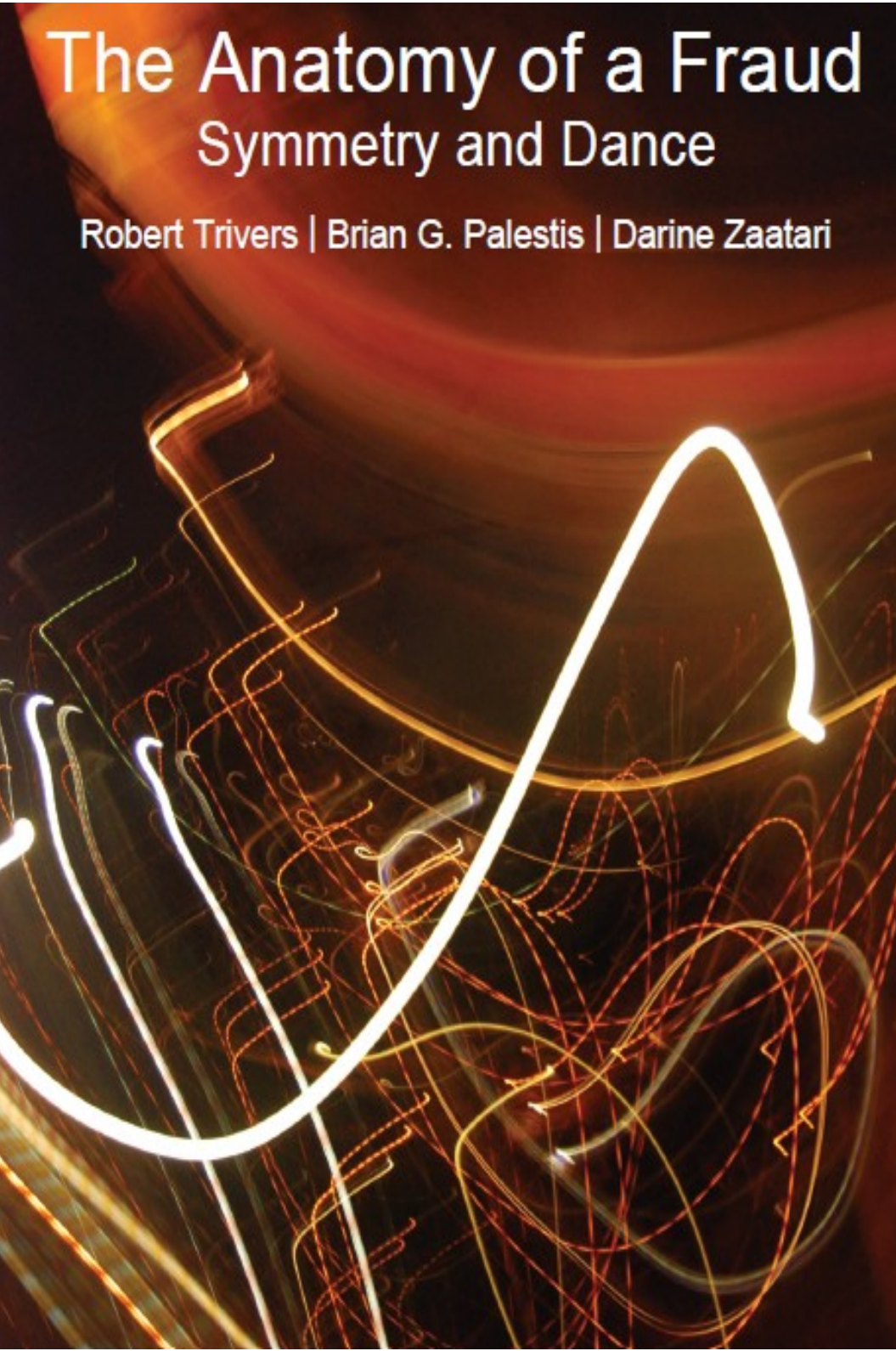
The Anatomy of a Fraud

Symmetry and Dance

Robert Trivers | Brian G. Palestis | Darine Zaatari



The Anatomy of a Fraud

The Anatomy of a Fraud

Symmetry and Dance

Robert Trivers

Anthropology & Biological Sciences, Rutgers University

Brian G. Palestis

Biological Sciences, Wagner College

Darine Zaatari

Antioch, California

TPZ Publishers
437 Brookside Ct
Antioch, Ca 94509

Manufactured in the United States of America

ISBN-13 978-0-615-28756-0
ISBN-10 0615287565

THE ANATOMY OF A FRAUD

CONTENTS

ABSTRACT

A thorough reanalysis of Brown et al. (2005) “Dance reveals symmetry especially in young men” shows that all of the major results appear to be based on hidden procedures designed to produce the results later derived. These procedures include the preselection of animations of Jamaicans dancing, apparently based on preliminary evaluation in New Jersey, so as to exclude symmetrical individuals who danced poorly and asymmetrical ones who danced well ($N = 10$ out of 10, $P < 0.001$). There are also systematic biases in averaging dance evaluations so as to produce significant results where none exist and more highly significant ones than do, in fact, exist. This appears primarily to have been achieved by reducing the variance in within-group dance evaluations thus making between-group comparisons more significant. How this reduction was achieved is obscure to us, as is the source of other biases in the data analysis, but all show the common pattern of making the evidence appear to be more striking than it really is. Using the same fluctuating asymmetry (FA) values used in Brown et al. one set of correlations is confirmed nearly exactly, namely, the sex difference in importance placed on symmetry in dance evaluations. This was a between-evaluator analysis that relied on the same grid of values used in the other analyses. This makes it all the stranger that the two sets of average dance evaluations do not match up. In addition, the significant negative correlation between male fluctuating asym-

metry and preference for the animations of relatively symmetrical females also disappears (even though this is also a between-evaluator analysis).

After conducting these analyses we were astonished to discover an additional, major source of bias in Brown et al. Values of FA were modified for 65 out of 80 cases of the dancers chosen (1996 and 2002 FAs combined) so as to place good dancers in the symmetrical category and poor dancers in the asymmetrical one. Meanwhile values for individuals not selected as dancers remain (with one exception) unchanged. Since the incorrect values were used in the between-evaluator comparisons of males and females, there is no way now to confirm these findings. The probability that all of these biases could have resulted from chance is well less than 1 in 10,000,000,000. Thus Brown et al.'s results appear to be entirely artificially constructed—that is, fraudulent. An analysis of the full data of (~2) Rutgers University evaluators, an unbiased set of data, reveals at best a weak positive relationship between symmetry and dancing ability, with no sex difference, using 2002 FA values only.

Finally we turn to an analysis of some of the factors that may have contributed to the fraud, especially decisions taken by Dr Trivers. We discuss the role that *Nature* played and we summarize the reactions of the other co-authors of Brown et al. (2005) to the discoveries described in this book. More generally, we mention some of the factors that reduce the chance that fraud will be discovered or (if discovered) revealed. We mention some recent cases and end with a summary of Fisher's famous reanalysis of Mendel's genetics work.

1. INTRODUCTION

In 2005 Brown et al. appeared to show a remarkable series of findings regarding dancing ability, sexual selection and fluctuating asymmetry (FA) in humans. Motion capture technology permitted a pure extraction of the phenotype of the dance from that of the dancer in a natural population of Jamaicans studied separately for degree of bodily asymmetry both in 1996 and 2002. The subjects were measured twice independently each time, the first time by some of the world's experts in measuring FA (Trivers et al. 1999) and the second time by trained graduate students from Rutgers. Brown et al. (2005) showed that more symmetrical individuals of both sexes were better dancers, but the effect was stronger for men than for women (Figure 1). Women, in turn, chose as good dancers individuals who were relatively more symmetrical compared to similar choice by men. Finally, more asymmetrical men tended to prefer the dances of relatively asymmetrical women. Ef-

fect sizes were strong and effects usually highly significant. As far as could be discerned, people were acting more or less exactly as predicted by Trivers (1972)—that is, dance revealed biological quality in both sexes, but more so in males than females, while females were more discriminating in choice.

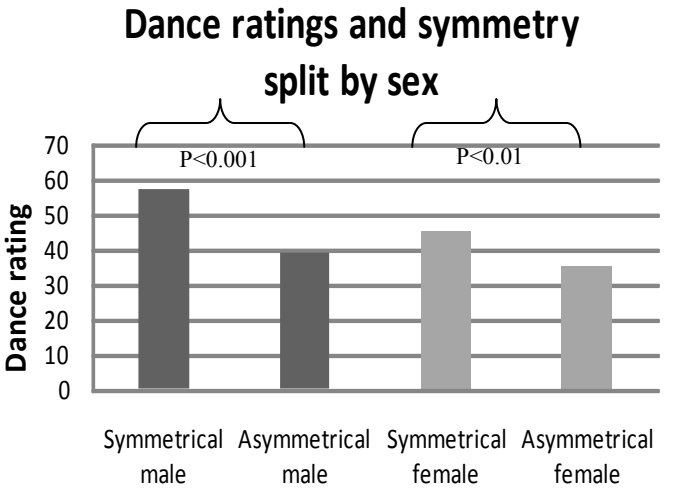


Figure 1. Mean dance ability of males and females by level of bodily symmetry (adapted from Brown et al. 2005).

The work also underscored the value of fluctuating asymmetry as a measure of biological quality, in particular, development stability—the ability of the genes, especially in the face of early stress, to create the phenotype they are aiming for. FA has a series

positive correlations across many species with such important variables as survival, speed, strength, resistance to parasites and physical attractiveness (Møller and Swaddle 1997, Gangestad et al 2001; Møller et al 2005). It is especially the association with sexual selection that led Brown et al. (2005) to expect positive associations between degree of bodily symmetry and dancing ability in Jamaica, a society in which such ability is strongly valued.

In what follows, we (who include one of the coauthors on Brown et al. 2005) will show that a series of biased procedures was introduced throughout the analysis of the data, apparently designed to achieve the striking set of results they apparently achieved. Three manipulations appear to have been used. One was to pre-select the sample so as to produce a prior association between symmetry and dancing ability. The second was apparently more complex, consisting of averaging the ~160 Jamaican evaluations per dancer so as to produce near-correct results which nevertheless showed less variance within groups (e.g. symmetrical females) than shown in our reanalysis and thus greater chance of finding significant between-group differences (e.g. between symmetrical and asymmetrical females). The third was only discovered when the rest of our analyses were completed: FA values of danc-

ers chosen were systematically altered to as to place good dancers in the symmetrical category and poor ones in the asymmetrical category, while FA values of dancers not chosen were (with one exception) never changed. It should be noted that (1) all statistical analyses were performed by Dr Brown and (2) he (as we later realized) used a novel set of FA values apparently generated by himself. We know of no deficiencies in the motion capture animations nor in the original dataset on fluctuating asymmetry that would produce the patterns we note below.

2. PROBLEMS WITH BROWN ET AL. (2005)

We first realized we had a problem when a student pointed out to us that some dancers were in fact outside the upper or lower FA tercile for which they were classified. The student had been taking a statistics course and thought to do an assignment using data on dance and FA from the Jamaican Symmetry Project. He was unable to replicate the statistics in Brown et al. (2005), and in the process noticed that some individuals appeared to be misclassified. We were performing similar analyses on related research and also noted inconsistencies in the data (or their analysis) that Dr. Brown appeared unable or unwilling to resolve. Even using the same SPSS files he said he used, we rarely got precisely the same statistical values, often differing in only minor details, but in some cases the discrepancies were large. He often responded that he was on the road away from his office, thought the differences might be due to differing statistical programs giving different results (e.g. SPSS vs

Excel) and that he would get back to us when he returned, which he typically did not do.

In addition, when analyzing dance ratings for a new group of dancers selected for high and low 2^{nd}:4^{th} digit ratio, we found results contradictory to those reported in Brown et al., at least regarding FA and dance ability (we find no correlation and a trend opposite the predicted direction), correlations between BMI and dance ability (a significant negative correlation, rather than no correlation), sex differences in dance ability (females rated higher than males, rather than males higher than females), and ability to recognize the sex of dancers in the motion-capture animations (71% correct identifications, rather than 62%). Although these discrepancies may, in part, result from the smaller range in FA of the dancers or greater age of the new evaluators, they also encouraged us to examine the results reported in Brown et al. more closely.

We were not reassured when other scientists such as Dr. Yanxi Liu and her student Mr. Seungkyu Lee (at Penn State) failed to replicate in many minor details the statistics found in Brown et al. and also discovered two notable errors, both in the "wrong" direction, that is, making the results look better than, in fact, they were:

$P < 0.05$ is the correct value for the comparison between symmetrical and asymmetrical females while $P < 0.01$ was the published value; similarly $P < 0.05$ is correct instead of $P < 0.005$ for symmetrical males compared to symmetrical females (see “Reanalysis of the Results”, below). Again, they got Dr. Brown’s usual response: he was on the road, different statistical programs often produced differing results, he would get back to them from his office, which he did not do for some time in spite of repeated follow-up requests. Eventually he replied that $P < 0.005$ may have been a typo. Given these disquieting events, we decided to undertake a full reanalysis of Brown et al (2005) using the computer files he sent us, including, where possible, repeating critical measures ourselves. Finally we thought to check his FA values against those in our master file.

3. WERE DANCERS CHOSEN RANDOMLY WITH REFERENCE TO DANCING ABILITY AND FA?

We concentrated initially on the way in which the 40 videos of dancers were chosen from the original sample of 167 (the actual sample size is smaller: N = 106, see below). The agreed upon criterion for Brown et al (2005) was that 10 males must have been in the upper 1/3rd of the asymmetry distribution (FA) in both 1996 and 2002, 10 males must have been in the lower 1/3rd in 1996 and 2002, and the same thing must hold true for the females chosen. This was not stated in the paper itself which said only that 20 were in the top third each time and 20 in the bottom—i.e. not split further by sex—but it is clear that the data would have been much more artificial if not initially split by sex because of an imbalance in the sex ratio (68 males, 38 females**)** and because females had slightly but significantly higher FA values than males in 1996. Also, it is clear from the equal sample sizes in Brown et al. that the data were split by sex.

If FA values were uncorrelated between the two periods, then one would expect exactly 10 individuals to qualify in the four categories out of an initial sample of 180. The actual sample is much smaller (N = 106). A list of 162 animations was supplied to us by Dr. Lee Cronk. In our total sample, 167 individuals were measured in both 1996 and 2002 but of these only 106 were also filmed for animations. Using the values in Dr. Brown's dataset, FA measures between the two periods are correlated (Pearson's $r = 0.402$, $P < 0.0001$) so that there were sufficient numbers to meet the stated criteria in all but one case, namely 11, 13, 9 and 16 (male symmetrical, then asymmetrical; female symmetrical, then asymmetrical). However, as discussed below (see "FA values of dancers were systematically altered") the FA values for many of the dancers in Dr. Brown's file are incorrect. Using the true values for FA there is actually no correlation between 1996 and 2002 FA ($r = 0.085$, $P = 0.275$). In fact, many of the dancers were not actually eligible for selection. Even using Dr. Brown's own numbers, we show in this section that the selection of dancers was performed in a biased manner.

The fact that one group contained fewer than 10 eligible dancers, and thus a subject needed to be added (who just missed meeting the stated criteria) was not mentioned in the paper, but

here we had already noted a more disturbing fact—in 2 additional cases an individual was chosen from *outside* the stated distribution in place of available ones within. Why? And when there were more than 10 to choose amongst, were they chosen at random or based on some criterion, such as one that would support a particular viewpoint? Since Dr. Brown chose the original dancers—and did all subsequent statistical analysis—the question becomes what criteria was he using, if any, and why?

First, how could he know which dancers were better or worse—if this was the criterion he was employing—given that the dances had not been rated yet by Jamaicans? All 165 dances, it turns out, had already been rated by (usually two but sometimes only one) Rutgers undergraduates majoring in dance, and William Brown had been in charge of analyzing these data. The data themselves and the analysis were available before the 40 videos were taken to Jamaica for scoring by the Jamaican youngsters in March of 2005. Of course if the Rutgers scoring and the Jamaican scoring were uncorrelated, then no bias would be introduced, but for the 40 animations scored in both places their values are, in fact, highly correlated ($r = 0.743$, $P < 0.0001$ with Dr. Brown's data for Jamaican dance ratings, and $r = 0.715$, $P < 0.0001$ with recalculated dance

ratings – see "Did Dr Brown average the rater evaluations correctly?", below). This, incidentally, is reassuring because it suggests that widely disparate people will rate the relative dancing ability of our sample in similar ways.

To give the data an initial bias, all that had to be done was to use the Rutgers data to help pre-select his sample so as to create positive correlations in the predicted direction. This appears to have happened—in every single case where individuals met the criteria but were not chosen (N = 10), the dances excluded were above or below the median dancing ability in the unpredicted direction (Table 1). For example, symmetrical (low FA) males had a median dance rating by Rutgers students of 120.5; all three excluded had lower values. In short, asymmetrical individuals who happened to be good dancers were invariably removed, as were symmetrical ones who happened to be poor dancers. This fact alone is significant**;** the binomial probability gives an estimated overall probability of 0.00098 and more detailed statistical analyses within categories also show striking deviations (Table 2). In the case of asymmetrical males, for example, three individuals who were better dancers than all those retained were removed and two by very wide margins. Adding in the three individuals who were selected despite not

meeting the criteria for selection to the ten excluded despite meeting the criteria, 12 of 13 “decisions” were in a direction favorable to the hypothesis (binomial probability = 0.0016). Further evidence that dancers were selected based on the scoring by Rutgers students comes from the fact that the only two eligible dancers lacking Rutgers dance scores were excluded. Note that the simple binomial test is conservative since with each removal of an individual from one half of the distribution, there is one fewer from which to choose the next removal, so that the real probability of multiple removals from one side of a distribution only is lower still.

As illustrated in Table 1, it appears that dancers were initially selected by being in the top 10 or bottom 10 in FA for one of the two years, rather than in the top third in both years, and then the lists were slightly adjusted in a biased manner. For example, among asymmetrical females, 9 of the 10 most asymmetrical subjects from 1996 were selected. However, the fourth most asymmetrical subject was excluded, and this individual’s dance was rated very highly by the Rutgers students, much higher than all those included. Altogether, selection was biased in every single case in the same direction, that is, congenial to theory, with two categories showing significant effects in themselves (Table 2). With this meth-

Table 1. Dancers chosen and not chosen, along with prior Rutgers dance ratings

Category	Dancer ID	Selected?	FA Rank 1996	FA Rank 2002	Rutgers Dance Score	Median Rutgers Score
Symmetrical male			Low to High	Low to High		120.5
	55	Yes	8	**1**	129.575	
	162	Yes	2	**2**	119.55	
	117	Yes	middle 3rd	**would be 3**	110.75	
	185	Yes	5	**4**	123.125	
	197	Yes	9	**5**	120.5	
	152	Yes	7	**6**	121.4	
	203	Yes	4	**7**	127.4	
	200	Yes	10	**9**	115.5	
	182	Yes	3	**10**	121.375	
	23	Yes	middle 3rd	would be 12	138.8	
						Below Median?
	189	No	6	**8**	113.975	Yes
	178	No	11	11	113.85	Yes
	70	No	1	13	120.475	Yes

Table 1. Dancers chosen and not chosen, along with prior Rutgers dance ratings*, continued*

Category	Dancer ID	Selected?	FA Rank 1996	FA Rank 2002	Rutgers Dance Score	Median Rutgers Score
Asymmetri-cal male			High to low	High to low		99.375
	206	Yes	**1**	12	82.6	
	115	Yes	**2**	4	95.275	
	33	Yes	**3**	7	87.275	
	192	Yes	**4**	9	99.375	
	103	Yes	**6**	2	75.45	
	222	Yes	**7**	8	99.4	
	21	Yes	**8**	10	98.15	
	113	Yes	**9**	3	109.475	
	139	Yes	**10**	13	87.5	
	94	Yes	11	5	105.8	
						Above Median?
	1	No	**5**	6	117.225	Yes
	216	No	12	1	123.65	Yes
	217	No	13	11	113.25	Yes

Table 1. Dancers chosen and not chosen, along with prior Rutgers dance ratings, *continued*

Category	Dancer ID	Selected?	FA Rank 1996	FA Rank 2002	Rutgers Dance Score	Median Rutgers Score
Symmetrical Female			Low to high	Low to high		117.75
	38	Yes	1	**1**	109.8	
	89	Yes	middle 3rd	**would be 2**	118.975	
	30	Yes	3	**3**	100.575	
	287	Yes	9	**4**	126.625	
	15	Yes	5	**5**	123.925	
	239	Yes	4	**6**	116.525	
	229	Yes	6	**7**	121.125	
	194	Yes	2	**8**	102.55	
	68	Yes	8	**9**	135.5	
	86	Yes	7	**10**	104.475	

For clarity, rankings are based only on eligible dancers (i.e. in top or bottom third for both years and having a usable dance video), except in cases where ineligible dancers were selected by Dr. Brown. Note that it appears the selection was instead based on the top or bottom 10 in a particular year (indicated by rankings in bold) and that selection was consistently biased relative to evaluations by Rutgers dance students.

Table 1. Dancers chosen and not chosen, along with prior Rutgers dance ratings, *continued*

Category	Dancer ID	Selected?	FA Rank 1996	FA Rank 2002	Rutgers Dancer Score	Median Rutgers Score
Asymmetri-cal female			High to low	High to low		110.65
	119	Yes	**1**	12	91.25	
	67	Yes	**2**	16	100.725	
	235	Yes	**3**	5	113.525	
	110	Yes	**5**	9	112.975	
	34	Yes	**6**	6	89.45	
	195	Yes	**7**	7	113.775	
	75	Yes	**8**	8	73.925	
	205	Yes	**9**	3	108.35	
	175	Yes	**10**	1	67.875	
	63	Yes	13	2	107.575	
						Above median?
	215	No	**4**	11	130.875	Yes
	123	No	11	10	117.325	Yes
	47	No	12	14	not rated	?
	51	No	14	15	not rated	?
	266	No	15	13	130.325	Yes
	210	No	16	4	112.95	Yes

Table 2. Significance of Deviations of Chosen Dancers by Category

Category	Selected dancers Mean Rutgers dance ratings	SE	N	Eligible dancers not selected Mean Rutgers dance ratings	SE	N	t	df	p*
Sym Males	122.80	2.45	10	116.10	2.19	3	1.41	11	0.185
Asymm Males	94.03	3.36	10	118.04	3.03	3	3.70	11	0.0035
Symm females	116.01	3.62	10	N/A					
Asymm females	97.43	5.30	10	122.87	4.55	4	2.77	12	0.017

*Using the non-parametric Mann-Whitney U Test, rather than the parametric t-test, does not change which comparisons are statistically significant, giving p-values of 0.091, 0.011, and 0.016, respectively.

odology, there need be no real correlation in nature in order to show that one exists on paper. Thus, in an analysis of covariance following the same methodology used in the paper, before we (co-authors of Brown et al. 2005) left for Jamaica we already had an over-all negative correlation between FA and dancing ability that was highly significant ($P < 0.0001$) and explained 43% of the variance in dancing ability (see Table 3C, below), pretty much precisely the same general results that we re-derived in Jamaica. But was this pre-selection of dancers sufficient—and was it actually based on the criteria we claim?

One possibility is that Dr. Brown rejected animations based on their quality and there happened to be a strong positive correlation between animation quality and support for our theory. To test for this, we had one of the original authors, Keith Grochow, rate the 52 animations for usability in ignorance of any information about dance ability or FA. He was asked if he would reject any based on inferior quality, which would they be? And if he found only minor defects in some, which were these? Eight he would have rejected for reasons of quality and of these, Dr. Brown rejected 3 (two of which opposed our theory and one lacked a Rutgers evaluation). The five Dr. Brown failed to reject all supported our theory. There-

fore in all seven cases where information on dance ability was already present, the selection favored the theory (binomial probability of 7 out of 7 = 0.0078). Thirteen animations were judged to have minor flaws, two of which Dr Brown rejected (both of which opposed our theory). Of the 11 he accepted, seven supported and four opposed. In short this analysis provides no support for the view that the correlations we uncovered resulted from the fact that poor quality animations happened to be those that opposed our theory. Separate from quality, animations appear to have been repeatedly chosen in a way that provided support for our theory. Note below that in correspondence with Dr Palmer, Dr Brown describes choosing the 40 from a larger sample without regard to animation quality but solely based on a random system.

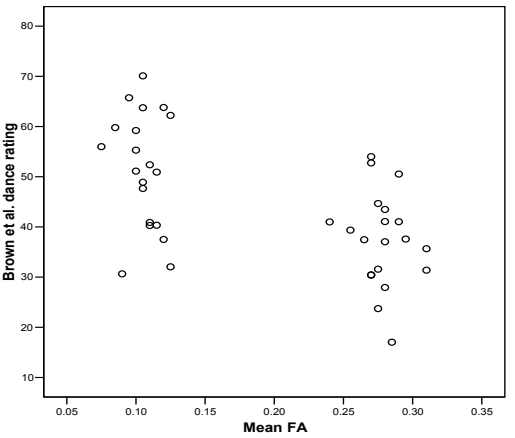
4. WERE RATER EVALUATIONS AVERAGED CORRECTLY?

We were surprised to discover that the answer to this question appears to be “no”. One would have thought that averaging a set of numbers correctly could produce only one result yet we fail to replicate Dr. Brown’s values. His and our values are highly correlated ($r = 0.818$, $P < 0.0001$) but of course they should be identical. The grand mean across all dancers is similar: we calculate 44.97 with our average dance ratings versus 44.41 with his, but the deviations are not trivial. Our sample shows a greater range in mean dance ratings (from 13.66 to 76.92) than does his (17.0 to 70.1) and a larger standard deviation (17.96 versus 12.57), and particular subjects often have quite different values. More to the point and most striking, using his values always produces stronger correlations in the predicted direction than do our data. For example, compare results based on his averages with our own for all dancers regardless of sex. Using Dr. Brown’s values in a simple regression, mean FA (averaged across 1996 and 2002) explains about 34% of the vari-

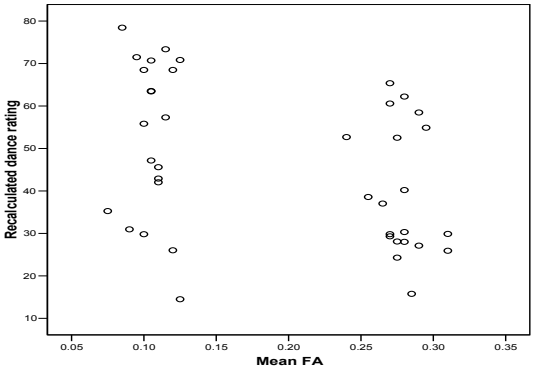
ance in average dance scores ($F_{1,38}$ = 19.442, P < 0.0001, r^2 = 0.338). Using our averages, the regression is still significant but much weaker ($F_{1,38}$ = 6.605, P = 0.014, r^2 = 0.148). Comparing scattergrams of the correlation, our computations show much more overlap in the dance scores of high and low FA individuals (Figure 2). When sex of dancer is included, effects in our reanalysis are only significant for males. Below we reanalyze the results presented in Brown et al. in more detail, using the same statistical tests and covariates he used, using both his and our average dance ratings.

How was this additional bias achieved? This is not obvious to us How do you get different values from averaging the same set of numbers? One possibility concerns zero values, which are ambi-

(a)



(b)



(c)

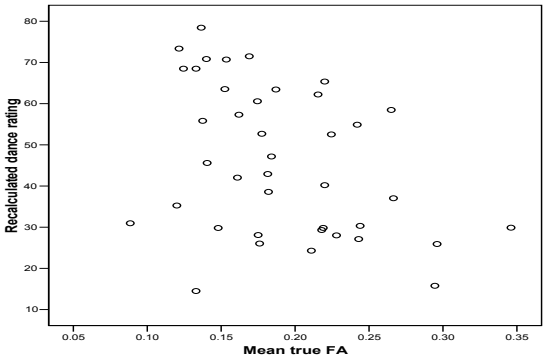


Figure 2. Scattergrams showing the relationship between mean FA (across 1996 and 2002) and mean dance ratings of the 40 selected dancers, separately (a) for mean dance ratings used by Brown et al, (b) for recalculated mean dance ratings, and (c) using the correct FA values (see below). Note the greater variability in (b) than (a) and the lack of distinct groups or any pattern in (c).

guous in the file sent to us by Dr. Brown. Zeros were present where missing values should be, so it is not always obvious which zeros

represent absent data and which represent awful dances. We included zero values, except in the obvious cases of entire rows of zeros (i.e. subjects who did not participate in the study). If one eliminates all zero values, one will skew dance evaluations consistently up, but since Dr. Brown appears to have skewed data down as well as up, this cannot be a sufficient explanation. Including or excluding zero values changes individual dance averages only slightly, because of the large number of evaluators and small number of zeros. For example, for subject #15, our average dance rating is 63.43 without zeros and 62.20 with zeros, while Brown's average is 48.9.

It does not appear that there was a bias in any particular direction, e.g. to increase the mean for symmetrical dancers and decrease it for asymmetrical ones. If we divide the subjects into categories (e.g. symmetrical subjects, asymmetrical subjects, symmetrical males, symmetrical females, etc.), we see the same pattern reported above for all subjects combined: grand means with the recalculated average dance scores are always similar to Dr. Brown's grand mean dance scores, but with higher standard deviations and higher ranges using the recalculated averages. Of course, one does not need to shift the mean to achieve the desired result—

decreasing the variability in the data alone would make the patterns appear more statistically significant than they really are.

Dr. Brown has recently claimed in correspondence (April 6, 2008) that he eliminated evaluations which were internally inconsistent, that is, in which sub-scales were scored in different directions (degree of energy, coordination, upper body ability, lower body etc). This was never the agreed-upon procedure, which was that all dance ratings would be based only on the overall evaluation (which was the only number included in the dataset). But in any case, Dr. Brown has not done what he said he did. Analysis of a subsample of evaluation forms shows no consistent link between internal variability and exclusions of data. An SPSS file sent to us by Dr. Brown includes a column next to each dancer's evaluations labeled "Variation" and composed of ones and zeros, thus having the appearance of a filter variable to exclude specific evaluations based on variation. However, when using the filter variable a very large number of evaluations are eliminated (dozens per dancer) and we still fail to replicate his average dance scores.

Finally, a very striking fact emerges from the reanalysis of a separate finding in Brown et al. (2005), namely that the dances that

young women prefer are those of relatively more symmetrical individuals than are the dances that males prefer (see below). We confirm Brown et al. (2005) almost exactly, using the same grid of values for asymmetry of dancer and average evaluated dance quality that produced results we were unable to duplicate above. In other words, when the columns of this grid are used for analysis, systematic deviations are found from Brown et al. (2005) but when the rows of the same grid are used, most findings are replicated almost exactly.

5. THE RE-MEASURE OF DANCE EVAULATIONS

To check whether biases may have been introduced in the process of scoring the youngster's dance evaluation sheets, we rescored 8 individuals chosen at random (number of evaluations per dancer: 151 to 155; total N = 1233). For each dance, each evaluator was asked to make a mark within a long thin rectangle ordered from 'very bad' on the left to 'very good' on the right. The mark was meant to indicate where along the continuum the evaluator thought this dance lay. Most marks consisted of checks. In rescoring these evaluations, we covered over the single summary number produced by Dr. Brown's assistant and chose the bottom of the check as the relevant point along the continuum. (In fact it would hardly have mattered had we chosen the left leading edge or the right, as long as we did it consistently throughout.)

In comparing our method to that of Dr. Brown's assistant, we noted several differences. We thought to count in centimeters and have one decimal place. They chose millimeters and two deci-

mal places, although in the paper measures were said to be made only to the nearest mm (on a 90mm scale). More importantly, we noted that the assistant sometimes used the bottom of the check mark and sometimes either leading edge. For example, rating dancer 38, subjects 210 and 224 used a check mark with bottom end touching the far left end of the scale indicating a choice of bad dancer. However, one was measured as 0 and the other as 2.69 mm. Moreover, double-checking measures reported by hand on the rating sheets across values in the dataset reveals some discrepancies. Some values were reported as 0, different to the actual measure, e.g. in rating dancer 94, ratings of subjects 54 and 58 (originally 86.35 and 85.52 respectively) were replaced by zeros. In other cases, the measured value was replaced by a different one in the dataset. Subject 128's rating of dancer 34 measured 88.66 mm on paper but reported as 0.8 in the dataset. These discrepancies did not seem to follow a particular pattern and constituted no more than 2 % of the data.

6. REANALYSIS OF THE RESULTS

Brown et al. (2005) tested the relationship between FA and dance ability using a 2 X 2 Analysis of Covariance (ANCOVA), with the high vs. low FA groups and sex as independent variables, BMI and age as covariates, and dance rating as the dependent variable. Below we repeat this analysis using the same software (SPSS 12.0) that Dr. Brown used with the same dataset he used and, separately, with our recalculated dance averages, which are also based on his own computer files. Although not stated in the paper, BMI was likely calculated as the average of 1996 and 2002 BMI and was likely square root transformed. We arrived at this conclusion based both on trial and error and on the description of a separate analysis by Dr. Brown performed in connection with other work. This reconstruction of the BMI variable led us to discover that 5 of the 40 selected dancers were missing 1996 BMI data, although all subjects were measured in 1996. In such cases, we used the 2002 BMI value.

The following results were reported; "a significant effect of symmetry ($F_{1,34}$ = 16.34, P < 0.001) and sex ($F_{1,34}$ = 10.99, P < 0.005), and there was a significant interaction between them ($F_{1,34}$ = 4.46, P < 0.042)…the dancers ranged in age from 14 to 19 years, but neither age nor BMI had an effect on dancing ability (both $F_{1,34}$ < 2.25, P > 0.15)." Using Dr. Brown's data we get similar, although not identical, results for symmetry, sex and the interaction effect (Table 3) but there is a nearly significant positive correlation between BMI and dance ratings and a significant positive correlation between age and dance ability. For symmetry we actually get exactly the same F-value but a lower P-value than reported, a rare case of an error causing the results to look *less*, rather than more, impressive**.** Disconcertingly, the range of ages of dancers in his own dataset is greater than what was reported in the paper: we find a range from 13 to 20, rather than 14 to 19**.** Perhaps by decreasing the range in dancer age, Dr. Brown was able to decrease the percent of variation in dance ability explained by age, and thus artificially increase the variation explained by FA (analyses of percent variation are described below). Brown et al. also report a mean age of 17.89 for symmetrical dancers and 17.40 for asymmetrical dancers. Using his own data, we find the same mean for asymmetrical dancers, but a mean age of 17.00 for symmetrical dancers.

Repeating the analysis again, this time using the recalculated dance averages, the overall ANCOVA is significant. However, none of the independent variables or covariates are significant predictors of dance ability, although FA, sex, BMI and age all approach significance (all $0.05 < P < 0.1$; Table 3).

TABLE 3. ANCOVAs using Brown's mean dance ratings, recalculated mean dance ratings and the original Rutgers ratings

A. Dependent Variable: **Brown dance ratings**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3417.086 (a)	5	683.417	8.455	<.0001	.554
Intercept	153.368	1	153.368	1.897	.177	.053
Sqrt BMI	273.989	1	273.989	3.390	.074	.091
Age	488.058	1	488.058	6.038	.019	.151
FA category (high or low)	1320.836	1	1320.836	16.341	<.0001	.325
Sex	691.022	1	691.022	8.549	.006	.201
FA category X Sex	379.625	1	379.625	4.697	.037	.121
Error	2748.186	34	80.829			
Total	85068.519	40				
Corrected Total	6165.272	39				

(a) $R^2 = 0.554$ (Adjusted $R^2 = 0.489$)

Table 3B. Dependent Variable: **Recalculated dance ratings**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4034.930 (a)	5	806.986	3.213	.018	.321
Intercept	8.401	1	8.401	.033	.856	.001
Sqrt BMI	959.178	1	959.178	3.819	.059	.101
Age	837.497	1	837.497	3.335	.077	.089
FA category (high or low)	795.978	1	795.978	3.169	.084	.085
Sex	796.013	1	796.013	3.169	.084	.085
FA category X Sex	543.018	1	543.018	2.162	.151	.060
Error	8539.351	34	251.157			
Total	93453.498	40				
Corrected Total	12574.281	39				

(a) $R^2 = 0.321$ (Adjusted $R^2 = 0.221$)

In addition to the ANCOVA, Dr. Brown also analyzed the main effects of symmetry and sex on dance ability with t-tests. Four comparisons were made, comparing the dance averages of symmetrical and asymmetrical dancers separately for males and females, and comparing the dance averages of males and females separately for symmetrical and asymmetrical dancers. In Table 4, we show the published results of these t-tests along with those we

Table 3C. Dependent Variable: **Rutgers dance ratings**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6119.711 (a)	5	1223.942	8.429	<.0001	.553
Intercept	786.151	1	786.151	5.414	.026	.137
Sqrt BMI	288.600	1	288.600	1.988	.168	.055
Age	66.894	1	66.894	.461	.502	.013
FA category (high or low)	3672.891	1	3672.891	25.295	<.0001	.427
Sex	62.781	1	62.781	.432	.515	.013
FA category X Sex	376.130	1	376.130	2.590	.117	.071
Error	4936.811	34	145.200			
Total	475002.274	40				
Corrected Total	11056.522	39				

(a) $R^2 = 0.553$ (Adjusted $R^2 = 0.488$)

calculate using Dr. Brown's data and, separately, the recalculated dance averages. In two of four cases the values in the paper show a higher level of statistical significance than we find using his data. In one case, we find the same value for t, but the wrong P-value was

TABLE 4. Comparison of t and p values for comparable data sets

Comparison	Published t and p	Dataset t and p	Recalculated t and p
Symm M vs Asymm M	4.06, <0.001	4.06, <0.001	2.44, 0.026
Symm F vs Asymm F	2.32, <0.01	2.32, 0.032	1.08, 0.30
Symm M vs Symm F	3.21, <0.005	2.61, 0.018	1.46, 0.16
Asymm M vs Asymm F	0.79, >0.45	0.86, 0.40	0.43, 0.68

Results of t-tests (all df = 18) comparing dance scores of subjects in different FA and sex categories are shown, indicating values reported in Brown et al., values obtained using Brown's dataset, and values obtained using recalculated dance averages.

was reported in Brown et al. In the other, we calculate a smaller value for t than reported, which also changes the P-value. These discrepancies had been previously noted by Yanxi Liu and Seungkyu Lee. Using the recalculated dance averages, the only significant difference in dance ability is between symmetrical and asymmetrical males, and in all four comparisons Dr. Brown's data give lower P-values than the recalculated dance averages.

Along with the results of the t-tests, Dr. Brown reported the percent of the variation in dance scores explained by the differences between symmetrical and asymmetrical males (48%) and between symmetrical and asymmetrical females (23%). Although not stated in the paper, it appears that these numbers come from partial eta-squared values in the ANCOVA when comparing only within one sex. The same pattern emerges as for the t-tests. Rather than 48% of the variance explained by male symmetry, we find 42.2% with Dr. Brown's data and 22.3% with recalculated dance averages. For female symmetry, rather than 23%, we get 12.8% with his data and 0.7% with recalculated dance averages.

In the "statistical analyses" section of the Methods, Brown et al. state that "dance ability variances were not significantly different, although a marginal difference was observed whereby males show greater variability in dance ability (180.19) than do females (113.03) (Levene test F = 3.29, P = 0.08)." We find essentially the same numbers using Brown's dataset: male variance = 180.18, female variance = 113.02, F = 3.299, P = 0.077, but very different numbers using the much more variable recalculated dance averages: 320.14, 313.11, F = 0.256, P = 0.616.

7. FA VALUES OF DANCERS WERE SYSTEMATICALLY ALTERED

When the analyses above were completed and we had already submitted this work for publication in *Nature*, we turned to resolve what we thought were two very minor discrepancies between Brown's dataset and our own: a female in one, male in another; no FA value in one data set, FA value in the other. We had not thought to compare FA values in the two datasets (for the two years) and were astonished to discover that they differed considerably (N=66) and according to a simple rule, 65 out of the 66 discrepancies concerned FA values of the 40 dancers whose animations were chosen to be shown in Jamaica. All other values (for non-chosen and those never filmed in the first place) were identical in all but one case. In 1996 117 out of 118 FAs for non-dancers were identical while 39 of 40 FAs of dancers in 1996 differed (2x2 Contingency Chi-Square with Continuity Correction, $X^2_1 = 142.53$, $p<0.0001$). In 2002 100 out of 100 non-dancers were identical while the FAs of 26 of 40 were changed ($X^2_1 = 75.58$, $p<0.0001$) (see Fig-

ure 3). These numbers are even more striking if we look at all individuals with FA data, rather than just those with Rutgers dance scores: there is agreement on the FA values of 221 of 222 non-dancers in 1996 and all 133 non-dancers in 2002. The system seems to have worked by first changing 2002 FA values so as to bring individuals into the appropriate categories and then 1996 data were afterwards made consistent. Since one began with the 2002 data one needed to make fewer changes there. That the FA values we use here are the correct values has been confirmed by both Dr. Amy Jacobson (for 1996 and 2002), a coauthor on Brown et al. (2005) and Dr. John Manning (for 1996), one of the original members of the Jamaican Symmetry Project and a co-author on Trivers et al. (1999).

The data are summarized in Table 5. Note that especially for 1996 many individuals changed the tercile in which they were located. Of these, 20 are above or below the median dancing ability in the predicted direction, 5 in the opposite direction and 1 right at the median (the binomial probability of 20 of 25 in the predicted direction = 0.0016). That the FA of dancers was altered means that the selection of dancers is even more strongly biased than we had originally thought (see above “Were dancers chosen randomly with

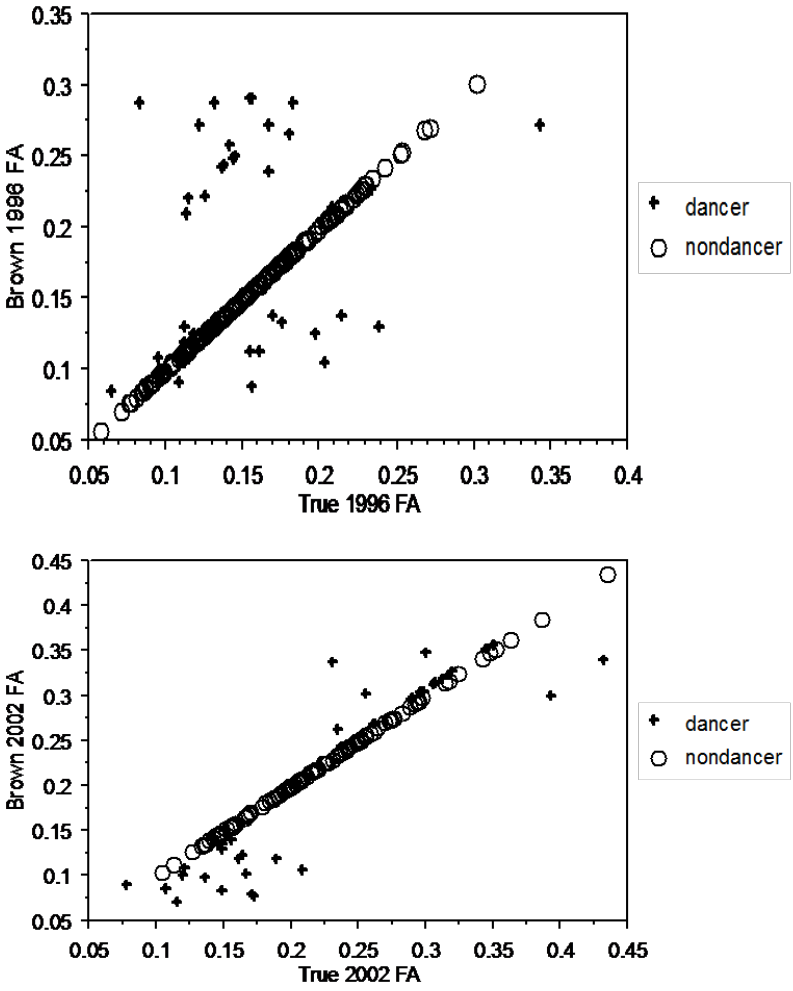


Figure 3. Scattergrams of Brown's FA values vs. the correct ones for dancers and non-dancers in (a) 1996 and (b) 2002. Note that non-dancers' values are almost always identical while dancers' are usually not.

Table 5. Comparison of True FA Values and Brown's FA Values for the 40 Dancers

Actual FA values for 1996 and 2002 are compared to those in Brown's dataset for each of the four categories of dancers used in Brown *et al*. Cases where an individual's actual tercile rank differs from the assigned category are indicated in bold. Rutgers dance scores are compared to the median for all males (116.813) in the dataset and all females (118.975).

Brown category	Actual FA 1996	Brown FA 1996	Actual Tercile 1996	Actual FA 2002	Brown FA 2002	Actual Tercile 2002	Rutgers Dance Score	Above median
Sym male								
	0.098	0.105	symm	0.175	0.073	symm	129.575	yes
	0.067	0.082	symm	0.173	0.075	symm	119.55	yes
	0.171	0.135	**asymm**	0.11	0.082	symm	110.75	**no**
	0.096	0.092	symm	0.081	0.086	symm	123.125	yes
	0.11	0.105	symm	0.139	0.093	symm	120.5	yes
	0.114	0.101	symm	0.191	0.115	symm	121.4	yes
	0.111	0.087	symm	0.164	0.115	symm	127.4	yes
	0.114	0.115	symm	0.152	0.132	symm	115.5	no
	0.158	0.085	**middle**	0.149	0.132	symm	121.375	**yes**
	0.121	0.122	**middle**	0.159	0.135	symm	138.8	**yes**

Brown category	Actual FA 1996	Brown FA 1996	Actual Tercile 1996	Actual FA 2002	Brown FA 2002	Actual Tercile 2002	Rutgers Dance Score	Below median
Asymm male								
	0.158	0.288	**middle**	0.264	0.264	asymm	82.6	**yes**
	0.157	0.287	**middle**	0.435	0.335	asymm	95.275	**yes**
	0.185	0.285	asymm	0.301	0.301	asymm	87.275	yes
	0.134	0.284	**middle**	0.396	0.296	asymm	99.375	**yes**
	0.345	0.269	asymm	0.347	0.347	asymm	75.45	yes
	0.182	0.262	asymm	0.258	0.298	asymm	99.4	yes
	0.144	0.254	**middle**	0.292	0.292	asymm	98.15	**yes**
	0.128	0.218	**middle**	0.303	0.343	asymm	109.475	**yes**
	0.117	0.217	**symm**	0.238	0.258	**middle**	87.5	**yes**
	0.116	0.206	**symm**	0.233	0.333	**middle**	105.8	**yes**

Brown category	Actual FA 1996	Brown FA 1996	Actual Tercile 1996	Actual FA 2002	Brown FA 2002	Actual Tercile 2002	Rutgers Dance Score	Above median
Sym female								
	0.178	0.130	**middle**	0.118	0.067	symm	109.8	**no**
	0.216	0.134	**asymm**	0.152	0.079	symm	118.975	**at median**
	0.241	0.126	**asymm**	0.122	0.095	symm	100.575	**no**
	no value	0.088	**no value**	0.169	0.097	**middle**	126.625	**yes**
	0.163	0.110	**middle**	0.211	0.103	**middle**	123.925	**yes**
	0.199	0.122	**asymm**	0.105	0.123	symm	116.525	**no**
	0.157	0.109	**middle**	0.167	0.119	symm	121.125	**yes**
	0.114	0.126	symm	0.152	0.124	symm	102.55	no
	0.099	0.090	symm	0.144	0.136	symm	135.5	yes
	0.206	0.102	**asymm**	0.146	0.137	symm	104.475	**no**

Brown category	Actual FA 1996	Brown FA 1996	Actual Tercile 1996	Actual FA 2002	Brown FA 2002	Actual Tercile 2002	Rutgers Dance Score	Below median
Asym female								
	0.085	0.285	**symm**	0.265	0.265	asymm	91.25	**yes**
	0.124	0.269	**symm**	0.24	0.24	**middle**	100.725	**yes**
	0.169	0.269	**middle**	0.315	0.315	asymm	113.525	**yes**
	0.147	0.247	**middle**	0.293	0.293	asymm	112.975	**yes**
	0.146	0.246	**middle**	0.31	0.31	asymm	89.45	**yes**
	0.14	0.240	**middle**	0.309	0.309	asymm	113.775	**yes**
	0.139	0.239	**middle**	0.299	0.299	asymm	73.925	**yes**
	0.169	0.236	**middle**	0.319	0.319	asymm	108.35	**yes**
	0.236	0.224	asymm	0.353	0.353	asymm	67.875	yes
	0.211	0.211	asymm	0.322	0.322	asymm	107.575	yes

reference to dancing ability and FA?").

It is noteworthy that after these false FA values were constructed, all asymmetrical males danced poorly as did all asymmetrical females. On the other hand, 8 out of 10 symmetrical males danced well, but only 4 out of 10 females did so. This suggests that not only was there an attempt to build in a positive association between symmetry and dancing ability from the beginning, but also the sex difference in its effect.

Note that because dancers were switched from one tercile to another in a complex pattern, it is not possible to compute the real sex difference in emphasis placed on bodily symmetry in dance evaluations (see below "Additional results") since we can no longer generate the binary groupings used in Brown et al. (2005). Thus, this result also disappears.

To see how individual FAs may have been manipulated, one dancer chosen and one not were compared for all 9 traits measured for FA in 1996. The non-chosen individual had identical values in our data set and Dr Brown's, but for the chosen dancer, all traits differed in relative FA, without any obvious pattern except that 8

out of 9 traits showed lower FA in our sample (4 by a factor under 5, 2 above 20 and 2 in between). Although relative FAs all differed, the numbers that are used to calculate relative FA (unsigned FA and trait size) were identical, except for one missing value. Relative FA for each trait is simply unsigned FA divided by trait size, thus it is clear that somehow new relative FA values were invented and used to calculate composite FA. As with mean dance ratings (see above, “Were rater evaluations averaged correctly?”), we get different numbers from Dr Brown even when using the same data in simple calculations.

It is of some interest to read Dr Brown’s description of what he claimed was his mode of choosing animations from the larger relevant category:

> “First I randomized subject numbers for the entire data set using web-based software (www.random.org). Afterwards random selection was done through a roll of a dice. Specifically if 14 males were in the top third percentile for time one (1996) and time two (2002) a dancer was eliminated if my dice rolled a "one" for any of those 14 males'. This was done until I reduced the sample to 10 for each category. I can provide the num-

> ber of people that were in the top and bottom thirds for time one and two by sex if you wish."

This is the very model of disinterested science—he throws a dice 6 times, on average, in order to make a single random elimination—but so far as we can tell it is pure fiction, since as as we have noted FA values were specifically created in order to place individuals in the top or bottom tercile. Dr Brown's description of his randomizing process is found in an e-mail (December 20, 2005) to one of the top experts in the analysis of FA data (Dr Richard Palmer, University of Alberta) who had written him a series of questions about the methodology of Brown et al. (2005).

We also examined the last two subjects in the dataset, subjects 287 and 288, because in both cases not all traits were measured in 1996, yet composite FA values are present in Dr Brown's dataset, which should be impossible. Subject 287 was a dancer, and 288 was the only nondancer with disagreement among datasets. Unsigned FAs and trait sizes match what is in the real dataset. Brown's dataset does not show any relative FA values in 1996 for these subjects, as if a reminder to not add up the scores due to missing traits. A summed composite FA is given nonetheless and

matches what would be calculated by simply adding up the correct relative FA values for all traits recorded. Of course, these individuals now automatically become “symmetrical”, because the missing sub-scores make the summed FA values artificially low. Both Brown’s dataset and the correct dataset show no summed FA in 2002 for subject 288. For subject 287 the relative (and thus also summed) FA scores in 2002 do not match, despite matching unsigned FAs and trait sizes. The mystery deepens for subject 287, because she is one of the 40 dancers and has Jamaican dance ratings in Dr Brown’s dataset, but is not on the list of subjects with animations or Rutgers dance scores. In analyzing Rutgers dance scores we have assumed that the animation for 287 was mislabeled as 281 – the alternative is that not only FA values for this subject were invented but also the Jamaican dance scores.

8. ADDITIONAL RESULTS

In addition to testing for the effects of FA on dance ability, Brown et al. also examined differences among the evaluators of the dances, based on sex and on their own FA. To compare the strength of the preference for symmetrical over asymmetrical dancers, a variable called “relative preference for symmetrical dancers” was constructed by subtracting mean ratings given to asymmetrical dancers from those given to symmetrical dancers, separately for sex of dancer and sex of rater. Here the results we calculate from the dataset are very close to those reported in the paper. Brown et al. report, “Female evaluators had a stronger relative preference for symmetrical male dancers (20.43 ± 13.54) than male evaluators (14.90 ± 17.55) (t_{154} = 2.21, P = 0.029)...” Our corresponding numbers are as follows: 20.43 ± 13.55, 14.94 ± 17.37, t_{154} = 2.15, P = 0.033. Additionally, “there was no sex difference in dance ratings of symmetrical females (t_{154} = 1.50, P = 0.137). However, male evaluators did give higher ratings to the dances of females (43.75 ± 17.67)

than did female evaluators (37.87 ± 15.08) (t_{154} = 2.19, P = 0.03)." Our corresponding numbers are as follows: t_{154} = 1.46, P = 0.145; 43.68 ± 17.58, 38.15 ± 15.28, t_{154} = 2.06, P = 0.041. As we noted above, what is so surprising about this near replication is that it is based on the same grid of numbers that give such discrepant values elsewhere.

Brown et al. also used multiple regression (with covariates BMI and age) to examine the influence of evaluator FA on the preference for symmetrical dancers, reporting that male FA was negatively correlated with their preference for symmetrical females ("partial R^2 = 0.11, P = 0.02"). When repeating this analysis using his own computer files, the regression model does not predict relative preferences for symmetrical females at all (overall regression R^2 = 0.010, P = 0.538; for FA, partial R^2 = 0.02, P = 0.174). While a partial regression plot in Brown et al. shows a strong relationship between male symmetry and preferences for female symmetry, we find none (compare our Figure 4 and Figure 2 in Brown et al. 2005). The regression model is also non-significant if we use the correct values for FA, rather than those in Brown's dataset (overall regression R^2 = 0.013, P = 0.726; for FA, partial R^2 = 0.010, P = 0.259). Using Brown's dataset, we do find that, as reported in Brown et al.,

"there was no significant association between female evaluator FA and preferences for symmetrical males' dances", and our numbers are similar. Brown et al. report: partial $R^2 = 0.02$, $P = 0.32$ for female FA, while we find partial $R^2 = 0.02$, $P = 0.23$. The overall regression shows no relationship ($R^2 \sim 0$, $P = 0.401$). If we use the correct FA values rather than Brown's values, our numbers are as follows: Female FA partial $R^2 = 0.025$, $P = 0.090$; overall model $R^2 = 0.016$, $P = 0.191$.

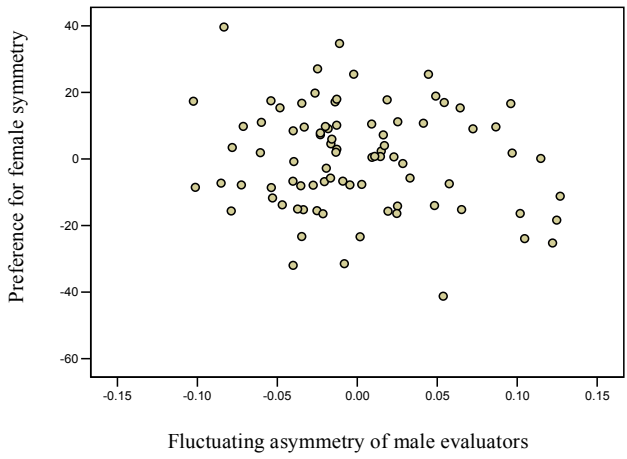


Figure 4. Scattergram showing the relationship between male evaluator FA and preferences for female symmetry using the numbers in Brown's dataset. This is a partial regression plot, controlling for BMI and age, and both variables are residuals. Compare to Figure 2 in Brown et al., which showed a strong negative relationship.

Unfortunately, Brown et al. do not report numbers for the overall regression model for either males or females, and also give little detail on how the model was constructed. The numbers we report here result from performing the regression separately for males and females. Performing one regression with both males and females combined, and adding sex and interaction effects as variables, does not improve the outcome.

Brown et al. report that the sex of the dancer in the animation was identified correctly only 62% of the time. As stated above, we found 71% in a more recent study involving many of the same subjects, but initially thought that this discrepancy could be explained by differences in evaluator age. However, the numbers in Dr. Brown's own dataset do not give 62% - instead the average is 68%, closer to the more recent study. We do calculate the same standard deviation (0.11) reported in Brown et al., and, as reported, females are significantly better than males at identifying sex. We find an even more striking sex difference than reported. In a Mann-Whitney test, $Z = 3.033$, $P = 0.002$, rather than $Z = 2.25$, $P < 0.03$, and the means for female and male evaluators respectively are 72% and 65%, rather than 64% and 60%. The differing numbers do not come from recalculating based on raw data, where a mistake in

scoring or data entry could have occurred, but instead from the column in the SPSS file sent to us by Dr. Brown which lists the proportion correct identifications for each evaluator**.** (Another column lists the number of correct identifications, and matches the proportions.) In other words, we are likely using exactly the same numbers he had but get very different results, even when simply calculating the mean.

Several additional analyses were reported by Brown et al. as online supplementary data, and are also reanalyzed here. 1) Mean and standard deviation of FA for 1996 and 2002 is reported for all four groups of dancers (symmetrical females, asymmetrical females, symmetrical males, asymmetrical males). Using Dr. Brown's FA values, we find the same eight values for mean FA (four categories across two years), and the standard deviations differ in only one of eight cases, likely due to an error in rounding off (calculated SD = 0.024, in the paper given as 0.03). That these values for FA (many of which are false) match what is reported in the paper is important, because it demonstrates that we are using the same data that Dr. Brown used. So it is an additional curiosity of reanalyzing Brown et al. (2005) that some results are confirmed almost exactly, while others collapse using the same numbers. 2)

Brown et al (2005) state that correct or incorrect sex identifications do not bias evaluations of dance ability. We are unable to replicate the 2 X 2 X 2 split-plot ANCOVA which added correct or incorrect sex identification as a within-subjects variable, because we do not have a file with the necessary data, although he claimed to have sent us all of his computer files related to the Jamaican Symmetry Project. 3) The 2 X 2 ANCOVA of sex and FA category included in the text of the paper includes age and BMI as covariates, as discussed above. In the supplementary analyses, ratings of facial attractiveness and self-esteem were separately tested as additional covariates and neither was significantly correlated with dance ratings (facial attractiveness, $F_{1,33} = 0.06$, $P = 0.81$; self-esteem, $F_{1,33} = 0.24$, $P = 0.63$). We do not have all of the data, but what we do have suggests that facial attractiveness is actually correlated with dance ability ($F_{1,12} = 6.59$, $P = 0.025$). These data come from attractiveness ratings by Jamaican peers only, and ratings are missing for many of the dancers. In Brown et al., the attractiveness variable came from averaging peer ratings with ratings by adult Jamaicans and Rutgers University students. We do not find a significant relationship between self-esteem and dance ratings, but again we are missing data ($F_{1,24} = 1.06$, $P = 0.31$).

9. IS THERE IN FACT A CORRELATION BETWEEN FA AND DANCE ABILITY?

To answer this question, we analyzed the full set of 165 dancers evaluated by the two Rutgers dance students (because of missing data, N = 162) and the correct values for FA. These are the only unbiased data we have. They could be improved by a larger sample of evaluators—preferably West Indians of the appropriate ages—but for the moment they can at least give us a rough sense of what actually may be true.

A simple regression of mean 1996 – 2002 FA and Rutgers dance scores shows no significant relationship between symmetry and dance ability ($F_{1,159} = 2.882$, $P = 0.092$, $r^2 = 0.018$; if split by sex, males: $F_{1,89} = 0.896$, $P = 0.346$, $r^2 = 0.010$, females: $F_{1,68} = 2.929$, $P = 0.092$, $r^2 = 0.041$). Performing this analysis instead as a multiple regression across the full data set with BMI, sex, FA, and age entered as covariates, as well as sex X FA and sex X BMI interactions, reveals an over-all model that is not significant ($F_{6,154} = 1.417$, $P =$

0.212, adjusted r^2 = 0.015). Only FA effects in females show any trend (P = 0.089; partial r = -0.12) while there is clearly no relationship in males (P = 0.402; partial r = -0.07). There is also a trend for BMI to be oppositely related to dancing ability in the two sexes**,** negative in females and positive in males (for sex X BMI interaction, P = 0.095, partial r = 0.13).

If instead of mean 1996-2002 FA we use 2002 FA only, there is a significant, but weak relationship between symmetry and dance (in a simple regression, $F_{1,138}$ = 8.744, P = 0.004, r^2 = 0.060; males: $F_{1,75}$ = 4.498, P = 0.037, r^2 = 0.057, females: $F_{1,61}$ = 3.852, P = 0.054, r^2 = 0.059). It makes sense to use 2002 data rather than the mean, because 1996 and 2002 values are not correlated when using the true values for FA (see above "Were dancers chosen randomly with reference to dancing ability and FA?") and 2002 is closer to the time the dances were recorded (year = 2004). But the full multiple regression model is not significant ($F_{6,133}$ = 1.730, P = 0.119, adjusted r^2 = 0.031). In this analysis a non-significant trend for FA and dance scores to be negatively correlated is apparent in both sexes (males: P = 0.071, partial r = -0.16; females: P = 0.064; partial r = -0.16), while none of the covariates show a relationship with dance scores (all P > 0.38, absolute value of all partial r < 0.1).

10. LESSONS TO BE LEARNED

Dr Trivers is fond of saying that he knows statistics "conceptually", i.e. he could not perform a t-test if you put a gun to his head, but he thinks he understands the general logic and function of the field. This is a pleasant enough joke in many contexts but not when you are in charge of a major project such as the Jamaican Symmetry Project in which yearly data sets are generated on the same set of ~160 individuals, split by sex, on a variety of biological, behavioral and social parameters, as a function of an underlying variable, degree of fluctuating asymmetry, itself a composite of 9 bodily measures, averaged and corrected for trait size, and, in turn, re-measured every ~5 years. In short, a very detailed and complex data set which requires not only ordinary statistics but the ability to search for numerous correlations in a complex manner while correcting for the frequency and the form of the multiple searches.

We believe this deficiency on the part of Dr Trivers contributed to the disaster in at least three ways. On the one hand, there was no one to oversee the statistical work, to spot suspicious patterns as they emerged and so on. Second, we imagine that Dr Brown may here have seen an opportunity that would not exist if Trivers were in statistical touch with his own data. In other words, Dr Brown may have presumed a very low likelihood that his data would ever be subject to any re-analysis or even chance contradiction.

Reward structures of course also encourage fraud (Montgomerie & Birkhead 2005). The primary benefits are those that flow from being first author on a major paper in *Nature* (featured on its cover). Although we believe it was published there because of the novel methodology used (motion capture work that isolated the phenotype of the dance from all other phenotypic characters of the dancer) it would hardly have been published without the striking results we appeared to achieve. Dr Brown also received $1000 extra pay for his analysis of the dance data and it seems unlikely that he would have received quite this amount had he failed to find such a striking pattern of data. It is also unlikely he would have been as warmly recommended for a university position

in September 2005 for, among other things, the "beautiful analysis he has just completed" of the dance and symmetry data, an opinion which in any case Dr Trivers was not qualified to offer.

Finally, the experimental design itself was a mistake: out of 106 that had FA values in both years, 40 animations were evaluated by 154 people. It would have been better done the other way around, 40 people evaluating all 106 dances—even if their work had to be spread over several days to avoid observer fatigue. The reason for this is well known. Different people are usually sufficiently well correlated in their evaluations that only a few (say, 10) are needed to get an accurate over-all measure, and 20 offer little in improvement. By contrast, if you only evaluate ¼ of the dancers you are throwing away ¾ of your data, with numerous unfortunate consequences (Preacher et al. 2005).

Social scientists often like a so-called extreme group design. You take only the 10 most symmetrical and the 10 least symmetrical males and females and, thus you are more likely to detect an effect than if you took 40 individuals at random. This is true and relevant if you are limited in subjects but we had the full sample available which permitted a more detailed analysis than the ex-

treme design, and by dropping the number of evaluators, over-all costs could remain the same. The only disadvantage is that we would have a harder time detecting a between-observer effect (e.g. asymmetrical males preferring the dances of relatively asymmetrical females) but these are secondary facts and the loss of the primary facts is more costly.

If we had evaluated all 106 animations, we would have a detailed regression across all FA values instead of only comparisons of two extreme categories, and we would also have a good measure of effect size (how much of dance ability is explained by the degree of bodily asymmetry of the dancer). In the extreme design you can measure an effect size but it requires a careful correction before you can interpret it (Preacher et al 2005). For example, in nature we hardly expect an effect size of even 15% for symmetry on dancing ability, perhaps 5 or at most 10 are more plausible— bodily asymmetry being itself poorly measured and an imperfect measure of the underlying variable presumed to be important (developmental stability) — only one of several variables expected to affect (or be associated with) dancing ability. Thus, if someone produced a data set with an effect size of 30% one would immediately have grounds for imagining that the data were “too good to

be true" (Gangestad et al 2001, Møller et al 2005). But in an extreme design it is hard to decide what is an extreme effect size. As we have noted, Brown et al. (2005) showed an effect size of 48% in males for FA and dancing ability and an effect size of 23% in females.

For present purposes, the most important effect of this poor design was that it more easily permitted fraud. If you use the full sample, you can not choose (or create) a sub-sample to fit your biases. Instead, you will have to create a full new data set, more difficult to achieve and much more likely to be detected. Also one would be less easily tempted to fraud if you knew that the full sample of Jamaican evaluations were going to be compared to the full sample of Rutgers ones. Dr Brown argued for the extreme design on the usual social science grounds but he probably already had in mind molding the sub-sample toward pre-conceived results. Certainly it was done within weeks of giving him the go-ahead.

11. THE ROLE OF *NATURE*

The primary role of *Nature* was to publish the paper and here they did a superb job, not just in printing the paper but in the high quality of referees they consulted. One pointed out an important flaw in our statistical analysis of dancing ability vs FA—we had done a between-observer analysis when we should have done a between-dancer one. Dr Brown hopped to the task of doing the correct analysis and (thankfully, or so it seemed) all the major results remained unchanged, although exact numbers of course changed. In retrospect, we believe this may have been the time when false dance averages were created, since it was easier to show a result with a between-observer analysis, given the large number of observers. In any case, the fact that everything stayed the same should have served as a warning that something was amiss. A second referee pointed out a series of problems that needed to be addressed before the paper could be accepted and in addressing these we strengthened the paper throughout.

As for publication of the work you are now reading, at first we submitted the work to *Evolution & Human Behavior*, the most relevant journal after *Nature*, we rationalized since it dealt with human behavior from an evolutionary standpoint. We thought *Nature* would not give us the space we needed and we could couple the publication of our paper in *E&HB* with a short retraction in *Nature* citing our published paper. After a month the editor wrote us and said they had never sent our paper out to review because (1) this was not their problem but *Nature*'s and (2) any difficulties arising belonged with *Nature* and not them. We liked their response because we thought it was honest and because we agreed with it.

We then rewrote the paper for *Nature* and submitted it in late November 2008. We suggested that a short retraction by us (and apology by Trivers) could be published, with the paper itself appearing on-line. *Nature* replied that they would first send our paper to the original co-authors for their comments and would share them with us, after which they wanted a revised manuscript as well as the full original data set, animations included, so that these could be sent to the original referees and possibly others, for independent analysis. We agreed to this and set about putting together all the files they would need. In the process, as we have

noted, we discovered that entirely fictitious FA values had been created for most of the 40 dancers. Thus, up to this point, we had an entirely favorable opinion of *Nature.* Their high standards had improved the paper in the first place and had now led us to discover the most damning evidence of fraud yet.

Eventually, after repeated requests*, Nature* coughed up the comments of the other co-authors. Dr Jacobson did not respond. Dr Brown responded but did not say whether his comments could be shared with us; he was queried by *Nature* on this point and so far has not responded. Drs Popovic, Grochow and Liu responded as a unit and said that perhaps a third person might evaluate the 52 animations for quality of animation in order to resolve the difference of opinion between Drs Cronk and Grochow. Since Dr Grochow had done a blind analysis of all 52 animations while Dr Cronk had done an evaluation of only the 12 he knew were rejected (see below), the University of Washington team was taking a very conservative position. From a scientific standpoint their data was the only useful information on animation quality. (We chose not to follow their advice here because by now the evidence for fraud is so overwhelming that even if one questions Dr Grochow's analysis, which we do not, it hardly changes the general conclusion.)

The truly interesting response was that of Dr Cronk. He originally conceived the project and he oversaw the Rutgers evaluations that both permitted the fraud and are now the only useful data we have (while we await additional dance evaluations by others). Dr Cronk sent only confidential comments not to be shared with us. This is curious on its face, since we had always shared all findings with him and Dr. Trivers had worked with him on the project from the beginning, but it was fully consistent with his pattern of behavior (as described below) from the moment the possibility of data fabrication first reared its ugly head.

What was stranger than the length of time it took *Nature* to respond was a set of new requirements that emerged. For one thing, because a "majority" opposed our paper, we would be limited to 600 words and one small figure or table. The logic of this was obscure to us. If there is dissension in the ranks, all the more reason to lay out our case in detail. If all co-authors (except one) agree, a relatively short retraction should be sufficient. Since Drs Brown and Cronk (at least) presumably opposed a retraction, considerable evidence, carefully analyzed, is necessary to make an airtight case. In short, *Nature* will not devote $1/10^{th}$ as much space to the correction of an error as to its launching even though the latter

is in print and the correction only online. As we argued in vain with *Nature*, cyberspace is not a limited resource, so what exactly is the problem?

Nature also developed a phobia about what we could say about what had transpired. We were only to focus on the "technical inaccuracies of the data" and not seek "any apportioning of blame". What did this mean? That statistical errors were made by unidentified objects so that every co-author must sit under a cloud of suspicion in order not to identify the person who did the statistics? Finally, from the time we submitted our paper until the time when we had received all the co-authors responses (or non-responses) was a full three months. *Nature* was taking its time to correct a paper that was now three years old (with 29 citations and counting). Indeed, we could look forward to another two months at least before we commenced negotiations with *Nature* over what kinds of assertions we would be permitted to make—assuming, that is, that they accepted the utility of publishing some kind of re-analysis of the original paper.

But this caution was perhaps fully justified from their stand-point. After all, it was not through any failure of their own that we

had come to this unfortunate impasse but a mess on our end. They could well say, you have already occupied prime real estate to publish these falsehoods, we can hardly give you endless space to correct them. But whatever the logic, their stance made it less likely that fraud will be fully and appropriately revealed and we decided it was best to publish on our own. As for the requirement that we not apportion blame, we believe this comes from a fear of having to defend a lawsuit, especially in the U.K., so that this factor also tends to discourage full treatment of fraudulent results.

12. TO ACKNOWLEDGE DATA FABRICATION OR NOT: THE CASE OF DR CRONK

Dr Cronk's initial position was that we should not bother with any reanalysis. Period. Perhaps there were minor problems, but they were best left aside. Then he repeatedly tried to get Dr Trivers to agree that if inconsistencies were discovered, we would not publish anything until we returned to Jamaica, repeated the work (at a cost of ~$10,000 U.S.) and only if we then failed to replicate Brown et al. (2005) should we publish. This, of course, is foolish on its face. Einstein once defined insanity as doing the same thing over and over again and expecting a different result. Having proven that all the findings in Brown et al. (2005) are manufactured, how could we possibly expect to generate these same false values through honest work—and why should the larger scientific community have to wait for this absurd exercise before they learned that none of the original findings could be corroborated *statistically* using the correct data set? In addition, as we noted at the beginning, an independent analysis of 40 other animations

studied for 2nd to 4th digit ratio, showed no relationship between FA and dancing ability, only a trend in the wrong direction.

Incidentally, in this new sample, and of course in the original full sample, Dr Cronk included some of the same animations that he declared were unusable for the Jamaican work. He insisted that in the U.S. people could easily overlook the minor flaws in these animations but he apparently felt that Jamaicans would not be up to this task. The basis for this belief is unknown to us.

Dr Cronk did say that he would let "the chips fall where they may" but this was before any chips fell. When we sent Dr Cronk the first draft of our paper, he claimed to find a series of statistical errors, all biased in our favor, and where there appeared to be no grounds for choosing our version over Brown's (e.g. averages of dancing ability), he preferred the more reliable Brown. He claimed that we had misclassified one individual as above the median in dancing ability instead of below (which was a mistake on his part). He pointed out that we had made an elementary statistical mistake in not changing the median dance evaluation as animations were eliminated. This is perfectly true. If Dr Brown starts with 14 asymmetrical females rated for dancing ability, and eliminates one

that is above the median in dancing ability, then you must recalculate the new median for the 13 that remain before you analyze the next elimination. But Dr Brown always eliminated from one side of the distribution only. By keeping the old median (a conservative but simple statistical test) we were helping Dr Brown's case, not hurting it. After 3 eliminations, for example, there were still 7 evaluations below the original median but now only 4 above, yet once again he eliminated one that was above. In short, Dr Cronk failed to see the implications of his own thinking. In addition, we chose this simple form of analysis because the fact that some individuals were present from outside the stated criteria made it difficult to determine precisely what distribution the comparisons should be based on.

Finally, Dr Cronk chided us for not taking the time to look at the 12 relevant animations to see if they were of sufficient quality. When Dr Cronk did so, he found 7 that were obviously deficient, hence we should only analyze the other 5 eliminated animations, a sample size too small to show anything. But here again, Dr Cronk appears to have misled himself. His 20 minute exercise had little to do with science since as the evaluator he was not blind to whether an animation had been chosen or not and he had, of course, a bias. We were already at work setting up the proper test in which some-

one expert at evaluating animations (Dr Keith Grochow) would do so in complete ignorance of which animations were chosen by Dr. Brown and which not. His work, as we have seen, showed no bias by quality of animation that would produce the results Dr Brown had generated.

Lest there be any doubt, concluded Dr. Cronk (May 20, 2008:

> There is no merit to their [T, P and Z's] claims against Will [Dr Brown]. It would therefore be appropriate for the authors of the TPZ document to formally apologize to him. If any of them have shared their suspicions with third parties, then they also have an obligation to seek out those third parties and to do whatever they can to restore Will's good name. Doing so would obviously be in Will's best interests and in the interests of fairness and justice. But it would also be in the interests of science. Will is a productive and imaginative young scholar who has already made important contributions to evolutionary psychology. To have his reputation sullied by these baseless accusations would make it difficult for him to continue to make these contributions.

The only thing we agree with in this paragraph is that Dr Brown is “imaginative”. Surely it is not safe to assume that his earlier published work can be taken at face value (nor his latter). To fail to acknowledge reality to us while holding tight to previous prejudices appears to be a policy of silence and denial—if one does not respond or acknowledge, hopefully the problem will go away. For a long time, we never heard from Dr Cronk, despite sending him the data and a detailed response to his letter and later versions of our paper with requests that he join us as co-author. Only after the paper had been submitted to *Nature* did he send a one-line note acknowledging that he had sent the full set of relevant animations to *Nature.* At the same time, his comments to *Nature* were held in confidence. When he was finally sent the evidence that well more than ½ of all of the FA values in the study had been fabricated, he promised to evaluate the findings very carefully as he said he had with our previous work. If he did so, he has not chosen to share any results with us. If we had taken his approach from the beginning, the fraud would remain undiscovered to this day with all its attendant and accelerating costs.

13. HOW COMMON IS FRAUD?

It is precisely for the reasons suggested above that we believe fraud of the sort documented in this book is more common than many would guess—not perhaps wholesale fabrication of data as in this case, but data manipulation and creation in the service of producing significant and noteworthy findings that do not in fact exist. In most cases fraud is unlikely to be detected due to a lack of replication, and if detected often goes unreported (Montgomerie and Birkhead 2005). Since circulating earlier versions of this book we have heard several stories from scientists of suspicious activity swept under the rug for the benefit of all concerned. To give but two notable examples:

A student's analysis of data required that there be no significant difference between two samples. A test of the null hypothesis that the student's samples came from the same population had a P-value of 0.99. This means that if the null hypothesis

were correct one would still obtain a greater difference than the observed difference 99% of the time. How did this unlikely event happen? Perhaps the student was lucky or perhaps the data had been concocted to make the samples similar (and the student had overdone the job). The matter was left uninvestigated.

A second student claimed that two related genes had been isolated because their DNA sequences had diverged. The genes were almost identical at non-synonymous sites but had almost complete divergence at synonymous sites even though one would expect many synonymous matches solely by chance. The departure from random expectations was very highly significant ($P < 0.0000000001$). The result cannot be explained by codon usage bias. The two sequences appeared to be actively avoiding each other at synonymous sites. Then what is going on? This may be an important discovery of an unexplained phenomenon that is worthy of publication—or the student may simply have been careless in fabricating data. The student cannot explain this anomaly, is not interested in pursuing it, and the professor shrugs it off as not worth pursuing. The dubious result is published without comment and the student goes on to a successful academic career. For an excellent treatment of scientific fraud suggesting that it is a general

problem of some importance, see Judson (2004).

In other cases, rather than outright fabrication of data, results may be presented in a misleading and biased manner. For example, ten Cate (2009) examined problems with Tinbergen's famous studies of the red spot on the herring gull's (*Larus argentatus*) beak acting as a releaser for pecking by chicks. Discrepancies between the actual results and the way Tinbergen described them accumulated in successive publications. At first a surprising finding was presented at face value, then (probably correctly) explained as resulting from a methodological problem, and a correction factor was created to adjust data for the methodological problem. In later publications the numbers were presented without mention of the correction factor, as if they were the real numbers and no methodological problems existed. Important details were omitted. Results from separate experiments were presented together and results from the same experiments presented separately in a misleading manner. Chicks were claimed to be naive when that could not be the case for the early experiments. As this example shows, even our most celebrated scientists may polish and manipulate the presentation of their work over time. Below we discuss the most famous case of all – that of Gregor Mendel.

14. FISHER'S REANALYSIS OF MENDEL'S CLASSIC WORK

In 1936 R.A. Fisher reanalyzed the classic paper of Mendel (1860) that forms the foundation of modern genetics and argued that Mendel had cooked the data. He had done this in two ways. When the predicted ratio of phenotypes was 1:3 Mendel found 1:3 but the individual values were too closely clustered around the expected value. In effect, Mendel had forgotten to include the variance. Although this could easily be explained away (see below), in Fisher's eyes Mendel also made a fatal mistake. In one case, Mendel expected a 1:2 ratio but his methodology by all logic should have produced a ratio of 1:1.7. He appeared to be unconscious of the bias in his methodology—the so-called 'ascertainment bias'—and he generated data that clustered around his expected value of 1:2, not the value his data should actually have generated. So far as we know this re-analysis is the first attempt to use statistics ex post facto to demonstrate a very improbable set of events (absent efforts to manipulate data, consciously or unconsciously).

What is the ascertainment bias? Mendel was crossing plants with themselves. If they were heterozygotes, then the double recessive phenotype should appear ¼ of the time. If it never appears you know that the plant is homozygous dominant. But 'never' takes a very long time; you must stop somewhere. Mendel stopped at 10: if 10 progeny were all dominant in phenotype then he assumed the parent was not a heterozygote—but of course by chance it could still be a heterozygote, since only 2.5 homozygous recessives are expected in a sample of 10 progeny. Stopping at 10 introduces a substantial bias that can be calculated exactly, just as Fisher (1936) did. In this case, 1:1.7 was the expected value yet Mendel reported ~1:2. In short, in Fisher's view, he cooked his data. When Fisher was done with him, the chance that Mendel had achieved these data by chance appeared to be less than 1 in a million.

Regarding the reduction in variance around the expected value, several possible explanations arise. Perhaps Mendel produced more exact ratios than expected because his hand-pollination, in fact, used much of the pollen available (instead of a random sub-sample of a much larger set). Much more likely is that he threw away extreme values as being biologically unreliable (or

some other rationalization, easily achieved) or stopped counting when close to his preferred result (also easy to achieve) or repeated experiments that gave deviant values thus tending toward more average values (Novitski 2004). The degree of consciousness of Mendel during any of these processes remains, of course, unknown.

It turns out that with particular assumptions one can easily derive Mendel's empirical results as the theoretically expected ones even in the face of an ascertainment bias. For example, it is possible to imagine that Mendel did not bother to score the full ten progeny if at any point a double recessive appeared but this could introduce a bias in the opposite direction of that of ascertainment if Mendel also replaced one of similar length lacking a recessive phenotype with a new sample of 10 (Novitski 2004). Or, under certain conditions it can be argued that Mendel naturally used samples greater than 10 and inclusion of these produced a countervailing bias, opposite to that of the ascertainment bias and stronger (Hartl and Fairbanks 2007). Of course, these excuses presume new behavior on the part of Mendel for which there is no evidence one way or another (Franklin et al. 2008).

In the case of the analysis presented here we are not dealing with the fundamental laws of genetics so that the picture of reality we paint is not easily contradicted on other grounds. Does FA have a strong or a weak effect on dance ability in Jamaican youngsters and is there a sex difference? It is hard to see how the rest of evolutionary theory is in any way affected by our answer to these questions.

In our case we had greater access to immediate data sets and analyses than did commentators on Mendel's work, including Fisher, who did so some 75 years after the fact. To us, this immediate response permits us sharply to limit competing hypotheses, e.g. in checking quality of eliminated and retained animations we eliminated quality of animation as a possible cause. It is as if we had evidence of Mendel pre-choosing his pea plants based on genetic differences in their tendency to segregate specific alleles — not to mention the creation of entirely fictitious segregation ratios.

Ours appears to be a simple case of conscious fraud, that is deliberately altering the data set to build in some of the very associations that were later discovered and then molding the later data to produce additional significant associations where none existed.

Since we are not here dealing with the laws of genetics but completely arbitrary facts regarding associations between the degree of bodily asymmetry and dancing ability in both sexes of one population of one species, there is no reason to suppose our fraud would easily have been detected (nobody would bother to check) unless, as happened, those involved came to suspect internal dishonesty. To date, we have received no response from Dr Brown to the manuscript you are reading. In that sense, we share something with Fisher: his subject was dead, ours is inert.

REFERENCES

Brown, W.M., Cronk, L., Grochow, K., Jacobson, A., Liu, C.K., Popovic, Z. and Trivers, R. 2005. Dance reveals symmetry especially in young men. *Nature* **438**: 1148-1150.

Fisher, R.A. 1936. Has Mendel's work been rediscovered? *Ann. Sci.* **1**: 115-137.

Franklin, A., Edwards, A.W.F., Fairbanks, D.J., Hartl, D.L., and Seidenfeld, T. 2008. *Ending the Mendel-Fisher Controversy*. University of Pittsburgh Press, Pittsburgh.

Gangestad, S.W., Bennett, K. and Thornhill, R. 2001.A latent variable model of developmental instability in relation to men's sexual behavior. *Proc. Roy. Soc. Lond. B* **268**: 1677-1684.

Hartl, D.L., and Fairbanks, D.J. 2007. Mud sticks: on the alleged falsification of Mendel's data. *Genetics* **175**: 975-979.

Judson, H.F. 2004. *The Great Betrayal: Fraud in Science*. Harcourt, New York.

Mendel, G. 1860. Versuche über pflantzen-hybriden. Verhandlungen des naturforschenden vereines. *Abh. Brünn* **4**: 3-47.

Møller, A.P. and Swaddle, J.F. 1997. *Developmental Stability and Evolution*. Oxford University Press, Oxford.

Møller, A.P., Thornhill, R and Gastestad, S.W. 2005. Direct and indi-

rect test for publication bias: asymmetry and sexual selection. *Anim. Behav.* **70**: 497-506.

Montgomerie, R. and Birkhead, T. 2005. A beginner's guide to scientific misconduct. *ISBE Newsletter* **17**: 16-24.

Novitski, E. 2004. On Fisher's criticism of Mendel's results with the garden pea. *Genetics* **166**: 1133-1136.

Preacher, K.J., Rucker, D.D., MacCallum, R.C. and Niewander, W.A. 2005. Use of extreme group approach: a critical examination and new recommendations. *Psychol. Methods* **10**: 178-192 .

ten Cate, C. 2009. Niko Tinbgergen and the red patch on the herring gull's beak. *Anim. Behav.* **77**: 785-794.

Trivers, R. 1972. Parental investment and sexual selection. In B. Campbell (Ed) *Sexual Selection and the Descent of Man 1871-1971*, Chicago, Aldine, Chicago pp. 136-179.

Trivers, R. Mannning, J.T., Thornhill, R., Singh, D., and McGuire, M. 1999. Jamaican Symmetry Project: long term study of fluctuating asymmetry in Jamaican children. *Hum. Biol.***71**: 417-430.

ACKNOWLEDGEMENTS

We thank Drs Cronk and Liu for access to unpublished data and Dr Grochow for evaluating dance animations for quality. We thank Drs Jacobson and Manning for providing back-up copies of our FA files. We thank Drs Cronk, Grochow, Haig, Lee, Liu, Palmer and Popovic for helpful comments. Dr Trivers thanks the Institute of Advanced Studies in Berlin for a fellowship during which this book was completed.

ABOUT THE AUTHORS

Robert Trivers is Professor of Anthropology and Biological Sciences, Rutgers University, New Brunswick. He is an authority on social theory based on natural selection and the evolution of selfish genetic elements. He is in charge of the Jamaican Symmetry Project.

Brian Palestis is Chair of the Department of Biological Sciences, Wagner College, Staten Island. He is an authority on tern behavior and ecology and on the evolutionary dynamics of B chromosomes. He has considerable experience analyzing fluctuating asymmetry in humans and terns.

Darine Zaatari is a Consultant in Antioch, California. She is an authority on the role of fluctuating asymmetry in economic games, especially the Ultimatum Game. She is the mother of Zeina.