Note added June 2009: for evidence that 'conditional entropy' is useless in distinguishing linguistic from nonlinguistic systems in general, see now http://www.safarmer.com/more.on.Rao.pdf (live link).

A Refutation of the Claimed Refutation of the Nonlinguistic Nature of Indus Symbols: Invented Data Sets in the Statistical Paper of Rao et al. (*Science*, 2009)

Steve Farmer[*], Richard Sproat[†*], and Michael Witzel[‡*]

The paper in *Science* on 23 April by Rao et al. [1] was written in response to an article that the three of us published five years ago that has led to heated polemics over India's oldest urban society [2, 3]. That paper argued that the short chains of symbols on Indus artifacts were not part of a writing system but of a simple nonlinguistic sign system of a type common in the ancient world. The vision of a nonliterate Indus society has solved a number of puzzles and now has many adherents, but it has also awakened resistance from Indian nationalists and researchers whose entire careers have been linked to the Indus-script thesis, one of whom is listed as a coauthor of this study. The question of the nature of India's first society in large part revolves around this issue, and discussion will continue at a conference in Japan this May.

There are many oddities in Rao et al. that undermine this newest attempt to back the traditional view of the symbols. The most obvious comes in their claim that the degree of order (or conditional entropy) in Indus inscriptions supposedly differs from that found in nonlinguistic systems. On pages 2-3 of the online Supplemental Information section of their paper, we find to our surprise — in contradistinction to what they say in the paper itself — that this claim is *not* based on a comparison of Indus signs with real-world nonlinguistic systems, but with two wholly artificial systems invented by the authors, one consisting of 200,000 randomly ordered signs and another of 200,000 fully ordered signs, that they spuriously claim represent the structures of all real-world nonlinguistic sign systems (which they refer to as Type 1 and Type 2). When they compare the Indus system with these artificial sets of random and ordered signs, they not unexpectedly find that the degree of order in the Indus system falls somewhere in between. It is important to realize that all their demonstration shows is that the Indus sign system has some kind of rough structure, which has been known since the 1920s. Similar results could be expected if they compared their artificial sign sets to any man-made symbol system, linguistic or nonlinguistic. Our paper in fact made much of this point and also gave examples of striking statistical overlaps between real-world (not invented) nonlinguistic and linguistic systems and argued that it is not possible to distinguish the two using statistical measures alone. Hence our paper made use of abundant archaeological as well as linguistic evidence in making our case [2].

Conditional entropy is not and has never before been claimed to be a statistical measure of whether or not a sign system is linguistic or nonlinguistic. Rao *et al*. only make it appear to be relevant to that end (as we find *only* in their online Supplemental Information section, but not in their paper itself) by inventing fictional sets of nonlinguistic systems that correspond (*pace* their claims) to nothing remotely resembling any ancient symbol system. If the paper had been properly peer reviewed it would not have been published.

---

[*] The Cultural Modeling Research Group, Palo Alto, California, saf@safarmer.com

[†] Center for Spoken Language Understanding, Division of Biomedical Computer Science, Oregon Health and Science University, Portland, Oregon, rws@xoba.com

[‡] Department of Sanskrit and Indian Studies, Harvard University, Cambridge, Massachusetts, witzel@fas.harvard.edu

The authors also emphasize what they perceive as a similarity between the Indus signs and Old Tamil, which they suggest supports the view espoused by Parpola, Mahadevan (listed as one of the authors), and Tamil/Dravidian nationalists in general that the Indus peoples spoke and wrote a Dravidian language. There are many problems here, including the fact that the first attestation of Old Tamil came nearly two thousand years after the Indus civilization disappeared, philological evidence that the Indus region was not Dravidian speaking in early historical times [4], and evidence based on Indus sign orders discussed in our paper that cannot be reconciled with purely suffixing languages like Old Tamil [2].

Finally on this topic, we can note that the authors compare the Indus signs with only four languages: English, Sumerian, Sanskrit and Old Tamil. Any claim for supposed similarity between Indus inscriptions and any Dravidian language would need to be based on comparisons with far more languages than this. Moreover, if they had compared Indus signs with any real-world (and not invented) sets of nonlinguistic symbols, we would expect as well that they would find similarities between the kind of order found in Indus inscriptions and those found in these systems as well. In our own paper, in fact, we showed that striking overlaps exist between Indus sign frequencies, frequencies in medieval heraldic signs, and in a variety of natural languages [2]. It can be demonstrated that many statistical overlaps exist in symbol systems in general, not just in those that encode speech.

The implausibility of the view that the so-called Indus script was true writing is suggested in many ways that do not require sophisticated analyses. The simplest argument is the best: the sheer brevity of the inscriptions. We possess thousands of inscribed Indus objects on a wide range of materials. The average inscription is 4-5 symbols long and the longest, found on a highly anomalous piece, carries 17. Before our paper, the lack of real texts was explained away by invoking the purely speculative image of lost perishable manuscripts. The speculation was spurious: we know of hundreds of literate societies, but not of one that wrote long texts on perishable materials but failed to do so as well on durable goods. It is interesting that simple arguments like this have been ignored by defenders of the traditional view, who often hold that view for reasons that have nothing to do with science, while questions involving the symbols are obfuscated with complex statistical arguments that when you read the fine print (and that not in the paper itself) turn out to depend on invented data.

1. Rajesh P.N. Rao et al., "Entropic Evidence for Linguistic Structure in the Indus Script," *Science* 23 April 2009.
2. S. Farmer, R. Sproat, and M. Witzel, "Collapse of the Indus Script Thesis," *EJVS* 11, 19, December 2004. Reprint at http://www.safarmer.com/fsw2.pdf
3. A. Lawler, "The Indus Script — Write or Wrong," *Science* 17 December 2004.
4. M. Witzel, "Substrate Languages in Old Indo-Aryan (Rgvedic, Middle, and Late Vedic)," EJVS 5, 1 August 1999.