Testing the Taxonomic Integrity of *Paranthropus boisei* sensu stricto

Nicole Silverman,1 Brian Richmond,2 and Bernard Wood2*

1Hominid Paleobiology Doctoral Program, Department of Anthropology, George Washington University, Washington, DC 20052
2Department of Anthropology, George Washington University, Washington, DC 20052

KEY WORDS *Paranthropus boisei*; mandible; bootstrapping; intraspecific variation

ABSTRACT The craniodental hypodigm of *Paranthropus boisei* sensu stricto is morphologically distinctive, but it has been suggested that the substantial variation in mandibular and dental size in that hypodigm may exceed that which is reasonable to subsume within a single hominin species. In this study, Fligner and Killeen, coefficient of variation (CV)-based and average taxonomic distance (ATD)-based bootstrap tests, were used to compare variation in size and shape of the mandibular corpus remains attributed to *P. boisei s.s.* with the variation observed in samples of great apes and modern humans. The degree of size variation in the *P. boisei s.s.* mandibular hypodigm is never observed in human and chimpanzee samples, is rare in gorillas, but is not uncommon in orangutans. However, the shape variation in the fossil group is comparable to the variation in the extant reference groups. Although the size variation in *P. boisei s.s.* is substantial, it is exaggerated by the effects of taphonomy. The small mandibles are more often abraded, whereas the large mandibles are more likely to have been infiltrated with matrix. On the basis of the results of this investigation of the mandibular corpus, there are no grounds for rejecting the “single-species” hypothesis for *P. boisei s.s.* When Sokal and Braumann’s adjusted CV values were used to predict the index of sexual dimorphism (ISD) for the *P. boisei s.s.*, despite the substantial geological time embraced by the mandibular corpus hypodigm, the predicted value of lnISD, when corrected for taphonomic factors, is comparable to the sexual dimorphism observed within *Gorilla*. Am J Phys Anthropol 115:167–178, 2001. © 2001 Wiley-Liss, Inc.

A common problem in paleontology is determining whether a collection of apparently morphologically homogeneous fossils samples one, or more than one, species. Within the hominin fossil record, *Paranthropus boisei* is arguably the most morphologically distinctive taxon. The hypodigm displays a substantial number of unique character states as well as unique combinations of symplesiomorphies (e.g., Strait et al., 1997; Wood and Collard, 1999); these factors should make members of this early hominin taxon relatively easy to identify. It is also a hominin taxon for which there is an especially good dental and mandibular fossil record (Wood, 1991; Wood et al., 1994), thus making it amenable to tests of its integrity.

The first evidence of *Paranthropus boisei sensu stricto* was recovered from Olduvai Gorge in 1955. The holotype, OH 3, a deciduous canine and molar, was discovered and initially described in 1959, was deemed sufficiently distinct to warrant allocation to not only a novel species, but also to a new genus, *Zinjanthropus boisei* (Leakey, 1959). Within a year, Robinson (1960) had suggested that *Zinjanthropus boisei* should be subsumed within *Australopithecus*, and now researchers generally refer to this taxon as either *Australopithecus boisei* (Leakey, 1959) Tobias, 1967, or *Paranthropus boisei* (Leakey, 1959) Robinson, 1960.

Subsequently, the hypodigm of *P. boisei* has been substantially enlarged, mainly by discoveries from Koobi Fora (Leakey and Walker, 1988; Wood, 1991), the Shungura Formation (Coppens, 1980; Howell and Coppens, 1976), and West Turkana (Leakey and Walker, 1988; Walker and Leakey, 1988). More inclusive interpretations of *P. boisei* subsume the 2.5-Myr-old KNM-WT 17000 within the hypodigm (e.g., Walker et al., 1986), but for the purposes of this study KNM-WT 17000 and KNM-WT 16005, to-
gether with the L338y-6 cranium, and the mandible Omo 18-1967-18 from the Shungura Formation, are assigned to a separate species, Paranthropus aethopicus (Arambourg and Coppens, 1968). In this paper the less inclusive interpretation of the taxon will be referred to as Paranthropus boisei sensu stricto, or P. boisei s.s. Fossil specimens allocated to P. boisei s.s. span the time period from approximately 2.2 to ca.1.3 Myr. They have all been recovered from sites associated with the Gregory Rift, extending from Konso, in Ethiopia (Suwa et al., 1997) in the north, to Malema, in Malawi (Kullmer et al., 1999) in the south. The oldest well-dated evidence comes from the Shungura Formation, in Ethiopia (Feibel et al., 1989), and temporally the most recent evidence is likely to be either OH 36 from Bed II, Olduvai Gorge (Aiello et al., 1999; but see Tamrat et al., 1995), or KNM-ER 1820, together with four large adult man-
eries, one on the face (Bilsborough and Wood, 1988) and Preuschoft, 1980; Dean and Wood, 1982; Chamberlain and Wood, 1985; Grine, 1988; Rak, 1988; Wood, 1991; Wood et al., 1999). However, two studies, one on the face (Bilsborough and Wood, 1988) and the other on the cranial base (Wood, 1991), have drawn attention to aspects of the cranial variation within P. boisei s.s. that do not match the degree and pattern of intraspecific variation in closely-related comparative taxa (Wood et al., 1991). However, it was the extent of variation within the mandibular component of the hypodigm P. boisei s.s. that attracted the interest of both Dean (1988) and Groves (1989). Dean (1988) examined the cross-sectional areas of juvenile mandibular corpora and noted that KNM-ER 1820 was the largest juvenile mandible attributed to any hominin taxon. He suggested that KNM-ER 1820, together with four large adult mandibles, Omo L7a-125, and KNM-ER 818, 1468 and 1469, that Chamberlain and Wood (1985) had suggested belonged to large-bodied males of P. boisei s.s., could be interpreted as evidence of a second Paranthropus species in East Africa. A year later, Groves (1989) also questioned the single-species interpretation of the P. boisei s.s. hypodigm. Specifically, he cited variability in the buccolingual breadth of P4 and calculated that there was only a 1% probability that the mandibular specimens L7a-125 and L74a-21 from the Omo Shungura Formation belonged to the same species (Groves, 1989, p. 244).

Morphometric data are now available for a sufficient number of mandibular specimens attributed to P. boisei s.s. to justify a formal test of the hypothesis that the adult mandibular fossil remains currently attributed to Paranthropus boisei sensu stricto represent a single taxon. In the event that the P. boisei sample is interpreted as representing one species, we also investigated the relationship between sexual dimorphism and within-sample variation. Specifically, we applied bootstrapping techniques to the estimation of the index of sexual dimorphism (ISD) from the reference samples’ adjusted coefficient of variation (CV*) (Sokal and Braumann, 1980). Least squares regression analysis was used to determine the ability of the CV* to predict the ISD for the reference samples. The predicted ISD value from the CV* for P. boisei s.s. would be expected to fall within the range of ISD values for the extant samples.

### MATERIALS AND METHODS

The comparative sample consists of four extant species, Homo sapiens, Pan troglodytes, Gorilla gorilla, and Pongo pygmaeus, comprising a pooled sample of 197 adult individuals (Table 1). The Gorilla sample includes two subspecies, Gorilla gorilla beringei and Gorilla gorilla gorilla, and the Pongo sample comprises Pongo pygmaeus abelii and Pongo pygmaeus pygmaeus. Data for Gorilla, Pongo, Homo, and approximately half of the Pan specimens were collected from specimens housed at the National Museum of Natural History, Smithsonian Institution; the balance of the Pan data was collected from the Cleveland Museum of Natural History. All the comparative specimens are adult as indicated by erupted M3s. Each comparative genus sample contains approximately equal numbers of males and females. The 27 fossil mandibles included in the analysis (Table 2) differ in their preservation. Some are complete corpora, but others preserve only parts of the mandibular corpus. Nonetheless, all specimens are complete enough to allow data to be collected for at least two of the variables detailed below.

Measurements were selected to capture size and shape information from parts of the mandible that are commonly preserved in the early hominin fossil record. Seventeen linear measurements (Table 3) were taken with digital calipers on the mandibular corpus of the extant specimens, using the measurement definitions described in Wood (1991). Data for all but one of the fossil specimens were taken from Wood (1991). The exception was the mandible belonging to the skull KGA10-525 from Konso; the

<table>
<thead>
<tr>
<th>Taxon</th>
<th>N</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paranthropus boisei sensu stricto</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens sapiens</td>
<td>52</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>55</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>Pongo</td>
<td>12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Pongo pygmaeus abelii</td>
<td>44</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>Gorilla</td>
<td>12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Gorilla gorilla beringei</td>
<td>22</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

TABLE 1. List and composition of comparative taxa and fossil taxon
measurements used for this specimen were obtained from Suwa et al. (1997). The fossil specimens exhibit varying degrees of post depositional degradation, ranging from being permeated by matrix-filled cracks to the loss of significant amounts of cortical bone, but the measurements used in the main part of this study made no allowance for postmortem damage. B.W. measured the fossil mandibles (except for Konso); N.S. measured the extant samples. Prior to measuring the extant samples, all measurements were taken repeatedly on four of the mandibles in the comparative sample by both B.W. and N.S. The method and technique were rehearsed until inter-(B.W. and N.S.) and intra- (N.S.) observer errors were <3%. Interobserver errors have negligible effects here, since variability measures are calculated intraspecifically.

### Analysis

Variation in both size and shape was assessed. The size of each individual mandible was calculated as the geometric mean (GM) (Mosimann, 1970) of the height and breadth of the mandibular corpus at M1 (i.e., the square root of computed corpus cross-sectional area). These variables were chosen in order to maximize the size of the fossil sample. Two methods were applied to these data, a bootstrapping procedure utilizing the coefficient of variation (CV)
as a measure of variability within each taxon, and the test of Fligner and Killeen (1976) (hereafter referred to as the FK test). The FK test was recently described by Donnelly and Kramer (1999) as one of the only methods for assessing equivalent variation between two groups with acceptable levels of type I and type II errors when the samples and source populations have different distributions. Donnelly and Kramer (1999) recommend the FK test as the most appropriate variation-based statistical test of a single species hypothesis in a fossil sample.

The GM data were used to compute Sokal and Braumann’s adjusted CV (hereafter referred to as CV*) to correct for bias due to small sample size (Sokal and Braumann, 1980; Sokal and Rohlf, 1995). Size variation, expressed as CV*, in the fossil sample was compared to variation within each of the extant populations using bootstrapping methods (Efron and Tibshirani, 1993; Lockwood et al., 1996). It would be inappropriate to compare the variation of a fossil sample of 27 individuals to the variation in a much larger extant sample, and there would be no assessment of whether the fossil sample variation lies within the confidence limits of the variation in the extant species. Bootstrapping is a technique for estimating standard errors and confidence limits for a variety of statistics (Efron and Tibshirani, 1993; Sokal and Rohlf, 1995). A form of bootstrapping suitable to comparing statistics of variation (here, CV*) is resampling from reference samples at sample sizes equal to the fossil sample (Lockwood et al., 1996). In this way, the random samples drawn from reference taxa approximate potential fossil samples. This enables the researcher to calculate the probability that the size and shape (see below) differences observed in the fossil sample are observed in the extant samples. Thus, 27 individuals were randomly sampled, with replacement (each individual has an equal opportunity of being selected more than once for a simulated sample), from each of the extant populations. After the initial subset of 27 individuals was randomly selected from the comparative sample, its CV was calculated, and the correction factor applied to compute CV* (Sokal and Rohlf, 1995). This process was repeated 999 times, for a total of 1,000 iterations. The 1,000 CV*s were then compared to the fossil CV* to determine the likelihood that the size variation present in the simulated samples exceeds that within the fossil sample.

The bootstrapping program also computes the ISD (Fleagle et al., 1980) for each simulated sample. The lnISD is calculated by taking the natural logarithm of the geometric mean of the CVs.

### TABLE 3. Measurements

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Definitions</th>
<th>CV test</th>
<th>ATD 16 variables</th>
<th>ATD 12 variables</th>
<th>ATD 4 variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus height at P₄</td>
<td>147-minimum distance between most inferior point on base and lingual alveolar margin at midpoint of P₄ (Wood, 1991, p 295).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Corpus width at P₄</td>
<td>148-maximum width at right angles to 147, taken at midpoint of P₄ (Wood, 1991, p 296).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Corpus height at M₁</td>
<td>150-same as 147 (Wood, 1991).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Corpus height at M₂</td>
<td>154-same as 147 (Wood, 1991).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Height of mental foramen (AL)</td>
<td>162-vertical distance between center of major mental foramen and alveolar border at that position (Wood, 1991, p 296).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>I₁–I₂ alveolar length</td>
<td>163-chord distance between infradentale and midpoint of alveolar septum between I₂ and C.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Mandibular canine socket (labiolingual length)</td>
<td>164-maximum internal breadth of canine alveolus in labiolingual axis (Wood, 1991, p 296).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Mandibular canine socket (mesiodistal length)</td>
<td>165-maximum internal breadth of canine alveolus in mesiodistal axis (Wood, 1991, p 296).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>P₃–P₄ alveolar length</td>
<td>167-minimum chord distance between midpoints of interalveolar septa between C/P₃ and P₄/M₁ (Wood, 1991, p 297).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>M₁–M₃ alveolar length</td>
<td>168-minimum chord distance between midpoint of interalveolar septum between P₄/M₁ and most posterior point of M₃ alveolus (Wood, 1991, p 297).</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

1 Numbers given at beginning of each definition refer to the measurement number in Wood (1991).
of the ratio of the average mandibular corpus size (GM) for the males within each simulated set, to the average GM for the females in the same simulated set (i.e., ln(average M/average F)). Plavcan (1994) stated that there is a nonlinear relationship between ISD and CV*, and therefore the ISD ratios should be log transformed to produce a linear relationship which facilitates interpretation of the regression. Smith (1999) also supports the use of log-transforming the ISD ratio for quantifying sexual size dimorphism in a population. Both CV* and lnISD were calculated for each randomly sampled set of 27 individuals, thereby enabling an assessment of the error associated with predicting the lnISD from the CV*s of each extant sample. If CV* is effective for predicting lnISD, and assuming that the specimens represent a single species, CV* may serve as a proxy for the degree of sexual dimorphism in the fossil taxon. Least squares regression analysis was used to determine the relationship between CV* and lnISD values for the 1,000 simulated samples for each extant taxon, because CV* is used to predict lnISD in the fossil sample.

The FK test is a nonparametric test that Donnelly and Kramer (1999, p. 512) describe as a “weighted scores test for equal dispersion.” The FK test offers a balance of type I and type II error rates irrespective of population distributions, and has been argued to provide a more accurate test of relative variation between the population in question and the reference population (Donnelly and Kramer, 1999). For each taxon, the GMs are ln-transformed, and the median value determined. Next, each ln-transformed value is subtracted from the median, and the absolute value of this difference is recorded [x’i = ln(xi) − median ln(x)]. The resulting values (x’i) are a measure of the absolute deviation from the median for each individual (Donnelly and Kramer, 1999). After these values are obtained, the individual within the sample that represents the median value is removed. At this point, the samples under investigation are combined (i.e., P. boisei s.s. + Gorilla gorilla, P. boisei s.s. + Pongo pygmaeus, etc.) and their transformed values (x’i) rank ordered. When ties in rank occur, the ranks are changed to follow sequential order rather than sharing the same rank. For example, if two individuals tied for rank 15, one individual would be ranked at 15, and the other at 16. The final step in the FK test is to convert the ranks to weighted scores (equation on p. 512 in Donnelly and Kramer, 1999). After the weighted scores are computed, ties are resolved. The average for the weighted scores is computed for the individuals sharing a rank, and this average value is used for those specimens throughout the rest of the statistical computation. The test statistic “T” is applied and then assessed for significance by computing the large sample approximation “z;” the P-value is determined from the z-value by referencing the table of the normal distribution, or by direct computation (Donnelly and Kramer, 1999). In essence, the FK test simply assesses whether the ranks (i.e., the dispersion about the median) are equivalent in the two samples, or whether one sample has a significantly greater representation of large ranks (large deviations from the median) relative to the other sample.

Shape variation was also assessed using a bootstrapping approach. Shape variables were computed by dividing each variable by the GM, where the GM for each specimen is the nth root of the product of n variables (Mosimann, 1970; Jungers et al., 1995). Therefore, each mandible was represented by n shape variables. Because the fossil mandibles vary in their states of preservation, three sampling strategies were used for the shape analysis. In the first, shape information about the fossil sample (i.e., the number of variables) was maximized while retaining at least three mandibles in the sample; this strategy resulted in the maximum of 16 variables to describe shape. In the second, sample size and information content from each mandibular corpus were balanced; this resulted in a sample comprising 12 variables and six fossil specimens. Lastly, fossil representation was maximized (10 individuals), but at the expense of the number of variables (four).

To simulate the fossil samples, three, six, and ten individuals were randomly drawn (without replacement) for each extant sample, to simulate the fossil sample. Characterizing the variation in shape in several variables and multiple individuals is not as straightforward as it is with a single size variable. In order to provide an approximate assessment of shape variation, the differences in shape were determined for each pair of individuals in the set by calculating the average Euclidian distance, also known as the average taxonomic distance (ATD). ATD is computed as the square root of the mean sum of the squares of differences between the values for n shape variables between two individuals (Sneath and Sokal, 1973; Aiello et al., 1999). After each set is randomly drawn, ATDs are computed between all possible pairs of individuals, and the greatest pairwise ATD is noted. Thus, 1,000 iterations yield 1,000 maximum ATDs for that taxon. The maximum fossil pairwise ATD is then compared to this distribution, and the probability of sampling the maximum fossil ATD is assessed. A multivariate version of the CV might provide a more robust assessment of shape variation in the fossil sample and randomly drawn extant sets. However, comparisons of the maximum fossil shape differences to the distributions of maximum ATDs in extant species are more likely to detect individuals that do not belong within a species sample. Thus, we argue that maximum pairwise shape contrasts might provide a more powerful (but potentially more susceptible to type I errors) test of a single-species hypothesis.

**RESULTS**

The degree of size variation in the P. boisei s.s. sample of 27 mandibles is reflected in the CV* of
The fossil sample is significantly more variable than ally comparable to the bootstrapped CV* (Table 4). Since the method is likely to err on the side of power and type I error, it is unlikely that the shape variation in the fossil sample has been underestimated. The greatest difference among the three possible pairings of the *P. boisei s.s.* mandibles that preserve all 16 variables was that between KNM-ER 729 and Peninj 1 (Fig. 2). These two specimens are presumed males, making a multiple-species interpretation of the fossil sample even less likely. If the fossil assemblage consisted of a mixture of two species, then the principle of competitive exclusion would suggest that morphologically similar, and presumably closely related, sympatric species are likely to be accessing different resources, and will thus be more likely to differ rather than be similar in size. Thus, one would expect the shape differences to be greater between the presumed taxa (i.e., between larger and smaller individuals) than within the taxa (i.e., within the same size range).

Unlike the situation for size, mandibular corpus shape variation in *P. boisei s.s.* is easily accommodated within every comparative sample. For the 12 and 4 variable analyses, including, respectively, 6 and 10 fossil specimens, the most disparate fossil pairings are also between two presumed males, KNM-ER 3230 and Omo L7a-125, even though presumed females are included in both samples (Figs. 3, 4). In both 12 and 4 variable analyses, variation in shape among the *P. boisei s.s.* fossil mandibles is comparable to that in the comparative groups (Table 4).

In an effort to estimate the degree of dimorphism from the variation in the fossil sample, the relationship between CV* and lnISD was examined. A high correlation ($r^2 = 0.91, P < 0.05$) is obtained between CV* and lnISD when a single CV* and lnISD are calculated for each taxon sample (Fig. 5a). However, the points representing the four extant taxa form what is essentially a two-point regression. More fundamentally, there is no information on the confidence limits surrounding the CV* and lnISD estimates for each species. In order to evaluate the potential error involved in estimating the lnISD in *P. boisei* from the sample CV*, we randomly drew 27 individuals from each balanced-sex extant sample and calculated the CV* and lnISD for each set. When this process was repeated 1,000 times for each species, it becomes clear that there is considerable variation in CV* and lnISD, depending on the individuals selected for the sample (Fig. 5b). This is almost certainly due to the fact that lnISD is biased depending on the sex composition of the simulated sample (Plavcan, 1994), and our random samples of 27 individuals occasionally sampled more males

![Image](63x399 to 300x723)

**Fig. 1.** Frequency distributions of corrected coefficient of variation values of mandibular corpus size simulated from extant reference samples. Note that CV*s as high as the fossil CV* (dashed line), and higher, are never observed in the simulations of *Pan* and *Homo*, and are rarely observed in the *Gorilla* simulations, but are not uncommon in *Pongo*.
than females, and vice versa. The correlation among the bootstrapped samples is lower \( (r^2 = 0.77, P < 0.01, \text{Fig. } 5b) \) than the correlation of the single CV*/lnISD values for each taxon population (compare Fig. 5a,b).

If the fossil hominin mandibular sample represents one species, the sample CV* of 12.3 predicts a lnISD of 0.184 (95% confidence limits, 0.183–0.185), but this prediction must be accompanied by an appreciation of its potential error, depending on the sex composition of the fossil sample (Fig. 5b). Therefore, even with a single fossil taxon sample of 27 individuals, a precise estimate of a species’ sexual dimorphism remains unavailable in the absence of information on the sex of the individuals. However, Figure 5b illustrates that the only extant hominoids with sexual dimorphism comparable to that observed in the \( P. \) boisei sample are \( G. \) gorilla and \( P. \) pygmaeus, a finding that is consistent with the bootstrapping results.

**DISCUSSION**

Debates about the taxonomy of \( P. \) boisei have focused on the significance of either allochronic, or synchronic, variation within the hypodigm. Discussions about the former concern the lower temporal boundary, or the FAD, of the taxon (e.g., Kimbel, 1995). More inclusive taxonomic inter-
pretations incorporate all East African “robust” australopith fossil evidence within *P. boisei* (e.g., Walker and Leakey, 1988; Brown et al., 1993). However, other analyses lay more stress on morphological discontinuities within the hypodigm immediately prior to, or within, Member G of the Shungura Formation (Howell et al., 1987; Suwa, 1988; Suwa et al., 1996; Wood et al., 1994). Thus, these workers refer the pre-ca. 2.3 Myr evidence to either *Australopithecus aethiopicus* (Suwa et al., 1996), or *Paranthropus aethiopicus* (Wood et al., 1994), and treat the major, post-ca. 2.3 Myr, more derived component of the *P. boisei sensu lato* hypodigm as *P. boisei sensu stricto*. There is little or no disagreement about what fossil evidence is relevant to this debate. The two interpretations simply reflect different philosophies about the taxonomic significance of the morphological changes that occur within the hypothetical *Paranthropus* lineage around 2.3 Myr. Those who espouse a more taxic interpretation of the fossil record see these changes as evidence for a speciation event (e.g., Wood et al., 1994). Supporters of a more gradualistic interpretation of the fossil record (e.g., Walker and Leakey, 1988) interpret the same data as evidence for the type of change one would expect to find within an evolving species (sensu Bown and Rose, 1987).

The debate about the taxonomic significance of allochronic variation within *P. boisei sensu lato* has tended to overshadow the concerns noted by some researchers (Bilsborough and Wood, 1988; Dean, 1988; Groves, 1989; Wood, 1991) about the taxonomic integrity of the post-ca. 2.3 Myr fossil evidence attributed to *P. boisei sensu stricto*. All these authors accept that members of the *P. boisei* s.s. hypodigm share a characteristic and distinctive morphology, but they have questioned the wisdom of subsuming within a single species fossil evidence that displays a level of variability that cannot be accounted for by either time or geography. This study addresses these concerns by using appropriate comparative evidence to assess the probability that either the shape, or the size, or both the shape and...
size, of the *P. boisei* s.s. mandibular corpus is more variable than would be expected for a single species. The results of the part of the current study that investigated variation in the shape of the mandibular corpus provide no support for the “excessive variation” hypothesis. Although the *P. boisei* s.s. hypodigm includes mandibular corpora spanning a substantial size range as well as close to a million years of time, and despite earlier demonstrations that the shape of the mandibular corpus of *P. boisei* s.s. is sensitive to differences in overall size (Chamberlain and Wood, 1985), the results of this study suggest that these factors do not result in any more variation in mandibular shape than is found within isochronic species-level samples of modern higher primate species. Indeed, the pairs of mandibles (Peninj 1/KNM-ER 729; KNM-ER 3230/Omo L7a-125) that are most different from one another in shape are among the largest individuals in the sample, and offer no support for the hypothesis that the fossil assemblage consists of a mixture of large-bodied and small-bodied species.

The present study provides more equivocal evidence about whether the *P. boisei* s.s. mandibular hypodigm subsumes excessive variation in corpus size. Both methods employed here to assess size variation, the CV* bootstrap and the FK test, indicate that the size variation observed in the fossil sample significantly exceeds the variation in *Pan* and *Homo*, but suggest that it is not significantly greater than variation in *Pongo* (Table 4). Only in comparisons with *Gorilla* do the two methods differ in their conclusions. The CV* bootstrap tests indicate that the fossil size variation is significantly greater than in *Gorilla*, while the FK test indicates that the fossil size variation is not significantly greater than *Gorilla* (Table 4). Since the FK test is robust to type I and type II errors (Donnelly and Kramer, 1999), the CV* bootstrapping method used here appears to have a higher type I error rate than the FK test. In other words, it is more likely to reject a true null hypothesis (here, that the fossil sample is comprised of one species). In this way, our CV* bootstrapping computation may be statistically more similar to the CV method of Cope and Lacy (1992) than their more recent (Cope and Lacey, 1994) version (see Donnelly and Kramer, 1999). Bootstrapping from both the fossil and extant samples and comparing the resulting distributions of CVs is generally more robust than the above bootstrapping alternatives (e.g., Cope and Lacy, 1992; and method employed in the present study), but this method has an undesirable type I error rate (Donnelly, personal communication).

Leaving aside the subtle differences in results between the two methods, the size variation observed in the fossil sample is substantial, exceeding that in humans, chimpanzees, and possibly gorillas. How should this result be interpreted? It is widely appreciated that the substantial time depth in most paleontological collections can inflate variation through temporal trends in size and shape, the effects of random drift, etc. These potential sources of variation must be considered when assessing the taxonomic significance of variation in a sample comprised of individuals pooled over considerable space and time. In this fossil sample, however, the shape differences in the pooled fossil sample are minimal compared to extant populations sampled from an instant in geological time. With respect to size, although there are no detectable trends in this taxon for increase or decrease in mandibular corpus size over time (Wood et al., 1994), this does not mean that mandibular corpus size might not have fluctuated through time, thus inflating the variation within a temporally extended sample. Perhaps the main reason for comparing variation in the fossil hypodigm with size variation in *Gorilla* and *Pongo* species is that the substantial within-species variation in these taxa, compared to that in *Pan* and *Homo*, helps compensate for the lack of a temporal dimension in the otherwise most appropriate extant comparators. However, the fact that the largest and smallest mandibles in the fossil sample are roughly contemporaneous in the middle of the known time range for *P. boisei* (Wood et al., 1994) suggests that the substantial size variation in the fossil sample cannot be attributed to collapsing time-related changes in size. It has been suggested that the inclusion of multiple subspecies in the comparative extant samples (e.g., *Pongo* and *Gorilla*) may help approximate the temporal and geographic variation in fossil assemblages (Lockwood et al., 1996; Richmond and Jungers, 1995), and it may be significant that these are comparative samples that come closest to the variation subsumed in the fossil sample. Thus, the balance of evidence suggests that the size and shape variation observed in the fossils attributed to *P. boisei* does not appear to have been heavily influenced by temporal factors.

Paleontological samples are prone to at least one other influence on within-species variation that does not affect extant comparative samples, namely, the processes that modify the size and shape of structures postmortem. For example, in this case, excessive abrasion can lead to the loss of surface bone and thus reduce the size of the mandibular corpus. Likewise, if a fossil is subjected to cycles of changes in temperature and humidity that cause the mandibular corpus to expand and contract, and if matrix permeates the cracks, the combined effect can artificially inflate the size of the mandibular corpus. The mandibles of *P. boisei* s.s. seem to have been particularly prone to abrasion and/or matrix-filled cracking (Wood, 1991), so taphonomic factors must be taken into account when assessing the significance of the results of comparative studies such as this one. We have attempted to assess the extent of these influences on the Koobi Fora component of the sample of *P. boisei* s.s. mandibular corpora by scoring each mandible on the 5-point scale set out in Figure 6. This relatively crude attempt to take into
account the affects of postmortem change suggests that there is an association between preservation score and overall size. The small mandibles all suffered from surface erosion, whereas the dominant effect on the large mandibles is matrix-filled cracking that inevitably artificially inflates the size of the mandibular corpus in these individuals (Fig. 6). Thus, taphonomic factors must account for some of the variation captured through the raw, uncorrected measurements of the *P. boisei* s.s. mandibular corpora. This must cast further doubt on the hypothesis that the juvenile mandible KNM-ER 1820, together with four large adult mandibles assigned to *P. boisei* s.s. (see above), provide sufficient evidence to justifying recognizing a second taxon of post-ca. 2.3 Myr East African *Paranthropus*. 

Although it is impossible to determine the exact amount of surface bone loss due to erosion or expansion resulting from infilled matrix, we have attempted to reconstruct the original dimensions of the corpus profile at M1 of the affected mandibles, based on our experience with unaffected fossil and extant mandibles. Measurements are increased by using the available preserved surface bone to estimate the height and width of the corpus in the pristine state. Likewise, the total widths of matrix-infilling are subtracted from the actual measurements of the corpus. This procedure is comparable with attempts to correct for interstitial wear (Wood and Abbott, 1983; Wood, 1991). The CV* of the “adjusted” sample drops from 12.3 to values between 10.6–11.5, depending on how conservative or generous the adjustments are. A CV* of 10.6 makes the size variation in the fossil hominin sample comparable to that in *Gorilla* (*P* = 0.25) and *Pongo* (*P* = 0.88), but it still significantly exceeds the variation in *Pan* (*P* = 0.00) and *Homo* (*P* = 0.01).

The “adjusted” CV* provides a more conservative estimate (0.14) of lnISD (Fig. 5b). Therefore, if the fossil sample is representative of the species, then the level of dimorphism is very close to the average dimorphism for our gorilla sample, and comparable to that in extant orangutans. This corroborates findings by Lockwood (1999) on the level of sexual dimorphism in the face of *P. boisei* (see also Leakey and Walker, 1988). Of course, this fossil sample could have been drawn from a species with higher or lower average dimorphism (Fig. 5b), but it is very...
improbable that it was as low as that observed in chimpanzees and humans. The level of sexual dimorphism in *P. boisei* is similar to those of *A. aferensis* (Johanson et al., 1978; Cole and Smith, 1987; McHenry, 1988; Richmond and Jungers, 1995; Lockwood et al., 1996; Lockwood, 1999) and possibly *A. anamensis* (Leakey et al., 1998), but may be slightly higher than the moderate to high levels apparent in the face of *A. africanus* (Lockwood and Tobias, 1999). Thus, it appears that when compared to extant higher primates, early hominins are characterized by moderately high, to high, levels of sexual dimorphism. *H. ergaster* currently provides the first clear evidence of a hominin that shows a decrease in sexual dimorphism to levels comparable to those of modern humans (Wrangham et al., 1999), apparently as a consequence of increase in female body mass (McHenry, 1994). Later members of the genus *Homo* also appear to exhibit a lower, more modern humanlike level of sexual dimorphism (Arsuaga et al., 1997), and is consistent with the hypothesis that an important adaptive shift occurred between early hominins and *Homo sensu stricto* (Wood and Colppard, 1999).

Thus, if there is a case for taxonomic heterogeneity within *P. boisei s.s.*, then it must rest on evidence other than that from the mandibular corpus. The finding that the effects of taphonomic processes may mimic excessive variation within a fossil hominin sample is salutary. Researchers should take taphonomy into account when investigating whether apparently excessive within-sample variation merits taxonomic recognition.

**ACKNOWLEDGMENTS**

We thank Leslie Aiello for providing the ATD bootstrapping program, Steve Donnelly for advice about the FK method, and the following individuals for allowing access to the primate collections in their care: Linda Gordon and David Hunt, Smithsonian Institution, National Museum of Natural History; and Bruce Latimer and Lymen Jellema, Cleveland Museum of Natural History. We also thank the three anonymous reviewers for their valuable comments.

**LITERATURE CITED**


