Bootstrap tests of significance and the case for humanlike skeletal-size dimorphism in *Australopithecus afarensis*

Jeremiah E. Scott*, Laura K. Stroik

School of Human Evolution and Social Change, Institute of Human Origins, Arizona State University, Tempe, AZ 85287-4101, USA

Received 28 December 2005; accepted 1 June 2006

Abstract

Most estimates of sexual size dimorphism in *Australopithecus afarensis* indicate that this early hominin was more dimorphic than modern humans. In contrast, a recent study reported that size variation in *A. afarensis*, as represented by postcranial remains from Hadar and Maka, Ethiopia, is statistically most similar to that of modern humans, indicating a humanlike level of sexual dimorphism. Here, we evaluate the evidence for humanlike dimorphism in *A. afarensis*. We argue that statistical support for this claim is not as robust as has been asserted for the following reasons: (1) the analysis from which the claim was derived does not distinguish the *A. afarensis* sample from either the human or chimpanzee samples; (2) for some of the comparisons made, the *A. afarensis* sample cannot be distinguished from the *Gorilla* sample using two-tailed tests; and (3) the *A. afarensis* postcranial sample used in the analysis may contain more male than female specimens, which precludes a straightforward interpretation of the statistical results. Thus, support for humanlike dimorphism is equivocal, and a greater level of dimorphism cannot be ruled out.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: A.L. 333 sample; Biased sex ratio; Resampling; Sexual dimorphism; Skeletal variation; Type I error

Introduction

Resampling-based methods such as bootstrapping and exact randomization have become ubiquitous in studies of fossil hominin variation (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; Reno et al., 2003, 2005; Harvati et al., 2004; Dobson, 2005; Lee, 2005; Villmoare, 2005). These methods are commonly used to test hypotheses of equal relative variation in fossil samples and samples of extant species in order to identify taxonomically heterogeneous fossil assemblages.

A recent study by Reno et al. (2003, 2005) used bootstrapping to test the hypothesis that *Australopithecus afarensis* was more sexually dimorphic than modern humans. Previous qualitative and quantitative assessments of size variation in this early hominin have suggested that it was probably more dimorphic than humans and chimpanzees, perhaps approaching gorilla and orangutan levels (e.g., Johanson et al., 1978; Johanson and White, 1979; McHenry, 1991, 1992, 1994; Lockwood et al., 1996, 2000). However, Reno et al. (2003, 2005) argued that the results of their analysis indicate that skeletal-size variation, and thus dimorphism, in *A. afarensis* was similar to that of modern humans. Although the use of bootstrapping would appear to give their conclusion strong statistical support, we argue here that the nature of the sample Reno et al. (2003, 2005) used and the manner in which they applied the bootstrap test render a straightforward interpretation of their results problematic.

The bootstrap is commonly used to generate distributions of sample statistics (e.g., the coefficient of variation) from comparative samples in order to determine the probability of obtaining a sample with a degree of variation identical to
that observed in a fossil assemblage from an extant species (e.g., Cope and Lacy, 1992, 1995; Lockwood et al., 1996). Typically, 1000 bootstrap samples are generated from each comparative sample, with each bootstrap sample containing a number of specimens equal to that in the fossil assemblage. In effect, the comparative samples are used to simulate fossil samples. If 5% or fewer of the bootstrap samples from a particular comparative taxon exhibit a level of variation that is greater than or equal to that in the fossil assemblage, then the null hypothesis of equal relative variation is falsified. This result is generally interpreted in one of two ways: (1) the species represented by the fossil sample is more sexually dimorphic than the comparative taxon or (2) the fossil sample comprises multiple species.

This bootstrap methodology is essentially the one used by Reno et al. (2003). Their analysis focused on postcranial skeletal-size variation in the A. afarensis sample, as represented by a “combined Afar” sample \( (n = 29 \text{ specimens from Hadar and Maka, Ethiopia}) \) and a more restricted subset of the former, the A.L. 333 sample \( (n = 22 \text{ specimens}) \). The A.L. 333 sample was of particular interest due to the fact that it is considered “a large geologically simultaneous death assemblage,” which eliminates the potentially confounding effects of temporal and ecogeographic variation (Reno et al., 2003: 9404). Because Reno et al.’s (2003) sample was composed of different anatomical elements (e.g., proximal femora, distal humeri, distal fibulae), they used ratios of femoral-head diameter (FHD) to other skeletal dimensions in the A.L. 288-1 partial skeleton to estimate FHD values for the remaining specimens, referring to this procedure as the “template method” (where A.L. 288-1 is the “template specimen”). Thus, Reno et al. used 22 specimens to calculate FHD-size variation in the A.L. 333 sample, even while recognizing that fewer than 22 individuals were represented.

According to Plavcan et al. (2005), the strict minimum number of individuals (MNI) for the A.L. 333 postcranial specimens used by Reno et al. is four; however, they estimated that the actual number of individuals represented by these specimens is probably between five and eight. Thus, some of the individuals in the A.L. 333 assemblage might have contributed disproportionately to the sample, which could artificially lower estimates of dimorphism (Plavcan et al., 2005). Moreover, Plavcan et al. (2005) noted that there appears to be a size bias in the sample, with a disproportionately high representation of large individuals.

Reno et al. (2003, 2005) attempted to deal with this issue by conducting two bootstrap analyses in which the number of individuals in the samples bootstrapped from the comparative taxa was set at \( n = 5 \) and \( n = 9 \). (The latter sample size is the MNI of the dental remains (Reno et al., 2003).) For each bootstrap sample, the same number and types of elements in the A.L. 333 and combined Afar samples \( (n = 22 \text{ and 29, respectively}) \) were selected from five or nine individuals from samples of humans, chimpanzees, and gorillas. The FHD values for these elements were then estimated as for the fossil specimens (i.e., using a randomly selected template specimen). Size variation in the simulated fossil samples, as well as in the combined Afar and A.L. 333 samples, was quantified using three statistics—the binomial dimorphism index (BDI), the coefficient of variation (CV), and the maximum/minimum ratio (MMR). Reno et al. (2003, 2005) found that both the combined Afar and A.L. 333 values for these statistics generally fell within the middle 95% of the distributions generated from the Pan and Homo samples, but the A.L. 333 BDI and CV fell in the lower 5% of the Gorilla distribution. These results led Reno et al. (2003, 2005) to conclude that the level of variation in the A. afarensis sample is too low to have come from a species as dimorphic as Gorilla, and thus they argued that dimorphism in this hominin was humanlike. Below, we discuss three issues that call into question Reno et al.’s (2003, 2005) interpretation of their results.

**Humanlike or chimpanzeelike?**

The first issue concerns the characterization of A. afarensis as specifically humanlike in skeletal-size variation. As pointed out by Plavcan et al. (2005), this conclusion ignores the fact that the A.L. 333 BDI, CV, and MMR are fully compatible with chimpanzeelike skeletal dimorphism. Reno et al. (2005: 284) rejected this interpretation, stating that “a chimpanzee level of dimorphism requires the assumption that any A. afarensis fossil has an almost equal probability of being either sex, regardless of its size; a position we do not hold.” Nevertheless, insofar as the statistical comparisons are concerned, variation in the A.L. 333 sample cannot be distinguished from that of humans or chimpanzees based on Reno et al.’s (2003, 2005) results.

The BDI and CV for the combined Afar sample did have low probabilities of being sampled from the chimpanzee sample (with variation in the fossil sample being greater than in the Pan sample), although the \( p \)-values were of borderline significance \((p = 0.05 \text{ and 0.073 for the BDI and CV, respectively; see Table 7 of Reno et al., 2003 (in the online supporting information)})\). However, Reno et al. (2003: 9406) argued that the elevated level of variation in the combined Afar sample relative to that observed in the A.L. 333 sample “reflects not only sexual dimorphism but ecogeographic and temporal factors as well” (see also Reno et al., 2005). Thus, according to their reasoning, the A.L. 333 results are the most reliable in terms of estimating skeletal-size dimorphism in A. afarensis. If this assumption is correct, then by Reno et al.’s logic, their results provide no basis for rejecting chimpanzeelike dimorphism in A. afarensis. Moreover, Reno et al.’s (2003) combined Afar analyses treated all A.L. 333 specimens as separate individuals \( (n = 22) \). As discussed below, such treatment renders the results of any statistical comparison between the combined Afar sample—or the A.L. 333 sample by itself—and the extant hominines difficult to interpret (see also Plavcan et al., 2005).

**Two-tailed or not two-tailed?**

The second issue concerns the bootstrap simulation analysis in which the simulated fossil samples were composed of
five individuals. For this analysis, the BDI, CV, and MMR for the A.L. 333 sample could not be statistically distinguished from the *Gorilla* values at an *α*-level of 0.05 using a two-tailed test (see Table 3 in Reno et al., 2005; note that the combined Afar sample was not examined in this analysis). However, Reno et al. (2005) preferred the results of the one-tailed tests for the A.L. 333—*Gorilla* BDI and CV comparisons, which had *p*-values less than 0.05. They (Reno et al., 2005: 281–282) justified the use of a directional test by stating:

However, a nondirectional test is inappropriate here for a number of reasons. First, Plavcan et al. (2005) adopt the conventional view (and thereby the hypothesis to be tested) that *A. afarensis* exhibits dimorphism greater than that of humans. This requires a directional test. Second, bootstrapping procedures are designed to determine to what degree extant samples are likely to yield variation equivalent to that observed in fossils. Those who have used these methods (e.g., Arsuaga et al., 1997; Silverman et al., 2001), including authors of the Plavcan et al. paper (Lockwood et al., 1996, 2000) have typically reported only one-tailed tests.

We disagree with this logic—the fact that previous analyses have used one-tailed tests does not make it correct in the context of Reno et al.’s (2003, 2005) study. A one-tailed test may be appropriate when examining whether the level of variation in a fossil sample is unusually high in comparison to extant species (e.g., Lockwood et al., 1996, 2000; Silverman et al., 2001; see also Cope and Lacy, 1992), but we question whether it is appropriate for addressing the issue of which extant hominine *A. afarensis* most resembled in skeletal-size variation. In the case of the A.L. 333—*Gorilla* comparison, a two-tailed test is required because there is no a priori reason for a directional alternative to the null hypothesis of equal relative variation in these two samples.\(^1\) The alternative hypothesis should be nondirectional difference.

Thus, according to Reno et al.’s (2005) results, if the A.L. 333 sample does contain only five individuals, then it cannot be statistically distinguished from the *Gorilla* sample—or the *Pan* and *Homo* samples—on the basis of skeletal-size variation at the *α* = 0.05 level. Admittedly, the probability of sampling a level of variation similar to that observed in the A.L. 333 sample from the *Gorilla* sample is still low. In the next section, we discuss an issue raised by Plavcan et al. (2005) that could account for this result.

---

\(^1\) It could be argued that a one-tailed test is justified on the grounds that levels of variation in the most dimorphic extant hominoids (*Gorilla* and *Pongo*) represent the upper limits for this clade, which implies that variation in *A. afarensis* should not exceed that in the *Gorilla* sample. However, Kelley and colleagues (Kelley and Xu, 1991; Kelley, 1993; Kelley and Plavcan, 1998) have argued that the most dimorphic extant primates are not necessarily the most dimorphic primates that have ever existed. If this is true, then it is possible that some fossil hominins might have been more dimorphic—and thus had a higher level of intraspecific variation—than *Gorilla* and *Pongo*.

### Sex-ratio skew and type I error

As noted above, Plavcan et al. (2005) observed that it is possible that small individuals (likely females if *A. afarensis* is moderately or strongly dimorphic) are underrepresented in the A.L. 333 postcranial sample, which would bias estimates of dimorphism toward monomorphism in this sample, as well as in the combined Afar sample. In response, Reno et al. (2005: 283) stated: “We wonder how such a male-dominated, yet infant and subadult laden, contemporaneous fossil assemblage could have accumulated (i.e., what kind of systematic bias yields the combination of many subadults and large males—be it social or taphonomic?).” This statement seems to be conflating two issues: (1) individual preservation, i.e., a single organism, and (2) specimen preservation, i.e., parts of a single organism. Although unlikely, it is possible that the sex ratio of the A.L. 333 individuals is balanced or somehow reflective of the social structure of *A. afarensis*. However, this does not mean that the sex ratio of the A.L. 333 *specimens* is balanced. It is possible that a few males contributed several specimens to the sample, whereas only a few female specimens survived to be recovered.

If the A.L. 333 sample is characterized by the type of bias postulated by Plavcan et al. (2005), then the level of variation in the sample—when each specimen is treated as a separate individual—could indicate a degree of dimorphism substantially and perhaps statistically significantly less than the true dimorphism. Thus, Reno et al.’s rejection of the hypothesis that skeletal-size variation in *A. afarensis* was similar to that of *Gorilla* could be a type I error (i.e., false rejection of the null hypothesis) resulting from the fact that a sex-balanced *Gorilla* sample was used to evaluate a potentially sex-skewed *A. afarensis* sample. Reno et al. (2005) argued that their resampling methodology accounted for any bias in the sample, but it is not entirely clear that it did. Inspection of Reno et al.’s (2003) Fig. 2 indicates that the distributions of CVs, BDIs, and MMRs for the bootstrapped *Gorilla* samples are not shifted toward monomorphism. Comparison of the statistics of variation for Reno et al.’s (2003) *Gorilla* sample (see their Table 2) to the means of those for the bootstrap distributions (see their Table 3) verifies this fact—these values are similar. This characteristic of the *Gorilla* distribution should not be surprising—Reno et al.’s (2003) *Gorilla* sample is sex-balanced, and since the distribution of the bootstrap samples will reflect the properties of the sample from which they were drawn, the resulting bootstrap distribution will not account for any type of sex bias in the A.L. 333 or combined Afar samples. In cases where the effect of sex bias is not great, this factor may not be an issue for the bootstrap test. However, if sex bias has a marked effect, statistical comparisons may be confounded.

Clearly, some of the simulated samples generated from the comparative samples will also be heavily sex-biased. However, size variation in *A. afarensis* was not evaluated using only these samples. The statistical significance of the differences observed between *A. afarensis* and the extant hominine samples was assessed using the *distributions* of the bootstrap samples, which necessarily reflect the sex-balanced nature of
the comparative samples. Thus, if the sex ratio of the A.L. 333 sample is skewed, such that the statistics of variation are biased downward, and the true level of dimorphism in the A.L. 333 sample is, in fact, closer to that of Gorilla, then comparison to the Gorilla distribution would be misleading. It would be analogous to comparing the mean for a sex-balanced sample of Gorilla individuals to the mean for a sample that contained mostly male Gorilla individuals.

To further investigate this issue, we used dental data from Plavcan (1990) to determine how type I error rate varies in relation to sex ratio. The data set comprised mandibular-canine-length data for Cercopithecus cephus (n = 36, sex balanced); these data were chosen because the index of sexual dimorphism (ISD = 1.31) and CV (15.54) for this trait are similar to the BDI and CV for Reno et al.’s (2003) Gorilla bootstrap distribution (the mean BDI and CV for the n = 9 Gorilla simulations are 1.30 and 15.52, respectively). We constructed seven sets of “sex-biased fossil samples,” with sex ratios ranging from 11 males/11 females to 17 males/5 females, by bootstrapping the male and female C. cephus subsamples separately, holding the number of male and female specimens constant. A sample size of n = 22 was chosen for these samples in order to simulate the conditions in the A.L. 333 sample.² Each set comprised 1000 bootstrap samples, and CVs for each sample were calculated. These CVs were compared to a distribution (the “null distribution”) of 1000 CVs generated by bootstrapping the C. cephus sample (at size n = 22) without regard to sex to test the hypothesis of equal relative variation between the C. cephus sample and the simulated sex-biased fossil samples.

Type I error rate for each set of simulated fossil samples was estimated by counting the number of these samples that had a CV less than or equal to the 5th percentile of the null distribution of CVs. Note that this procedure is the same as the one used by Reno et al. (2003, 2005) and others (e.g., Lockwood et al., 1996; Silverman et al., 2001; Villmoare, 2005) to test whether variation in fossil samples differs significantly from that in extant samples. We focused specifically on false rejections of the null hypothesis in which the simulated fossil samples were found to have a significantly lower level of variation than the C. cephus sample (i.e., a one-tailed test) so that our results could be compared to Reno et al.’s (2003, 2005) A.L. 333—Gorilla comparison. We stress, however, that we are not using the results derived from the C. cephus data to directly evaluate the A.L. 333—Gorilla comparison; rather, we are examining the effect of skewed sex ratios on bootstrap tests of significance. Specifically, we are examining whether type I error is affected by the type of bias that may characterize the A.L. 333 sample.

The type I error rates for each sex-ratio set are shown in Figure 1, and Figure 2 illustrates three of the bootstrap distributions: the null distribution and the distributions for the 14/8 and 17/5 sex-ratio sets. As expected, the samples with the most balanced sex ratios (11/11 and 12/10) have type I error rates below 5%. For the samples with a sex ratio of 13/9 (41% female), the null hypothesis of equal relative variation is rejected 6.6% of the time. Type I error rate increases dramatically from this point, so that when only five of the specimens in the sample are female (23%), the type I error rate is 55.7%. Thus, the hypothesis of equal relative variation is rejected for over half of the 17/5 sex-ratio samples, even though they were drawn from the same sample as the null distribution.

It is important to point out that the effect of sex-ratio skew on type I error rate should vary in relation to the true level of dimorphism, such that as dimorphism decreases, type I error rate is less affected, and as dimorphism increases, the effect on type I error rate becomes more pronounced. In addition, since the level of variation in sex-balanced samples is heavily influenced by the degree of intrasexual variation in the more heavily represented sex, intrasexual variation should affect type I error as well. In the case of the C. cephus sample, the male CV is 6.63 (compared to an overall CV of 15.54). A larger intrasexual CV would have a less dramatic effect on type I error, whereas a lower intrasexual CV would have a more marked effect. Because the intrasexual CVs of A. afarensis are essentially unknowable, at least with present samples, we cannot assess their effect in this case.

What do these results mean for the A.L. 333—Gorilla comparison? Obviously, the sex ratio of the A.L. 333 postcranial sample will never be known, but it is possible that large, presumably male, individuals contributed disproportionately to the assemblage (Plavcan et al., 2005). Thus, a sex ratio for the specimens—not the individuals—as skewed as 17/5 is

² Although Reno et al. (2003, 2005) performed simulations in which the number of individuals was set at n = 5 and n = 9, they still used n = 22 specimens to estimate variation in each analysis.
null distribution. Towards a lower mean level of variation (i.e., “monomorphism”) relative to the sample from which they were drawn. For the sample (CV = 15.54). Over half (55.7%) of the bootstrap samples with a sex ratio of 17 males/5 females are considered significantly different from the sample from which they were drawn. For the samples with a sex ratio of 14 males/8 females, 11.1% are considered significantly different. Note that the distributions generated using sex-biased samples are shifted toward a lower mean level of variation (i.e., “monomorphism”) relative to the null distribution.

not out of the question, particularly if a single partial skeleton of a male individual constituted the bulk of the A.L. 333 sample. If this bias is accepted as a possibility, then it undermines the strength of Reno et al.’s (2003, 2005) interpretation of the statistical significance of the difference between the A.L. 333 and Gorilla samples. The bootstrap test does not fully account for sex bias. Thus, the low probability of sampling the level of variation observed in the A.L. 333 assemblage from the Gorilla sample might be due to an imbalance in the representation of male/female specimens (Plavcan et al., 2005), not because A. afarensis was less dimorphic than Gorilla.

Within the A.L. 333 sample, there are two levels of potential bias: (1) bias due to unequal representation of the sexes at the individual level and (2) bias due to disproportionate contribution of some individuals to the number of specimens, which could be sex-based (e.g., males contributed more specimens than females). Recognition that the number of individuals in the A.L. 333 sample is probably much less than the number of specimens (22 specimens vs. perhaps nine or fewer individuals; Reno et al., 2003; Plavcan et al., 2005) suggests that, if there is bias in the sample, it might be possible to control for at least some of it by grouping specimens of similar size (as measured by estimated FHD) into “individuals,” thereby eliminating redundant data points. While the information gained from such a procedure can never be used to formally test a hypothesis, it is instructive in terms of documenting the variational properties of the assemblage at different sample sizes.

We explored the effect of eliminating bias due to the potential overrepresentation of certain individuals by grouping the A.L. 333 specimens into “individuals” using cluster analysis (Ward’s hierarchical clustering method) on the FHD values provided by Reno et al. (2003; see their Table 1). This analysis was performed using SAS version 9.1 (SAS Institute, Inc., 2002–2003). We started with n = 22 individuals [i.e., the total number of A.L. 333 specimens used by Reno et al. (2003)] and grouped specimens into a decreasing number of individuals down to n = 8 [i.e., Plavcan et al.’s (2005) upper estimate of the number of A.L. 333 individuals]. In cases where groupings were equally probable for a given sample size, we arbitrarily chose one. Alternate groupings did not affect the results.

None of the individuals contained duplicated anatomical elements (e.g., two right proximal humeri). Each individual’s FHD value was calculated as the mean of the FHD values for the specimens included in that individual. It cannot be overemphasized that we do not claim that the cluster analysis has divided up the specimens in a manner that accurately reflects the true composition of the sample (i.e., if there are eight individuals, we acknowledge that our procedure may not have assigned the specimens correctly to each of the eight individuals). Rather, this exercise is heuristic.

For each “reduced” A.L. 333 sample, we calculated the BDI and CV; the MMR was not examined because it is a poor estimator of dimorphism [as demonstrated by Reno et al.’s (2003) results] and is generally not used for this purpose. Bootstrapping was used to construct 95% confidence limits for each statistic at each sample size (Efron and Tibshirani, 1993; Manly, 1997). This procedure involved resampling with replacement from the individuals in each reduced A.L. 333 sample 2000 times.3 The BDI and CV were calculated for each bootstrap sample, producing a distribution of sample estimates. Numerous approaches to generating bootstrap confidence limits are available for this analysis, we used the simple percentile method, which defines the 95% confidence interval as the middle 95% of the values in the bootstrap distribution (Efron and Tibshirani, 1993; Manly, 1997). In cases where

3 Manly (1997) recommends at least 2000 iterations for constructing 95% confidence intervals.
the bootstrap estimator was found to be excessively biased,\(^4\) we used bias-corrected percentile limits (e.g., Manly, 1997: 46–48). The confidence limits were used to evaluate Reno et al.’s (2003) claim that the degree of skeletal-size variation in the A.L. 333 sample is significantly different from that of *Gorilla*. If this hypothesis is correct, then the confidence intervals for each measure of variation in the A.L. 333 sample should not include the means for the BDI and CV from the *Gorilla* bootstrap distribution (we used the values from the \(n = 9\) simulations; see Table 3 in Reno et al., 2003). Note that this test is two-tailed.

The BDIs, CVs, and confidence intervals for each reduced A.L. 333 sample are shown in Figure 3. The BDIs and CVs for the *Gorilla, Homo*, and *Pan* samples are also displayed. When the number of individuals in the A.L. 333 sample is 12 or more, the confidence intervals for both statistics encompass only the *Homo* and *Pan* values—they do not overlap the *Gorilla* values, supporting Reno et al.’s (2003, 2005) conclusion that A.L. 333 differs significantly from *Gorilla*. However, at a sample size of \(n \leq 11\) individuals, the A.L. 333 confidence intervals overlap all of the comparative taxa. This result is due in part to wider confidence intervals, but the level of sample variation also increases as the sample size decreases, shifting the confidence intervals upward so that they include the *Gorilla* values. This pattern of change should not be surprising—collapsing the specimens into fewer and fewer data points (i.e., from \(n = 22\) to \(n = 8\)) reduces potentially redundant values, mostly from larger individuals, which gives greater weight to the smaller specimens, thus increasing the level of sample variation.

The upshot of this exercise is that rejection of the null hypothesis of equal relative variation in the A.L. 333 and *Gorilla* samples seems to depend on the assumptions made about the A.L. 333 sample. According to our exercise, if it is assumed that there are 12 or more individuals in the assemblage, then Reno et al.’s characterization of skeletal-size variation in *A. afarensis* as humanlike receives very limited support—variation in the A.L. 333 sample still cannot be distinguished from that of *Pan*. In contrast, if it is assumed that there are fewer than 12 individuals, then the hypothesis of equal relative variation cannot be refuted for any of the extant hominine–A.L. 333 comparisons. Unfortunately, the number of individuals in the A.L. 333 sample remains a matter of debate (e.g., Plavcan et al., 2005; Reno et al., 2005), and thus this issue cannot be resolved.

---

\(^4\) Bias in this case refers to bias in the bootstrap estimator and should not be confused with the discussion of bias in terms of the sex ratio of the A.L. 333 sample. Bias in a bootstrap estimator occurs when the mean of the bootstrap distribution differs substantially from the observed statistic (Efron and Tibshirani, 1993; Manly, 1997). In such cases, the bootstrap confidence limits will be shifted in the direction of this difference (e.g., when the bootstrap mean is greater than the observed statistic, the confidence limits will be shifted “up” relative to the observed statistic). According to Efron and Tibshirani (1993), when bias is less than 0.25 standard errors of the bootstrap mean, it can be ignored; otherwise, the confidence limits should be corrected. This guideline was followed in the analyses presented here.

---

**Fig. 3.** Binomial dimorphism indices (upper graph) and coefficients of variation (lower graph) for the reduced A.L. 333 samples. The vertical lines represent the 95% bootstrap confidence intervals for each statistic. The horizontal lines represent the BDIs and CVs for Reno et al.’s (2003) *Gorilla* (solid), *Homo* (dashed), and *Pan* (dotted) samples.

**Conclusions**

In sum, the degree of dimorphism in *A. afarensis* remains an open question. The A.L. 333 postcranial sample cannot be statistically distinguished from the *Pan* or *Homo* samples at any level of analysis. Claims for a statistically significant difference between the A.L. 333 and *Gorilla* samples are problematic due to the fact that the composition of the A.L. 333 sample is unknown. Consequently, the A.L. 333 sample by itself cannot be used to characterize skeletal-size dimorphism in *A. afarensis* as humanlike, chimpanzeelike, or gorillalike. Moreover, because Reno et al.’s (2003, 2005) combined Afar sample consists largely of A.L. 333 specimens (22 out of 29
elements, or 76%), the results derived from analyses of this larger sample are also suspect. Thus, although it is possible that this early hominin had more in common with modern humans in terms of social behavior—and the physiological and anatomical correlates thereof—that is generally accepted, this hypothesis receives little or no statistical support.

Acknowledgements

We thank Bill Kimbel, Charlie Lockwood, Michael Plavcan, and Caley Orr for critical readings of an early version of this manuscript; their comments contributed greatly to the final product. Many thanks also go to JHE editor Susan Antón, the associate editor, and the anonymous referees for their insightful criticisms and extremely useful suggestions. Some of the analyses performed herein were conducted using macros written by Charlie Lockwood.

References


