AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 90:381–384 (1993)

Notes and Comments

Reply to Dr. Foote

Milford H. Wolpoff

Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109

For the 1991 AAPA meetings in Milwaukee, I organized a symposium with Dr. A. Mann entitled "A New Definition of Neandertal." The symposium was featured in a *Science* review article on the meetings (Culotta, 1991), which contrasted two conflicting views of the place of Neandertals in human evolution. One was the view, attributed to Y. Rak, that "the modern and the Neandertal specimens from the Near East are so different morphologically that they couldn't possibly fit into a single species." Culotta summarized my own, opposing position on the status of Neandertal populations in the following paragraph:

At the symposium Wolpoff and Baruch Arensburg of Tel Aviv argued that bones from the Israeli sites are not from two separate species but from one population that interbred. They acknowledge that this population was quite variable anatomically, but they attribute such variability to the flow of genes from Africa mingling with Neandertal-like European genomes. Furthermore, Wolpoff claimed, *the degree of variability in anatomical form among the specimens wasn't all that great—no greater than in today's Detroit, with its population of European Americans, African Americans, Amerindians, and Asians.*" As Wolpoff put it succinctly: "The separate species concept won't fly."

The Culotta report was reasonably accurate. The italicized sentence was not precisely what I said, but it was a reasonable inference from something I did say—namely, that mixed populations like Detroit provide an example of population admixture that is similar to the variation in Middle Paleolithic Levantine samples. When van Vark and Bilsborough (1991) wrote to *Science* criticizing Culotta's paraphrase of my own views (which they had not heard and did not bother to enquire about), I accordingly decided to defend it against their criticisms.

Van Vark and Bilsborough set out to test the italicized sentence using the D^2 statistic. They calculated an average D^2 distance for all possible pairwise comparisons of 19 cranial variables in a worldwide sample (taken from Howells, 1989) of 2,216 modern males and females. They compared this figure with a D^2 calculated from the same measurements for four Middle Paleolithic Levant crania (Amud, Qafzeh 6, Skhul 5, Tabun 1), and concluded that the worldwide sample is "appreciably less diverse" than the four ancient Levantines.

My response (Wolpoff, 1992) was prompted by van Vark's (1974) own estimation of the value of multivariate procedures in population comparisons (pp. 67–68):

> "The more data that are missing and the smaller the samples, the more the value of all these methods decreases . . . these methods can have their uses in an investigation where individuals originating from only one population are involved, such as, for example, sex diagnosis . . . This might theoretically also hold for the comparative population investigation if the comparison concerns exclusively known recent populations. However, for all intents and purposes, they prove to be useless for the comparison of prehistoric populations."

If in spite of these cautions, comparisons are sought, to validly compare D^2 statistics between any samples it must be assumed "that the hominid populations being compared have the same variance-covariance matrix as that computed for the recent population" (van Vark, 1984, p. 336). However, knowledge of the variance-covariance matrix for the four Levantine crania is a requirement unlikely to be met. Obviously the matrix cannot be directly calculated from a

Received October 16, 1992; accepted November 14, 1992.

sample of four. Moreover, the matrix cannot be estimated from other samples; the population affinities of these specimens are not only unknown, but cannot be assumed (since this is the very aspect of their relationship that is to be tested).

The D^2 comparison raises an additional problem. To determine the average pairwise D^2 for 2,216 crania, van Vark and Bilsborough (1991) made 2,454,220 different matches of modern crania and compared them with six different matches of the Levant specimens. This vast disparity in sample sizes depresses the magnitude of the grand variation measured by the pairwise D^2 average for the larger sample, and this is even further decreased by the naturally lower magnitude of variation within each of the intrapopulation pairs (the Howells's woldwide data set spans only 28 populations) incorporated in the average. Whatever their relationship, the fossil sample is composed of one individual from each of four biological populations.

I was also concerned about missing data and their estimation. Since no set of 19 (of Howells's) measurements is preserved for all four Levant crania they used, the authors must have estimated some data points. Estimating missing data requires knowledge of the correlation matrix—knowledge that is impossible to obtain for the same reason that the variance-covariance matrix is unknown. And there are other problems with the fossil sample. There are nine additional crania from the Middle Paleolithic of the Levant that were not included, some of which (e.g., Skhul 4 and Qafzeh 9) are almost as complete as those used. Van Vark and Bilsborough's (1991) disregard of these specimens is inexplicable; and leaving more incomplete specimens out of the picture is ill advised, because it discards valuable information about variation, which we cannot afford to disregard given the restricted nature of the fossil sample.

The matter at issue in this debate is whether phenetic evidence shows there are two human species in the Levantine Middle Paleolithic. This brief history is intended to make two points: 1) we cannot resolve this issue by relying on statistics of variation that require knowledge of the variance-covariance matrix of the Levant sample, and 2) any resolution must incorporate all of the specimens so as to make maximum use of the limited fossil data set. These two points could not be made in my subsequent reply letter to *Science* that is the focus of M. Foote's comments—there was simply not the space—but they are explicit in a complete reading of the three articles that comprise the discussion.

To make best use of the Levantine fossil sample, and to avoid the seemingly insurmountable problems in the D^2 comparisons, I used the sample ranges as a much simpler and less assumption-ridden measure of variation in the individual measurements. (All parties accept the disclaimer that this debate is over the *magnitude* of variation in individuals' features, and not their *pattern* of variation.) The sample range is not problem free, as the present discussion shows, but there is much to be said for using a technique with a single problem having known effects on the conclusion, as compared with the morass of difficulties involved in applying an average D^2 comparison. The most outstanding difficulty raised by using observed range as a measure of sample variation is its dependence on sample size—a dependence that I, and I am sure every reader of this journal, was aware of before reading Foote's comment. This problem *does not* necessarily make the results "misleading" or warrant Foote's insistence that comparing ranges in samples of different sizes "should be categorically avoided." It *does* require discretion in interpreting the consequences, and it makes some potential results of the comparisons uninterpretable (see below).

I examined the variation question by comparing ranges of variation for 14 measurments that could be taken in both the Levant sample of 13 individuals (or less, depending on the measurement) and a published 18th-century London cemetery sample of 388. These comparisons were done twice, once as a direct comparison of observed ranges and again as a comparison omitting the upper and lower extremes from the London cemetery range (to eliminate the influence of outliers). In both cases the London sample was more variable, for some measurements dramatically so. I concluded

that "The amount of variation in measurements from the Middle Paleolithic people from the Levant appears to be less than in a modern population." If so, then, this comparison provides no support for the contention attributed to Rak, that the two supposedly different groups comprising the Levantine Middle Paleolithic—Neandertals and "anatomically modern *Homo sapiens*" according to some authors (cf. Stringer, 1988)—are too morphologically different to fit into a single species.

Foote objects to these comparisons. Proving the obvious, he shows that larger normally distributed samples will always have larger ranges, and that omitting the outliers will not alter this relation as long as the outliers are predictable extremes of the normal samples and not biological oddities. He calculates the ratio of expected ranges in normally distributed samples of $n = 13$ and $n = 388$ to demonstrate that with or without the outliers in the normal distribution, the larger sample will have the wider range by a predictable factor.

The difference in sample sizes is clearly a concern in comparing observed ranges, although it could be pointed out that comparing samples of 13 and 388 probably creates fewer, and certainly more predictable problems, than the comparison of the average of six D^2 matches and 2,454,220 D^2 matches. Yet, there are several reasons why I was (and remain) unwilling to make the assumptions that underlie his demonstration, reasons that undermine its relevance to the problems of making comparisons between real biological samples. Two of these are the unknown pattern of internal variation in the Levant sample (i.e., the biological group from which the sampling population was drawn), and the potential difference between statistical and biological outliers.

The Middle Paleolithic Levant sample is not a sample of a single biological population, even if its constituent members are in the same species (likely, in my view) or in the same race (unlikely, in my view). The sample is too small and incomplete for any internal determination of its underlying distribution, and to assume normality is to assume the answer to the question addressed, i.e., whether this sample shows too much variation to represent a single species. Foote asserts:

> Wolpoff's claim, that "the amount of variation in measurements from the middle Paleolithic . . . appears to be less than in a modern population" is unjustified, unless 1) his emphasis is on the appearance rather than on the existence of a true difference in variation, or 2) his emphasis is on the samples, rather than on the underlying populations these samples represent.

In fact, I meant what I said. The results do not provide as much information about the underlying distribution as we would like to know (and cannot without making further assumptions), and my emphasis is indeed on the samples. Foote misunderstands the question, which is not whether one sample has less variation that the other, but whether the Levant sample is too variable to be a single species. I am tempted to say that in this case he has provided the right answer, but to the wrong question.

My second problem with Foote's comment has to do with the fact that real biological samples are not mathematically ideal. I am not willing to assume that the outliers of the larger London sample were necessarily artifacts of random sampling. Biological oddities are artifacts of biological processes, not of normally distributed bell-shaped idealized samples, and omitting biological outliers has, potentially, very different consequences from omitting outliers on a random curve. Foote's demonstration only has relevance if we are willing to assume normality, and not a particularly biological basis, for the outliers in the London cemetery sample.

In the end, if the range comparison had revealed the Levantines to be more variable, it would have been a powerful argument in support of Rak's ideas about its taxonomy because that result is in the opposite direction of the bias that might be created by different sample sizes in the comparison. The actuality of a lower magnitudes of variation in the Levant sample can only be interpreted in the context of the Rak proposal; they do not necessarily support a single-species interpretation because they could be a consequence of sample size bias. The true situation, however, is that we cannot be certain what the magnitude of the bias is, be-

cause different biological and statistical entities are sampled in the comparison.

Making the best of a bad sampling situation is something that we who study fossils are invariably obliged to do. I would never have tried to solve the question of human species in the Levant by comparing phenetic variation between the Levantine and modern samples because they are samples of different things. This was Rak's approach, which set off this cascade of comments but provided no real insight into the solution of the issue. One thing I have been doing (with A. Kramer and T. Crummett) to try to resolve this issue is examining cluster and PAUP analyses of nonmetric traits in the Levantine sample, to attempt to refute the hypothesis that there are distinguishable "Neandertal" and "non-Neandertal" groupings. Preliminary results, presented by Kramer at the 3rd (1992) International Congress of Human Paleontology, do not support the hypothesis. There are ways of proceeding with paleontological research without assuming the results.

ACKNOWLEDGMENTS

I am deeply indebted to E. Giles for the insight he provided in understanding the pitfalls of the van Vark and Bilsborough analysis.

LITERATURE CITED

Culotta E (1991) Pulling Neandertals back into our family tree. Science *252:*376.

Howells WW (1989) Skull shapes and the map: Craniometric analyses in the dispersion of modern Homo. Papers of the Peabody Museum of Archaeology and Ethnology *79:*1–189.

Stringer CB (1988) The dates of Eden. Nature *331:*565–566.

van Vark GN (1974) The investigation of human cremated skeletal material by multivariate statistical methods. I. Methodology. Ossa *1:*63–95.

van Vark GN (1984) On the determination of hominid affinities. In GN van Vark and WW Howells (eds).: Multivariate Statistical Methods in Physical Anthropology. Dordrecht: D. Reidel, pp. 323–349.

van Vark GN, and Bilsborough A (1991) Shaking the family tree. Science *253:*834.

Wolpoff MH (1992) Levantines and Londoners. Science *255:*142.