Journal of Negative Results
— Ecology & Evolutionary Biology —
vol. 9: 1–14 • Helsinki, 30 August 2012
ISSN 1459-4625
http://www.jnr-eeb.org/

Are smarter people (a bit) more symmetrical? A matter of how to adjust for publication bias?

Stefan Van Dongen

Group of Evolutionary Ecology, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium (stefan.vandongen@ua.ac.be)

Received 23 September 2011 accepted 5 June 2012

Variation in general mental ability (*g*) may be due to a general fitness factor, leading to the prediction that *g* should relate to indicators of fitness. One such an indicator may be fluctuating asymmetry (FA), the morphological outcome of developmental instability. But the general association between FA and fitness components has been debated for several decades. Therefore, the meta-analysis by Banks *et al*. (2010) on the association between FA and intelligence is very timely and relevant. However, I argue that Banks *et al*. (2010) did not take the opportunity to fully address the possible issue of publication bias thoroughly enough, providing no estimates of average effect sizes adjusting for publication bias. In doing so in this paper, I show that adjusting for publication bias leads to markedly lower average effect sizes that were in some cases no longer statistically significant (between –0.06 and 0.12 *vs*. 0.14). I emphasize though that, at present, the assumptions behind the use of the corrections for publication bias cannot be tested explicitly leading to the, perhaps, disappointing conclusion that in spite of the 14 samples across nearly 2000 individuals, it is at present impossible to come to any robust conclusions. The evidence in favour of publication bias presented here, however, arguably suggests that associations between FA and *g* are at best very weak, but may also be non-existent. Further research, based on larger samples in combination with unbiased reporting, is undoubtedly indispensable to come to robust conclusions about the association between FA and intelligence.

Introduction

In a recent paper, Banks *et al*. (2010) present a meta-analysis of the association between fluctuating asymmetry and general mental ability (*g*) or intelligence, concluding that on average a weak but statistically significant negative relationship exists. Consequently, the authors conclude that smarter people are on average a bit more symmetrical, which is in line with the idea that reduced intelligence may be associated with lower fitness and (the increased expression of) developmental disturbances or developmental instability (DI). The most common measure of DI is subtle morphological asymmetry, called fluctuating asymmetry (FA), of which the general robust link with fitness is often debated. Van Dongen and Gangestad (2011), for example, performed a large scale meta-analysis of the relationship between FA and fitness and health in humans, concluding that in spite of an on-average robust significant negative correlation, the strength of the variation varied strongly among studies. An important factor contributing to this

Editor in charge of this manuscript: Bob O'Hara

variation was publication bias, and Van Dongen and Gangestad (2011) stressed that — albeit not easy to achieve — it is important to study and understand the influence of this publication bias to come to meaningful conclusions. Although not always statistically significantly so, direct comparisons of effect sizes from published and unpublished studies all point into the direction of stronger effects in published studies, where effect sizes were between 40% and 130% smaller in the unpublished studies (Møller *et al*. 2005, Banks *et al*. 2010). In addition, Van Dongen (2011) showed that associations between FA and attractiveness were smaller if they were reported as a side result and not as the main hypothesis of the paper. It is also important to realize that the exact impact of publication bias is very difficult to assess and depends on assumptions that are hard to test, especially if indirect methods are applied. Therefore, it is crucial to provide a kind of sensitivity analysis approaching the problem from difficult angles to provide a range of possible outcomes (Van Dongen & Gangestad 2011). In that way, it is possible to calculate adjusted effect sizes in an attempt to correct for publication bias.

Meta-analyses are commonly used to summarize the available information from different sources, where the strength of association is expressed as an effect size (often a Pearson's correlation coefficient, which has a straightforward interpretation). The most crucial assumptions are that the different studies estimate the same biological phenomenon (the association between FA and intelligence in this case), and that the available effect sizes represent an unbiased sample of all possible effect sizes. It has been emphasized repeatedly that studies of FA in particular may be prone to publication bias, where negative/non-significant results are less likely to become published (e.g., Palmer 1999, 2000, Møller & Jennions 2001, Van Dongen 2011, Van Dongen & Gangestad 2011). It has also repeatedly been emphasized that the possible effects of publication bias need to be studied explicitly and that publication bias may affect conclusions profoundly (e.g., Jennions & Møller 2002a, b, Van Dongen 2011, Van Dongen & Gangestad 2011). Two common ways to detect problems of publication bias are to compare effect sizes between published and unpublished studies and to test for associations between sample size and effect size (i.e., constructing funnel plots). Banks *et al*. (2010) applied both methods and found evidence for publication bias since effect sizes were smaller in the unpublished studies, and published (but not unpublished) effect sizes decreased with sample sizes (*see* also below). Thus, the authors emphasized the presence of publication bias, yet, unfortunately did not thoroughly attempt to calculate adjusted effect sizes. Banks *et al*. (2010) based their conclusions on the available effect sizes because the 'funnel plot looked symmetrical'. However, in this way, they implicitly assume that they have included all unpublished results in their analysis as well, which may not be the case. Given that the effect sizes of the unpublished results were markedly lower and usually opposite to what was predicted by theory (*see* below), one or a few additional unpublished results may alter the outcome drastically.

In this paper I am not able to proof or disproof the conclusions by Banks *et al*. (2010). Yet, I found it a missed opportunity to explore effects of selective reporting and to compare different ways to adjust for it in a set of studies with similar expected strengths of association. Therefore, I present a reanalysis of their work to emphasize that, depending on how attempts are made to adjust for publication bias, different conclusions can be reached. More specifically I will present different estimates of effect sizes adjusting for publication bias and will conclude that a good understanding of the mechanism and magnitude of publication bias is required (and currently lacking) to draw firm conclusions. While the review by Banks *et al*. (2010) might be referred to in the literature as providing evidence for an association between FA and intelligence (despite the authors formulating their conclusions very carefully), I argue that at this point it is impossible to conclude whether there is an association between asymmetry and intelligence or not. In fact, at this point evidence is more in favour of weaker or no associations than those from Banks *et al*. (2010). More research is clearly needed. But most importantly, I emphasize the potential importance of publication bias and how it can alter conclusions profoundly.

Material and methods

Effect sizes were obtained from table 1 in Banks *et al*. (2010) and I analyzed them at the level of Pearson's correlations. All effect sizes were changed in sign, such that positive values were in line with the prediction that more intelligent people are more symmetrical. Thus, a positive average effect size is expected from theory (as suggested by Banks *et al*. 2010).

I then constructed a funnel plot and tested for an association between sample size and effect size in the published results. Average effect sizes were obtained for all estimates, the published and the unpublished estimates. Obviously, these analyses led to the conclusion that publication bias was likely (as emphasized by Banks *et al*. 2010).

Next, I attempt to investigate if there are reasons to assume that other missing/unpublished studies exist, except those retrieved by Banks *et al*. (2010). To this end, I simulated effect sizes for a range of sample sizes assuming an average effect size of 0.14 (the average observed across all available estimates). This pattern is then compared to both the published and the unpublished effect sizes to explore the likelihood that 0.14 is close to the true average effect size (estimate based on all available effect sizes, *see* also Banks *et al*. 2010). Next to this qualitative visual comparison of observed and simulated effect sizes, a more formal evaluation of the pattern will be performed. Based on a visual examination, I assumed that selective reporting was likely to be present for studies with sample size below 200 (*see* also below for further arguments). To find evidence that selective reporting biased the average effect size of 0.14 upward, and thus that other unpublished studies could exist, I compared the observed proportion of significant effect sizes with the power to detect a true effects size of 0.14. These levels of statistical power were calculated as the proportion of significant simulated effect sizes for $n < 200$ and $n \geq 200$. If the average effect size of 0.14 is biased upward due to more unpublished results, I expect a higher proportion of observed effect sizes to be significant than expected by chance (i.e., the power of the analysis) for $n < 200$, and the opposite for $n \geq 200$ (although the latter will be difficult to detect given the few studies in this category). If I find evidence for more missing studies, I will adjust average effect sizes for selective reporting in three ways, as outlined next.

The simplest way to correct for publication bias is to ignore the estimates based on small sample sizes. The effect of publication bias can be assumed to diminish with increasing sample sizes such that larger studies are more likely to yield unbiased effect sizes. Therefore, the association between effect size and sample size is expected to asymptote towards the unbiased average effect size. I therefore explored the association between sample size and effect size non-parametrically using a loess approach to visually inspect the point where the association levels off (*see* e.g., Van Dongen & Gangestad 2011). In addition, I calculated the average effect size of the largest studies only ($n > 200$, *see* below for arguments) assuming a random effects model (but a fixed effects model gave very similar results). It has been emphasized repeatedly that negative associations between effect sizes and sample sizes may reflect covariation between study quality and sample size. If, for example, smaller studies have measured FA more accurately (by measuring more traits), higher effects sizes are expected on average (but still with higher variation though). To explore this possibility, the robustness of the association between effect size and sample size was studied by adding possible moderator variables (i.e., number of traits, variation in IQ, year of publication and sex ratio, all available from Banks *et al*. 2010).

Publication bias can also result in a negative association between effect size and year of publication. This has often been observed (Jennions & Møller 2002a, Van Dongen & Gangestad 2011), and could most parsimoniously be attributed to '… a publication bias against non-significant or weaker findings' (Jennions & Møller 2002a). Simmons *et al*. (1999) attribute the association between year and effect size to a paradigm shift. If so, overall conclusions can only be drawn if there has been a sufficient time since the initial idea was proposed and published. One could also argue that on average, later studies may better reflect the true strength of association. Therefore,

I also explored the association between effect size and year of publication and provided an average effect size of the most recent published studies.

Jennions and Møller (2002b) emphasized that 'literature reviews [should] assess the robustness of their main conclusions by correcting for potential publication bias using the trim and fill method'. Therefore, I applied the trim and fill method to symmetrize the funnel graph and obtained a corrected average effect size (*see* Duval & Tweedie 2000 for details) again assuming a random effects model. However, it is important to note here that this approach may underestimate the number of missing studies when only few effect sizes are available (Duval & Tweedie 2000).

Finally, and because the trim and fill method may not accurately adjust for publication bias in such a small meta-analysis, I applied and extended a recently developed model-based approach to obtain an unbiased estimate of the average effect size. The trim and fill method may fail especially in relatively small meta-analyses because it imputes missing/unobserved unpublished effect sizes based on the available published effect sizes using the symmetry of the funnel plot as a criterion. The model-based approach attempts to model the process of publication bias by estimating an area in the funnel plot where estimates are not published because of their negative outcome. I based my model on the model presented by Formann (2008). In addition, I assumed that for effect sizes above a threshold sample size ($n > 200$ here as well, *see* below), the outcome would no longer affect the likelihood of publication. The model is based on a truncated normal distribution. The mean of this distribution ($\mu_{unb.}$) represents the unbiased estimate of the average effect size. The standard deviation of this normal distribution is the sampling variation of a correlation coefficient and is a function of the sample size: $\sigma = \sqrt{(1-r^2)/(n-2)}$. The truncation parameter α represents the level of truncation, or in other words the effect size below which result will not get published (if $n \leq 200$). Estimates were obtained in a Bayesian framework with weak uniform priors for the average effect size and α (between –1 and 1). I ran five independent MCMC's of 25 000 iterations and discarded the first 5000 as burn in.

Since the time of publication of the Banks *et al*. (2010) study, two of the unpublished effect sizes were published (Gangestad *et al*. 2010). Therefore, next to performing the analysis on the basis of the data as presented by Banks *et al*. (2010) I also provided average effect sizes and applied the trim and fill method to this dataset. This had no effect on the results from the model-based approach because the estimates in Gangestad *et al*. (2010) were reported as a side result, not the main objective of the paper. Van Dongen (2011) already showed that results reported 'on the side' had much lower effect sizes, and it is reasonable to assume that it will not affect the likelihood of getting published. Below I will refer to the unpublished studies as reported in Banks *et al*. (2010) and specifically mention separate analyses whenever the two published estimates in Gangestad *et al*. (2010) are referred to. All analyses were performed in the package Meta (version 1.6) in R version 2.10.1 (R Development Core Team 2009) and the model-based estimates were obtained in OPENBUGS (http://www.openbugs.info/w/).

Results

Average effect sizes and funnel plots

Average effect sizes across all available studies were significantly positive and also showed statistically significant heterogeneity among estimates (Table 1). The positive average effect size, however, was entirely due to the published estimates as, on average, the unpublished estimates showed a slightly negative average effect size that did not differ from zero (Table 1). Thus, there is a clear difference between published and unpublished results suggesting the presence of publication bias (*see* also Banks *et al*. 2010). The fact that two additional effect sizes were published recently did not change this overall result (Table 1).

The funnel plot of the association between effect size and sample size showed a strong negative association for the published studies ($r = -0.90$, df = 8, $p = 0.0005$), but not for the unpublished studies ($r = 0.70$, df = 2, $p = 0.30$). The

association differs significantly between published and unpublished studies (ANCOVA test of interaction: $F_{1,10} = 21.1$, $p = 0.001$). Remarkably, all published estimates from studies with sample size below 200 were statistically significant (except for the two estimates that were published recently by Gangestad *et al*. 2010), while only one was marginally so for the estimates from larger samples (Fig. 1). All unpublished results were not statistically significant. Clearly, and as acknowledged by Banks *et al*. (2010), there appears to be strong evidence of publication bias. The fact that the two effect sizes from Gangestad *et al*. (2010) were not statistically significant is not very surprising, as Gangestad and co-workers reported this result not as the main hypothesis of their paper (*see* also Van Dongen 2011), while it was the main hypothesis in all other published ones. Nevertheless, including them to test for the association between effect size and sample size resulted in a smaller and non-significant correlation ($r = –0.40$, df = 10, $p = 0.18$).

The random model showed significant heterogeneity among estimates (Table 1). This heterogeneity, however, was only significant in the published studies and appeared to be due to the studies with smaller sample sizes ($n < 200$, Table 1).

Is there evidence for more missing studies?

Figure 1 presents simulated effect sizes in the absence of publication bias (i.e., the symmetric funnel plot that would be observed if each study would have equal probability of being published). In addition, the observed effect sizes of published and unpublished studies were added. Note that all published effect sizes of studies with sample size below 200 were significant and above the average effect of 0.14, while the unpublished results and the published results from studies with $n \geq 200$ were below 0.14 and only one was marginally significant. With a true effect size of 0.14, the power is 26% for $n < 200$, and 60% for $n \geq 200$. So in the absence of publication bias, only 26% of the effect sizes are expected to be significant when $n < 200$, and 60% are expected to be significant if $n \geq 200$. The observed proportions were 70% and 25% respectively. The markedly higher observed number of significant results for $n < 200$ (the seven out of 10 significant effect sizes for $n < 200$ was significantly higher than the expected 26%, binomial test: $p = 0.004$; and this test was also significant after including the two recently published effect

Table 1. Average effect sizes (sample size) and their 95% confidence intervals based on different approaches and attempts to correct for publication bias (*see* text for details). Statistically significant averages are set in boldface. A test for heterogeneity is added. 'Published estimates 2' refers to the dataset where two additional estimates were published by Gangestad *et al.* (2010). Estimates in italics are considered to reflect estimates adjusted for publication bias (*see* text for details and arguments). *** $p < 0.001$.

	Average effect size [95% CI]	Heterogeneity
All estimates ($n = 14$):	**0.14** [**0.045** to **0.238**]	$Q_{13} =$ **53.0*****
All published estimates ($n = 10$):	**0.22** [**0.120** to **0.318**]	$Q_{9} =$ **31.5*****
All published estimates 2 ($n = 12$):	**0.17** [**0.08** to **0.27**]	$Q_{11} =$ **41.7*****
All unpublished estimates ($n = 4$):	–0.06 [–0.158 to 0.043]	$Q_{3} = 3.3$
All estimates before 2008 ($n = 8$):	**0.27** [**0.19** to **0.35**]	$Q_{7} = 11.2$
All estimates after 2007 ($n = 6$):	*–0.02* [*–0.09* to *0.05*]	$Q_{5} =$ *5.6*
All estimates where sample size > 200 ($n = 4$):	*0.03* [*–0.041* to *–0.109*]	$Q_{3} =$ *4.4*
All published + trim and fill ($n = 10 + 4$):	**0.12** [**0.013** to **0.232**]	$Q_{13} =$ **65.0*****
All published 2 + trim and fill ($n = 12 + 3$):	**0.11** [**0** to **0.212**]	$Q_{14} =$ **65.6*****
Model-based (threshold $n < 200$):	*0.04* [*–0.013* to *0.096*]	

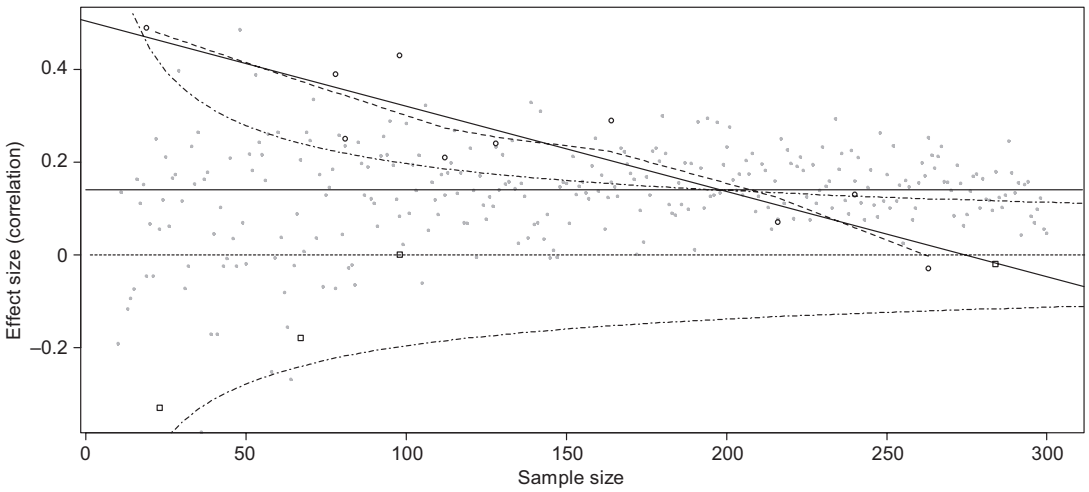

Fig. 1. Funnel plot of simulated and observed data. Data were simulated (grey dots) for a true average effect size of 0.14 (the estimated average effect size of all observed studies, published and unpublished). Published (open circles) and unpublished (open squares) effect sizes of the association between fluctuating asymmetry and intelligence were also added to the plot. The horizontal solid line represents the average effect size of 0.14, the dashed line that of no effect. The dashed-point lines represent the limits above which the Pearson correlation coefficients are statistically significant. The solid line represents the linear regression line of the association between published effect sizes and sample size, and the dashed line the loess curve. Note that the association between effect size and sample size does not level off at higher sample sizes, suggesting that the asymptote is not reached.

sizes by Gangestad *et al*. (2010): seven out of 12 successes with an expected success of 26%, binomial test: $p = 0.02$) suggests that there may be more unpublished results available than those retrieved by Banks *et al*. (2010). For the effect sizes of studies with $n \geq 200$, the discrepancy between the observed significance and that expected by power did not differ significantly (one out of four significant with a power of 60%: binomial test: $p = 0.31$), yet, only little data were available. Nevertheless, fewer significant results were observed than expected if the average effect size would be 0.14. Thus, the few unpublished studies available were probably not sufficient to result in unbiased estimates of average effect sizes, and that more unpublished material likely exists, except if for some reason smaller studies would have higher expected effect sizes.

Effects of moderator variables on the negative association between sample size and effect size

As indicated above, the negative association between sample size and effect size may be explained by the fact that smaller studies have higher expected effect sizes. To explore this alternative explanation a regression model was fitted with effect size as dependent variable and sample size as explanatory variable (slope = –0.18, SE = 0.03, $t_8 = -5.60$, $p = 0.0005$). In addition, number of traits, variation in IQ, sex ratio and year of publication were also added. None of these additional variables explained significant variation in effect sizes (all $p > 0.08$), and the effect of sample size was still statistically significant in the full model (slope = –0.15, SE = 0.04, $t_4 = -3.29$, $p = 0.03$). Thus, these moderator variables do not seem to cause the negative association between effect sizes and sample sizes. Associations of the moderator variables with sample size are depicted in Fig. 2. There appears to be a correlation between sample size and proportion of females in the sample, where the smallest studies only provided estimates for males (Spearman's rank correlation: $r = 0.67$, $p = 0.009$). However, the proportion of females in the sample was not correlated with effect size (Spearman's rank correlation: $r = 0.13$, $p = 0.65$; Fig. 3).

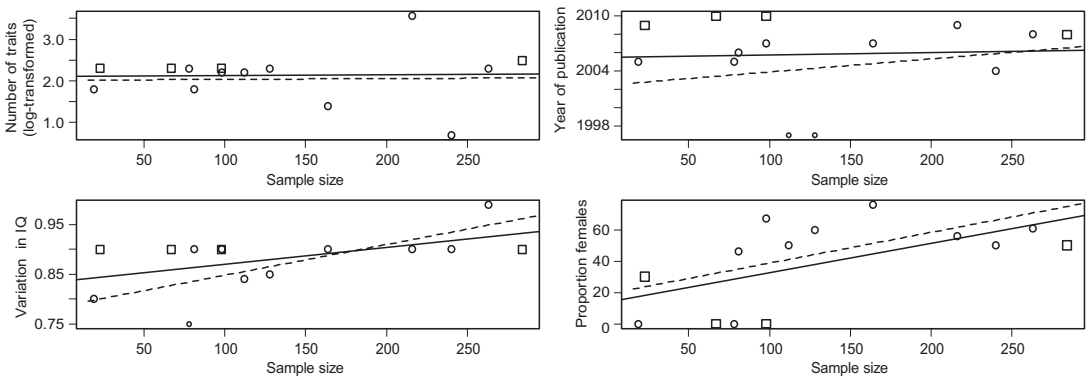

Fig. 2. Associations between sample size and possible moderator variables that could confound the association between sample size and effect size. Associations are given for published (circles and dashed line) and unpublished (squares and solid line) estimates.

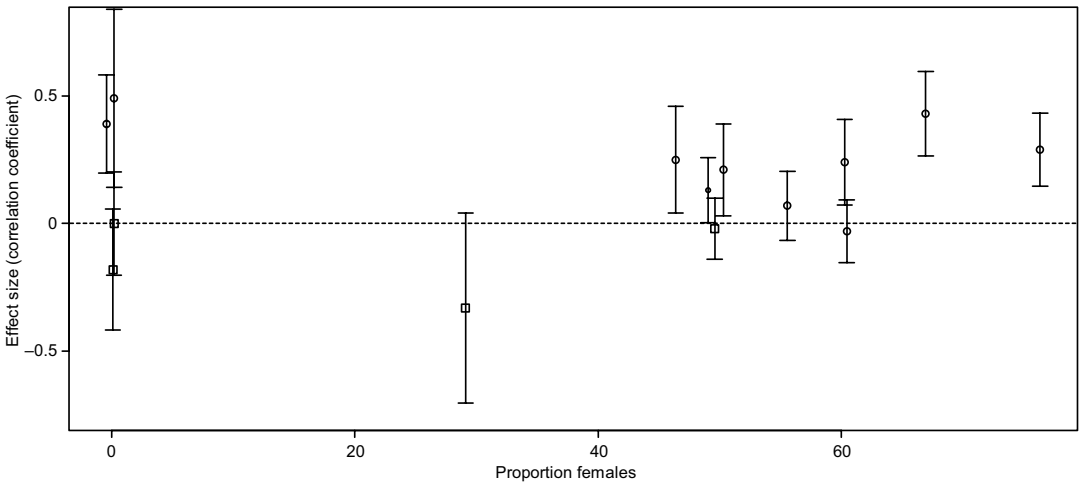

Fig. 3. Association between effect sizes and proportion of females in the samples. Published (circles) and unpublished (squares) effect sizes of the association between fluctuating asymmetry and intelligence (±95% confidence intervals).

Adjusted average effect sizes

As indicated above, there are several ways to adjust for this possible publication bias and some are applied and contrasted here. The association between effect size and sample size appeared nearly linear and did not level off. Indeed, the loess curve followed the linear regression line quite well (Fig. 1). It is thus impossible to determine the asymptote, but it seems reasonable to assume that the negative association between effect sizes and sample size did not level off for sample sizes below 200 (Fig. 4). Therefore, I assumed that above a sample size of 200, the likelihood of publication is independent of the observed outcome, and an unbiased estimate of the average effect size can be obtained by excluding all studies with $n < 200$. In fact, if the curve would level off at even higher sample sizes, larger studies ($n > 300$) would have significantly negative effect sizes, which seem very unlikely. Thus, the first adjusted average effect size, based on only four estimates (three published and one not), was now close to zero and no longer significant. The four studies did not show significant heterogeneity (Table 1).

Second, I explored the association between effect size and year as publication bias may cause negative studies to be published later (Jennions & Møller 2002a) or publication bias may

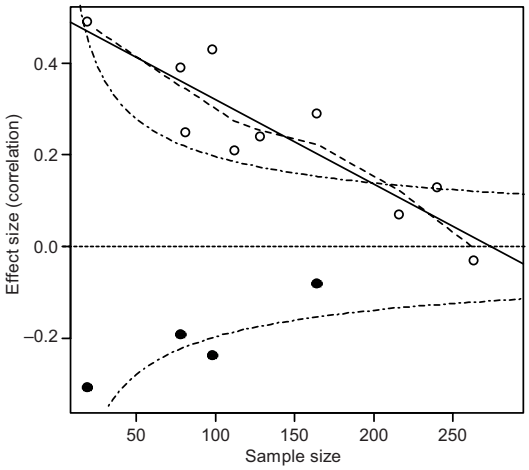

Fig. 4. Funnel plot of published (circles) effect sizes and effect sizes imputed by the trim and fill methods (dots) of the association between fluctuating asymmetry and intelligence. The solid line represents the linear regression line of the association between published effect sizes and sample size, and the dashed line the loess curve. The dashed-point lines represent the limits above which the Pearson correlation coefficients are statistically significant. Note that the funnel plot is now looking symmetric [as it did when plotting the published and unpublished effect sizes (Fig. 1)], but that effect sizes around the average effect size of 0.12 (*see* Table 1) are missing for the small studies.

be weaker some time after the publication of initial papers (Simmons *et al*. 1999, Van Dongen & Gangestad 2011). Effect sizes decreased with year (Spearman's rank correlation: $r = –0.63$, $p = 0.015$), and this decrease was not gradual (Fig. 5). All estimates before 2008 were positive and significant, while all estimates from 2008 onwards, were not statistically significant, nearly zero and not showing significant heterogeneity (Fig. 5 and Table 1).

Thirdly, I applied the trim and fill method to correct for possible publication bias. This method estimated that the four missing studies were sufficient to symmetrize the funnel plot (Fig. 4), leading to an average effect size of 0.12 which was significantly larger than zero and close to the estimate proposed by Banks *et al*. (2010). However, the pattern of the funnel plot was not as expected (*see* e.g., Fig. 1) and lacked effect sizes close to the average effect size for $n < 200$. Applying the trim and fill method to the published data including the two estimates from Gangestad *et al*. (2010) estimated three missing estimates and an average effect size of 0.11, which was only just significant (Table 1). The results of the trim and fill method may, however, be biased because of the low number of estimates available (Duval & Tweedie 2000).

Finally, the model-based approach estimated an area where results would not be published when $n < 200$ at the truncation parameter $\alpha = 0.20$ (95%CI = 0.18–0.21) (Fig. 6). The average effect size of this truncated normal distribution equalled 0.04 and did not significantly differ from zero (Table 1).

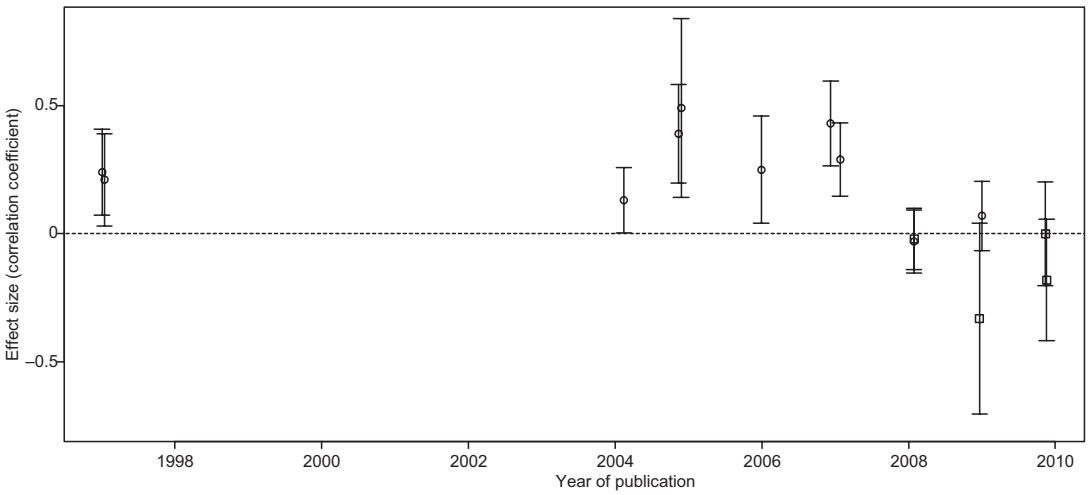

Fig. 5. Association between effect sizes and year publication. Published (circles) and unpublished (squares) effect sizes of the association between fluctuating asymmetry and intelligence (±95% confidence intervals).

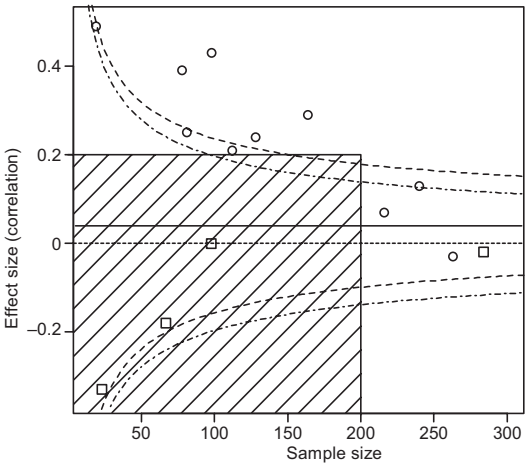

Fig. 6. Funnel plot of published (circles) and unpublished (squares) effect sizes of the association between fluctuating asymmetry and intelligence. The rectangle represents the area where effect sizes are not published because of the small sample size and non-significant outcome. The dashed-point lines represent the limits above which the Pearson correlation coefficients are statistically significant. The solid horizontal line represents the model-based average effect size of 0.04 (Table 1), and the dashed lines the 95% bands in which 95% of the estimates are expected to fall. Given that four out of 14 available estimates (i.e., 29%) fall outside these bands, this model suggests the occurrence of a substantial number of unpublished results.

Discussion

What is the evidence for publication bias in studies of fluctuating asymmetry in general?

Publication bias occurs when researchers, editors, reviewers or companies/institutes handle the reporting of positive results (i.e., statistically significant and/or in the expected direction) differently from negative results (i.e., statistically insignificant and/or opposite to what is commonly expected). This leads to bias in the overall literature, and there is evidence from the medical, social and biological sciences that publication bias does exist. As such, this should not be dramatic as long as it does not change the overall picture of the strength of association or effect sizes in a particular field. Thus, the central question is '*how important is publication bias in a particular field and does adjusting for it change the overall conclusion?*'. Unfortunately, and inherent to the problem, the studies that are not published are not readily available for analysis and adjustments, and one often has to rely on indirect approaches to adjust effect sizes for possible publication bias. This requires us to make additional assumptions, which are often difficult to test. In the field of the evolutionary biology of fluctuating asymmetry, the possible effects of publication bias were emphasized by Palmer (1999) using funnel plots. However, Thornhill *et al*. (1999), for example, suggested that negative associations between effect size and sample size could also emerge for other reasons. First, they correctly argue that sample size may co-vary with design. For example, experimental studies may inherently have smaller sample sizes *and* higher effect sizes. Yet, this explanation does not apply to the analysis here because all studies are observational. In addition, Van Dongen and Gangestad (2011) still found an — albeit weaker — negative association between effect size and sample size in studies with comparable design and expected outcome (including IQ). They concluded that, at least in part, the negative association between sample sizes and effect sizes in the human FA literature was due to publication bias. Second, Thornhill *et al*. (1999) argue that when seeing small effect sizes initially, researchers may increase sample sizes in subsequent studies 'realizing that null findings with small sample sizes are not highly informative', and that these larger studies will show on average smaller effect sizes. However, this argument implicitly assumes that the smaller studies are only published if significant and showing high effect sizes on average. This explanation may be correct, but implicitly assumes publication bias. Finally, study quality may co-vary with sample size, such that smaller studies have higher expected effect sizes, but I am unaware of any study supporting this idea. Thus, none of the abovementioned alternative explanations appear very likely here. In addition, several studies have suggested moderate to strong effects of publication bias in the area of FA research.

First, the best way to document publication bias is to directly compare effect sizes between published and unpublished results. Møller *et al*. (2005) reviewed studies on FA and sexual selec-

tion and in their most powerful test, the unpublished studies showed significant effect sizes that were about 40% smaller (table 3 in Møller *et al*. 2005). Although Møller *et al*. (2005) argued that the *p* value of 0.02 was not statistically significant after Bonferonni correction, no such correction is really needed because it is the main hypothesis tested (and with the highest power). Effect sizes of the unpublished studies in Banks *et al*. (2010) were even 130% smaller than the published ones.

Second, the relevance of a negative association between effect size and sample size to detect publication bias can be explored by comparing this association between estimates from studies where effects are put forward as the main hypothesis tested and studies where the estimate is presented as a side result (Van Dongen 2011). If effect sizes and significance affect the likelihood of being published (as predicted by the publication bias idea), this should primarily emerge if the result is the primary hypothesis of the paper. It can thus be predicted that, if publication bias is important, negative associations between effect sizes and sample sizes should only emerge if effects sizes reflect the primary hypothesis and not if effect sizes are reported as a side result (Van Dongen 2011). This is exactly what Van Dongen (2011) found for studies of the association between FA and attractiveness, again supporting publication bias.

Third, publication bias may also cause negative associations between year and effect size. Negative studies may take longer to get published (Jennions & Møller 2002a) and/or a so-called paradigm shift, where following a scientific revolution, a 'bandwagon effect' causes corroborative studies to be published relatively easily, while scepticism and improved methodology may lag behind, leading to an initial set of studies with high average effect sizes and a decrease in effect sizes after some time (i.e., a revolution of patterns).

Thus, publication bias seems very plausible in the area of fluctuating asymmetry, which has led to general conclusions about its importance and very specific recommendations by various authors. Jennions and Møller (2002a) conclude that '… publication bias, whatever the underlying cause, appears to be a problem in biology because both year of publication and sample size are correlated with effect size'; Jennions and Møller (2002b), end their abstract as 'We suggest that future literature reviews assess the robustness of their main conclusions by correcting for potential publication bias using the 'trim and fill' method'; and Simmons *et al*. (1999) caution about 'general conclusions from meta-analyses conducted before revolutions have settled'. These recommendations are followed in this paper.

Publication bias in studies of the association between fluctuating asymmetry and intelligence

I present a reanalysis of a recent meta-analysis on the association between fluctuating asymmetry (FA) and intelligence in humans (Banks *et al*. 2010). While the authors of the original analysis were thorough in the collection of data and its presentation, I would like to focus on the aspect of selective reporting, which, I show, may have caused substantial bias leading to different conclusions. I provide several lines of evidence which suggest publication bias. Importantly, these are all indirect effects, so it is not possible to actually prove that publication bias exists or to provide fully unbiased adjusted estimates. However, because several earlier studies suggest that publication bias may have important implications for overall conclusions and sensitivity analysis were generally advocated, I found this exploration relevant and timely for the association of FA with intelligence.

Effect sizes correlated negatively with sample size and year of publication, and unpublished studies showed weaker and non-significant effect sizes. In spite of the near symmetry of the funnel plot (Banks *et al*. 2010), visual inspection of the graph showed a lack of effect sizes around the average effect size (somewhere around 0.12–0.14). For the smaller sample sizes ($n < 200$), effect sizes were generally above this level, for the larger sample sizes ($n > 200$), they were below this average estimate from Banks *et al*. (2010). Indeed, the observed distribution of effect sizes did not match that expected to be generated by an overall effect size of 0.14

(and randomly missing effect sizes) (Fig. 1). In small studies ($n < 200$) more effect sizes were statistically significant than expected based on the power of the analyses. This suggests that more unpublished studies could exist. In fact, the required sample size to detect a true correlation of 0.14 with a power of 80% is 400, so none of the available studies had sufficient power to detect it. One could argue that the pattern is not a result of publication bias, but that for some unknown reason, smaller studies had higher average effect sizes (e.g., Thornhill *et al*. 1999). This would indeed explain the observed heterogeneity in effect sizes among published studies (Table 1). However, the association between effect size and sample size was absent for the unpublished studies (and the two published effect sizes which were reported as a side result by Gangestad *et al*. 2010). Thus, if smaller studies would indeed have larger expected effect sizes, one would have to come up with a reason why this is not the case for unpublished results (*see* also Van Dongen 2011 for similar arguments). Furthermore, several moderator variables did not explain the negative association between effect size and sample size in the published studies. Thus, no evidence for this alternative explanation could be found here and the presence of publication bias in the study of the association between intelligence and FA appears likely.

Since Banks *et al*. (2010) published their analysis, two effect sizes indicated as unpublished, were published by Gangestad *et al*. (2010). As a consequence, the association between effect size and sample size became weaker and not statistically significant after including these two estimates as published. However, the two estimates in Gangestad *et al*. (2010) were reported as a side result, while all other published estimates were reported as main results. Van Dongen (2011) showed in the context of associations between FA and attractiveness, that if estimates are published as a side result, they show on average weaker effects and are independent of sample size. Consequently, Van Dongen (2011) concluded that such results reported 'on the side' are likely to be unaffected by publication bias. Thus, it is likely that the estimates in Gangestad *et al*. (2010) are also unaffected by publication bias. Thus, no association between effect size and sample size is expected and the fact that their estimates got published in the meanwhile does not reduce the evidence for publication bias. Furthermore, while the association between effect size and sample size was no longer significant after including the two estimates from Gangestad *et al*. (2010) as published, the second line of evidence for publication bias remained significant. Indeed, the proportion of significant effect sizes was also higher than expected on the basis of statistical power after including the two previously unpublished effect sizes. So at least in terms of statistical significance, this result appears robust.

Adjusted effect sizes

All applied methods to detect publication bias appear to support its presence. That leaves us with the question: has publication bias led to erroneous conclusions? The trim and fill method was unable to reconstruct the expected distribution, because all published effect sizes based on a relatively small sample ($n < 200$) were unrealistically high (i.e., all above 0.2 and statistically significant, while power was only 26% assuming that all studies measure the same effect). A simulation study by Duval and Tweedie (2000) showed that when few (< 25) effect sizes are available, the estimated number of missing studies is biased downwards. It can therefore be argued that in this particular case, simply relying on the symmetry of the funnel plots did not lead to an appropriate/unbiased estimate of the average effect size.

Since the trim and fill method may not provide an adequate estimate of the number of missing studies here (due to the small number of available effect sizes), it can be argued that the average effect sizes based on the larger studies only ($n \geq 200$), the model-based approach might more closely reflect the true effect size. In each case, the estimate was small ($r < 0.05$) and not statistically significant. In addition, the fact that the most recent estimates were all close to zero, possibly after the revolution following a paradigm shift (Simmons *et al*. 1999), suggests low average effect sizes. Thus, publication bias

may very well have affected the main conclusion of this study, as it appears to do in at least 15%–21% of other meta-analyses (Jennions & Møller 2002b).

It remains difficult to assess the actual number of unpublished results, especially in this relatively small study. It can be expected that the number is not huge, given that relatively few research groups are studying intelligence and FA. Nevertheless, only few additional missing studies could have profound effects on the average effect size, since overall only 14 estimates are available. I do want to reemphasize here that there is no guarantee or solid proof that my adjusted estimates are truly unbiased Nevertheless, they are likely to reflect 'the truth' more closely, assuming, as argued here, that the skew in the funnel plot is due to publication bias. The main objective of this exploration is to look at the data from different angles and this led to the conclusion that any attempt to adjust estimates for publication bias, in spite of their inherent untestable assumptions, causes association between FA and intelligence to disappear.

What associations could one expect to find?

All in all, one can ask the question what strength of association one would expect to emerge between FA and general intelligence if they would indeed exist. As Banks *et al*. (2010) noted, all research so far has been performed in Western societies where variation in both fitness (due to genetics, malnutrition, severe illness, etc.) and general intelligence is likely to be relatively small. While Van Dongen and Gangestad (2011) estimated that on average the association between developmental instability (DI) and a wide range of measures of health and quality in humans may be around 0.30 (but the uncertainty of this estimate is still very wide), one could argue that the expected association between DI and intelligence would be somewhat lower (for example, the estimate of Van Dongen & Gangestad 2011 also includes studies of severe and lethal congenital abnormalities in fetuses), perhaps 0.20. In addition, it is well known that FA only weakly reflects DI, which would reduce this correlation in DI by another factor of 2 to 3 (*see* also Van Dongen & Gangestad 2011). Thus, at best, if any, the expected correlation between FA and general intelligence cannot be expected to be higher than 0.05–0.1. Thus, my, arguably, unbiased estimate of –0.02, 0.03 and 0.04 may in fact be quite accurate and closer to expectation. However, at this point no study is large enough to detect this. The required sample size to detect a correlation of 0.1 with a power of 80% is 620, and to detect a correlation of 0.05 it is 3150, a size that is difficult to reach in empirical studies. Clearly, meta-analyses are very well suited to combine evidence from different studies, but the biased reporting as suggested in this study, hamper its application and interpretation of the outcome.

Next to the fact that FA only weakly reflects developmental instability, its magnitude is usually small and FA can become easily confounded with measurement error and directional asymmetry. On the one hand, most studies of FA report repeated measurements and traits that are measured accurately enough. On the other hand, directional asymmetries in extremities of vertebrates have been suggested to be of environmental origin and are invoked to demonstrate behavioural lateralization (e.g., Galatius & Jespersen 2005) and thus asymmetry may not reflect DI. In humans with extreme lateralization of behaviour, handedness as a cause of morphological asymmetries has been suggested repeatedly and studied intensively for over a century (e.g., Plato *et al*. 1980, Purves *et al*. 1994, Roy *et al*. 1994). Because skeletal elements are remodelled during development (Lang *et al*. 2006, Auerbach & Ruff 2006), asymmetrical loading of limbs likely causes morphological asymmetries (Krahl *et al*. 1994, Pettersson *et al*. 2000, Auerbach & Ruff 2006, Kanchan *et al*. 2008). Cranial or facial asymmetries, in turn, may originate from lateralizations in behaviour and the brain (Smith 1907, LeMay 1977, Steele 2000, Kizilkaya *et al*. 2006). Furthermore, in humans, directional asymmetry appears to increase with age, possible due to sustained mechanic loading (Blackburn 2011), DA of the upper extremities increases with years of heavy working (Ozener 2010), DA appears larger in upper extremities (Sarringhaus *et al*. 2005) and handedness cor-

related with hand asymmetry (Van Dongen *et al*. 2009), all emphasizing the importance of environmentally determined DA. While upper limbs often show a right biased asymmetry — consistent with a majority of humans being right-handed — often, but to a lesser extent, the legs show DA in the opposite direction, thought to be a reflection of a compensatory action of legs in right-handed individuals. This so-called cross-asymmetry is also observed at the individual level through negative correlations in asymmetry between bones of the upper and lower extremities (Van Dongen *et al*. 2010). On the other hand, some have found directional asymmetry in foetal limbs free of mechanic loading (although lateralization in movements also occur in human foetuses) suggesting a pre-adaptation to handedness during adult life (but results are mixed; *see* Van Dongen *et al*. 2010). Although further research is clearly needed, there is quite some evidence accumulating recently that behavioural lateralization affects asymmetry in extremities, which would complicate the interpretation of patterns in FA even more as the observable asymmetry does not (entirely) reflect DI.

Concluding remarks

In summary, I found evidence for the existence of additional unknown missing studies, assuming, as argued here, that the skew in the funnel plot is due to publication bias. However, the actual number is very difficult to estimate. Adjusting for this publication bias led to much lower and non-significant average effect sizes compared to the average of 0.14 as put forward by Banks *et al*. (2010). Adjustments for publication bias rely on strong assumptions behind the mechanism and importance of publication bias. Unfortunately, and as is often the case, these assumptions are difficult to test. The strong association between sample size and effect size (r = –0.9, Fig. 4) and the marked lower effect sizes in more recent studies and studies with higher sample sizes, suggests that it is unlikely to be a by-product of another factor that co-varies with sample size. Furthermore, the negative association between effect size and sample size was not present in the unpublished studies. It, therefore, seems very premature to conclude that there exists a robust association between DI and intelligence. I, therefore, suggest that future studies should assure sufficiently high sample sizes and number of traits measure, recommendations also provided by Banks *et al*. (2010), among many other useful suggestions. Until then, I can only conclude that at present there is no strong evidence that smarter people are a bit more symmetrical, but more research may prove otherwise.

References

Auerbach, B. M. & Ruff, C. B. 2006: Limb bone bilateral asymmetry: variability and commonality among modern humans. — *Journal of Human Evolution* 50: 203–218.

Banks, G. C., Batchelor, J. H. & McDaniel, M. A. 2010: Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. — *Intelligence* 38: 393–401.

Blackburn, A. 2011: Bilateral asymmetry of the humerus during growth and development. — *American journal of physical anthropology* 145: 639–646.

Duval, S. & Tweedie, R. 2000: A non-parametric "Trim and Fill" method of accounting for publication bias in meta-analysis. — *Journal of the American Statistical Association* 95: 89–98.

Formann, A. K. 2008: Estimating the proportion of studies missing for meta-analysis due to publication bias. — *Contemporary Clinical Trials* 29: 732–739.

Galatius, A. & Jespersen A. 2005: Bilateral directional asymmetry of the appendicular skeleton of the harbor porpoise (*Phocoena phocoena*). — *Marine Mammal Science* 21: 401–410.

Gangestad, S. W., Thornhill, R. & Garver-Apgar, C. E. 2010: Men's facial masculinity predicts changes in their female partners' sexual interest across the ovulatory cycle, whereas men's intelligence does not. — *Evolution and Human Behavior* 31: 412–424.

Jennions, M. D. & Møller, A. P. 2002a: Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. — *Proceedings of the Royal Society London Series B* 269: 43–48.

Jennions, M. D. & Møller, A. P. 2002b: Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. — *Biological reviews* 77: 211–222.

Kanchan, T., Kumar, M., Kumar, P. & Yoganarasimha, K. 2008: Skeletal asymmetry. — *Journal of Forensic and Legal Medicine* 15: 177–179.

Kizilkaya, E., Kantarci, M., Basekim, C. C., Mutlu, H. & Karaman, B. 2006: Asymmetry of the height of the ethmoid roof in relationship to handedness. — *Laterality* 11: 297–303.

Krahl, H., Michaelis, U., Pieper, H., Quack, G. & Montag,

M. 1994: Stimulation of bone growth through sports. — *American Journal of Sports Medicine* 22: 751–757.

Lang, T. F., Leblanc, A. D., Evans, H. J. & Lu, Y. 2006: Adaptation of the proximal femur to skeletal reloading after long-duration spaceflight. — *Journal of Bone and Mineral Research* 21: 1224–1230.

LeMay M. 1977: Asymmetries of the skull and handedness. Phrenology revisited. — *Journal of the Neurological Sciences* 32: 243–253.

Møller, A. P. & Jennions, M. D. 2001: Testing and adjusting for publication bias. — *Trends in Ecology and Evolution* 16: 580–586.

Møller, A. P., Thornhill, R. & Gangestad, S. W. 2005: Direct and indirect tests for publication bias: asymmetry and sexual selection. — *Animal behavior* 70: 497–506.

Ozener, B. 2010: Fluctuating and directional asymmetry in young human males: effect of heavy working condition and socioeconomic status. — *American Journal of Physical Anthropology* 143: 112–120.

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.

Pettersson, U., Alfredson, H., Nordstrom, P., Henriksson-Larsen, K. & Lorentzon, R. 2000: Bone mass in female cross-country skiers: relationship between muscle strength and different BMD sites. — *Calcified Tissue International* 67: 199–206.

Plato, C. C., Wood, J. L. & Norris, A. H. 1980: Bilateral asymmetry in bone measurements of the hand and lateral hand dominance. — *American Journal of Physical Anthropology* 52: 27–31.

Purves, D., White, L. E. & Andrews, T. J. 1994: Manual asymmetry and handedness. — *Proceedings of the National Academy of Sciences USA* 91: 5030–5032.

R Development Core Team 2011: R: *A language and environment for statistical computing*. — R Foundation for Statistical Computing, Vienna, Austria [URL http://www.R-project.org/].

Roy, T. A., Ruff, C. B. & Plato, C. C. 1994: Hand dominance and bilateral asymmetry in the structure of the second metacarpal. — *American Journal of Physical Anthropology* 94: 203–211.

Sarringhaus, L. A., Stock J. T., Marchant L. F. & McGrew W. C. 2005: Bilateral asymmetry in the limb bones of the chimpanzee (*Pan troglodytes*). — *American Journal of Physical Anthropology* 128: 840–845.

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

Smith, G. E. 1907: Asymmetry of the brain and skull. — *Journal of Anatomy and Physiology* 41: 236.

Steele, J. 2000: Handedness in past human populations: skeletal markers. — *Laterality* 5: 193–220.

Thornhill, R., Møller, A. P. & Gangestad, S. W. 1999: The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. — *American Naturalist* 154: 234–241.

Van Dongen, S. 2011: Associations between asymmetry and human attractiveness: possible direct effects of asymmetry and signatures of publication bias. — *Annals of Human Biology* 38: 317–323.

Van Dongen, S., Cornille, R. & Lens, L. 2009: Sex and asymmetry in humans: what is the role of developmental instability? — *Journal of Evolutionary Biology* 22: 612–622.

Van Dongen, S. & Gangestad, S. W. 2011: Human fluctuating asymmetry in relation to health and quality: a meta-analysis. — *Evolution and Human Behavior* 32: 380–398.

Van Dongen, S., Wijnaendts, L. C. D., ten Broek, C. M. A. & Galis, F. 2010: Human fetuses and limb asymmetry: No evidence for directional asymmetry and support for fluctuating asymmetry as a measure of developmental instability. — *Animal Biology* 60: 169–182.