AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 135:311–328 (2008)

Strong Postcranial Size Dimorphism in *Australopithecus afarensis*: Results From Two New Resampling Methods for Multivariate Data Sets With Missing Data

Adam D. Gordon,[1]* David J. Green,[1,2] and Brian G. Richmond[1,3]

[1]*Department of Anthropology, Center for the Advanced Study of Hominid Paleobiology, The George Washington University, Washington, DC 20052*
[2]*Department of Anthropology, Hominid Paleobiology Doctoral Program, The George Washington University, Washington, DC 20052*
[3]*Human Origins Program, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560*

KEY WORDS resampling methods; geometric mean; hominin evolution; Monte Carlo methods

ABSTRACT There is considerable debate over the level of size dimorphism and inferred social behavior of *Australopithecus afarensis*. Most previous studies have analyzed size variation in single variables or multiple variables drawn from single elements. These approaches suffer from small sample sizes, underscoring the need for new techniques that incorporate measurements from multiple unassociated elements, reducing the influence of random sampling on size variation in fossil samples. One such technique, the template method, has recently been proposed but is limited to samples with a template specimen and is sensitive to a number of assumptions. Here we present two new resampling methods that do not require a template specimen, allow measurements from multiple unassociated elements to be included in a single analysis, and allow for significance tests between comparative and fossil multivariate data sets with missing data. Using these new methods, multivariate postcranial size dimorphism is measured using eight measurements of the femur, tibia, humerus, and radius in samples of *A. afarensis*, modern humans, chimpanzees, gorillas, and orangutans. Postcranial dimorphism in *A. afarensis* is similar to that of gorillas and orangutans, and significantly greater than in modern humans and chimpanzees. Because studies in living primates have examined the association of behavior with dimorphism in body mass and craniodental measurements, not postcrania, relationships between postcranial dimorphism and social behavior must be established to make robust behavioral inferences for *A. afarensis*. However, the results of this and past studies strongly suggest behavioral and mating strategies differed between *A. afarensis* and modern humans. Am J Phys Anthropol 135:311–328, 2008.

The degree of size dimorphism (SD) present in fossil taxa is of interest because of its potential in reconstructing social behavior and ecological stress in paleontological settings (Plavcan and van Schaik, 1997a; Plavcan, 2000, 2002; Gordon, 2006). In particular, as an early hominin with a relatively large fossil record, the degree of dimorphism present in *Australopithecus afarensis* has been under study for some time (e.g., Kimbel and White, 1988; McHenry, 1991a, 1996; Richmond and Jungers, 1995; Lague and Jungers, 1996; Lockwood et al., 1996; Lague, 2002; Reno et al., 2003, 2005; Plavcan et al., 2005; Harmon, 2006). Such studies usually compare the degree of size variation present in a fossil sample with a comparable sample of material from extant great apes and modern humans. The nature of the *A. afarensis* hypodigm is such that no single measurement can be collected from all known specimens to measure the amount of size variation present in the entire hypodigm or subsets divided solely on the basis of age or locality. Thus researchers have often chosen to examine individual measurements or groups of measurements from elements that are relatively well represented in the fossil record (e.g., proximal femur measurements, mandibular measurements). Unfortunately, at small sample sizes (including the relatively large samples for fossil hominins that these types of studies have used), random sampling effects in terms of which individuals and elements are fossilized and later discovered may cause a collection of elements of atypically low or high variation to be preserved in the fossil record, producing misleading results in comparative tests of relative variation. Most current methods do not take full advantage of the information already available in the fossil record. New techniques are needed that allow information to be combined from multiple unassociated elements, increasing sample size in terms of both specimens and measurements, and thus reducing the influence of random sampling on measured fossil variation (e.g. Reno et al., 2003; Gordon, 2004).

However, every type of measurement available in a fossil sample should not be uncritically lumped together in a single analysis, because dimorphism levels may dif-

Grant sponsors: The George Washington University's Selective Excellence Initiative; NSF Integrative Graduate Education and Research Traineeship program; Grant number: NSF DGE-9987590.

*Correspondence to: Adam Gordon, Department of Anthropology, Center for the Advanced Study of Hominid Paleobiology, The George Washington University, 2110 G St. NW, Washington, DC 20052, USA. E-mail: agordon@gwu.edu

Received 13 November 2006; accepted 18 September 2007

DOI 10.1002/ajpa.20745
Published online 28 November 2007 in Wiley InterScience (www.interscience.wiley.com).

fer between measurements both within and between taxa. For example, levels of dimorphism vary substantially across the skull in primates, with greater dimorphism generally found in facial measurements than in neurocranial or orbital measurements (Plavcan, 2003). To consider a more extreme example, Tague (2005) found that in species with high body mass dimorphism, some pelvic dimensions are actually larger in females than in males. To include such measurements with dimensions from the rest of the skeleton in an analysis of overall skeletal dimorphism would produce misleading results. Similarly, patterns of dimorphism (i.e., the ranking of SD between measurements within a taxon) for mandibular variables have been found to vary greatly between hominoid species (Taylor, 2006). Including a set of measurements with different patterns of dimorphism between study taxa (e.g., measurements that show a consistent level of SD within one study species but a wide range of SD levels for a second study species) would also produce results that would be difficult to interpret at best. But when measurements show similar levels of dimorphism within a species and similar patterns across species, it is useful to be able to consider multiple types of measurements in a single analysis. In this way, one can maximize information from the fossil record and reduce the influence of random sampling effects.

To date, only two methods have been developed to assess the degree of SD present in a multivariate data set: measuring SD for a geometric mean of all measurements for each specimen (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Harmon, 2006), or measuring SD for estimates of one measurement, where values are estimated from other measurements using a set of ratios derived from a template specimen (Reno et al., 2003). In both cases, multivariate data are reduced to univariate data which can then be analyzed using a variety of techniques (e.g., max/min ratio, mean method ratio, method of moments, coefficient of variation, binomial dimorphism index, etc.).

The geometric mean has been shown to be effective in combining multiple measurements into a single measure of overall size (Mosimann, 1970; Jungers et al., 1995), but previous uses of the geometric mean in measuring SD have been limited to data sets in which every measurement is available for every specimen, comparative and fossil, so that a single measure of size can be calculated for each specimen as the geometric mean of all measurements for that specimen. The template method allows for the analysis of multivariate data sets with missing data by estimating the value of one measurement absent in a particular fossil specimen (e.g., femoral head diameter) based on the ratio between that measurement and another measurement that is present in the fossil (e.g., humeral head diameter) as measured in a template individual of the fossil species for which both measurements are available (Reno et al., 2003). The template method has been criticized on various grounds, particularly its susceptibility to error due to including multiple measurements from the same individual (Plavcan et al., 2005; Scott and Stroik, 2006), whereas geometric mean methods are designed for such multiple measurements. In addition, by using a ratio to estimate missing data, the template method assumes that ratios are equal in all members of a species regardless of size, and thus that all measurements included in an analysis scale isometrically with each other. This assumption may or may not be true for any given data set, and the consequences of violating this assumption have not yet been tested. Regardless of whether these criticisms have a major effect on test results, two constraints are certain: the template method can only be used when a fossil specimen complete enough to be used as a template is available, and only measurements present in the template individual can be included in an analysis.

TABLE 1. Sample sizes for each comparative taxon

Species	Male	Female
Gorilla gorilla	**25**	**25**
G. g. beringei	9	4
G. g. gorilla	14	21
G. g. ssp.	2	0
Pongo pygmaeus	**12**	**12**
P. p. abelli	2	4
P. p. pygmaeus	6	4
P. p. ssp.	4	4
Homo sapiens	**25**	**25**
black	13	12
white	12	13
Pan troglodytes	**25**	**25**
P. t. schweinfurthii	5	2
P. t. troglodytes	14	17
P. t. verus	4	6
P. t. ssp.	2	0

Total number of specimens per species are shown in bold; number of specimens by subspecies or ethnicity are unbolded. Ethnic categories for *Homo sapiens* follow labels in the Hamann–Todd Collection notes.

Here we present two new resampling methods for assessing taxonomic differences in multivariate SD based on the geometric mean, both of which 1) allow for missing data (e.g., individuals without all measurements), 2) do not assume a particular scaling relationship between included measurements, and 3) do not require the presence of a template specimen. We use these methods to address the question of how dimorphic the postcranium of *A. afarensis* was relative to that of modern humans and great apes. We further discuss the distinction between postcranial SD and body mass SD, and suggest additional techniques for reconstructing mass dimorphism in *A. afarensis*.

MATERIALS AND METHODS

Sample

Multivariate postcranial SD was compared between a fossil sample of *A. afarensis* and a comparative sample drawn from skeletal collections of *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus* from the National Museum of Natural History, the American Museum of Natural History, and the Cleveland Museum of Natural History (Table 1). Only adult individuals of known sex were included in the sample, and both sexes were represented equally within each extant taxon. Multiple subspecies are present for each of the ape taxa: *P. t. troglodytes*, *P. t. schweinfurthii*, *P. t. verus*, *G. g. gorilla*, *G. g. beringei*, *P. p. pygmaeus*, and *P. p. abelli*. Multiple ethnic groups are present in the modern human sample.

Measurements for *A. afarensis* were published measurements available in McHenry (1992) and McHenry and Berger (1998). Measurements were included if at least two different specimens preserving that element were available and if the measurements were taken from articular surfaces or transverse sections of long bones, since geometric means of these types of measure-

TABLE 2. A. afarensis *specimens and measurements included in analysis*

Specimen	HUMHEAD	ELBOW$^{0.5}$	RADTV	FEMHEAD	FEMSHAFT$^{0.5}$	DISTFEM$^{0.5}$	PROXTIB$^{0.5}$	DISTTIB$^{0.5}$
A.L. 288-1	27.3	20.5	15.0	28.6	20.9	–	40.3	18.2
A.L. 128-1/129-1	–	–	–	–	21.6	37.5	39.9	–
A.L. 137-48a	–	22.9	–	–	–	–	–	–
A.L. 211-1	–	–	–	–	28.2	–	–	–
A.L. 322-1	–	22.9	–	–	–	–	–	–
A.L. 333-3	–	–	–	40.2	31.3	–	–	–
A.L. 333-4	–	–	–	–	–	45.6	–	–
A.L. 333-6	–	–	–	–	–	–	–	21.7
A.L. 333-7	–	–	–	–	–	–	–	24.8
A.L. 333-42	–	–	–	–	–	–	50.6	–
A.L. 333-95	–	–	–	–	29.1	–	–	–
A.L. 333-96	–	–	–	–	–	–	–	21.0
A.L. 333-107	35.1	–	–	–	–	–	–	–
A.L. 333w-40	–	–	–	–	30.8	–	–	–
A.L. 333w-56	–	–	–	–	–	45.0	–	–
A.L. 333x-14	–	–	22.2	–	–	–	–	–
A.L. 333x-26	–	–	–	–	–	–	52.3	–

Fossil measurements in mm, taken from McHenry (1992) and McHenry and Berger (1998).

ments have been shown to scale isometrically with respect to body mass in primates in general and great apes in particular (Gordon, 2004). On the basis of these criteria, the following measurements were used for *A. afarensis* and were collected for the comparative sample: humeral head (HUMHEAD), distal humerus (ELBOW), transverse radial head (RADTV), femoral head (FEMHEAD), femoral shaft (FEMSHAFT), distal femur (DISTFEM), proximal tibia (PROXTIB), and distal tibia (DISTTIB). Measurements were taken without regard to side (right or left); when differences in preservation quality were present between sides (e.g. due to cortical bone erosion or presence of dried connective tissue), data collection occurred on the side which allowed for a more accurate measurement. Descriptions of these measurements, which follow the guidelines in McHenry and Corruccini (1978), McHenry (1992), and McHenry and Berger (1998) are given below:

1. HUMHEAD—Maximum anteroposterior diameter of humeral head taken perpendicular to the shaft axis.
2. ELBOW—The product of capitular height and articular width of the distal humerus. Capitular height is taken from the anteroproximal border of capitulum to the distoposterior border along the midline. Articular width is taken across the anterior aspect of articular surface from the lateral border of the capitulum to the medial edge of the articular surface.
3. RADTV—The mediolateral diameter of the radial head.
4. FEMHEAD—Maximum superoinferior diameter of the femoral head.
5. FEMSHAFT—The product of the anteroposterior and transverse diameters of the femoral shaft taken just inferior to the lesser trochanter.
6. DISTFEM—The product of the biepicondylar and shaft anteroposterior diameter of the distal femur.
7. PROXTIB—The product of the anteroposterior and transverse diameters of the proximal tibia. A-P is taken with one jaw of the calipers on the line connecting the posterior surfaces of the medial and lateral condyles and the other jaw on the most distant point on the medial condyle. Transverse diameter is the distance between the most lateral point on the lateral condyle to the most medial point on the medial condyle (perpendicular to the A-P diameter).
8. DISTTIB—The product of the anteroposterior and transverse diameters of the distal tibia. A-P diameter is the distance between the most anterior and posterior points of the talar facet in the A-P plane. Transverse diameter is the distance between the midline of the medial malleolus and the midline of the most medial point of the talar facet before the fibular facet begins.

To ensure that measurements taken on the extant materials were collected in the same manner as had been described in the previous publications (McHenry, 1992; McHenry and Berger, 1998), they were first replicated on a high quality cast of A.L. 288-1 housed at NMNH. Only published fossil measurements were used for the analyses in the present study (McHenry, 1992; McHenry and Berger, 1998), and all specimens are from Hadar. The square root was taken of area measurements so that all measurements used in this study are of the same dimensionality (i.e., linear). The resulting sample is complete with respect to measurements for the comparative species, but is missing data for *A. afarensis* (Tables 1 and 2).

Measuring dimorphism

SD has been measured in many ways, but typically when the sex of each specimen is known it is measured as the ratio of mean male size to mean female size, or as the logarithm of that ratio if it is to be used in statistical analyses (Smith, 1999). When sexes are unknown, a variety of methods have been used to estimate SD, including max/min ratio (e.g., Richmond and Jungers, 1995), mean method ratio (e.g., Simons et al., 1999), method of moments (e.g., Josephson et al., 1996), coefficient of variation (e.g., Leutenegger and Shell, 1987; Lockwood et al., 1996), and the binomial dimorphism index (e.g., Reno et al., 2003). Each of these techniques is susceptible to error under various conditions, although simulation studies and studies of actual primate data have shown max/min ratios to be particularly poor estimators while mean method ratios are relatively good estimators (Plavcan, 1994; Rehg and Leigh, 1999).

Figure 1 presents SD and 95% confidence intervals for body mass, a variety of univariate linear measurements, and the geometric mean (GM) of those measurements in modern humans and great apes. Visual inspection of the logged male:female ratios (SD values in the left-hand,

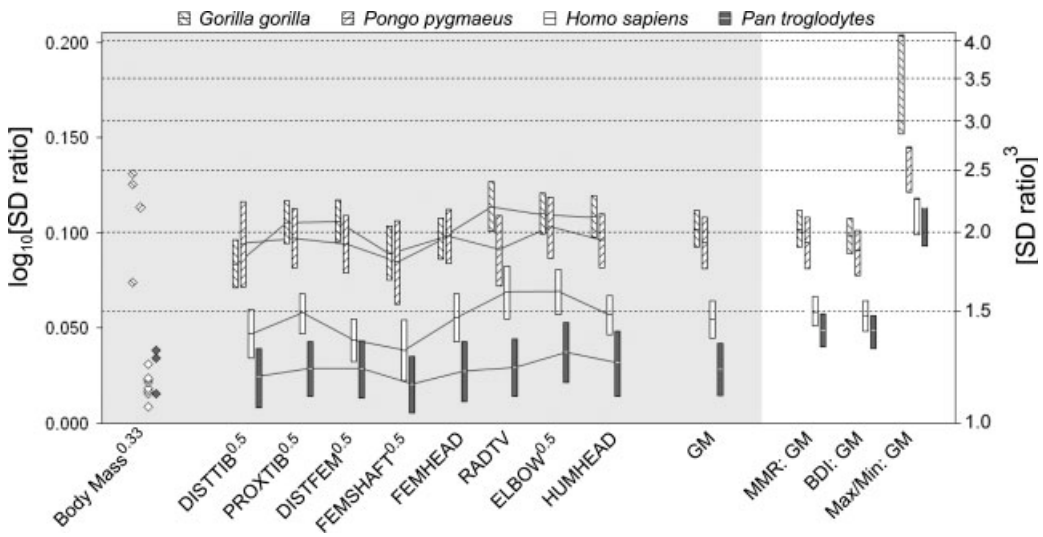

Fig. 1. Univariate and multivariate postcranial dimorphism for extant hominoids. Two standard measurements of sexual dimorphism (SD) are represented on the vertical axis: logged ratios on the left, and the cube of raw ratios on the right (cubing converts linear ratios to volumetric ratios for easier comparison to mass dimorphism). Horizontal solid lines within vertical bars represent observed values and bootstrapped 95% confidence intervals, respectively, for dimorphism ratios for each measurement (confidence intervals for max/min ratios have not been adjusted upwards for estimator bias). Dimorphism values in the gray left-hand portion of the figure are male:female ratios, values in the white right-hand portion of the figure are estimates of SD that do not presuppose knowledge of the sex of any specimen. Diamonds at far left represent population/subspecific body mass dimorphism calculated from Smith and Jungers (1997). The eight linear variables connected by line segments within each species are standard fossil hominin measurements (McHenry, 1992; McHenry and Berger, 1998) that were collected from extant taxa representing samples of the tibia, femur, radius, and humerus. GM is the geometric mean of those eight variables as calculated for each extant specimen. MMR:GM is the mean method ratio calculated for GM, BDI:GM is the binomial dimorphism index, and Max/Min:GM is the maximum/minimum ratio.

gray portion of the figure) shows that the postcranial measurements selected for this study tend to track similar levels of dimorphism within a taxon, and that the GM tracks the central tendency of each taxon. Notice that postcranial SD, whether considered univariately or multivariately (in the form of the GM) is similar to body mass SD in the great apes, but is markedly different from body mass SD in modern humans. This discrepancy will be addressed further in the Discussion.

Also shown in Figure 1 are three different estimates of logged SD for GM using mean method ratios, the binomial dimorphism index, and max/min ratios (SD values in the right-hand, white portion of the graph). Mean method ratios (MMR) estimate SD by finding the arithmetic mean of a set of values, then dividing the average of the values greater than the mean by the average of the values smaller than the mean. The binomial dimorphism index (BDI) calculates a weighted mean of all ratios for n measurements running from the mean of the $n - 1$ largest values divided by the smallest value on the one hand to the largest value divided by the mean of the $n - 1$ smallest value on the other. Max/min ratios estimate SD by dividing the largest value by the smallest value, ignoring the rest of the measurements in the data set.

Max/min ratios clearly overestimate actual postcranial SD, while the other two techniques perform well for gorillas and orangutans, slightly overestimate postcranial SD in modern humans, and greatly overestimate postcranial SD in chimpanzees (coefficient of variation estimates, not shown here because they do not produce ratio values directly, perform similarly to MMR and BDI). In general, techniques which split a data set and divide the mean of the larger values by the mean of the smaller values (such as MMR and BDI) perform well when all adult males are larger than all adult females, but tend to overestimate SD when male and female size ranges overlap (Fig. 1; Plavcan, 1994; Rehg and Leigh, 1999). However, estimates of SD for mildly dimorphic or monomorphic taxa are never as high as SD estimates for taxa as dimorphic as gorillas and orangutans using MMR or BDI. In addition, because these estimates act to decrease rather than increase the difference in SD between taxa, statistical tests of difference in SD based on either MMR or BDI will if anything be more conservative than tests based on actual SD.

Multivariate SD in this study is represented as the base 10 logarithm of MMR for the fossil and extant samples, although the methods described here are flexible and can also be used with any other technique for estimating SD ratios when sex is unknown (e.g., BDI or max/min ratios). Smith (1999) noted that distributions built from ratios of two positive values can be problematic in that they are bounded by zero on the low end and are unbounded at the high end; as such the distributions cannot be symmetrical and thus cannot be normal. In the same study, Smith (1999) also showed that log-transforming the ratios removes this constraint because ratios greater than one become positive, ratios less than one become negative and the logged ratio values are unbounded at both the high and low ends. Because MMR values are calculated by dividing a larger number by a smaller (as are BDI and max/min values), these ratios cannot be less than one and thus logged ratios cannot be negative. However, this study uses Monte Carlo procedures for significance testing (see below) and since these types of methods do not make any assumptions regarding the distribution of data (Manly, 1997), normality of the distributions is not required. In this

TABLE 3. Example showing mathematical equivalence of ratio of GMs and GM of ratios

Sex	HUMHEAD	ELBOW$^{0.5}$	RADTV	FEMHEAD	FEMSHAFT$^{0.5}$	DISTFEM$^{0.5}$	PROXTIB$^{0.5}$	DISTTIB$^{0.5}$	**GM**
F	45.6	34.9	21.1	37.5	29.6	44.7	49.0	24.2	*34.4*
F	49.3	35.4	26.4	40.0	28.0	48.6	53.2	25.1	*36.8*
F	47.2	37.9	27.1	40.6	27.0	50.3	56.0	27.1	*37.7*
F	51.6	38.7	26.7	40.9	31.3	50.3	55.1	27.5	*38.9*
F	50.7	37.7	28.9	43.6	31.0	52.2	58.9	29.0	*40.1*
M	54.4	43.1	29.2	47.7	33.9	57.1	67.3	29.5	*43.4*
M	62.1	46.3	32.1	48.5	37.3	58.5	64.8	34.3	*46.5*
M	63.0	46.0	36.6	50.6	36.5	62.1	69.4	30.7	*47.5*
M	65.1	46.5	35.0	52.1	40.0	62.9	71.4	31.0	*48.5*
M	64.3	49.3	36.6	54.1	39.9	64.9	74.4	34.4	*50.4*
SD	*1.26*	*1.25*	*1.30*	*1.25*	*1.28*	*1.24*	*1.28*	*1.20*	***1.26***

Male:female ratios (SD) for each linear measurement and GMs of all measurements for each individual have been calculated for a subset of the *G. gorilla* data. All measurements are in mm; ratios are unitless. Multivariate dimorphism for this data set can be calculated as either the ratio of average male GM divided by average female GM, or as the geometric mean of the male:female ratios for each linear measurement. In either case the result is the same, the ratio of 1.26 shown in bold italics. Note that in all cases sex-specific means are calculated as geometric means, not arithmetic means; however, ratios of sex-specific arithmetic means are identical to the ratios of sex-specific geometric means shown here at three significant digits.

study we are using logged ratios because logging the data means that the magnitude of the difference between two ratios would be independent of whether the larger or the smaller value was in the numerator (Smith, 1999). Additionally, equal differences between log-transformed ratios are the same as equal proportional differences of the raw ratios (e.g., log[2] − log[1] = log[1] − log[0.5], just as 2/1 = 1/0.5); this allows comparison of proportional differences across the full range of observed logged ratios.

Multivariate dimorphism and missing data

Until now, GM approaches to multivariate SD have been limited to studies with no missing data because a comparison between two or more GM is only meaningful if all of the means are based on the same type and number of measurements. For example, it is meaningless to compare the GM of femoral head size, humeral head size, and tibial plateau size from one individual to the GM of only femoral head size and tibial plateau size from another. Thus a measure of SD cannot be calculated for a GM based on multiple measurements unless all measurements have been collected for all specimens in a sample. However, the identical value of multivariate SD can be calculated without generating GM values for each specimen.

It can be shown that the ratio of GM values (e.g., a ratio of average male GM size to average female GM size) is mathematically equivalent to the GM of ratios (e.g., the GM of the male:female ratio of femoral head size, the male: female ratio of humeral head size, and the male:female ratio of tibial plateau size) (Appendix). This relationship is demonstrated using a subset of the gorilla sample in Table 3. As shown here, the multivariate SD ratio of 1.26 can be arrived at in either of two ways: by first calculating the GM for each specimen and then calculating the ratio of male and female mean GM, or by first calculating male:female mean ratios for each linear variable and then calculating the GM of the ratios (Table 3). The only difference from traditional measures of SD is that ratios of means (such as mean male and female values for actual SD, or mean large and small values for MMR) must be based on geometric means rather than arithmetic means (Appendix). Ratios of geometric means tend to be extremely similar to ratios of arithmetic means; for our extant sample the ratios are identical within species to three significant digits.

The mathematical equivalence of a ratio of GM with the GM of ratios allows for a multivariate measure of SD to be calculated even when a taxon has missing data. Monte Carlo resampling approaches can be developed which reduce the number of observations in comparative taxa to those of incomplete fossil data sets and provide significance tests for difference in SD between taxa. Monte Carlo resampling methods such as randomization and bootstrapping have become common in recent years for significance tests of difference in size and shape variation between fossil hominins and extant taxa (e.g., Grine et al., 1993, 1996; Kramer, 1993; Kramer et al., 1995; Richmond and Jungers, 1995; Lague and Jungers, 1996; Lockwood et al., 1996, 2000; Arsuaga et al., 1997; Lockwood, 1999; Silverman et al., 2001; Lague, 2002; Richmond et al., 2002; Reno et al., 2003, 2005; Harvati et al., 2004; Dobson, 2005; Villmoare, 2005; Harmon, 2006; Skinner et al., 2006; Green et al., 2007). Two such resampling approaches are described below: one in which the observed measure of SD in a fossil sample is tested against resampled distributions of SD for comparative data, and one in which SD values are resampled in both comparative and fossil data sets for significance testing. Both procedures were coded by ADG for the statistical programming language *R*, version 2.4.1 (Ihaka and Gentleman, 1996).

Method 1: Single observation of dimorphism in fossil taxon

For each comparative species, an iterative resampling procedure was repeated 10,000 times (see Fig. 2). In the first step of each iteration, the set of individuals in the comparative species is resampled without replacement and without regard to sex, a number of times equal to the number of "individuals" in the fossil sample (Fig. 2, Step 1). For the purposes of this procedure, elements found in association are considered to belong to one individual (e.g., the A.L. 288-1 skeleton, the A.L. 128-1/129-1 partial leg) and all other elements are considered to belong to separate individuals. In some cases this procedure likely overestimates the number of individuals present in the fossil sample. However, dimorphism will be calculated separately for each measurement, then combined as the GM of those values; thus non-antimere multiple elements from a single individual will not have an "artificial" effect on the observed level of dimorphism

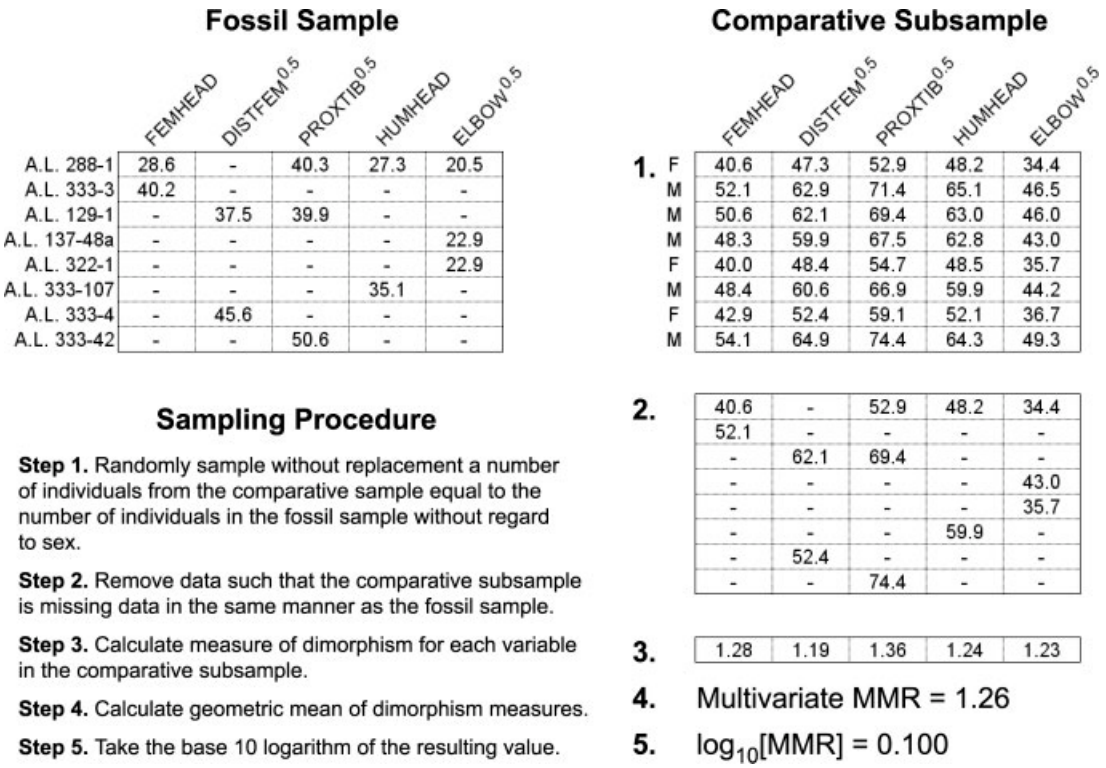

Fig. 2. Resampling procedure, single extinct observation method. The above process (illustrated with a subset of the measurements and the fossil sample due to space restrictions) is repeated 10,000 times for each comparative taxon to produce randomized distributions of log MMR values for significance tests against the observed value of log MMR in the *A. afarensis* sample.

present in a fossil sample, in that no individual will be represented more than once in the calculation of dimorphism for any given measurement.

After the individuals had been sampled from the comparative sample, their data matrix was reduced to resemble that of the fossil sample (Fig. 2, Step 2). For example, A.L. 288-1 in the fossil sample has complete data except for $DISTFEM^{0.5}$, so one individual in the comparative subsample retains all of its measurements except for $DISTFEM^{0.5}$. Similarly, A.L. 333-42 only has data for $PROXTIB^{0.5}$, so one individual in the comparative subsample retains data only for $PROXTIB^{0.5}$ (Fig. 2, Step 2). After data reduction, MMR was calculated for each measurement (Step 3), multivariate MMR was calculated as the GM of MMR values (Step 4), and then logged (Step 5). Steps 1 through 5 were then repeated 10,000 times for each comparative species to produce a distribution of resampled logged multivariate MMR values for each taxon.

The resampling procedure produced log MMR values based on identical sample sizes for all taxa, fossil and comparative. This allowed for distributions of log MMR to be calculated for extant taxa that are directly comparable with log MMR as calculated for the fossil sample. Reducing the sample size of comparative taxa in each iteration to that of fossil samples and then further reducing comparative samples by removing data to match missing data in the fossil sample led to this procedure producing distributions of log MMR with greater variance than would a procedure that simply resampled complete individuals a number of times equal to the fossil sample. Accordingly, the procedure used in this study was a more conservative test of differences in dimorphism than a randomization test that did not account for differences in sample size and data structure.

Because we are interested in whether or not *A. afarensis* is more dimorphic than extant taxa, significance tests were one-sided. The number of resampled log MMR values for a particular comparative species that were equal to or greater than the observed value of log MMR in the fossil sample were divided by the total number of resampled values (i.e., 10,000) to produce the *P* value. As noted above, this Monte Carlo method makes no assumptions regarding the distribution of the data (i.e., it is a nonparametric test) (Manly, 1997). This method shall be referred to hereafter as the single extinct observation method.

Method 2: Resampled distribution of dimorphism in fossil taxon

Like previous studies (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Reno et al., 2003; Harmon, 2006), the single extinct observation method compares the observed level of dimorphism in the fossil sample to resampled distributions of dimorphism in comparative samples. However, because SD is measured independently for each linear measurement before the GM of ratios is calculated, fossil samples are large enough to generate resampled distributions themselves. For example, using the *A. afarensis* sample in this study, multivariate MMR can be calculated for over 1.5 billion unique combinations of fossil measurements using a bootstrap procedure on the fossil sample. Comparing

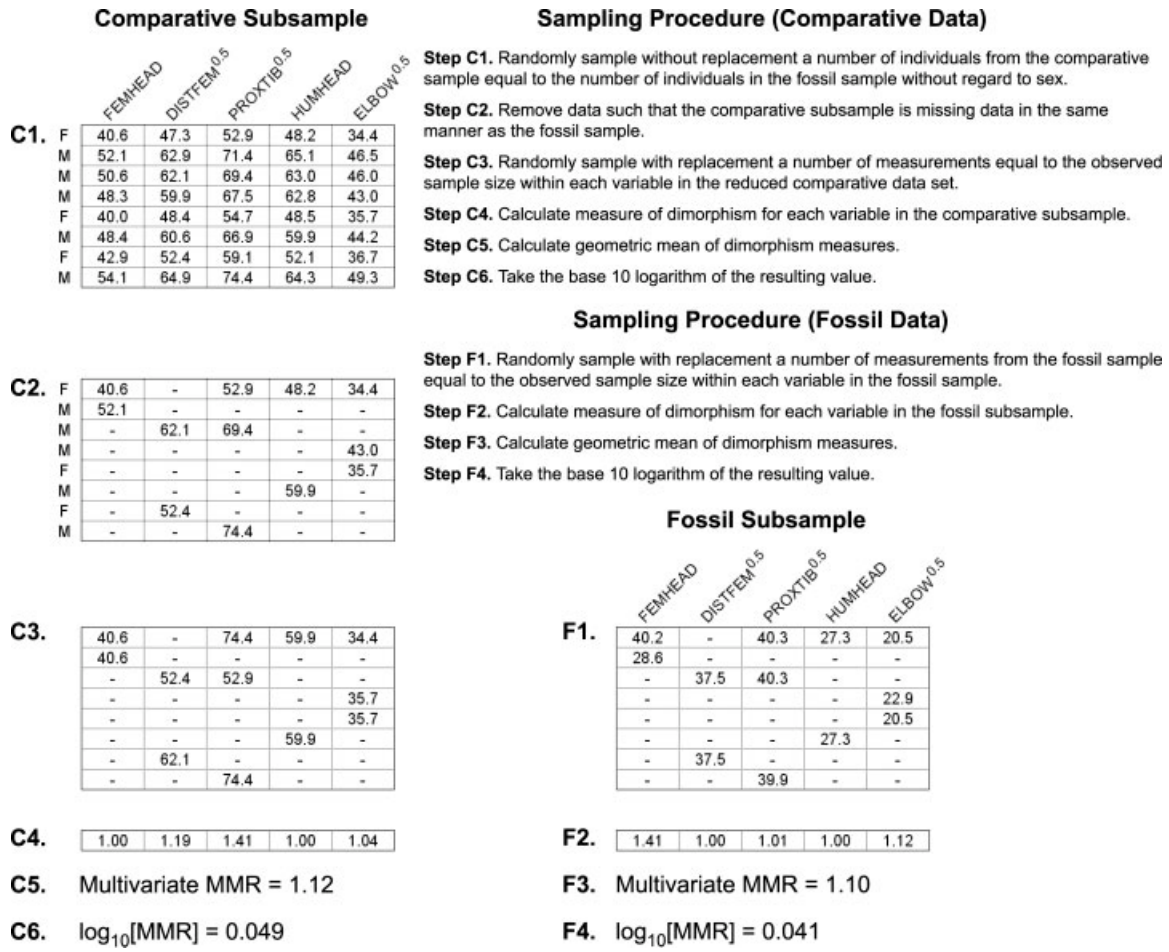

Fig. 3. Resampling procedure, resampled extinct distribution method. The above process (illustrated with a subset of the measurements and the fossil sample due to space restrictions) is repeated 10,000 times for each comparative taxon (Steps C1–C6) and the fossil sample (Steps F1–F4) to produce randomized distributions of log MMR values for significance tests of difference in log MMR between extant species samples and the *A. afarensis* sample.

resampled distributions of extant values against resampled distributions rather than single observations of fossil values provides a much more conservative test of difference in multivariate SD and is more likely to find overlap in fossil and extant dimorphism between taxa. Comparable distributions for comparative and fossil samples are generated using a bootstrap technique, in which resampling occurs with replacement (Efron and Tibshirani, 1993; Manly, 1997).

As above, for each comparative species an iterative resampling procedure was used (see Fig. 3). Steps C1 and C2 are identical to Steps 1 and 2 in the single extinct observation method, in which the comparative sample was sampled a number of times equal to the number of "individuals" in the fossil sample, and the resulting data matrix was reduced to resemble the fossil data matrix. At this point a bootstrap step was introduced in which values were resampled with replacement within each measurement a number of times equal to the total number of values present for that measurement in the sparse data matrix (Fig. 3, Step C3). MMR was then calculated for each measurement (Step C4), multivariate MMR was calculated as the GM of MMR values (Step C5), and that value was then logged (Step C6). Steps C1 through C6 were repeated 10,000 times for each comparative species to produce a distribution of resampled logged multivariate MMR values for each taxon.

In addition, a bootstrap procedure was applied to the fossil sample. Just as in Step C3 for the comparative sample, values were resampled with replacement within each measurement a number of times equal to the total number of values present for that measurement (Fig. 3, Step F1). As above, MMR was calculated for each measurement (Step F2), multivariate MMR was calculated as the GM of MMR values (Step F3), and that value was then logged (Step F4). Steps F1 through F4 were then repeated 10,000 times to produce a distribution of resampled logged multivariate MMR values for the fossil sample. This procedure thus generated a total of 10,000 resampled logged MMR values for each species in the analysis, fossil and comparative alike.

As with the single extinct observation method, this method produced log MMR values based on identical sample sizes for all taxa. In addition, this method used identical procedures for comparative and fossil samples to resample with replacement within sparse data matrices, allowing for directly comparable distributions of log MMR between all taxa. Significance tests for pairwise comparisons of log MMR between the fossil sample and

comparative taxa were conducted using a randomization procedure that calculated the difference between randomly paired log MMR values from each pair of taxon-specific distributions, where the difference was always calculated as [*A. afarensis*-comparative species] and thus could take on positive or negative values. One-tailed *P* values were calculated as the number of differences equal to or less than zero (i.e., *A. afarensis* equally dimorphic or less dimorphic than the comparative species) divided by the total number of differences (i.e., 10,000). As above, this randomization method is a nonparametric test that makes no assumptions regarding the distribution of the data (Manly, 1997). This method shall be referred to hereafter as the resampled extinct distribution method.

Constrained analysis for smallest specimen

Some may think that because A.L. 288-1 and A.L. 128-1/129-1 are probably the two smallest specimens in our sample and are represented by multiple measurements, those measurements should be constrained to be represented by similarly small members of the comparative extant taxon for each iteration of the resampling procedure. Although the resampled extinct distribution method addresses this concern (see "Methodological issues" in the Discussion), we repeated our analyses for both methods using the following constraint.

First, the overall size for each comparative specimen was calculated as the GM of all measurements for each specimen. Then, in each iteration of the resampling procedure after Step 1 for the single extinct observation method (see Fig. 2) and Step C1 for the resampled extinct distribution method (see Fig. 3), the two smallest specimens in each resampled subset were assigned to the spots in the data matrix corresponding to A.L. 288-1 and A.L. 128-1/129-1 (i.e., the first two rows in Matrix 1 in Fig. 2 and in Matrix C1 in Fig. 3). Afterwards, Steps 2–5 (see Fig. 2) were followed for the single extinct observation method and Steps C2–C6 and F1–F4 (see Fig. 3) were followed for the resampled extinct distribution method.

The rationale for this procedure is as follows: A.L. 288-1 and A.L. 128-1/129-1 are probably the smallest individuals in our fossil sample of *A. afarensis*, which in turn was drawn from a larger population of living *A. afarensis* individuals, most of whom were not fossilized (i.e., the populations of *A. afarensis* living in the same general time and place as our fossil sample). Similarly, the revised procedure imposed a constraint so that the comparative specimens that represented the measurements found in A.L. 288-1 and A.L. 128-1/129-1 in each iteration were the smallest specimens in a sample drawn from a larger population of measurements for that species (i.e., the full data set for each comparative taxon). We present results for both the constrained and unconstrained (hereafter referred to as "standard") analyses using the single extinct observation method and the resampled extinct distribution method.

Multiple representation of single specimens

Another concern is that results might be affected by multiple measurements from single individuals that are mistakenly thought to represent multiple individuals (as distinct from known multiple measurements which are already accounted for by the standard sampling procedure as in the case of A.L. 288-1 or A.L. 128-1/129-1). For example, there is a high probability that multiple measurements from the A.L. 333 site come from fewer individuals than there are elements, perhaps as few as five individuals (Plavcan et al., 2005), and the strict minimum number of individuals for the A.L. 333 sample in this analysis is three. The methods described above should be less sensitive than other methods to this problem of unknown number of individuals because dimorphism is first calculated separately for each measurement, which prevents multiple contributions from single individuals to the dimorphism calculated for a particular measurement except in the case of antimeres. Here we describe a procedure developed to test this assertion.

Because the measurements from A.L. 333 may come from as few as five individuals, we randomly generated 1,000 distinct fossil data matrices in which the twelve specimens from A.L. 333 were combined into five hypothetical specimens. In other words, there are many ways that five individuals could have contributed the particular elements found in the A.L. 333 fossil assemblage, and we generated 1,000 such possibilities. For example, the HUMHEAD measurement of A.L. 333-107 and the FEMHEAD and FEMSHAFT measurements of A.L. 333-3 could be combined, producing a single hypothetical specimen with those three measurements. The A.L. 333 measurements (as five individuals) were then combined with the five from other Hadar sites, thereby including all the fossil measurements, but reducing the total number of fossil individuals from 17 to 10. Next, 10 individuals were randomly selected from each extant sample, and the measurements were selected from them in the same pattern as they were for the fossil configuration (e.g., HUMHEAD, FEMHEAD, and FEMSHAFT taken from one individual representing the hypothetical "individual" A.L. 333-107/333-3), and the various methods described previously were applied.

Although it is very computationally intensive (i.e., 1,000 data sets resampled 10,000 times producing millions of sampling events), the process of repeating the analysis for 1,000 different ways in which five individuals could have contributed the thirteen measurements of the A.L. 333 sample is necessary for the following reason. Consider an example in which measurements from three fossil elements (e.g., humerus, femur, and tibia) are known to be from two fossil individuals, but it is not known which two elements came from a single individual. There are three possible pairs of individuals that could produce this sample:

1. Individual A: humerus and femur; Individual B: tibia
2. Individual A: humerus and tibia; Individual B: femur
3. Individual A: femur and tibia; Individual B: humerus

When comparing this sample with extant taxa, the particular case (1, 2, or 3) is important because the value of measurements sampled from any pair of extant individuals will differ depending on the case. For example, for the same pair of extant individuals, different values for these measurements will results when elements are divided between individuals as in Case 1 (humerus and femur measurements come from one individual and tibia from the other) or Case 2 (humerus and tibia measurements come from one individual and femur from the other). Thus test results are generated for a wide range of the possible ways in which the *A. afarensis* sample could include thirteen A.L. 333 measurements drawn

TABLE 4. Multivariate log_{10} dimorphism and P*-values for test of difference between* A. afarensis *and comparative taxa*

			Single extinct observation method				Resampled extinct distribution method			
Species	Actual SD	Actual MMR	Standard MMR mean	Standard MMR *P*-value	Constrained MMR mean	Constrained MMR *P*-value	Standard MMR mean	Standard MMR *P*-value	Constrained MMR mean	Constrained MMR *P*-value
A. afarensis	–	–	0.111 (–)	–	0.111 (–)	–	0.076 (0.018)	–	0.076 (0.018)	–
Gorilla gorilla	0.101	0.101	0.080 (0.016)	**0.034**	0.095 (0.016)	0.149	0.057 (0.017)	0.219	0.067 (0.018)	0.360
Pongo pygmaeus	0.095	0.095	0.076 (0.013)	**0.003**	0.087 (0.012)	**0.019**	0.055 (0.015)	0.194	0.063 (0.016)	0.297
Homo sapiens	0.055	0.058	0.050 (0.010)	**<0.001**	0.059 (0.010)	**<0.001**	0.036 (0.010)	**0.028**	0.042 (0.011)	0.055
Pan troglodytes	0.028	0.049	0.044 (0.010)	**<0.001**	0.058 (0.010)	**<0.001**	0.032 (0.010)	**0.017**	0.041 (0.011)	0.054

Actual SD is the logged male:female ratio and actual MMR is the logged mean method ratio measured for each complete comparative taxon. For the resampling tests, the means of resampled distributions and *P*-values for significance of difference in MMR from *A. afarensis* are reported. Standard deviations of resampled distributions are given in parentheses next to the means. The single extinct observation method compares a single measurement of dimorphism from the fossil sample to distributions from comparative taxa as previous studies have done (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Reno et al., 2003; Harmon, 2006), while the resampled extinct distribution method is a more conservative method that generates a distribution of values from the fossil sample, unlike previous studies. Results using both versions of these methods (standard tests and tests constrained to use the smallest specimens in the sample to represent A.L. 288-1 and A.L. 128-1/129-1) are shown. The value given as the *A. afarensis* MMR mean for the single extinct observation method is the single observed MMR value in the fossil sample. Differences significant at $\alpha = 0.05$ are indicated in bold.

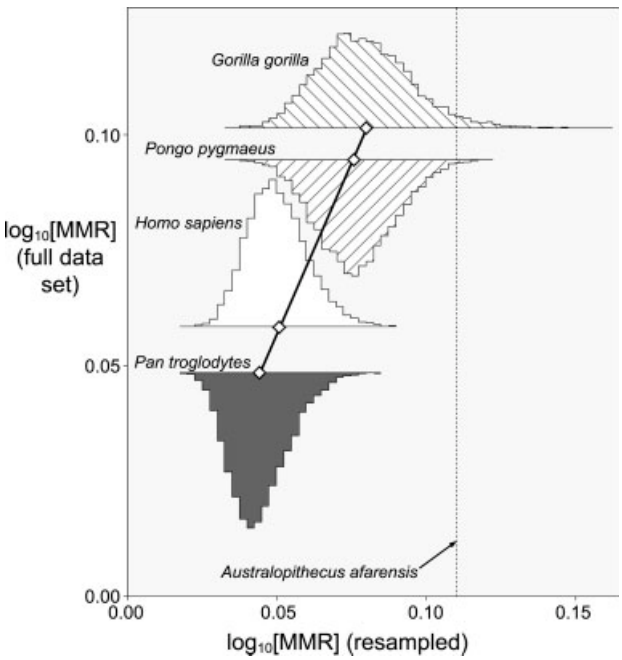

Fig. 4. Means of randomized log MMR distributions versus observed log MMR values in the comparative sample using the standard version of the single extinct observation method (open diamonds). Solid diagonal line gives regression line. Superimposed are histograms showing the distribution of randomized log MMR values. Some histograms are inverted for better visibility. The observed value of log MMR in the *A. afarensis* sample is given as a dashed line to show how much of each of the resampled distributions exceed it.

from five individuals. All of the methods described above (i.e., standard and constrained versions of both the single extinct observation method and the resampled extinct distribution method) were repeated for each of these 1,000 fossil data matrices, and *P*-value ranges were compared with the observed *P*-value for tests based on the original fossil data matrix to determine whether the number of individuals at A.L. 333 has a marked effect on inferred dimorphism in *A. afarensis*.

RESULTS

Single extinct observation method

Distributions of log_{10}[MMR] were calculated for each extant species using the single extinct observation method. Regardless of whether the standard or constrained version of this method is used, results are qualitatively similar (Table 4). Among comparative samples, gorillas have the highest mean, followed by orangutans, humans, and chimpanzees, the same ranking as when MMR or SD is measured for the full comparative samples (Table 4). Although resampled MMR means differ from observed MMR values for the full sample due to the missing data resampling procedure, the correlation between the two sets of values is quite high (standard, $R^2 = 1.000$; constrained, $R^2 = 0.980$). Relative dimorphism difference between taxa is preserved for the standard analysis: as shown in Figure 4, values fall directly on the reduced major axis regression line of

TABLE 5. Ranges of P*-values for analyses of reduced number of fossil individuals*

Species	Single extinct observation method		Resampled extinct distribution method	
	Standard MMR *P*-value	Constrained MMR *P*-value	Standard MMR *P*-value	Constrained MMR *P*-value
G. gorilla	0.037–0.108 (0.034)	0.135–0.276 (0.149)	0.215–0.244 (0.219)	0.322–0.350 (0.360)
P. pygmaeus	0.004–0.052 (0.003)	0.030–0.183 (0.019)	0.187–0.218 (0.194)	0.269–0.303 (0.297)
H. sapiens	<0.001–<0.001 (<0.001)	<0.001–<0.001 (<0.001)	0.029–0.041 (0.028)	0.048–0.069 (0.055)
P. troglodytes	<0.001–<0.001 (<0.001)	<0.001–<0.001 (<0.001)	0.015–0.025 (0.017)	0.038–0.053 (0.054)

Measurements from A.L. 333 were combined randomly to produce 1,000 distinct fossil data matrices in which those measurements represent five individuals rather than twelve; all four analyses (with 10,000 resampling events each) were performed for each of the 1,000 fossil matrices. All measurements were included in each data matrix. Minimum and maximum *P*-values are reported, with *P*-values for the original fossil sample as reported in Table 4 given in parentheses.

actual $\log_{10}$[MMR] values against resampled standard $\log_{10}$[MMR] means for the comparative sample.

Also shown in Figure 4 are histograms for the resampled distributions of $\log_{10}$[MMR] for the comparative taxa, as well as a vertical line indicating the single measurement of $\log_{10}$[MMR] observed in the *A. afarensis* sample. The degree of multivariate dimorphism observed in the fossil sample is higher than all of the resampled values for *P. troglodytes* and *H. sapiens*, and is equaled or exceeded only by the uppermost portions of the *P. pygmaeus* and *G. gorilla* distributions. This is also true for the constrained analysis. One-tailed significance tests for the single extinct observation method find that this sample of *A. afarensis* is significantly more dimorphic than all of the comparative species in the standard analysis and all but *G. gorilla* in the constrained analysis (Table 4).

When the number of individuals contributing measurements to the A.L. 333 site is adjusted from twelve to five for 1,000 randomly combined fossil data sets, the range of *P*-values provides close agreement with the results found for the original fossil data set (Table 5). While in some cases dimorphism is not significantly greater in *A. afarensis* than in *G. gorilla* or *P. pygmaeus*, in all cases (for both standard and constrained analyses) *P*-values for significance tests of dimorphism in *A. afarensis* versus both *H. sapiens* and *P. troglodytes* are all below 0.001 (Table 5).

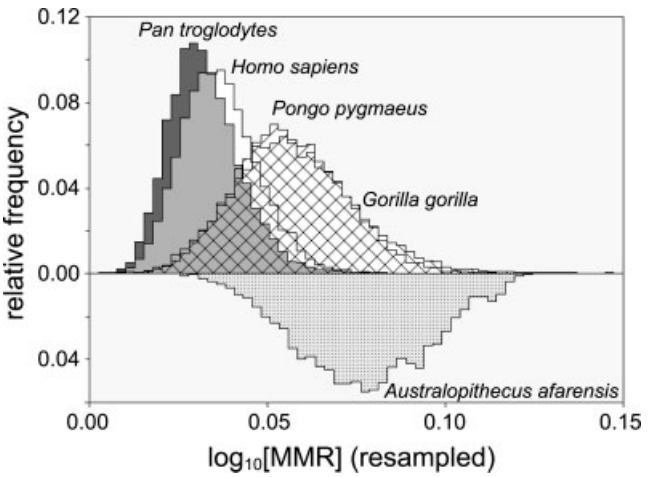

Fig. 5. Histograms of Monte Carlo distributions of log MMR for the *A. afarensis* and comparative samples using the standard version of the resampled extinct distribution method. The *A. afarensis* histogram is inverted for ease of comparison with extant histograms.

Resampled extinct distribution method

Using the resampled extinct distribution method, distributions of $\log_{10}$[MMR] were calculated for each extant species as well as for *A. afarensis* (see Fig. 5). Dimorphism rank follows that of MMR or SD as measured in the full comparative sample, the correlation between full sample $\log_{10}$[MMR] and resampled $\log_{10}$[MMR] is quite high (standard, $R^2 = 0.999$; constrained, $R^2 = 0.987$), and relative dimorphism difference between taxa is preserved in the standard analysis. As expected, generating a resampled distribution of *A. afarensis* dimorphism rather than using a single observation dramatically increases the overlap of possible dimorphism values between taxa. Whereas the observed value of $\log_{10}$[MMR] in *A. afarensis* overlapped with only a small proportion of the gorilla and orangutan distributions and not at all with human and chimpanzee distributions using the single extinct observation method, all distributions overlap substantially using the resampled extinct distribution method (see Fig. 5). This more conservative test finds no significant difference in dimorphism between *A. afarensis* and *G. gorilla* or *P. pygmaeus*; however, dimorphism in *A. afarensis* remains significantly higher than that of both *H. sapiens* and *P. troglodytes* for the standard analysis and borders on significance ($P = 0.055$ and 0.054, respectively) for the constrained analysis (Table 4).

It is useful to note that in the constrained analysis, the *P*-value for a one-tailed test of dimorphism in *G. gorilla* greater than that of *H. sapiens* is 0.126, and the *P*-value for *P. pygmaeus* greater than *H. sapiens* is 0.142. Thus, at the $\alpha = 0.05$ level, the constrained version of the resampled extinct distribution method is so conservative that it is unable to differentiate the degree of dimorphism present between gorillas and humans. A useful common statistical concept in this case is that of relative risk: the probability that *A. afarensis* is more dimorphic than *H. sapiens* (i.e., $1 - 0.055 = 0.945$), divided by the probability that another species (e.g., *G. gorilla*) is more dimorphic than *H. sapiens* (i.e., $1 - 0.126 = 0.874$). Values of one indicate that *A. afarensis* and *G. gorilla* are equally likely to be more dimorphic than *H. sapiens*, values greater than one indicate that *A. afarensis* is more likely than *G. gorilla* to be more dimorphic than humans, and values less than one indicate the reverse.

In all four combinations of standard and constrained applications of the single extinct observation method and the resampled extinct distribution method, relative risk as applied here is greater than one for comparisons with both *G. gorilla* and *P. pygmaeus* (Table 6). Therefore, all four tests agree that the probability that *A. afarensis* is

TABLE 6. Relative risk values for comparisons of dimorphism between A. afarensis *and* G. gorilla*, and* A. afarensis *and* P. troglodytes

	Single extinct observation method		Resampled extinct distribution method	
Relative risk comparison	Standard	Constrained	Standard	Constrained
$P[MMR_{A.\ afarensis} > MMR_{H.\ sapiens}]/P[MMR_{G.\ gorilla} > MMR_{H.\ sapiens}]$	1.055	1.025	1.136	1.082
$P[MMR_{A.\ afarensis} > MMR_{P.\ troglodytes}]/P[MMR_{G.\ gorilla} > MMR_{P.\ troglodytes}]$	1.024	1.022	1.089	1.076
$P[MMR_{A.\ afarensis} > MMR_{H.\ sapiens}]/P[MMR_{P.\ pygmaeus} > MMR_{H.\ sapiens}]$	1.054	1.031	1.133	1.102
$P[MMR_{A.\ afarensis} > MMR_{P.\ troglodytes}]/P[MMR_{P.\ pygmaeus} > MMR_{P.\ troglodytes}]$	1.022	1.027	1.081	1.089

Relative risk values are the probability that species *A* is more dimorphic than species *C* divided by the probability that species *B* is more dimorphic than species *C*. For all analyses, *A. afarensis* is more likely than both *G. gorilla* and *P. pygmaeus* to be more dimorphic than *H. sapiens* and *P. troglodytes* (i.e., all relative risk values are greater than one).

more postcranially dimorphic than modern humans is greater than the probability that gorillas and orangutans are more dimorphic than modern humans.

Furthermore, reducing the number of individuals contributing measurements to the A.L. 333 site from twelve to five has no qualitative effect on these results. The range of *P*-values generated for 1,000 randomly combined fossil data sets show that dimorphism in *A. afarensis* is always significantly greater than that of *H. sapiens* and *P. troglodytes* in the standard analysis, and is close to significance for the constrained analysis (Table 5). In all cases, relative risk is greater than one.

DISCUSSION

Methodological issues

The methods described and applied here significantly advance our ability to extract information preserved in the fossil record. These methods improve our ability to incorporate data from multiple disparate fossil elements into a single analysis of relative SD. Resampling tests such as those developed here and those used in earlier studies (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Reno et al., 2003; Harmon, 2006) do not attempt to assign confidence intervals to the actual degree of dimorphism for a fossil taxon; instead, they test whether or not the degree of dimorphism observed in a fossil sample is greater than that seen in comparable samples from comparative taxa. Here the resampling process reduces the comparative samples to sparse data matrices mimicking the fossil samples so that multiple measurements from single fossil individuals are also represented by single individuals from the comparative taxa, making SD measures directly comparable between fossil and comparative samples. As shown above, means of resampled distributions of log MMR using either standard method are not equal to actual log MMR but are nearly perfectly correlated with actual log MMR observed for full data samples, and thus proportionality in log SD has been preserved (i.e., if the difference in log MMR between species *A* and *B* is twice the difference between species *A* and *C* for the full comparative sample, then the same is true for means of the resampled distributions for those species). This is because the procedure of reducing the comparative data to match the structure of the sparse data matrix for *A. afarensis* alters the observed level of dimorphism, but does so in a consistent manner across all taxa. Thus any altering of dimorphism ratios that occurs in the fossil sample due to the particulars of missing data occurs equally in all comparative samples that it is being compared with, and has no impact on significance tests.

An additional property of these new methods is that data can be included from any fossil element, regardless of whether a template specimen exists that preserves that element or not. However, scaling considerations should be taken into account before an element is included in any analysis of multivariate SD, regardless of whether multivariate SD is estimated using the methods described here or other techniques such as the template method. We have limited this study to measurements that tend to scale isometrically with body mass (and thus also with each other) in primates; however, see "Postcranial dimorphism vs. body mass dimorphism" below for further discussion.

The methodological issues surrounding the various procedures used in this study are now considered in detail below.

Single extinct observation method vs. resampled extinct distribution method. All previous resampling studies of SD in *A. afarensis* have compared single estimates of SD as observed in the fossil sample with distributions for comparative taxa (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Reno et al., 2003; Harmon, 2006). We provide a comparable method here (single extinct observation method). However, because such methods do not account for the prior probability that a sample drawn from *A. afarensis* as a whole would produce the level of dimorphism observed in a particular fossil sample, comparing a single measure of SD in a fossil sample against distributions for extant samples effectively assumes that the observed SD value for the fossil sample is the exact SD value for the larger population of living *A. afarensis* individuals that the sample was drawn from, a highly unlikely prospect. When sample sizes are large enough, extant data and fossil data can be resampled to generate Monte Carlo distributions for all taxa, which allows for more robust statistical tests than comparing against a single extinct observation (resampled extinct distribution method). A rule of thumb regarding when it is feasible to use a resampling procedure such as our resampled extinct distribution method is when the number of unique combinations of the data is at least 10 times as large as the number of iterations performed (Chernick, 1999). When that condition is met, resampling with replacement from a sample provides a distribution for the central tendency of the parameter in question, which in this case is log MMR. Therefore, if the procedure is iterated 10,000 times, it is feasible to use the resampled extinct distribution method if a fossil

sample can produce at least 100,000 unique combinations of the data. To determine the number of unique combinations for a data set, first calculate the number of unique combinations for each measurement as

$$(2n - 1)!/(n![n - 1]!),$$

where n is the number of specimens for a given measurement, then find the product of these values for all measurements. Because of the multiplicative nature of this value, the resampled extinct distribution method will be applicable in most cases except those where the fossil data set is very small and made up of only a few types of measurements. As noted earlier, the number of unique combinations possible for the sample in this study is more than 1.5 billion.

In general, the resampled extinct distribution method should be preferred to the single extinct observation method because it provides a more conservative test of difference in multivariate SD. However, the advantages of using the resampled extinct distribution method diminish when sample sizes are low because variances of resampled distributions for both fossil and comparative data are responsive to several factors, which will affect significance tests. For example, resampling will produce levels of dimorphism in the resampled distributions that are both very high (when resampling randomly selects only the highest and lowest measurements in a data set) and very low (when resampling randomly selects only a single value multiple times, resulting in apparent monomorphism), as well as more moderate values. Extreme values will occur more often when sample sizes are low for a particular measurement in the fossil sample. For example, if the fossil sample has only two values for a given measurement, the same value will be selected 2 out of 3 times when sampling with replacement, producing a monomorphic ratio of one. This applies equally to resampled distributions for both fossil and comparative samples, so variance of all of these distributions will increase as sample size decreases in one or more measurements in the fossil sample. Therefore, at low sample sizes the resampled extinct distribution method may not find significant difference in SD between a fossil sample and any comparative taxon; however, it may also not find a significant difference between any of the comparative taxa (e.g., *Gorilla gorilla* and *Homo sapiens*). To assess the probability that a Type II error has occurred in the event that a fossil sample is found not to differ significantly from any of the comparative taxa, the significance test can be repeated for the two extant species showing the most extreme difference in SD. This test is performed in the same way it is calculated for the fossil sample. For example, randomly pair values from the African ape resampled distributions and calculate [*G. gorilla-P. troglodytes*] for each pair, then divide the number of differences that are equal to or less than zero by the total number of differences. If the resulting P value is not significant then a Type II error has been committed for these two taxa and may also have been committed when testing for difference from the fossil sample. In such a case the concept of relative risk is particularly useful, as it compares the relative probabilities of high dimorphism between *A. afarensis* and extant apes of known high dimorphism.

In this study, a Type II error has occurred using the most conservative test, the constrained version of the resampled extinct distribution method, in that both *G. gorilla* and *P. pygmaeus* were not found to have significantly higher dimorphism than either *H. sapiens* or *P. troglodytes*. Relative risk analysis in this case shows that although all P-values exceed 0.05 for comparisons of dimorphism for gorillas, orangutans, and *A. afarensis* against that of modern humans and chimpanzees, the probability that *A. afarensis* is more dimorphic than humans and chimpanzees exceeds the probability that gorillas and orangutans are more dimorphic than the other modern taxa.

Standard vs. constrained methods. To assess the potential effect of A.L. 288-1 and A.L. 128-1/129-1 being well-preserved compared with the other fossils, we restricted the comparative specimens which contribute those same multiple measurements to be two of the smallest individuals from our comparative sample in each iteration. However, randomization procedures are explicitly designed to test observed cases against the full range of possibilities inherent in the data structure of a sample (Manly, 1997). For example, previous applications of randomization techniques to studies of SD in *A. afarensis* (e.g., Richmond and Jungers, 1995; Lockwood et al., 1996; Reno et al., 2003; Harmon, 2006) do not constrain extant sampling procedures by size considerations. In all of these studies, resampling procedures allow each specimen equal probability of being included in any subsample, and in some iterations the subsample is made up of only females or only males. Thus fossil SD is compared against distributions of extant SD which include values calculated from subsamples made up of only one sex, even though such subsamples are clearly not representative of the sex ratio or size range present in the fossil sample. In the case of the procedures presented here, this means allowing all specimens in the sample equal probability of being selected as the specimens which contribute multiple elements in the same manner as A.L. 288-1 and A.L. 128-1/129-1.

For the *A. afarensis* sample used in this study, the presence of multiple measurements from A.L. 288-1 and A.L.128-1/129-1 makes it likely that the single observed value of dimorphism for our fossil sample does not represent the mean of possible *A. afarensis* dimorphism values. Two responses to this possibility present themselves: compare the single SD observation from the fossil record to distributions of SD values from comparative taxa that are forced to resemble the fossil sample in preservation of specimens of various size (e.g., the constrained single extinct observation method), or compare a distribution of SD values generated from the fossil sample to distributions of SD values from the comparative taxa (e.g., the standard resampled extinct distribution method). Combining both the size constraint and the resampled distribution for the fossil sample (e.g., the constrained resampled extinct distribution method) violates the principles of a randomization test and is overly conservative, as demonstrated by the failure of the constrained resampled extinct distribution method to find significant difference between the level of dimorphism present in gorillas and modern humans.

In general, we prefer the standard resampled extinct distribution method over the constrained single extinct observation method because we do not know the overall size (GM of all variables) of each specimen in our fossil sample due to missing data. For example, it is not clear if A.L. 288-1, A.L. 128-1/129-1, or perhaps even another specimen such as A.L. 322-1 or A.L. 137-48a has the

smallest overall size (GM) among the fossil specimens in our sample because the specimen with the smallest measurement for one variable does not necessarily have the smallest measurement for another variable (e.g., A.L.288-1 has a smaller FEMSHAFT measure but larger PROXTIB measure compared to A.L. 128-1/129-1). Instead of placing a constraint, based on incomplete knowledge of the fossil record, on which comparative specimens can contribute measurements, it is methodologically more appropriate for comparisons to be made between a distribution of dimorphism values for our fossil sample, rather than a single observation, and distributions from comparative taxa (Manly, 1997). This is what the standard version of the resampled extinct distribution method does. By resampling with replacement from within the fossil sample, in most iterations A.L. 288-1 and A.L. 128-1/129-1 are present in fewer than the 10 measurements they represent in the full sample, and for some iterations they are not present in the resampled population at all. Comparison of standard deviations among the distributions of resampled log MMR for *A. afarensis* and the comparative taxa using the standard version of the resampled extinct distribution method (Table 4) highlights this point: *A. afarensis* in fact has the highest standard deviation of all five species, indicating that the presence of multiple measurements from A.L. 288-1 and A.L. 128-1/129-1 is not restricting the distribution of resampled dimorphism values to a narrow range.

Of all methods presented here, the standard version of the resampled extinct distribution method is the most appropriate technique for application to multivariate data sets with missing data. That said, relative risk analyses (Table 6) demonstrate that regardless of method, all tests agree that *A. afarensis* is more likely than both *G. gorilla* and *P. pygmaeus* to be more dimorphic than both *H. sapiens* and *P. troglodytes*.

Comparison of results with previous studies

The results demonstrate a degree of postcranial multivariate SD in *A. afarensis* most consistent with the level of dimorphism present in gorillas and orangutans, a result that has been found in several previous studies of univariate and multivariate SD in the postcranium and elsewhere (e.g., Kimbel and White, 1988; Lockwood et al., 1996; Plavcan et al., 2005; Harmon, 2006). When the results here are compared with the most relevant work in the literature, studies of SD of actual postcranial dimensions [as opposed to dimorphism in estimated mass (e.g., McHenry, 1991a) or estimated postcranial dimensions (e.g., Reno et al., 2003)], we find complete agreement with earlier work: *A. afarensis* is observed to have postcranial SD greater than the mean postcranial SD observed in gorillas, orangutans, modern humans, and chimpanzees (Lockwood et al., 1996; Harmon, 2006). These results do not support a modern human-like or chimpanzee-like level of dimorphism as some other studies have suggested (e.g., Reno et al., 2003, 2005).

The difference in findings between this study and that of Reno et al. (2003, 2005) is based on at least two factors. First, sample inclusion criteria and measurements differed between studies, producing different fossil samples. For example, due to the requirement that variables scale isometrically with each other, fewer measurements were used in this study (26 measurements from 17 specimens in this study as opposed to 29 measurements from 29 specimens in Reno et al., 2003) which may account for differences in results, although it should be noted that the present study includes a much larger number of specimens than earlier studies of actual postcranial dimorphism which were restricted in sample size due to the requirement of only including specimens with complete data (e.g., Lockwood et al., 1996; Harmon, 2006). However, the inclusion of variables that do not scale isometrically with each other probably has a larger effect than differences in sample size, at least above very small sample sizes (see discussion of scaling in "postcranial dimorphism vs. body mass dimorphism" below).

Additionally, the two types of methods differ in their sensitivity to individuals being represented multiple times in the fossil sample (e.g., from the A.L. 333 site). As previous studies have shown, the template method is highly sensitive to this issue (Plavcan et al., 2005; Scott and Stroik, 2006; however, see Reno et al., 2005). This is because multiple measurements from the same individual produce similar values for the measurement being estimated (e.g., femoral head size). When dimorphism is assessed for that measurement, the value is artificially low because of the repeated measurements from one individual. The methods described in this article are less sensitive to such issues: although antimeres will cause the same problem as occurs in the template method, any other multiple elements from the same individual will not because dimorphism is evaluated separately for each measurement before overall SD is calculated as the GM of dimorphism ratios for each measurement. Thus, even when multiple elements are present, each individual contributes only one measurement to any one particular measure of dimorphism (e.g., dimorphism in HUMHEAD, FEMHEAD, etc.). The effect of such multiple representation on significance tests was demonstrated empirically by performing the entire set of analyses for 1,000 fossil data matrices in which the twelve specimens from A.L. 333 were combined into five hypothetical specimens. While there were some small differences in *P*-values between the original analysis and the analysis of the 1,000 alternative fossil matrices, results did not differ qualitatively from those of the original analysis with regard to the question of *A. afarensis* SD as compared with that of modern humans and chimpanzees; i.e., differences significant in the original analysis at $\alpha = 0.05$ were also significant for all 1,000 alternative fossil matrices (Table 5).

It has also been suggested that differences between Reno and others' work and earlier studies that found high values of SD in *A. afarensis* were due to bias resulting from variation in body size of the species over time and space (Reno et al., 2003). Indeed, recent studies have found temporal trends in dental and mandibular size between Hadar and Laetoli specimens (Kimbel et al., 2006) and within the younger portion of the Hadar fossil record (between 3.17 and 3.00 mya; Lockwood et al., 2000). In this study we have limited the sample spatially to the Hadar hominins, and temporally to specimens in the age range where Lockwood and others found no temporal trend in mandibular size (between 3.4 and 3.18 mya). In addition, our comparative sample is made up of multiple subspecies for all of the ape taxa, and multiple ethnic groups for the modern human sample. Thus there is greater variation in the comparative sample than there would be otherwise, which may or may not equal or exceed the effects of temporal variation on the fossil sample. What can be said definitively is the following: if in fact at any given time *A. afarensis* was only

as dimorphic as modern humans or chimpanzees, then the effect of temporal variation on *A. afarensis* postcranial size would have to be extreme relative to variation in postcranial size between subspecies and/or populations of extant great apes and humans to make the *A. afarensis* sample studied here appear as dimorphic as modern gorillas and orangutans. Future applications of the methods described here can be used to explicitly test whether even more temporally constrained samples of *A. afarensis* show similarly high levels of multivariate SD.

Given that all of the analyses in this study showed that our *A. afarensis* sample was more likely than both the gorilla and orangutan samples to be more postcranially dimorphic than both chimpanzees and modern humans (Table 6), we follow previous studies in inferring a moderate to high degree of male–male competition for mating opportunities in *A. afarensis* (e.g., Plavcan and van Schaik, 1997a; Plavcan, 2000), with the caveat that reconstructed behaviors in fossil taxa are complicated by a variety of factors (Plavcan, 2002). However, it is important to bear in mind that this and previous studies of dimorphism in *A. afarensis* examine postcranial SD. Many previous studies on living primates have identified behavioral and ecological correlates of body mass dimorphism (Clutton-Brock et al., 1977; Leutenegger and Kelly, 1977; Gaulin and Sailer, 1984; Cheverud et al., 1985; Kappeler, 1990, 1991; Leigh, 1992, 1995; Ford, 1994; Martin et al., 1994; Leigh and Shea, 1995; Mitani et al., 1996; Plavcan and van Schaik, 1997a,b; Smith and Cheverud, 2002; Gordon, 2004, 2006; Plavcan, 2004) and craniodental dimorphism (Leutenegger and Kelly, 1977; Kay et al., 1988; Greenfield, 1992; Plavcan and van Schaik, 1992, 1997a; Plavcan et al., 1995; Plavcan, 2004), but to date no studies have examined the relationship between postcranial dimorphism and variables such as mating behavior and social group construction. And although one study has examined the relationship between craniofacial and body mass dimorphism in primates (Plavcan, 2003), no such study has been conducted for the relationship between postcranial and body mass dimorphism.

Postcranial dimorphism vs. body mass dimorphism

As previous researchers have noted, skeletal dimorphism is not equivalent to body mass dimorphism (e.g., Plavcan, 2003; Plavcan et al., 2005; Harmon, 2006). Richmond and Jungers (1995) point out, and we show in Figure 1, modern human SD exceeds that of chimpanzee SD in most postcranial dimensions, but the reverse pattern is true for body mass. The relationship between SD in any two variables (e.g., body mass and femoral head diameter) is directly related to their scaling relationship. Consider Figure 6, in which simulated logged data for one variable (*Y*) has been plotted against simulated logged data for three other variables (*X*1, *X*2, and *X*3). Since log[M/F] = log[M] − log [F], log SD for any variable is equivalent to the mean of the logged data for males minus the mean of the logged data for females. As shown in Figure 6, when variables are of the same dimension and scale isometrically with regard to each other, they will have the same amount of SD. However, when the scaling relationship deviates from isometry, levels of SD will differ between variables measured from the same set of specimens (see Fig. 6). Thus if postcranial dimensions do not scale isometrically with body mass for all of the taxa and measurements included in a study, the resulting analyses of dimorphism will not provide reliable estimates of difference in body mass dimorphism regardless of whether the test is based on univariate or multivariate data sets, complete or missing data, GM or template values.

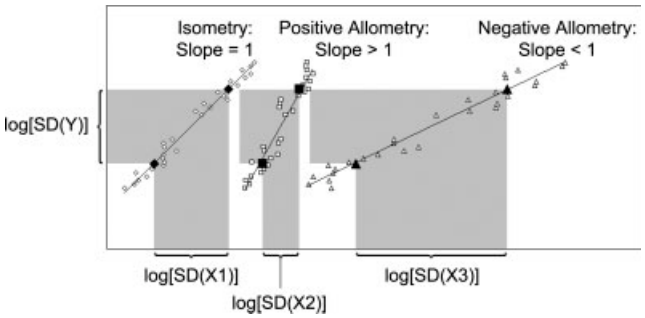

Fig. 6. Relationship between scaling and size dimorphism for pairs of variables (hypothetical data). When plotted in log space, the log of sexual dimorphism (SD) is the difference between the mean of male values and the mean of female values. For example, the length of the bracket along the *Y*-axis is equal to the log of the male:female ratio for variable *Y*, where the bracket indicates the distance between the sex-specific means (shown as closed symbols). When two variables of the same dimensionality (i.e., linear, area, or volume measurements) scale isometrically with each other they will have highly similar SD values (compare the length of the brackets for *Y* and *X*1); these SD values would be identical if there were no variation about the regression line. When there is positive allometry for the scaling of *Y* on *X*, the *X* variable will have a lower SD value than *Y* (compare brackets for *Y* and *X*2), while the reverse is true for negative allometry (compare brackets for *Y* and *X*3). Note that the slope of the scaling relationship can be estimated by the slope of a line passing through the female and male means, which is equivalent to log[SD(*Y*)]/log[SD(*X*)].

That said, not all methods respond in the same way to this problem. For example, it might be thought that deviations from isometric scaling between variables would affect the methods described here and the template method equally. However, a simple example will show that this is not the case. Consider 100 observations divided between 10 variables, where variables *X*1 through *X*9 have a dimorphism of 2.2 and variable *X*10 has a dimorphism of 1.2. If observations are distributed equally among all 10 variables (i.e., 10 measurements per variable) then both methods will perform similarly and return a dimorphism value near 2.07, the GM of the 10 variables. But if the variables are unevenly distributed, e.g., two measurements each for variables *X*1 through *X*9 and 82 measurements for variable *X*10, then the methods will provide different results.

The template method, which estimates the value of each measurement by using a ratio (which necessarily assumes an isometric scaling relationship) between the predictor and predicted variables, will be dominated by estimates from variable *X*10 and produce an overall level of dimorphism close to the dimorphism of variable *X*10 (i.e., close to 1.2). This is effectively a weighted average of dimorphism values among the 10 variables, where weights are the number of measurements per variable. The GM, which averages dimorphism across variables and is not weighted by the number of measurements per variable, will still produce a level of dimorphism close to 2.07. In general, in cases where some variables do not

scale isometrically with others in the data set, the template method results will be heavily influenced by the distribution of data through the various variables while the methods described here will produce consistent (although also inaccurate) results regardless of how the measurements are divided among the variables.

In an attempt to address the problem of deviations from isometry, measurements included in this study were restricted to long bone articular surfaces and cross-sectional properties of long bone shafts. These types of measurements have been shown to scale isometrically with body mass in primates in general and nonhuman hominoids in particular, but not in humans (Jungers, 1988, 1990; Godfrey et al., 1995). Additionally, Gordon (2004) found that GM of these types of variables tend to scale isometrically with body mass in nonhuman hominoids, but not in humans. Thus, just as the levels of extant ape postcranial dimorphism in Figure 1 closely resemble the level of body mass dimorphism seen in their various subspecies, the performance of log MMR based on this postcranial data set in the resampling procedure is expected to closely resemble log MMR based on body mass for the ape species. However, this expectation does not hold for the human sample. Because the variables in this data set do not scale isometrically with body mass in humans, postcranial SD levels do not match those of body mass in Figure 1 and are not expected to perform similarly to body mass in the resampling analysis.

The difference between patterns of SD in postcrania and body mass are frustrating, but do not necessarily pose an insurmountable problem. If *A. afarensis* has a scaling pattern for each variable that is identical to the modern human pattern, identical to the modern ape pattern, or intermediate between these extremes, then are there at least two possible approaches that can be used to test for significant difference in body mass dimorphism between fossil taxa and modern apes and humans. The first approach is to restrict the postcranial data set to only those measurements that scale isometrically with body mass in all comparative taxa, or even equivalently allometric in all taxa such that proportionality in SD is preserved between all taxa. However, this approach is unlikely to be very productive because most postcranial variables appear to scale differently with body mass between humans and apes (Ruff, 1988; Jungers, 1990; Hartwig-Scherer and Martin, 1992; Ruff and Runestad, 1992; Hartwig-Scherer, 1993). The second approach is to scale SD for each measurement by the allometric scaling relationship between that measurement and mass for the comparative sample, and to apply all of the various extant scaling patterns to the fossil sample to determine whether there is a qualitatively consistent result (e.g., is *A. afarensis* always more dimorphic than *H. sapiens*).

A variant of this approach has been in use for some time in McHenry's studies of size and dimorphism in *A. afarensis* and other early hominins (e.g., McHenry, 1986, 1991a, 1991b, 1992, 1994; McHenry and Berger, 1998), where allometric scaling relationships between skeletal dimensions and body mass in both apes and humans are used to generate two sets of estimates of body mass in fossil hominins, and the results are considered together to identify consistent patterns. Similarly, the methods described in this article can easily be modified so that rather than calculating a simple GM of ratios for each measurement within a multivariate data set, a weighted GM is calculated where the weights are equal to the scaling relationship between each measurement and body mass. By rerunning the analysis a number of times equal to the number of comparative species, each time applying a different comparative taxon's weightings to the fossil sample, it can be determined whether or not a consistent result regarding the ranking of and significant difference between SD of the various taxa exists. These and other methods should be further developed to obtain a better assessment of the primary variable of interest for social group and mating system reconstruction that skeletal dimorphism reflects; i.e., body mass dimorphism.

CONCLUSION

New methods presented here allow for tests of significant difference in dimorphism between extant and fossil multivariate data sets, even when data are missing from the fossil sample (i.e., individuals do not preserve all measurements). Furthermore, these new methods are not limited to fossil samples with specimens that preserve all of the measurements used in the study (i.e., a template specimen), and avoid or ameliorate some of the assumptions made by the template method. Results show that the *A. afarensis* sample studied here, composed of measurements representing the humerus, radius, femur, and tibia from Hadar fossils from 3.4 to 3.18 mya, is significantly more postcranially dimorphic than modern humans and chimpanzees and is more similar in overall postcranial dimorphism to extant gorillas and orangutans. These results support previous studies which have found similar levels of dimorphism in isolated regions of the postcranium, mandible, and dentition (e.g., Kimbel and White, 1988; Lague and Jungers, 1996; Lockwood et al., 1996; Lague, 2002; Harmon, 2006). Modern human-like dimorphism as suggested by Reno et al. (2003; 2005) is not supported.

A few cautions have been raised here regarding the application of any tests of dimorphism differences. First, just as sampling error can skew a small sample to produce Type I errors, Type II errors are also likely with small sample sizes because variances of test parameters tend to increase with decreasing sample size. When no significant differences between fossil and comparative samples are found, comparative samples should be compared in the same manner as fossil samples to determine whether tests erroneously find no difference in samples known to differ in SD such as *G. gorilla* and *H. sapiens*, and relative risk should be calculated. Second, the effect on a particular test of multiple representation of single individuals in a fossil sample must be considered and addressed when the possibility exists for such representation in a given data set (Reno et al., 2003; Plavcan et al., 2005; Scott and Stroik, 2006). Third, patterns of dimorphism (i.e., the similarity and ranking of SD levels for various measurements within a taxon) must be taken into account because including multiple measurements whose patterns vary between taxa will alter results depending on which measurements have greater representation in the fossil record, increasing the probability of Type I or II errors occurring. Finally, documented associations between dimorphism and behavior in living primates have been based on body mass dimorphism and craniodental dimorphism, not postcranial dimorphism. Taxon-specific scaling relationships between postcranial measurements and body mass must be considered before the results of tests of difference in postcranial SD can be

used to confidently infer particular types of social behavior and mating systems in fossil species. Given the results of this and past studies of postcranial SD, what can be stated with confidence is that such behaviors differed between *A. afarensis* and modern humans.

ACKNOWLEDGMENTS

We wish to thank Jennifer Clark and Richard Potts (Human Origins Program at the Smithsonian Institution), Linda Gordon and David Hunt (National Museum of Natural History), Eileen Westwig (American Museum of Natural History), and Lyman Jellema and Yohannes Haile-Selassie (Cleveland Museum of Natural History) for access to extant specimens and fossil casts under their care. We would also like to thank two anonymous reviewers for their helpful comments on this work, and Phillip Williams for helping to organize DJG's research trip to Cleveland.

APPENDIX

Proof: Ratios of GM of measurements are equivalent to GM of ratios of measurements.

For n measurements, their GM is calculated as

$$\mathrm{GM} = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}} \tag{A.1}$$

Overall size for a single specimen with no missing data can be calculated as the GM of those n measurements, so

$$\mathrm{Size}_{\mathrm{Male}} = \left(\prod_{i=1}^{n} M_i\right)^{\frac{1}{n}} \tag{A.2}$$

where M indicates a measurement from a male individual in this case. A measure of average male size can then be calculated as the GM of overall size for a sample of males, such that

$$\mathrm{Average\ Size}_{\mathrm{Male}} = \left[\prod_{j=1}^{m}\left(\prod_{i=1}^{n} M_{ij}\right)^{\frac{1}{n}}\right]^{\frac{1}{m}} \tag{A.3}$$

where m is the number of male individuals in the sample so that M_{ij} is the ith measurement from the jth male in the sample. Equation (A.3) can be further reworked as follows:

$$\mathrm{Average\ Size}_{\mathrm{Male}} = \prod_{j=1}^{m}\prod_{i=1}^{n} M_{ij}^{\frac{1}{nm}} \tag{A.4}$$

$$\mathrm{Average\ Size}_{\mathrm{Male}} = \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right)^{\frac{1}{n}} \tag{A.5}$$

Similarly, average female size can be expressed as

$$\mathrm{Average\ Size}_{\mathrm{Female}} = \prod_{i=1}^{n}\left(\prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}} \tag{A.6}$$

where p is the number of female individuals in the sample so that F_{ik} is the ith measurement from the kth female in the sample. Sexual dimorphism (SD) can then be expressed as a simple ratio of male and female size,

$$\mathrm{SD} = \mathrm{Average\ Size}_{\mathrm{Male}} / \mathrm{Average\ Size}_{\mathrm{Female}} \tag{A.7}$$

Substituting Eqs. (A.5) and (A.6) into (A.7) yields

$$\mathrm{SD} = \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right)^{\frac{1}{n}} \Bigg/ \prod_{i=1}^{n}\left(\prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}} \tag{A.8}$$

Taking the logarithm of both sides of the equation produces

$$\log \mathrm{SD} = \log\left[\prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right)^{\frac{1}{n}} \Bigg/ \prod_{i=1}^{n}\left(\prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}}\right] \tag{A.9}$$

Equation (A.9) can be reworked as follows:

$$\log \mathrm{SD} = \log \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right)^{\frac{1}{n}} - \log \prod_{i=1}^{n}\left(\prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}} \tag{A.10}$$

$$\log \mathrm{SD} = \frac{1}{n}\log \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right) - \frac{1}{n}\log \prod_{i=1}^{n}\left(\prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right) \tag{A.11}$$

$$\log \mathrm{SD} = \frac{1}{n}\sum_{i=1}^{n}\left(\log \prod_{j=1}^{m} M_{ij}^{\frac{1}{m}}\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\log \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right) \tag{A.12}$$

$$\log \mathrm{SD} = \frac{1}{n}\sum_{i=1}^{n}\left(\log \prod_{j=1}^{m} M_{ij}^{\frac{1}{m}} - \log \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right) \tag{A.13}$$

$$\log \mathrm{SD} = \frac{1}{n}\sum_{i=1}^{n}\log\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}} \Bigg/ \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right) \tag{A.14}$$

$$\log \mathrm{SD} = \frac{1}{n}\log \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}} \Bigg/ \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right) \tag{A.15}$$

$$\log \mathrm{SD} = \log \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}} \Bigg/ \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}} \tag{A.16}$$

Transforming both sides of the equation yields

$$\mathrm{SD} = \prod_{i=1}^{n}\left(\prod_{j=1}^{m} M_{ij}^{\frac{1}{m}} \Bigg/ \prod_{k=1}^{p} F_{ik}^{\frac{1}{p}}\right)^{\frac{1}{n}} \tag{A17}$$

Equation (A.17) states that SD is equal to the GM of male:female ratios for each of the n original variables. In other words, calculating SD as the ratio of the average of two groups of GM of individual measurements (males and females in this case) produces identical results to calculating SD as the GM of ratios for each of the measurements (see Table 3 in the text for an example).

LITERATURE CITED

Arsuaga JL, Carretero JM, Lorenzo C, Gracia A, Martinez I, Bermudez de Castro JM, Carbonell E. 1997. Size variation in middle Pleistocene humans. Science 277:1086–1088.

Chernick MR. 1999. Bootstrap methods: a practitioner's guide. New York: Wiley.

Cheverud JM, Dow MM, Leutenegger W. 1985. The quantitative assessment of phylogenetic constraints in comparative

analyses: sexual dimorphism in body weight among primates. Evolution 39:1335–1351.
Clutton-Brock TH, Harvey PH, Rudder B. 1977. Sexual dimorphism, socionomic sex ratio and body weight in primates. Nature 269:797–800.
Dobson SD. 2005. Are the differences between Stw 431 (*Australopithecus afarensis*) and A.L. 288-1 (*A. afarensis*) significant? J Hum Evol 49:143–154.
Efron B, Tibshirani RJ. 1993. An introduction to the bootstrap. London: Chapman & Hall.
Ford SM. 1994. Evolution of sexual dimorphism in body weight in platyrrhines. Am J Primatol 34:221–244.
Gaulin SJC, Sailer LD. 1984. Sexual dimorphism in weight among the primates: the relative impact of allometry and sexual selection. Int J Primatol 5:515–535.
Godfrey LR, Sutherland MR, Paine RR, Williams FL, Boy DS, Vuillaume-Randriamanantena M. 1995. Limb joint surface areas and their ratios in Malagasy lemurs and other mammals. Am J Phys Anthropol 97:11–36.
Gordon AD. 2004. Evolution of body size and sexual size dimorphism in the order primates: Rensch's rule, quantitative genetics, and phylogenetic effects. Ph.D. dissertation, University of Texas at Austin.
Gordon AD. 2006. Scaling of size and dimorphism in primates II: Macroevolution. Int J Primatol 27:63–105.
Green DJ, Gordon AD, Richmond BG. 2007. Limb-size proportions in *Australopithecus afarensis* and *Australopithecus africanus*. J Hum Evol 52:187–200.
Greenfield LO. 1992. Relative canine size, behavior, and diet in male ceboids. J Hum Evol 23:469–480.
Grine FE, Demes B, Jungers WL, Cole TM. 1993. Taxonomic affinity of the early *Homo cranium* from Swartkrans, South Africa. Am J Phys Anthropol 92:411–426.
Grine FE, Jungers WL, Schultz J. 1996. Phenetic affinities among early *Homo crania* from East and South Africa. J Hum Evol 30:189–225.
Harmon EH. 2006. Size and shape variation in *Australopithecus afarensis* proximal femora. J Hum Evol 51:217–227.
Hartwig-Scherer S. 1993. Body weight prediction in early fossil hominids: towards a taxon-"independent" approach. Am J Phys Anthropol 92:17–36.
Hartwig-Scherer S, Martin RD. 1992. Allometry and prediction in hominoids: a solution to the problem of intervening variables. Am J Phys Anthropol 88:37–57.
Harvati K, Frost SR, McNulty KP. 2004. Neanderthal taxonomy reconsidered: implications of 3D primate models of intra- and interspecific differences. Proc Natl Acad Sci USA 101:1147–1152.
Ihaka R, Gentleman R. 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314.
Josephson SC, Juell KE, Rogers AR. 1996. Estimating sexual dimorphism by method-of-moments. Am J Phys Anthropol 100:191–206.
Jungers WL. 1988. Relative joint size and hominoid locomotor adaptations with implications for the evolution of hominid bipedalism. J Hum Evol 17:247–265.
Jungers WL. 1990. Scaling of postcranial joint size in hominoid primates. In: Jouffroy FK, Stack MH, Niemitz C, editors. Gravity, posture and locomotion in primates. Firenze: Il Sedicesimo. p 87–95.
Jungers WL, Falsetti AB, Wall CE. 1995. Shape, relative size, and size-adjustments in morphometrics. Yearb Phys Anthropol 38:137–161.
Kappeler PM. 1990. The evolution of sexual size dimorphism in prosimian primates. Am J Primatol 21:201–214.
Kappeler PM. 1991. Patterns of sexual dimorphism in body weight among prosimian primates. Folia Primatol 57:132–146.
Kay RF, Plavcan JM, Glander KE, Wright PC. 1988. Sexual selection and canine dimorphism in New World monkeys. Am J Phys Anthropol 77:385–397.
Kimbel WH, Lockwood CA, Ward CV, Leakey MG, Rak Y, Johanson DC. 2006. Was *Australopithecus anamensis* ancestral to *A. afarensis*? A case of anagenesis in the hominin fossil record. J Hum Evol 51:134–152.
Kimbel WH, White TD. 1988. Variation, sexual dimorphism and the taxonomy of *Australopithecus*. In: Grine F, editor. Evolutionary history of the "robust" *Australopithecines*. New York: Aldine de Gruyter. p 175–191.
Kramer A. 1993. Human taxonomic diversity in the Pleistocene: does *Homo erectus* represent multiple hominid species? Am J Phys Anthropol 91:161–171.
Kramer A, Donnelly SM, Kidder JH, Ousley SD, Olah SM. 1995. Craniometric variation in large-bodied hominoids: testing the single-species hypothesis for *Homo habilis*. J Hum Evol 29:443–462.
Lague MR. 2002. Another look at shape variation in the distal femur of *Australopithecus afarensis*: implications for taxonomic and functional diversity at Hadar. J Hum Evol 42:609–626.
Lague MR, Jungers WL. 1996. Morphometric variation in Plio-Pleistocene hominid distal humeri. Am J Phys Anthropol 101:401–427.
Leigh SR. 1992. Patterns of variation in the ontogeny of primate body size dimorphism. J Hum Evol 23:27–50.
Leigh SR. 1995. Socioecology and the ontogeny of sexual size dimorphism in anthropoid primates. Am J Phys Anthropol 97:339–356.
Leigh SR, Shea BT. 1995. Ontogeny and the evolution of adult body size dimorphism in apes. Am J Primatol 36:37–60.
Leutenegger W, Kelly JT. 1977. Relationship of sexual dimorphism in canine size and body size to social, behavioral, and ecological correlates in anthropoid primates. Primates 18:117–136.
Leutenegger W, Shell B. 1987. Variability and sexual dimorphism in canine size of *Australopithecus* and extant hominoids. J Hum Evol 16:359–367.
Lockwood CA. 1999. Sexual dimorphism in the face of *Australopithecus afarensis*. Am J Phys Anthropol 108:97–127.
Lockwood CA, Kimbel WH, Johanson DC. 2000. Temproal trends and metric variation in the mandibles and dentition of *Australopithecus afarensis*. J Hum Evol 39:23–55.
Lockwood CA, Richmond BG, Jungers WL, Kimbel WH. 1996. Randomization procedures and sexual dimorphism in *Australopithecus afarensis*. J Hum Evol 31:537–548.
Manly BFJ. 1997. Randomization, bootstrap, and Monte Carlo methods in biology. London: Chapman & Hall.
Martin RD, Willner LA, and Dettling A. 1994. The evolution of sexual size dimorphism in primates. In: Short RV, Balaban E, editors. The differences between the sexes. Cambridge: Cambridge University Press. p 159–200.
McHenry HM. 1986. Size variation in the postcranium of *Australopithecus afarensis* and extant species of Hominoidea. Hum Evol 1:149–156.
McHenry HM. 1991a. Sexual dimorphism in *Australopithecus afarensis*. J Hum Evol 20:21–32.
McHenry HM. 1991b. Petite bodies of the "robust" australopithecines. Am J Phys Anthropol 86:445–454.
McHenry HM. 1992. Body size and proportions in early hominids. Am J Phys Anthropol 87:407–431.
McHenry HM. 1994. Behavioral ecological implications of early hominid body size. J Hum Evol 27:77–87.
McHenry HM. 1996. Sexual dimorphism in fossil hominids and its sociological implications. In: Shennan S, Steele J, editors. Power, sex and tradition: the archeology of human ancestry. London: Routledge and Keegan Paul. p 91–109.
McHenry HM, Berger LR. 1998. Body proportions in *Australopithecus afarensis* and *A. africanus* and the origin of the genus *Homo*. J Hum Evol 35:1–22.
McHenry HM, Corruccini RS. 1978. The femur in early human evolution. Am J Phys Anthropol 49:473–488.
Mitani JC, Gros-Louis J, Richards AF. 1996. Sexual dimorphism, the operational sex ratio, and the intensity of male competition in polygynous primates. Am Nat 147:966–980.
Mosimann JE. 1970. Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. J Am Stat Assoc 65:930–945.
Plavcan JM. 1994. Comparison of 4 simple methods for estimating sexual dimorphism in fossils. Am J Phys Anthropol 94: 465–476.

Plavcan JM. 2000. Inferring social behavior from sexual dimorphism in the fossil record. J Hum Evol 39:327–344.
Plavcan JM. 2002. Reconstructing social behavior from dimorphism in the fossil record. In: Plavcan JM, Kay RF, Jungers WL, van Schaik CP, editors. Reconstructing behavior in the primate fossil record. New York:Kluwer. p 297–338.
Plavcan JM. 2003. Scaling relationships between craniofacial sexual dimorphism and body mass dimorphism in primates: implications for the fossil record. Am J Phys Anthropol 120: 38–60.
Plavcan JM. 2004. Sexual selection, measures of sexual selection, and sexual dimorphism in primates. In: Kappeler PM, van Schaik CP, editors. Sexual selection in primates: new and comparative perspectives. Cambridge: Cambridge University Press. p 230–252.
Plavcan JM, Lockwood CA, Kimbel WH, Lague MR, Harmon EH. 2005. Sexual dimorphism in *Australopithecus afarensis* revisited: how strong is the case for a human-like pattern of dimorphism? J Hum Evol 48:313–320.
Plavcan JM, van Schaik CP. 1992. Intrasexual competition and canine dimorphism in anthropoid primates. Am J Phys Anthropol 87:461–477.
Plavcan JM, van Schaik CP. 1997a. Interpreting hominid behavior on the basis of sexual dimorphism. J Hum Evol 32:346–374.
Plavcan JM, van Schaik CP. 1997b. Intrasexual competition and body weight dimorphism in anthropoid primates. Am J Phys Anthropol 103:37–68.
Plavcan JM, van Schaik CP, Kappeler PM. 1995. Competition, coalitions and canine size in primates. J Hum Evol 28:245–276.
Rehg JA, Leigh SR. 1999. Estimating sexual dimorphism and size differences in the fossil record: a test of methods. Am J Phys Anthropol 110:95–104.
Reno PL, Meindl RS, McCollum MA, Lovejoy CO. 2003. Sexual dimorphism in Australopithecus afarensis was similar to that of modern humans. Proc Natl Acad Sci USA 100:9404–9409.
Reno PL, Meindl RS, McCollum MA, Lovejoy CO. 2005. The case is unchanged and remains robust: *Australopithecus afarensis* exhibits only moderate skeletal dimorphism. J Hum Evol 49:279–288.
Richmond BG, Aiello LC, Wood BA. 2002. Early hominin limb proportions. J Hum Evol 43:529–548.
Richmond BG, Jungers WL. 1995. Size variation and sexual dimorphism in *Australopithecus afarensis* and living hominoids. J Hum Evol 29:229–245.
Ruff CB. 1988. Hindlimb articular surface allometry in *Hominoidea* and *Macaca*, with comparisons to diaphyseal scaling. J Hum Evol 17:687–714.
Ruff CB, Runestad JA. 1992. Primate limb bone structural adaptations. Annu Rev Anthropol 21:407–433.
Scott JE, Stroik LK. 2006. Bootstrap tests of significance and the case for humanlike skeletal-size dimorphism in *Australopithecus afarensis*. J Hum Evol 51:422–428.
Silverman N, Richmond B, Wood B. 2001. Testing the taxonomic integrity of *Paranthropus boisei* sensu stricto. Am J Phys Anthropol 115:167–178.
Simons EL, Plavcan JM, Fleagle JG. 1999. Canine sexual dimorphism in Egyptian Eocene anthropoid primates: *Catopithecus* and *Proteopithecus*.. Proc Natl Acad Sci USA 96:2559–2562.
Skinner MM, Gordon AD, Collard NJ. 2006. Mandibular size and shape variation in the hominins at Dmanisi, Republic of Georgia. J Hum Evol 51:36–49.
Smith RJ. 1999. Statistics of sexual size dimorphism. J Hum Evol 36:423–459.
Smith RJ, Cheverud JM. 2002. Scaling of sexual dimorphism in body mass: a phylogenetic analysis of Rensch's rule in primates. Int J Primatol 23:1095–1135.
Smith RJ, Jungers WL. 1997. Body mass in comparative primatology. J Hum Evol 32:523–559.
Tague RG. 2005. Big-bodied males help us recognize that females have big pelves. Am J Phys Anthropol 127:392–405.
Taylor AB. 2006. Size and shape dimorphism in great ape mandibles and implications for fossil species recognition. Am J Phys Anthropol 129:82–98.
Villmoare B. 2005. Metric and non-metric randomization methods, geographic variation, and the single-species hypothesis for Asian and African *Homo erectus*. J Hum Evol 49:680–701.