Volume 86, No. 1 March 2011

The Quarterly Review of Biology

MEASUREMENT AND MEANING IN BIOLOGY

David Houle*
Centre for Ecological & Evolutionary Synthesis, University of Oslo, 0316 Oslo, Norway and Department of Biological Science, Florida State University
Tallahassee, Florida 32306-4295 USA
e-mail: dhoule@bio.fsu.edu

Christophe Pélabon
Department of Biology, Centre for Conservation Biology, NTNU
N-7491 Trondheim, Norway

Günter P. Wagner
Department of Ecology & Evolutionary Biology, Yale University
New Haven, Connecticut 06405 USA

Thomas F. Hansen
Centre for Ecological & Evolutionary Synthesis, Department of Biology, University of Oslo
0316 Oslo, Norway

keywords
measurement theory, meaningfulness, dimensional analysis, scale type, scaling, significance testing, modeling

abstract

Measurement—the assignment of numbers to attributes of the natural world—is central to all scientific inference. Measurement theory concerns the relationship between measurements and reality;

*Authorship is on a nominal scale.

The Quarterly Review of Biology, March 2011, Vol. 86, No. 1

0033-5770/2011/8601-0001$15.00

its goal is ensuring that inferences about measurements reflect the underlying reality we intend to represent. The key principle of measurement theory is that theoretical context, the rationale for collecting measurements, is essential to defining appropriate measurements and interpreting their values. Theoretical context determines the scale type of measurements and which transformations of those measurements can be made without compromising their meaningfulness. Despite this central role, measurement theory is almost unknown in biology, and its principles are frequently violated. In this review, we present the basic ideas of measurement theory and show how it applies to theoretical as well as empirical work. We then consider examples of empirical and theoretical evolutionary studies whose meaningfulness have been compromised by violations of measurement-theoretic principles. Common errors include not paying attention to theoretical context, inappropriate transformations of data, and inadequate reporting of units, effect sizes, or estimation error. The frequency of such violations reveals the importance of raising awareness of measurement theory among biologists.

Introduction

PROGRESS IN science often involves quantification. Ideas progress from loose verbal accounts to become rigorous mathematical models. At the same time, concepts and entities progress from incomplete verbal definitions to become variables and parameters that derive their meaning from a precise theoretical context. For example, ecology developed from an unconnected set of verbal ideas in 1920 to a unified science with a rigorous foundation in mathematical population ecology by about 1970 (Kingsland 1985), and similar changes took place during the modern synthesis in evolutionary biology. Arguably, empirical progress results mainly from better measurement. Measurements improve either because of more rigorous theory, which defines what is important to measure, or better instruments, which allow more accurate measurements of familiar quantities or make the previously unknown measurable. The high status of quantification in science is therefore no surprise, nor is the desire of researchers to support their work with quantitative measurements.

For measurements to be meaningful, however, they must retain their connection to the theoretical and instrumental context from which they were derived. Measurement theory concerns the relationship between measurements and reality; its goal is ensuring that inferences about measurements reflect the underlying reality we intend to represent. Unfortunately, in biology, the connection between concepts and measurements is often lost during the measurement process. Quantitative measurements flourish, but they are often used and manipulated in ways that render the conclusions drawn from them meaningless. This conclusion is familiar to any participant in journal clubs or discussion groups.

Consider some brief examples: Dunn et al. (1999) devised an index of concern for the conservation of Canadian land birds by averaging ordinal indices for abundance, breadth of range, and evidence of population decline. Wolman (2006) pointed out that these averages are meaningless because ordinal indices do not reflect the magnitudes of the attributes they measure. Post and Forchhammer (2002) reported a negative correlation between climate and variation in abundance of caribou and musk ox on Greenland. Vik et al. (2004) showed that this result is meaningless because changing the units used to measure abundance could reverse this correlation. Diaz and Rützler (2001) reviewed studies of the importance of organisms on coral reefs, all of which used the area covered to measure abundance. Wulff (2001) pointed out that the most relevant attribute is biomass, and measuring area dramatically underestimates the biomass of sponges relative to encrusting organisms because sponge biomass scales with volume. Harvey and Clutton-Brock (1985) compiled a widely used source of primate body-size data. They presented species means without standard errors, sample sizes, or references to the source of the data. Smith and Jungers (1997) traced the source of these data and found so many errors and inaccuracies that they concluded that “the data table . . . should *never* have been acceptable as a source" (p. 549). We will make the case that such errors are not isolated mistakes or differences of opinion, but systemic errors in the scientific process that result from the absence of a theory of measurement and meaning in biology.

Scientific representation of biological reality also extends to the use of mathematical models to capture biological relations. Mathematical models are representations of hypotheses about reality, and their form and manipulation must be consistent with the reality they are meant to represent. Measurement in the usual sense assigns numbers to aspects of reality, whereas mathematical modeling assigns functions to aspects of reality. This representational aspect of modeling is often forgotten when specific models are used to represent general hypotheses. A typical example is Maynard Smith's (1976) claim, based on a highly specific population-genetic model, that Zahavi's handicap principle, in which individuals display costly traits to enforce honest signals of quality, cannot work. This claim was not a mathematical error but a representational error, because the specific functional forms assumed by Maynard Smith were not a complete representation of the universe of cases inherent in the handicap hypothesis. Pomiankowski (1987, 1988) and Grafen (1990a,b) showed that handicaps can evolve under reasonable conditions on the basis of models with more general functional forms.

Our purpose here is to bring a broad measurement theory framework for understanding the relationship between meaning and measurement to the attention of biologists. We believe that awareness of measurement theory helps us to do better science by providing tools to ensure the meaningfulness of our work. Our starting point is the field of representational measurement theory, a well-established mathematical approach that enables us to determine whether the relationships among numerical measurements are consistent with the relationships among the attributes they are meant to represent. Representational measurement theory is virtually unknown in biology, so our education in the theory has come mostly from nonbiological sources (Stevens 1946, 1968; Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990; Hand 1996, 2004; Sarle 1997; Michell 1999). In a few exceptional cases, biologists have taken an explicitly measurement-theoretic approach (e.g., Stahl 1962; Rosen 1978b; Wolman 2006), but these seem to have had little general impact.

We also argue that measurement theory can be used more broadly as the basis of a theory of scientific meaning that extends beyond the domain of the mathematical theory of representational measurement to include explicit consideration of what to measure and what conclusions can be drawn from those measurements. We suggest the term "conceptual measurement theory" for this broader approach to the relationship between measurements and meaning. This broader aspect of extracting meaning is more familiar to biologists, as we are relatively well-versed in how to think about hypotheses and their testing. Our claim is that thinking about these as part of the measurement process is tremendously useful. We have found that this broader conception of measurement has followed naturally as we have begun to incorporate explicit representational measurement theory into our own work (Wagner et al. 1998; Hansen and Wagner 2001a,b; Wagner and Laubichler 2001; Hansen and Houle 2008; Wagner 2010). Measurement theory provides a language for discussion of a part of scientific practice that is both critical to good science and frequently done poorly. Consequently, we will use the term measurement theory to refer to all aspects of extracting meaning from observations, experiments, and models.

In this review, we first summarize the basic themes of formal, representational measurement theory. We then lay out broader conceptual measurement principles. We drive home the importance of these points by making the case that violations of these basic measurement principles are widespread in biology, as in the examples briefly touched on above. Finally, we discuss what can be done to incorporate measurement theory into the education and day-to-day thinking of biologists.

Measurement

We naturally identify entities that exist in the world (with varying degrees of distinctness) and conceive of attributes of those entities that seem important to us. The idea that some attributes are more deserving of attention than others immediately makes clear that a priori ideas about what is important and interesting underlie all measurement. For example, if our entities are individual organisms, we might be interested in the attribute "size," which might predict which individuals will leave more offspring. Implicit in this simple statement are at least five complex concepts. First is the theoretical context that leads us to care about the number of offspring individuals produce, which might be, among many other possibilities, an evolutionary, an ecological, or an economic context. Second is at least a hypothesis and possibly prior evidence that size might be an important predictor of number of offspring. Third is an implicit definition of size. Fourth is a notion of what constitutes a countable offspring. Finally, we must circumscribe the set of organisms to which our hypothesis is applied.

Measurement consists of an assignment of numbers to attributes of entities so that the relations between the numbers can capture empirical relations among the attributes (Krantz et al. 1971; Luce et al. 1990; Hand 1996, 2004). When what is true about the relations of the numbers is true about the relations of the attributes, the conclusions we draw from the numbers are meaningful conclusions about nature. Many numerical relations could help to capture empirical realities, including order, differences, ratios, and equivalence. Which of these is appropriate depends on our hypothesis about what might actually matter, the actual empirical relations, and how measurement is done.

A simple example of the measurement process is shown in Figure 1. The theoretical context of sexual selection has led to a hypothesis that the length of a male guppy predicts his attractiveness to potential mates. The empirical relationship of lining up two fish side-by-side and recording which is longer is captured instead by measurement of length in standard units. The conclusions about length that can be drawn from the pair-wise empirical comparisons can also be drawn from numbers properly assigned to lengths. Drawing conclusions about the relationship between male lengths and the larger theoretical context of male attractiveness to females then proceeds by the usual scientific process, incorporating additional measurements and experiments. Note that whether the original hypothesis is true or not is not important to the measurement process. What is important is that each of the steps from hypothesis to identification of entities and attributes to measurements and conclusions be well motivated and consistent. False hypotheses will tend to be rejected when all these connections hold.

Representational Measurement Theory

Representational measurement theory is a mathematical system for determining when the relations among the numerical measurements assigned to attributes reflect the corresponding empirical reality (Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990; Hand 1996, 2004; Sarle 1997). The clearest introduction to representational measurement theory is Chapter 1 of Krantz et al. (1971), and we adopt their terminology here. The relationship of attributes in the world is an *empirical relational structure* and consists of a set of entities (e.g., organisms, genotypes, genes, or proteins), the empirical operations that can be made with those entities (e.g., comparing, combining, mutating, or placing in an environment), the attributes that arise from those operations (e.g., contest outcome, trait value, or fitness), and the inferences that can be made from comparison of those attributes. The empirical relational structure is then represented by a *numerical relational structure* in which at-

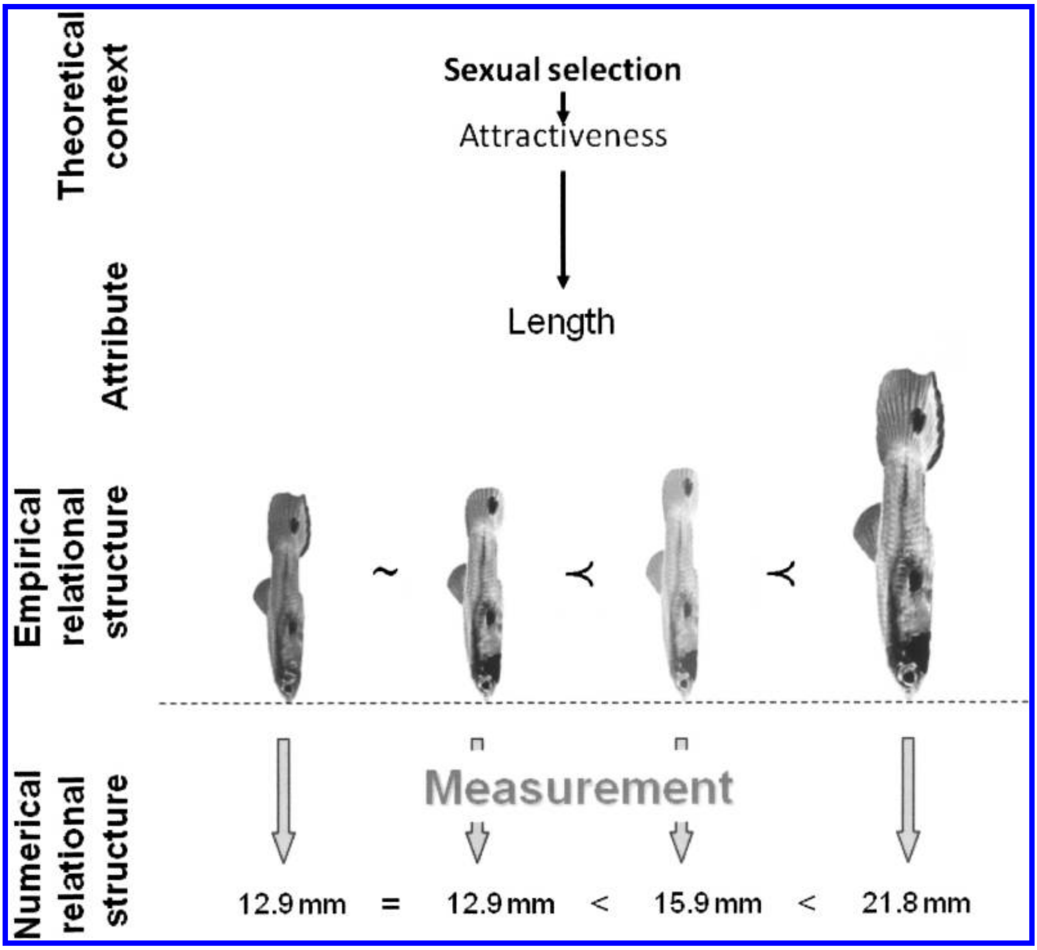

Theoretical context
Attribute
Empirical relational structure
Numerical relational structure
Sexual selection
Attractiveness
Length
~
≺
≺
Measurement
12.9 mm = 12.9 mm < 15.9 mm < 21.8 mm

FIGURE 1. THE MEASUREMENT PROCESS

We imagine a study where the theoretical context is sexual selection. Within this context, we focus on attractiveness of males and then on the hypothesis that size influences attractiveness in the guppy. Size can be measured in many ways, so the concept of "size" was referred to the attribute "overall length of the fish" under the hypothesis that females prefer large males and treat the tail as part of the body. The middle third of the figure represents the empirical relations that are at the root of measurement. For empirical comparison of length, fish could be aligned with their noses against a flat surface, and the identity of the fish that extends farther noted. For a pair of fish A and B, the possible results are that A extends farther than B, which we can represent as $A \succ B$; that we cannot decide whether A extends farther than B, $A \sim B$; and that A extends less far than B, $A \prec B$. Representational measurement theory proves that the conclusions that can be drawn about length on the basis of the pairwise empirical comparisons can also be drawn from a numerical system consisting of numbers (a and b) assigned to lengths of A and B plus a mapping of the empirical relations $\succ$, $\sim$, $\prec$ among fish to the relations $>$, $=$, and $<$ among the numbers.

tributes are mapped to numbers and empirical operations and comparisons are mapped to mathematical relations (greater or less than) and operators (addition or multiplication). A numerical relational structure has the property of meaningfulness when inferences about numbers can be translated into inferences about the original entities (Weitzenhoffer 1951; Luce et al. 1990). Which kinds of numerical relationships are meaningful in this sense is a natural way to define *scale type*. Thus, scale type encapsulates what properties of a set of measurements could be used to draw empirically meaningful conclusions.

Assignment of numbers to attributes—the act of measurement—usually depends on just a few empirical relations (reviewed in Krantz et al. 1971, Chapter 1; Hand 2004). Primary among these is a procedure by which the order of attributes can be established—that an elephant is bigger than a mouse, that 17 May 1814 is later than 4 July 1776, or that animal A is socially dominant to animal B. For example, in Figure 1, the empirical procedure is to place fish next to each other in a standard orientation, with their snouts against a flat object, then see which fish's tail extends farther. This simple operation gives us both the order of lengths and operational equivalence, when we judge that we cannot tell which of the two extends farther. A second empirical operation critical to many measurements is combining entities so that their attributes are also combined in a meaningful way. This is termed *concatenation*. For the fish in Figure 1, we could cut rods equivalent to the length of each fish, then lay the rods end to end to ask questions such as: "Is twice the length of fish A greater or less than the length of fish B?" When concatenation operations are possible, we can also construct a *standard sequence* of concatenated entities (a ruler, for example) that permits counting of attributes in standard units of measurement. A second example is measurement of mass. The empirical operation of placing an elephant and a mouse on a balance will tell us that the elephant is heavier because the scale tips toward the elephant. We can then "concatenate" mice by placing more than one mouse (or other more convenient objects that are each equivalent in weight to a mouse) on the scale, to determine how many mouse equivalents are necessary to make something as heavy as an elephant, that is to measure mass of the elephant in mouse units. The empirical relational structure includes the animals we wish to compare and the operations of concatenation (e.g., placing more than one mouse on a scale) and comparison (does the scale tip toward the elephant or the mice?).

Measurement then proceeds by assignment of numbers to attributes and mathematical operations (such as addition) to empirical operations (such as placing objects on a scale); together the numbers and the operations define a numerical relational structure. When both order determination and concatenation are empirically relevant, these operations can be mapped to a numerical relational structure that uses positive real numbers to represent the attribute, addition to represent concatenation, and the > operator for comparisons. Measurement in these cases is termed *extensive measurement*; both lengths and weights are therefore extensively measurable attributes.

A second important type of empirical relational structure arises when measurement depends on paired comparisons because no natural concatenation operation is possible; this is termed *intensive measurement*. This approach was developed in psychology for measurement of attributes such as attractiveness or utility of objects to test subjects (see, e.g., Luce 1959). Subjects are confronted with pairs of objects, such as faces of humans, and asked to rank them. With repeated trials and under certain well-defined conditions, an overall judgment of the quality of the individual can be made, for example, the "intrinsic attractiveness" of a face. Recently, a similar approach has proven useful for defining a measure of fitness based on pairwise competition experiments, for example, among a set of bacterial strains (Wagner 2010 and see below).

Representational measurement theory proves which numerical relational structures can correspond precisely to the empirical results one would obtain using the actual entities, such as fish or mice and elephants. Narens (1981, 1985) and others (Luce et al. 1990, Chapter 20) have shown that only a small number of such numerical relational structures can preserve meaningfulness given these types of simple empirical relations. More familiarly, these structures define the scale types first

TABLE 1

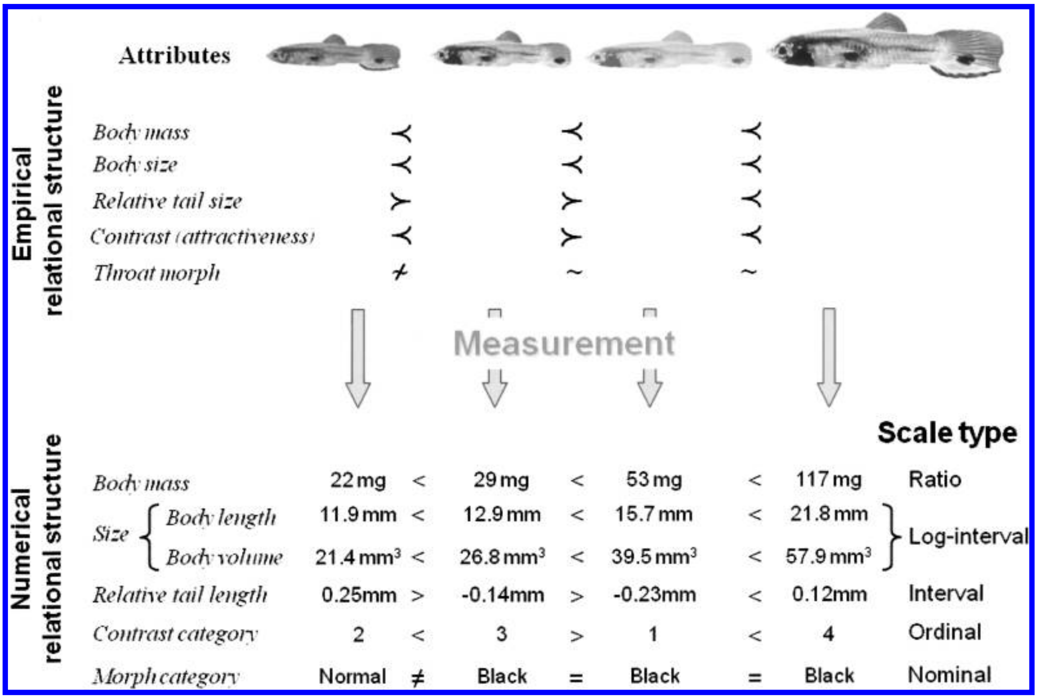
Classification of scale types (after Stevens 1946, 1959, 1968; Luce et al. 1990:113)

Scale type	Permissible transformations	Domain	Arbitrary parameters	Meaningful comparisons	Biological examples
Nominal	Any one-to-one mapping	Any set of symbols	Countable	Equivalence	Species, genes
Ordinal	Any monotonically increasing function	Ordered symbols	Countable	Order	Social dominance
Interval	x -> $ax+b$	Real numbers	2	Order, differences	Dates, Malthusian fitness
Log-interval	x -> ax^b, a, $b>0$	Positive real numbers	2	Order, ratios	Body size
Difference	x -> $x+a$	Real numbers	1	Order, differences	Log-transformed ratio-scale variables
Ratio	x -> ax	Positive real numbers	1	Order, ratios, differences	Length, mass, duration
Signed ratio*	x -> ax	Real numbers	1	Order, ratios, differences	Signed asymmetry, intrinsic growth rate (r)
Absolute	None	Defined	0	Any	Probability

* Luce et al. (1990) defined this ratio scale but did not discuss or name it. Stevens did not consider this scale.

identified by Stevens (1946, 1959, 1968). The scale types relevant to biological systems are listed in Table 1 with examples of each. Figure 2 shows examples of scale types resulting from measurements on a set of fish. We emphasize that scale types can be characterized in several different ways: most fundamentally, they are determined by the empirical operations that the scientist wishes to represent using numerical operations. It is very important to note that the theoretical context is an irreducible part of this formulation. The same data (for example, lengths) may reside on a different scale type depending on the question asked. We give specific examples of this relationship between scale type and hypothesis below. A second convenient characterization of scale type refers to the kinds of mathematical relationships among the measurements that are potentially meaningful, and this has led to the names of the scale types: nominal, ordinal, interval, log-interval, difference, ratio, signed ratio, and absolute. Third, the scale types can be formally characterized by three related properties: the permissible transformations that preserve the relevant relationships among measurements, the number of arbitrary parameters that must be adopted to establish the numerical system, and the domain of numbers (or symbols) to which they apply.

An understanding of scale types is most easily developed from specific examples. For the examples of length and weight developed above, note that the empirical relations can be used to establish the order of attributes (fish A is longer than fish B) and their differences (elephant A is heavier than elephant B by 20 mouse units) and ratios (the concatenation of two fishes the length of A is equivalent to the length of fish B). Any numerical relational system that reflects all of those relationships is by definition on a ratio scale. Positive real numbers are the domain of the measurements, as a negative weight or length has no physical equivalent. To map attributes to numbers, we had to specify one arbitrary parameter in these cases—a unit of length or mass. *Permissible transformations* of the numerical data are defined as those transformations that preserve the correspondence of the numerical relationships to all the empirical relationships that could be measured. For ratio scales the only permissible transformation is multiplication by a constant. To see this, assume that we have four entities and that the empirical attributes, say lengths, are represented by the letters A, B, C, and D. We then measure the four lengths and map the nonnumerical attribute A to a number *a*, the attribute B to *b*, and so forth. If we



Body mass		22 mg	<	29 mg	<	53 mg	<	117 mg	Ratio
Size	Body length	11.9 mm	<	12.9 mm	<	15.7 mm	<	21.8 mm	Log-interval
	Body volume	21.4 mm^3	<	26.8 mm^3	<	39.5 mm^3	<	57.9 mm^3	
Relative tail length		0.25mm	>	-0.14mm	>	-0.23mm	<	0.12mm	Interval
Contrast category		2	<	3	>	1	<	4	Ordinal
Morph category		Normal	≠	Black	=	Black	=	Black	Nominal

FIGURE 2. REPRESENTATIONAL MEASUREMENT THEORY

A set of entities, individual fish, can be associated with different empirical relational structures depending on what attributes we focus on and the context we are interested in. Each attribute induces more or less strict order and equivalence relations among the fish. Representational measurement consists of mapping these relations to a numerical system such that the relations among the attributes are preserved in the relations among the numbers assigned to them. The scale types refer to mappings that preserve different relations. The attribute of mass illustrates a ratio scale type. In many conceptual contexts, for example, studies of metabolic rate, a compelling logical reason favors adoption of mass as the appropriate attribute to measure. In this case, we require a mapping that preserves the order of all fish according to mass (as illustrated) and also the order of differences and ratios between their masses. We can use a standard unit (say, the gram) that we could add up to describe the mass of any of the fishes. The only operation we could apply to our measurements that would preserve these properties would be multiplying by a constant (for example, changing the unit from gram to pound). If we were interested in size, but had no reason to choose mass as a measure of size over other possible measures such as length or area, then we would have to admit raising to a power to the permissible transformations, yielding the log-interval scale. On this scale type, the order of differences in fish size is meaningless because the order of differences can change depending on the choice of exponent, but the order of their ratios is meaningful because the order of ratios does not change with the choice of exponent. Note that whether each of these measures of body size is on a ratio or log-interval scale depends on the conceptual context, not on the attribute itself. The residuals of a regression of tail length on body length measure relative tail size. This example illustrates an interval scale type, where differences between relative tail sizes are meaningful (i.e., a natural unit exists), but their ratios are meaningless. The differences themselves reside on a signed ratio scale, so we can say that the relative tail area of the left fish differs from that of the second fish (0.39) by 11% more than that of the last fish differs from that of the third fish (0.35). If we seek only to represent order, as can be illustrated by contrast in the context of male attractiveness judged by females, we have an ordinal scale type. In that case, statements such as "the difference between fish A and fish B is larger than the difference between fish B and fish C" are meaningless. Finally, a nominal scale type, as illustrated by imaginary throat morphs, preserves only equivalence relations; order is meaningless.

can show, with an empirical operation, that the number of As that must be concatenated to be equivalent to B is less than the number of Cs that must be concatenated to be equivalent to D, we want our numerical measures to have the properties $a/b > c/d$

and $a-b > c-d$. Multiplying all of the numerical measurements by the same constant obviously preserves the truth of these statements; any other type of transformation, such as exponentiation or adding a constant to each number, can readily lead to violation of one or both statements.

Log-transformation of ratio-scale variables places the data on a new scale type, the difference scale, with the domain of the real numbers. On this scale type, differences are equivalent to ratios of the corresponding exponentially transformed values. The only permissible transformation is addition of a constant.

Body size is fundamental to much of biology, and a considerable literature addresses allometry and scaling relationships between size and a wide variety of other phenotypes (e.g., Schmidt-Nielsen 1984). Inspection of this literature, however, reveals two features. First, in most cases, measures of size are log-transformed before analysis. Second, measures of size can be measured on a linear, area, or volume scale, and these are treated interchangeably (Solow and Wang 2008). For example, to investigate Cope's Rule that body size tends to increase within lineages, Alroy (1998) studied the log body mass of fossil mammals predicted from linear tooth dimensions, whereas Hone et al. (2005) used log bone length. The interchangeability of these scales in practice suggests that biologists consider raising measures of size to a power to be a permissible transformation. Choice of (for example) linear over volume measures is seen as a matter of convenience or convention. A key implication of this practice is that ratios are meaningful—the order of ratios, say $a/b > c/d$, is preserved when one raises all the measurements to a power—but that the order of differences is not. For example, say that a = 5, b = 1, c = 10, and d = 7. On this scale, $a/b > c/d$ ($5 > 1.43$) and $a\text{-}b > c\text{-}d$ ($4 > 3$). When the measurements are raised to the power of 2, the ratio a^2/b^2 is still greater than c^2/d^2 ($25 < 2.04$), but the order of the differences is reversed: $a^2 - b^2 < c^2 - d^2$ ($24 < 51$). The scale type with these permissible transformations is the log-interval—so named because once the data are log transformed the ratios that are meaningful on the original scale become differences residing on an interval scale type.

A key example of the context dependence of scale type now presents itself: above, we claimed that body lengths and body mass are on a ratio scale, and now we say that they are on a log-interval scale. The difference is that different theoretical contexts are applied in the two cases. When a measure of length is treated as a measure of length it is on a ratio scale. For example, if, as in Figure 1, we hypothesize (or have evidence) that female guppies really choose their mates on the basis of body length, rather than volume, then the choice is dictated by this hypothesis. In an allometric context, we choose to regard powers of body size as potentially relevant to the questions. This distinction is a key part of our argument that measurement theory must incorporate theoretical context and is therefore relevant to the wider issues of scientific reasoning.

If our data consisted of dates of occurrences, the interval scale type would be appropriate. The key difference between interval and ratio scales is that the empirical relation of concatenation does not apply: one cannot directly concatenate dates. How would you decide how many 4 July 1776s add up to give you a 17 May 1814? Without this operation, the magnitudes of dates cannot be directly judged; the ratio of dates does not correspond to an empirical operation and is therefore not meaningful. We can, however, use the difference between two dates, say January 1 and January 2 as our concatenatable unit—how long we wait until the date changes. Then the difference between 4 July 1776 and 17 May 1814 can be expressed as the number of days on top of 4 July 1776 that are needed to reach 17 May 1814. The permissible transformations therefore include addition of a constant, for example, adopting the Judaic or the Islamic calendar instead of the Gregorian calendar, as well as multiplication by a constant (adopting the Julian year, rather than Gregorian one). To arrive at this system, we must specify two arbitrary parameters, the starting and stopping dates we use as a standard unit of time. The use of concatenation in measuring the difference between two dates or temperatures means that the differences between dates are extensively measurable and therefore on a ratio scale.

The case of an interval scale allows us to illustrate the important distinction between scale and scale type. Other familiar examples of interval scale type are the Celsius and Fahrenheit temperature scales. All temperature scales based on arbitrary numerical assignments to two states, such as freezing and boiling points of water, are on the interval scale type regardless of the unit of temperature used, but once a unit of temperature has been chosen, we have a scale on which units are essential to the interpretation of the numbers. When we say, on the basis of scale type, that differences can be meaningfully compared, what we mean is that the relationships of the differences are preserved under the permissible transformations and not that the numerical values of the differences are preserved. For example, if temperature changes in one hour from 8 to 10 in degrees Celsius, and then in the second hour to 11 degrees C, we can conclude that the temperature rose twice as fast in the first hour as in the second hour. If we convert to the Fahrenheit scale, the temperatures (46.4 to 50 to 51.8 degrees F) and the differences (3.6 and 1.8 degrees F) change, but temperature can still be inferred to have risen twice as fast in hour one as in hour two.

Now consider the case of social dominance, in which animals can be placed on a simple linear hierarchy. The single relevant empirical operation is to place two animals together and observe which one is dominant. The outcome of such a contest does not directly tell us anything about the magnitude of the difference between animals. No empirical operation corresponds to concatenation, either of the animals themselves or of their differences. Even if we could duplicate animal A and put two As with one B, the result would no longer tell us anything about pairwise dominance. Similarly, knowing that both A and B are dominant to C does not necessarily tell us anything about the relation between A and B. Therefore, any numerical system that preserves the order is permissible, and the scale type is ordinal. The assignment of numbers to submissive and dominant individuals could as well be 1 and 2 as 1 and 1000 or 999 and 1000.

The absolute scale type differs from the others in that no transformation is permissible. For example, the axiomatic definition of a probability constrains its domain to the interval 0 and 1 and specifies the precise meaning of any value on that range. An arcsine square-root transformed probability is not a probability.

In Table 1, we also consider two examples of a relatively unknown scale type, the signed ratio scale. This type had previously been identified on mathematical grounds (Luce et al. 1990), but to our knowledge no real examples have been discussed. This scale differs from the ratio scale in that the domain includes all real numbers, so ratios have a sign that can be negative or positive. For example, for a measurement of signed asymmetry (length of structure on the left side of the body minus length of the same structure on the right side), both the ratio of the asymmetry and whether the asymmetry is in the same direction are meaningful. Similarly, the intrinsic rate of population increase has both a magnitude and a sign.

PRAGMATIC MEASUREMENT THEORY

In many situations, the relationship between attributes and the numerical values we use to represent them is not entirely clear. At one extreme, these uncertainties stem from quantifiable factors such as measurement error that need not challenge our understanding of the empirical relational system. At the other extreme, we may be studying an attribute about which we cannot be sure what measurements can actually represent it or even whether a hypothesized attribute actually exists. Numerous examples of such attributes exist in behavioral research, and these have been much discussed in the application of measurement to psychological research (see, e.g., Michell 1999).

For example, it is still controversial whether the perception of sensory stimuli leads to a quantitative attribute called sensation, or whether intellectual ability is really a quantitative trait. These uncertainties arise particularly when the goal of measurement is to predict some future outcome, without necessarily having any underlying model of the relationship between the attributes that are measured and the outcome (Breiman 2000).

These considerations have occasioned considerable debate about the usefulness of representational measurement theory. In the extreme, some claim that representation is an illusion and that measurements capture only what the measurement procedure measures. This view leads to the philosophical position of operationalism, in which scientific concepts and variables are defined solely in terms of their measurement procedures (Bridgman 1927). In this view, intelligence is nothing more than what is measured by intelligence tests (Boring 1945). Sneath and Sokal (1973), for example, used operationalism to argue for quantification and algorithmic decision-making in biological classification because it would be repeatable and objective, consciously giving up the goal that such classifications would reflect evolutionary history. Such positions contrast sharply with representational measurement theory, which assumes that the point of measurement is to ensure meaningful statements about an empirical relational structure that may exist whether we are there to measure it or not.

Today, pure operationalism is largely discounted by most scientists, but many authors acknowledge that many measurements are not purely representational, even when representation of an empirical system is the goal (Hand 1996, 2004). Hand (2004) called these nonrepresentational aspects of measurement *pragmatic*. The more pragmatic the measurements are, the greater the uncertainty about whether the measurements actually represent the attribute we want to characterize and, therefore, about the potential meaning of the results. Pragmatic considerations often arise when researchers must choose between several alternative measurements of the same underlying entity. For example, we might wish to measure the attribute "sexual attractiveness to females" of a sample of males in a population, but unless we fully understand what goes on in the brain and nervous system of those females, we cannot know how best to assess attractiveness.

McGhee et al. (2007) measured two aspects of male-female mating interactions in the bluefin killifish (*Lucania goodei*): the time a female spent associating with each of two confined males and which of two males was successful when two males and one female were placed in a single tank with no barriers. Association and mating success were poorly correlated, and which of them is a better measure of attractiveness is unclear. One interpretation of McGhee et al.'s finding is that female interest in the confined males measures attractiveness, which then gets overwhelmed by male-male interactions. A second is that male-male interactions furnish additional information on mate suitability to females, changing their preferences (Wiley and Poston 1996). A measurement-theoretical response would suggest additional investigation of the link between the measurement and the underlying entity (e.g., Michell 1999), but this is often a very difficult task. Most studies make the pragmatic choice of one measure of an attribute such as attractiveness, without verifying that it is the best such measure. This pragmatic approach may be the best way forward because a standard assay may at least be comparable across different experiments. The possible shortcomings of one's measure should be borne in mind and reevaluated when the opportunity arises. Humility and common sense are necessary complements to representational measurement theory.

Discussion of these issues can be facilitated by the concept of validity, which "describes how well the measured variable represents the attribute being measured" (Hand 2004:129). Validity is widely discussed in the social and behavioral sciences, and to some extent in medicine, particularly with respect to the representation of mental states. It can be usefully applied in evolutionary biology as well.

TAKING A BROAD VIEW OF MEASUREMENT THEORY

Although the formal results of representational measurement theory are not familiar to most biologists, the fundamental issues that they address are the concern of every working scientist: how we can best understand reality. All measurement stems from a theoretical or conceptual context; good science depends on preserving the connection of measurement, data handling, analysis, and interpretation to that conceptual context. Our provisional ideas about the study entities specify the relevant attribute and, thus, the hypothesized empirical relational structure we should study. Once this step is taken, representational measurement theory ensures that our numbers reflect the necessary empirical relations. Specifying an empirical relation system to study and drawing conclusions from the observed empirical relations back to the concepts—i.e., "conceptual measurement theory"—is part of the measurement process.

The importance of concepts and hypotheses for measurement can be illustrated with the question: Why are giraffes (*Giraffa camelopardalis*) so tall? The obvious explanation proposed by many (e.g., Darwin 1871, Chapter 7) is that giraffes are tall so that they can browse on tall trees. When this hypothesis is investigated, height is clearly the relevant attribute. Alternatively, Simmons and Scheepers (1996) proposed that giraffe necks are an adaptation for male-male competition. Male giraffes compete for dominance by "necking," combat in which males swing their armored heads at one another. This hypothesis suggests that height is not the most informative aspect of form to measure; instead measurements might focus on the amount of force that can be delivered by the swinging head, which must be a nonlinear function of the mass of the head and length of the neck. The cause of the elongated shape of giraffes has not been resolved (Cameron and du Toit 2007), but different hypotheses clearly assign different meanings to the same phenomenon.

The importance of multiple hypotheses in this example suggests a measurement-theoretic explanation for the effectiveness of "strong inference" (Platt 1964) and the "method of multiple working hypotheses" (Chamberlin 1890, 1965). The common element to these approaches is that the researcher should seek to consider all reasonable hypotheses simultaneously rather than to focus on only one. In Chamberlin's evocative terms, "the investigator thus becomes the parent of a family of hypotheses: and, by his parental relation to all, he is forbidden to fasten his affections unduly upon any one" (Chamberlin 1890, 1965:756). In the measurement context, the investigator considers that the same phenomenon may reflect a variety of underlying empirical relational systems and, thus, that a variety of measures may be desirable. The same data may have a different meaning under each hypothesis.

MEASUREMENT THEORY AS AN AID TO MODELING AND STATISTICS

The premise of measurement theory is that we make measurements to learn about an empirical relational structure. Hypothesized empirical relational structures are also the basis for theoretical models, so an important but underappreciated role for measurement theory is as a guide to the definition and identification of parameters in models. A good definition must identify forms and parameters that are operational in that they represent the relevant empirical relations and can also be related to obtainable data. This principle is linked to Lewontin's (1974) conceptions of a *dynamically sufficient* system as one that would allow prediction if its parameters were known and an *empirically sufficient* system as a dynamically sufficient system whose parameters are estimable with sufficient accuracy to allow prediction. Dynamically sufficient parameters are meaningful from a measurement-theory perspective. Empirically sufficient ones lead to actual measurements that are meaningful.

The quantitative genetic concepts of additive effects and additive genetic variance are good examples of variables that are both dynamically meaningful and empirically operational. Fisher (1918) developed an abstract model of the genotype-phenotype map and proposed population-level measures of the effect of allele substitutions, the average excess, and the additive effect (Fisher 1941). The additive genetic variance is the population variance of these effects (summed over loci) and, because the response to selection depends on variance (Fisher 1930; Price 1970), the additive variance emerges as a key parameter for prediction of the short-term evolvability—capacity of a population to evolve (Houle 1992; Wagner and Altenberg 1996). Fisher's measures are approximately dynamically sufficient in that they capture that part of the effect of alleles on the phenotype that determines how their frequencies change under an episode of selection, although they are not dynamically sufficient for long-term evolution (Frank 1995). In addition, Fisher also showed that these parameters are all estimable from data on the phenotypes of relatives and thus empirically sufficient for the task of predicting short-term response to selection.

DIMENSIONAL ANALYSIS

Dimensional analysis is perhaps the most important practical application of measurement theory in theoretical physics and should be integral to all model building. In its simplest form, it requires that mathematical models be dimensionally consistent (i.e., that the units on the left- and the right-hand sides of an equation be equal; apples cannot equal oranges), and it provides a technique for reducing the model to a minimum number of dimensionless essential parameters. Despite sporadic applications (e.g., by Stahl 1961, 1962; Rosen 1962, 1978a,b; Gunther 1975; Heusner 1982, 1983, 1984; McMahon and Bonner 1983; Prothero 1986, 2002; Stephens and Dunbar 1993; Gunther and Morgado 2003; Frank 2009), more formal dimensional analysis has played little role in biological theory or modeling. This is reflected in the arbitrary assumptions about functional forms that are commonly found in biological models.

The key result in formal dimensional analysis is Buckingham's π theorem (Bridgman 1922, Chapter 4; Krantz et al. 1971, Theorem 10.4). This theorem considers a putative law or model that relates a set of ratio-scale variables, x_i, as $\mathrm{f}(x_1,\ldots x_n) = 0$, where f is an arbitrary function. The units of these variables can be used to identify a set of fundamental variables (and scales) that span the n x variables (for the technical meaning of "span" see Krantz et al. 1971). If exactly m fundamental scales exist, the π theorem guarantees that the law can be reformulated as a function of n - m new variables that all are products of powers of the original variables as $\mathrm{g}(\pi_1, \ldots, \pi_{\text{n-m}}) = 0$, where $\pi_i = x_1{}^{a_1} \cdot x_2{}^{a_2} \cdot \ldots \cdot x_n{}^{a_n}$, where each π_i has a different combination of exponents, a_j.

If it is known what variables enter a problem, this theorem can be surprisingly powerful in simplifying and solving models. As an example, it can be used to derive the basic exponential growth equation of population ecology from minimal assumptions. Assume that all we know is that population growth involves the three ratio-scale variables population number at time t, $N(t)$; population number at time zero, $N(0)$; time, t; and a signed ratio-scale variable, the rate parameter, r. Our model or "law" of population growth then takes the form $\mathrm{f}(N(t), N(0), r, t) = 0$, for an unspecified function f, where we also assume that $N(t)$ can be uniquely expressed by the rest of the variables. The units of $N(t)$ and $N(0)$ are individuals (or individuals per area), the unit of time is a time interval (e.g., seconds or generations), and the unit of r is the inverse of the time interval. We therefore have four variables and two fundamental scales, which are individuals and time. The π theorem implies we can write the model in terms of two dimensionless variables that will have to involve powers of $N(t)/N(0)$ and powers of rt. The result is $\mathrm{g}(N(t)/N(0), rt) = 0$ for some function g. By the assumption that $N(t)$ is a function of the other variables, this function can then be rewritten as $N(t) = N(0)\mathrm{h}(rt)$ for some function h. Because h does not depend on any parameters beyond rt, this implies that $N(t) = N(s)\mathrm{h}(r(t-s))$ for any time $s < t$. Hence, we must have $N(0)h(rt) = N(s)\mathrm{h}(r(t-s))$, and that implies

$$\begin{aligned} \mathrm{h}(rt) &= (N(s)/N(0))\mathrm{h}(rt - rs) \\ &= \mathrm{h}(rs)\mathrm{h}(rt - rs), \end{aligned}$$

which gives us the functional equation $\mathrm{h}(rt) = \mathrm{h}(rs)\mathrm{h}(rt - rs)$. Let $\mathrm{k}(x) = \ln(\mathrm{h}(x))$ and write the equation as $\mathrm{k}(rt) = \mathrm{k}(rs) + \mathrm{k}(rt - rs)$. This shows that $\mathrm{k}(x) = ax$ for some constant, a, and therefore that $\mathrm{h}(x) = \mathrm{e}^{ax}$. By subsuming the constant a into r, we have derived the exponential growth equation:

$$N(t) = N(0)\mathrm{e}^{rt}$$

for $r \geq 0$ (a similar argument based on $r < 0$ completes the derivation). Note that we did this without specifying any model of population growth. The law of exponential growth follows entirely from knowing the relevant variables and their scales. Such results are, at first, very surprising, but they signal the general power of dimensional analysis. Knowing which variables are relevant to a problem conveys a huge amount of information. Of course, the law of exponential growth may not hold if other variables or parameters were involved. If, for example, we add a carrying capacity, K, with the same units as N, the π theorem would stipulate three dimensionless variables in g, and additional assumptions would be needed to produce a specific solution. The approach will give misleading results if some relevant variable or parameter is omitted.

Among earlier applications of dimensional analysis in biology, we particularly note Robert Rosen's program for theoretical biology, in which dimensional analysis played a central role (Rosen 1978b). For example, Rosen (1962) derived D'Arcy Thompson's (1917) theory of transformations of biological form from dimensional arguments based on fitness optimization. His development was, however, highly abstract and, as far as we know, it has not led to empirical research. Dimensional analysis has also played a minor role in discussion of physiological scaling relationships (Stahl 1961, 1962; Heusner 1982, 1983, 1984; but see Butler et al. 1987 for a devastating critique of Heusner's use of dimensional arguments), and such arguments are also central to Charnov's (1993) theory of life-history invariants, although Charnov does not make the connection to measurement theory explicit. Stephens and Dunbar (1993) developed applications of dimensional analysis in behavioral ecology with admirable clarity. They showed how the marginal-value theorem (Charnov 1976) can be partially derived and illuminated by formal dimensional analysis, and they showed that certain models of optimal territory size from the literature are in fact dimensionally inconsistent.

We believe that dimensional analysis has been underused in biology. It has the potential to clarify the necessary and sufficient assumptions for fundamental models and concepts in biology along the lines we have briefly illustrated for the exponential growth law.

MEANINGFUL MODELING

The formal derivation of the exponential growth law from dimensional considerations shows that it has a status similar to the familiar "laws" of physics, such as Newton's law of gravitation. To see the analogy one must appreciate that Newton's law is also an abstraction for the most symmetrical case. The gravitational law applies only to a rotationally symmetrical gravitational field, which does not exist in the solar system because of variations in the density of matter in the planets and the presence of other bodies. A similar argument can be made for the general validity of Wright's selection equation, which can also be directly derived from measurement-theoretical considerations (Wagner 2010). Biology includes rather few such laws, however, and we must accept that most models in biology are representations of qualitative relationships that are too complex to capture in simple law-like relations. The complexity and high dimensionality of the relevant empirical relational structures require that some aspects of representation must necessarily be given up.

Levins (1966) described three modeling strategies based on which aspects of reality are left out. In type I models, generality is sacrificed in favor of precision and realism, as in highly complex fisheries models in which as many specific aspects as possible of the stock in question are included. In type II models, realism is sacrificed to generality and precision, as in the Lotka-Volterra models of population ecology. In type III models, precision is sacrificed to generality and realism, as in Levins's own models of selection in heterogeneous environments based on general qualitative assumptions about the fitness function (e.g. Levins 1962).

Levins himself favored a research strategy in which type III models or multiple type II models are used to reach robust qualitative insights and predictions. This approach contrasts sharply with that taken in most modeling papers in the fields with which we are familiar. Typically, a question is investigated by analysis of a single specific model of type I or II. In many cases, results are obtained solely by simulations based on algorithms including numerous auxiliary specifications, so the results necessarily have limited applicability. Type I models can only make predictions for highly specific circumstances, and type II models can only demonstrate the possibility of a phenomenon, not its plausibility.

These limitations are often forgotten in theoretical biology, where general conclusions are routinely drawn from specific models. In the introduction, we mentioned Maynard Smith's (1976) claim, based on a type II model, that the handicap principle could not work. Maynard Smith's claim was later shown to be incorrect (Pomiankowski 1987, 1988; Grafen, 1990a,b) and the handicap principle is now a cornerstone of many sexual-selection models. Using measurement theory to diagnose the problem, we can see that arbitrary specific model assumptions, such as linearity, put constraints on the empirical relational structure that prevent it from capturing the full range of possibilities in the idea being modeled.

This dynamic of overly broad claims from narrow models being overturned by type III models is quite common. For example, Broom et al. (2005) claimed on the basis of specific families of cost and benefit functions that automimicry, a situation in which some defenseless (e.g., nonpoisonous) individuals exist within a generally aposematic species, cannot be maintained as a stable genetic polymorphism. Svennungsen and Holen (2007) considered a wider variety of functions and showed that automimicry can indeed be evolutionarily stable under a number of realistic conditions. A second example is the large literature on whether plasticity, learning, and variation increase or decrease a response to selection, known as the Baldwin effect (Baldwin 1896). Multiple models have been published showing that either the Baldwin effect of acceleration was predicted (e.g., Hinton and Nowlan 1987) or favoring the opposite result of slower evolution (e.g., Borenstein et al. 2006). Paenke et al. (2007) demonstrated that the conclusions of these models were dictated by the change in the slope of the relationship between fitness and the focal trait. When plasticity increases this slope, the variance in fitness increases accelerating evolution; the converse occurs when the slope decreases. The general model focuses our attention on a key biological aspect of plastic systems that was not previously considered explicitly.

MEANINGFUL STATISTICS

Measurement theory requires that numbers only be manipulated in ways that retain their representation of the empirical relational structure of interest. Statistical models, however, only make assumptions about the distributional properties of the data and can be applied to any set of numbers that fulfill the distributional criteria. Consequently, statistical education and practice often recommend transformations of the data that make the numbers fit the statistical model. This practice often brings measurement principles and statistical practice into conflict. Measurement theory puts severe constraints on the statistical manipulations that can be done without loss of some or all of the meaning present in the data.

This fundamental truth has not had an important place in the biostatistical literature, but has been the subject of considerable debate in psychology (reviewed by

Hand 1996, 2004; Michell 1999). Some hold that measurement theory is irrelevant to statistics because the "numbers do not remember where they came from" (Lord 1953:751), and much statistical practice in biology seems based on this viewpoint. For example, the discussion of transformations in the context of ANOVA in the widely used biostatistics textbook by Sokal and Rohlf (1995) focuses on convincing the reader that "the scale of measurement is arbitrary, you simply have to look at the distributions of transformed variates to decide which transformation most closely satisfies the assumptions of the analysis of variance" (p. 412), a sentiment echoed by others (Dytham 2003; Logan 2010). If that is statistics, we want no part of it, as science is about nature, not numbers. We follow Adams et al. (1965) and define a meaningful statistical statement as one whose truth is invariant to permissible scale transformations (in the measurement-theoretic sense) of the underlying data. The sad results of ignoring scale during statistical analysis are legion. For example, if we have ordinal-scale variables, such as ranks of items, the statement that the arithmetic mean of one sample is larger than that of another is not meaningful, because rankings reflect perceived order only, whereas the act of averaging assumes that the values reflect magnitude. Permissible transformations of ordinal data will usually exist that alter the order of the means. For example, if we have two entities in category B, and three in category A, and the Bs are ranked 2 and 3 out of five, then mean rank of Bs is 2.5 while the mean rank of the As is 3.3. Now assume that more entities were included in the ranking, so that the As are now ranked as 1, 9, and 10, and the Bs as 7 and 8, yielding a lower mean score for category A (6.7) than for category B (7.5). Despite the reversal of the order of the means, both rankings are equally valid. A mean is a meaningless statistic for ordinal variables. Wolman (2006) pointed out that conservation biologists make precisely this error when making recommendations based on the average of expert rankings of, for example, species vulnerabilities or conservation values.

Once an impermissible transformation is used during the analysis of the data, the aspect of reality to which the measurements and the associated statistical results apply is, by definition, changed. Similarly, nonparametric approaches usually test a hypothesis about the sample median rather than the means. This point is often ignored in the interpretation of the results of statistical tests, potentially resulting in meaningless conclusions. Sometimes transformations of data lead to understandable changes in meaning, as when a log transformation maps ratio relations into difference relations and changes a log-interval scale type into an interval-scale type, but transformations commonly lead to fundamental changes in the meaning of the numbers. More often than not, such changes are not communicated by the authors of the study. We present two such examples later in this paper. Fortunately, the generalization of statistical models to encompass distributions other than the normal and the rise of numerical methods of analysis usually obviate the need for unprincipled transformations of data. In many cases, statistical tests are surprisingly robust to violations of assumptions (see e.g., Whitlock and Schluter 2009:403). Nevertheless, difficult cases will remain in which the assumptions of the available statistical analyses are not met, but the transformations that would correct the problem would divorce the measurements from the empirical relations. Collaborations between biologists and statisticians may be necessary in such cases.

The divorce of statistics from meaning is perhaps most apparent when only the qualitative results of statistical tests are given or interpreted, ignoring the numerical values of the estimated parameters. In such cases, a nominal or ordinal conclusion is drawn from estimates that are on a much stronger scale type. For example, the conclusion of Takahashi et al.'s (2008) study of sexual selection in peacocks, which is summed up in their paper's title Peahens Do Not Prefer Peacocks with More Elaborate Trains, was based on a failure to reject the null hypothesis of no selection. Their best estimate of the strength of selection was that a peacock gained 3% more matings for every additional eye spot in his tail. This is extremely strong selection, three times the strength of selection on fitness itself (Hereford et al. 2004). The range in the number of eyespots among males was 42, suggesting that the peacock with the most eyespots had an expected mating success $1.03^{41} \approx 3.4$ times that of the one with the fewest. Consideration of the estimate of the strength of selection, rather than its lack of statistical significance would properly lead to the conclusion that this study lacks the power to detect even extremely strong female preference rather than the authors' misleading conclusion that peahens are indifferent to the plumage of peacocks.

The confusion of biological with statistical significance is exceedingly common in ecology and evolution (see, e.g., Yoccoz 1991; Anderson et al. 2000). The seriousness of the problem is underscored by the more than 300 criticisms of this practice that Anderson et al. found in the scientific literature through the year 2000; many more have surely accumulated by now. The prevalence of such errors in the face of so many efforts to prevent them indicates a systemic problem; we believe the problem is the lack of awareness of what measurement theory tells us about the meaning of numbers. Biologists accept P-values as measures of effects for the same reason that they commonly neglect to report and consider units and transformations. The skill of interpreting numbers is neither taught nor practiced: too often quantification is window dressing for qualitative arguments.

Measurement theory is a description of what meaningful quantification entails. It is precisely the medicine needed to restore meaning to statistical practice. Fulfilling the assumptions of the statistical model is important, but it is by itself useless if the statistical model does not respect the empirical content of the data.

Measurement in Biological Practice

The principles we have outlined above are so near to platitudes that it may seem odd that we think them worth emphasizing. There are, however, numerous examples of errors arising from violations of these principles—sometimes for representational errors, but just as frequently for simple conceptual errors that lead to irrelevant measurements. Even more troubling is that the perceived best practice in a particular field may incorporate a measurement error that renders an entire class of studies meaningless. We present a sampling of such errors here.

example 1: remember context

The actual theoretical context and the measurements performed can sometimes be mismatched. The most spectacular errors of this type occur when the investigator loses track of the theoretical context that is the ostensible purpose of a study. A striking example of this can be found in much recent work on the evolution of allometry. Julian Huxley (1924, 1932) introduced allometry as a simple scaling relationship between a trait, Y, and overall body size, X, as a power law, $Y = aX^b$, which is usually and more conveniently expressed as the linear relationship $\log(Y) = \log(a) + b \log(X)$. Empirically, such linear log-log relations are common both within species (static allometry) and between species (evolutionary allometry). Huxley showed that static allometries between two traits will result when they are under common growth regulation, and the idea that evolution may be constrained to follow static allometries has received considerable attention (e.g., from Gould 1977). On the basis of his model, Huxley (1932:5–6) felt that the intercept $\log(a)$ was "of no particular biological significance," but the exponent, b, "has an important meaning." Indeed, in physiology, biomechanics, and life-history theory much theoretical and empirical work has gone into establishing the exact value of the allometric exponent for different traits (e.g., Schmidt-Nielsen 1984; Charnov 1993). When the overall relationship between size and another trait varies, the intercept is usually far more variable than the slope (Teissier 1936; White and Gould 1965; Jerison 1969; Greenewalt 1975; Dudley 2000).

Although many authors follow Huxley's conception of allometry as "the slope of a . . . log-log regression of the size of a structure on body size" (Eberhard 2009: 48), some have adopted a "broad" sense of allometry as "[c]hanges in shape which accompany change in size" (Mosimann 1970: 930). In this usage, all kinds of nonlinear and even sigmoid and discontinuous (threshold) relationships are referred to as allometric (Frankino et al. 2010). Thus, broad-sense allometry is essentially synonymous with shape and, therefore, without precise connection to any of the theoretical models that have motivated the interest in the allometric slope as an important evolutionary constraint, whether Huxley's hypothesis of common growth regulation or more sophisticated models from physiology, biomechanics, or life-history theory.

Several investigators over the last 20 years have claimed to alter allometry by artificial selection on trait indices (Weber 1990, 1992; Wilkinson 1993; Frankino et al. 2005, 2007). Unfortunately, they make only passing reference to the theoretical concepts that have defined the interest of allometry to biology. None, for example, log-transformed their data. The study by Frankino et al. (2005) is illustrative. The title of the study, the abstract, and the text formulate the problem as explaining the conservatism of "allometry" or "scaling relationships," but include no discussion of specific allometric models. The implication, for those familiar with the usual conception of allometry as slope on a log-log scale, is that the study will be about the evolution of this slope. In their experiment, Frankino et al. (2005) sought to alter wing loading, the relationship between wing area and body mass, in the butterfly *Bicyclus anynana*. To do so, they selected for increased or decreased individual deviations from the major axis of variation between forewing area and body mass. They achieved statistically significant responses in this index.

What do these results mean for allometry, the stated subject of the study? The answer is entirely unclear, because the type of selection used by Frankino et al. will favor changes in both intercept ($\log(a)$) and allometric coefficient b. The only window through which the reader can assess the nature of the responses is in a figure (Frankino et al. 2005, Figure 2A) that presents the distributions at the end of the experiment on an arithmetic scale. No analyses are made on the log scale, however, and no attempt to determine whether the differences are in the slope or the intercept. The theoretical context of allometry is not incorporated into the design or analysis of this study, despite the author's invocation of the concept of allometry.

All of these experiments (Weber 1990, 1992; Wilkinson 1993; Frankino et al. 2005, 2007) alter something about the relationship between traits and clearly demonstrate that body shape shows genetic variation, a valuable conclusion. All of them invoke the concept of allometry, but how to interpret their results in terms of allometry is perfectly ambiguous. One extreme interpretation is that all of the response is in the parameter a, suggesting that, contrary to the conclusions of these papers, the allometric slope is constrained by a lack of genetic variation. A second interpretation is that some part of the response is in the exponent b and that natural selection can indeed reshape allometric relationships easily.

Meaning is lost in this class of studies because a theoretical context is invoked but then ignored. The design of these experiments precluded testing the claim that allometry can readily be altered, because the measurements and analyses did not reflect hypotheses about allometry. This is an error in specifying the attribute of interest as any aspect of the relationship between pairs of traits rather than as the slope of their relationship measured from log-transformed data.

EXAMPLE 2: DO NOT USE A RUBBER RULER

Consider an animal foraging in an environment containing two equally common types of resource patches, where one type of patch has twice the payoff of the other. The animal chooses one patch at random and finds it has payoff x, but it cannot tell whether it is a good or a bad patch. Should it then switch patches? If it switches, it has a 50% chance of encountering the same payoff x but also a 50% chance of encountering a different payoff. In the case in which the payoff is different, the chances are equal that the new patch will have twice the value ($2x$) or half the value ($x/2$) of the original patch. From optimal foraging theory, we might then argue that the animal should switch patches, because the expected payoff for the new patch would be $2x{\cdot}1/2+x/2{\cdot}1/2 = 5x/4 > x$. Notice that this conclusion would be the same whether the animal chose the good patch or the poor patch on the first attempt. The grass is always greener on the other side!

This conclusion is clearly absurd, but identifying the error is not trivial. The cause is failure to adopt a common scale. To see the problem, first define the payoff in the poor patch as z and that in the better patch as $2z$. In terms of this fixed scale, we easily see that the payoff of the first patch is $x_1 = 2z{\cdot}1/2 + z{\cdot}1/2 = 3z/2$, whereas the payoff of the second is $x_2 = z{\cdot}1/2 + 2z{\cdot}1/2 = 3z/2$, giving the obviously correct answer that $x_1 = x_2$. With this definition in mind, we can return to the first equation, and diagnose the problem; x is used to equal both z and $2z$ in the same equation.

This example, derived from the envelope paradox in probability theory (Hand 2004), illustrates the dangers of a lack of awareness of scale. In fact, scales that depend on the entity to be measured are common in evolutionary biology. The use of heritability to measure evolutionary potential is one important example. We can define short-term evolvability as the expected response to a given selection gradient (Houle 1992; Hansen et al. 2003; Hansen and Houle 2008), and under certain assumptions it equals the additive genetic variance (Lande 1979). To compare additive variances across traits and populations, we need a common scale. The standard practice has been to use the trait's population variance as the scaling unit. Dividing the additive variance by the population variance yields the heritability, h^2, which is a dimensionless number. Note, however, that the population variance contains the additive variance as a component and, furthermore, that its other components, such as epistatic and environmental variances, tend to be correlated with the additive genetic variance (Houle 1992). Therefore, when heritabilities are used for comparison of evolvability across traits and populations, we use a scale that depends strongly on the entity to be measured. The naive use of h^2 as a measure of evolutionary potential has lead to some highly dubious generalizations, such as the idea that life-history traits and fitness components are less evolvable than morphological traits (see, e.g., Roff and Mousseau 1987). Houle (1992, 1998; Houle et al. 1996) proposed using a mean-standardized scale (such as a coefficient of variation or its square I_A) because it is not necessarily a function of the attribute to be measured; on this scale, the lower h^2 of life-history traits is clearly due to high levels of environmental variance. The use of the variance-standardized scale gives a misleading conclusion for reasons strikingly similar to the one that generates the patch paradox. The naïve expectation that evolvability can be measured equally well by either h^2 or mean-standardized additive variance is wrong; in fact they are almost uncorrelated (Houle 1992).

In this case, the measurement error is in selecting an attribute to measure, h^2, that is unsuited to the theoretical context. Heritability is unsuitable because it standardizes a quantity of interest with something to which it is autocorrelated. This example shows that meaning is drastically altered by seemingly innocent choices of scale. The choice of scale is a fundamental and nontrivial part of both model building and statistical estimation. A lack of awareness of such issues is a major source of misreporting and misinterpretation of biological results.

EXAMPLE 3: INTERPRET YOUR NUMBERS

Yet another way of robbing a study of meaning is to fail to specify a theoretical context. A good example is Kingsolver et al.'s (2001) paper, The Strength of Phenotypic Selection in Natural Populations. One of their conclusions was that "directional selection on most traits and in most systems is quite weak" (p. 253), yet Kingsolver et al. never specified a theoretical context that would help to determine the strength of selection. As pointed out by Conner (2001), this omission leaves readers wondering whether selection is really strong or weak.

Kingsolver et al. reviewed a large sample of the available estimates of linear selection gradients obtained in wild populations (Lande and Arnold 1983; Endler 1986). The linear selection gradient measures the change in relative fitness for a unit change in the value of a trait. They followed Lande and Arnold (1983) in adopting the unit of trait standard deviation as the basis for comparison across different traits and populations. Despite the claim about weak selection in the abstract, the text of the paper contained just two brief statements about selection strength, on pages 250–251 and 253. The median strength of selection of 0.16 is termed "rather modest," and values greater than 0.5 are "very strong," but no discussion is included of criteria for deciding what constitutes strong or weak selection. Their argument for the rarity of strong selection is instead based the approximately exponential distribution of the absolute values of the gradients with a mode at 0. This change of emphasis from that implied by the title and abstract is made explicit in the Methods section, where Kingsolver et al. state "the overall 'average' strength of selection is unlikely to be very informative. We focus our analyses on the distributions of selection strengths" (p. 248). The conclusion drawn from this analysis would be more accurately stated as "most selection estimates are much smaller than the extreme estimates."

With respect to the strength of selection, therefore, Kingsolver et al. (2001) imply that they have a theoretical context, but then decline to apply any theory or concepts related to the units of measurement. Symptomatic of this problem is that the units of measurement, although clearly stated in the Methods section, are never attached to any of the actual measurements in the paper. If the units are stated, Kingsolver et al. (2001) statements that the median strength of selection, a 16% change in relative fitness with a one-standard-deviation change in the trait, is modest, whereas a change of 50% with a one-standard-deviation change is very strong selection, sound rather contradictory.

Clearly, for comparison of the strengths of selection on different traits in different organisms, some standard scale must be adopted, but which scale? Hereford et al. (2004) proposed two criteria for judging the strength of selection and showed that each is naturally addressed on a different scale. Under frequency-independent selection, the strength of selection is naturally viewed as a function of the adaptive landscape, which is external to the population subject to selection. Keeping this theoretical context in mind makes clear that standardizing the selection gradient by the trait standard deviation has problems similar to those of basing a measure of evolvability on heritability. To get at this concept of the steepness of the adaptive landscape in the neighborhood of the population, the variance-standardized measure immediately requires a second measure of the variation in the trait and in fitness itself to reveal whether a large value of the standardized gradient is due to a steep landscape in a population with small variation or to a population with a large amount of variation on a relatively flat landscape.

Hereford et al. (2004) instead standardized the selection gradient by the trait mean, which gives the change in relative fitness for a proportional change in trait. On a mean-standardized scale, the strength of selection on fitness itself is one, providing a benchmark for strong selection (Hansen et al. 2003). Hereford et al. (2004) showed that the median mean-standardized selection gradient was 0.31, or 31% of the strength of selection on fitness, and on this basis concluded that the results were "notable for the extremely strong selection observed" (p. 2141).

As noted above, however, an alternative idea of strength of selection could be based on the change of fitness within the range of the population—that is, strong selection occurs when the expected fitnesses of individuals within the population are typically very different (Hereford et al. 2004). In this theoretical context, a variance-standardized ruler is no longer rubber, but tells us just what we need to know. For example, in a population with a range of phenotypes of four standard deviations (readily found within a normally distributed population of modest population size), the median variance-standardized selection gradient from Kingsolver et al. (2001) of 0.16 predicts that a typical least-fit individual two standard deviations below the mean has a fitness only 52% that of the typical most-fit individual, two standard deviations above the mean. This again sounds like strong selection, contrary to the conclusions of Kingsolver et al. (2001), but for completely different reasons than in the mean-standardized case. Which scale is used really does matter: if the fitness landscapes were left unchanged, but the traits studied had much smaller coefficients of variation, the differences in fitness on a variance-standardized scale would also have been much smaller.

EXAMPLE 4: RESPECT SCALE TYPE

Fitness is one of the most fundamental quantitative concepts in biology. It predicts the evolutionary outcome of competition among genotypes or phenotypes for representation in the next generation. Two types of fitness measures are commonly used in biology, Wrightian fitness w and Malthusian fitness m. Wrightian fitness naturally arises in models where changes in gene or genotype frequencies are predicted by the equation

$$p' = \frac{pw}{\overline{w}},$$

where p is the relative frequency of a genotype before selection, p' the frequency after selection, and $\overline{w}$ the mean fitness of the population. The biological meaning of Wrightian fitness is contained in the ratio of fitness measures, as reflected in the use of "relative fitness" in population genetic theory. Wagner (2010) showed that Wrightian fitness, w, is a ratio-scale measure under the assumption that the time scale is fixed. We relax this assumption here because conclusions about fitness can be drawn on any time scale. Fitness, therefore, becomes a log-interval-scale variable, as the conclusions are invariant to power transformations of fitness, so we are equally entitled to draw conclusions about two or more episodes of selection. In this case, the exponent is determined by the time scale. The result of selection is therefore invariant to a transformation: $w \rightarrow aw^b, a, b > 0,$ but not to the transformation $w \rightarrow w+c$.

The Malthusian parameter m measures fitness in continuous time. If the age structure of the population is stable, the instantaneous rate of change in genotype frequency is given by the Crow-Kimura differential equation:

$$\dot{p} = p(m - \overline{m}).$$

The Malthusian fitness measure m is approximately related to Wrightian fitness (under weak selection) by $m = \ln w$. In this case, the biological meaning of the fitness measures is contained in the differences of the fitness values, rather than the ratios, and thus m is an interval-scale variable. The permissible scale transformations are therefore of the form $m \rightarrow am + b$, where a determines the change in time scale. This conclusion is of course unsurprising as m is the log of w and w is a log-interval-scale variable. A somewhat counterintuitive consequence of this mathematical fact is that only the differences between ms have evolutionary meaning; neither the sign nor the absolute magnitude is meaningful for evolution, although they have ecological meaning as growth rates.

The difference in scale types of fitness measures has important consequences in experimental practice that have not always been respected. An example of the problem arises in Remold and Lenski's (2001) study of the causes of genotype-by-environment interaction in *Escherichia coli*. Remold and Lenski studied the effects of single mutations on fitness when bacteria were grown in different environments. They used a fitness assay in which a mutant strain, M, competed against a standard strain, S, starting at low density and equal frequencies of the two genotypes (Lenski 1988). They estimated the initial and final frequencies by spreading samples on agar plates and recording the number of colonies of each type. These counts are used to estimate the Wrightian fitness of genotype x as $N'_x/N_x = w_x$ where N_x is the initial count and N'_x is the final count. To estimate the Malthusian fitness, m, Remold and Lenski treated the changes in bacterial density as the result of an exponential growth process

$$N_x(t) = N_x(0)e^{mt}$$

(Lenski 1988). One can therefore use the counts of bacterial cultures N_x and N'_x to estimate m for each genotype as

$$\ln\left[\frac{N'_M}{N_M}\right] = \ln[w_M] = m_M$$

.

$$\ln\left[\frac{N'_S}{N_S}\right] = \ln[w_S] = m_S.$$

Recall that the biological meaning is in the magnitude of the difference, $m_M - m_S$, rather than in their ratios. Remold and Lenski (Lenski et al. 1991; Remold and Lenski 2001), however, took their ratio to define a "relative fitness":

$$f_{M,S} = \frac{m_M}{m_S}.$$

The measure $f_{M,S}$ is still a measure of relative fitness in the sense that, if genotype M has higher fitness than S, then $f_{M,S}>1$, and if a third genotype R has higher fitness than M, then $f_{R,S}>f_{M,S}$. This measure preserves the order of fitness values between genotypes and can therefore be used as an ordinal-scale variable, but the magnitude of these $f_{M,S}$ values has no biological meaning. This problem is largely innocuous if $f_{M,S}$ values are used only to test for the existence of differences in fitness in a single environment, say between an ancestral and a derived genotype, but $f_{M,S}$ does not measure the magnitude of the fitness difference and is therefore meaningless when we compare fitness in different environments, as Remold and Lenski (2001) did.

To make the problem concrete, imagine experiments where genotypic fitnesses are measured in only two environments and the fitnesses of the two genotypes in the environments are different, $m_{M\cdot1} \neq m_{M\cdot2}$ and $m_{S\cdot1} \neq m_{S\cdot2}$. First assume that the differences between the fitness values in the two environments are the same, $m_{M\cdot1} - m_{S\cdot1} = m_{M\cdot2} - m_{S\cdot2}$, so that the allele frequencies follow exactly the same trajectory in both environments. If we compare the alternative relative fitness measures based on ratios of ms, we find that $f_{M,S\cdot1} \neq f_{M,S\cdot2}$, erroneously suggesting the presence of genotype-by-environment interaction. For example, if we assume that the Malthusian fitnesses are $m_{M\cdot1} = 0.02$, $m_{S\cdot1} = 0.03$, $m_{M\cdot2} = 0.01$, and $m_{S\cdot2} = 0.02$, the fitness differences within each environment is 0.01, and genotype and environment do not interact because selection proceeds exactly the same in each environment. Using these numbers, however, $f_{M,S\cdot1} = 2/3$ and $f_{M,S\cdot2} = 1/2$, leads us to conclude that a genotype-by-environment interaction is present. If, on the other hand, $m_{M\cdot1} = 0.015$, $m_{S\cdot1} = 0.03$, $m_{M\cdot2} = 0.01$, and $m_{S\cdot2} = 0.02$, the ratios are constant, leading one to infer an absence of interactions, when the differences show a true genotype-by-environment interaction. Using this approach, Remold and Lenski (2001) concluded that no evidence supported genotype-by-environment interactions involving temperature, but that interactions involving the type of carbon resource offered were abundant. Neither of these conclusions is justified.

Remold and Lenski (2001) came to meaningless conclusions because they used ratios to interpret data on an interval scale type. Respect for scale type is essential for drawing inferences about an empirical system from the measurements obtained by experiment. In this case, the actual measurements were appropriate to the question, and the measurement error came from a mismatch between the summary statistic chosen and the scale type.

example 5: do not let statistics overrule meaning

An example of the conflict between statistics and meaning is a study of the relationship between morphological divergence and divergence in genetic variance matrices performed by Podolsky et al. (1997). The theoretical context for this study is the prediction of longer-term evolutionary potential from estimates of the within-population genetic variation. That genetic variation is summarized in a matrix, **G**, which contains the additive genetic variances for each traits and the covariances between them. The Lande model (Lande 1979) predicts how variation affects divergence over many generations if **G** remains sufficiently stable over the relevant time scale. The question of whether stability is expected or observed is still largely unresolved (Steppan et al. 2002; Arnold et al. 2008), as **G** matrices are complex objects that are difficult to estimate precisely. The Podolsky et al. study was a relatively early attempt to address this question in a maximum-likelihood framework and had a positive impact by highlighting the problems of statistical power in such studies.

Podolsky et al. (1997) compared population means and **G** matrices for lengths of nine structures in 11 populations of the plant *Clarkia dudleyana.* They judged disparity between matrices using the average squared difference between their elements. Divergence between populations was measured as Mahalanobis distance, D^2, which is the distance in variance units in the direction between the two samples in multivariate space. The problematic aspect of their analysis is that they used a wide variety of transformations of the underlying data before estimating the **G** matrices; five of the nine lengths were untransformed, three were square-root transformed, and one was transformed as $y = \ln(\ln(x) + 1)$. Podolsky et al.'s (1997) stated goal in choosing transformations was to transform "to normality as feasible" (p. 1787). Departures from normality can cause serious estimation problems for maximum-likelihood algorithms based on the Gaussian likelihood, and the transformations therefore have a good statistical justification given that the method of analysis is tailored to normally distributed data.

On the other hand, transformations also alter the relationships between means and variances, which were the subject of Podolsky et al.'s (1997) analysis. Equalization of variances among groups in an ANOVA is an alternative reason for choosing particular transformations and one that most statisticians regard as more important than transformation to normality. For example, size, such as the lengths of body parts used in this study, are often approximately lognormally distributed, so variance and covariances scale with the mean on the linear scale, but would be independent on the log scale. By choosing a variety of transformations of the variables with regard to normality of residuals, Podolsky et al. altered the mean-variance relationships that they wished to study in different ways for different traits. These transformations affected both the distance used to judge divergence of means and the estimates of **G** whose disparity was predicted. Calculated from the transformed data, D^2 combined data from many incompatible scales, rendering any test of mean-variance relationships meaningless; if different transformations had been applied, the results could very well have been different (Adams et al. 1965).

We do not mean to suggest that no transformations should ever be used. A common and principled type of transformation would be to log transform all of the data, converting to a different but equally valid scale type (interval or difference), where differences play the role of ratios on the original scale. In other theoretical contexts, transformations to still other scales would be appropriate. For example, if the investigator had valid reasons to believe that natural selection had a simple relationship with the square root of length, square-root transformation would be justified in a study of natural selection. Rarely, however, are such empirically based arguments offered for the use of anything other than a logarithmic transformation.

Transformations very frequently change the scale type and, therefore, the meaning of measurements. In many cases, the result is meaningless tests of hypotheses. In this case, the error was a general one of comparing quantities that were incommensurate because transformation placed them on different scales.

Example 6: Treat Measurements as Measurements

An exceedingly common measurement error in biology is the neglect of units, which changes the carefully gathered measurements into a mere set of numbers. This error is obvious to the prepared mind when a table of naked numbers is shamelessly paraded in front of us. Sometimes the units of each measurement can be divined, by laborious shuffling from methods to appendices to tables to results, but this is a chore no reader should be subjected to. A deeper symptom of the insidious neglect of units is evident from the curious case of quadratic selection gradients.

Lande and Arnold (1983) showed that the relationship between fitness and trait values could be approximated from linear and quadratic regression coefficients. If z is the deviation of an individual trait from the population mean, then the Lande-Arnold model rewrites relative fitness as

$$w = \alpha + \beta z + \frac{1}{2}\gamma z^2 + \varepsilon$$

where α is mean fitness and ε is residual variation. Extension to the multivariate case is straightforward. The linear gradient, β, is necessary for prediction of the short-term response of the mean to natural selection, whereas the more complete description of the selection surface including γ is important for understanding how variance is changed by selection (Lande 1980), for visualizing the selective surface (Phillips and Arnold 1989), and for understanding whether the one-generation prediction of the response to selection is likely to hold over longer time scales. The quadratic version of the Lande-Arnold model has been widely used. Kingsolver et al. (2001) found 573 estimates of γ in a sample of 63 studies, and Stinchcombe et al.'s (2008) much more limited survey of one journal from 2002 through 2007 turned up 32 additional studies with 673 estimates of γ.

Surprisingly, Stinchcombe et al. (2008) discovered that the majority of published estimates of univariate γ were off by a factor of 2. The source of this error is that two different models are used in the Lande-Arnold approach: conceptually, Lande and Arnold (1983) preferred to think of the average curvature of the selective surface around the population mean and so defined the parameter γ as the second derivative of the fitness landscape; fitness is a function of (½)γ. On the practical side, Lande and Arnold showed that the multiple regression of trait values on relative fitness can be used to estimate the parameters β and γ in the above model. Regression models are parameterized

$$w = \alpha + bz + qz^2 + \varepsilon,$$

so that, although $b{=}\beta$, $q{=}1/2\gamma$. Stinchcombe et al. (2008) estimated the proportion of studies that reported q as γ by asking the authors of a sample of papers how they calculated γ. A stunning 78% of the authors reported using q as γ.

How is it that hundreds of papers containing thousands of estimates can be published over 27 years before anyone notices that the majority of the estimates are wrong? The answer seems to be a systemic lack of respect for measurement and models in biology. Despite the wide use of the Lande-Arnold method, the interest of most users has been in determining whether linear or quadratic selection can be detected by a statistical test for selection, not in the actual strength or shape of selection. Exceedingly few authors have used the Lande-Arnold parameters β or γ to make predictions about evolution or variation, and even fewer have tried to test those predictions (for exceptions see Postma et al. 2007). The numerical estimates have served a mainly decorative purpose.

Biology is in general poised between being a descriptive, qualitative science and a quantitative one. Many questions have only qualitative answers: Is this organism known to science or undescribed? Which piece of land should be purchased to further conservation goals? Other attributes of nature may productively be treated as qualitative, even though underlain by quantitative reality. For example, knowing whether we can reject the idea that a particular bit of DNA is evolving at the neutral rate is useful. Consequently, we sometimes forget that there are cases where quantity matters, such as quadratic selection gradients. The lack of attention to the definition of γ has damaged the meaning of 20 years of published estimates. This problem matters. Stinchcombe et al. demonstrated that this numerical error can suggest that multivariate fitness landscapes have a shape qualitatively different from their real one. Many hypotheses about the maintenance of genetic variation depend on the actual value of γ (see, e.g., Turelli 1984).

example 7: know what your parameters mean

Our final two examples are cases where parameters of models are assigned a meaning that they do not actually have. A good example is the idea that negative genetic correlations were the necessary consequence of an evolutionary theory of limits on life history, which somehow became widespread during the 1980s (see, e.g., Bell and Koufopanou 1986). This misconception occasioned considerable debate and additional experiments when the data did not conform to this expectation (see, e.g., Rose 1984; Reznick 1985). The entire controversy rests on a misinterpretation of what a genetic correlation, and the genetic covariance it standardizes, means. Genetic covariance quantifies the degree of dependence of one trait on another. No discontinuous change in the response to selection accompanies a change in sign of a genetic correlation (Via and Lande 1985; Charlesworth 1990; Houle 1991; Fry 1993). When the covariance goes from negative to positive, the correlated effects of selection simply go from slightly negative to slightly positive. Instead, the conclusion that tradeoffs are not perfect (because correlations are >-1) should have been apparent from the beginning. Half a generation of biologists working on the genetics of life histories spent too much of their time worrying that their estimates of correlations were not negative enough because some referred a reasonable hypothesis (that tradeoffs are important) to an irrelevant attribute (the sign of the genetic correlation). The hypothesis of constraint directly predicts a boundary beyond which fitness cannot evolve. It does not predict genetic correlations without subsidiary assumptions, such as a lack of deleterious mutations.

example 8: make meaningful measures

One of our roads to discovering measurement theory came through a desire to represent the evolutionary effects of epistasis (Wagner et al. 1998; Hansen and Wagner 2001b). Fisher's definition of additive effects has proven extremely fruitful because it was defined to quantify the importance of parent-offspring resemblance in the response to selection. In contrast, the statistical measures of gene interaction (Fisher 1918; Cockerham 1954; Kempthorne 1954)—dominance variance, additive-by-additive variance, and additive-by-dominance variance, among others—have not proven useful for understanding of the dynamical role of gene interaction in evolution because they represent a statistical measure of departures from simpler models in an ANOVA framework rather than a measure of the dynamical consequences of epistasis. In spite of this, the use of statistical measures of epistasis has persisted over many years because these were for a long time the only measures of epistasis that had been suggested. As a result, when biologists wished to study the interesting phenomenon of epistasis, for example, inspired by Wright's shifting balance theory, they naturally turned to the statistical measures, even in the absence of explicit theory to justify their relevance. As a result, the statistical measures of epistasis have been assigned various meanings that they do not possess, with effects similar to the case of the sign of genetic correlations. For example, the "statistical" conceptualization excludes systematic directional effects by definition and has lead to the widespread but incorrect notion that the influence of epistasis on the dynamics of quantitative characters under selection can be ignored (see Carter et al. 2005; Hansen et al. 2006; Pavlicev et al. 2010). The only explicit dynamical role suggested for statistical epistatic variance was that it could be converted to additive variance by genetic drift during population bottlenecks and thus boost evolvability during founder events (see, e.g., Goodnight 1987; Cheverud and Routman 1996). This interpretation is misleading because, although epistasis causes additive effects to change when allele frequencies change, the increase in additive variance is not proportional to any decrease in epistatic variance components (Barton and Turelli 2004). The change in additive variance under genetic drift depends strongly on the pattern of epistasis and the additive variance may even decrease in expectation under some kinds of epistasis (Hansen and Wagner 2001b).

Seeing this, however, required defining new measures of epistasis that capture dynamically important properties. A nonstatistical representation of epistasis had been proposed by Cheverud and Routman (1995), and Wagner et al. (1998) followed this proposition by defining epistasis in terms of changes in the effect of allele substitutions with changes in the genetic background (i.e., substitutions at other loci). Hansen and Wagner (2001b) developed the multilinear representation of the genotype-phenotype map along these lines and showed how the new representation of epistasis can be related to the statistical representation used in quantitative genetics (for further development see Álvarez-Castro and Carlborg 2007). Using this model, we could show how different patterns of epistasis affect evolutionary dynamics (Hansen and Wagner 2001a,b; Hermisson et al. 2003; Carter et al. 2005; Hansen et al. 2006; Fierst and Hansen 2010). An important result is that epistasis can accelerate a response to selection in the presence of a systematic pattern of positive interactions between the effects of substitutions at different loci. Conversely, a systematic pattern of negative interactions leads to canalization. Carter et al. (2005) identified a "directional" epistatic parameter that measures these effects. This directional parameter is dynamically meaningful in a way similar to that of Fisher's additive effect and additive variance. It captures the second-order effect of evolutionary changes in the additive effects, as shown in Figure 3.

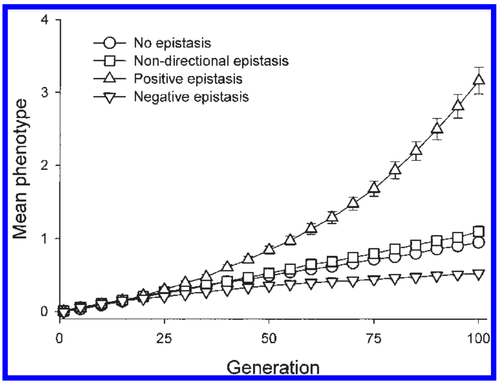


Figure 3. Effect of Pairwise Epistasis on response to Directional Selection

Each trajectory is the mean of 100 replicate individual-based simulations; the bars show ± 1 S.E. The error bars are smaller than the symbols for most of the cases. Each simulation models populations of 1000 individuals with 20 loci influencing a single trait with identical allele frequencies and allelic effects that furnish an initial additive genetic variance of 1.0. Mutation is absent. Populations were subjected to 100 generations of positive directional selection of strength 1% change in relative fitness per trait unit. Directional epistasis is defined as in Carter et al. (2005), where μ_ε is the mean strength of directional epistasis, and σ_ε is the standard deviation of the directional epistasis between pairs of loci. The four cases shown are: No epistasis $\mu_\varepsilon = 0$, $\sigma_\varepsilon = 0$; Nondirectional epistasis $\mu_\varepsilon = 0$, $\sigma_\varepsilon = 1.41$; Positive epistasis $\mu_\varepsilon = 1$, $\sigma_\varepsilon = 1$; and Negative epistasis $\mu_\varepsilon = -1$, $\sigma_\varepsilon = 1$.

What Can We Do?

The problems that we have described are widespread in biology. We have chosen examples mostly from the evolutionary biology literature because that is our own area of expertise, and those areas are the ones in which we can more confidently diagnose the issues raised by each example. We strongly suspect that ecologists, for example, will find similar problems with the measurements meant to instantiate such concepts as population size, density dependence, interaction strength, and productivity (see, e.g., Wulff 2001; Vik et al. 2004).

Some readers may feel that our examples are simply unfortunate isolated errors by individual researchers. Such errors are common, but ultimately inconsequential because of the self-correcting nature of science. We emphasize that the examples we discuss are symptomatic of *systemic* errors that reflect widely held beliefs and attitudes. Many of our examples involve highly competent, indeed leading, researchers, and the problems are repeated in many other studies. For example, the much-criticized errors in Harvey and Clutton-Brock's (1985) compilation of primate body-size data might be treated as their responsibility alone, but Smith and Jungers (1997) traced the way these data were transferred from source to source and often lost their meaning and context along the way. Eventually, in their final destination, the result was absurdities such as assigning six "means" to males and females of three species of *Ateles* on the basis of what were originally measurements from a total of three individuals. The shortcomings and errors by Harvey and Clutton-Brock (1985) arose because the "means" were repeatedly presented without standard errors, sample sizes, or references to original sources, and the errors were thus allowed to propagate (Smith and Jungers 1997). This case is a scandal because of the long chain of authors who succumbed to the temptation to use data without checking original sources, and all of the reviewers and editors who did not object. Similarly, the ideas that the sign of genetic correlations was an important attribute and that heritability predicts evolvability became problematic not when the first author made this claim, but as the number of those who accepted the incorrect premise grew; these are systemic problems.

A related objection is that our extension of measurement theory to incorporate the theoretical context and hypothesis generation is not helpful because we are discussing "good science, clear thinking, or an appreciation for the subtlety of an argument"—things that good scientists already know and value. We agree that this is precisely what we are doing, but disagree that taking a measurement-theory stance is not helpful. We have known for years about many of the examples we discuss above, but we find that the terminology and mindset of measurement-theoretic thinking allow us to diagnose and discuss common features of these cases that combine to make them examples of flawed or useless science. For example, we find clarifying the ability to declare that using the sign of a genetic correlation to indicate whether a tradeoff exists is an attribute error. Our claim is not that conceptual measurement theory is revolutionary, but that it aids clear thinking about good science.

The most practical counterweight to the many possible pitfalls that accompany the task of measurement is awareness. If our experience is any indication, as soon as one becomes explicitly aware of measurement as a discrete topic, much unsatisfactory science is readily diagnosed in those terms. In this way, measurement theory can become a routine background to reading and reviewing the literature, discussing it with colleagues, and designing research strategies. Once a culture of thinking about meaning and measurement is in place, many of the types of errors that we have documented will be more readily avoided during the design and analysis of experiments, caught by reviewers before they are published, or quickly noticed when they are.

A second way to address the measurement problem is through education. Most graduate programs include a required background of at least one course in statistics, but we have never come across a course in meaningfulness or even measurement theory—it is curious that so little effort is spent learning what our conclusions might mean, but substantial effort is spent on how to quantify those conclusions. We hope that someday the theory of measurement and meaning will have its place in scientific education alongside, and as a partial counterweight to, statistics. Looking farther down the road, if explicit measurement theory takes root in biology, we can look forward to a time when a backlog of well-studied biological examples will be available as a familiar entrée to the study of measurement and meaning.

In the meantime, we offer a modest list of measurement principles as an aid to thinking about both good and bad measurement:

1. Keep theoretical context in mind.
2. Honor your family of hypotheses.
3. Make meaningful definitions.
4. Know what the numbers mean.
5. Remember where the numbers come from.
6. Respect scale type.
7. Know the limits of your model.
8. Never substitute a test for an estimate.
9. Clothe estimates in the modest raiment of uncertainty.
10. Never separate a number from its unit.

ACKNOWLEDGMENTS

Our understanding of the ideas in this paper was developed in discussion with members of the Hansen and Houle laboratories and with numerous visitors to the Centre for Ecological and Evolutionary Synthesis (CEES) at the University of Oslo, Norway, and the National Evolutionary Synthesis Center (NESCENT) at Duke University, U.S. We thank Stevan J. Arnold, Donald Griffin, Øistein Holen, Mihaela Pavlicev, Jura Pintar, Claus Rueffler, Tore Schweder, Patrick Phillips, Michael Wade, and two anonymous reviewers for their comments on the manuscript. Arnaud LeRouzic performed the simulations in Figure 3. David Houle was supported in this research by a visiting professorship at CEES and a visiting scientist appointment at NESCENT.

REFERENCES

Adams E. W., Fagot R. F., Robinson R. E. 1965. A theory of appropriate statistics. *Psychometrika* 30: 99–127.

Alroy J. 1998. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. *Science* 280:731–734.

Álvarez-Castro J. M., Carlborg Ö. 2007. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* 176:1151–1167.

Anderson D. R., Burnham K. P., Thompson W. L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* 64:912–923.

Arnold S. J., Bürger R., Hohenlohe P. A., Ajie B. C., Jones A. G. 2008. Understanding the evolution and stability of the **G**-matrix. *Evolution* 62:2451–2461.

Baldwin J. M. 1896. A new factor in evolution. *American Naturalist* 30:441–451, 536–553.

Barton N. H., Turelli M. 2004. Effects of genetic drift on variance components under a general model of epistasis. *Evolution* 58:2111–2132.

Bell G., Koufopanou V. 1986. The cost of reproduction. Pages 83–131 in *Oxford Surveys in Evolutionary Biology*, Volume 3, edited by R. Dawkins and M. Ridley. Oxford (UK): Oxford University Press.

Borenstein E., Meilijson I., Ruppin E. 2006. The effect of phenotypic plasticity on evolution in multipeaked fitness landscapes. *Journal of Evolutionary Biology* 19:1555–1570.

Boring E. G. 1945. The use of operational definitions in science. *Psychological Review* 52:243–245.

Breiman L. 2000. Statistical modeling: the two cultures. *Statistical Science* 16:199–215.

Bridgman P. W. 1922. *Dimensional Analysis.* New Haven (CT): Yale University Press.

Bridgman P. W. 1927. *The Logic of Modern Physics.* New York: MacMillan Company.

Broom M., Speed M. P., Ruxton G. D. 2005. Evolutionarily stable investment in secondary defences. *Functional Ecology* 19:836–843.

Butler J. P., Feldman H. A., Fredberg J. J. 1987. Dimensional analysis does not determine a mass exponent for metabolic scaling. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 253: R195–R199.

Cameron E. Z., du Toit J. T. 2007. Winning by a neck: tall giraffes avoid competing with shorter browsers. *American Naturalist* 169:130–135.

Carter A. J. R., Hermisson J., Hansen T. F. 2005. The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theoretical Population Biology* 68:179–196.

Chamberlin T. C. 1890. The method of multiple working hypotheses. *Science* 15:92–96.

Chamberlin T. C. 1965. The method of multiple working hypotheses: with this method the dangers of parental affection for a favorite theory can be circumvented. *Science* 148:754–759.

Charlesworth B. 1990. Optimization models, quantitative genetics, and mutation. *Evolution* 44:520–538.

Charnov E. L. 1976. Optimal foraging, the marginal value theorem. *Theoretical Population Biology* 9:129–136.

Charnov E. L. 1993. *Life History Invariants: Some Explo-*

rations of Symmetry in Evolutionary Ecology. Oxford (UK): Oxford University Press.

Cheverud J. M., Routman E. J. 1995. Epistasis and its contribution to genetic variance components. *Genetics* 139:1455–1461.

Cheverud J. M., Routman E. J. 1996. Epistasis as a source of increased additive genetic variance at population bottlenecks. *Evolution* 50:1042–1051.

Cockerham C. C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics* 39:859–882.

Conner J. K. 2001. How strong is natural selection? *Trends in Ecology and Evolution* 16:215–217.

Darwin C. 1871. *The Descent of Man, and Selection in Relation to Sex*. London: John Murray and Sons.

Diaz M. C., Rützler K. 2001. Sponges: an essential component of Caribbean coral reefs. *Bulletin of Marine Science* 69:535–546.

Dudley R. 2000. *The Biomechanics of Insect Flight: Form, Function, Evolution*. Princeton (NJ): Princeton University Press.

Dunn E. H., Hussell D. J. T., Welsh D. 1999. Priority-setting tool applied to Canada's landbirds based on concern and responsibility for species. *Conservation Biology* 13:1404–1415.

Dytham C. 2003. *Choosing and Using Statistics: A Biologist's Guide*. Second Edition. Malden (MA): Blackwell Publishing.

Eberhard W. G. 2009. Static allometry and animal genitalia. *Evolution* 63:48–66.

Endler J. A. 1986. *Natural Selection in the Wild*. Princeton (NJ): Princeton University Press.

Fierst J. L., Hansen T. F. 2010. Genetic architecture and postzygotic reproductive isolation: evolution of Bateson-Dobzhansky-Muller incompatibilities in a polygenic model. *Evolution* 64:675–693.

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 (UK): Clarendon Press.

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

Frank S. A. 1995. George Price's contributions to evolutionary genetics. *Journal of Theoretical Biology* 175:373–388.

Frank S. A. 2009. The common patterns of nature. *Journal of Evolutionary Biology* 22:1563–1585.

Frankino W. A., Emlen D. J., Shingleton A. W. 2010. Experimental approaches to studying the evolution of animal form: the shape of things to come. Pages 419–478 in *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments*, edited by T. Garland, Jr. and M. R. Rose. Berkeley (CA): University of California Press.

Frankino W. A., Zwaan B. J., Stern D. L., Brakefield P. M. 2005. Natural selection and developmental constraints in the evolution of allometries. *Science* 307:718–720.

Frankino W. A., Zwaan B. J., Stern D. L., Brakefield P. M. 2007. Internal and external constraints in the evolution of morphological allometries in a butterfly. *Evolution* 61:2958–2970.

Fry J. D. 1993. The "general vigor" problem: can antagonistic pleiotropy be detected when genetic covariances are positive? *Evolution* 47:327–333.

Goodnight C. J. 1987. On the effect of founder events on epistatic genetic variance. *Evolution* 41:80–91.

Gould S. J. 1977. *Ontogeny and Phylogeny*. Cambridge (MA): Belknap Press.

Grafen A. 1990a. Biological signals as handicaps. *Journal of Theoretical Biology* 144:517–546.

Grafen A. 1990b. Sexual selection unhandicapped by the Fisher process. *Journal of Theoretical Biology* 144: 473–516.

Greenewalt C. H. 1975. The flight of birds: the significant dimensions, their departure from the requirements for dimensional similarity, and the effect on flight aerodynamics of that departure. *Transactions of the American Philosophical Society* 65: 1–67.

Gunther B. 1975. Dimensional analysis and theory of biological similarity. *Physiological Reviews* 55:659–699.

Gunther B., Morgado E. 2003. Dimensional analysis revisited. *Biological Research* 36:405–410.

Hand D. J. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 159:445–473.

Hand D. J. 2004. *Measurement Theory and Practice: The World Through Quantification*. London (UK): Arnold.

Hansen T. F., Álvarez-Castro J. M., Carter A. J. R., Hermisson J., Wagner G. P. 2006. Evolution of genetic architecture under directional selection. *Evolution* 60:1523–1536.

Hansen T. F., Houle D. 2008. Measuring and comparing evolvability and constraint in multivariate characters. *Journal of Evolutionary Biology* 21:1201–1219.

Hansen T. F., Pélabon C., Armbruster W. S., Carlson M. L. 2003. Evolvability and genetic constraint in *Dalechampia* blossoms: components of variance and measures of evolvability. *Journal of Evolutionary Biology* 16:754–766.

Hansen T. F., Wagner G. P. 2001a. Epistasis and the mutation load: a measurement-theoretical approach. *Genetics* 158:477–485.

Hansen T. F., Wagner G. P. 2001b. Modeling genetic architecture: a multilinear theory of gene interaction. *Theoretical Population Biology* 59:61–86.

Harvey P. H., Clutton-Brock T. H. 1985. Life history variation in primates. *Evolution* 39:559–581.

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

Hermisson J., Hansen T. F., Wagner G. P. 2003. Epistasis in polygenic traits and the evolution of genetic architecture under stabilizing selection. *American Naturalist* 161:708–734.

Heusner A. A. 1982. Energy-metabolism and body size. II. Dimensional analysis and energetic nonsimilarity. *Respiration Physiology* 48:13–25.

Heusner A. A. 1983. Body size, energy-metabolism, and the lungs. *Journal of Applied Physiology* 54:867–873.

Heusner A. A. 1984. Biological similitude: statistical and functional relationships in comparative physiology. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 246:R839–R846.

Hinton G. E., Nowlan S. J. 1987. How learning can guide evolution. *Complex Systems* 1:495–502.

Hone D. W. E., Keesey T. M., Pisani D., Purvis A. 2005. Macroevolutionary trends in the dinosauria: Cope's rule. *Journal of Evolutionary Biology* 18:587–595.

Houle D. 1991. Genetic covariance of fitness correlates: what genetic correlations are made of and why it matters. *Evolution* 45:630–648.

Houle D. 1992. Comparing evolvability and variability of quantitative traits. *Genetics* 130:195–204.

Houle D. 1998. How should we explain variation in the genetic variance of traits? *Genetica* 102–103: 241–253.

Houle D., Morikawa B., Lynch M. 1996. Comparing mutational variabilities. *Genetics* 143:1467–1483.

Huxley J. S. 1924. Constant differential growth-ratios and their significance. *Nature* 114:895–896.

Huxley J. S. 1932. *Problems of Relative Growth.* New York: L. MacVeagh, The Dial Press.

Jerison H. J. 1969. Brain evolution and dinosaur brains. *American Naturalist* 103:575–588.

Kempthorne O. 1954. The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 143:103–113.

Kingsland S. E. 1985. *Modeling Nature: Episodes in the History of Population Ecology.* Chicago (IL): University of Chicago Press.

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

Krantz D. H., Luce R. D., Suppes P., Tversky A. 1971. *Foundations of Measurement, Volume I: Additive and Polynomial Representations.* New York: Academic Press.

Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* 33:402–416.

Lande R. 1980. The genetic covariance between characters maintained by pleiotropic mutations. *Genetics* 94:203–215.

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

Lenski R. E. 1988. Experimental studies of pleiotropy and epistasis in *Escherichia coli.* I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution* 42:425–432.

Lenski R. E., Rose M. R., Simpson S. C., Tadler S. C. 1991. Long-term experimental evolution in *Escherichia coli.* I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138:1315–1341.

Levins R. 1962. Theory of fitness in a heterogeneous environment. I. Fitness set and adaptive function. *American Naturalist* 96:361–373.

Levins R. 1966. Strategy of model building in population biology. *American Scientist* 54:421–431.

Lewontin R. C. 1974. *The Genetic Basis of Evolutionary Change.* New York: Columbia University Press.

Logan M. 2010. *Biostatistical Design and Analysis Using R: A Practical Guide.* Hoboken (NJ): Wiley-Blackwell.

Lord F. M. 1953. On the statistical treatment of football numbers. *American Psychologist* 8:750–751.

Luce R. D. 1959. *Individual Choice Behavior: A Theoretical Analysis.* New York: Wiley and Sons.

Luce R. D., Krantz D. H., Suppes P., Tversky A. 1990. *Foundations of Measurement, Volume III: Representation, Axiomatization, and Invariance.* New York: Academic Press.

Maynard Smith J. 1976. Sexual selection and the handicap principle. *Journal of Theoretical Biology* 57: 239–242.

McGhee K. E., Fuller R. C., Travis J. 2007. Male competition and female choice interact to determine mating success in the bluefin killifish. *Behavioral Ecology* 18:822–830.

McMahon T. A., Bonner J. T. 1983. *On Size and Life.* New York: W. H. Freeman.

Michell J. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept.* Cambridge (UK): Cambridge University Press.

Mosimann J. E. 1970. Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. *Journal of the American Statistical Association* 65:930–945.

Narens L. 1981. On the scales of measurement. *Journal of Mathematical Psychology* 24:249–275.

Narens L. 1985. *Abstract Measurement Theory.* Cambridge (MA): MIT Press.

Paenke I., Sendhoff B., Kawecki T. J. 2007. Influence of plasticity and learning on evolution under directional selection. *American Naturalist* 170:E47–E58.

Pavlicev M., Le Rouzic A., Cheverud J. M., Wagner G. P., Hansen T. F. 2010. Directionality of epistasis in a murine intercross population. *Genetics* 185: 1489–1505.

Phillips P. C., Arnold S. J. 1989. Visualizing multivariate selection. *Evolution* 43:1209–1222.

Platt J. R. 1964. Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* 146:347–353.

Podolsky R. H., Shaw R. G., Shaw F. H. 1997. Population structure of morphological traits in *Clarkia dudleyana*. II. Constancy of within-population genetic variance. *Evolution* 51:1785–1796.

Pomiankowski A. 1987. Sexual selection: the handicap principle does work-sometimes. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 231:123–145.

Pomiankowski A. N. 1988. The evolution of female mate preferences for male genetic quality. Pages 136–184 in *Oxford Surveys in Evolutionary Biology*, Volume 5, edited by P. H. Harvey and L. Partridge. Oxford (UK): Oxford University Press.

Post E., Forchhammer M. C. 2002. Synchronization of animal population dynamics by large-scale climate. *Nature* 420:168–171.

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

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

Prothero J. 1986. Methodological aspects of scaling in biology. *Journal of Theoretical Biology* 118:259–286.

Prothero J. 2002. Perspectives on dimensional analysis in scaling studies. *Perspectives in Biology and Medicine* 45:175–189.

Remold S. K., Lenski R. E. 2001. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *Proceedings of the National Academy of Sciences USA* 98: 11388–11393.

Reznick D. 1985. Costs of reproduction: an evaluation of the empirical evidence. *Oikos* 44:257–267.

Roff D. A., Mousseau T. A. 1987. Quantitative genetics and fitness: lessons from *Drosophila*. *Heredity* 58: 103–118.

Rose M. R. 1984. Genetic covariation in *Drosophila* life history: untangling the data. *American Naturalist* 123:565–569.

Rosen R. 1962. The derivation of D'Arcy Thompson's theory of transformations from theory of optimal design. *Bulletin of Mathematical Biophysics* 24:279–290.

Rosen R. 1978a. Dynamical similarity and theory of biological transformations. *Bulletin of Mathematical Biology* 40:549–579.

Rosen R. 1978b. *Fundamentals of Measurement and Representation of Natural Systems*. New York: Elsevier North-Holland.

Sarle W. S. 14 September 1997. Measurement Theory: Frequently Asked Questions. ftp://ftp.sas.com/pub/neural/measurement.html. (Accessed 10 January 2008).

Schmidt-Nielsen K. 1984. *Scaling: Why is Animal Size So Important?* Cambridge (UK): Cambridge University Press.

Simmons R. E., Scheepers L. 1996. Winning by a neck: sexual selection in the evolution of giraffe. *American Naturalist* 148:771–786.

Smith R. J., Jungers W. L. 1997. Body mass in comparative primatology. *Journal of Human Evolution* 32:523–559.

Sneath P. H. A., Sokal R. R. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco (CA): W. H. Freeman.

Sokal R. R., Rohlf F. J. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. Third Edition. New York: W. H. Freeman.

Solow A. R., Wang S. C. 2008. Some problems with assessing Cope's Rule. *Evolution* 62:2092–2096.

Stahl W. R. 1961. Dimensional analysis in mathematical biology. I. General discussion. *Bulletin of Mathematical Biology* 23:355–376.

Stahl W. R. 1962. Similarity and dimensional methods in biology. *Science* 137:205–212.

Stephens D. W., Dunbar S. R. 1993. Dimensional analysis in behavioral ecology. *Behavioral Ecology* 4:172–183.

Steppan S. J., Phillips P. C., Houle D. 2002. Comparative quantitative genetics: evolution of the G matrix. *Trends in Ecology and Evolution* 17:320–327.

Stevens S. S. 1946. On the theory of scales of measurement. *Science* 103:677–680.

Stevens S. S. 1959. Measurement, psychophysics and utility. Pages 18–63 in *Measurement: Definitions and Theories*, edited by C. W. Churchman and P. Ratoosh. New York: Wiley.

Stevens S. S. 1968. Measurement, statistics, and the schemapiric view. *Science* 161:849–856.

Stinchcombe J. R., Agrawal A. F., Hohenlohe P. A., Arnold S. J., Blows M. W. 2008. Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing? *Evolution* 62:2435–2440.

Suppes P., Krantz D. H., Luce R. D., Tversky A. 1989. *Foundations of Measurement, Volume II: Geometrical, Threshold, and Probabilistic Respresentations*. New York: Academic Press.

Svennungsen T. O., Holen Ø. H. 2007. The evolutionary stability of automimicry. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 274: 2055–2063.

Takahashi M., Arita H., Hiraiwa-Hasegawa M., Hasegawa T. 2008. Peahens do not prefer peacocks with more elaborate trains. *Animal Behaviour* 75: 1209–1219.

Teissier G. 1936. Croissance comparée des formes locales d'une même espèce. *Mémoires du Musée Royal D'Histoire Naturelle de Belgique* 3:627–634.

Thompson D. W. 1917. *On Growth and Form.* Cambridge (UK): Cambridge University Press.

Turelli M. 1984. Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical Population Biology* 25: 138–193.

Via S., Lande R. 1985. Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution* 39:505–522.

Vik J. O., Stenseth N. C., Tavecchia G., Mysterud A., Lingjærde O. C. 2004. Ecology (communication arising): living in synchrony on Greenland coasts? *Nature* 427:697–698.

Wagner G. P. 2010. The measurement theory of fitness. *Evolution* 64:1358–1376.

Wagner G. P., Altenberg L. 1996. Perspective: complex adaptations and the evolution of evolvability. *Evolution* 50:967–976.

Wagner G. P., Laubichler M. D. 2001. Character identification: the role of the organsm. Pages 141–164 in *The Character Concept in Evolutionary Biology*, edited by G. P. Wagner. San Diego (CA): Academic Press.

Wagner G. P., Laubichler M. D., Bagheri-Chaichian H. 1998. Genetic measurement theory of epistatic effects. *Genetica* 102–103:569–580.

Weber K. E. 1990. Selection on wing allometry in *Drosophila melanogaster*. *Genetics* 126:975–989.

Weber K. E. 1992. How small are the smallest selectable domains of form? *Genetics* 130:345–353.

Weitzenhoffer A. M. 1951. Mathematical structures and psychological measurements. *Psychometrika* 16: 387–406.

White J. F., Gould S. J. 1965. Interpretation of the coefficient in the allometric equation. *American Naturalist* 99:5–18.

Whitlock M. C., Schluter D. 2009. *The Analysis of Biological Data.* Greenwood Village (CO): Roberts and Company.

Wiley R. H., Poston J. 1996. Perspective: indirect mate choice, competition for mates, and coevolution of the sexes. *Evolution* 50:1371–1381.

Wilkinson G. S. 1993. Artificial sexual selection alters allometry in the stalk-eyed fly *Cyrtodiopsis dalmanni* (Diptera: Diopsidae). *Genetical Research* 62:213–222.

Wolman A. G. 2006. Measurement and meaningfulness in conservation science. *Conservation Biology* 20:1626–1634.

Wulff J. 2001. Assessing and monitoring coral reef sponges: why and how? *Bulletin of Marine Science* 69:831–846.

Yoccoz N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72:106–111.