How Many Variables are Too Few? Effect of Sample Size in STET, a Method to Test Conspecificity for Pairs of Unknown Species

SANG-HEE LEE
Department of Anthropology, University of California at Riverside, Riverside,CA 92521-0418, USA; sang-hee.lee@ucr.edu

ABSTRACT

One of the questions most often asked in paleoanthropology is whether the amount of variation in a fossil sample is too much to be from a single species. STET (STandard Error Test) uses standard error of the coefficient from a linear regression model relating a pair of specimens to ask if variation in a fossil sample needs to be explained as the presence of multiple species. Previous studies have used this method to test the null hypothesis of no difference in various hominid samples and showed that the variation is not too great to reject the hypothesis of single species. In this paper, properties and limits of STET are explored using skeletal data of known sex and species.

Data include 53 cranial and postcranial measurements of modern humans (n=87) and chimpanzees (n=43). STET values are calculated for all possible interspecific (n=3,741) and intraspecific pairs (n=4,644) to generate distributions of STET. Results show that interspecific STET values are always greater than intraspecific STET values when all 53 variables are used. When variable numbers are arbitrarily decreased by random sampling, STET becomes less effective. Random sampling of variables is repeated 1,000 times to assess the effectiveness of STET. The amount of over-lap between interspecific and intraspecific STET values is used as error rates from 0.01 to 0.10. Minimum number of variables necessary for STET to be effective ranges from 30 (0.10 error rate) to 48 (0.01 error rate).

INTRODUCTION

STET (**ST**andard **E**rror **T**est) assesses whether the variation shown in a pair of specimens measured by dispersion from a regression line is too great for a conspecific pair (Wolpoff and Lee 2001). In various hominid fossil samples, including habilines, australopithecines, and Neandertals, it was shown that the variation observed is not larger than what is expected in a single species (Lee and Wolpoff 2005, 2007; Wolpoff and Lee 2001, 2006).

Missing data constitute a major problem in analyzing fossil data. In a multivariate analysis, all individual specimens need measurements for all variables. When a data matrix includes missing data, if there are not too much missing data and if data are missing at random, missing data values can be estimated using various methods such as maximum likelihood or multiple imputation (Little and Rubin 2002). The exact proportion of missing data for imputation to be valid varies on a case-by-case basis. However, it is often the case that the proportion of missing data in a fossil data matrix is more than 20%, which is far greater than what is generally suggested (Cohen and Cohen 1984). Alternatively, either the specimen that lacks those variables or the variable that does not have all the values may be excluded (list- or pair-wise deletion). However, this approach is considered poor and unadvisable in modern statistics (Little and Rubin 2002). Because STET compares each pair of specimens using the measurements that are available for that particular pair, not all specimens in a dataset need to have all the variables preserved. Therefore, STET holds promise in analyzing fossil data, allowing studies of several fossil specimens using a larger number of variables.

However, properties of STET under various conditions have not been examined. Because STET employs the number of variables in its calculation (see below), it was suspected that there might be a threshold of minimum number of measurements that is necessary to use STET accurately. In previous studies, a conservative number of 40 was determined to be the cut-off for the number of measurements from visual inspection—Lee and Wolpoff inspected more than 300 STET values with various number of variables, and concluded that STET values were stable when the number of variables used was 40 or more (Lee and Wolpoff 2005; Wolpoff and Lee 2001, 2006). However, this threshold of 40 variables was not investigated systematically. The purpose of this paper is to derive the minimum number of variables needed to apply STET, using skeletal samples of known sex and species.

MATERIALS AND METHODS

Metric data on modern humans and chimpanzees were collected by the author using the Hamann-Todd Collection housed at the Cleveland Museum of Natural History, which includes a large selection of non-human primates and humans of known sex. Only complete, adult individuals with fully erupted permanent dentition were used. For humans, only those individuals between 20 and 50 years

PaleoAnthropology 2011: 260–267. ISSN 1545-0031
doi:10.4207/PA.2011.ART56

of age were included in the data set. Using digital sliding calipers, spreading calipers, tape measure, and osteometric board, measurements were taken on the left side of each individual for consistency. When the left side was absent, the right side was measured in substitution. Only cases with reasonable bilateral symmetry were included in this study.

Fifty-three variables were measured for humans and chimpanzees, including 32 cranial and 21 post-cranial measurements. Variables include standard osteometric measurements (Martin and Saller 1957) and non-standard measurements often used in hominid fossil studies (Appendix 1). With postcranial measurements, the range of measurements is greater than that observed in cranial measurements only. Therefore, the regression slope will be influenced by the postcranial measurements, which are larger than cranial measurements, raising the possibility of the pull of the outliers (Figure 1). However, since it is the dispersion of the points around the regression line, and not the significance of the slope itself that is being examined in the method, the inclusion of postcranial measurements is not expected to influence the results of the analysis.

The data set consists of 130 individuals: 87 modern humans (50 males, 37 females) and 43 chimpanzees (18 males, 25 females). Each specimen from the dataset was paired with each other, leading to 8,385 pairs. These pairs were grouped into three—two intraspecific groups ("HH" including all human-human pairs [n=3,741] and "PP" including all chimpanzee-chimpanzee pairs [n=903]) and one interspecific group ("HP" including all human-chimpanzee pairs [n=3,741]). STET values for pairs in these three groups (HH, PP, and HP) were calculated under various conditions to examine the effect of sample size in applying the method.

STET is similar in approach to the methods suggested by Lovejoy (1979) and formulated by Thackeray and colleagues (1997; 1995). All address the question of if the observed variation in a sample can be interpreted to reflect species differences, using a statistic based on the standard error of the slope of regressions between pairs of specimens that relates all of the homologous measurements each pair shares. The rationale of these methods lies with the argument that when a number of measurements from a pair of individuals are plotted, the points would be more dispersed if the pair were sampled from two different species. Thus a measurement of dispersion, such as STET, would indicate whether the pair were conspecific or not (see Figure 1).

As in the method proposed by Thackeray and colleagues, STET uses the standard error of the slope from the least-squares regression in a bivariate plot of linear measurements between two specimens. In these comparisons each specimen is plotted against another specimen with the measurement values of one specimen acting as the x-axis coordinates and the values of the same measurements of the other acting as the y-axis coordinates. It is the dispersion of variables around the regression line that is important for this test, not the slope of the line itself. The rationale of this approach is that the dispersion would be low if the variation is due to geometric and allometric shape similarities.

STET is different from the method used by Thackeray and colleagues. The modifications were made to address two issues. First, the method used by J.F. Thackeray and colleagues compares a pair of female-male specimens, for each axis. However, this approach had a limitation for examining fossil specimens, where sex is not known for the majority (if not all) of the specimens. Second, as the regression method minimizes the deviation of the dependent variable from the regression line (and the independent variable is assumed to have no variance), the standard error (and the regression slope) differs when the independent and dependent variables (in this case, specimens) switch. There is no basis to choose which specimen is dependent and/or which is independent. One solution to this problem would be to calculate the reduced major axis regression, which minimizes the perpendicular distance of each point from the regression line. The disadvantage of a reduced major axis approach is that there is no direct way to calculate the standard error of its slope (Sokal and Rohlf 1995).

For these reasons, an alternative was developed—standard errors of the mean for each comparison ($s.e._{mx}$ for the linear regression of X on Y and $s.e._{my}$ for Y on X) was calculated, squared, and added to each other. STET was the square root of that measure, multiplied by 100:

$$STET = 100*[(s.e._{mx})^2+(s.e._{my})^2]^{½}.$$

One could think of STET as a hypotenuse joining the sides of a triangle determined by the two orthogonal standard errors.

STET values would reflect similarity (and conspecificity) between a pair of specimens. A pair of specimens from two different species would have a STET value that is higher than a STET value from a pair of specimens from a single species (see Figure 1). The rationale is that when a number of measurements from a pair of individuals are plotted, the points would be more dispersed if the pair was sampled from two different species. Thus a measurement of dispersion, STET value, would indicate whether the pair were conspecific or not. That is, a low STET value indicates a single species represented in a pair, while a high STET value indicates that the two specimens are from two different species.

Using the data described earlier, this paper explored the properties and behavior of the STET method by asking two questions. First, it was asked if STET values reflect variation above or below the species level. For all the pairs of the three groups (two intraspecific, HH and PP, and one interspecific, HP), the STET value was calculated, and the distributions of the STET values for the three groups were compared. It was expected that there would be a clear and discrete difference between interspecific STET values and intraspecific STET values, with the former being consistently larger than the latter.

The second question examined was the effect of the number of variables used for STET values to be meaningful. Thackeray and colleagues developed and tested their

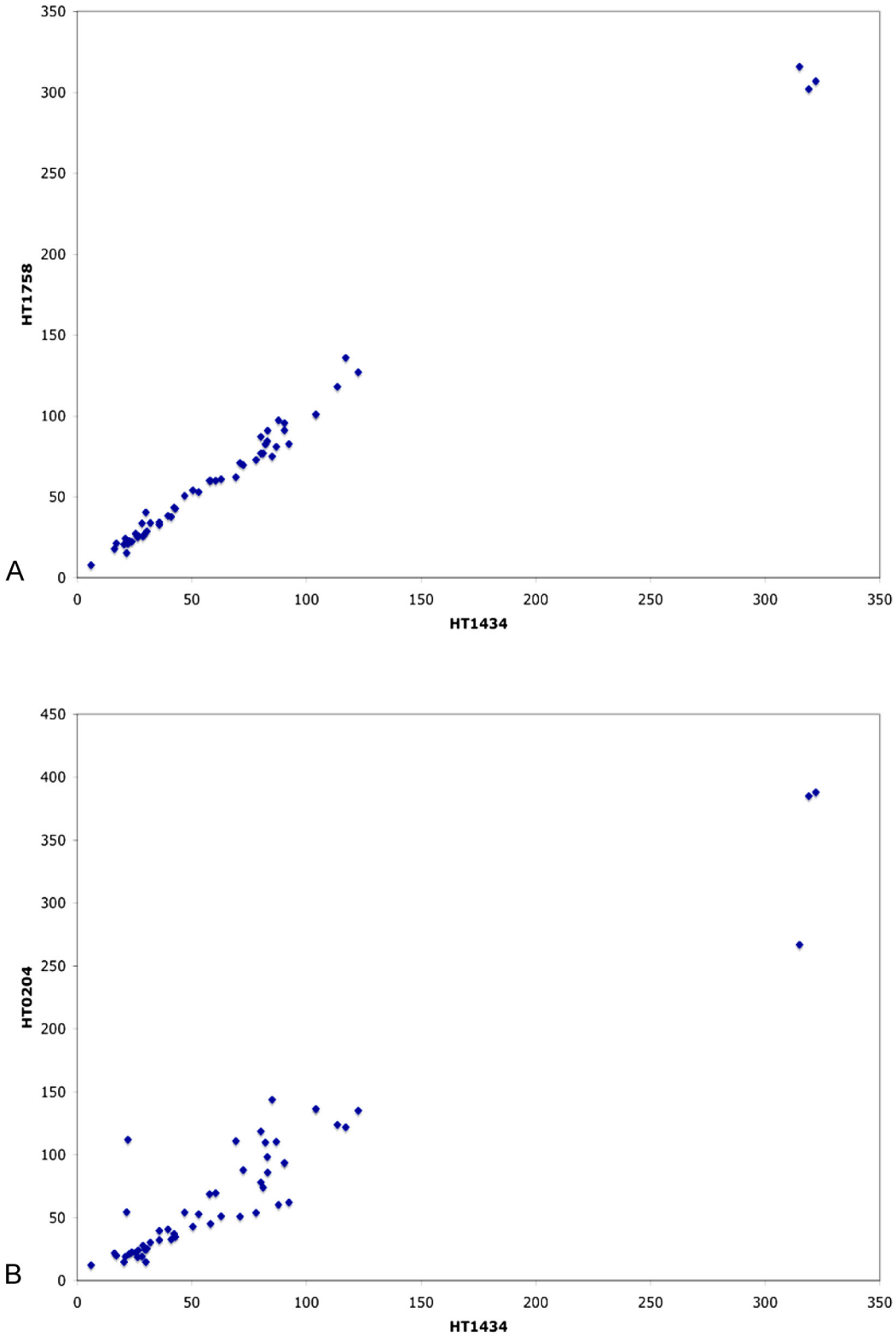

Figure 1. Scatter plots of 53 variables: a) chimpanzee (female) and chimpanzee (male) (STET=1.65); b) chimpanzee (female) and human (female) (STET=5.43). The different STET values reflect difference in dispersion from the best fitting straight line. The three outliers of the scatterplots are long bone lengths. The STET values do not differ whether they are included or not.

method on 10 cranial and mandibular variables from a sample of 1260 specimens representing 70 taxa (Thackeray et al. 1997; Thackeray et al. 1995; Thackeray et al. 2005), while Aiello and colleagues provided a test of their method

using 20 craniodental variables in eight extant non-human primate species and seven fossil hominid specimens (Aiello et al. 2000). However, it is suspected that these values are influenced by sample size, i.e., the number of variables, and 10 or 20 may not be a sufficient number for a valid application of the method. Figure 2 shows two different situations when 10 variables are randomly selected. In one case (see Figure 2a), there is no overlap between the intraspecific STET values and the inter-specific STET values. That is, interspecific STET values are always larger than intraspecific STET values. In this case, a division point will be without error, where a STET value higher than the division point will always indicate an interspecific pair and a STET value lower than the division point will always indicate an intraspecific pair. However, in another case (see Figure 2b), there is a significant overlap between the intraspecific STET values and the interspecific STET values. Within the overlap, intraspecific STET may be greater than interspecific STET values. The amount of overlap, then, can be used to assess the error of STET.

It is expected that, as the number of variables decreases, the distribution of STET values increases in variance and the amount of overlap increases, and thus the reliability of STET decreases. In this paper, error of STET is estimated by generating distributions of STET under various sample sizes and examining the degree of overlap between intraspecific and interspecific groups. For example, the sample size where 99.5% of interspecific STET values exceed 99.5% of intraspecific STET values, with 1% of overlap, is the minimum sample size with 1% error. In the same way, different levels of errors were examined—2%, 5%, and 10%. This error is one-tailed, because the overlap occurs in one side of a distribution (see Figure 2, where the overlap occurs between the high STET values in the intraspecific samples and the low STET values in the interspecific sample).

For a systematic examination of the effect of the number of variables used, the original dataset was modified into subsamples after deleting variables; the number of variables varied from a minimum of 5 (with 48 variables deleted) to a maximum of 53 (no variables deleted). For example, there is only one way to include all 53 variables, 53 ways to sample 52 variables, and 1378 ways to sample 51 variables. For variables size 51, 52, and 53, all possible subsamples have been examined.For variables of size 50 and less, there is an extremely large number of different ways to draw a subsample of a specific number of variables, and the number of possible sampling exponentially increases with more variables to be deleted. For instancde, there are 23,426 ways to delete 3 out of 53 variables (and sampling 50 variables), and the number increases to 292,825 ways for deleting 4 out of 53 variables (and sampling 49 variables). Performing an exhaustive permutation for all possible sampling scenarios is unnecessary and would take a very long time. In this paper, sampling was repeated 1,000 times as a sufficient number of runs for each number of variables sampled from 5 to 50.

RESULTS

SPECIES AND VARIATION

STET distributions show distinctively different values between interspecific and intraspecific pairs. The STET distributions for the two intraspecific samples, HH (*Homo*) and PP (*Pan*), are virtually identical (Figure 3). The resulting HH distribution has a mean of 1.50, a median of 1.40, and a standard deviation of 0.49. The resulting PP distribution has the mean of 1.68, median of 1.65, and the standard deviation of 0.37. When each of the human data was paired with each of the chimpanzee data to generate a distribution of interspecific STET values (n=3,741), the HP distribution was distinctly different from either the human or the chimpanzee data (see Figure 3). The resulting HP distribution has a mean of 5.98, a median of 5.94, and a standard deviation of 0.60. There is almost a four-fold increase in the mean STET value from an intraspecific pair to an interspecific pair.

There is no overlap between the PP distribution and the HP distribution—the maximum STET value for the PP distribution 3.10, which is less than the minimum STET value for the HP distribution, 3.86 (Table 1). The maximum STET value for the HH distribution (4.11) is larger than the minimum value for HP (3.86), and within the range of HP. However, it is a statistically insignificant amount—values 3.86 or greater are observed in the frequency of 0.0008 (0.08%) from the HH distribution. Also, values 4.11 or less are observed less than 0.001 (0.1%) from the HP distribution.

EFFECT OF SAMPLE SIZE

STET performs better when a large number of variables are used. When the number of variables was 50 or more, interspecific STET values always exceeded intraspecific STET values. Allowing for 1% error, interspecific STET values exceeded either of the intraspecific STET values when sample size was 39. Sample size of 48 was necessary for 99% of interspecific STET values to exceed both intraspecific samples.

When fewer variables are used, STET distributions increased in variance. In other words, when a small number of variables were used, there was a greater chance that pairs of the same species yielded STET values higher than those generated by pairs of different species. Consequently, there was a greater degree of overlap between interspecific and intraspecific STET values. Table 2 lists the degree of overlap when different numbers of variables were used. When 21 variables were used, STET values from pairs of different species were larger than those from pairs of the same species 90% of the time, indicating a 10% error. When 39 variables were used, STET values from pairs of different species were larger than those from pairs of the same species more than 98% of the time, with 2% error.

DISCUSSION AND CONCLUSION

STET provides a promising approach in addressing the taxonomy issue in terms of morphological variation. The problem of missing data can be addressed by examining

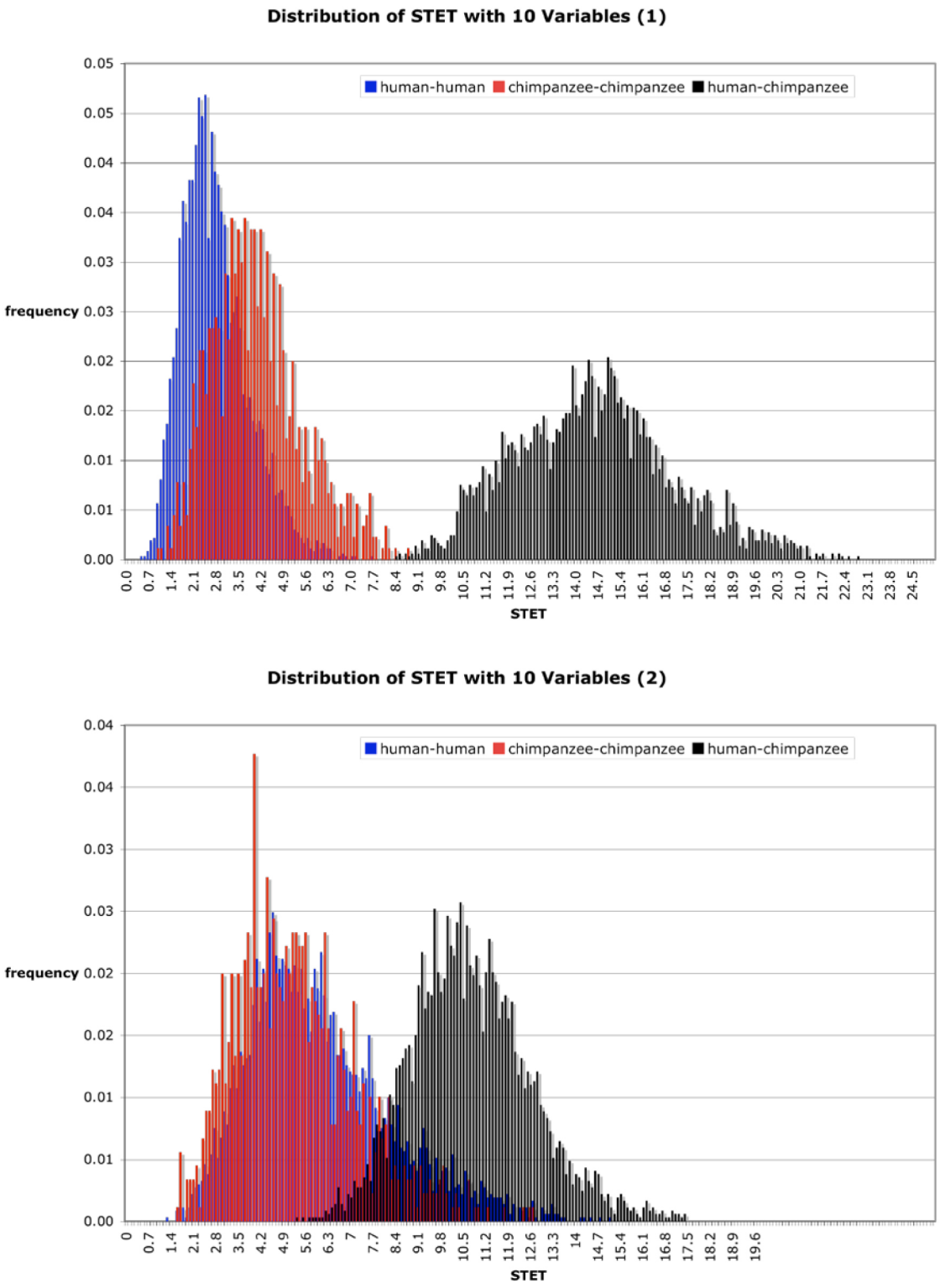

Figure 2. Distribution of STET values with 10 variables. Above, STET distributions using 10 randomly selected variables. Interspecific STET values (in black) always exceed intraspecific STET values (in red and blue). Below, STET distributions using another set of 10 randomly selected variables. The variables are different from the above case. Interspecific STET values (in black) overlap with intraspecific STET values (in red and blue).

metric variables shared by a pair of specimens. Results of this study provide support for the validity of previous studies using STET where only STET values based on 40 or more variables were included in the analyses (Lee and Wolpoff 2005, 2007; Wolpoff and Lee 2001, 2006). The reliability of STET is quite high when 40 variables are used,

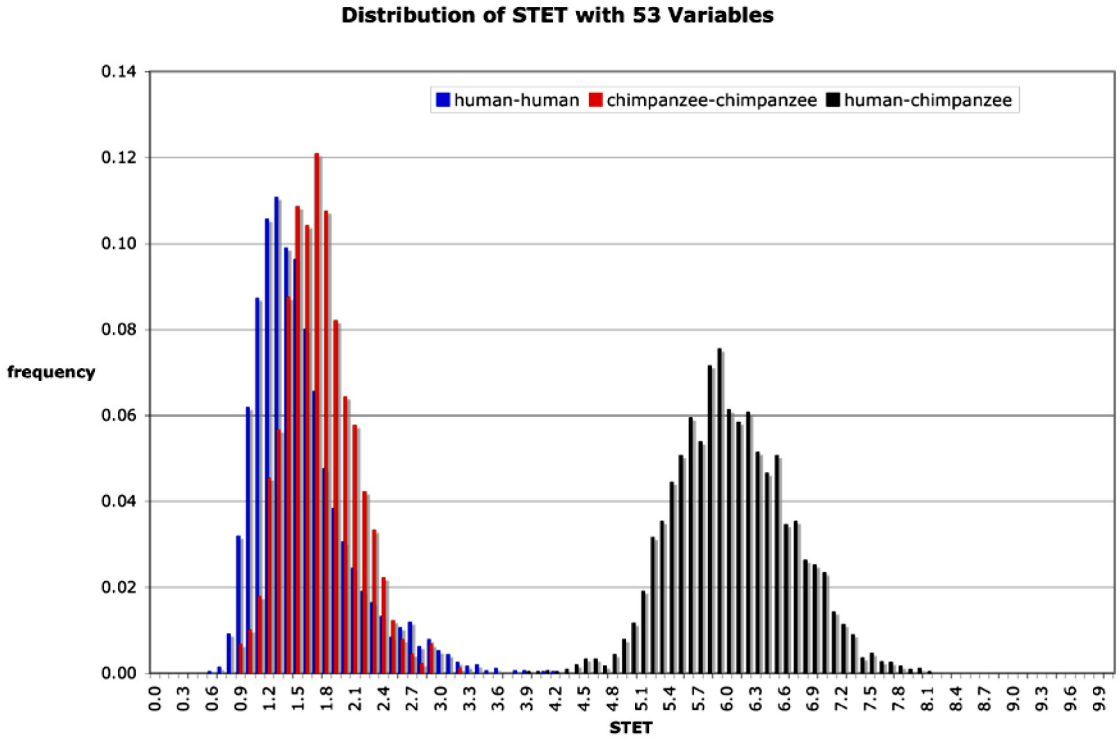

Figure 3. STET values for inter-species and intra-species comparisons of chimpanzee and human samples. Data for this figure are summarized in Table 1. The interspecific STET values (in black) and intraspecific STET values (in red and blue) comparisons do not overlap.

with an error rate of 2%.

This study is limited in its scope, both in the composition of the comparative sample and the number of variables used. A further study including other non-human apes such as orangutans and gorillas, using a larger set of measurements is necessary in particular to address allometric effect of the method.

ACKNOWLEDGEMENTS

The author thanks Yohannes Haile-Selassie and Lyman Jellema at the Cleveland Museum of Natural History for the permission and support to study the Hamann-Todd Osteological Collection, and John Hawks for his help in improving the manuscript. This research was partly supported by the University of California at Riverside.

REFERENCES

Aiello, L.C., Collard, M., Thackeray, J.F., and Wood, B.A. 2000. Assessing exact randomization-based methods for determining the taxonomic significance of variability in the human fossil record. *South African Journal of Science* 96, 179–183.

Cohen, J. and Cohen, P. 1984. *Applied multiple regression/correlation analysis for the behavioral sciences*, Second Edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

TABLE 1. STET VALUES FOR INTER- AND INTRASPACIFIC COMPARISIONS IN FIGURE 3.*

	Pan-Pan	*Homo-Homo*	*Pan-Homo*
Number of pairs	903	3741	3741
Median	1.65	1.39	5.94
Mean	1.68	1.50	5.98
Standard Deviation	0.37	0.49	0.60
Minimum	0.82	0.59	3.86
Maximum	3.10	4.11	8.03

*Data based on *Pan* (n=43) and *Homo* (n=87) sample using 53 homologous measurements.

TABLE 2. NUMBER OF VARIABLES COMPARED AND DEGREE OF OVERLAP BETWEEN INTRA- AND INTERSPECIFIC STET DISTRIBTIONS.

Number of variables	Interspecific STET is larger than either intraspecific STET	Interspecific STET is larger than both intraspecific STET
21	90%	
28	95%	
30		90%
33	98%	
39		95%
45		98%
50	100%	100%

Lee, S.-H. and Wolpoff, M.H. 2005. Habiline variation: a new approach using STET. *Theory in Biosciences* 124, 25–40.

Lee, S.-H. and Wolpoff, M.H. 2007. Herto and the Neandertals: what can a 160,000-year-old African tell us about European Neandertal evolution?, in: Sankhyan, A.R. and Rao, V.R. (eds.), *Human origins, genome and people of India: genomic, palaeontological and archaeological perspectives*. Allied Publishers, New Delhi (India), pp. 329–336.

Little, R.J.A. and Rubin, D.B. 2002. *Statistical analysis with missing data*, Second Edition. Wiley-Interscience. Hoboken, New Jersey.

Lovejoy, C.O. 1979. Contemporary methodological approaches to individual primate fossil analysis, in: Morbeck, M.E., Preuschoft, H., and Gomberg, N. (eds.), *Environment, behavior, and morphology: dynamic interactions in primates*. Gustav Fischer, New York, pp. 229–243.

Martin, R. and Saller, K. 1957. *Lehrbuch der anthropologie*. Gustav Fischer Verlag, Stuttgart.

Sokal, R.R. and Rohlf, F.J. 1995. *Biometry*, 3rd Edition. W. H. Freeman, New York.

Thackeray, J.F., Bellamy, C.L., Bellars, D., Bronner, G., Bronner, L., Chimimba, C., Fourie, H., Kemp, A., Krüger, M., Plug, I., Prinsloo, S., Toms, R., Van Zyl, A.J., and Whiting, M.J. 1997. Probabilities of conspecificity: application of a morphometric technique to modern taxa and fossil specimens attributed to *Australopithecus* and *Homo*. *South African Journal of Science* 93, 195–196.

Thackeray, J.F., Helbig, J., and Moss, S. 1995. Quantifying morphological variability within extant mammalian species. *Palaeontologia Africana* 31, 23–25.

Thackeray, J.F., Maureille, B., Vandermeersch, B., Braga, J., and Chaix, R. 2005. Morphometric comparisons between Neanderthals and 'anatomically modern' *Homo sapiens* from Europe and the Near East. *Annals of the Transvaal Museum* 42, 47–51.

Wolpoff, M.H. and Lee, S.-H. 2001. The late Pleistocene human species of Israel. *Bulletins et Mémoires de la Société d'Anthropologie de Paris* 13, 291–310.

Wolpoff, M.H. and Lee, S.-H. 2006. Variation in the habiline crania—must it be taxonomic? *Human Evolution* 21, 71–84.

APPENDIX 1

	Description	*Martin*
	Cranium	
1	Maximum cranial breadth	M8
2	Maximum bi-parietal breadth	B
3	Minimum frontal breadth	M9(1)
4	Maximum frontal breadth	M10 (B″)
5	Biauricular breadth	M11
6	Biasterionic breadth	M12
7	Infratemporal breadth	M14
8	Bregma-Inion	
9	Bregma-Lambda	M30 (S′2)
10	Auricular point-Bregma	
11	Auricular point-Inion	
12	Auricular point-Asterion	
13	Lambda-Inion	
14	Inner biorbital breadth	M43(1) (EOW)
15	Biorbital breadth	M44
16	Zygomatic bone length along zygomaxillary suture	
17	Bizygomatic breadth	M45 (J)
18	Facial breadth	M46(a)
19	Facial height	M48
20	Nasion-Prosthion	M48
21	Nasospinale-Prosthion	M48(1)
22	Supraorbital torus thickness, central	
23	Anterior interorbital breadth	M50 (IOW)
24	Orbital breadth	M51 (O1)
25	Orbit height	M52 (O2)
26	Nasal breadth	M54 (NB)
27	Nasal height	M55 (NH′)
28	Mandibular fossa breadth	
29	Mandibular fossa length	
30	Auricular point-mastoidale	
31	Supramastoid crest breadth	M8(c)
	Mandible	
32	Mandible corpus height at P4/M1	
	Humerus	
33	Humerus length	M1
34	Maximum head diameter	
35	Midpoint shaft diameter (A-P)	M5
36	Midpoint shaft diameter (M-L)	M6
37	Biepicondylar breadth (maximum)	M4
38	Minimum shaft circumference	M7
39	Midshaft circumference	M7(a)
40	Breadth of the trochlea posterior face (ridge-ridge) at the base of fossa	
41	Breadth of the articular surface of the anterior face	
	Femur	
42	Maximum femur length	M1
43	Morphological length (standing on the condyles)	
44	Shaft circumference below lesser trochanter	
45	Midshaft AP diameter	M6
46	Midshaft ML diameter	M7(a)
47	Midshaft circumference	M8
48	ML diameter below lesser trochanter	M9
49	AP diameter below lesser trochanter	M10
50	Perpendicular head diameter	M18
51	Biepicondylar breadth	M21
52	Distal articular surface breadth	
53	Lateral condyle breadth at base midpoint	M21(e)