A standard property of residuals (Cox and Hinkley, 1974) is that the inner product with any linear combination of regressors equals zero. In this case, the constant (1 for every datapoint) equals the sum of the z_{tg} for all the alleles at any given locus. Thus the inner product of the residuals with a vector of 1s must be zero. $\sum p_g \tilde{x}_g$ is just that inner product, and so the means in Panels 1 and 2 are equal. Fisher had worked on additive and non-additive genetic variation (Fisher, 1918) and then spent the 1920s doing fundamental work in statistics including multiple regression, and presumably this equality was clear to him when he published the fundamental theorem in 1930, but to few others at the time.

Allison, A. C. (1954). Notes on sickle-cell polymorphism. *Annals of Human Genetics*, 19:39–57.

Batty, C. J. K., Crewe, P., Grafen, A., and Gratwick, R. (2014). Foundations of a mathematical theory of darwinism. *Journal of Mathematical Biology*, 69:295–334. doi: 10.1007/s00285-013-0706-2.

Bennett, J. H., editor (1983). *Natural Selection, Heredity and Eugenics (Includ-*

ing Selected Correspondence of R.A. Fisher with Leonard Darwin and others). Oxford University Press, Oxford.

Cosmides, L. M. and Tooby, J. (1981). Cytoplasmic inheritance and intragenomic conflict. *Journal of Theoretical Biology*, 89:83–129.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Crewe, P., Gratwick, R., and Grafen, A. (2017). Defining fitness in an uncertain world. *Journal of Mathematical Biology*. doi: 10.1007/s00285-017-1164-z.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Edwards, A. W. F. (1994). The fundamental theorem of natural selection. *Biological Reviews*, 69:443–474.

Edwards, A. W. F. (2014). R.A. Fisher's gene-centred view of evolution and the Fundamental Theorem of Natural Selection. *Biological Reviews*, 89:135–147.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Springer, Berlin, Heidelberg, New York.

Ewens, W. J. (1989). An interpretation and proof of the fundamental theorem of natural selection. *Theoretical Population Biology*, 36:167–180.

Ewens, W. J. (1992). An optimizing principle of natural selection in evolutionary population genetics. *Theoretical Population Biology*, 42:333–346.

Ewens, W. J. (2004). *Mathematical Population Genetics I. Theoretical Introduction*. Springer, Berlin, Heidelberg, New York.

Ewens, W. J. and Lessard, S. (2015). On the interpretation and relevance of the fundamental theorem of natural selection. *Theoretical Population Biology*, 104:59–67.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press. See Fisher (1999) for a version in print.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, 11:53–63.

Fisher, R. A. (1999). *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford, UK. Variorum Edition of 1930 OUP edition and 1958 Dover edition, edited by J. Henry Bennett.

Frank, S. A. (1997). The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution*, 51:1712–1729.

Frank, S. A. (1998). *The Foundations of Social Evolution*. Princeton University Press.

Frank, S. A. (2012). Natural selection. IV. The Price Equation. *Journal of Evolutionary Biology*, 25:1002–1019.

Frank, S. A. and Slatkin, M. (1992). Fisher's Fundamental Theorem of Natural Selection. *Trends in Ecology and Evolution*, 7:92–95.

Gayon, J. (1998). *Darwinism's Struggle for Survival*. Cambridge University Press, Cambridge, UK.

Grafen, A. (2015a). Biological fitness and the fundamental theorem of natural selection. *American Naturalist*, 186:1–14.

Grafen, A. (2015b). Biological fitness and the Price Equation in class-structured populations. *Journal of Theoretical Biology*, 373:62–72.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7:1–52.

Kimura, M. (1958). On the change of population fitness by natural selection. *Heredity*, 12:145–167.

Lehmann, L. and Rousset, F. (2014). Fitness, inclusive fitness and optimization. *Biology and Philosophy*, 29:181–195.

Lessard, S. (1997). Fisher's fundamental theorem of natural selection revisited. *Theoretical Population Biology*, 52:119–136.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Maynard Smith, J. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246:15–18.

Orzack, S. H. (2014). A commentary on "The Formal Darwinism Project": there is no grandeur in this view of life. *Biology and Philosophy*, 29:259–270.

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

Price, G. R. (1972a). Extension of covariance selection mathematics. *Annals of Human Genetics*, 35:485–490.

Price, G. R. (1972b). Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics*, 36:129–140.

Prigogine, I. (1978). Time, structure and fluctuations. *Science*, 201:777–785.

Williams, G. C. (1966). Natural selection, the costs of reproduction, and a refinement of Lack's principle. *American Naturalist*, 100:687–690.

Williams, G. C. (1975). *Sex and Evolution*, volume 8 of *Monographs in Population Biology (eds Simon A. Levin and Henry S. Horn)*. Princeton University Press, Princeton, New Jersey.