Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theory in Biosciences 124 (2005) 25–40
Theory in Biosciences
www.elsevier.de/thbio

Habiline variation: A new approach using STET

Sang-Hee Lee[a,*], Milford H. Wolpoff[b]

[a]*Department of Anthropology, University of California at Riverside, Riverside, CA 92521-0418, USA*
[b]*Department of Anthropology, University of Michigan, Ann Arbor, MI 48109-1382, USA*

Received 10 June 2004; accepted 14 January 2005

Abstract

The problem of whether the hominid fossil sample of habiline specimens is comprised of more than one species has received much attention in paleoanthropology. The core of this debate has critical implications about when and how variation can be explained by taxonomy. In this paper, we examine the problem of whether the observed variation in habiline samples reflects species differences. We test the null hypothesis of no difference by examining the degree of variability in habiline sample in comparison with other single-species early hominid fossil samples from Sterkfontein and Swartkrans (Sterkfontein is earlier than the habiline sample, Swartkrans may be within the habiline time span). We developed a new method for this examination, which we call STandard Error Test of the null hypothesis of no difference (STET). Our sampling statistic is based on the standard error of the slope of regressions between pairs of specimens, relating all of the homologous measurements that each pair shares. We show that the null hypothesis for the habiline sample cannot be rejected. The similarities of specimen pairs within the habiline sample are not more than those observed between the specimens in the australopithecine samples we analyzed.

*Corresponding author. Tel./fax: +1 951 827 5409.
E-mail addresses: sang-hee.lee@ucr.edu (S.-H. Lee), wolpoff@umich.edu (M.H. Wolpoff).

1431-7613/$ - see front matter © 2005 Elsevier GmbH. All rights reserved.
doi:10.1016/j.thbio.2005.01.004

Introduction

When the East African discoveries first began to accumulate in large numbers in the early 1960s, attempts were made to interpret them within the framework of Robinson's dietary hypothesis, developed to explain anatomical variation in the South African australopithecines in an adaptive context (Robinson, 1963). The Leakeys' discovery of the first *Homo habilis* specimen, the juvenile OH 7, set the stage for the interpretation of what was then viewed as a smaller jawed, larger-vaulted East African gracile australopithecine variety, but one that linked *Australopithecus africanus* with *Homo erectus* (Leakey et al., 1964). Most of the evolutionary models proposed to explain this were variants of the dietary hypothesis. It seemed as though South African *A. robustus* evolved into East African *A. boisei* by continuing the trend for dental robustness, while *A. africanus* evolved into *H. habilis* by continuing the trend for encephalization and dental reduction (in Robinson's scheme, *A. robustus* expressed the ancestral condition for *A. africanus*, even though its remains were dated later). But there were also questions raised about whether *A. africanus* and *H. habilis* were all that distinct. A second related, perhaps more significant worry was about whether *H. habilis* was composed of more than one species: was there too much variation in the habiline sample to be attributable to a single species, and how one might tell?

This issue was first raised by Robinson (1965, 1966), who contended that *H. habilis*, as described at Olduvai Gorge (Leakey et al., 1964; Tobias and von Koenigswald, 1964), was a mix of two species. For Robinson, the specimens from Olduvai Bed I were an advanced *A. africanus,* and from Olduvai Bed II were early *H. erectus.* This was the first splitting of the habiline sample, and at a time when the sample was at its smallest. Clearly, its dramatic expansion in size has not put the issue to rest.

The issue of multiple habiline species was revisited with the discovery of the Lake Turkana specimen ER 1470, and other Turkana discoveries, where it was specifically cast as a question of whether sex or species accounted for the differences between specimens such as ER 1470 and ER 1813 (Fig. 1; Chamberlain, 1989; Kramer et al., 1995; Lieberman et al., 1988, 1996; Miller, 1991, 2000; Tattersall, 1992; Wood and Collard, 1999a). Of the individual specimens, the largest cranium, ER 1470 seemed the most problematic; with Walker (1976) considering it an australopithecine, while Alexeyev (1986) placed it in *Pithecanthropus rudolfensis*. This subsequently became the "*H. rudolfensis* Aleveyev" of Wood (1987, 1999). The habiline specimens have been extremely well published (Tobias, 1991; Wood, 1991), and this may account for the many questions about their variation and the new approaches that have been used to resolve them. In the end, Tobias, who started it all, maintains the contention that a single taxon best describes the variation (2003).

The explanation of variability is not just a problem for paleoanthropologists, but is the fundamental problem of all biology. The specific issue raised in the habiline case is when variability *must* be explained by taxonomy. This is quite different than asking whether variation *could* be explained by taxonomy. Surely the past decades have shown that the taxonomic explanation of variation is so easy to apply

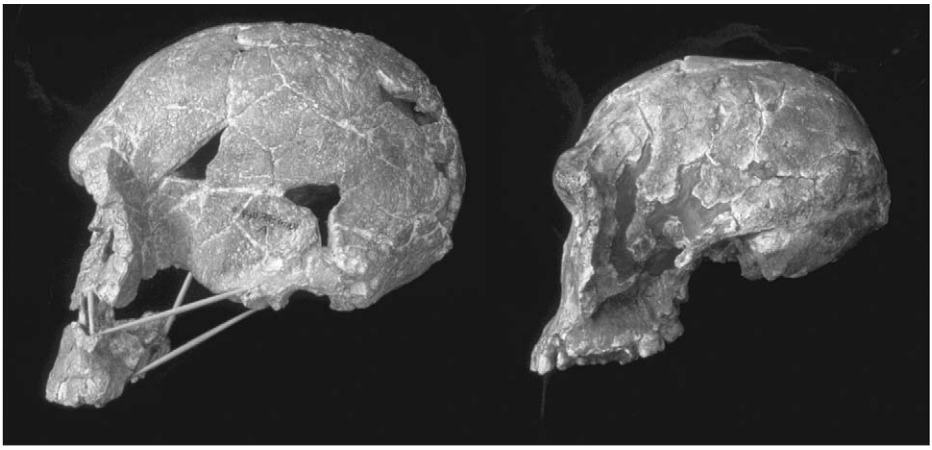

Fig. 1. The two habiline specimens causing most current interpretative problems are these Turkana crania, ER 1813 (right) and ER 1470. The cranial capacities for these specimens are, respectively, 509 and 752 cc. Are these different sexes or different taxa? Both specimens illustrated are casts.

(Tattersall, 2000) that the potential number of hominid taxa seems without end, even *within* sites such as Skhul (Schwartz and Tattersall, 2000), or Dmanisi, where Dmanisi crania and mandibles were distinguished (Schwartz, 2000) that later prove to belong to the same specimen. The "could" approach is wrong-minded, and inevitably implies that speciation is the main cause of evolutionary change and therefore of variation, and that adaptation is an illusory consequence of other processes (Tattersall, 1999, 2000). Accept a speciose taxonomy and one must buy into the theory explaining it. Tobias (2003) muses that Louis Leakey, the great splitter of the mid-20th century, would feel comfortable with this taxonomic approach and its consequences. We do not think so. We believe that in the context of many of today's publications Leakey would now appear to be one of the more prominent lumpers.

Let us return to the problem of whether variability *must* be explained by taxonomy or variation *could* be explained by taxonomy. "Must" is a hypothesis that can be disproved, "could" is a self-fulfilling prophesy that sidetracks many ways of understanding the evolutionary process. Little wonder that Tattersall (2000, p. 2) could describe the evolutionary synthesis as orthogenesis and complain: "the synthesis was doomed to harden … into a dogma: a dogma whose heavy hand continues to oppress the science of human origins a half-century later." So much for Huxley, Mayr, Simpson, and Wright.

The issue of how to approach the role of taxonomy in explaining variation is not inconsequential, since it not only reflects how the evolutionary process is understood, but also addresses the very basis of how science works. And taxonomy is not independent of biological process, since variation at or above the species level exists in protected gene pools that may be mixed up with each other but cannot mix

together, while below the species level mixture is expected and normal, and its effect on the pattern and magnitude of variation is significant.

In this paper, we ask whether habiline variability *must* be described at the level of species. We take a null hypothesis approach to this question, and propose a new technique to examine whether the magnitude and pattern of variation are sufficient to disprove the contention that there is a single species in the habiline sample.

Materials and methods

Null hypothesis approach

Differences within species can be quite large, and it is also well known that differences between species can be subtle and minimal, as shown by sibling species (Mayr, 1963), cases of mimicry (Jiggins et al., 2001) and certain other cases of closely related groups (Kimbel and Martin, 1993). One exacerbating aspect of the habiline sample is that the similarities among the specimens are not so small, nor are the differences between them so great, as to allow an obvious well-supported interpretation of whether interspecific or intraspecific variation is represented. However, we might expect to develop tests of a null hypothesis for the entire sample, tests based on the magnitude and pattern of variation *within* the sample that could disprove the contention that the *sample* characteristics require the interpretation of a single species. It is always possible that such tests could fail because there are species differences that cannot be quantified or otherwise observed, this possibility is beyond the capacity of any statistic to address.[1]

Standard error of the paired specimen slope

A new test of this null hypothesis was proposed by Thackeray et al. (1995, 1997), and subsequently modified by us (Wolpoff and Lee, 2001). Thackeray and colleagues quantified their expectations for conspecific variation by examining the standard error of the LMS regression slopes for a series of linear measurements shared by pairs of specimens. In these comparisons each specimen was plotted against another, with the measurement values of one acting as the x-axis coordinates and the values for the same measurements of the other acting as the y-axis coordinates. It is the dispersion of variables around the regression line through the bispecimen plot that is important for this test, not the slope of the line itself. This modification of Q-mode analysis compares specimens rather than measurements, similar to an approach suggested by Lovejoy (1979). Thackeray and colleagues developed and tested the approach on a large sample of 1260 specimens representing 70 extant vertebrate and

[1]Discussing this issue, Aiello et al. (2000) suggest that a method of multiple working hypotheses (in the sense proposed by Chamberlain, 1965) be used instead of trying to reject the null hypothesis. Apart from the epistemological issues, it is challenging to apply such a method when the difficulties of assessing multiple species hypotheses are considered (Milius, 2001).

invertebrate taxa (Thackeray et al., 1995; Thackeray et al., 1997). Ten disperse measurements of the cranium and mandible were used in an exact randomizations, comparing males to females known to be within the same species within each of the 70 taxa. The slope for each comparison was calculated, along with its standard error. The test statistic these studies examined is the log of the standard error of the slope.

The log standard error of the slope is a measure of the dispersion around the regression line, which is the observation of interest.

> High s.e.$_m$ values relate to high morphological variability when measurements of any two specimens are compared, reflected also by a high degree of scatter of measurements around a regression line. Relatively low s.e.$_m$ values can be expected in situations where there is only a small degree of morphological difference, associated with limited scatter around a regression line, reflecting similarities in shape of two specimens being compared (Thackeray et al., 1997, p. 196).

The standard error was logged because this has a normal distribution for the distribution of values for all 1260 specimens. It was argued that this test statistic reflects the consequences of variation in both size and shape, in that it shows "both geometric and allometric shape similarities between crania rather than just geometric shape similarity" (Aiello et al., 2000, p. 180). Allometry, of course, describes shape differences as a function of size, and we take this statement to mean that this standard error test statistic is sensitive to variation in both size and shape.

The average log of the standard error for the linear regression slope was reported to be -1.78 for the 70 species reference sample, with a standard deviation of 0.27. Thackeray et al. (1997) found that the interval of ± 2 standard deviations around the mean (-1.24 to -2.32) encompasses 95% of the logs of the standard error for the linear regression slope for bispecimen comparisons within the 70 species they studied. Subsequent analysis (Aiello et al., 2000) based on 20 measurements resulted in a 95% upper limit for 15 specimens of *Pan* of -1.32, and a 95% upper limit for 8 non-human primate species (including *Pan*) with sample sizes ranging between 8 and 24, of -1.05.

Strictly speaking, the lower (more negative) confidence intervals are not relevant since comparisons with even smaller values are nonetheless within the same species. The confidence interval determination is one-sided, and therefore all the figures given in these papers are somewhat too large – they actually represent the 97.5% confidence interval. A more serious problem, discussed below, lies in the small number of comparisons used for their determinations.

STET

We have modified the test statistic proposed by Thackeray and colleagues, and proposed changes in the procedure for using it (Wolpoff and Lee, 2001). There are a number of reasons for this. The argument for using a log transform was not compelling, and we did not want to dampen the effects of larger values of our test statistic without sufficient reason. The test procedure did not take advantage of the

significant potential of the approach to compare samples of specimens with uncertain or unknown sex (unless sex determination was the object of the comparison, as in Thackeray et al. (2000)).

The original testing was based on the same set of 10 measurements in 70 species. There were subsequent comparisons of those results with bispecimen analyses of different sets of measurements, as dictated by the state of preservation of the different specimens (Aiello et al., 2000; Thackeray et al., 2000). These comparisons raise problems of sample size that must be addressed. Our modifications respond to these concerns, and to the issues raised in the discussion of error detailed below. We call our modified approach the *ST*andard *E*rror *T*est of the null hypothesis of no difference – *STET*.

Calculation of STET

The linear regression analysis Thackeray and colleagues used minimizes the deviation of the dependent variable from the regression line. For cases where the bispecimen comparison is not symmetric around a linear regression line, the regression of X on Y differs from the regression of Y on X, and the standard errors of the regression slopes differ as well. We find it desirable not to specify pairs of different sex, and so we have no *a priori* reason to choose independent and dependent variables in the bispecimen comparisons. The problem is that the choice of independent and dependent variables has some influence on the standard errors for the linear regression slopes. Thackeray and colleagues always did regressions of females on males, so the assignment of independent and dependent variables was not a problem for them.

One solution to this problem could be to calculate the reduced major axis regression, which minimizes the orthogonal distance of each point from the regression line instead of minimizing the distance along the X- or the Y-axis (the orthogonal distance is the square root of the sum of the squares of the X- and Y-axis deviations). 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, 1981). For this reason we chose a different solution, calculating standard errors of the mean for both comparisons (s.e.$_{mx}$ for the linear regression of X on Y and s.e.$_{my}$ for Y on X) and reporting combined value as the square root of the sum of the squares of the two. One could think of STET as a hypotenuse joining the sides of a triangle determined by the two orthogonal standard errors.

$$\mathrm{STET} = 100[(\mathrm{s.e.}_{mx})^2 + (\mathrm{s.e.}_{my})^2]^{1/2}. \tag{1}$$

This test statistic is not directly comparable to the standard error-based statistics published by Thackeray et al. (1995, 1997), and Aiello et al. (2000). This may not be a problem, however, because as noted below, there are compelling reasons not to make such comparisons.

Error related to sample size

We were concerned about the influence of sample size on these comparisons, how does sample size influence the magnitude of STET? Such an influence is suggested by the differences in the 95% intervals calculated for the two studies cited above: Reporting the statistic Thackeray and colleagues used (log (s.e.$_m$)), 70 species compared with 10 measurements had an upper limit of -1.24, but 8 species compared with 20 measurements had an upper limit of -1.05. One would have expected the larger number of species to encompass a broader range of variation, but it did not. The opposite was the case. We were concerned because our procedure has the potential for different numbers of measurements in every bispecimen comparison. This was not a problem in earlier studies, where the number of measurements was held to a small but constant value – often 10 – but limiting the comparisons to 10 measurements artificially and perhaps detrimentally reduces the power of the comparisons.

We addressed the issue of whether the number of measurements was important by calculating STET values, and seriating bispecimen plots in order of the values for this statistic. We plotted a sample of 311 bispecimen comparisons of hominid crania ranging in time from the Pliocene to approximately 50,000 years ago. For the most part a visual seriation of the dispersion in these plots closely fit the ordering based on STET. The smaller the value for STET, the closer the fit to a straight line. Only a few of the bispecimen comparisons did not seriate with the STET values. These stood out as being either much less or much more disperse than their neighbors in the ordering. In every similar case the sample size for these exceptions was less than 40.

To further specify the influence of sample size on the magnitude of STET, we plotted the STET values as a function of the number of measurements underlying each comparison, and calculated a LMS regression. The effect of sample size becomes random when the sample size gets larger, but the effect is evident in smaller samples whose values for the statistic were sample size dependent. To find when the sample size is large enough to avoid this effect, we examined the residuals from the regression slopes. For the comparisons of specimens with 40 or more measurements we could discern no relationship between the number of measurements and the size of the residuals. Moreover, none of the residuals exceeded 0.8 standard deviations from the expected value. Therefore, we do not report comparisons for specimens with fewer than 40 measurements in common.

Procedure

To examine the habiline sample we calculated STET from paired specimen comparisons of crania that were described by between 45 and 147 homologous linear measurements systematically taken on the original specimens by one of the authors (MHW) over the course of several decades (Fig. 2). The actual measurements used in each comparison, of course, depend on the preservation of the two specimens being compared. These data were recorded for the four complete or mostly complete East

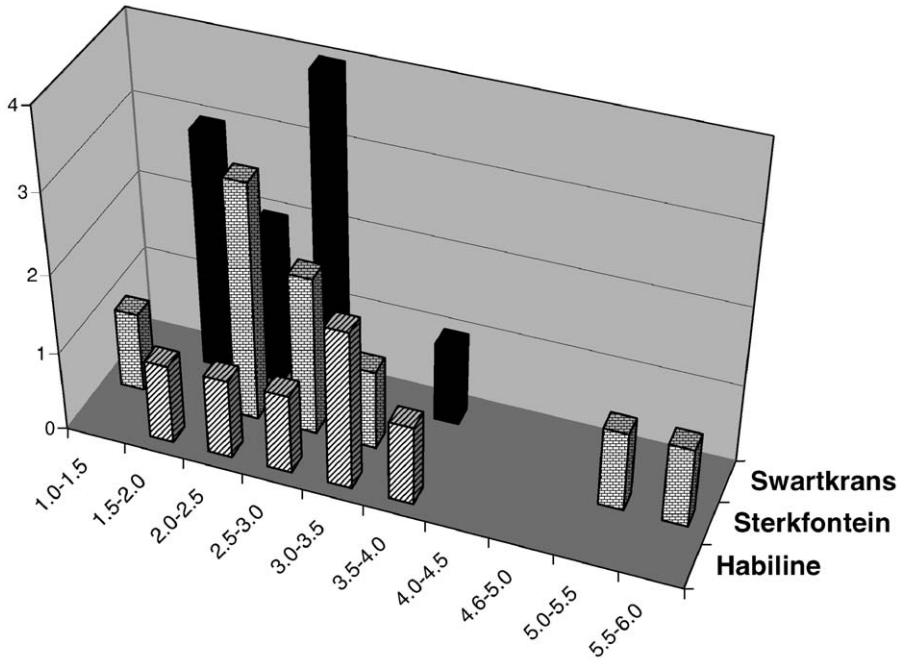

Fig. 2. STET values for the habiline, Sterkfontein, and Swartkrans samples compared (from Tables 1–3). The distributions shown are of all comparisons of individuals within each sample. The Sterkfontein sample is most variable for this statistic, and its maximum STET values greatly exceed the habiline (and Swartkrans) maximum. Both Sterkfontein and Swartkrans samples are comprised of individuals widely if not universally believed to represent a single taxon (regardless of the views of the authors; for instance, the Swartkrans sample does *not* include SK 80/847). Habiline STET variation across the range of comparisons is quite like the other two samples, and certainly not more extreme, as one could expect if the habilines were a mixture of two different taxa. Of course this does not prove the habilines are not such a mixture, is shows only, but significantly, that the null hypothesis cannot be disproved.

African crania that have received most attention in the habiline debate, KNM-ER 1470, KNM-ER 1813, OH 16, and OH 24 (Table 1).

For comparison we examined the dispersion for the most complete *A. africanus* crania have been used: STS 5, STS 19, STS 71, STW 505, and TM 1511 (Table 2). We selected *A. africanus* specimens from a single site, Sterkfontein, to eliminate additional sources of variation. Sterkfontein is older than the entire habiline sample. We chose Swartkrans as a second early hominid site for comparison, and determined STET for the best-preserved crania: SK 46, 48, 52, 79, and 83 (Table 3). Swartkrans overlaps in age with the habiline sample. Both of these Transvaal australopithecine sites have crania that some authors attribute to a habiline species; for instance, STW 53 and SK 80/847. To be conservative we did not include any of these crania in our

Table 1. Comparisons[a] of STET values for habiline specimens

	ER 1470	ER 1813	OH 24	OH 16
ER 1470		140	133	62
ER 1813	1.72		147	46
OH 24	2.58	2.19		45
OH 16	3.41	3.68	3.11	

[a]Comparisons in the lower left portion of the table show STET, the upper right comparisons show the sample sizes for each comparison. All comparisons are based on a sufficient number of measurements.

Table 2. Comparisons[a] of STET values for Sterkfontein specimens

	STS 71	TM 1511	STS 5	STW 505	STS 19	STS 25
STS 71		70	207	93	42	41
TM 1511	1.34		85	*	*	*
STS 5	2.19	2.48		102	59	47
STW 505	2.30	*	3.25		*	*
STS 19	2.76	*	2.83	*		*
STS 25	5.49	*	5.62	*	*	

[a]Comparisons in the lower left portion of the table show STET, the upper right comparisons show the sample sizes for each comparison. An asterisk means the comparison is based on an insufficient number of measurements. The two outliers in the comparisons (see Fig. 2) do not involve the large vault STW 505, but rather the diminutive STS 25.

Table 3. Comparisons[a] of STET values for Swartkrans specimens, all considered "*A. robustus*"

	SK 46	SK 48	SK 83	SK 52	SK 79
SK 46		73	82	78	66
SK 48	1.91		105	118	61
SK 83	2.27	1.77		124	69
SK 52	2.55	1.84	1.98		62
SK 79	2.83	2.69	3.89	2.90	

[a]Comparisons in the lower left portion of the table show STET, the upper right comparisons show the sample sizes for each comparison.

australopithecine samples, although in doing so we do not necessarily agree with these taxonomic assessments.

The size of the measurement set reflects the standardized measurements from R. Martin, the Biometrika school, W.W. Howells data set, and other normally used sources. Added to these were additional measurements developed to allow comparisons of fragmentary cranial remains too incomplete for standard measurements to be

possible. This provides a much larger measurement base than has ever been used before, and one particularly designed to maximize comparisons of specimens that are not complete. While incomplete specimens were used, in all of the cases considered, most (or all) of the vault and at least part of the face was present. We did not consider bispecimen comparisons based on single cranial bones, or on specimens missing entire portions of the cranium. No dental measurements were included, although measurements along the palate defined by tooth positions were in the sample pool. We also did not include mandibular measurements as only a few of the crania have associated mandibles.

Results

STET values indicate the dispersion around the best fitting line for paired comparisons within each sample (Fig. 3), and are a measure of variation that includes both size and shape. Thus, low values for STET such as 1.34, based on the 70 comparisons of the Sterkfontein specimens STS 71 and TM 1511 (Table 2, Fig. 3), are reflected in a linear array of the measurements for the two specimens that shows little scatter around a straight line. The higher STET value of 2.19 found in the comparison of the habiline crania OH 24 and ER 1813 (Table 2, Fig. 3) reflects a more scattered distribution.

The mean STET value for the habilines is 2.78 ($\sigma = 0.75$). This is larger than the mean STET value for Swartkrans (2.46, $\sigma = 0.66$), but smaller than STET for Sterkfontein (3.14, $\sigma = 1.47$). The observed range of STET values for the habilines is virtually identical to the Swartkrans range, and both of these are markedly less than the Sterkfontein range (Fig. 2).

The three STET distributions are compared in Fig. 2. Both the habiline sample and the more variable Sterkfontein sample have a large male and a small female whose visual comparisons are quite similar: ER 1470 and ER 1813 (Fig. 1), and STW 505 and STS 5 (Fig. 4). The paired comparisons of these two are quite similar to each other (Fig. 5).

Discussion

Interpreting the taxonomy of the habiline sample was a problem from the beginning, even when the sample was limited to the Olduvai remains, and did not improve when fairly complete specimens such as ER 1470 and 1813 were discovered. The fact remained that even these specimens were given diverse individual affinities, ER 1470 was affiliated with australopithecines (Walker, 1976), or described as a habiline (Leakey and Walker, 1980) and ER 1813 was affiliated with the *Homo* lineage (Johanson and White, 1979), or described as an australopithecine (Leakey and Walker, 1980). A specimen very similar to ER 1470, ER 3732, was described as *H. ergaster* (Schwartz and Tattersall, 2000).

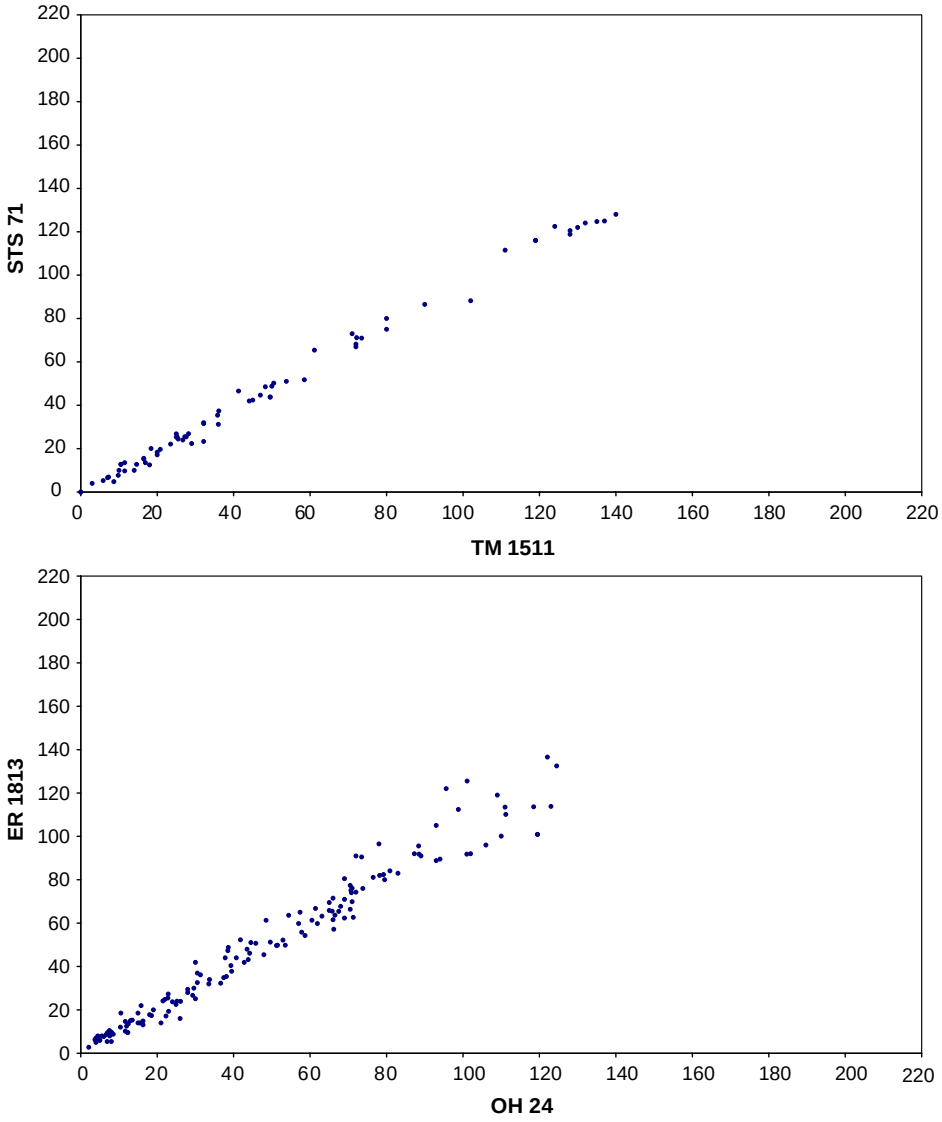

Fig. 3. Comparison of a larger STET (ER 1813 and OH 24 with STET = 2.19 ($n = 147$), below) and a smaller one (STS 71 and TM 1511 with STET = 1.34 ($n = 70$), above). The different STET values reflect difference in dispersion from the best fitting straight line.

When quantitative assessment was finally brought into the picture, authors such as Stringer (1986) using the magnitude of the coefficients of variation (CV), and Lieberman et al. (1988, 1996) using pairwise analyses, continued to argue that something was wrong: habiline variation was excessive and greater than the

Fig. 4. Variation in Sterkfontein *Australopithecus africanus* is seen in this comparison of STW 505 (left) and STS 5, both casts. Note that STW 505 lacks the posterior portion of its cranial vault; the specimen was longer than the comparison suggests, and this is reflected in the estimated cranial capacity of 600 cc (Hawks and Wolpoff, 1999), compared with the 485 cc determination for STS 5. As in the comparison of the Turkana habilines (Fig. 4), the larger specimen has a flatter face with a more anterior positioned bas for the zygomatic process of the maxilla. These comparisons are compatible with Rak's (1985) observation that larger specimens at Sterkfontein are more *Paranthropus*-like. Photograph courtesy of J. Hawks.

expectation for a single species. Kramer et al. (1995) were critical of this interpretation, noting that a more systematic analysis of pairwise comparisons showed that habiline differences were not unexpected in a species with gorilla-like variation. They regarded their results as equivocal, however, for two reasons: they questioned whether gorilla variation was the best model for comparison, and because their comparison of habiline CVs for the measurements they studied revealed a different pattern than the CVs of other large-bodied hominoids. From this they concluded that a rejection of the single species hypothesis was reasonable. However, a different interpretation is that gorillas *may not be dimorphic enough* to be a good model. Living large-sized hominoids are rare and there is reason to believe that some large-bodied fossil hominoids were even more variable and had a greater magnitude of sexual dimorphism than gorillas (Kelley, 1993). Moreover, as Miller (1991) noted earlier, the habiline CVs are based on sample sizes so small that they are unreliable indicators of the true variation, with a bizarrely broad 95-percentile range for each coefficient. Whether because of sample size or because of the true underlying distribution, these approaches could not disprove the null hypothesis.

We chose an alternative approach for comparisons with penecontemporary or slightly earlier australopithecines, the taxa most closely related to the habilines (Wolpoff, 1999; Wood and Collard, 1999b). Using STET cannot improve the sample size, but it does provide a way to maximize the information in the sample by examining all of the possible comparisons based on as many data as are preserved. For us, convincing disproof of the null hypothesis would be the demonstration that

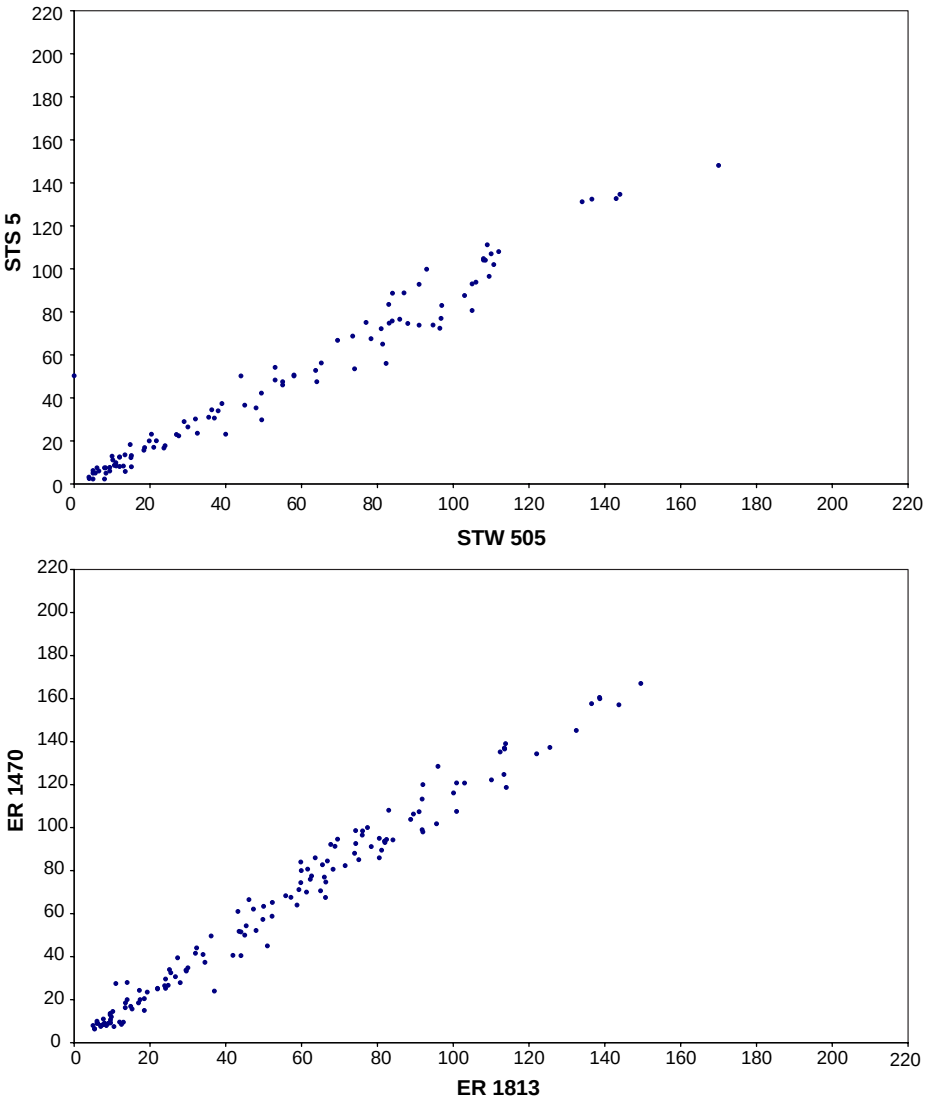

Fig. 5. Plots for the specimens illustrated in Figs. 1 and 4, a habiline male and female (above) with a STET of 1.72, and a male and female from Sterkfontein with a STET of 3.25. In each case the measurements shared by both specimens are plotted against each other. The habiline pair, ER 1813 and 1470, has the lowest STET value within the habiline sample.

the habiline specimens differ more from each other than either the Sterkfontein or the Swartkrans australopithecine specimens do, as reflected in the STET statistics for each sample. However, this is not the case. The magnitude and pattern of dispersions in the habiline comparisons are quite like those found in each of the australopithecine samples, and there are no grounds for rejecting the null hypothesis.

Conclusion

Here we present a new sampling statistic designed to address the null hypothesis for a controversial hominid sample, and lay out the procedure for its use. STET describes size and shape variation with the estimates of dispersion around the slopes of linear regressions through pairs of specimens. It has the advantages of maximizing fossil data with the comparison of specimens of uncertain or unknown sex and allows bispecimen analyses based on both different numbers of measurements and different measurement sets to be combined.

We use STET to test the null hypothesis for the habiline sample, and calibrated the results by comparison with samples of known taxonomic affiliation from two different australopithecine sites. The question is whether some of the habiline variation can reasonably be ascribed to a phylogenetic basis: are we required to accept the interpretation that there are different human species in the fossil habiline sample? To address this, we asked whether the habilines are more different from one another than specimens within other single-species samples of australopithecines from single sites. Because our test statistic, STET, is the dispersion around the paired specimen slopes, it reflects variation in both size and shape and so there is reason to believe STET has the potential to address this issue.

Our STET comparisons do not refute the null hypothesis for the habiline sample, a sample some have argued to contain two species, most recently "*A. habilis*" and "*A. rudolfensis*." These results reinforce several others in concluding that a null hypothesis – taxonomy does not underlie the size and shape variation of the sample – cannot be disproved. The paired specimen comparisons within the habilines we examined show a dispersion similar to the Swartkrans comparisons, and less dispersion than the Sterkfontein comparisons. These are two relevant samples of single species represented by a small number of crania in closely related taxa from sites with similar issues of taphonomic and temporal sampling.

One might still argue that there are two habiline species, with specimens equally similar or more similar to each other than the australopithecine specimens at Sterkfontein; so similar, in other words, that they cannot be clearly distinguished. Perhaps so, no matter what statistics are used the small sample sizes, site taphonomy, and temporal variation in the samples must leave the door open for the possibility of such an interpretation. But if there is no way to distinguish species differences in a mixed sample, the null hypothesis cannot be disproved for it, and at the moment this is where things stand.

Acknowledgments

We thank the curators of the fossil specimens from Koobi Fora and Olduvai Gorge at the National Museums of Kenya, and of Swartkrans and Sterkfontein specimens at the Transvaal Museum and the Anatomy Department at the University of the Witwatersrand, for permission to study the specimens in their care.

References

Aiello, L.C., Collard, M., Thackeray, J.F., Wood, B.A., 2000. Assessing exact randomization-based methods for determining the taxonomic significance of variability in the human fossil record. S. Afr. J. Sci. 96, 179–183.

Alexeyev, V.P., 1986. The Origin of the Human Race. Progress Publishers, Moscow.

Chamberlain, A.T., 1989. Variations within *Homo habilis*. In: Giacobini, G. (Ed.), Hominidae. Proceedings of the Second International Congress of Human Paleoanthropology. Editoriale Jaca Books, Milan, pp. 175–181.

Chamberlain, T.C., 1965. The method of multiple working hypotheses. J. Geol. 5, 837–848.

Hawks, J.D., Wolpoff, M.H., 1999. Endocranial capacity of early hominids. Science 283, 9.

Jiggins, C.D., Naisbit, R.E., Coe, R.L., Mallet, J., 2001. Reproductive isolation caused by colour pattern mimicry. Nature 411, 302–305.

Johanson, D.C., White, T.D., 1979. A systematic assessment of early African hominids. Science 203, 321–330.

Kelley, J., 1993. Taxonomic implications of sexual dimorphism in Lufengpithecus. In: Kimbel, W.H., Martin, L.B. (Eds.), Species, Species Concepts, and Primate Evolution. Plenum, New York, pp. 429–458.

Kimbel, W.H., Martin, L.B. (Eds.), 1993. Species, Species Concepts, and Primate Evolution. Plenum, New York.

Kramer, A., Donnelly, S.M., Kidder, J.H., Ousley, S.D., Olah, S.M., 1995. Craniometric variation in large-bodied hominoids: testing the single-species hypothesis for *Homo habilis*. J. Hum. Evol. 29, 443–462.

Leakey, R.E.F., Walker, A.C., 1980. On the status of *Australopithecus afarensis*. Science 207, 1103.

Leakey, L.S.B., Tobias, P.V., Napier, J.R., 1964. A new species of the genus *Homo* from Olduvai Gorge. Nature 202, 7–9.

Lieberman, D.E., Pilbeam, D.R., Wood, B.A., 1988. A probabilistic approach to the problem of sexual dimorphism in *Homo habilis*: a comparison of KNM-ER 1470 and KNM-ER 1813. J. Hum. Evol. 17, 503–511.

Lieberman, D.E., Wood, B.A., Pilbeam, D.R., 1996. Homoplasy and early *Homo*: an analysis of the evolutionary relationships of the *H. habilis sensu stricto* and *H. rudolfensis*. J. Hum. Evol. 30, 97–120.

Lovejoy, C.O., 1979. Contemporary methodological approaches to individual primate fossil analysis. In: Morbeck, M.E., Preuschoft, H., Gomberg, N. (Eds.), Environment, Behavior, and Morphology: Dynamic Interactions in Primates. Gustav Fischer, New York, pp. 229–243.

Mayr, E., 1963. Animal Species and Evolution. Belknap Press of Harvard University Press, Cambridge.

Milius, S., 2001. Alarming butterflies and go-getter fish: overlooked ways to invent new species. Sci. News 160, 42–45.

Miller, J.M.A., 1991. Does brain size variability provide evidence of multiple species in *Homo habilis*? Am. J. Phys. Anthropol. 84, 385–398.

Miller, J.M.A., 2000. Craniofacial variation in *Homo habilis*: an analysis of the evidence for multiple species. Am. J. Phys. Anthropol. 112, 103–128.

Rak, Y., 1985. Sexual dimorphism, ontogeny, and the beginning of differentiation of the robust australopithecine clade. In: Tobias, P.V. (Ed.), Hominid Evolution: Past, Present, and Future. Proceedings of the Taung Diamond Jubilee International Symposium. Alan R. Liss, New York, pp. 233–237.

Robinson, J.T., 1963. Adaptive radiation in the australopithecines and the origin of man. In: Howell, F.C., Bourlière, F. (Eds.), African Ecology and Human Evolution, vol. 36. Viking Fund Publication in Anthropology, pp. 385–416.

Robinson, J.T., 1965. *Homo 'habilis'* and the australopithecines. Nature 205, 121–124.

Robinson, J.T., 1966. The distinctiveness of *Homo habilis*. Nature 209, 957–960.

Schwartz, J.H., 2000. Taxonomy of the Dmanisi crania. Science 289, 55–56.

Schwartz, J.H., Tattersall, I., 2000. The human chin revisited: what is it and who has it? J. Hum. Evol. 38, 367–409.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, second ed. W.H. Freeman, San Francisco.

Stringer, C.B., 1986. The credibility of *Homo habilis*. In: Wood, B.A., Martin, L.B., Andrews, P.J. (Eds.), Major Topics in Primate and Human Evolution. Cambridge University Press, Cambridge, pp. 266–294.

Tattersall, I., 1992. The many faces of *Homo habilis*. Evol. Anthropol. 1, 33–37.

Tattersall, I., 1999. The abuse of adaptation. Evol. Anthropol. 7, 115–116.
Tattersall, I., 2000. Paleoanthropology: the last half-century. Evol. Anthropol. 9, 2–16.
Thackeray, J.F., Helbig, J., Moss, S., 1995. Quantifying morphological variability within extant mammalian species. Palaeontol. Afr. 31, 23–25.
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., Whiting, M.J., 1997. Probabilities of conspecificity: application of a morphometric technique to modern taxa and fossil specimens attributed to *Australopithecus* and *Homo*. S. Afr. J. Sci. 93, 195–196.
Thackeray, J.F., Mdaka, S., Navsa, N., Moshau, R., Singo, S., 2000. Morphometric analyses of conspecific males and females: an exploratory study of extant primate and extinct hominid taxa. S. Afr. J. Sci. 96, 534–536.
Tobias, P.V., 1991. *Homo habilis*: Skulls, Endocasts, and Teeth. Olduvai Gorge IV. Cambridge University Press, New York.
Tobias, P.V., 2003. Encore Olduvai. Science 299, 1193–1194.
Tobias, P.V., von Koenigswald, G.H.R., 1964. A comparison between the Olduvai hominines and those of Java, and some implications for hominid phylogeny. Nature 204, 515–518.
Walker, A.C., 1976. Remains attributable to *Australopithecus* in the East Rudolf succession. In: Coppens, Y., Howell, F.C., Isaac, G.L., Leakey, R.E.F. (Eds.), Earliest Man and Environments in the Lake Rudolf Basin. University of Chicago Press, Chicago, pp. 484–489.
Wolpoff, M.H., 1999. Paleoanthropology, second ed. McGraw-Hill, New York.
Wolpoff, M.H., Lee, S.-H., 2001. The late Pleistocene human species of Israel. Bull. Mém. Soc. Anthropol. Paris 13, 291–310.
Wood, B.A., 1987. Who is the real *Homo habilis*? Nature 327, 187–188.
Wood, B.A., 1991. Hominid Cranial Remains. Koobi Fora Research Project 4. Oxford University Press, Oxford.
Wood, B.A., 1999. '*Homo rudolfensis*' Alexeev, 1986 – fact or phantom? J. Hum. Evol. 36, 115–118.
Wood, B.A., Collard, M., 1999a. The changing face of genus *Homo*. Evol. Anthropol. 8, 195–207.
Wood, B.A., Collard, M., 1999b. The human genus. Science 284, 65–71.