

How to compare fluctuating asymmetry of different traits

J. J. WINDIG & S. NYLIN

Department of Zoology, University of Stockholm, 106 91 Stockholm, Sweden

Keywords:

fluctuating asymmetry;
Monte Carlo simulations;
statistical tests.

Abstract

Comparing fluctuating asymmetry (FA) between different traits can be difficult because traits vary at different scales. FA is generally quantified either as the variance of the difference between left and right (σ_{L-R}^2) or the mean of the absolute value of this difference ($\mu_{|R-L|}$). Corrections for scale differences are obtained by dividing by trait size mean. We show that a third index, one minus the correlation coefficient between left and right ($1 - r_{L,R}$), is equivalent to σ_{L-R}^2 standardized by trait size variance. The indices are compared with Monte-Carlo simulations. All achieve the expected correction for scale differences. High type I error rates (false indication of differences) occur only for σ_{L-R}^2 and $\mu_{|R-L|}$ if trait sizes close to or below 0 occur. $1 - r_{L,R}$ with a bootstrap test has always low error rates. Recommendation of the index to be used should be based on whether standardization of FA by trait size mean or trait size variance is preferred. A survey of 36 traits in the Speckled Wood Butterfly (*Pararge aegeria*) indicated that σ_{L-R}^2 is slightly higher correlated to trait size variance than to trait size mean. Thus $1 - r_{L,R}$ seems to be the superior index and should be reported when FA of different traits is compared.

Introduction

Fluctuating asymmetry (FA), small random departures from bilateral symmetry (Van Valen, 1962; Markow, 1995), is often seen as an indicator of 'overall quality' or 'general health' (e.g. Zakharov, 1992; Møller, 1997). The idea behind this concept is that individuals of low quality cannot control their development precisely, and consequently more often develop different phenotypes on both sides. But the relationship between FA and quality is not that straightforward. Although a large number of studies report positive relationships between aspects of fitness and FA (e.g. Møller, 1992; Møller *et al.*, 1995, 1996; Møller & Zamora-Munoz, 1997), absence of such a relationship is quite frequent (e.g. Markow & Ricker, 1992; Fowler & Whitlock, 1994; Tomkins & Simmons, 1995; Dufour & Weatherhead, 1998a; Dufour & Weatherhead, 1998b) or ambiguous. FA may differ between traits depending on, for example, whether they are involved in sexual selection (Møller & Pomiankowski, 1995; Dufour & Weatherhead, 1996) or whether natural selection on trait size or symmetry itself is intense (Balmford *et al.*, 1993; Brakefield & Breuker, 1996; Crespi, 1997; Brown & Bomberger Brown, 1998). Clearly it would be desirable to compare FA, and its heritability, of different (kind of) traits (Leamy, 1997; Markow, 1997; Palmer & Strobeck, 1997; Pomiankowski, 1997).

How to test for differences in FA of different traits is not straightforward. Different traits vary at different scales and a certain difference between left and right can be large or small depending on the trait considered. Dufour & Weatherhead (1996), for example, compared several traits of the red winged blackbird. They noted that the absolute asymmetry of different traits was related to trait size, and that the asymmetry of the epaulet coloration was especially high, but did not employ statistical tests to evaluate the significance of the difference. Møller & Hoglund (1991) divided the difference between left and

erhead, 1998b) or ambiguous. FA may differ between traits depending on, for example, whether they are involved in sexual selection (Møller & Pomiankowski, 1995; Dufour & Weatherhead, 1996) or whether natural selection on trait size or symmetry itself is intense (Balmford *et al.*, 1993; Brakefield & Breuker, 1996; Crespi, 1997; Brown & Bomberger Brown, 1998). Clearly it would be desirable to compare FA, and its heritability, of different (kind of) traits (Leamy, 1997; Markow, 1997; Palmer & Strobeck, 1997; Pomiankowski, 1997).

Correspondence: Sören Nylin, Department of Zoology, University of Stockholm, 106 91 Stockholm, Sweden.
Tel.: + 46 8164033; fax: +46 8167715; e-mail: soren.nylin@zoologi.su.se

right by the mean trait size to compare wing and tail ornaments of different birds, but other corrections for scale effects are possible. Clearly a comparison of the different measures of FA that take into account scale differences is desirable.

Palmer & Strobeck (1986) and Palmer (1994) thoroughly analysed how to compare FA *within* traits. They reviewed different indices used for quantifying FA and simulated populations with different levels of FA for comparisons. There exists, however, no review of how to analyse FA across traits. This paper fills in the gap by comparing different methods to analyse FA *between* traits. To do so, Monte-Carlo simulations (Crowley, 1992) are performed. This is a powerful method to determine bias and error rates of different statistical methods. These analyses also check some of the results of Palmer & Strobeck (1986) for variation in FA within traits, and additionally provide type I error rates (how often is a significant difference indicated, when there is none), something not done by the simulations in Palmer & Strobeck (1986). The different methods are also evaluated using an empirical data-set of 36 traits in the Speckled Wood Butterfly (*Pararge aegeria*).

Methods and results

Indices for comparing FA across traits

Fluctuating asymmetry of single traits is quantified by measuring the left and right side and calculating an FA index. Palmer & Strobeck (1986) list 10 possible FA indices, and recommend two of these, **FA3** and **FA7**, if samples have to be compared when FA is related to trait size. **FA3** is the mean of the absolute value of the difference between left and right (=unsigned difference), divided by the population trait size mean, while **FA7** is the variance of the (signed) difference, divided by the population trait size mean (Table 1). **FA2** and **FA6** are very similar to **FA3** and **FA7**, respectively, and can also be used to correct for differences caused by trait size. The difference is that **FA2** and **FA6** correct by dividing

Table 1 Measures that can be used to compare levels of FA between traits measured at different scales. Subscript R – L denotes difference in value measured on left and right side, TS denotes trait size = average of values of left and right side.

Recommended by				
Palmer & Strobeck (1986)		Index used in this paper		
No.	Index	No.	Index	Significance test
FA3	$\mu_{ R-L } / \mu_{TS}$	FA2	$\mu_{ R-L } / TS$	<i>t</i> -test
		FA3	$\mu_{ R-L } / \mu_{TS}$	<i>t</i> -test
FA7	$\sigma_{R-L}^2 / \mu_{TS}^2$	FA6	σ_{R-L}^2 / TS^2	<i>F</i> -test
		FA7a	$\sigma_{R-L}^2 / \mu_{TS}^2$	<i>F</i> -test
FA9	$1 - \rho_{R,L}^2$	FA9a	$1 - \rho_{R,L}$	confidence intervals bootstrap

the FA value of each individual with its own trait size (mean of left and right). Dividing by the mean can be problematic if the mean is an arbitrary value, or where it is for example 0 or a negative value, such as the position of an element measured as a distance from a landmark.

A 5th index, **FA9**, is also insensitive to trait size differences, and independent of the mean. This is 1 minus the square of the correlation coefficient between left and right. This index is not often used, probably because it is not recommended by Palmer & Strobeck (1986). Procrustes methods, which use landmark points and the squared distance between them to quantify shape variation, have recently been used for the analysis of FA (Auffray *et al.*, 1996; Smith *et al.*, 1997; Klingenberg & McIntyre, 1998). These methods are related to **FA9** (Smith *et al.*, 1997).

A statistical test is needed to compare FA of two traits. A *t*-test is widely used to compare unsigned FA values (**FA2** and **FA3**) between two samples (e.g. Møller, 1995; Dufour & Weatherhead, 1998a). Often values are transformed to obtain a normal distribution, prior to applying a *t*-test (Swaddle *et al.*, 1994), but Palmer (1994) states that this is not needed. *F*-tests are used to compare two variances such as **FA6** and **FA7**. **FA7** must, however, be modified slightly in order to apply an *F*-test. **FA7** is a variance divided by trait size mean. But the mean varies at a different scale from the variance, which varies at a quadratic scale. Thus the ratios of two **FA7** indices do not follow a *F*-distribution. It is thus better to divide the variance of signed differences by the square of the mean rather than the mean. This index will be referred to in the rest of the paper as **FA7a**. In a Monte-Carlo simulation (details of the procedure are given below) of two samples with equal FA and equal means, **FA7** ratios indicated a significant difference in 35.4% instead of the expected 5%, while **FA7a** ratios indicated a difference in 4.8%. The same reasoning does not apply to **FA6**, for which the difference between left and right is divided by trait size (mean of left and right) before calculating a variance, and at that stage both trait size and the signed difference vary at the same scale.

The procedure proposed by Palmer (1994) to determine which of the two FA-indices is larger and then to take the ratio of the larger index over the smaller is not correct either. This procedure assumes that the *F*-distribution is symmetrical, which it is not. The correct way is first to calculate the FA-indices and their ratio and then to apply a two-sided *F*-test. Thus for a test at the 5% level the 2.5% and 97.5% percentiles are looked up in a statistical table at the appropriate degrees of freedom.

FA9 contains the square of a correlation coefficient. One can use correlation coefficients to test for differences between traits. In this paper we will use $1 - r$ (termed **FA9a**) instead of $1 - r^2$, since this measure is also related more straightforwardly to **FA7** and **FA2** (see below). Ninety-five per cent confidence intervals for correlation coefficients are constructed by *z*-transforming *r*, adding/

subtracting $1.96/\sqrt{(N-3)}$ and then back-transforming (Sokal & Rohlf, 1981). Two correlation coefficients are significantly different at the 5% level if their 95% confidence intervals do not overlap, although this test is often conservative. An alternative is to bootstrap the difference between two correlation coefficients (Efron & Tibshirani, 1993). A bootstrap proceeds as follows: draw at random one individual for each trait, place them back in the data-set, repeat the two preceding steps until the size of the original data-set is reached, calculate the estimate for the new data-set, repeat the whole procedure a large number of times (e.g. 1000×), rank the bootstrap estimates, and the 95% confidence intervals are given by the 2.5 and 97.5 percentiles (e.g. the 25th and 976th estimate for 1000 runs).

Relationships between the indices

Mathematical relationships between the indices are straightforward under certain assumptions. These assumptions are that (i) the signed difference between left and right is normally distributed (ii) with a mean of 0, and (iii) both left and right are normally distributed with equal variances. The relationship between the mean of the unsigned difference ($\mu_{|R-L|} = \mathbf{FA1}$) and the standard deviation of the signed difference between left and right ($\sigma_{R-L} = \mathbf{FA4}^{1/2}$) is given by Palmer (1994) and Houle (1997):

$$\mu_{|R-L|} = \sigma_{R-L} \sqrt{(2/\pi)} \quad (1)$$

where R and L denote right and left. Rewriting gives

$$\sigma_{R-L}^2 = \mu_{|R-L|}^2 (\pi/2). \quad (2)$$

FA7a is thus the square of **FA3** multiplied by a constant.

The relationship between **FA9** and **FA7** is also straightforward. The variance of the difference of two variables is the sum of their variances minus twice their covariance. This gives the following formula for **FA3**:

$$\sigma_{R-L}^2 = \sigma_L^2 + \sigma_R^2 - 2\text{COV}_{R,L}. \quad (3)$$

Rewriting and keeping in mind that $\sigma_L^2 = \sigma_R^2 = \sigma_{TS}^2$ (the subscript TS denoting trait size) gives

$$\text{COV}_{R,L} = \sigma_{TS}^2 - (\sigma_{R-L}^2/2). \quad (4)$$

The correlation coefficient between left and right is $\text{COV}_{R,L}$ divided by the product of σ_L and σ_R . This product is equal to the variance of the trait size (σ_{TS}^2). Division by σ_{TS}^2 thus gives the correlation coefficient on the left and thus

$$1 - \rho_{R-L} = (\sigma_{R-L}^2/2\sigma_{TS}^2). \quad (5)$$

FA9a is thus simply **FA4** standardized by (twice the) the variance of trait size, whereas **FA7a** is **FA4** standardized by the square of the mean of trait size. Because trait size variance is related to trait size range, **FA9a** is also related to trait size range. Because of this Palmer & Strobeck (1986) recommend not using **FA9**.

Monte Carlo simulations: which index is the best?

Methods

Monte-Carlo simulations were performed to determine the statistical power of the different methods to compare FA among traits. Samples were generated using the following formula:

$$TS_{is} = \mu_{TS} + ts_i + FA_s$$

where TS_{is} is the phenotype of the i th individual at side s (= left or right), μ_{TS} = the mean trait size, ts_i is the deviation of the trait size of individual i from the population mean caused by genetic and environmental influences common to both sides, drawn from a normal distribution with mean 0 and variance σ_{ts}^2 , and FA_s is the deviation of side s caused by fluctuating asymmetry, drawn from a normal distribution with mean 0 and variance $0.5\sigma_{R-L}^2$ (the variance of R - L is twice the variance due to the FA contribution of the individual sides). Simulations were run with values for $\mu_{|R-L|}$ and σ_{R-L}^2 comparable to those reported in the literature (e.g. Houle, 1992; Palmer & Strobeck, 1997; Pomiankowski, 1997). To obtain these, use was made of the relationships derived in the previous section. The value of σ_{R-L}^2 is related to the mean unsigned asymmetry ($\mu_{|R-L|}$) by eqn 2 and the total variance of the trait size (σ_{TS}^2) is the sum of σ_{ts}^2 and σ_{R-L}^2 . Thus, for example, for a trait with $\mu_{|R-L|} = 0.1$ and $\sigma_{TS}^2 = 1$, $0.5\sigma_{R-L}^2$ was 0.007854 and σ_{ts}^2 0.992146. The simulations here are basically similar to those in Palmer & Strobeck (1986) except that ts_i is here drawn from a normal distribution whereas Palmer & Strobeck (1986) used a uniform distribution. The important assumptions here are that no measurement error is involved (or that it has been removed by, for example, repeated measurements), no directional or antisymmetry exists, within traits FA is not dependent on trait size, and that FA is normally distributed (see Van Dongen (1998) for a discussion of non-normal distributions in FA).

Simulations were run with different combinations of μ_{TS} , σ_{ts}^2 and σ_{R-L}^2 . For each combination 500 samples of pairs of traits, each with two sides, were simulated. Each sample consisted of 50 individuals. For each sample the indices were calculated and whether they differed significantly at the 5% level between the two traits. This means that for each run of 500 samples 25 (= 5%) were expected to differ significantly, a lower percentage indicating a conservative test. Values for μ_{TS} were 1, 2, 5 or 10; values for μ_{FA} were 0.01, 0.02, 0.05 and 0.10, thus ranging from 0.1% to 10%, and values for σ_{TS} were 0.1, 0.2, 0.5 and 1, the coefficient of variation (CV) thus ranging from 1% to 100%.

Can the tests handle scale differences correctly?

To determine whether the three methods work properly when scale differences are involved, traits were

Table 2 Performance of indices (see Table 1) comparing traits with fluctuating asymmetry varying relative to trait size mean and variance. Parameter values for trait 1 were always the same; for trait 2 they were 1, 2, 5 or 10 times as large as for trait 1, so that the relative asymmetry remained the same. Mean is the average of 500 simulation runs; bias is given by the difference between expected and observed means. Error rate indicates the percentage of tests that indicate a significant difference out of 500 runs. **FA1** and **FA4** are tests uncorrected for differences in trait size means.

Parameters			FA2/3				FA6/7a				FA9a						
			Mean	Error rate <i>t</i> -test			Mean	Error rate <i>F</i> -test			Mean	Error rate					
μ_{TS}	σ_{TS}	μ_{R-L}	FA3 untrans	FA3 box-cox	FA2	FA1	FA7a	FA6	FA4		CI	Bootstrap					
expected			0.01000	5.0%	5.0%	5.0%	5.0%	0.01253	5.0%	5.0%	5.0%	0.00785	5.0%	5.0%			
trait 1			1	0.1	0.01	0.00998						0.00804					
trait 2			1	0.1	0.01	0.01007	5.6%	5.6%	5.2%	5.2%	0.01255	4.6%	5.8%	4.6%	0.00818	0.4%	0.0%
			2	0.2	0.02	0.01001	5.8%	6.8%	6.6%	99.0%	0.01247	4.6%	5.0%	99.4%	0.00800	0.6%	0.0%
			5	0.5	0.05	0.01001	5.4%	4.2%	5.0%	100.0%	0.01251	4.6%	4.8%	100.0%	0.00799	0.2%	0.0%
			10	1.0	0.10	0.00998	6.6%	6.2%	5.2%	100.0%	0.01244	4.8%	7.0%	100.0%	0.00799	0.4%	0.0%

compared that all had a μ_{FA} which was 1% of μ_{TS} and 10% of σ_{TS} . The first trait had always a μ_{TS} of 1 and a σ_{TS} of 0.1; for the second trait these parameters were 1, 2, 5 or 10 times as large.

All indices perform well (Table 2). The average value is always close to the expected value, and thus the bias (difference between expected and observed mean) is small. For all indices, except **FA9a**, about or slightly more than the expected 5% of the runs indicate a significant difference. Box-Cox transformations hardly made a difference for **FA3** (Table 2) and **FA2** (not shown), so in subsequent tests we do not show error rates for Box-Cox transformed values. Differences are found in less than 1% of the runs for **FA9a**, both when using confidence intervals or bootstrapping, and these tests are thus conservative.

How often are differences indicated when they are not real?

To determine how often the indices indicate differences when they are not real (type I error), simulations were run with two traits that were equal for μ_{TS} , μ_{R-L} and σ_{R-L}^2 . Simulations were run with $\mu_{TS} = 1$, $\mu_{FA} = 0.01$ and $\sigma_{R-L} = 0.2$. To determine whether variation in one of the parameters influenced the error rate, more simulations were run with one parameter varying independently from the others.

Differences are found in less than 1% of the runs for **FA9a** using either confidence intervals or bootstrapping (Table 3). Differences are found for slightly more than the expected 5% in most cases for **FA2** and **FA7a**. When the variance in trait size is high both **FA2** and **FA7a** perform

Table 3 Type I error. Comparison of two traits that have equal trait size means and variances, and equal levels of fluctuating asymmetry. Error rate gives the percentage of simulations that indicate a significant difference between the two traits.

Parameters			FA2/3				FA6/7a				FA9a			
			Trait 1		Trait 2		<i>t</i> -test error rate		Trait 1		Trait 2		<i>F</i> -test error rate	
μ_{TS}	μ_{R-L}	σ_{TS}			FA3	FA2			FA7a	FA6	Trait 1	Trait 2	CI	Bootstrap
1	0.01	0.2	0.0100	0.0101	7.2%	5.6%	0.01249	0.01256	7.8%	9.4%	0.00199	0.00208	0.8%	0.0%
1	0.02	0.2	0.0200	0.0198	8.2%	6.8%	0.02494	0.02466	7.8%	13.8%	0.00901	0.00776	0.6%	0.0%
1	0.05	0.2	0.0501	0.0495	4.6%	4.4%	0.06252	0.06193	5.0%	7.2%	0.04979	0.04973	0.6%	0.0%
1	0.10	0.2	0.0997	0.0996	5.8%	5.0%	0.12427	0.12412	6.6%	8.4%	0.20207	0.20135	1.0%	0.0%
1	0.01	0.1	0.0100	0.0101	5.6%	5.2%	0.01249	0.01255	4.8%	5.8%	0.00804	0.00818	0.4%	0.0%
1	0.01	0.5	0.0100	0.0101	12.0%	1.6%	0.01257	0.01259	11.4%	77.0%	0.00033	0.00032	0.8%	0.0%
1	0.01	1.0	0.0104	0.0103	25.4%	2.2%	0.01295	0.01285	27.2%	78.8%	0.00008	0.00008	0.8%	0.0%
1	0.10	1.0	0.1024	0.1017	24.4%	2.2%	0.12754	0.12669	26.0%	80.6%	0.00899	0.00897	0.6%	0.0%
10	0.10	1.0	0.0100	0.0099	6.8%	6.4%	0.01244	0.01243	7.8%	7.6%	0.00899	0.00897	0.6%	0.0%
2	0.01	0.2	0.0049	0.0050	6.0%	5.6%	0.00625	0.00628	6.4%	6.2%	0.00199	0.00208	0.8%	0.0%
5	0.01	0.2	0.0020	0.0020	5.4%	5.8%	0.00250	0.00251	5.4%	5.0%	0.00199	0.00208	0.8%	0.0%
10	0.01	0.2	0.0010	0.0010	5.2%	5.6%	0.00125	0.00125	4.8%	4.6%	0.00199	0.00208	0.8%	0.0%

Table 4 Type II error. Comparison of two traits with different levels of fluctuating asymmetry, but equal mean and variances for trait size. Error rate: percentage of runs that indicated no difference. Mean trait size was always 1 for both traits.

Parameters			FA2/3				FA6/7a				FA9a			
μ_{TS}		σ_{TS}	<i>t</i> -test error rate				<i>F</i> -test error rate						Error rate	
Trait 1	Trait 2		Trait 1	Trait 2	FA3	FA2	Trait 1	Trait 2	FA7a	FA6	Trait 1	Trait 2	CI	Bootstrap
0.01	0.02	0.1	0.0100	0.0201	0.8%	1.2%	0.01249	0.02495	0.2%	0.8%	0.00904	0.03192	27.6%	3.6%
0.01	0.02	0.2	0.0100	0.0200	1.4%	1.8%	0.01249	0.02494	0.8%	0.8%	0.00199	0.00800	25.6%	4.8%
0.01	0.02	0.5	0.0101	0.0202	1.2%	75.2%	0.01257	0.02520	0.8%	17.2%	0.00033	0.00129	27.0%	3.6%
0.01	0.02	1.0	0.0104	0.0203	6.2%	96.2%	0.01295	0.02531	6.2%	14.2%	0.00008	0.00032	28.8%	3.6%
0.01	0.05	0.1	0.0100	0.0502	0.0%	0.0%	0.01249	0.06266	0.0%	0.0%	0.00904	0.20209	0.0%	0.0%
0.01	0.05	0.2	0.0100	0.0501	0.0%	0.0%	0.01249	0.06252	0.0%	0.0%	0.00199	0.04979	0.0%	0.0%
0.01	0.05	0.5	0.0101	0.0502	0.0%	47.4%	0.01257	0.06274	0.0%	6.0%	0.00033	0.00798	0.0%	0.0%
0.01	0.05	1.0	0.0104	0.0513	0.0%	91.0%	0.01295	0.06405	0.0%	6.8%	0.00008	0.00200	0.0%	0.0%
0.01	0.10	0.1	0.0100	0.0997	0.0%	0.0%	0.01249	0.12431	0.0%	0.0%	0.00904	0.77763	0.0%	0.0%
0.01	0.10	0.2	0.0100	0.0996	0.0%	0.0%	0.01249	0.12427	0.0%	0.0%	0.00199	0.20107	0.0%	0.0%
0.01	0.10	0.5	0.0101	0.1004	0.0%	39.2%	0.01257	0.12509	0.0%	2.8%	0.00033	0.03179	0.0%	0.0%
0.01	0.10	1.0	0.0104	0.1024	0.0%	85.2%	0.01295	0.12754	0.0%	3.8%	0.00008	0.00799	0.0%	0.0%
0.02	0.05	0.1	0.0201	0.0502	0.2%	0.2%	0.02495	0.06266	0.0%	0.0%	0.03192	0.20209	4.2%	0.0%
0.02	0.05	0.2	0.0200	0.0501	0.4%	0.4%	0.02494	0.06252	0.0%	0.0%	0.00800	0.04979	4.2%	0.0%
0.02	0.05	0.5	0.0202	0.0502	0.2%	64.2%	0.02520	0.06274	0.0%	11.2%	0.00130	0.00799	4.8%	0.8%
0.02	0.05	1.0	0.0203	0.0513	0.8%	94.2%	0.02531	0.06405	0.4%	14.6%	0.00032	0.00200	6.0%	0.6%
0.02	0.10	0.1	0.0200	0.0997	0.0%	0.0%	0.02495	0.12431	0.0%	0.0%	0.03192	0.77794	0.0%	0.0%
0.02	0.10	0.2	0.0200	0.0997	0.0%	0.0%	0.02494	0.12427	0.0%	0.0%	0.00800	0.20107	0.0%	0.0%
0.02	0.10	0.5	0.0202	0.1004	0.0%	46.2%	0.02520	0.12509	0.0%	7.2%	0.00130	0.03179	0.0%	0.0%
0.02	0.10	1.0	0.0203	0.1024	0.0%	90.4%	0.02531	0.12754	0.0%	6.2%	0.00032	0.00799	0.0%	0.0%
0.05	0.10	0.1	0.0502	0.0997	0.6%	1.2%	0.06266	0.12431	0.2%	0.4%	0.20209	0.77763	6.2%	0.4%
0.05	0.10	0.2	0.0501	0.0996	1.4%	1.4%	0.06252	0.12427	0.4%	0.2%	0.04979	0.20107	20.4%	3.2%
0.05	0.10	0.5	0.0502	0.1004	1.8%	74.4%	0.06274	0.12509	0.8%	13.0%	0.00709	0.03179	28.4%	4.0%
0.05	0.10	1.0	0.0513	0.1024	6.2%	94.4%	0.06405	0.12754	5.2%	15.0%	0.00200	0.00799	26.2%	4.0%

worse: up to more than 25% of the time differences are indicated. Checking individual results indicates that this happens when the estimated population mean trait size differs considerably (while in fact they are equal). When the CV is low, high error rates disappear (compare rows 8 and 9 in Table 3). In most traits measured for FA these high CVs will not occur because this means that negative values (if CV = 100%) or values close to 0 (if CV = 50%) can occur. Because of trait sizes close to 0, correcting on an individual basis, as done by **FA2** and **FA6**, does not improve the error rates. Values close to zero give extremely large positive or negative values. Consequently, the variance in samples containing such values increases, and this results in conservatism for *t*-tests (used for **FA2**) and liberalism for *F*-tests (used for **FA6**).

How often are differences not indicated when they do exist?

To test whether the three indices differ in their ability to distinguish between different levels of FA, two traits with equal mean and variances for trait size but with different levels of fluctuating asymmetry were compared. All six combinations of $\mu_{R-L} = 0.01, 0.02, 0.05$ and 0.1 were tested. These simulations were run at four levels of σ_{R-L}

(0.1, 0.2, 0.5 and 1.0), to see if the type II error rate depended on the CV, as was the case for type I error rates for **FA2/3** and **FA6/7a**.

Generally, type II error rates are low for **FA3**, **FA7a** and **FA9a** when bootstrapping is used (Table 4). Highest type II error rates for **FA3** and **FA7a** (up to 6.2%) are found when CV is high (>50%). Type II error rates for **FA9a** with bootstrapping are highest when the difference between μ_{R-L} is equal or less than two-fold, but they are always lower than 5%. When **FA9a** with confidence intervals is used, error rates increase to well above 20%. **FA6** and especially **FA2** have high type II error rates when trait size variance is relatively large. This is caused by the presence of trait sizes close to 0.

Is FA related to trait size mean or variance?

The decision of which index to apply must also depend on whether one wants to standardize by the trait size variance (**FA9a**) or mean (all other indices). In our opinion there is no principal reason why one should be preferred over the other. Trait size variance tends to increase with trait size mean, and FA is thus related to both. It is not known whether FA tends to vary more

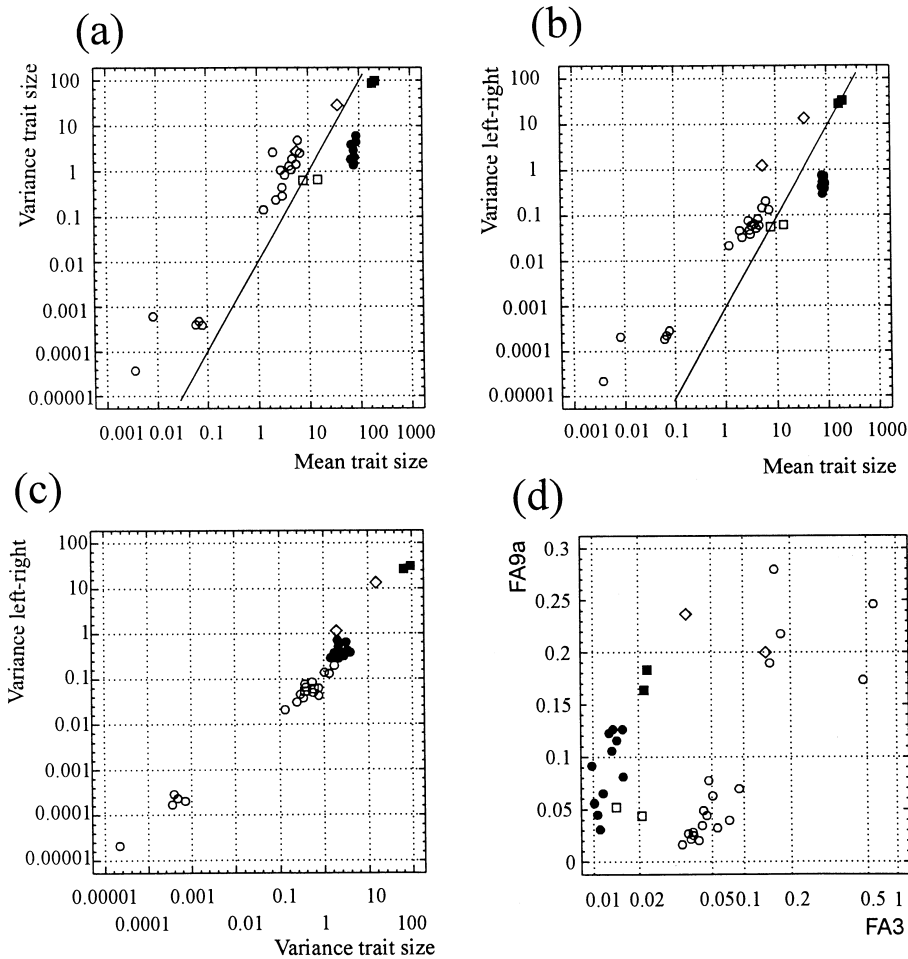


Fig. 1 Relationships between variance and mean of trait size in 36 traits of the Speckled wood butterfly (*Pararge aegeria*). \circ = areas of wing spots, \bullet = colour of wings spots, \square = size of wing, \blacksquare = darkness of wings, \diamond = pupal traits. (a) Variance and mean of trait sizes. Line indicates coefficient of variation = 10%. (b) Variance of the difference between left and right (FA4) and mean trait size. Line connects points where the mean absolute difference between left and right = 2%. (c) FA4 and trait size variance. (d) FA9a (=FA standardized by trait size variance) and FA3 (FA standardized by trait size mean).

with trait size mean or variance. To begin to answer this we have compared 36 traits in a laboratory population of the Speckled Wood Butterfly *Pararge aegeria*.

Traits were measured on 184 individuals, offspring of 16 wild caught females. Traits consisted of two pupal traits, 32 measurements on wing pattern elements, and length and width of hind wing. Traits were measured with an image analyser (Windig, 1991). A detailed description of measurement procedures and measurement error can be found in the Appendix. For all traits FA was considerably larger than measurement error. Directional asymmetry was involved for two traits (areas of yellow spots) on the hind wing.

Relationships between mean trait size, variance of trait size and σ_{R-L}^2 are presented in Fig. 1. As expected, there is a narrow relationship between mean and variance of trait size ($r = 0.8577$; Fig. 1A), but with some variation. Wing

size measurements and especially measurement of spot colour have relatively lower trait size variance, while small spots have relatively higher variance. σ_{R-L}^2 follows a remarkably similar pattern (Fig. 1B). When compared with the trait size mean it is relatively small for wing size and spot colour measurements and large for small spots. Consequently, it is slightly better correlated to the variance of trait size ($r = 0.9645$; Fig. 1C) than to the mean of trait size ($r = 0.9367$), though the difference between the correlations is not significant (bootstrap of difference between correlations: $P = 0.196$). The correlation between \log FA3 and FA9a is 0.5075 (Fig. 1D).

Discussion

FA9a with bootstrapping always has low error rates. FA3 and FA7a are next best; they tend to indicate too often

significant differences if trait size values close to or below 0 occur. The other three tests perform less well. All indices behave well in that they correct as expected for trait size differences. **FA9a** does this by standardizing to trait size variance, a fundamentally different way from the other indices, which standardize by trait size mean.

Palmer & Strobeck (1986) recommend not using **FA9**. The main reason for this recommendation is that the index depends on trait size range. As we can see here the reason for this is that **FA9** standardizes σ_{R-L}^2 by trait size variance, which is related to trait size range, even more so in the uniform distributions used by Palmer & Strobeck (1986). They tested the discriminatory ability by comparing the results of five different levels of FA with an ANOVA. For **FA9** this leads to a low value because the absolute difference between two estimated values can be very small (see, e.g. rows 3 and 4 in Table 4), while in fact they are significantly different.

Measurement error is of particular importance in studies of FA (Palmer, 1994). **FA9a** can be corrected for measurement error in a similar way as Palmer (1994) explain for the calculation of **FA10**. In a repeated measures ANOVA with side, individual and their interaction as factors, the error mean square is an estimate of the variance caused by measurement error (V_{ME}). The correlation between left and right is the ratio of the covariance of left and right and the product of the standard deviations of left and right (which is equal to the variance of trait size). The covariance is not affected by measurement error, but the trait size variance is increased. To correct for measurement error one has to subtract $2 \times V_{ME}$ from the nominator of the correlation coefficient (two times because measurement error is added to both the left and the right sides).

FA9a can thus be corrected for measurement error. It also has the advantage over other indices that it can be readily used for traits with an arbitrary mean, or a mean less than or equal to 0. **FA9a** is also easily computed and it thus seems a very useful index for comparing FA among traits. At the genetic level usually a heritability of the unsigned difference (such as **FA2**) is calculated (e.g. Windig, 1998; Woods *et al.*, 1998), but here the same recommendation can be made. It seems a sensible strategy to report an estimate similar to **FA9a**, the genetic correlation between the left and right sides, as suggested by Roff (1997).

FA9a seems to be the best index to compare FA across traits. If the closer relationship between FA and the variance of trait size compared to trait size mean that was found for the *Pararge aegeria* data is a general phenomenon it is indeed to be preferred above the other traits. Standardizing by trait size variance does, however, obscure a possible relationship between trait size variance and FA, a relationship predicted by the hypothesis that more variable traits have higher developmental instability (Livshits *et al.*, 1998). The other indices that standardize by trait size obscure, however, a possible

relationship of FA with trait size. **FA3** and **FA7a** can be applied if values close to or below 0 do not occur. **FA3** has the advantage that FA as a percentage of the mean is intuitively easy to understand. **FA7a** can easily be corrected for measurement error, but the same is possible for **FA9a**. Since most studies up to now have published only **FA2/3** it seems wise to report both **FA9a** and **FA3** in future studies that compare FA between traits.

Acknowledgments

We are grateful for comments on an earlier version by Stefan van Dongen (Antwerp). Support was provided by the Swedish Natural Science Research Council.

References

- Auffray, J., Alibert, P., Renaud, S. & Bonhomme, F. 1996. Fluctuating asymmetry in *Mus musculus* subspecific hybridization: traditional and procrustes comparative approaches. In: *Advances in Morphometrics* (L. Marcus, M. Corti, A. Loy, G. Naylor & D. Slice, eds), pp. 275–283. Plenum Press, NY.
- Balmford, A., Jones, I.J. & Thomas, A.L.R. 1993. On avian asymmetry: evidence of natural selection for symmetrical tails and wings in birds. *Proc. R. Soc. Lond. B* **252**: 245–251.
- Brakefield, P.M. & Breuker, C.J. 1996. The genetical basis of fluctuating asymmetry for developmentally integrated traits in a butterfly eyespot pattern. *Proc. R. Soc. Lond. B* **263**: 1557–1563.
- Brown, C.R. & Bomberger Brown, M. 1998. Intense natural selection on body size and wing and tail asymmetry in cliff swallows during severe weather. *Evolution* **52**: 1461–1475.
- Crespi, B.J. 1997. Fluctuating asymmetry in vestigial and functional traits of a haplodiploid insect. *Heredity* **79**: 624–630.
- Crowley, P.H. 1992. Resampling methods for computation-intensive data-analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.* **23**: 405–447.
- Dufour, K.W. & Weatherhead, P.J. 1996. Estimation of organism-wide asymmetry in red-winged blackbirds and its relation to studies of mate selection. *Proc. R. Soc. Lond. B* **263**: 769–775.
- Dufour, K.W. & Weatherhead, P.J. 1998a. Bilateral symmetry and social dominance in captive male red-winged blackbirds. *Behav. Ecol. Sociobiol.* **42**: 71–76.
- Dufour, K.W. & Weatherhead, P.J. 1998b. Bilateral symmetry as an indicator of male quality in red-winged blackbirds: associations with measures of health, viability, and parental effort. *Behav. Ecol.* **9**: 220–231.
- Efron, B. & Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, NY.
- Fowler, K. & Whitlock, M.C. 1994. Fluctuating asymmetry does not increase with moderate inbreeding in *Drosophila melanogaster*. *Heredity* **73**: 373–376.
- Houle, D. 1992. Comparing evolvability and variability of quantitative traits. *Genetics* **130**: 195–204.
- Houle, D. 1997. Comment on 'a meta-analysis of the heritability of developmental stability' by Møller and Thornhill. *J. Evol. Biol.* **10**: 17–20.
- Klingenberg, P.L. & McIntyre, G.S. 1998. Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with procrustes methods. *Evolution* **52**: 1363–1375.

- Leamy, L.J. 1997. Is developmental stability heritable? *J. Evol. Biol.* **10**: 21–29.
- Livshits, G., Yakovenko, K., Kletselman, L., Karasik, D. & Kobylansky, E. 1998. Fluctuating asymmetry and morphometric variation of hand bones. *Am. J. Phys. Anthropol.* **107**: 125–136.
- Markow, T.A. 1995. Evolutionary ecology and developmental stability. *Ann. Rev. Ent.* **40**: 105–120.
- Markow, T.A. 1997. Meta-analysis of the heritability of developmental stability: a giant step backward. *J. Evol. Biol.* **10**: 31–37.
- Markow, T.A. & Ricker, J.P. 1992. Male size, developmental stability, and mating success in natural populations of three *Drosophila* species. *Heredity* **69**: 122–127.
- Møller, A.P. 1992. Patterns of fluctuating asymmetry in weapons: evidence for reliable signaling of quality in beetle horns and bird spurs. *Proc. R. Soc. Lond. B* **248**: 199–208.
- Møller, A.P. 1995. Leaf-mining insects and fluctuating asymmetry in elm *Ulmus glabrus* leaves. *J. Anim. Biol.* **64**: 697–707.
- Møller, A.P. 1997. Developmental stability and fitness: a review. *Am. Nat.* **149**: 916–932.
- Møller, A.P., Cuervo, J.J., Soler, J.J. & Zamora-Munoz, C. 1996. Horn asymmetry and fitness in gemsbok, *Oryx g. gazella*. *Behav. Ecol.* **7**: 247–253.
- Møller, A. & Hoglund, J. 1991. Patterns of fluctuating asymmetry in avian feather ornaments: implications for models of sexual selection. *Proc. R. Soc. Lond. B* **245**: 1–5.
- Møller, A.P. & Pomiankowski, A. 1995. Fluctuating asymmetry and sexual selection. *Genetica* **89**: 267–279.
- Møller, A.P., Soler, M. & Thornhill, R. 1995. Breast asymmetry sexual selection, and human reproductive success. *Ethol. Sociobiol.* **16**: 207–219.
- Møller, A.P. & Zamora-Munoz, C. 1997. Antennal asymmetry and sexual selection in a cerambycid beetle. *Anim. Behav.* **54**: 1509–1515.
- Palmer, A.R. 1994. Fluctuating asymmetry analyses: a primer. *Developmental Instability: its Origins and Evolutionary Implications* (T. A. Markow, ed.), pp. 335–364. Kluwer Academic Publishers, Dordrecht.
- Palmer, A.R. & Strobeck, C. 1986. Fluctuating asymmetry: measurements, analysis and pattern. *Ann. Rev. Ecol. Syst.* **17**: 391–421.
- Palmer, A.R. & Strobeck, C. 1997. Fluctuating asymmetry and developmental stability: heritability of observable variation vs. heritability of inferred cause. *J. Evol. Biol.* **10**: 39–49.
- Pomiankowski, A. 1997. Genetic variation in fluctuating asymmetry. *J. Evol. Biol.* **10**: 51–55.
- Roff, D.A. 1997. *Evolutionary Quantitative Genetics*, Chapman & Hall, New York.
- Smith, D.R., Crespi, B.J. & Bookstein, F.L. 1997. Fluctuating asymmetry in the honey bee, *Apis mellifera*: effects of ploidy and hybridization. *J. Evol. Biol.* **10**: 551–574.
- Sokal, R.R. & Rohlf, F.J. 1981. *Biometry*, Freeman, San Francisco.
- Swaddle, J.P., Witter, M.S. & Cuthill, I.C. 1994. The analysis of fluctuating asymmetry. *Anim. Behav.* **48**: 686–689.
- Tomkins, J.L. & Simmons, L.W. 1995. Patterns of fluctuating asymmetry in earwig forceps: no evidence for reliable signaling. *Proc. R. Soc. Lond. B* **259**: 89–96.
- Van Dongen, S. 1998. How repeatable is the estimation of individual fluctuating asymmetry? *Proc. R. Soc. Lond. B* **263**: 1423–1427.
- Van Valen, L. 1962. A study of fluctuating asymmetry. *Evolution* **16**: 125–142.

- Windig, J.J. 1991. Quantification of Lepidoptera wing patterns using an image analyser. *J. Res. Lepid.* **30**: 82–94.
- Windig, J.J. 1998. Evolutionary genetics of fluctuating asymmetry in the peacock butterfly (*Inachis io*). *Heredity* **80**: 382–392.
- Woods, R.E., Hercus, M.J. & Hoffmann, A.A. 1998. Estimating the heritability of fluctuating asymmetry in field *Drosophila*. *Evolution* **52**: 816–824.
- Zakharov, V.M. 1992. Population phenogenetics: analysis of developmental stability in natural populations. *Act. Zool. Fenn.* **191**: 7–30.

Received 8 April 1999; accepted 27 May 1999

Appendix

Measurement of traits

Traits were measured with an image analyser. The image analyser consisted of a video camera (JVC TK 5109), connected to a Pentium PC. Qwin software from Leica was used to grab 24-bit colour images (578 × 764 pixels), and measuring traits. Thirty-four traits were analysed (see Fig. A1) with the help of a software program written by one of us (J.J.W.). For the yellow spots both the area and colour (or hue) were analysed. Colour was measured on a scale from red (= 0) to yellow, green, blue and again red (= 255). The yellow spots varied in colour from dark

Traits compared

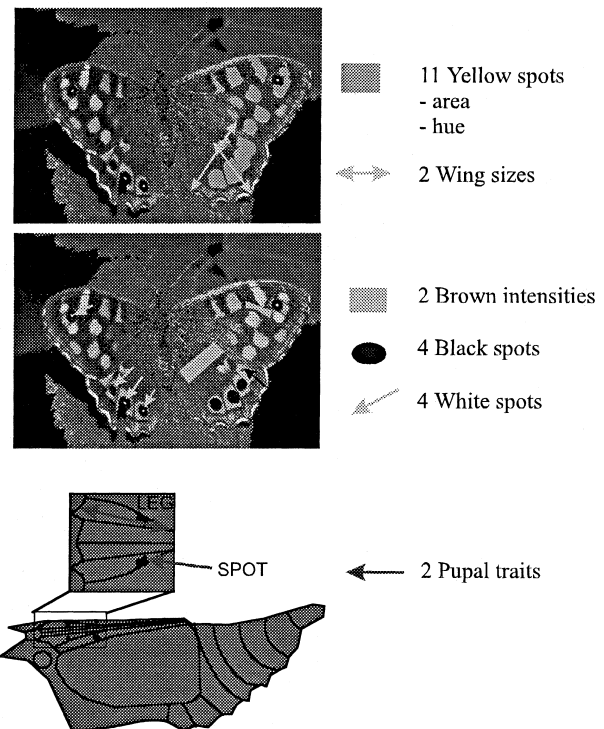


Fig. A1 Traits measured for comparing relative size of FA.

Table A1 Results of repeated measures ANOVA. F indicates traits on the forewing, H on the hind wing. Degrees of freedom were 3 for individual, 99 for side and 1 for the interaction term. The MS for side was divided by the MS interaction to obtain the *F*-value for side.

Trait	Individual		Side		Individual*Side	
	<i>F</i>	<i>P</i>	<i>F</i>	<i>P</i>	<i>F</i>	<i>P</i>
Pupal traits						
Leg segment	129.40	****	2.15	0.148	4.23	****
Pupal spot	22.17	****	0.12	0.733	25.84	****
Adult traits						
White spot F	307.10	****	0.27	0.609	26.62	****
Yellow spot 1 F Area	252.97	****	0.10	0.010	5.59	****
Yellow spot 2 F Area	866.98	****	1.22	0.277	10.73	****
Yellow spot 3 F Area	545.50	****	10.67	0.003	13.25	****
Yellow spot 4 F Area	634.19	****	0.34	0.570	9.90	****
Yellow spot 5 F Area	1923.2	****	0.49	0.493	12.91	****
Yellow spot 6 F Area	385.21	****	0.08	0.772	3.84	****
Yellow spot 7 F Area	3293.7	****	1.02	0.319	6.81	****
Yellow spot 1 F Hue	190.76	****	7.29	0.011	9.08	****
Yellow spot 2 F Hue	469.62	****	1.91	0.176	20.70	****
Yellow spot 3 F Hue	314.21	****	1.62	0.211	5.88	****
Yellow spot 4 F Hue	503.12	****	3.87	0.057	9.90	****
Yellow spot 5 F Hue	342.62	****	2.02	0.164	12.91	****
Yellow spot 6 F Hue	395.00	****	2.90	0.098	3.84	****
Yellow spot 7 F Hue	163.42	****	6.46	0.016	6.81	****
Brown Intensity F	6.14	****	0.02	0.876	1.75	0.024*
Wing length H	43.34	****	0.00	0.927	4.52	****
Wing width H	22.01	****	0.05	0.825	5.90	****
Yellow spot 1 H Area	956.77	****	8.54	0.006	16.49	****
Yellow spot 2 H Area	511.71	****	3.93	0.055	8.27	****
Yellow spot 3 H Area	780.31	****	27.61	0.0001**	7.97	****
Yellow spot 4 H Area	266.56	****	12.57	0.001*	7.28	****
Yellow spot 1 H Hue	541.36	****	0.73	0.406	15.74	****
Yellow spot 2 H Hue	793.35	****	3.40	0.074	25.61	****
Yellow spot 3 H Hue	720.18	****	0.13	0.718	15.48	****
Yellow spot 4 H Hue	337.05	****	9.58	0.004	11.39	****
Black spot 1 H Area	3201.02	****	0.61	0.446	740.31	****
Black spot 2 H Area	2065.33	****	0.96	0.344	90.22	****
Black spot 3 H Area	1158.07	****	7.77	0.009	40.88	****
Black spot 4 H Area	1564.00	****	2.45	0.126	68.69	****
White spot 2 H Area	416.14	****	0.86	0.369	19.78	****
White spot 3 H Area	250.80	****	0.75	0.623	26.75	****
White spot 4 H Area	401.64	****	1.54	0.223	63.82	****
Brown Intensity H	97.45	****	3.90	0.056	2.22	0.002**

* $P < 0.05$ (after Bonferroni correction), ** $P < 0.01$, **** $P < 0.0001$. For all the traits the asymmetry was significantly larger than measurement error. Directional asymmetry was detected after the sequential Bonferroni correction for multiple tests for two traits.

orange to bright red, approximately from 25 to 50 on the scale mentioned above. Intensity of brown coloration was measured on a grey scale varying from black (= 0) to white (= 255). Results of measurements were directly sent to a spreadsheet and could not be seen at the time of measurement. Two pupal traits were measured prior to the emergence of butterflies, at a 50× magnification with a microscope fitted with a micrometer.

Forty-three of the 169 individuals measured were measured a second time. Second measurements were

performed on newly grabbed images to include all the sources of measurement error. To analyse whether asymmetry was significantly larger than measurement error and whether directional asymmetry was present, an ANOVA of repeated measures was performed, with individual, side and their interaction as factors. Significance of the interaction term indicates that the asymmetry is larger than the measurement error, significant of the side term that there is directional asymmetry present.