May 13, 2001 (with corrections added 5/5/03)

CH 17. Fluctuating Asymmetry Analyses Revisited

A. Richard Palmer and Curtis Strobeck

Department of Biological Sciences

University of Alberta

Edmonton, Alberta T6G 2E9 CANADA

SUGGESTED RUNNING HEAD: Fluctuating asymmetry analyses

In: Developmental Instability (DI): Causes and Consequences (2003)

(Editor: Michal Polak, Publisher: Oxford University Press) pp. 279-319

EDITORIAL CORRESPONDENCE:

Until June 22, 2001:
A. Richard Palmer
Department of Biological Sciences
University of Alberta
Edmonton, Alberta T6G 2E9
CANADA

phone: 403-492-3633, 3308
FAX: 403-492-9234
E-mail: Rich.Palmer@UAlberta.CA

June 22 - Sept. 1, 2001:
A. Richard Palmer
Bamfield Marine Station
Bamfield, British Columbia V0R 1B0
CANADA

250-728-3301
250-728-3452
Rich.Palmer@UAlberta.CA

DEDICATION

This contribution is dedicated to F. James Rohlf on the occasion of his 65th birthday. We should all emulate his passion for understanding the tools of his trade and for sharing that understanding so generously with students, colleagues and friends.

ABSTRACT

In spite of a decade of furious activity and an increasingly bewildering array of analytical methods, essential requirements for a robust study of fluctuating asymmetry have not changed: judicious choice of traits, meticulous attention to measurement precision, visual inspection and tests for dubious data, appropriate tests and corrections for size-dependence, confirmation that subtle deviations from symmetry both exceed those expected due to measurement error and meet the criteria for ideal fluctuating asymmetry, and an open mind about alternative hypotheses. We review these requirements and try to clarify why they are so essential.

Studies of fluctuating asymmetry face a number of serious challenges: a) random phenotypic variation arises for reasons other than developmental instability, b) all descriptors of FA estimate a variance and variances are estimated with much lower confidence than means (i.e., repeatability is lower), c) subtle departures from symmetry are typically so minute they are exceedingly difficult to measure reliably, d) measurement error and trait size interact in complex and mischievous ways, and e) tests for departures from normality are uncomfortably weak for small to modest sample sizes. We outline the foundations of these challenges and some of the ways they may be addressed.

Persistent efforts to improve analytical tools nonetheless have yielded some useful advances: a) log transformations help remove the size-dependence of subtle asymmetries and the heterogeneity of variance that can arise from this size-dependence, b) proper critical values for the kurtosis statistic provide more reliable statistical tests, c) indexes that combine information from multiple traits yield more reliable estimates of individual developmental instability, and d) a generalization of Levene's test improves both the ease and the power of analyses testing for differences in fluctuating asymmetry among individuals, traits or groups.

Finally, as an appendix, we provide a detailed worked example, with commentary, of a complete FA analysis. It outlines how care and common sense in preliminary analyses greatly improve the rigor of the final results. This appendix, and the data files for the analyses, are available as web supplements.

ABBREVIATIONS

ANOVA: analysis of variance

DA: directional asymmetry

df: degrees of freedom

DI: developmental instability

FA: fluctuating asymmetry, average unsigned deviation from symmetry.

FA1, FA4, FA10, etc.: various fluctuating asymmetry indexes as numbered in Palmer (1994); see also Tables 1, 2.

M_1, M_2, M_3, etc.: a series of replicate measurements of a given trait on a given individual

ME: measurement error

ME1, ME2, ME3, etc.: various measurement error indexes as numbered in Table 3.

MS: mean squares

MS_{SI}: between-sides mean squares from a sides by individuals ANOVA

SD: standard deviation

R - L: right minus left

X_i: observation X on individual i

$\overline{X}$: mean of a sample of individual observations (X_i)

OUTLINE

IV.B) MULTIVARIATE MEASURES OF INDIVIDUAL FA

IV.B1) Why combine information from multiple traits?

IV.B2) Combining information from multiple traits- prior methods

IV.B3) Combining information from multiple traits- average proportional FA

IV.B4) Landmark methods

V) ANALYSES OF FA VARIATION: VALIDATING THE DATA

V.A) MEASUREMENT ERROR & REPEATABILITY

V.A1) Measurement error- the problem

a) Misinterpreting ME as DI

b) ME and FA are often comparable

c) ME artificially inflates FA

d) ME can not be partitioned out of individual FA for single trait

e) ME outliers create leptokurtosis

f) High ME can create artificial size-dependent FA

g) ME obscures FA variation

V.A2) Measurement error- description

V.A3) Perils of repeatability

V.A4) Hypothetical repeatability (*R*) and among-individual variation in DI

a) Tests for DI heterogeneity

b) Corrections for bias

c) Problems with hypothetical repeatability

V.A5) ANOVA procedure testing the significance of FA relative to ME

V.A6) What to do when replicate measures of all individuals is not practical

V.B) DEPARTURES FROM IDEAL FA

V.B1) Directional asymmetry- the problem & tests

V.B2) Departures from normality- the problem

a) Skew

b) Leptokurtosis

c) Platykurtosis

V.B3) Tests for kurtosis

VI) ANALYSES OF FA VARIATION: TESTING FOR DIFFERENCES

VI.A) LEVENE'S TEST FOR HETEROGENEITY OF VARIANCE

VI.B) DIFFERENCES AMONG INDIVIDUALS (MULTIPLE TRAITS)

VI.C) DIFFERENCES BETWEEN TWO SAMPLES (MULTIPLE TRAITS)

I) INTRODUCTION

Fluctuating asymmetries are small, random deviations from symmetry of bilaterally symmetrical traits (Ludwig, 1932). They presumably reflect the residual variation after all the direct effects of genotype and environment on trait form have been removed (Mather, 1953). As a consequence, the average unsigned deviation from symmetry, to which the term fluctuating asymmetry (FA) typically refers, has achieved prominence as a measure of developmental precision (Palmer, 1996): the ability of a given genotype to produce the same *target phenotype* (Nijhout and Davidowitz, this volume) repeatedly — on opposite sides of the body — under well-defined environmental conditions (Zakharov, 1992). Both environmental and genetic stress appear to increase FA (Leung and Forbes, 1996; Palmer, 1996; Vøllestad *et al*., 1999; but see critique by Bjorksten *et al*., 2000). In addition, the subtle deviations from symmetry that yield FA may also relate to individual quality or fitness (Møller, 1997; Brown and Brown, 1998; and Houle, 1998; but see comments by Clarke, 1998). For these reasons, FA has been widely studied in many ecological and evolutionary contexts (Møller and Swaddle, 1997), although hints of discomfort about the validity of published claims have arisen on several fronts (Houle, 1998; Palmer, 1999; Simmons *et al*., 1999; Palmer, 2000; Palmer and Hammond, 2000).

Over the last decade, the literature on FA has exploded (Palmer, 2000). In addition to a flood of data papers (Møller and Swaddle, 1997), many new analytical methods have been advanced, along with several critiques of methodological issues. Will refined analytical methods improve the quality of FA data and analysis? In some situations they may. However, more sophisticated analyses will never compensate for poor data or sloppy thinking. Below we try to bring some common sense to bear on problems typically encountered in FA analyses.

I.A) DISQUIETING REVELATIONS

Two recent reports suggest some areas of the FA literature may have been compromised by a large number of 'false positive' results. First, as more studies tested whether FA exceeded measurement error (ME), fewer and fewer detected significant associations between individual FA and sexual selection (Simmons *et al*., 1999). This suggests a) many earlier studies may have reported false positive results, and b) minimizing ME remains a significant challenge for FA studies, not only because it weakens results but also because it actually introduces bias in several insidious ways (*Section V.A*). Second, a meta-analysis of published associations between FA and sexual selection revealed direct evidence of selective reporting (Palmer, 1999): the tendency to publish preferentially results that are either significant statistically or consistent with expectation (Palmer, 2000). In addition, some of the more remarkable published reports of correlations with individual asymmetry in humans — such as with IQ, attractiveness, sexual satisfaction, physical prowess, individual fecundity, and the timing of ovulation — have invited pointed criticism (Palmer and Hammond, 2000, see also http://www.biology.ualberta.ca/palmer.hp/asym/FA/FA-Refs.htm).

While the reputation of FA has been tarnished by these critiques, it is too early to dismiss it as a useful tool for inferring developmental instability. Other meta-analyses have revealed little or no evidence of selective reporting among studies of other relations, either between FA and stress (Leung and Forbes, 1997 as re-analyzed in Palmer, 2000) or between FA and heterozygosity (Vøllestad *et al*., 1999), even though they did reveal low mean effect sizes and high variability. Formal replications of prior FA studies would go a long way toward returning respectability to the field (Palmer, 2000).

I.B) SOME ESSENTIAL TERMINOLOGY

The terminology associated with subtle asymmetries is a challenge for both newcomers and veterans. Appropriate use of terms for patterns, which are observable, and terms for presumed processes, which are inferred, is critical to avoid perpetuating sloppy thinking. Our use of the terms we use frequently is outlined below. A more complete set of definitions is available in Palmer (1994), Nijhout and Davidowitz (this volume), and the glossary to this book.

1) Terms for (observable) patterns

fluctuating asymmetry (FA)- a pattern of variation of the difference between the right and left sides (R - L) where the variation is normally distributed about a mean of zero.

antisymmetry- a pattern of variation of (R - L) where the variation is distributed about a mean of zero, but the frequency distribution departs from normality in the direction of platykurtosis or bimodality.

directional asymmetry (DA)- a pattern of variation of (R - L) where the variation is normally distributed about a mean that is significantly different from zero.

developmental precision- a general, neutral term for describing how closely a structure approaches its ideal or *target phenotype* (Nijhout and Davidowitz, this volume) for a particular genotype and growth environment. It implies nothing about causation and is not restricted to bilateral traits. Size-independent (dimensionless) measures of FA (*Section IV.A*) offer one measure of developmental precision. The coefficient of variation among serially homologous parts of an individual (e.g., legs of an individual millipede), or among genetically identical individuals reared under identical conditions, would be another. It is also dimensionless.

2) Terms for (inferred) processes or causes

developmental noise- random variation in a suite of developmental factors that are the ultimate cause of subtle deviations from symmetry, including metabolic rates, concentrations of regulatory molecules, diffusion, thermal noise, and rates of cell division, cell growth and cell death (see also Nijhout and Davidowitz, this volume). Increased developmental noise yields

lower developmental precision and increased FA.

developmental stability- the capacity of an individual to correct for random perturbations caused by developmental noise. Increased developmental stability yields higher developmental precision and decreased FA.

developmental instability (DI)- the combined contributions of developmental noise and developmental stability that define the expected or hypothetical variance of R-L. This term is equivalent to the *asymmetry potential* of M.E. Soule (pers. comm.). Increased DI yields lower developmental precision and increased FA, but DI may increase due either to increases in developmental noise or to decreases in developmental stability. This usage differs from Palmer (1994) but avoids unintended implications that observed differences in FA are due to differences in developmental noise versus developmental stability when nothing is known about the actual causes.

I.C) ANXIETY ABOUT METHODS

A proliferation of methods sometimes suggests a discipline in turmoil. Where the biological signal is weak, but the questions alluring, hope springs eternal that increasingly sophisticated analytical tricks will somehow extract more reliable results from recalcitrant data. Unfortunately, a surfeit of methods may also discourage new studies or leave those unfamiliar with the detailed pros and cons confused about how best to proceed.

We would be the last to diminish the importance of methodology to studies of FA. Nonetheless, we believe firmly that the greatest increase in quality of results will come not from increasingly sophisticated analyses, but rather from a greater awareness that the little things count. Careful attention to choice of traits (*Section III*), measurement protocol and analysis of ME (*Section V.A*), detection of outliers (Appendix V), and tests for departures from normality (*Section V.B*), coupled with the use of multiple independent simple tests to confirm that results are not analysis-dependent, will yield more convincing results than recourse to sophisticated methods after the data have been collected, with the hope that oversights in design and protocol can somehow be ameliorated.

Below, we summarize recent methodological and conceptual refinements, in the hopes of providing a sounder foundation to FA analyses and interpretation. This chapter supplements, rather than replaces, an earlier FA analysis primer (Palmer, 1994). We hope the worked example provided in Appendix V will reinforce appreciation of how simple graphical inspections of the data, and a few elementary tests, are all that are required for a well-conducted study.

II) FIVE CORE CONCEPTS

II.A) RANDOM VARIATION ≠ DEVELOPMENTAL INSTABILITY

FA refers to random (normally distributed) variation of the difference between sides (R - L) about a mean of zero. Most biologists interpret this random variation as evidence of developmental instability (DI), because the effects of genotype and environment should be the same for both sides and therefore cancel out of the difference (Mather, 1953; Van Valen, 1962; Palmer, 1996).

Unfortunately, even for traits that exhibit ideal FA (see Fig. 1a below), subtle departures from symmetry may not be due solely to DI in the conventional meaning of this phrase (Palmer, 1994). We believe this to be one of *the* most troubling aspects of FA studies. If departures from symmetry due to DI can not be distinguished from those due to random — but repeatable! — environmental effects on form, FA can not serve as an index of DI.

The problem is simple: observable random variation in a trait may have more than one cause. Developmental noise undoubtedly contributes to deviations from symmetry, and in some cases it may be the primary cause. However, random deviations from symmetry may also arise due to random effects of the environment on phenotypically plastic traits (*Section III.B*) or to random effects of wear and tear (*Section III.B*).

For biologically sensible estimates of DI, traits must be selected judiciously to avoid those confounding factors (*Section III.D*).

II.B) DEPARTURES FROM SYMMETRY ESTIMATE A VARIANCE . . .

Descriptors of FA estimate a variance, not a mean. The greater the underlying DI, the greater the observed *variance* of R - L. Therefore, tests for differences in FA among individuals, traits or samples are fundamentally tests for heterogeneity of variance. They are not tests for differences in means, in the sense that most biologists understand this. *Average asymmetry,* meaning average |R - L|, is just one convenient way to describe the variance of R - L (see *Section IV.A2* below).

This simple fact makes many of the problems associated with FA analyses painfully obvious: a) ME increases the variance but not the mean of a sample (*Section V.A1*), b) variances are harder to estimate with confidence than means (Smith *et al*., 1982), c) differences in distribution shape can have large effects on the variance (*Section V.A1*), d) a single outlier datum will have a larger effect on the variance than on the mean (*Section V.A1*), and e) many tests for heterogeneity of variance are quite sensitive to distribution shape (*Section VI.A*).

II.C) . . . WITH ONE DEGREE OF FREEDOM

Asymmetry in a single trait of a bilaterally symmetrical individual yields limited information about underlying DI because the difference between sides estimates the variance due to DI with

only one degree of freedom (Palmer, 1994, p. 360; Van Dongen, 1998; Whitlock, 1998), and estimates of a variance with one degree of freedom have limited statistical power (Smith *et al.*, 1982). For example, even among a hypothetical group of individuals of identical DI raised under identical conditions, R - L will still vary simply due to sampling error: by chance, some individuals will be nearly symmetrical and some may be very asymmetrical. This likely accounts for the widespread observation that subtle asymmetries in one trait rarely correlate with subtle asymmetries of other traits on the same individuals (Van Valen, 1962; Soulé and Couzin-Roudy, 1982; Palmer and Strobeck, 1986; Dufour and Weatherhead, 1996; Møller and Swaddle, 1997; Houle, 1998).§

Nonetheless, if DI affects all traits in an individual similarly, then incorporating deviations from symmetry from multiple traits should yield greater power to detect differences among individuals (*Section VI.B*). Each trait provides an independent estimate of the underlying DI of that individual and therefore adds another degree of freedom to the estimate, so long as the potentially confounding effects of variation in trait size are removed (*Section IV.A3*).

II.D) TRAIT COVARIATION AND TRAIT SIZE AFFECT FA

FA is influenced not only by the underlying DI, which affects both the right and left sides, but also by negative covariation between the sides and by interactions with trait size.

If we knew the exact size a trait should be for a particular genotype and growth environment — the *target phenotype* of Nijhout and Davidowitz (this volume) — the coefficient of variation of that trait for a group of genetically identical individuals raised under identical environmental conditions would describe that trait's DI, because all genetic and environmental effects on trait size were eliminated. This would be true for single, medial traits (e.g., bill length in birds), as well as for one side of a paired, bilateral trait. Under these conditions, all that would be gained by taking the difference between the sides would be an estimate of DI based on two traits instead of one.

In the real world, of course, both genotype and environment affect trait size, so the variance in a medial trait like bill length in a sample of individuals arises from a complex mixture of the effects of genotype, environment, and DI. Fortunately, for bilateral traits, genotype and environment typically affect both sides similarly, so the right and left sides exhibit positive covariation. This is why the variance of the difference between two sides is such a convenient number:

$$\mathrm{var}(R_i - L_i) = \mathrm{var}(R_i) + \mathrm{var}(L_i) - 2\,\mathrm{covar}(R_i L_i) \quad (1)$$

where $\mathrm{covar}(R_i L_i)$ is the covariance between R_i and L_i

$$\mathrm{covar}(R_i L_i) = \sum[(R_i - \bar{R})(L_i - \bar{L})] / (N - 1) \quad (2)$$

and where $\bar{R}$ and $\bar{L}$ are the population means of the right (R) and left (L) sides, and N is the number of individuals. In theory, the term $2\,\mathrm{covar}(R_i L_i)$ removes all of the positive covariation

§ It also accounts for why the repeatability of FA can be as low as 20% even when the coefficient of variation of DI among individuals is 100% (Houle 2000 J Ev Biol 13:720).

between R_i and L_i due to genotype and environment, leaving only the uncorrelated random variation of R_i and L_i due to DI.

This statistical trick yields a biologically meaningful descriptor of DI only so long as *all* of the interdependencies between R and L due to genotype and environment are both positive and captured by the term 2 covar(RL). Unfortunately for studies of FA, if the covariation between sides is negative, or if the variance of R or L depends on trait size, then var(R - L) becomes a complex mixture of the effects of genotype, environment and DI, and cannot be interpreted as a simple measure of DI.

Equation 1 yields some useful insights into why certain idiosyncrasies of FA analyses are so important: a) antisymmetry is effectively negative covariation between the sides (Van Valen, 1962), therefore subtle antisymmetry will inflate var(R - L), b) if a morphogen that influences trait growth is limiting, such that an excess on one side yields a deficit on the other (Klingenberg and Nijhout, 1998), this may also yield a negative covariation (normal covariant asymmetry, Palmer *et al.*, 1993) that would inflate var(R - L), and c) if the range of body sizes in a sample is large, and var(R_i) and var(L_i) increase with the trait means, then var(R_i - L_i) no longer estimates a single DI variance but rather a whole family of DI variances that depend on the trait's size distribution (*Section IV.A3*).

Tests for antisymmetry (*Section V.B3*) and size dependence (*Section IV.A4*) are therefore essential elements of a FA analysis.

II.E) LOG TRANSFORMATION YIELDS SIMPLE, SCALE-FREE ANALYSES

Where ME is relatively small (*Section V.A1*), log-transformation of raw measurements offers a versatile and attractive solution to many elements of FA analyses.

1) A conventional result via an unconventional route $|R - L| / [(R + L) / 2] \quad |\ln(R/L)| = |\ln(R) - \ln(L)|$ (*Section IV.A7*). In other words, the difference between the natural logs is effectively equivalent to the difference between the sides divided by the mean. Both describe FA as a proportion of the trait mean, and therefore yield dimensionless (scale free) indexes that allow the FA variation of very different sized traits to be compared directly (*Section IV.A7*).

2) Avoiding undesirable size-dependent heterogeneity Where FA increases with trait size, and where considerable size variation exists within a sample, the frequency distribution of (R - L) will be leptokurtic because it represents a mixture of individuals with different variances (Wright, 1968; *Section V.B2b*). This potentially confounds tests for FA relative to ME (*Section V.A*) and for departures from normality (*Section V.B2*). The frequency distribution of ln(R) - ln(L), however, is not influenced by simple size-dependence, so any remaining leptokurtosis must be due to other factors, such as outliers (*Sections IV.A1e*), the insidious effects of ME (*Sections IV.A6*), true heterogeneity in underlying DI (*Section V.B2b*), etc.

3) Improved power for measures of DI in an individual Deviation from symmetry in a single trait in an individual estimates the underlying DI variance with not much confidence (*Section II.C*). Log-transformed measurements allow deviations from symmetry to be averaged for multiple traits in an individual, thereby increasing the ability to detect DI differences among individuals (*Sections IV.B3, VI.B*). Statistical evidence for DI heterogeneity within a sample is an essential prerequisite to tests for correlations between FA and individual quality, fitness, attractiveness, etc. With no evidence for DI heterogeneity, such tests are pointless.

4) Testing for differences among groups using multiple traits per individual When comparing samples of individuals, each trait provides an independent estimate of DI. Unfortunately, because FA is often proportional to trait size and traits typically differ in size, a simple pooling of traits may yield misleading results. However, ln(R) - ln(L) is not influenced by simple size-dependence. When combined with a multi-way Levene's test (*Section VI.A*), multiple traits may be combined in a single analysis to test for differences among groups of interest as well as interactions between groups (*Section VI*).

III CHOICE OF TRAITS

Several fundamental concerns should govern the choice of traits for a FA analysis.

III.A) BEWARE DEPARTURES FROM IDEAL FA

As noted ritually in discussions of FA variation, subtle departures from bilateral symmetry generally take three forms, each defined by a unique combination of mean and variance of right-left (R - L) differences in a sample: *fluctuating asymmetry* (mean = 0, normal), *directional asymmetry* (DA; mean ≠ 0, normal), and *antisymmetry* (mean = 0, platykurtic or bimodal). Differences between the sides of individuals in traits that exhibit either DA or antisymmetry likely arise from a complex mixture of genetic and non-genetic causes (see Palmer and Strobeck, 1992, for a detailed graphical explanation). Therefore, traits exhibiting DA or antisymmetry may not yield reliable measures of DI (Palmer and Strobeck, 1992, 1996; but see Graham *et al*., 1993, 1998, for a minority opinion).

III.B) BEWARE PHENOTYPIC PLASTICITY

Distinguishing departures from symmetry due to DI from those due to predictable environmental effects is most troublesome for traits known to exhibit phenotypic plasticity. Three examples illustrate the problem.

First, many plants exhibit pronounced phenotypic plasticity (Bradshaw, 1965). Significantly, measures of variability often differ among regions of single plants, as observed in

tobacco (Paxman, 1956; Sakai and Shimamoto, 1965), *Clarkia* (Sherry and Lord, 1996), iris (Tarasjev, 1995), and trees (Bagchi *et al*., 1989). In addition, intraplant variability may vary over time in individual plants (Roy, 1958), directions of deviation from symmetry in leaves may be related to phyllotaxis (Dormer and Hucker, 1957), and directional departures from symmetry in plants may be induced experimentally (Solangaarachchi and Harper, 1989; Desbiez *et al*., 1991). Similarly, in foliose lichens local micro-climates may induce departures from radial symmetry in individual thalli (Armstrong and Smith, 1992). These observations suggest deviations from symmetry, or from some organ-specific invariant (Freeman *et al*., 1993), arise due a complex mixture of direct — presumably repeatable — effects of the environment, along with the random effects of DI. Therefore, departures from symmetry or other invariants (Freeman *et al*., 1993) in plants seem like unreliable measures of DI.

Second, vertebrate bones grow by accretionary growth and are capable of significant remodelling (reviewed in Olsen *et al*., 2000). In human limb bones, differential use may increase asymmetry (Malina, 1983; Trinkaus, 1994). In addition, the right and left legs of humans exhibit compensatory growth during ontogeny so that deviations from symmetry in an individual vary over time (Hermanussen *et al*., 1989). As a consequence, deviations from symmetry in vertebrate bones may be difficult to use as an index of DI.

Third, many animals exhibit use-induced differences in structures used for food handling or processing (Travis, 1994). Where paired structures exist for manipulating prey, differential use of one side may induce morphological asymmetries. For example, in a shell-crushing crab, a harder diet induces relatively larger claws (Smith and Palmer, 1994). In the same species, serial observations revealed that individual crabs forced to crush prey developed a more pronounced preference to crush with one claw (either R or L) than those fed soft food (Palmer and Harrison, unpublished). Such a learned handedness may induce morphological differences between the sides. Therefore, in paired structures where one side may be used more than the other, departures from symmetry may arise due to both differential use and DI.

For these reasons, traits known to be very plastic should be avoided: environmentally-induced asymmetries (Nijhout and Davidowitz, this volume) tell us nothing about DI.

III.C) BEWARE TRAITS VULNERABLE TO WEAR

Traits vulnerable to breakage or wear complicate interpretations of FA variation. First, even if both sides wear at the same rate on average, the amount of wear will rarely be identical. Differences in wear between sides will likely be normally distributed about a mean of zero, and therefore indistinguishable from variation due to DI (*Section V.A1a*). Second, deviations from symmetry due to breakage or wear can create an artificial dependence of |R - L| on trait size, since loss of material from one side increases the asymmetry but decreases average trait size (Sullivan *et al*., 1993). Most importantly, departures from symmetry due to differential wear tell us nothing

about DI.

III.D) PREFERRED TRAITS

Reliable traits for studies of FA should share several qualities.

III.D1) High repeatability They should be easily and repeatably measured (metrical) or scored (meristic; but see Palmer, 1994, for an extensive discussion of idiosyncrasies of meristic traits). Considerable analytical angst may be avoided if several traits are examined for FA relative to ME at the start of a study, and only those where mean $|R - L|$ exceeds mean $|M_1 - M_2|$ by at least twofold are used. Similarly, time invested early in a study to reduce ME by refining the measurement protocol will be more than repaid by increased statistical power and confidence in the final results.

III.D2) Low plasticity They should not exhibit significant plasticity or remodelling (see *Section III.B*). They should not be traits where one side may be used more often than the other, or where one side might experience different micro-environmental conditions.

III.D3) Low vulnerability to wear They should not be vulnerable to wear or injury (see *Section III.C*).

III.D4) Geometric independence Where multiple traits are examined per individual, they should be both geometrically and developmentally independent. Linear measurements that share a common endpoint, for example, are not geometrically independent: variation in the position of the shared endpoint will affect both dimensions. Similarly, different dimensions of the same structure (e.g., length and width of a single leg segment, or wing vein-lengths on the same wing) may not yield independent estimates of DI because perturbations early in development affect the entire structure or because they are more highly integrated developmentally (Leamy, 1993; Klingenberg and Zaklan, 2000; Klingenberg 'integration' this volume).

III.D5) No or predictable size-dependence The difference between sides should either be completely independent of the mean, as in some meristic traits (Berry, 1968; Angus and Schultz, 1983; but see Palmer, 1994, for exceptions), or it should increase in proportion to trait size due to simple allometry so that transformations removing trait size effects are valid (see *Section IV.A7*).

(- - - - Table 1 approximately here - - - -)

IV) DESCRIPTORS OF FA

IV.A) UNIVARIATE MEASURES OF FA

IV.A1) Standardized conventional FA indexes

Our earlier summary of FA indexes (Palmer and Strobeck, 1986; Palmer, 1994) allowed relations among indexes to be seen more clearly. It suffered from one unfortunate disadvantage: numerical values for the different indexes were not directly comparable because some were mean differences (FA1 - FA3), some were variances of untransformed differences (FA4 - FA7, FA10), and some were log-transformed differences (FA8). The indexes in each row of Table 1 here yield descriptors of FA that are directly comparable numerically (see *Sections IV.A2* and *IV.A3* for justification). Unfortunately, FA3 is equivalent to FA2 (and FA6a equivalent to FA7a) only when FA is proportional to trait size; if FA is independent of trait size FA3 underestimates FA2 by an amount related to the size variation (see Appendix IV). For trait size CV < 20%, FA3 deviates from FA2 by less than 5%, but for a trait size CV of 40%, FA3 deviates from FA2 by nearly 20%.

The pros and cons of these indexes are discussed at length in Palmer (1994). FA1 and FA2 are the most popular indexes by far, which is fortunate, because they are less affected by departures from normality (skew or leptokurtosis) than are FA4a to FA6a. The indexes in rows 2 and 3, along with index FA10b, are dimensionless and express FA as a proportion of trait size. This allows FA to be compared directly among traits of very different overall size.

Although rather more cumbersome to compute, FA10a and FA10b have one major advantage over all other FA indexes: they describe the average difference between sides after ME has been factored out. Because of the biases ME introduces (*Section V.A1*), one or both of them are worth computing to confirm that differences in FA among groups persist after ME has been partitioned out, even if other indexes are used for more sophisticated tests of differences among individuals or samples (*Section VI*).

Unfortunately, antisymmetry will artificially inflate all of these indexes and DA will inflate those based on unsigned deviations (FA1-3, FA8a) (Palmer, 1994). So tests for platykurtosis (*Section V.B3*) and DA (*Section V.B1*) must precede any tests for differences in FA among individuals or groups.

(- - - - Figure 1 approximately here - - - -)

IV.A2) Why an average deviation estimates a variance and an asymmetrical distribution of |R - L| is not to be feared

For traits that exhibit ideal FA, R - L differences exhibit a normal distribution about a mean of zero, and the standard deviation SD_{R-L} describes the spread of R - L differences about that mean (Fig. 1a). indexes FA4a, FA5a, FA6a and FA7a all estimate $SD_{(R-L)}$ directly.

Taking the absolute value of (R - L) differences amounts to flipping the left side of the distribution over onto the right (Fig. 1b), so now twice as many observations exist to the right of zero but none exist to the left. This has two very important and useful consequences. First, the truncated normal distribution (Fig. 1b) is strongly asymmetrical (skewed to the right), so its mean and variance are inextricably linked. In fact, for a truly normal distribution, the $CV_{|R - L|} = SD_{|R - L|} / mean_{|R - L|}$ is a constant: $\sqrt{((\pi - 2) / 2)} = 0.756$ (Houle, 1997). If DI varies among individuals, though, the resulting distribution of R - L is no longer normal (*Section V.B2b*), and this constant no longer applies. Second, the expected value (i.e., mean) of this distribution (Fig. 1b) differs from the expected standard deviation of the signed asymmetry distribution (Fig. 1a) by a simple constant, $0.798 = \sqrt{(2/\pi)}$ (Kendall and Stuart, 1951). Therefore, the mean|R - L| provides an unbiased estimate of SD(R-L), although it is somewhat less efficient statistically (87.6%, Kendall and Stuart, 1951; Palmer and Strobeck, 1992).

Because the expected value of $mean|R - L| = \sqrt{(2/\pi)}\ SD_{R-L} = 0.798\ SD_{R-L}$ (Fig. 1), indexes based on variances (FA4, FA5, FA6, FA7, and FA10 of Palmer, 1994) may be easily modified to make them directly comparable to indexes based on average difference (FA1, FA2, FA3). The modified indexes of Table 1 show the appropriate modification.

Many biologists new to FA analyses are troubled by the highly skewed distribution of |R - L|. Some even try transformations to correct for this skew because they have been so rigidly trained to correct for departures from normality before conducting any statistical tests. But this fear is unwarranted. The skew of the |R - L| distribution is precisely why this index has its useful properties! Over forty years ago, Levene (1960) recognized that the difference between the means of two truncated normal distributions (e.g., Fig. 1b) provided a robust and unbiased estimate of the difference between the variances of the untransformed normal distributions (Fig. 1a). This is the basis of Levene's test for heterogeneity of variance, which is perhaps the most straightforward and versatile test available for FA variation (*Sections II.B* and *VI.A*).

IV.A3) Trait size variation: the problem

Generalizations about patterns of FA variation have been seriously hampered by the impact of trait size variation (Palmer and Strobeck, 1986). During normal growth, the variability of a trait tends to increase with trait size (Lande, 1977; Van Valen, 1978): the long bones of an elephant's hind legs are more variable in absolute terms than the homologous bones of a mouse. The real question, of course, is whether one is *proportionally* more variable than the other. Several corrections for the size-dependence of variability have been proposed (Palmer and Strobeck, 1986; Leung, 1998), including some rather peculiar ones (Evans and Hatchwell, 1993). One index — trait difference divided by trait mean [|R-L|/((R+L)/2)] — is widely used in many studies of FA variation, but it has been criticized because of the apparent lack of independence of the numerator and denominator (see Evans and Hatchwell, 1993, and the exchange between Sullivan *et al*., 1993, and Cuthill *et al*., 1993). Furthermore, this widely used index does not lend itself easily to tests for

the significance of FA variation relative to ME because the average of the replicate measurements must be computed first. Clearly, a method that avoided these shortcomings would be preferred, particularly if it were easier to use.

Dependencies of subtle asymmetries on trait size complicate the analysis and interpretation of FA differences in several commonly encountered situations:

a) Inferred differences in DI among populations. FA variation offers a valuable tool for estimating the effects of genetic or environmental stress on different populations, and therefore has many promising applications in biomonitoring and conservation (Parsons, 1992; Clarke, 1995; but see Heard *et al*., 1999, for a critical commentary). However, for many organisms overall body size, or the relative size of particular traits, also differ among populations due to genetic or environmental effects on growth and form (Futuyma, 1986). If FA varies with trait size, and average trait size differs among populations, then inferred differences in DI among populations may be either enhanced or obscured by size-dependent variability (Palmer and Strobeck, 1986).

b) Inferred differences in DI among taxa or traits. Comparative (e.g., Gummer and Brigham, 1995; Brakefield and Breuker, 1996; Crespi and Vanderkist, 1997; Bromberg and Jaros, 1998) or historical studies (e.g., see Smith, 1998) of FA variation can yield significant insights into the evolution of DI. Many other questions remain to be addressed: Are some categories of taxa (e.g., homeotherms vs poikilotherms, arthropods vs vertebrates) or some categories of traits (e.g., locomotory vs feeding vs reproductive, endoskeletons vs exoskeletons) more developmentally predictable than others? Unfortunately, estimates of DI based on FA are greatly complicated by differences in overall trait size or dimensionality.

c) Estimating organism-wide DI based on multiple traits. Deviations from symmetry in a single trait provide at best a weak estimate of the underlying DI variance (*Section II.C*). But averaging asymmetries of multiple traits (*Section VI.B*) should increase the ability to detect differences in DI among individuals because each trait provides an independent estimate of the underlying DI (Palmer, 1994). However, a composite measure of organism-wide DI based on asymmetries of multiple traits must take into account the potentially confounding effects of differences in trait size.

d) Within-sample heterogeneity in FA and leptokurtosis. If FA varies with trait size, and trait size varies within a sample, then the average FA for the sample will reflect a mixture of underlying DI and size-dependent asymmetry variation (e.g., see Rowe *et al*., 1997). This has the same effect as combining groups of individuals with different variances: both yield leptokurtosis in the pooled sample (Wright, 1968; Palmer and Strobeck, 1992). Therefore, interpreting within-

population leptokurtosis as direct evidence for within-sample variation in DI (e.g., Gangestad and Thornhill, 1999) seems risky at best, simply because leptokurtosis may arise in so many different ways (*Section V.B2b*).

e) Correlations between trait or body size and individual quality. In addition to the statistical complications noted above, body size or trait size differences may reflect real differences in individual quality. Arbitrary statistical 'removal' of trait size effects may therefore potentially obscure biologically significant differences in FA among individuals or groups (Palmer and Strobeck, 1986; Leung, 1998). Clearly, elimination of all size-dependent FA variation is not a desirable outcome, since some may reflect true size-dependence of DI, so correction for 'size' effects should be based on biologically sound, *a priori* models of growth.

IV.A4) Tests for size-dependence

Tests for size-dependence of FA may be done several ways. Recall that |R - L| estimates the SD of (R - L) with one degree of freedom (*Section II.C*). Therefore a test of the association between trait asymmetry |R - L| and trait size [(R+L)/2] is effectively a Levene's test for heterogeneity of variance that tests for association between two continuous variables for each individual — one estimating the variance, the other estimating the mean — rather than the conventional test between two or more groups where only variance estimates are used.

Tests for size-dependence are best conducted *before* testing for ideal FA (*Section V.B*) because size-dependent heterogeneity in asymmetry variation can yield leptokurtosis in the frequency distributions of R - L or obscure subtle antisymmetry (*Section IV.A3*).

A parametric, least-squares linear regression of trait asymmetry |R - L| vs trait size [(R+L)/2] is one potential test, but this test assumes homogeneity of variance (the variance in Y should be independent of the value for X). Clearly if the average |R - L| differs between traits of different size, so will the variance because the mean and variance of absolute deviations are inextricably intertwined (*Section IV.A2*). Fortunately, heterogeneity of variances generally decreases the power of a regression analysis, so the result of this test is conservative. However, parametric tests of association may be strongly influenced by one or two extreme observations, and so are more likely to yield a spurious positive result if data are unusually distributed.

Non-parametric tests of association (Spearman and Kendall coefficients of rank correlation) are preferred for this analysis because they do not assume homogeneity of variance and are not influenced by a few extreme observations. These two tests differ in how they weight pairs of ranks. Spearman's ρ is preferred where the reliability of closely ranked values is uncertain (Sokal and Rohlf 1995, p. 600), and is therefore somewhat more appropriate to FA data because of the uncertainty of FA estimates. Both yield similar values, though, for real FA data (see Step 8, Appendix V).

IV.A5) Correcting for trait size variation: the standard solutions

The growth of most body parts arises from a simple multiplicative process: replication of cells. As a consequence, trait variability (as measured by its standard deviation) increases in proportion to the mean (i.e., the coefficient of variation, CV, is independent of the mean). This is why log transformations so nicely linearize relations between dimensions of almost any two traits (Huxley, 1924; Huxley, 1932) and standardize the variances (Lewontin, 1966).

Several FA indexes correct for trait size effects by expressing deviations from symmetry as a proportion of trait size (Table 1). At the level of individuals, dividing the difference |R - L| or (R-L) by the mean= (R + L)/2 is the most common transformation (FA2 and FA6a of Table 1). This transformation yields a convenient dimensionless index of FA, and therefore allows differences in proportional FA to be compared directly among traits of very different sizes. A similar transformation may also be applied at the level of the entire sample (FA3 and FA7a of Table 1).

As justified elsewhere (Palmer and Strobeck, 1986; Palmer, 1994; Leung, 1998), however, this transformation should not be applied blindly. Depending on the pattern of size-dependence, other transformations may be more appropriate (Leung, 1998). In addition, where FA is fixed but trait size varies considerably, a correction for size-dependence can generate spurious differences in FA (*Section IV.A6*). Fortunately, most morphological traits do exhibit simple multiplicative growth, so these standard transformations seem appropriate in most cases.

(- - - - Figure 2 approximately here - - - -)

IV.A6) Correcting for trait size variation: a fundamental concern

Measurement error (ME) seriously complicates tests for differences in FA among traits of different size. For normal studies of morphological variation, variation due to ME is a small percentage (1 to 5%) of the true biological variation, therefore the increase in biological variation relative to the mean is not seriously affected by ME. Unfortunately, in FA studies, ME (e.g., as ME1 of Table 3) may be a sizeable fraction (25 to 100%) of the true average difference between sides, FA1= mean |R - L|. So while the true biological variation may increase in proportion to the mean (Fig. 2a), ME tends to be constant and independent of the mean, both for the same trait of individuals of different sizes and among traits where the same protocol was used on each trait. As a consequence, larger individuals or larger traits will appear to exhibit proportionally lower FA than smaller ones simply because ME is a smaller proportion of the between-sides variation (compare Fig. 2b to Fig. 2c)!

Care must therefore be taken when testing for the size-dependence of FA, where larger size is thought to reflect higher quality (reviewed in Møller and Swaddle, 1997). If unscaled FA (e.g., FA1) declines with increasing trait size, this can only occur if DI is smaller in larger individuals, since ME is typically constant. Therefore larger individuals will exhibit lower FA.

However, if a size-scaled index (e.g., FA2) declines with increasing trait size, an additional

test is required to determine whether this decline is greater than expected given a constant ME. Two approaches seem reasonable. First, ask: is $SLOPE_{FA}$ — the slope of proportional FA (e.g, $FA2_i$) vs trait size $(R_i + L_i)/2$ — significantly steeper than $SLOPE_{ME}$ — the slope of proportional ME (e.g., mean(ME1) / $[(R_i + L_i)/2]$ vs trait size $(R_i + L_i)/2$? Because $SLOPE_{ME}$ has negligible error (relative ME declines roughly linearly with increasing trait size if the size range is less than twofold, Fig. 2c), the statistical test is a simple one-sample t-test: $t_s = (SLOPE_{FA} - SLOPE_{ME}) / SE_{FA}$, where SE_{FA} is the observed standard error of $SLOPE_{FA}$. t_s is then compared to critical values of the Student's t distribution for N-1 degrees of freedom. If the size range is relatively small (less than two-fold), and two measurements have been taken per side, the expected $SLOPE_{ME} = -ME1 / S^2$, where ME1 is as in Table 3 and trait size S = mean[(R + L)/2] (see Appendix II for derivation).

Alternatively, divide the size range into three or more size categories. For each size category, compute FA10a (Table 1), which factors ME out. Then, ask: does the ratio FA10a / $SIZE_c$ decline significantly with increasing $SIZE_c$, where $SIZE_c$ = mean $[(R_i + L_i)/2]$ for each size category. Any decline in the ratio FA10a / $SIZE_c$ must be due to a true decline in proportional FA.

Such tests should be conducted whenever ME (as ME1, see Table 3) exceeds 10% of the best estimate of the true average difference between sides (e.g., FA10a, Table 1).

(- - - - Figure 3 approximately here - - - -)

IV.A7) Correcting for trait size variation: a new and versatile solution based on ln(R/L)

If ME (as ME1, Table 3) is small (e.g., <10% of FA1, Table 1), so that concerns about the bias ME introduces to size-adjusted estimates of FA are minimal (*Section IV.A6*), then an alternative approach to quantifying size-adjusted of FA based on FA8 (Palmer, 1994) offers several advantages.

FA8 scales out size variation by taking the ratio R/L. This ratio may have been the very first index of FA variation ever used (Sumner and Huestis, 1921), well before the phenomenon of FA was given a name (Ludwig, 1932). Where DA and antisymmetry are absent, variation in this ratio reflects the proportional variation about the expected mean of 1.0 (Sumner and Huestis, 1921). Fear of ratios (e.g., Atchley *et al*., 1976), however, seems to have discouraged use of this index.

Indeed, variation in the ratio (R/L) does have one unfortunate property (Fig. 3). Even if R_i and L_i are normally distributed, the frequency distribution of R_i / L_i is skewed (Fig. 3a). This skew becomes more pronounced the greater the difference |R - L| as a proportion of the mean, (R+L)/2 (Fig. 2c). Fortunately, the frequency distribution of $\log(R_i / L_i)$ is no longer skewed, no matter how large FA is relative to trait size (Figs. 3b, c). Therefore FA8 of Palmer (1994), and its more useful descendent FA8a (Table 1), are perfectly reasonable descriptors of FA.

Because both FA2 and FA8a (Table 1) estimate the size-scaled, between-sides variance, one might ask how these two indexes are related. Surprisingly, for all practical purposes, they are numerically equivalent, if $\log_e$ = ln (natural or Napierian logarithms) is used instead of $\log_{10}$ (Briggsian logarithms), because

$$|\ln(R/L)| \quad |R - L| / [(R + L)/2] \tag{3}$$

More precisely, letting $d_1 = (R - L) / [(R + L)/2]$, d_1 is simply the first term of an expansion series (see Appendix I for proof):

$$|\ln(R/L)| = |d_1 + d_1^3 / 12 + d_1^5 / 80 + \ldots|$$

Significantly the second and all subsequent terms in this series can be ignored in studies of FA because d_1 is almost always less than 0.1 and typically closer to 0.01 (Palmer, 1996). So even if deviations from symmetry approach 10% of trait size (i.e., $d_1 = 0.1$), the second term in this series would be less than 0.001, and all higher order terms would be even smaller. Therefore, FA2 and FA8a (Table 1) are equivalent to at least three decimal places.

Furthermore, a trick from first-year calculus reveals a most useful relationship:

$$\ln(R/L) = \ln(R) - \ln(L), \text{ so } |\ln(R/L)| = |\ln(R) - \ln(L)| \tag{4}$$

Best of all, this equivalence means that numerical values of $|\ln(R/L)|$ actually describe FA as a proportion of the trait mean, so no back-transformation is needed to obtain biological meaning. A simple $\log_e$ transformation of all measurements opens up a whole spectrum of versatile yet straightforward tests, because standard FA analyses applied to ln-transformed data are analyses of size-independent or scale-free FA variation.

- Tests for FA relative to ME will be less sensitive to within-sample FA heterogeneity due to size dependence (*Section IV.A3d*).
- Tests for departures from ideal FA (*Section V.B*) will not be confounded by FA heterogeneity due to size dependence.
- Deviations from symmetry in multiple traits of an individual can be combined to yield a more reliable estimate of individual, organism-wide DI (*Section IV.B3*).
- Multiple traits may be incorporated into a single analysis, thereby increasing the power of tests for FA differences among individuals or groups (*Section VI*) without concerns about unwanted effects of size-dependence (*Section IV.A3*).

Finally, since $\ln(X) = 2.303 \log_{10}(X)$, the same analyses may be done with either natural or base 10 logarithms. The only advantage to natural logs is that they yield numerical estimates of FA that are directly comparable to FA2 (Eq. 3). Of course, the confounding effects of ME apply to FA8a just as they do to all size-standardized indexes (*Section IV.A6*).

IV.B) MULTIVARIATE MEASURES OF INDIVIDUAL FA

IV.B1) Why combine information from multiple traits?

If an organism-wide level of DI exists, then each trait should provide some information about it. Unfortunately, the deviation from symmetry in a single trait estimates the underlying DI variance of that individual with limited confidence (*Section II.C*). However, if each trait provides an independent estimate of the underlying DI variance, then combining information from multiple traits should increase confidence in estimates of individual DI (Leary and Allendorf, 1989; Palmer, 1994; Leung *et al*., 2000). Effectively, each additional trait adds one additional degree of freedom to the estimate.

(- - - - Table 2 approximately here - - - -)

IV.B2) Combining information from multiple traits- prior methods

Pooling information from multiple traits must be done with care for two reasons. First, if FA is not significantly larger than ME for some traits, differences among individuals may be obscured by pooling traits. Therefore, it is wise to exclude traits where FA is not significantly larger than ME before computing these composite indexes. Second, where FA is measured as a proportion of trait size, these multivariate indexes are biased by differences in ME in the same way as individual traits (*Section IV.A*): larger individuals or traits will appear proportionally less variable because ME makes up a smaller proportion of the between-sides variation (Fig. 2c).

Many multivariate indexes have now been advanced (Table 2). Unfortunately, some are vulnerable to size-dependent differences in FA (FA11, FA13), some are only meaningfully applied to meristic traits (FA12), and some simply lack much statistical power (FA16). Nonetheless, others show real promise as general multivariate indexes of individual DI.

Leung et al. (2000) suggest two intriguing new indexes (Table 2). One (FA14) divides each value of |R - L| for a trait of a given individual by the mean |R - L| of that trait for the entire sample. The second (FA15) is a purely nonparametric index based on rank orderings of |R - L|. |R - L| is ranked from high to low independently for each trait in a sample, and the composite measure of individual asymmetry is the sum of these ranks for all traits of an individual.

Both indexes avoid the problems that arise when average FA differs considerably among traits, either due to size-dependence or some other factor, because both express the asymmetry in a single trait of an individual relative to the asymmetry in that trait for the entire sample. But both suffer from two limitations. First, multiple computational steps are required to compute each index, so they are more cumbersome to apply. Second, both yield numerical descriptors of average FA that are not directly comparable among studies. So, although FA14 and FA15 may not be useful descriptors of organism-wide DI, they do offer interesting alternative tests for differences in

organism-wide DI among samples because they are independent of *both* trait size and average trait FA. Therefore, both seem useful as tests of statistical significance.

IV.B3) Combining information from multiple traits- average proportional FA

The simple transformation ln(R) - ln(L) removes scale effects by expressing the difference between sides as a proportion of trait size (*Section IV.A6*). This not only removes within-sample heterogeneity due to size variation that can lead to leptokurtosis of (R - L) for individual traits (*Section IV.A3d*), but it allows a composite measure of FA for an individual to be computed simply as the average of the proportional deviations from symmetry of multiple traits (FA17, Table 2). In addition to being easier to compute, FA17 expresses the FA of an individual in numerical values that are directly comparable to those obtained for single traits (FA2 and FA8a, Table 1).

Finally, a traits x individuals ANOVA on |ln(R) - ln(L)| allows a more powerful test for heterogeneity of DI among individuals (*Section VI.B*), since it pools the information from multiple traits to get a better estimate of the DI of each individual.

IV.B4) Landmark methods

The revolution that is sweeping morphometrics (Rohlf, 1993) offers some intriguing new ways to examine the DI of both size and shape variation of complex structures (Auffray *et al*., 1996; Smith *et al*., 1997; Arnqvist and Martensson, 1998; Klingenberg *et al*., 1998; Klingenberg and McIntyre, 1998; Auffray *et al*., 1999). This revolution focuses on landmarks — developmentally homologous points in either 2D or 3D space (Bookstein, 1992) — rather than conventional measures of distance used in traditional morphometric studies.

For a multivariate method, the procedure is not terribly complex. In a nutshell, the analysis involves four steps (see Klingenberg and McIntyre, 1998, for a nice graphical illustration):

a) Record landmark data. Digitized XY coordinates of multiple landmarks of a single structure (e.g., jaw bone, insect wing) are replicated at least twice independently for each side.

b) Align landmark sets. All the constellations of landmarks for both replicates of both sides are aligned relative to each other as closely as possible using a least-squares Procrustes fitting procedure (Rohlf and Slice, 1990), after all the left-side landmarks have been reflected. The landmark constellations are first centered on their respective centroids, scaled to a common centroid size (average deviation of landmarks from the centroid), and then rotated about the centroids so as to minimize the squared deviations of all landmarks from their respective means.

This transformation yields Procrustes coordinates, where all trait size variation has been removed and only shape variation remains. The distribution of observed landmarks about the mean of each landmark for the entire sample is typically bivariate normal, so the deviation of an aligned right-side landmark from the homologous landmark on the left side of an

individual specimen (XY_{iR} - XY_{iL}) is conceptually the same as (R_i - L_i)/[(R_i + L_i)/2]. Each landmark therefore contains information about DI.

c) Test for differences in FA relative to ME. As in any FA analysis, the X and Y coordinates of all the Procrustes coordinates now contain information about ME (the difference in locations between replicate sets of landmarks) and asymmetry (true difference in location between landmarks of the right and left sides), and the significance of asymmetry relative to ME may be tested using a modification of the standard sides x individuals ANOVA procedure (Palmer and Strobeck, 1986). As with conventional descriptors of asymmetry, the frequency distribution of (XY_{iR} - XY_{iL}) for each landmark, as well as the total shape difference between sides [$\sum$(XY_{iR} - XY_{iL}) / k, where k= the number of landmarks] must be tested for antisymmetry (platykurtosis; *Section V.B3*).

d) Test for differences in FA among individuals or among groups. The average right-left difference of all the aligned landmarks yields a single, multivariate estimate of the deviation from symmetry in an individual (FA18, Table 2), and these can then be analyzed using any of the standard tests for FA differences among individuals or groups (*Section VI*).

Landmark analyses offer two significant advantages over conventional distance analyses. First, trait size FA, overall trait shape FA, and the FA of individual landmarks, may all be compared among individuals or groups. This allows a much more detailed exploration of the effects of the local vs global effects of DI on a structure (Klingenberg and McIntyre, 1998). Second, FA is estimated from multiple traits (landmarks), so it has the potential to give a more robust index of individual DI.

But landmark analyses also have some shortcomings. First, they are limited to single, relatively rigid, elements (e.g., vertebrate bones, arthropod limb segments or wings, fish body outlines). Second, if FA is not greater than ME for all landmarks, then real FA differences at a few landmarks may be swamped out by the noise of ME at others. Third, the Procrustes alignment procedure necessarily makes variation in any one landmark dependent on the variation of all others. In other words, one highly variable landmark will induce apparent variation in the remainder via the least-squares fitting algorithm. Finally, corrections for allometry are not possible, so size-dependent changes in shape, or variability, may confound the interpretation of FA variation.

Landmark analyses seem like a promising new approach to FA variation, but their full potential (and limitations) have yet to be explored.

V) ANALYSES OF FA: VALIDATING THE DATA

V.A) MEASUREMENT ERROR AND REPEATABILITY

Although better than no predictor at all, deviation from symmetry is still a poor predictor of underlying DI of an individual because of two sources of error: measurement error (ME) and sampling error. First, deviations from symmetry are so small that they are typically similar in magnitude to ME (Palmer, 1996). Therefore, ME often contributes a high percentage of the total between-sides variation (Fields *et al*., 1995; Van Dongen and Lens, 2000), and reduces the correlation between observed FA and inferred underlying DI. Second, the deviation from symmetry of a single trait in an individual — even if it were measured without error — only estimates the underlying DI of that individual with one degree of freedom (*Section II.C*).

To have confidence that differences in R - L among individuals are not simply an artifact of ME, the significance of FA relative to ME must be tested. To have confidence that differences in R - L among individuals reflect real differences in underlying DI, and not just sampling error, requires an estimate of the *hypothetical repeatability* (Van Dongen, 1998).

V.A1) Measurement error- the problem

Boring as it may be, attention to ME is perhaps more important than any other aspect of a FA study (Greene, 1984; Palmer and Strobeck, 1986; Swaddle *et al*., 1994; Fields *et al*., 1995; Merilä and Björklund, 1995). If this simple fact were better appreciated, many misleading conclusions in the FA literature or failed studies of FA variation (see review by Simmons *et al*., 1999) might have been avoided.

The claim by some FA enthusiasts that ME cannot generate interesting patterns reflects a remarkable ignorance of elementary statistics. Unlike conventional analyses, where ME simply reduces the signal relative to the noise, ME poses serious problems for FA analyses.

a) Misinterpreting ME as DI. Deviations from symmetry due to ME are indistinguishable from those due to DI (Palmer and Strobeck, 1986; Palmer, 1994), because they are random, independent, and normally distributed about a mean of zero. Therefore, just as for random deviations from symmetry induced by other causes (*Section II.A*), normally distributed deviations about a mean of zero, by themselves, are not unambiguous evidence of DI.

b) ME and FA are often comparable. Deviations from symmetry are often so small that they are similar in size to typical errors in measurement (Greene, 1984; Palmer, 1996), therefore measurements must be taken exceedingly carefully to have any hope of detecting real differences in FA among samples.

c) ME artificially inflates FA. Increasing ME actually increases FA for all indexes of FA variation except those that factor out ME (e.g., FA10a, FA10b, Table 1). Therefore differences in ME, for example among different samples of the same trait measured on different days, can yield

differences in FA that are entirely artificial (see Fig. 7 of Palmer, 1994). More seriously, ME is likely to differ consistently among traits for a variety of reasons. Therefore artificial differences in FA among traits could arise entirely due to differences in ME.

d) ME can not be partitioned out of individual FA for a single trait. Although average ME may be partitioned out of the between-sides variation for a sample via ANOVA (*Section V.A5*), it can not be partitioned out of deviations from symmetry in a single trait in an individual.

e) ME outliers create leptokurtosis. Outlier measurements, due to causes other than ME such as recording errors, transcription errors, data entry errors, calibration errors, sorting errors, etc., are a common cause of leptokurtosis in frequency distributions (see Steps 1 & 2, and Table V.6, of Appendix V). Therefore, leptokurtosis may not be as reliable an indicator of within-sample heterogeneity of DI as some would like (Gangestad and Thornhill, 1999).

f) High ME can create artificial size-dependent FA. If ME (as ME1, see Table 3) is comparable to FA (as FA1, Table 1), and if considerable variation in trait size exists within a sample, FA as a proportion of trait size (e.g., FA2, Table 1) will decline with increasing trait size (*Section IV.A6;* Fig. 2). Also, for traits with the same ME, smaller traits will appear to exhibit lower FA as a proportion of trait size (e.g., FA2) than larger traits.

(- - - - Figure 4 approximately here - - - -)

g) ME obscures FA variation. Finally, as it does in any conventional analysis, ME potentially obscures differences in underlying DI. Even in the absence of ME, statistical support for parallel variation in |R - L| between pairs of traits on the same individuals may be low (Fig. 4), simply because |R - L| estimates the underlying DI variance with only one degree of freedom (*Section II.C*). For example, even with high DI (16-fold range) and with no ME, asymmetries in one trait will only show a correlation of < 0.3 with asymmetries of other traits on the same individuals. Introducing ME reduces these correlations even further (Fig. 4).

(- - - - Table 3 approximately here - - - -)

V.A2) Measurement error- description

Discussions of ME can be quite confusing if underlying error *variances* such as σ^2_{ME} (the variance of repeat measurements on a single side, Table 3a) are not distinguished from numerical *descriptors* of ME, like ME1 (Table 3b). Therefore when ME is referred to in general, it should mean σ^2_{ME}. However, when referring to a specific descriptor of ME some convention is required to indicate which one is being used (e.g., Table 3).

Measurement error may be quantified in several ways, some of which are more informative than others (Table 3). Some descriptors (ME1, ME2) report ME in the original units of

measurement. Because they describe the actual ME for a trait, at least one of these should be reported for each trait in every FA analysis (see Appendix V). One descriptor (ME3) simply expresses the average difference between replicate measurements as a percent of average difference between sides. Others (ME4, ME5) don't describe ME directly, but rather express FA variation as a proportion of the total between sides variation, which includes ME. ME4 and ME5 are simply different ways of measuring repeatability, and are often reported in FA studies because they provide a standardized measure that is easy to understand. The larger the repeatability, the smaller the ME relative to FA.

Some might be puzzled by the equations for ME4 and ME5 (Table 3), both of which yield repeatabilities (r_I) but appear to differ from the more familiar equation (Lessels and Boag, 1987; r_I is sometimes referred to as the "intraclass correlation coefficient", thought it doesn't mean correlation in the sense that the term is used now, Sokal and Rohlf, 1995, p. 214):

$$r_I = \frac{s_x^2}{s_x^2 + s_e^2} \qquad (5)$$

where s_x^2 is the best estimate of the true underlying variance in some variable *x* and s_e^2 is the error variance (note that the denominator [$s_x^2 + s_e^2$] is actually the observed variation, which includes both the underlying and error variation). The equations differ, though, only because ME4 and ME5 refer to MS from ANOVA, which may include one or more variance components, whereas Eq. 5 refers to the actual underlying variances. ME4 and ME5 yield the same number as Eq. 5 because the expected value for $MS_{individuals}$ or $MS_{interaction}$ is $s_e^2 + n\, s_x^2$ and that for MS_{error} is s_e^2 (Palmer and Strobeck, 1986; Sokal and Rohlf, 1995, p. 214). Substituting these in the equations for ME4 or ME5 yields

$$r_I = \frac{(s_e^2 + n\, s_x^2) - s_e^2}{(s_e^2 + n\, s_x^2) + (n-1)\, s_e^2} = \frac{n\, s_x^2}{n\, s_x^2 + n\, s_e^2} = \frac{s_x^2}{s_x^2 + s_e^2}.$$

ME4 and ME5 are simpler because they may be computed easily using the MS from ANOVA. No fiddling is needed to estimate the variance components of Eq. 5. Furthermore, this approach avoids the confusion that arises commonly when MS from ANOVA, which are indeed variances, are mistaken for the individual variance components that contribute to them (Lessels and Boag, 1987).

V.A3) Perils of repeatability

Many studies of FA variation express ME as a repeatability coefficient (ME4, ME5, Table

3). These coefficients are appealing because they express the variation among traits as a proportion of the total variation (including ME). So, for example, a repeatability of 0.9 or 90% implies that 90% of the total observed variation among a set of replicate measurements is due to underlying variation in the trait being measured and 10% is due to error. In studies of FA, it is critical to remember that the 'trait' whose repeatability should be measured is *deviation from symmetry* (R - L), not trait *size* (R or L). Even if the repeatability of trait size measurements is very high, the repeatability of FA may be extremely low, simply because FA is typically such a small percentage of trait size (Fields *et al*., 1995).

Measures of %ME (ME3) or repeatability (ME4, ME5), however, all suffer from the same limitation. On the one hand, as dimensionless numbers, they are convenient and easy to interpret. On the other hand, true ME can not be obtained from them with confidence because differences in repeatability or reliability can increase due either to a decrease in ME or to an increase in FA (Table 3d). More seriously, values for repeatability may be greatly inflated, or %ME greatly decreased, if an investigator specifically chooses individuals with the widest possible range of asymmetry variation to estimate ME. In other words, if the subsample of individuals on which repeatability is estimated exhibits a wider range of FA than other samples in the study, then the repeatability will be artificially inflated. Finally, to obtain an actual measure of ME (with units) requires substituting some estimate of FA into the above equations. Fortunately, this problem does not arise where repeat measurements have been taken on all individuals in a study, and where ME3, ME4, or ME5 are computed using all the data.

As a consequence, whenever statistical measures of repeatability are used to describe the size of FA relative to ME, the ME should also be given, either as ME1 or ME2. Either will be in the units of original measurement, and the two are easily interconverted (Table 3).

V.A4) Hypothetical repeatability (*R*) and among-individual variation in DI

Two issues arise when testing for associations between individual FA and a particular phenomenon of interest, or when estimating the heritability of FA. First, grounds must exist for believing that DI truly varies among individuals. Second, correlations with FA will underestimate correlations with DI.

a) Tests for DI heterogeneity. The total observed FA variation within a sample, V_{R-L} = var(R-L) = FA4 (Palmer and Strobeck, 1986), arises from at least three sources (modified slightly from the notation of Whitlock, 1998):

i) V_{DI} = true variation in the DI among individuals within a sample (V_{DI} is therefore the variance of a set of DI variances).

ii) V_{err} = variation due to the intrinsic uncertainty of deviations from symmetry as predictors of the true DI variance in a given individual (i.e., observed values of (R - L) will still vary considerably even among individuals of identical DI; *Section II.C*), and

iii) V_{me} = variation due to ME.

In a sample where DI truly varies among individuals, the total asymmetry variation among individuals V_{R-L} is some function of V_{DI}, V_{err}, and V_{me}. Therefore, before estimating the heritability of DI — effectively a correlation between parents and offspring — or testing for correlations between individual DI and other phenomena of interest, the following must be true: $V_{R-L} > V_{err} + V_{me}$. If V_{R-L} is not significantly greater than expected for a given $V_{err} + V_{me}$ then no justification exists for testing for correlations with individual DI (i.e., V_{DI} is negligible).

The simplest statistical test for DI heterogeneity within a single sample is a test for leptokurtosis of a single trait (*Section V.B3*). If V_{DI} contributes significantly to var(R-L), then the frequency distribution of (R-L) *must* be leptokurtic (Wright, 1968), because it represents a mixture of individuals each with different DI (e.g., see Fig. 1 of Van Dongen, 1998). Alternatively, where multiple traits have been measured, a traits x individuals ANOVA provides a more powerful test for DI heterogeneity among individuals (*Section VI.B*).

b) Corrections for bias. Even if DI varies significantly among individuals, correlations with individual |R-L| will consistently underestimate correlations with DI, because of uncertainty introduced by V_{me} and V_{err} (Whitlock, 1996). The *hypothetical repeatability* (Van Dongen, 1998) attempts to correct for such a bias.

Equations for hypothetical repeatability (***R***, Table 3) have been derived by Whitlock (1996, 1998), Björklund and Merilä (1997) and Van Dongen (1998). Just as it does for a conventional repeatability (e.g., Eq. 5), ***R*** attempts to estimate the true variation in DI among individuals as a proportion of the total observed FA variation in a sample: $\boldsymbol{R} = V_{DI} / V_{R-L}$. In addition, ***R*** may be used to adjust heritabilities, or correlations with |R - L|, to better reflect correlations with underlying DI (Whitlock, 1996).

When the results of several published studies were examined more closely, |R - L| of one trait correlated more strongly with |R - L| of a second trait as ***R*** increased (Van Dongen and Lens, 2000), reinforcing the belief that differences in DI among individuals are expressed organism-wide and that pooling information from several traits should provide a better estimate of underlying individual DI (see also Lens and van Dongen, 1999, and Section VI.B below). However, in a similar analysis of published heritability estimates, heritability of FA did not increase with increasing ***R***, suggesting that reports of significant heritability of FA may be spurious (Van Dongen and Lens, 2000).

c) Problems with hypothetical repeatability. The concept of hypothetical repeatability is an important one for studies of FA variation among individuals. If ***R*** could be estimated reliably it would be a valuable tool for determining when correlations with |R - L| should be expected, and for adjusting those correlations to provide a better estimate of correlations with underlying DI

(Whitlock, 1996). We therefore applaud the attempts to derive a quantitative descriptor (Whitlock, 1996; Björklund and Merilä, 1997; Van Dongen, 1998; Whitlock, 1998).

Unfortunately, existing derivations of ***R*** all appear flawed in one way or another. Björklund and Merilä (1997) point out that the $CV_{|R-L|} = SD_{|R-L|} / mean_{|R-L|}$ is a constant. But this is only true if DI is invariant. The more DI varies among individuals within a sample, the more leptokurtic the distribution of (R-L) becomes (*Section V.A4a*), and the greater SD|R-L| becomes relative to the mean|R-L|. The suggestion that many observed $CV_{|R-L|}$ are much greater than expected due to large ME (Björklund and Merilä, 1997) is simply not correct. Leptokurtosis is likely responsible since the value for $CV_{|R-L|}$ depends on the leptokurtosis of (R-L), not on ME.

The derivations of Whitlock (1996; 1998) and Van Dongen (1998) appear to make a different mistake. Both derive ***R*** by summing variances among individuals of different underlying DI. But variances are additive only if the mean is the same, not if the means differ. The expected mean of (R-L) is zero, and therefore constant, so summing var(R-L) is entirely appropriate. However, the expected mean of |R-L| clearly depends on the variance (see Fig. 1). Therefore, for two samples of different variance, $var|R-L|_{pooled} \neq var|R-L|_1 + var|R-L|_2$. The impact of this error on derivations of ***R*** is unclear.

In addition, the numerical value for ***R*** is still an *estimate* of the true ***R*** of a sample, so it is subject to uncertainty. Before too much faith is placed in ***R*** as tool to correct for sampling error and ME, it would be helpful to know the standard error of ***R*** for a variety of combinations of V_{DI}, V_{err}, and V_{me}. If ***R*** differs less than 2 SE from zero, then its value as a correction is dubious at best. Valuable as ***R*** may be, we still need a reliable derivation and a standard error. Conclusions based upon existing derivations are therefore difficult to judge.

Finally, although likely obvious to many, a simple observation is worth repeating: the low correlations reported for many associations with FA may simply reflect truly low correlations, not an artificial lowering due to ME and sampling error.

V.A5) ANOVA procedure testing the significance of FA relative to ME

The two-way, mixed model ANOVA procedure (sides= fixed, individuals= random) advanced by Palmer and Strobeck (1986) to test for the significance of FA relative to ME is easy to conduct and easy to interpret (explained in detail in Palmer, 1994). This procedure also:

- tests for the significance of DA
- allows an estimate of repeatability to be computed (ME5, Table 3), and
- permits the only estimates of FA where ME has been factored out (FA10a, FA10b, Table 1),

For these reasons, it remains a valuable tool for studying FA variation using either conventional (Palmer, 1994; Swaddle *et al*., 1994; Merilä and Björklund, 1995; Björklund and Merilä, 1997; Van Dongen, 1999) or multivariate (Klingenberg and McIntyre, 1998) measures of asymmetry.

Appendix V illustrates a complete worked example of the procedure.

V.A6) What to do when replicate measures of all individuals is not practical

Sometimes the sheer size of a study is so large that replicate measurements of all traits in all individuals is impractical. Also, in unusual cases where ME is small (e.g., where ME1 of Table 3 is <10% of FA1), replicate measurements of all traits in all individuals may not be necessary. Under these circumstances, it is sufficient to take repeat measurements on only a subset of individuals. A sides-x individuals ANOVA (*Section V.A5*) on this subset is still necessary to confirm that FA exceeds ME, as is a report of the level of ME (as ME1 or ME2, *Section V.A2*).

It should be obvious that if greater care is taken during repeat measurements than during the rest of the study, the true ME will be underestimated and the conclusions about FA variation (or its absence) may be meaningless. Therefore, every effort must be made to ensure the measurement protocol when estimating ME on a subsample includes *all* the sources of error present in the main study (day-to-day variation, among-observer variation, wear and tear on specimens, calibration errors in digitizing systems, effects of inexperience, etc.). One solution would be to conduct the first set of replicate measurements of a subsample of specimens at the beginning of the study and the second set at the end.

V.B) DEPARTURES FROM IDEAL FA

V.B1) Directional asymmetry

Traits where one side is consistently larger than the other in the same direction (DA) complicate both the analysis and interpretation of FA variation. Analyses are complicated, because a number of FA indexes, including FA1, FA2, FA3 and FA8a (Table 1), are artificially inflated by DA (Palmer, 1994). Interpretation is complicated because even if DA is factored out statistically (Graham *et al*., 1998), the remaining between-sides variation is likely a complex mix of directional genetic effects, directional environmental effects (likely via the effects of growth rate on allometry), and DI. Therefore, as a rule, if traits exhibit significant DA they are best excluded from FA analyses (Palmer and Strobeck, 1992; Palmer, 1994).

Unfortunately, sometimes even very slight DA may become significant statistically in studies involving large samples. Under these circumstances, too many data might be lost if these traits were excluded, and factoring out DA would seem desirable. The critical question here is: At what point is DA so small that it is unlikely to confound interpretations of FA variation? Any rule is arbitrary, but a potentially useful rule of thumb may help. If DA, as mean(R - L), is no larger than FA4a (Table 1), then the predisposition towards one side is less than the average deviation about mean(R - L). Therefore, since the underlying variation in DA would likely be 10 - 20% of the mean DA — the CV for many traits is commonly in this range (Lande, 1977) — deviations about the mean DA would be due largely to DI.

One insidious way that spurious DA may creep into a FA study is via human handedness. Because most humans are strongly handed (Perelle and Ehrman, 1994), measurements on the right side of an organism might be made slightly, but consistently differently from those on the left. This is a potentially serious problem where measurements require considerable manual dexterity. Helm and Albrecht (2000) report a striking example where statistically significant DA arose exclusively due to the handedness of the observers and suggest ways to avoid this problem.

Tests for DA are an essential step in any FA analysis. They ask nothing more than whether the mean (R-L) differs significantly from zero. Many conventional tests may be applied, including one-sample t-tests of mean(R - L) vs zero, paired t-test of R vs L, the sides x individuals ANOVA procedure (*Section V.A5*), as well as others (Palmer, 1994).

(- - - - Table 4 approximately here - - - -)

V.B2) Departures from normality- the problem

Many factors may cause the distribution of R - L to depart from normality (Table 4). Some departures are mere inconveniences that have nothing to do with DI (Table 4, a.i, a.ii, b.i, and b.ii, b.iii). Others reflect unusual mixtures of different kinds of asymmetry variation (Table 4, a.iii, a.iv, a.v, and b.iv) that may or may not be detectable via mixture analysis (Van Dongen *et al*., 1999). However, some (Table 4, b.v, b.vi, and c.i) are likely occurrences in studies of subtle asymmetries and have significant implications for analysis and interpretation.

a) Skew. Skew — the third central moment of a frequency distribution (Sokal and Rohlf, 1995) — refers to departures from normality that are asymmetrically distributed about the mean. It ranges from $-\infty$ (an elongate tail to the left) to $+\infty$ (an elongate tail to the right). For a normal distribution, skew is zero.

The most common causes of skew are either anomalous data or mixtures of different types of asymmetry variation (Table 4a). Fortunately, the former are readily fixed via careful inspection of the data (see Appendix V for an example of detection and correction), and the latter are largely hypothetical, and therefore not likely a common problem.

b) Kurtosis, general. Kurtosis — the fourth central moment of a frequency distribution (Sokal and Rohlf, 1995) — refers to departures from normality that are symmetrically distributed about the mean. Values of kurtosis range from -2 (extreme platykurtosis) to $+\infty$ (extreme leptokurtosis). Unfortunately, the kurtosis statistic has no simple verbal description (Balanda and MacGillivray, 1988). As a consequence, the history of its interpretation abounds with color and controversy (Balanda and MacGillivray, 1988; Dodge and Rousson, 1999). Reviewing this history helps reveal why a given value of kurtosis is hard to interpret as revealing something specific about the shape of a frequency distribution.

Until the dawn of computers, few biologists or statisticians were interested in verbal interpretations of kurtosis. Originally the sign of the kurtosis statistic was interpreted to reveal a joint excess (positive- or lepto-kurtosis), or a joint deficit (negative or platy-kurtosis), in both the peak *and* the tails of a frequency distribution (see Finucan, 1964). Later, it was interpreted to indicate merely the direction that the frequencies at the *center* of a distribution departed from a normal distribution: positive kurtosis meant an excess and negative kurtosis a deficit. But this was confirmed to be incorrect in four specific examples (Kaplansky, 1945). Ali (1974) then suggested kurtosis was best interpreted only as the *tailedness* of a distribution regardless of what was happening at the peak — positive indicating heavy tails and negative indicating light tails — because observations in the tails contribute disproportionately to kurtosis. Darlington (1970, p. 19) argued that "kurtosis is best described … as a measure of unimodality versus bimodality", or the *tendency toward bimodality* of a distribution: the more negative the kurtosis the more pronounced the bimodality. But Chissom (1970) showed that purely rectangular 'unimodal' distributions yielded negative kurtosis, and Hildebrand (1971) showed that not all bimodal distributions yielded negative kurtosis. Double gamma bimodal distributions (where, at each mode, the distribution of observations may be quite asymmetrical or skewed) could yield kurtosis values that ranged from -2 to 3 (Hildebrand, 1971). Perhaps most seriously of all, Balanda & MacGillivray (1988, p. 114) showed how a single value of kurtosis that is considered mesokurtic (normal) could nonetheless arise from distributions that were either bimodal (double gamma) or narrow-peaked (TukeyLambda; their Fig. 2)! Clearly, the kurtosis statistic by itself tells us nothing specific about distribution shape. Perhaps this should come as no surprise, since a single number is unlikely to capture reliably the many possible ways a symmetrical distribution might depart from normality.

Moors (1986) advanced perhaps the most useful interpretation of kurtosis. Kurtosis describes reasonably well the density of observations at two specific locations on a frequency distribution: one standard deviation above and one standard deviation below the mean (see Fig. 1a). An excess at these two locations yields negative kurtosis (platykurtosis) whereas deficiencies there yield positive kurtosis (leptokurtosis). This view nicely explains how a bimodal distribution can yield either negative or positive values of kurtosis depending on how close the peaks are to the mean. A bimodal distribution, where the distribution about each mode is normal and close to ± 1 SD from the mean, will yield the most extreme negative kurtosis.

Fortunately for studies of FA, most platykurtic distributions of right-left differences appear to be composed of two peaks each of which is roughly normal and similar in size (see Fig. 6 below). Under these conditions, each mode will lie very close to ± 1 SD from the mean, particularly as the bimodality becomes more pronounced (Chissom, 1970). Therefore the kurtosis statistic should be a reasonably sensitive measure of the kind of bimodality likely to be observed in studies of bilateral variation.

That kurtosis describes the concentration of observations around ± 1 SD from the mean

may be seen easily in the descriptive formula for kurtosis (Darlington, 1970; Sokal and Rohlf, 1995):

$$k= [\sum(X_i - \bar{X})^4 / (N*SD^4)] - 3 \quad (6)$$

where N is the sample size, $\bar{X}$ is the sample mean, X_i is the value of X for individual i, and SD is the standard deviation of the sample computed using N rather than N-1. For a given value of the numerator, the larger the SD, the smaller the kurtosis.

The constant three is an arbitrary correction because, if uncorrected, k= 3 for a normal distribution. This correction term, which ensures k= 0 for a normal distribution, makes the values for kurtosis parallel those for skew, which is zero for a normal distribution with no correction.

c) Leptokurtosis. Leptokurtosis can arise from many causes (Table 4b). One or a few extreme measurement errors, or a few injured or damaged individuals, will increase the length of the tails of a frequency distribution of (R - L) and yield leptokurtosis (Table 4, b.i, b.ii). Often these can be detected and eliminated from an analysis by standard outlier tests (see Step 2, Appendix V for an illustration of detection and correction). Variation in ME can also yield leptokurtosis (Table 4, b.iii), but careful records of when or by whom data were recorded can reject this possibility. These latter three causes are likely much more common than generally acknowledged because few studies present data in such a way as to check for them. Leptokurtosis due to a mixture of FA and antisymmetry (Table 4, b.iv) is certainly possible, but largely hypothetical.

The two remaining causes of leptokurtosis, both of which involve within-sample heterogeneity in FA, are widespread and significant to studies of FA. First, heterogeneity will arise if |R - L| increases with trait size (*Section IV.A3d*) and considerable size variation exists within a sample. This heterogeneity doesn't necessarily reflect heterogeneity in underlying DI, it may simply represent size-dependence of variability. Fortunately, if |R - L| scales isometrically with trait size, (R+L)/2, and ME is not too large (*Sections IV.A6, IV.A7*), size-adjusted indexes of FA (FA2, FA6a, FA8a, Table 1) will eliminate this source of heterogeneity.

Second, if two samples of different FA are pooled, the resulting mixture will be leptokurtic: the more extreme the differences in FA between the two samples the greater the leptokurtosis (Palmer and Strobeck, 1992). This occurs any time samples with different variances are pooled (Wright, 1968). Gangestad & Thornhill (1999) and Van Dongen (1998) generalized this observation further, showing that a mixture of individuals with many different levels of DI also exhibits leptokurtosis. This type of heterogeneity — real among-individual variation in DI — is not only likely in studies of FA, but must exist any time significant correlations are found or anticipated between individual |R - L| and some phenomenon of interest (*Section VI.B*), since such correlations, if not spurious, absolutely depend on the existence of among-individual variation in DI.

Unfortunately, as should be apparent from Table 4b, significant leptokurtosis by itself is

not strong evidence for within-sample variation in DI (b.vi) unless other potential causes (b.i - b.v) have been rejected. Therefore the claim that widespread leptokurtosis in studies of FA reveals strong evidence of widespread within-sample heterogeneity in DI (Gangestad and Thornhill, 1999) seems premature. Once again, just because results are consistent with an interesting biological explanation does not mean that explanation is the correct one.

d) Platykurtosis. Platykurtosis arises primarily due to antisymmetry (Table 4c).

The current meaning of the term antisymmetry was advanced by Van Valen (1962, p. 126) as "a bimodal distribution of the signed differences between the sides … or, in less extreme cases, a tendency toward platykurtosis as compared with a normal distribution of the same variance". This definition is more general than Timoféeff-Ressovsky's (1934, p. 79), which referred to an absence, or virtual absence, of symmetrical individuals in a sample.

Curiously, the etymology of the term 'antisymmetry' seems widely unappreciated. For traits that exhibit antisymmetry, the frequency distribution of R - L differences shows a distinct valley between two, typically equally-sized, peaks that are equidistant from zero (see Fig. 6). This valley of 'missing' observations is centered on zero the same way the peak is centered on zero for symmetrical traits, hence the 'anti' in antisymmetry!

Both DA (*Section V.B*) and antisymmetry reflect an innate predisposition towards asymmetry (Palmer and Strobeck, 1992). For all practical purposes, the difference between them is simply the predictability of the *direction* of that predisposition. For DA that predisposition is always toward the same side, but for antisymmetry that predisposition is random in direction (Palmer *et al*., 1993). Unfortunately, although not all that different in underlying cause, DA is easy to detect statistically (*Section V.B1*) but antisymmetry is not.

V.B4) Tests for kurtosis

Tests for significant kurtosis (k) are complicated by three factors. First, different statistical packages compute kurtosis in different ways. Some (Statview, Systat) compute k using Eq. 6 above. Others (SPSS, Excel) compute an 'unbiased estimate' of k (Sokal and Rohlf, 1995):

$$k = \frac{(n+1)n\sum y^4}{(n-1)(n-2)(n-3)\,s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \qquad (7)$$

where y refers to a deviation from the mean. Therefore care must be taken to ensure the proper tests are conducted when assessing the statistical significance of a kurtosis estimate. Fortunately, the following simple test reveals which formula is being used. For a sample of four points (-1, -1, +1, +1) k= -2.0 with Eq. 6 and k= -6.0 with Eq. 7.

The relative merits of Eq. 6 compared to Eq. 7 depend entirely on how the kurtosis statistic

is used. If only a *description* of kurtosis is required, so that it may be tested for statistical significance, then Eq. 6 is preferred. However, to *predict* the kurtosis of a population based on the observed kurtosis of a subsample, Eq. 7 is preferred. Eq. 6 yields a biased estimate of k for small sample sizes which Eq. 7 avoids (Sokal and Rohlf, 1995). As sample size increases the two forms converge (Table 5).

(- - - - Figure 5 & Table 5 approximately here - - - -)

A second complication is distribution symmetry. The frequency distribution of the kurtosis statistic, regardless of how it is computed, is highly skewed unless sample sizes are large (>200, Fig. 5). Therefore the single standard error sometimes suggested for testing the significance of kurtosis (Sokal and Rohlf, 1995; Zar, 1999), or yielded by some statistics packages, will yield very misleading conclusions, particularly about the significance of platykurtosis at small sample sizes (N < 100). To assist with studies of FA, we tabulate critical values for both platykurtosis and leptokurtosis over the range of sample sizes normally encountered in studies of FA (Table 5). More extensive critical values for kurtosis based on Eq. 6 may be found in the original sources (Pearson and Hartley, 1966; D'Agostino, 1986) and for Eq. 7 (leptokurtosis only) in Zar (1999, Table B.23).

(- - - - Figure 6 approximately here - - - -)

A third complication is limited statistical power. Unless sample sizes are rather large, the kurtosis statistic has limited power to detect antisymmetry (Fig. 6), even when the proper critical values are used (Table 5). For N= 10, the power of k — the percent of trials that reached statistical significance — barely surpassed 60% even when the distance between peaks (2D) was ten times the SD of (R-L) about each peak. Similarly, for N= 20, the power of k was only 80% ($\alpha = 0.05$, Fig. 6a) or 60% ($\alpha = 0.01$, Fig. 6b) when the distance between peaks (2D) was five times the SD of (R-L) about each peak. When antisymmetry is weak, such as when the distance between peaks is twice the SD about each peak, the power is severely limited: sample sizes of 200 only achieve 60% and about 30% power for $\alpha = 0.05$ and $\alpha = 0.01$, respectively. Mixture analysis may provide some help here (e.g., see Van Dongen *et al*., 1999), particularly if antisymmetry is assumed to arise as a mixture of two normal distributions that are equidistant from zero, but to our knowledge, no power analysis comparable to that of Fig. 6 has been conducted.

Because leptokurtosis arises from a heterogeneous mixture of DI variances, a more powerful and direct test for heterogeneity of DI among individuals is a traits x individuals ANOVA on $|\ln(R) - \ln(L)|$ (*Section VI.B*). This test pools the information from multiple traits to get a better estimate of the actual underlying DI of an individual. Clearly, in the absence of statistical evidence for DI heterogeneity among individuals, estimates of the heritability of DI or reports of correlations between individual DI and fitness, quality, attractiveness or other traits of interest, are not very

informative (*Section VI.B*).

VI) ANALYSES OF FA VARIATION: TESTING FOR DIFFERENCES

VI.A) LEVENE'S TEST FOR HETEROGENEITY OF VARIANCE

Tests for differences in FA at any level, among individuals, traits, or samples, are fundamentally tests for heterogeneity of variance because FA estimates a variance (*Section II.A*).

Levene's test (Levene, 1960) is a widely under-appreciated but versatile and easy to use test for heterogeneity of variance (Van Valen, 1978; Palmer, 1994). Although not the most powerful test if distributions are truly normal, more powerful tests are so sensitive to departures from normality that their use is strongly discouraged (Van Valen, 1978). Levene's test is, however, the most powerful test of the two common tests that are least sensitive to departures from normality (Palmer and Strobeck, 1992).

Levene's test works by transforming signed deviations from the mean into absolute deviations. This transforms a symmetrical normal distribution into a highly asymmetrical, truncated normal distribution that is skewed to the right (Fig. 1). As a consequence, the mean of the absolute deviations estimates the SD of the untransformed normal distribution (*Section IV.A2*).

A significant advantage to Levene's test, compared to other tests for heterogeneity of variance, is the ease with which it may be applied in a variety of ANOVA designs (Yezerinac *et al*., 1992; Palmer, 1994; Crespi and Vanderkist, 1997), thereby avoiding the knotty problem of multiple single-factor tests. In addition, by transforming signed deviations to absolute deviations, Levene's test may therefore simultaneously test for differences between samples or traits, and also for interactions between samples or traits. Furthermore, it is conceptually and computationally straightforward, and may be conducted with any conventional statistical package.

Below, we illustrate three applications of Levene's test to situations commonly encountered in FA analyses.

(- - - - Tables 6 & 7 approximately here - - - -)

VI.B) DIFFERENCES AMONG INDIVIDUALS (MULTIPLE TRAITS)

Where asymmetry has been measured for multiple traits, each trait provides an independent estimate of the underlying DI of an individual (*Section II.C*). To take advantage of this, though, the information from multiple traits must somehow be combined. In addition, differences in FA due purely to trait size need to be removed (*Sections IV.A3 -IV.A7*).

A two-way ANOVA (traits x individuals) on replicate measurements of $|\ln(R) - \ln(L)|$ achieves this test nicely (Tables 6, 7). This is a fully model II ANOVA. This test may be expanded easily to include as many traits, individuals, and replicate measurements as desired. If neither traits

nor individuals vary much in size, the same analysis could be conducted with |R - L| instead of |ln(R) - ln(L)|. The MS_{err} here is a measure of the average within-group variance among replicate measurements.

Note that this traits x individuals Levene's test is a more powerful test for differences in DI among individuals within a sample than tests for leptokurtosis (*Section V.B3*) for two reasons. First, for sample sizes typically found in studies of FA, kurtosis is not a very powerful statistic for detecting departures from normality (Fig. 6). Second, this Levene's test combines information from multiple traits, thereby yielding a better estimate of the DI of an individual: each additional trait adds a degree of freedom (*Section II.C*). Therefore if a traits x individuals Levene's test yields no significant effect due to individuals, no statistical support exists for variation in DI among individuals. Clearly, with no such statistical support, attempts to estimate the heritability of DI, or correlations between DI and individual fitness or quality, are pointless.

(- - - - Tables 8 & 9 approximately here - - - -)

VI.C) DIFFERENCES BETWEEN TWO SAMPLES (MULTIPLE TRAITS)

In the same way that additional traits provide more power when testing for heterogeneity of DI among individuals (*Section VI.B*) they also provide more power to a Levene's test when testing for differences in FA among groups. If traits differ in size, and FA depends on trait size, then, as before, a size-corrected measure of FA should be used (*Sections IV.A3 -IV.A7*).

Tables 8 and 9 illustrate a hypothetical two-way ANOVA testing for differences in FA between two groups (e.g., sexes) and among multiple traits. This is a mixed-model ANOVA (group= fixed, trait= random). This test may also be expanded easily to include as many traits, groups, or individuals as desired. If neither groups nor traits differ much in size, the same analysis could be conducted with |R - L| instead of |ln(R) - ln(L)|. See Appendix V for some worked examples. The MS_{err} here is the average within-group variation in FA.

(- - - - Tables 10 & 11 approximately here - - - -)

VI.D) THREE-WAY AND HIGHER ORDER INTERACTIONS

Levene's test is a particularly attractive test for FA variation because it may be generalized to any ANOVA design imaginable. In this manner, information from different traits, different subsets of individuals (e.g., different sexes), and different treatments of interest (e.g., stress levels) may be combined into a single analysis.

Tables 10 and 11 illustrate a hypothetical three-way ANOVA. See Step 10 (Appendix V) for a fully worked example using published data. The MS_{err} here (Table 11) is the average within-group variation in FA (i.e., among-individual variation).

VII) CONCLUSIONS

Fluctuating asymmetry analyses are neither conceptually difficult nor computationally complex. However, attention to a few fundamental details will greatly improve the quality of FA studies. First, choose traits carefully: examine many traits initially then choose those most appropriate for the study of FA. Avoid traits a) that do not exhibit ideal FA (i.e., that exhibit significant DA or antisymmetry; *Section III.A*), b) that are vulnerable to plasticity or wear (*Section III.B*), c) where ME is a high percentage of FA (*Section V.A*), and d) where size-dependence of FA does not exhibit simple allometry (*Section IV.A3*). Second, inspect the data carefully via scatterplots and frequency distributions to ensure outlier measurements or outlier individuals are not confounding estimates of FA (Appendix V, Steps 1-5). Third, use multiple traits per individual wherever possible. These provide improved power for detecting differences in DI among individuals (*Section VI.B*) and among populations (*Section VI.C*). Fourth, when testing for correlations between individual FA and some factor of interest, or when estimating the heritability of FA, confirm that FA varies significantly among individuals first (*Sections V.A4, VI.B*). Fourth, use a single multi-way analysis rather than several simpler analyses to avoid the problems that arise when conducting multiple statistical tests (*Section VI.D*). Finally, where alternate tests of the same hypothesis make different assumptions, and where these assumptions are hard to validate, multiple tests are advised. If different tests yield the same or similar results, then clearly the results are robust even if assumptions are violated.

As is so often the case, these rules are very similar to those for any well-conducted study. A wider adherence to them would significantly improve studies of FA variation.

ACKNOWLEDGEMENTS

This research was supported by NSERC Operating Grants A7245 (to ARP) and A0502 (to CS). We thank Michal Polak for his invitation to contribute, for his comments on the MS, and for his generous efforts to bring this volume together. We are deeply grateful to Berni Crespi for providing the raw data used to illustrate an FA analysis (Appendix V). An anonymous reviewer offered numerous helpful and lexicologically dextrous comments on the MS.

LITERATURE CITED

Ali, M. M. 1974. Stochastic ordering and kurtosis measure. Journal of the American Statistical Association **69**: 543-545.

Angus, R. A. 1982. Quantifying fluctuating asymmetry: not all methods are equivalent. Growth **46**: 337-342.

Angus, R. A., and R. H. Schultz. 1983. Meristic variation in homozygous and heterozygous fish. Copeia **1983**: 287-299.

Armstrong, R. A., and S. N. Smith. 1992. Lobe growth variation and the maintenance of symmetry in foliose lichen thalli. Symbiosis **12**: 145-158.

Arnqvist, G., and T. Martensson. 1998. Measurement error in geometric morphometrics: Empirical strategies to assess and reduce its impact on measures of shape. Acta Zoologica Academiae Scientiarum Hungaricae **44**: 73-96.

Atchley, W. R., C. T. Gaskins, and D. Anderson. 1976. Statistical properties of ratios. I. empirical results. Systematic Zoology **25(2)**: 137-148.

Auffray, J.-C., P. Alibert, S. Renaud, A. Orth, and F. Bonhommeet. 1996. Fluctuating asymmetry in *Mus musculus* subspecific hybridization: Traditional and Procrustes comparative approach. Pp. 275-283 in Advances in Morphometrics, L. F. Marcus, M. Corti, A. Loy, G. Naylor and D. Slice, eds. Plenum, New York.

Auffray, J. C., V. Debat, and P. Alibert. 1999. Shape asymmetry and developmental stability. Pp. 309-324 in On Growth and Form: Spatio-temporal Pattern Formation in Biology, M. A. J. Chaplain, G. D. Singh and J. C. McLachlan, eds. Wiley, New York.

Bagchi, S. K., V. P. Sharma, and P. K. Gupta. 1989. Developmental instability in leaves of *Tectona grandis.* Silvae Genetica **38**: 1-6.

Balanda, K. P., and MacGillivray. 1988. Kurtosis: A critical review. American Statistician **42**: 111-119.

Berry, R. J. 1968. The biology of non-metrical variation in mice and men. Pp. 103-133 in The Skeletal Biology of Earlier Human Populations, D. R. Brothwell, ed. Pergamon Press, New York.

Björklund, M., and J. Merilä. 1997. Why some measures of fluctuating asymmetry are so sensitive to measurement error. Annales Zoologici Fennici **34**: 133-137.

Bjorksten, T. A., K. Fowler, and A. Pomiankowski. 2000. What does sexual trait FA tell us about stress? Trends in Ecology and Evolution **15**: 163-166.

Bookstein, F. L. 1992. Morphometric Tools for Landmark Data: Geometry and Biology. Cambridge Univ. Pr., New York.

Bradshaw, A. D. 1965. Evolutionary significance of phenotypic plasticity in plants. Advances in Genetics **13**: 115-155.

Brakefield, P. M., and C. J. Breuker. 1996. The genetical basis of fluctuating asymmetry for developmentally integrated traits in a butterfly eyespot pattern. Proceedings of the Royal Society of London. Series B **263**: 1557-1563.

Bromberg, M. B., and L. Jaros. 1998. Symmetry of normal motor and sensory nerve conduction measurements. Muscle and Nerve **21**: 498-503.

Brown, C. R., and M. B. Brown. 1998. Intense natural selection on body size and wing and tail asymmetry in cliff swallows during severe weather. Evolution **52**: 1461-1475.

Chissom, B. S. 1970. Interpretation of the kurtosis statistic. American Statistician **24(Oct)**: 19-22.

Clarke, G. M. 1995. Relationships between developmental stability and fitness: Application for conservation biology. Conservation Biology **9**: 18-24.

Clarke, G. M. 1998. Developmental stability and fitness: The evidence is not quite so clear. American Naturalist **152**: 762-766.

Crespi, B. J., and B. A. Vanderkist. 1997. Fluctuating asymmetry in vestigial and functional traits of a haplodiploid insect. Heredity **79**: 624-630.

Cuthill, I. C., J. P. Swaddle, and M. S. Witter. 1993. Fluctuating asymmetry. Nature **363**: 217-218.

D'Agostino, R. B. 1986. Tests for the normal distribution. Pp. 367-419 in Goodness-of-Fit Techniques, R. B. D'Agostino and M. A. Stephens, eds. Marcel Dekker Inc., New York.

Darlington, R. B. 1970. Is kurtosis really peakedness? American Statistician **24(Apr)**: 19-22.

Desbiez, M. O., M. Tort, and M. Thellier. 1991. Control of a symmetry-breaking process in the course of the morphogenesis of plantlets of *Bidens pilosa* L. Planta **184**: 397-402.

Dodge, Y., and V. Rousson. 1999. The complications of the fourth central moment. American Statistician **53**: 267-269.

Dormer, K. J., and J. Hucker. 1957. Observations on the occurrence of prickles on the leaves of *Ilex aquifolium*. Annals of Botany (New Series) **21**: 385-398.

Dufour, K., and P. J. Weatherhead. 1996. Estimation of organism-wide asymmetry in red-winged blackbirds and its relation to studies of mate selection. Proceedings of the Royal Society of London. Series B **263**: 769-775.

Evans, M. R., and B. J. Hatchwell. 1993. New slants on ornament asymmetry. Proceedings of the Royal Society of London. Series B **251**: 171-177.

Fields, S. J., M. Spiers, I. Herschkovitz, and G. Livshits. 1995. Reliability of reliability coefficients in the estimation of asymmetry. American Journal of Physical Anthropology **96**: 83-87.

Finucan, H. M. 1964. A note on kurtosis. Journal of the Royal Statistical Society **B26**: 111-112.

Freeman, D. C., J. H. Graham, and J. M. Emlen. 1993. Developmental stability in plants: Symmetries, stress and epigenesis. Genetica **89**: 97-119.

Futuyma, D. J. 1986. Evolutionary Biology. Sinauer, Sunderland, MA. pp.

Gangestad, S. W., and R. Thornhill. 1999. Individual differences in developmental precision and fluctuating asymmetry: a model and its implications. Journal of Evolutionary Biology **12**: 402-416.

Graham, J. H., J. M. Emlen, D. C. Freeman, L. J. Leamy, and J. A. Kieser. 1998. Directional asymmetry and the measurement of developmental instability. Biological Journal of the Linnean Society **64**: 1-16.

Graham, J. H., D. C. Freeman, and J. M. Emlen. 1993. Antisymmetry, directional asymmetry, and dynamic morphogenesis. Genetica **89**: 121-137.

Greene, D. L. 1984. Fluctuating dental asymmetry and measurement error. American Journal of Physical Anthropology **65**: 283-289.

Gummer, D. L., and R. M. Brigham. 1995. Does fluctuating asymmetry reflect the importance of traits in little brown bats (*Myotis lucifugus*). Canadian Journal of Zoology **73**: 990-992.

Heard, S. B., M. A. Campbell, M. L. Bonine, and S. D. Hendrix. 1999. Developmental instability in fragmented populations of prairie phlox: A cautionary tale. Conservation Biology **13**: 274-281.

Helm, B., and H. Albrecht. 2000. Human handedness causes directional asymmetry in avian wing length measurements. Animal Behaviour **60**: 899-902.

Hermanussen, M., K. Geiger-Benoit, and J. Burmeister. 1989. Analysis of differential growth of the right and the left leg. Human Biology **61**: 133-141.

Hildebrand, D. K. 1971. Kurtosis measures bimodality? American Statistician **25(Feb)**: 42-43.

Houle, D. 1997. Comment on "A meta-analysis of the heritability of developmental stability" by Møller and Thornhill. Journal of Evolutionary Biology **10**: 17-20.

Houle, D. 1998. High enthusiasm and low *R*-squared. Evolution **52**: 1872-1876.

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

Huxley, J. S. 1932. Problems of Relative Growth. Methuen & Co. Ltd., London.

Kaplansky, I. 1945. A common error concerning kurtosis. Journal of the American Statistical Association **40**: 259.

Kendall, M. G., and A. Stuart. 1951. The Advanced Theory of Statistics. Hafner, London.

Klingenberg, C. P., and G. S. McIntyre. 1998. Geometric morphometrics of developmental instability: Analyzing patterns of fluctuating asymmetry with Procrustes methods. Evolution **52**: 1363-1375.

Klingenberg, C. P., G. S. McIntyre, and S. D. Zaklan. 1998. Left-right asymmetry of fly wings and the evolution of body axes. Proceedings of the Royal Society of London. Series B **265**: 1255-1259.

Klingenberg, C. P., and H. F. Nijhout. 1998. Competition among growing organs and developmental control of morphological asymmetry. Proceedings of the Royal Society of

London. Series B **265**: 1135-1139.

Klingenberg, C. P., and S. D. Zaklan. 2000. Morphological integration between developmental compartments in the *Drosophila* wing. Evolution **54**: 1273-1285.

Lande, R. 1977. On comparing coefficients of variation. Systematic Zoology **26**: 214-217.

Leamy, L. 1993. Morphological integration of fluctuating asymmetry in the mouse mandible. Genetica **89**: 139-153.

Leary, R. F., and F. W. Allendorf. 1989. Fluctuating asymmetry as an indicator of stress in conservation biology. Trends in Ecology and Evolution **4**: 214-217.

Lens, L., and S. van Dongen. 1999. Evidence for organism-wide asymmetry in five bird species of a fragmented afrotropical forest. Proceedings of the Royal Society of London. Series B **266**: 1055-1060.

Lessels, C. M., and P. T. Boag. 1987. Unrepeatable repeatabilities: A common mistake. Auk **104**: 116-121.

Leung, B. 1998. Correcting for allometry in studies of fluctuating asymmetry and quality within samples. Proceedings of the Royal Society of London. Series B **265**: 1623-1629.

Leung, B., and M. R. Forbes. 1996. Fluctuating asymmetry in relation to stress and fitness: Effects of trait type as revealed by meta-analysis. Ecoscience **3**: 400-413.

Leung, B., and M. R. Forbes. 1997. Modelling fluctuating asymmetry in relation to stress and fitness. Oikos **78**: 397-405.

Leung, B., M. R. Forbes, and D. Houle. 2000. Fluctuating asymmetry as a bioindicator of stress: Comparing efficacy of analyses involving multiple traits. American Naturalist **155**: 101-115.

Levene, H. 1960. Robust tests for equality of variances. Pp. 278-292 in Contributions to Probability an Statistics, I. Olkin, ed. Stanford Univ. Press, Stanford.

Lewontin, R. C. 1966. On the measurement of relative variability. Systematic Zoology **15**: 141-142.

Ludwig, W. 1932. Das Rechts-Links Problem im Teirreich und beim Menschen. Springer, Berlin. 496 pp.

Malina, R. M. 1983. Human growth, maturation, and regular physical activity. Acta Medica Auxologica **15**: 5-27.

Mather, K. 1953. Genetical control of stability in development. Heredity **7**: 297-336.

Merilä, J., and M. Björklund. 1995. Fluctuating asymmetry and measurement error. Systematic Biology **44**: 97-101.

Møller, A. P. 1997. Developmental stability and fitness: A review. American Naturalist **149**: 916-932.

Møller, A. P., and J. P. Swaddle. 1997. Developmental Stability and Evolution. Oxford Univ. Press, Oxford.

Moors, J. J. A. 1986. The meaning of kurtosis: Darlington revisited. American Statistician **40**:

283-284.

Olsen, B. R., A. M. Reginato, and W. Wang. 2000. Bone development. Annual Review of Cell and Developmental Biology **16**: 191-220.

Palmer, A. R. 1994. Fluctuating asymmetry analyses: A primer. Pp. 335-364 in Developmental Instability: Its Origins and Evolutionary Implications, T. A. Markow, ed. Kluwer, Dordrecht, Netherlands.

Palmer, A. R. 1996. Waltzing with asymmetry. BioScience **46**: 518-532.

Palmer, A. R. 1999. Detecting publication bias in meta-analyses: A case study of fluctuating asymmetry and sexual selection. American Naturalist **154**: 220-233.

Palmer, A. R. 2000. Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. Annual Review of Ecology and Systematics **31**: 441–480.

Palmer, A. R., and L. M. Hammond. 2000. The Emperor's codpiece: A post-modern perspective on biological asymmetries. International Society of Behavioral Ecology Newsletter **12**: 13-20.

Palmer, A. R., and C. Strobeck. 1986. Fluctuating asymmetry: measurement, analysis, patterns. Annual Review of Ecology and Systematics **17**: 391-421.

Palmer, A. R., and C. Strobeck. 1992. Fluctuating asymmetry as a measure of developmental stability: Implications of non-normal distributions and power of statistical tests. Acta Zoologica Fennica **191**: 57-72.

Palmer, A. R., C. Strobeck, and A. K. Chippindale. 1993. Bilateral variation and the evolutionary origin of macroscopic asymmetries. Genetica **89**: 201-218.

Parsons, P. A. 1992. Fluctuating asymmetry: A biological monitor of environmental and genomic stress. Heredity **68**: 361-364.

Paxman, G. J. 1956. Differentiation and stability in the development of *Nicotiana rustica*. Annals of Botany **20**: 331-347.

Pearson, E. S., and H. O. Hartley. eds. **1966.** Biometrika Tables for Statisticians. Cambridge Univ. Press, Cambridge, UK.

Perelle, I. B., and L. Ehrman. 1994. An international study of human handedness: The data. Behavior Genetics **24**: 217-227.

Rohlf, F. J. 1993. A revolution in morphometrics. Trends in Ecology and Evolution **8**: 129-132.

Rohlf, F. J., and D. Slice. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. Systematic Zoology **39**: 40-59.

Rowe, L., R. R. Repasky, and A. R. Palmer. 1997. Size-dependent asymmetry: Fluctuating asymmetry versus antisymmetry and its relevance to condition-dependent signaling. Evolution **51**: 1401-1408.

Roy, S. K. 1958. The regulation of petal number. Current Science **27**: 134-135.

Sakai, K. I., and Y. Shimamoto. 1965. Developmental instability in leaves and flowers of Nicotiana tabacum. Genetics **51**: 801-813.

Sherry, R. A., and E. M. Lord. 1996. Developmental stability in leaves of *Clarkia temblorіensis* (Onagraceae) as related to population outcrossing rates and heterozygosity. Evolution **50**: 80-91.

Simmons, L. W., J. L. Tomkins, J. S. Kotiaho, and J. Hunt. 1999. Fluctuating paradigm. Proceedings of the Royal Society of London. Series B **266**: 593-595.

Smith, B. H., S. M. Garn, and P. E. Cole. 1982. Problems of sampling and inference in the study of fluctuating dental asymmetry. American Journal of Physical Anthropology **58**: 281-289.

Smith, D. R., B. J. Crespi, and F. L. Bookstein. 1997. Fluctuating asymmetry in the honey bee, *Apis mellifera*: Effects of ploidy and hybridization. Journal of Evolutionary Biology **10**: 551-574.

Smith, L. D., and A. R. Palmer. 1994. Effects of manipulated diet on size and performance of Brachyuran crab claws. Science **264**: 710-712.

Smith, L. H. 1998. Asymmetry of Early Paleozoic trilobites. Lethaia **31**: 99-112.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Freeman, New York.

Solangaarachchi, S. M., and J. L. Harper. 1989. The growth and asymmetry of neighbouring plants of white clover (*Trifolium repens* L.). Oecologia (Berlin) **78**: 208-213.

Soulé, M. E., and J. Couzin-Roudy. 1982. Allomeric variation. 2. Developmental instability of extreme phenotypes. American Naturalist **120**: 765-786.

Sullivan, M. S., P. A. Robertson, and N. A. Aebischer. 1993. Fluctuating asymmetry measurement. Nature **361**: 409-410.

Sumner, J. L., and R. R. Huestis. 1921. Bilateral asymmetry and its relation to certain problems in genetics. Genetics **6**: 445-485.

Swaddle, J. P., M. S. Witter, and I. C. Cuthill. 1994. The analysis of fluctuating asymmetry. Animal Behaviour **48**: 986-989.

Tarasjev, A. 1995. Relationship between phenotypic plasticity and developmental instability in Iris pumila L. Genetika **31**: 1655-1663.

Timoféeff-Ressovsky, N. W. 1934. Über der Einfluss des genotypischen Milieus und der Aussenbedingungen auf die Realisation des Genotypes. Mach. Ges. Wiss. Göttingen, Math.-Physik. Klasse, Fachgruppe 6 **1**: 53-106.

Travis, J. 1994. Evaluating the adaptive role of morphological plasticity. Pp. 99-122 in Ecological Morphology. Integrative Organismal Biology, P. C. Wainwright and S. M. Reilly, eds. Univ. Chicago Press, Chicago.

Trinkaus, E. 1994. Postcranial robusticity in *Homo*. II: Humeral bilateral asymmetry and bone plasticity. American Journal of Physical Anthropology **93**: 1-34.

Van Dongen, S. 1998. How repeatable is the estimation of developmental stability by fluctuating asymmetry? Proceedings of the Royal Society of London. Series B **265**: 1423-1427.

Van Dongen, S. 1999. Accuracy and power in fluctuating asymmetry studies: Effects of sample size and number of within-subject repeats. Journal of Evolutionary Biology **12**: 547-550.

Van Dongen, S., and L. Lens. 2000. The evolutionary potential of developmental stability. Journal of Evolutionary Biology **13**: 326-335.

Van Dongen, S., L. Lens, and G. Molenberghs. 1999. Mixture analysis of asymmetry: modelling directional asymmetry, antisymmetry and heterogeneity in fluctuating asymmetry. Ecology Letters **2**: 387-396.

Van Valen, L. 1962. A study of fluctuating asymmetry. Evolution **16**: 125-142.

Van Valen, L. 1978. The statistics of variation. Evolutionary Theory **4**: 33-43.

Vøllestad, L. A., K. Hindar, and A. P. Møller. 1999. A meta-analysis of fluctuating asymmetry in relation to heterozygosity. Heredity **83**: 206-218.

Whitlock, M. 1996. The heritability of fluctuating asymmetry and the genetic control of developmental stability. Proceedings of the Royal Society of London. Series B **263**: 849-853.

Whitlock, M. 1998. The repeatability of fluctuating asymmetry: A revision and extension. Proceedings of the Royal Society of London. Series B **265**: 1429-1431.

Wright, S. 1968. Evolution and the Genetics of Populations. Vol.1. Genetics and Biometrical Foundations. Univ. of Chicago Press, Chicago.

Yezerinac, S. M., S. C. Lougheed, and P. Handford. 1992. Morphological variability and enzyme heterozygosity: Individual and population level correlations. Evolution **46**: 1959-1964.

Zakharov, V. M. 1992. Population phenogenetics: Analysis of developmental stability in natural populations. Acta Zoologica Fennica **191**: 7-30.

Zar, J. H. 1999. Biostatistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Table 1. Conventional FA indexes for a sample of individuals based on one trait per individual, standardized so that numerical values of related indexes are directly comparable (modified from Table 1 of Palmer, 1994).

Trait-size correction	Measure of asymmetry for a given trait of individual i		
	Unsigned asymmetry $\lvert R_i-L_i\rvert$	Signed asymmetry† (R_i-L_i)	Ratio between sides $\ln(R_i/L_i)$§
none	**FA1:** mean $\lvert R-L\rvert$	**FA4a:** $0.798\sqrt{\text{var}(R-L)}$ **FA5a:** $0.798\sqrt{[\sum(R-L)^2/N]}$	
by individual	**FA2:** $\text{mean}\left[\frac{\lvert R-L\rvert}{(R+L)/2}\right]$	**FA6a:** $0.798\sqrt{\text{var}\left[\frac{(R-L)}{(R+L)/2}\right]}$	**FA8a:** mean $\lvert\ln(R/L)\rvert$
by sample	**FA3:** $\frac{\text{mean}\lvert R-L\rvert}{\text{mean}[(R+L)/2]}$	**FA7a:** $\frac{0.798\sqrt{\text{var}\,(R-L)}}{\text{mean}[(R+L)/2]}$	

Other indexes for single traits:

FA9: $1 - r^2$ of correlation between R and L (i.e., % bilateral variation not due to positive covariation); a potentially misleading index (Angus, 1982; Palmer, 1994).

FA10a: $0.798\sqrt{2\sigma^2_i}$, where $\sigma^2_i = (MS_{sj} - MS_m)/M$ = the estimated underlying DI variance of a given side of individual i, and where MS_{sj} = sides x individuals interaction MS, MS_m = measurement error MS, M= number of replicate measurements per side, from a sides x individuals ANOVA on *untransformed* replicate measurements of R and L (see Table 3 of Palmer & Strobeck, 1986). When the number of replicate measurements per side is two, this simplifies to: $0.798\sqrt{(MS_{sj} - MS_m)}$. Describes the magnitude of total non-directional asymmetry for a trait after ME has been partitioned out. For traits exhibiting ideal FA (Fig. 1a), it may be compared directly with FA1 to view the decline in FA1 after removing ME.

FA10b: $0.798\sqrt{2\sigma^2_i}$, where σ^2_i is computed as for FA10a, but the data analysed are *log transformed* replicate measurements ln(R) and ln(L). Describes the magnitude of total non-directional asymmetry *as a proportion of the trait mean* for a trait after ME has been partitioned out. For traits exhibiting ideal FA (Fig. 1a), it may be compared directly with FA2 to view the decline in FA2 after removing ME. Only recommended where size variation is small.

† See *Section IV.A2* for an explanation of how variances can be transformed into an estimate of average deviation.

§ See *Section IV.A7* for an explanation why FA8a and FA2 are equivalent to three decimal places.

Table 2. Indexes for individual FA based on multiple traits per individual.

Previous indexes (Palmer, 1994).

FA11: asymmetry in an individual (A_i)= $\sum|R_i-L_i|$ for all traits of an individual; the index for a sample is $\sum A_i$ /N where N= number of individuals in the sample.
CONS: only meaningful where mean (A_i) is comparable for all traits (Palmer, 1994).

FA12: a non-parametric index; asymmetry in an individual (A_i)= total number of asymmetrical traits in an individual, independent of how large the deviation is between sides; the index for a sample is $\sum A_i$ /N where N= number of individuals in the sample.
CONS: only meaningful for meristic traits (Palmer, 1994).

FA13: Generalized index of overall FA (GFA); a multivariate measure of average deviation from symmetry for multiple metrical traits (see Leung *et al*., 2000, for detailed explanation).
CONS: complex and difficult to apply.

New indexes

FA14: asymmetry in individual i is $\sum[|FA_{ij}| / \overline{|FA_j|}] / N_t$, where FA_{ij} is the deviation from symmetry of trait j in individual i, and $\overline{|FA_j|}$ is the average absolute deviation from symmetry of trait j for the entire sample (index CFA 2 of Leung *et al*., 2000).
PROS: removes size-dependent differences in FA among traits; removes among-trait differences in underlying DI; more powerful than FA15 where leptokurtosis is minor.
CONS: potentially yields biased values if ME is constant but trait size varies (e.g., see Fig. 2c); not comparable quantitatively to other studies, so it is more useful as a test of significance than for describing FA differences; less powerful than FA15 in the presence of moderate leptokurtosis (Leung *et al*., 2000);

FA15: a non-parametric index; asymmetry in individual i is $\sum RFA_{ij}$, where RFA_{ij} is the rank value of |R - L| for trait j of individual i and the |R - L| values are ranked separately for each trait in the sample (index CFA 3 of Leung *et al*., 2000).
PROS: removes size-dependent differences in FA among traits; removes among-trait differences in underlying DI; more powerful than FA15 in the presence of moderate leptokurtosis (Leung *et al*., 2000).
CONS: potentially yields biased values if ME is constant but trait size varies (e.g., see Fig. 2c); not comparable quantitatively among studies, so it is more useful for significance testing than

for describing FA differences; less powerful than FA14 where leptokurtosis is minor.

FA16: MANOVA on $|FA_{ij}|$ (index CFA 6 of Leung *et al*., 2000)

PROS: not vulnerable to departures from normality.

CONS: consistently lower power than related multivariate indexes (Leung *et al*., 2000).

FA17: $\sum|\ln(R_j/L_j)| / T = \sum|\ln(R_j) - \ln(L_j)| / T$, where R_j and L_j are measurements of the R and L side for trait j and T is the number of traits per individual.

PROS: expresses the average proportional deviation from symmetry of all traits of an individual combined; directly comparable with indexes based on single traits (FA2 and FA8a, Table 1).

CONS: yields biased values if ME is constant but trait size varies (e.g., see Fig. 2c).

FA18: landmark based index; $\sqrt{\sum (XY_{iR} - XY_{iL})^2}$ for i= 1 to k, the total number of landmarks per specimen, of Procrustes aligned landmarks from structures on the right (XY_{iR}) and left (XY_{iL}) sides of an individual (see Appendix of Klingenberg and McIntyre, 1998).

PROS & CONS: see *Section IV.B4*.

Table 3. Measurement error (ME) and repeatability in studies of FA.†

---

a) true underlying error in measurement

σ^2_{ME} = variance of repeat measurements of a single side due to ME. In the absence of DI (i.e., var(R - L) is due solely to ME) var(R - L) = 2 σ^2_{ME} / n (see Eq. III.7, Appendix III).

b) descriptors of ME that include units of measurement

ME1: average difference between pairs of measurements on one side, ME1 = $\sum|M_1 - M_2| / N$.
PROS: ME1 may be compared directly to FA1 (Table 1) when FA1 is computed using two measurements per side (i.e., ME1 = FA1 in the absence of DI; see Eq. III.8, Appendix III); provides an independent estimate of ME2 (ME2 = ME1 / 0.798; see *Section IV.A2*).
CONS: limited to pairs of repeat measurements.

ME2: SD of repeated measurements, ME2 = $\sqrt{[\sum var(M_1, M_2, M_3, \ldots M_n) / N]} = \sqrt{MS_m}$ where MS_m is the error MS from a sides x individuals ANOVA (Palmer, 1994).
PROS: estimates the true underlying ME (σ^2_{ME}); not limited to two measurements per trait; may be compared directly to FA1: in the absence of DI FA1 = $(0.798 \sqrt{(2 / n)}) * ME2 = 0.798 \sqrt{((2 / n) * MS_m)}$ (see Eq. III.8, Appendix III).
CONS: none.

c) descriptor of ME that is independent of units of measurement

ME3: %ME = 100*ME1 / FA1 = 100*MS_m / $MS_{interaction}$ where FA1 is measured as in Table 1.
PROS: easy to compute; easy to interpret.
CONS: cannot estimate true ME without knowledge of FA1 or $MS_{interaction}$.

d) repeatability of FA, independent of units of measurement

ME4: repeatability, $r_I = \dfrac{MS_{individuals} - MS_{error}}{MS_{individuals} + (n - 1)\, MS_{error}}$

where $MS_{individuals}$ is the among- and MS_{error} is the within-individual MS, from a one-way ANOVA (repeat measurements of R-L nested within individuals, Zar, 1999, p. 405).
PROS: a dimensionless number that estimates the true FA variation as a proportion of the total between-sides variation including ME; easy to interpret (ranges from -1 to +1).
CONS: requires another analysis in addition to the standard test of the significance of FA relative to ME; cannot be used to estimate ME with units (e.g., ME1) without knowing $MS_{individuals}$.

ME5: repeatability, $r_A = \dfrac{MS_{interaction} - MS_m}{MS_{interaction} + (n - 1)\, MS_m}$

PROS: a dimensionless number that estimates the true FA variation as a proportion of the total between-sides variation including ME; easy to interpret (ranges from -1 to +1); readily computed from MS obtained in the standard test of significance of FA relative to ME (Section V.A5, Palmer and Strobeck, 1986).

CONS: cannot be used to estimate ME with units (e.g., ME1) without knowing $MS_{individuals}$.

e) repeatability of DI among individuals ('hypothetical repeatability' of Van Dongen, 1998)

R: hypothetical repeatability = $1-(((V_{FA} + V_{ME}) \times (\pi-2)/\pi)/V_{|FA|})$, where V_{FA}= var(R-L) = FA4, $V_{|FA|}$= var|R-L|, and V_{ME}= $(ME2)^2 = MS_m$.

PROS: may potentially be used to correct for the downward bias of correlations with FA1 due to sampling error and measurement error (Whitlock, 1996).

CONS: derivation potentially flawed (*Section V.A4c*); the standard error of ***R*** is unknown so its ability to reveal true variation in DI among individuals is unclear.

---

† M_1, M_2, … M_n= repeat measurements on the same side in the same individual, n= number of repeat measurements, N= total number of objects measured (normally twice the number individuals since one object is measured on each side), MS= mean squares. $MS_{interaction}$ = sides x individuals interaction MS and MS_m= the error MS from the standard sides x individuals ANOVA used to test the significance of FA relative to ME (Section V.A5, Palmer and Strobeck, 1986).

Table 4. Underlying causes of departures from normality in studies of FA.

a) causes of skew

i) one or more individuals were damaged on the same side, yielding extreme values of (R-L).

ii) one or more individuals exhibit extreme values of (R-L) in the same direction because of measurement or recording errors.

iii) a mixture of individuals where some exhibit ideal FA (Fig. 1a) and others exhibit weak DA.

iv) a mixture of individuals where some exhibit DA and others exhibit antisymmetry (Palmer and Strobeck, 1992).

v) bimodal variation in R - L where the two modes are of different height; likely due to a mixture of antisymmetry and DA.

b) causes of leptokurtosis

i) outlier measurements or other causes of heterogeneity of ME (*Section V.A* and Steps 1 and 2, Appendix V).

ii) outlier values of (R-L) for a few individuals, due to wear, injury or some type of error (*Section V.A* and Steps 3-5, Appendix V).

iii) a mixture of individuals where some were measured with one level of ME and others were measured with another level of ME (e.g., due to changes in ME with experience, to session-to-session differences in ME, or to differences in ME among measurers).

iv) a mixture of individuals where some exhibit ideal FA (Fig. 1a) and others exhibit antisymmetry (Palmer and Strobeck, 1992).

v) heterogeneity of (R-L) variation within a sample due to size-dependence of (R-L) (*Section IV.A3d*).

vi) heterogeneity of (R-L) variation within a sample due to true variation in underlying DI among individuals (*Section IV.A3d*).

c) causes of platykurtosis

i) antisymmetry, consistent deviations of (R - L) from zero, but the side that is larger varies at random (Van Valen, 1962).

Table 5. Critical values of the kurtosis test statistic for deviations of frequency distributions from normality in the direction of platykurtosis (broad-peaked or bimodal) and leptokurtosis (narrow-peaked and long-tailed). Significant platykurtosis may signal the presence of antisymmetry.

	Critical values for Eq. 6†				Critical values for Eq. 7§			
	Platykurtosis		Leptokurtosis		Platykurtosis		Leptokurtosis	
Sample Size	5% level	1% level	5% level	1% level	5% level	1% level	5% level	1% level
7	-1.59	-1.75	0.55	1.23	-1.997	-2.395	3.109	4.710
8	-1.54	-1.69	0.70	1.53	-1.814	-2.132	2.899	4.617
9	-1.47	-1.65	0.86	1.82	-1.674	-2.030	2.829	4.639
10	-1.44[a]	-1.61[a]	0.95[a]	2.00[a]	-1.575	-1.881	2.624	4.480
12	-1.36	-1.54	1.05	2.20	-1.442	-1.720	2.416	4.248
15	-1.28	-1.45	1.13	2.30	-1.284	-1.563	2.152	3.973
20	-1.18[b]	-1.36[b]	1.18[b]	2.38[b]	-1.161	-1.403	1.869	3.471
25	-1.09	-1.28	1.15	2.29	-1.052	-1.288	1.735	3.196
30	-1.02	-1.21	1.12	2.20	-0.992	-1.220	1.549	2.862
35	-0.97	-1.16	1.09	2.12	-0.936	-1.147	1.440	2.651
40	-0.93	-1.11	1.06	2.04	-0.886	-1.098	1.333	2.512
45	-0.89	-1.07	1.02	1.96	-0.848	-1.049	1.301	2.313
50	-0.85	-1.05	1.00	1.88	-0.817	-1.016	1.217	2.268
60	-0.79	-0.97	0.94	1.75	-0.767	-0.954	1.132	2.005
70	-0.75	-0.93	0.89	1.64	-0.717	-0.901	1.029	1.804
80	-0.71	-0.88	0.85	1.54	-0.682	-0.870	0.983	1.744
90	-0.68	-0.84	0.81	1.46	-0.658	-0.821	0.916	1.611
100	-0.65	-0.82	0.78	1.39	-0.631	-0.803	0.869	1.510
120	-0.61	-0.78	0.75	1.26	-0.583	-0.742	0.791	1.386
130	-0.59	-0.74	0.70	1.21				
140	-0.57	-0.72	0.67	1.17	-0.554	-0.707	0.748	1.262
150	-0.55	-0.71	0.65	1.13				
160	-0.54	-0.68	0.63	1.09	-0.521	-0.674	0.692	1.178
180	-0.51	-0.65	0.60	1.03	-0.501	-0.640	0.643	1.102
200	-0.49	-0.63	0.57	0.98	-0.478	-0.620	0.617	1.020
250	-0.45	-0.58	0.52	0.87			0.560	0.909
300	-0.41	-0.54	0.47	0.79			0.510	0.819
400	-0.36	-0.48	0.41	0.67			0.439	0.694
500	-0.33	-0.43	0.37	0.60			0.391	0.610

† Critical values for sample sizes $\leq$ 200 for Eq. 6 were obtained from D'Agostino (1986, his Table 9.5), and for sample sizes $\geq$ 200 were obtained from Pearson and Hartley (1966, Table 34). All were obtained by subtracting 3 from the original values, to make them comparable to the skew statistic (see *Section V.B2c*).

[a] Confirmed to be within 1% using 50,000 replications.

[b] Confirmed to be within 1% using 20,000 replications.

§ Critical values obtained by simulation. Kurtosis was computed using Eq. 7 on 30,000 replicates of normal(0,1). For sample sizes $\geq$ 200 critical values were obtained from Table B.23 of Zar (1999), which are only valid for leptokurtosis.

Table 6. The structure of a hypothetical Levene's test for differences in FA among individuals and traits.†

	Trait 1	Trait 2 …	Trait k
Indiv. 1	$\lvert\ln(R_1) - \ln(L_1)\rvert$	$\lvert\ln(R_1) - \ln(L_1)\rvert$	etc.
	$\lvert\ln(R_2) - \ln(L_2)\rvert$	$\lvert\ln(R_2) - \ln(L_2)\rvert$	
	$\lvert\ln(R_i) - \ln(L_i)\rvert$	$\lvert\ln(R_i) - \ln(L_i)\rvert$	
Indiv. 2	$\lvert\ln(R_1) - \ln(L_1)\rvert$	$\lvert\ln(R_1) - \ln(L_1)\rvert$	
	$\lvert\ln(R_2) - \ln(L_2)\rvert$	$\lvert\ln(R_2) - \ln(L_2)\rvert$	
…	$\lvert\ln(R_i) - \ln(L_i)\rvert$	$\lvert\ln(R_i) - \ln(L_i)\rvert$	
Indiv. j	etc.		

† R_1, R_2, and R_i are replicate measurements of the right side and L_1, L_2, and L_i are replicate measurements of the left side of a single trait in an individual. i= total number of replicate measurements, j= total number of individuals, k= total number of traits.

Table 7. Outcome and interpretation of the hypothetical Levene's test for differences in FA among individuals and traits. This is a fully model II ANOVA, since both traits and individuals are random effects. If specific traits are selected *a priori* to test for different levels of DI, then traits may be considered a fixed effect, but the expected MS and therefore tests of significance change (Sokal and Rohlf, 1995; p. 333-334).†

Source of variation	Observed MS	Expected MS	Denominator MS for F test	Interpretation if significant
Individuals (I, random)	MS_I	$\sigma^2_e + n\,\sigma^2_{IT} + n\,\sigma^2_I$	MS_{IT}	FA_{rel} varies among individuals
Traits (T, random)	MS_T	$\sigma^2_e + n\,\sigma^2_{IT} + n\,\sigma^2_T$	MS_{IT}	FA_{rel} varies among traits
IxT Interaction	MS_{IT}	$\sigma^2_e + n\,\sigma^2_{IT}$	MS_{err}	Difference in FA_{rel} among traits depends on individual
Error (due to measurements)	MS_{err}	σ^2_e		

† Expected MS from Sokal (1995). n= number of replicate measurements. σ^2_e= measurement error variance, σ^2_{IT}= variance component due to interaction, σ^2_T= variance component due to traits, σ^2_I= variance component due to individuals, FA_{rel} = relative FA = FA as a proportion of trait size (FA8a of Table 1).

Table 8. The structure of a hypothetical Levene's test for differences in FA among traits and between two groups (e.g., sex), based on multiple individuals per group.†

	Trait 1	Trait 2 …	Trait k
Male	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$	etc.
	$\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$	$\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$	
	$\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$	$\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$	
	…	…	
	$\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	$\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	
Female	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$	
	$\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$	$\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$	
	$\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$	$\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$	
	…	…	
	$\lvert\ln(R_{if}) - \ln(L_{if})\rvert$	$\lvert\ln(R_{if}) - \ln(L_{if})\rvert$	

† R_1= average of all replicate measurements of the right side for individual 1, L_1= average of all replicate measurements of the left side for individual 1, etc. i= total number of individuals, k= total number of traits.

Table 9. Outcome and interpretation of the hypothetical two-way Levene's test for differences in FA between sexes and among traits. This is a mixed-model ANOVA, since sex is a fixed effect and traits is a random effect. If specific traits are selected *a priori* to test for different levels of DI, then traits may be considered a fixed effect, but the expected MS and therefore tests of significance change (Sokal and Rohlf, 1995; p. 333-334).†

Source of variation	Observed MS	Expected MS	Denominator MS for F test	Interpretation if significant
Sex (S, fixed)	MS_S	$\sigma^2_e + n\,\sigma^2_{ST} + S^*$	MS_{ST}	FA_{rel} differs between sexes
Traits (T, random)	MS_T	$\sigma^2_e + na\,\sigma^2_T$	MS_{err}	FA_{rel} varies among traits
SxT Interaction	MS_{ST}	$\sigma^2_e + n\,\sigma^2_{ST}$	MS_{err}	Difference in FA_{rel} between sexes depends on trait
Error (due to individuals)	MS_{err}	σ^2_e		

† Expected MS from Sokal and Rohlf (1995), where σ^2_e= residual variation among individuals, σ^2_{ST}= variance component due to interaction, σ^2_T= variance component due to traits, S*= variance component due to sex, n= number of individuals per sex, a= number of sexes. See footnote to Table 7 for remaining terms.

Table 10. The structure of a hypothetical Levene's test for differences in FA among traits, between two groups (e.g., sex), and between two habitats (e.g., high stress, low stress), based on multiple individuals per group.†

	Trait 1		Trait 2		… Trait k
	High stress	Low stress	High stress	Low stress	
Male	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$ $\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$ $\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$ … $\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$ $\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$ $\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$ … $\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$ $\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$ $\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$ … $\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	$\lvert\ln(R_{1m}) - \ln(L_{1m})\rvert$ $\lvert\ln(R_{2m}) - \ln(L_{2m})\rvert$ $\lvert\ln(R_{3m}) - \ln(L_{3m})\rvert$ … $\lvert\ln(R_{im}) - \ln(L_{im})\rvert$	etc.
Female	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$ $\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$ $\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$ … $\lvert\ln(R_{jf}) - \ln(L_{jf})\rvert$	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$ $\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$ $\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$ … $\lvert\ln(R_{jf}) - \ln(L_{jf})\rvert$	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$ $\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$ $\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$ … $\lvert\ln(R_{jf}) - \ln(L_{jf})\rvert$	$\lvert\ln(R_{1f}) - \ln(L_{1f})\rvert$ $\lvert\ln(R_{2f}) - \ln(L_{2f})\rvert$ $\lvert\ln(R_{3f}) - \ln(L_{3f})\rvert$ … $\lvert\ln(R_{jf}) - \ln(L_{jf})\rvert$	

† R_{1m}= average of all replicate measurements of the right side for male individual 1, L_{1m}= average of all replicate measurements of the left side for male individual 1, R_{1f}= average of all replicate measurements of the right side for female individual 1, L_{1f}= average of all replicate measurements of the left side for female individual 1, etc. i= total number of males; j= total number of females, k= total number of traits.

Table 11. Outcome and interpretation of the hypothetical three-way Levene's test for differences in FA between sexes (male, female), between habitats (high stress, low stress), and among traits. This is a mixed model ANOVA, since sex and habitat are fixed effects and traits is a random effect. If specific traits are expected to show different levels of DI, then traits may be considered a fixed effect also, but see Sokal & Rohlf (1995; p. 376-377) for expected MS and proper tests.†

Source of variation	Observed MS	Expected MS	Denominator MS for F test	Interpretation if significant
Sex (S, fixed)	MS_S	$\sigma^2_e + nb\ \sigma^2_{ST} + S^*$	MS_{ST}	FA_{rel} differs between sexes
Habitat (H, fixed)	MS_H	$\sigma^2_e + na\ \sigma^2_{HT} + H^*$	MS_{HT}	FA_{rel} differs between habitats
Traits (T, random)	MS_T	$\sigma^2_e + nab\ \sigma^2_T$	MS_{err}	FA_{rel} varies among traits
SxH Interaction	MS_{SH}	$\sigma^2_e + n\ \sigma^2_{SHT} + SH^*$	MS_{SHT}	FA_{rel} difference between sexes depends on habitat
SxT Interaction	MS_{ST}	$\sigma^2_e + nb\ \sigma^2_{ST}$	MS_{err}	FA_{rel} difference between sexes depends on trait
HxT Interaction	MS_{HT}	$\sigma^2_e + na\ \sigma^2_{HT}$	MS_{err}	FA_{rel} difference between habitats depends on trait
SxHxT Interaction	MS_{SHT}	$\sigma^2_e + n\ \sigma^2_{SHT}$	MS_{err}	One 2-way interaction depends on the state off the third factor
Error (due to individuals)	MS_{err}	σ^2_e		

† Expected MS from Sokal and Rohlf (1995), where σ^2_e= residual variation among individuals, S*, H*, and SH* are the variance components due to Sex and Habitat, and Sex x Habitat interaction (including their df), σ^2_{ST}, σ^2_{HT}, and σ^2_{SHT} are the variance components of the two- and three-way interactions, n= number of individuals per sex, a= number of sexes (2), b= number of traits, and FA_{rel} = relative FA = FA as a proportion of trait size (FA8a of Table 1).

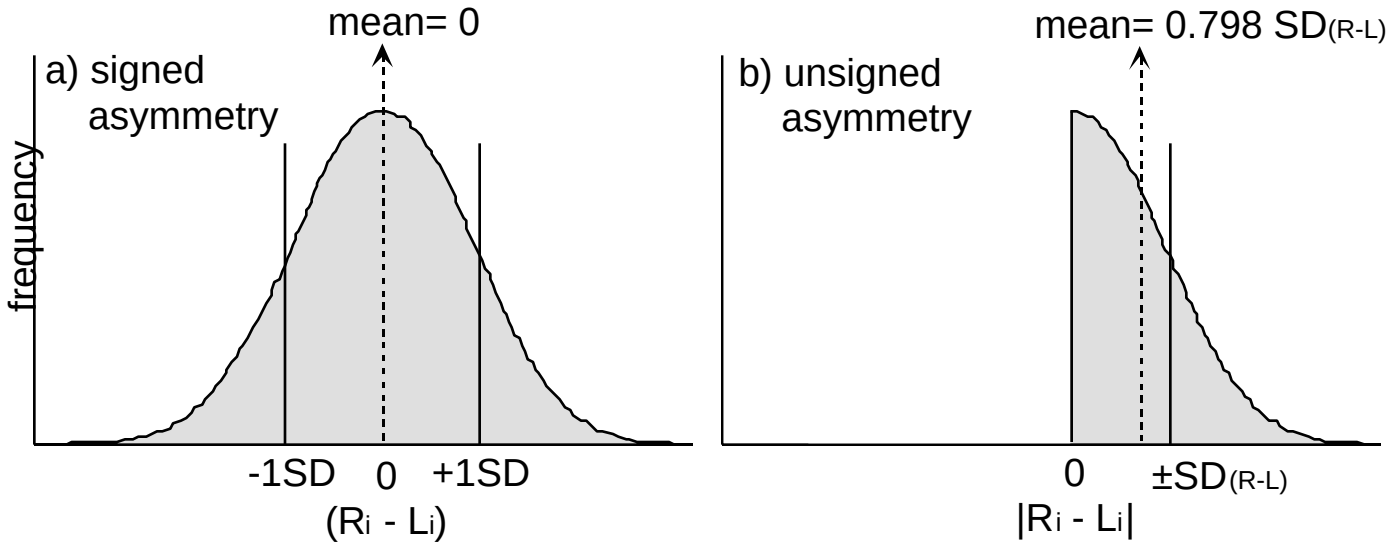

Figure 1. Hypothetical frequency distributions of a) signed (R - L) and b) unsigned |R - L| departures from symmetry for a trait that exhibits ideal FA (mean zero, normal). SD= standard deviation.

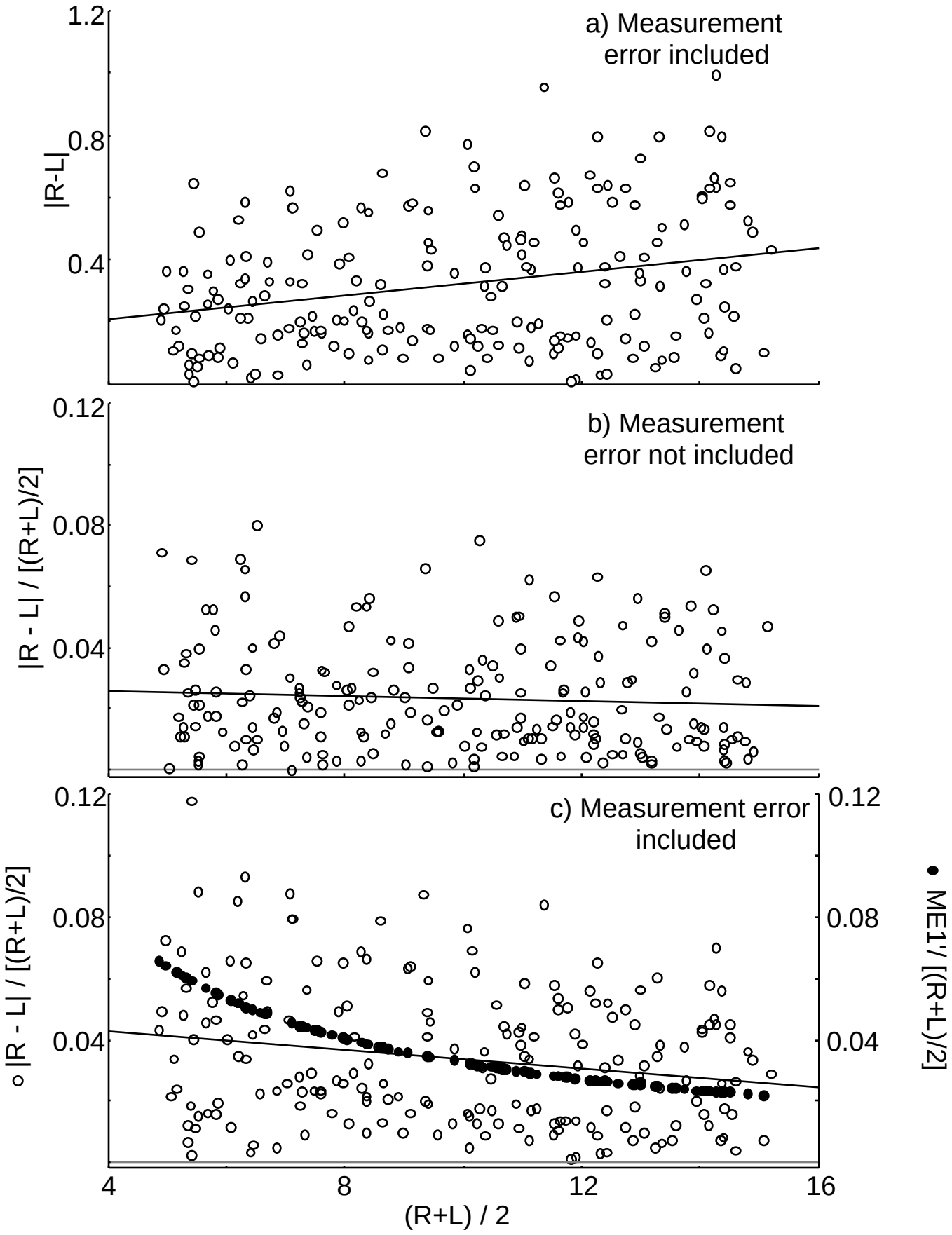
a) Measurement error included
|R-L|
1.2
0.8
0.4
b) Measurement error not included
|R - L| / [(R+L)/2]
0.12
0.08
0.04
c) Measurement error included
o |R - L| / [(R+L)/2]
• ME1' / [(R+L)/2]
0.12
0.08
0.04
4
8
12
16
(R+L) / 2

Figure 2. Simulated variation illustrating the dependence of asymmetry on trait size, (R + L)/2, for three cases: a) unscaled asymmetry with ME included (Spearman ρ= 0.228, P= 0.0013), b) size-scaled asymmetry, ME not included (Spearman ρ= –0.078, P= 0.27), c) size-scaled asymmetry, ME included (Spearman ρ= –0.174, P= 0.014). Solid lines indicate least-square linear regressions. In this simulation, both the underlying DI variance and the variance due to ME were set to 2% of trait size. Variation for the right side was simulated as $R_i = S' + S_i + DI_i + ME_i$ or $R_i = S' + S_i + DI_i$ depending whether or not ME was included (a) or not (b): trait size variation (S_i) = U_i * S', where U_i = UniformRandom(-0.5, 0.5) and S' = 10; size-dependent developmental instability (DI_i)= 0.02 * S_i * d_i, where d_i = RandomNormal(0,1); constant measurement error (ME_i)= 0.02 * S' * e_i, where e_i = RandomNormal(0, 1). Variation for the left side was also simulated this way, but with independent draws of DI_i and ME_i. Solid symbols in (c) indicate the expected |R - L| due solely to ME if only one measurement was taken per side ($ME1' = 0.798\ \sigma_{ME}\ \sqrt{2} = 0.798 * 0.2\ \sqrt{2} = 0.226$; see Appendix III.a for derivation) divided by trait size, $(R_i + L_i)/2$.

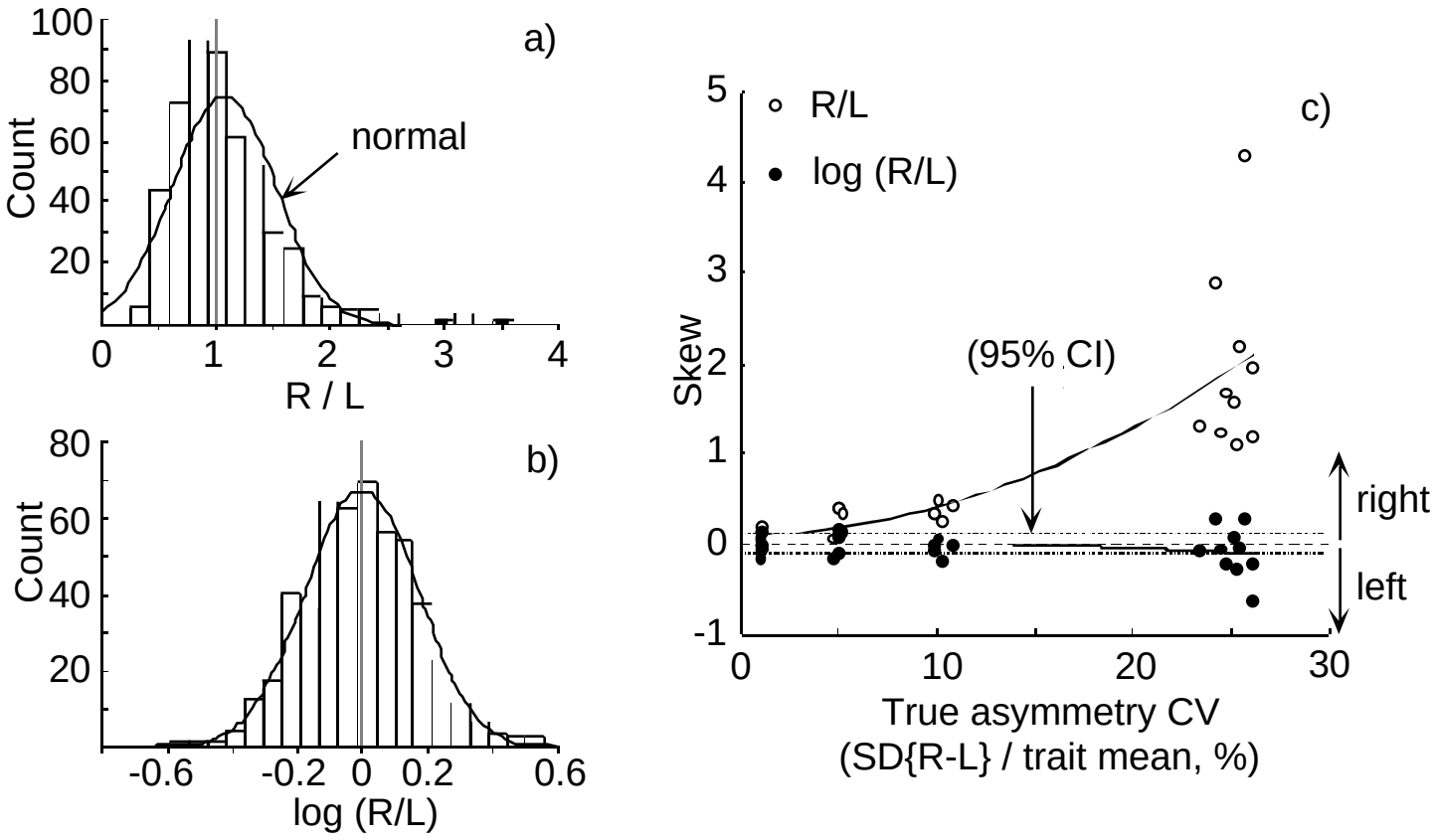

Figure 3. (a,b) Frequency distributions of the ratio (R/L) and of log (R/L) obtained from computer simulations [both right (R) and left (L) were normally distributed random deviates, mean= 10, SD= 2.5, N= 500]. (c) Effect of increasing asymmetry variation on the skew of ratios (for each point, N= 500, both R and L were normally distributed random deviates, mean= 10, SD= 0.1, 0.5, 1.0, 2.5; asymmetry CV refers to 100 (SD_{R-L}/trait mean); 10 simulations were conducted for each SD; solid lines indicate least-squares first-order polynomial regressions; small dashed lines indicate 5% significance levels for skew based on a sample size of N= 500.

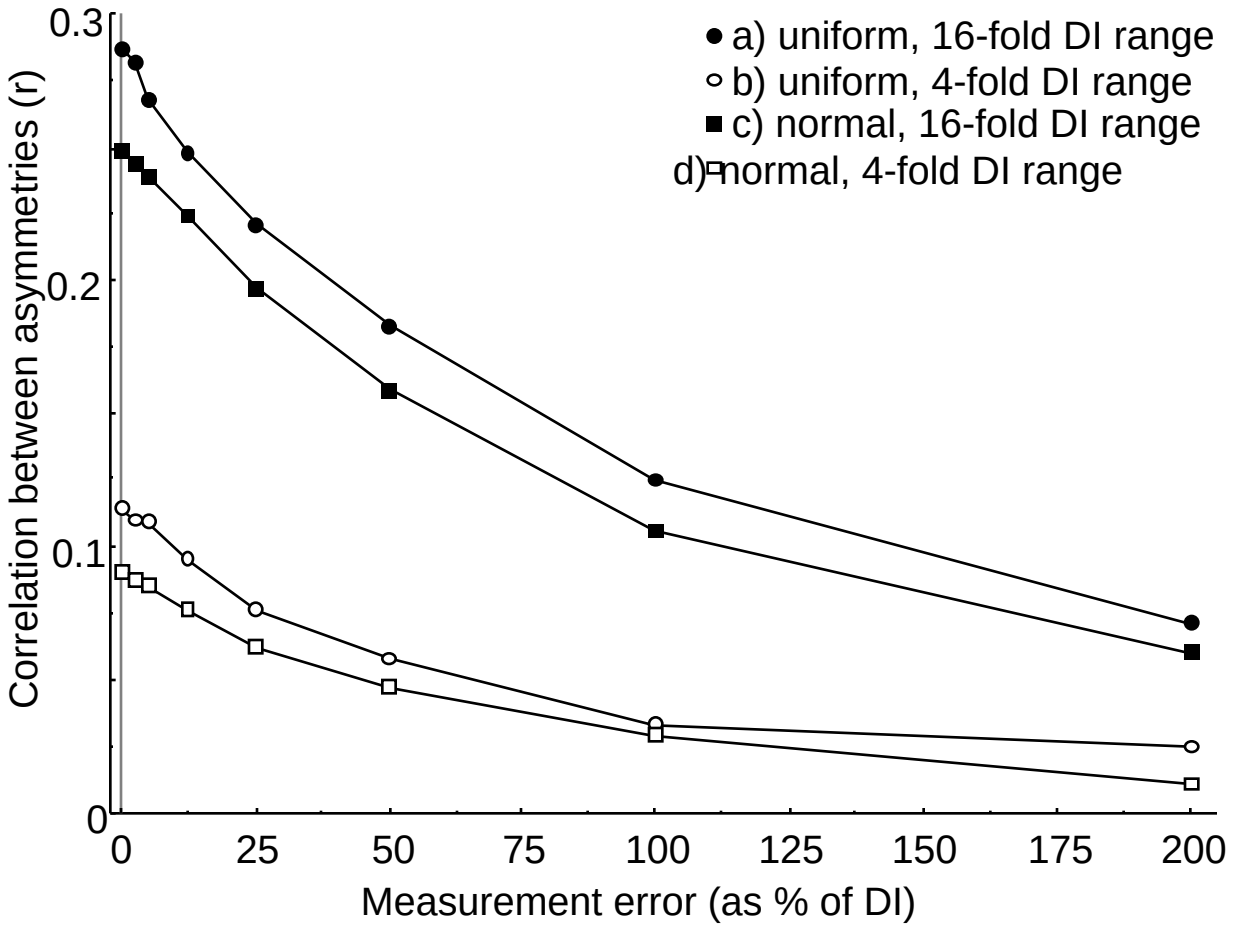

Figure 4. Effect of measurement error (ME) on the strength of the correlation among individuals in a single sample between |R-L| in one trait and |R-L| of a second trait in the same individual. Each population consisted of a mixture of individuals exhibiting three different levels of underlying DI variance: expected var(R - L) = DI = 1/x, 1, x. Two populations were simulated, as were two different distributions of DI variation: (●) x= 4, proportions of all three DI levels equal, (○) x= 2, proportions of all three DI levels equal, (■) x= 4, proportions of DI levels 1:2:1, (□) x= 2, proportions of DI levels 1:2:1. The ME variance var(M_1 - M_2) is expressed as a percent of the median DI variance (i.e., a value of 100 means the variance of replicate measurements equals the median DI variance between sides). The simulations were conducted by S. Van Dongen using the model described in Van Dongen (1998) (figure modified from Palmer 2000).

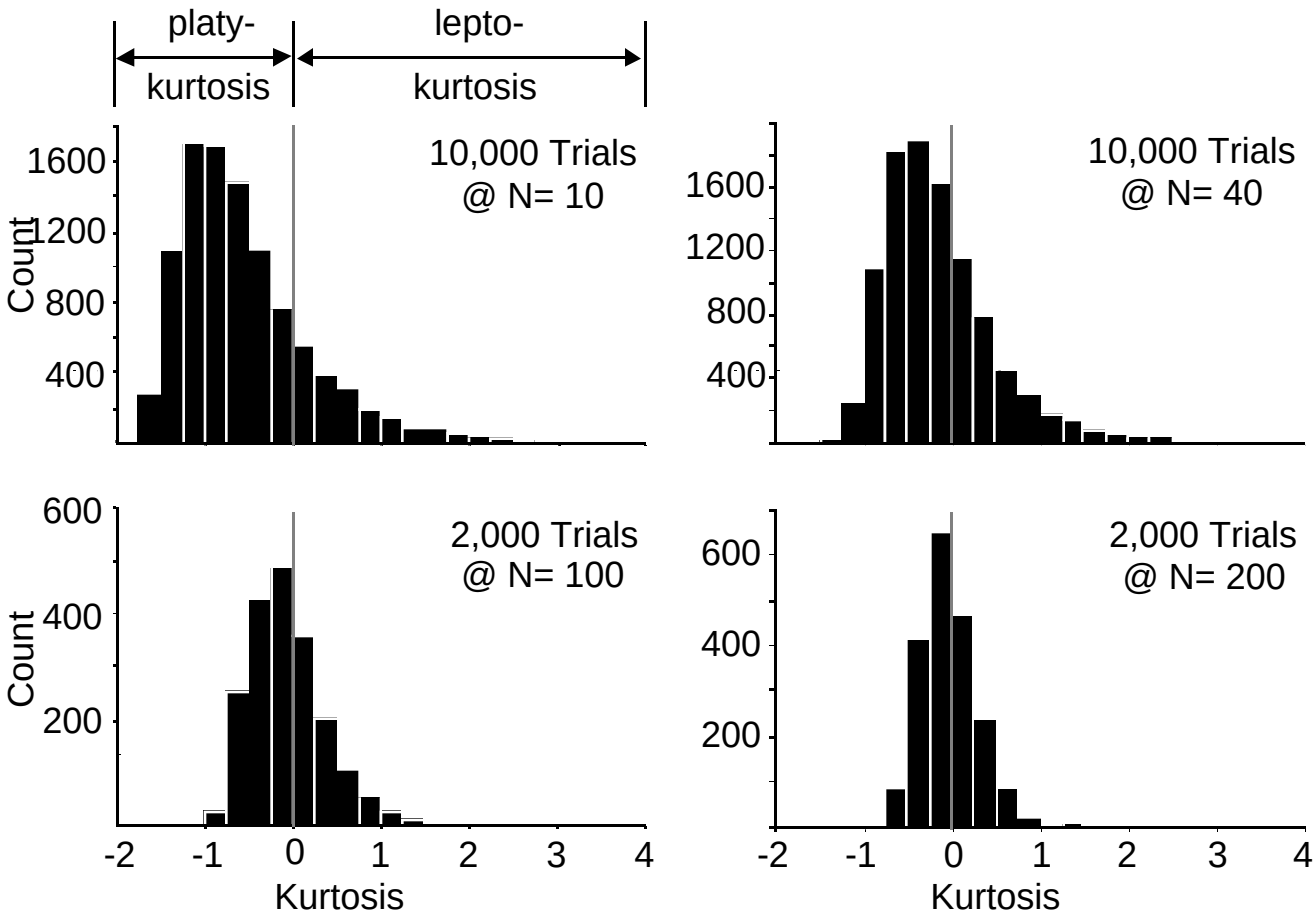

Figure 5. Frequency distributions of the kurtosis statistic (computed from Eq. 6) as a function of sample size. For each trial kurtosis was computed for a distribution of random normal deviates (mean= 0, SD= 1).

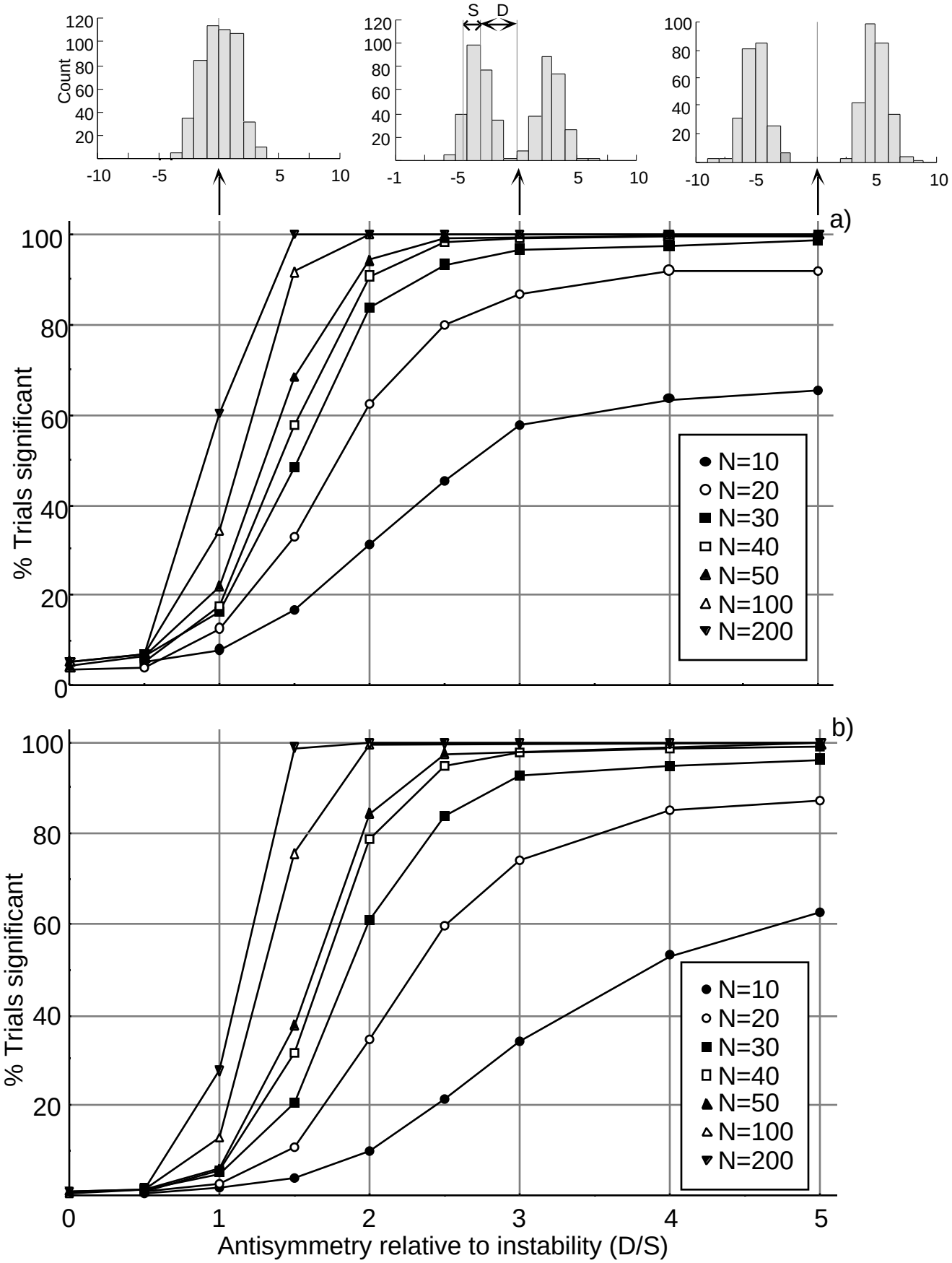
S
D
Count
a)
b)
% Trials significant
Antisymmetry relative to instability (D/S)
N=10
N=20
N=30
N=40
N=50
N=100
N=200

Figure 6. Power curves for the kurtosis statistic as bimodality increases: a) $\alpha = 5\%$ significance, b) $\alpha = 1\%$ significance. Kurtosis was computed using Eq. 6. Critical values for kurtosis were obtained from Table 5. Antisymmetry was simulated by varying the value for D (the distance between one peak and zero, middle panel of frequency distributions at top). S (the standard deviation of the variation about each peak, middle panel of frequency distributions at top), was held constant at 1.0. Frequency distributions at the top illustrate a single sample of N= 500 simulated observations.

APPENDIX I
Relations among FA indexes that scale out trait size

Several indexes express subtle asymmetry as a proportion of trait size in an individual (Palmer and Strobeck, 1986):

$$d_1 = (R - L) / ((R + L) / 2), \tag{I.1}$$

$$d_2 = |d_1| = |R - L| / ((R + L) / 2). \tag{I.2}$$

$$d_3 = \ln (R / L) = \ln (R) - \ln (L). \tag{I.3}$$

$$d_4 = |d_3| = |\ln (R / L)| = |\ln (R) - \ln (L)| \tag{I.4}$$

d_1 is used to compute index FA6, d_2 is used to compute index FA2, d_3 is used to compute index FA8, and d_4 is used to compute index FA8a (Palmer and Strobeck, 1986, and Table 1).

The relations between d_1 and d_2, and between d_3 and d_4, are obvious. The relations between d_1 and d_3, and between d_2 and d_4, are not, but these indexes can be shown to be equivalent, for all practical purposes, via a Taylor expansion series approximation.

First, consider an approximation to the natural log of one particular ratio:

$$\ln\left(\frac{1+x}{1-x}\right) = 2\left[x + \frac{x^3}{3} + \frac{x^5}{5} + ...\right] \tag{I.5}$$

This ratio can be shown to be equivalent to $\ln\left(\frac{R}{L}\right)$ as follows. First,

$$\ln\left(\frac{R}{L}\right) = \ln\left(\frac{\frac{R}{(R+L)/2}}{\frac{L}{(R+L)/2}}\right) = \ln\left(\frac{1+\frac{R}{(R+L)/2}-\frac{(R+L)/2}{(R+L)/2}}{1+\frac{L}{(R+L)/2}-\frac{(R+L)/2}{(R+L)/2}}\right) = \ln\left(\frac{1+\frac{R-L}{R+L}}{1-\frac{R-L}{R+L}}\right) = \ln\left(\frac{1+x}{1-x}\right) \tag{I.6}$$

where $x = \frac{R-L}{R+L}$

Second, substituting $\ln\left(\frac{R}{L}\right)$ for $\ln\left(\frac{1+x}{1-x}\right)$, and $\frac{R-L}{R+L}$ for x, in equation (I.5) yields:

$$\ln\left(\frac{R}{L}\right) = 2\left[\frac{R-L}{R+L} + \frac{\left(\frac{R-L}{R+L}\right)^3}{3} + \ldots\right] = \frac{R-L}{(R+L)/2} + \frac{\left(\frac{R-L}{(R+L)/2}\right)^3}{12} + \ldots. \quad \text{(I.7)}$$

Substituting from equation (I.1 and I.3) yields:

$$d_3 = d_1 + d_1^3/12 + d_1^5/80 + \ldots \quad \text{(I.8)}$$

Significantly, for studies of FA variation, the second and all subsequent terms in this series can be ignored because d_1 is almost always less than 0.1 and typically closer to 0.01 (Palmer, 1996). So even if deviations from symmetry approach 10% of trait size ($d_1 = 0.1$), the second term in this series would be less than 0.0001 and all higher order terms would be even smaller. Therefore, to at least three decimal places, $d_1 = d_3$ and $d_2 = d_4$.

LITERATURE CITED (APPENDIX I)

Palmer, A. R. 1996. Waltzing with asymmetry. BioScience **46**: 518-532.

Palmer, A. R., and C. Strobeck. 1986. Fluctuating asymmetry: measurement, analysis, patterns. Annual Review of Ecology and Systematics **17**: 391-421.

APPENDIX II
Expected size-dependence of ME for size-scaled FA indexes

If measurement error (ME) is constant, but trait size varies, size-scaled measures of FA (e.g., FA2, FA3, FA6a, FA7a, FA8a) will yield a negative association between apparent FA and trait size (*Section IV.A6*). The expected slope of FA2 vs trait size, due simply to ME where only a single measurement is taken per side, can be predicted as follows.

Definitions

μ = overall mean trait size, to which ME is proportional.

$\mu+x$ = size of a trait in an individual, to which DI is proportional (x refers to the deviation of a trait in an individual from the population mean.

$b\mu$ = standard deviation of repeat measurements of one side (b expresses ME as a proportion of overall mean trait size, μ)

$a(\mu+x)$ = standard deviation of one side due to DI (a expresses DI as a proportion of individual trait size, $\mu+x$, therefore FA= SD(R - L) = $a(\mu+x)\sqrt{2}$ for one measurement per side; see Appendix III.a for derivation),

$k = 0.798 = \sqrt{(2/\pi)}$ = the constant to convert SD(R - L) to mean|R - L| (Kendall and Stuart, 1951).

Derivation

$$\mathrm{FA2} = \frac{|R-L|}{(R+L)/2} = \frac{k\sqrt{2a^2(\mu+x)^2+2b^2\mu^2}}{\mu+x} = k\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2} \qquad \text{(II.1)}$$

The derivative of FA2 relative to x is

$$\frac{d\,\frac{|R-L|}{(R+L)/2}}{dx} = \frac{dk\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2}}{dx} = \frac{k\,\frac{1}{2}}{\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2}}\,\frac{d2b^2\left(\frac{\mu}{\mu+x}\right)^2}{dx} =$$

$$\frac{k\,\frac{1}{2}}{\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2}}\,2b^2\left(\frac{2\mu}{\mu+x}\right)\frac{d\left(\frac{\mu}{\mu+x}\right)}{dx} = \frac{2kb^2\left(\frac{\mu}{\mu+x}\right)}{\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2}}\,\frac{(-1)\mu}{(\mu+x)^2} =$$

$$\frac{-2kb^2\left(\frac{\mu}{\mu+x}\right)^2}{(\mu+x)\sqrt{2a^2+2b^2\left(\frac{\mu}{\mu+x}\right)^2}} \qquad \text{(II.2)}$$

Evaluated at x=0 (i.e., at the overall mean trait size μ), the expected slope of FA2 versus trait size is

$$\left.\frac{d\dfrac{|R-L|}{(R+L)/2}}{dx}\right|_{x=0} = \frac{-2kb^2}{u\sqrt{2a^2+2b^2}} = \frac{-2kb^2u^2}{u^2\sqrt{2a^2u^2+2b^2u^2}} \qquad \text{(II.3)}$$

If DI is absent (i.e., the difference between sides is due exclusively to ME) then Eq. II.3 simplifies to

$$\left.\frac{d\dfrac{|R-L|}{(R+L)/2}}{dx}\right|_{x=0} = \frac{-k\sqrt{2b^2u^2}}{u^2} = \frac{-k\sqrt{2}bu}{u^2} = \frac{-\text{FA1}}{u^2} \qquad \text{(II.4)}$$

where FA1 is defined as in Table 1. For n measurements per side (see derivation in Appendix III), the expected slope of FA2 vs trait size at $x = 0$ is

$$\left.\frac{d\dfrac{|R-L|}{(R+L)/2}}{dx}\right|_{x=0} = \frac{-k\sqrt{(2/n)}bu}{u^2} = \frac{-\text{FA1}}{u^2} \qquad \text{(II.5)}$$

where FA1 is defined as in Table 1 and the difference between sides is due exclusively to ME.

Eq. II.4 may also obtained as follows. The derivative

$$\frac{d(Cx^n)}{dx} = nCx^{n-1}$$

therefore

$$\frac{d(C/x)}{dx} = \frac{d(Cx^{-1})}{dx} = -Cx^{-2} = -C/x^2 \qquad \text{(II.6)}$$

where C is any constant. For studies of FA, C would refer to FA1 when DI variation was absent (see derivation in Appendix III) and x to trait size (R + L) / 2.

APPENDIX III
Expected contribution of ME to FA

Measurement error (ME) inflates all descriptors of FA except those that partition out ME (*Sections IV.A1, V.A*). Discussions of ME can be quite confusing if underlying error *variances* such as σ^2_{ME} below are not distinguished from numerical *descriptors* of ME, like ME1 (Table 3). Therefore when ME is discussed in general, it refers to σ^2_{ME}. Specific descriptors of ME are referred to using the convention in Table 3.

Definitions

μ = overall mean trait size.

$\varepsilon_R, \varepsilon_L$ = deviations of the size of the right and left side respectively from the mean trait size μ due to DI [from a normal distribution $(0, \sigma^2_I)$]; note that all variation in trait size in the derivations below is due solely to DI (i.e., underlying body size variation in absent).

$\delta_{R1}, \delta_{R2}, \delta_{L1}, \delta_{L2}$ = deviations of measurements 1 and 2 from the right and left sides ($\mu + \varepsilon_R$, $\mu + \varepsilon_L$, respectively) [from a normal distribution $(0, \sigma^2_{ME})$].

σ^2_I = variance of trait size on one side among individuals due to DI.

σ^2_{ME} = variance of replicate measurements of a single side due to ME.

$M_{1R}, M_{2R}, M_{1L}, M_{2L}$, etc. = actual first, second, etc. measurements of the right and left sides.

Preliminaries

First, recall the relationships between sums and differences of variances:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{covar}(XY) \quad \text{(III.1)}$$

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\,\text{covar}(XY)$$

Where the covariance is zero and the expected means are identical — as would be expected between independent replicate measurements or between sides that differ only due to DI — the term 2 covar(XY) disappears yielding:

$$\text{var}(X + Y) = \text{var}(X - Y) = \text{var}(X) + \text{var}(Y) \quad \text{(III.2)}$$

Second, recall that:

$$\text{var}(aX) = a^2\text{var}(X) \text{ and therefore } \text{var}(X / a) = \text{var}(X) / a^2 \quad \text{(III.3)}$$

Derivations

The amount that ME inflates FA may be computed as follows.

a) For one measurement per side:

$$(R_i - L_i) = M_{1R} - M_{1L}$$
$$= (\mu + \varepsilon_R + \delta_{R1}) - (\mu + \varepsilon_L + \delta_{L1})$$
$$= (\varepsilon_R - \varepsilon_L) + (\delta_{R1} - \delta_{L1})$$

Because the DI variances and the ME variances are the same for the right and left sides:

$$\text{var}(R - L) = (\sigma^2_I + \sigma^2_I) + (\sigma^2_{ME} + \sigma^2_{ME})$$
$$= 2\sigma^2_I + 2\sigma^2_{ME} \quad \text{(III.4)}$$

In the absence of DI (i.e., $\sigma^2_I = 0$ so any difference between sides is due solely to ME), this simplifies to:

$$\text{var}(R - L) = 2\sigma^2_{ME}$$
$$\text{SD}(R - L) = \sqrt{2}\sigma_{ME}$$
$$\text{FA1} = \text{mean } |R - L| = 0.798\sqrt{2}\,\sigma_{ME} = 2\sigma_{ME} / \sqrt{\pi} \quad \text{(III.5)}$$

b) For multiple (n) measurement per side:

$$(R_i - L_i) = (M_{1R} + M_{2R} + \ldots + Mn_R) / n - (M_{1L} + M_{2L} + \ldots + Mn_L) / n$$
$$= (\mu + \varepsilon_R + \delta_{R1} + \mu + \varepsilon_R + \delta_{R2} + \ldots + \mu + \varepsilon_R + \delta_{Rn}) / n$$
$$- (\mu + \varepsilon_L + \delta_{L1} + \mu + \varepsilon_L + \delta_{L2} + \ldots + \mu + \varepsilon_L + \delta_{Ln}) / n$$
$$= (\varepsilon_R - \varepsilon_L) + (\delta_{R1} + \delta_{R2} + \ldots + \delta_{Rn} + \delta_{L1} + \delta_{L2} + \ldots + \delta_{Ln}) / n$$

Because the DI variances and the ME variances are the same for the right and left sides:

$$\text{var}(R - L) = (\sigma^2_I + \sigma^2_I) + 2\,(n\sigma^2_{ME}) / n^2$$
$$= 2\,\sigma^2_I + (2n / n^2)\,\sigma^2_{ME}$$
$$= 2\,(\sigma^2_I + \sigma^2_{ME} / n) \quad \text{(III.6)}$$

In the absence of DI (i.e., $\sigma^2_I = 0$ so any difference between sides is due solely to ME), this simplifies to:

$$\text{var}(R - L) = 2\,\sigma^2_{ME} / n \quad \text{(III.7)}$$

$$\text{SD}(R - L) = \sigma_{ME}\,\sqrt{(2 / n)}$$

$$\text{FA1} = \text{mean}\,|R - L| = 0.798\,\sigma_{ME}\,\sqrt{(2 / n)} = 2\sigma_{ME} / \sqrt{(n\pi)} \quad \text{(III.8)}$$

Equation III.6 has the desirable property that as $n \rightarrow \infty$ $\text{var}(R - L) \rightarrow 2\,\sigma^2_I$, the observable FA due exclusively to DI. Similarly, Equation III.7 has the desirable property that as $n \rightarrow \infty$ FA1 -> 0.

APPENDIX IV
Relation between FA2 and FA3

FA2 and FA3 both describe FA as a proportion of trait size (Table 1). In FA2, trait asymmetry in each individual is standardized by the trait size of that individual. In FA3, the average trait asymmetry of the entire sample is standardized by the average trait size of the entire sample. How different are these indices?

Simulation

Trait size and FA variation were simulated as follows:

μ = overall mean trait size (set to a constant of 10).

x_i= trait size variation (UniformRandom (-s,s), where s was 0.1, 0.3, 0.5, 0.7, or 0.9).

ε_{Ri}, ε_{Li} = variation due to DI [RandomNormal(0, DI), where DI was 0.01, 0.02, 0.05; because μ= 10, this means the SD of DI was 1%, 2% or 5% of the overall trait mean].

When DI was independent of trait size, right and left were simulated as:

$R_i = \mu + \mu x_i + \mu\varepsilon_{Ri}$ and $L_i = \mu + \mu x_i + \mu\varepsilon_{Li}$

When DI was proportional to trait size, right and left were simulated as:

$R_i = \mu + \mu x_i + \mu x_i\varepsilon_{Ri}$ and $L_i = \mu + \mu x_i + \mu x_i\varepsilon_{Li}$

FA2 and FA3 were computed as in Table 1. The CV of trait size was computed as SD[$(R_i+L_i)/2$] / mean [$(R_i+L_i)/2$], but note that trait size exhibited a uniform distribution. For each trial, N= 10,000.

Results

When DI was proportional to trait size, FA2 and FA3 yielded the same values, regardless of the size range or the level of variation due to DI (Figure IV.1). However, when DI was constant, FA3 underestimated FA2. The amount of this underestimate depended on the trait size range, but not on the level of variation due to DI. If trait size CV was < 20%, FA3 deviated from FA2 by less

than 5%. However, for a trait size CV of 40%, FA3 deviated from FA2 by nearly 20%, and this deviation became more pronounced with increasing trait size CV.

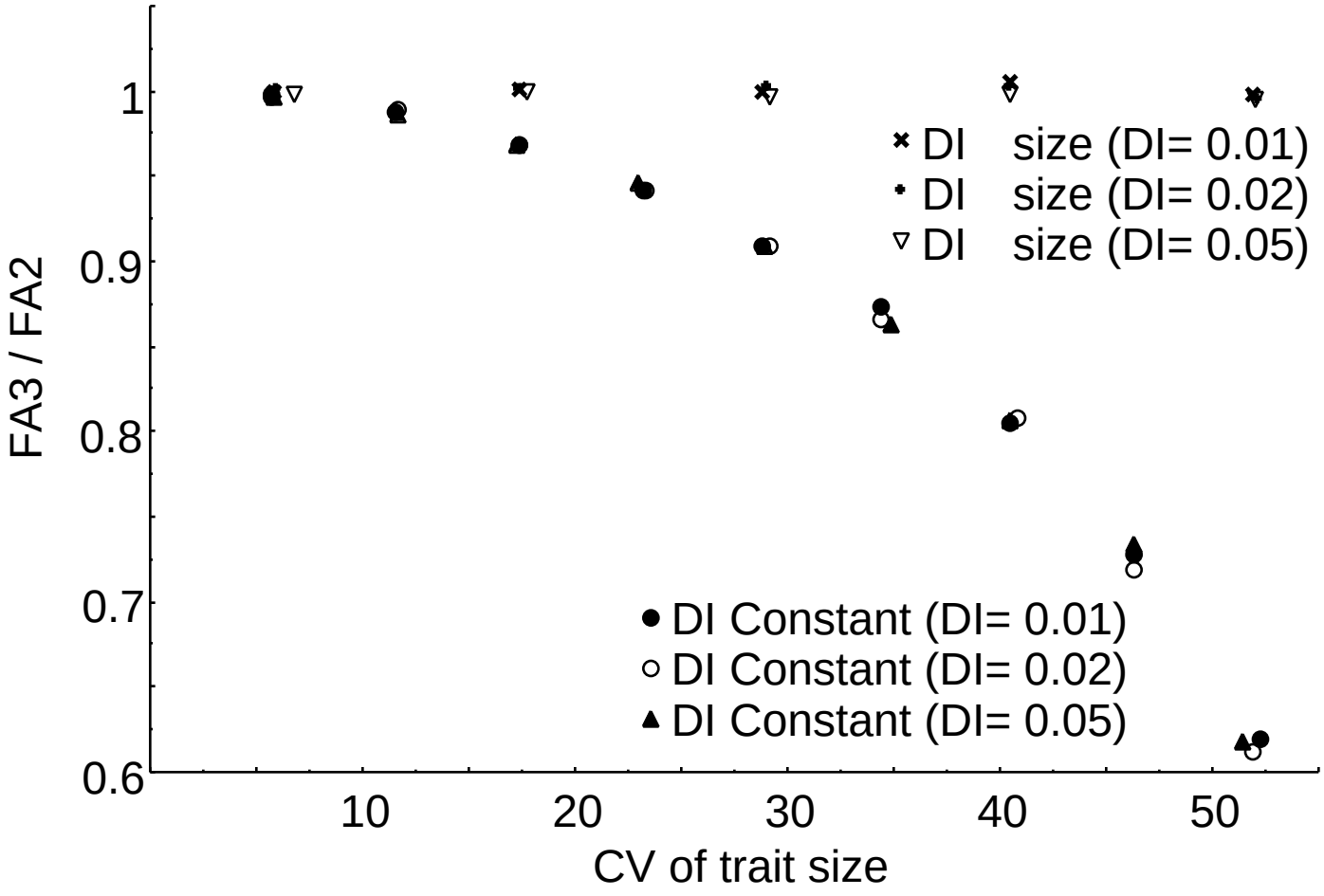

Figure IV.1. The ratio of two size-scaled FA indices (FA3/FA2) as a function of trait size variation for three levels of DI (SD of DI as a proportion of the overall trait mean). Each point was obtained from a simulated sample size of 10,000.

APPENDIX V
Fluctuating-asymmetry analysis: A step-by-step example

This appendix and its associated data files are available as web supplements from:

http://www.oup-usa.org/sc/0195143450

and

http://www.biology.ualberta.ca/palmer.hp/DataFiles.htm.