How WEIRD are WALS languages?

Östen Dahl
Stockholm University

”WEIRD societies”

- Henrich et al. (2010, 61) note that behavioural scientists tend to make ”broad claims about human psychology and behavior” based on samples from ”Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies” at the same time as, in their opinion, these societies ”are among the least representative populations one could find for generalizing about humans.”

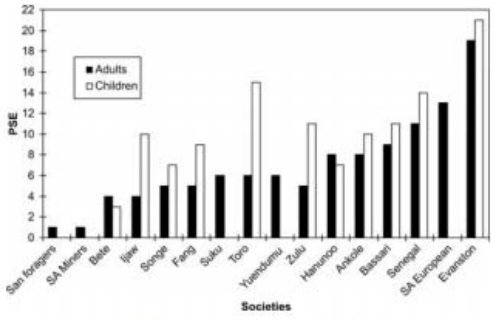


Figure 2. Müller-Lyer results for Segall et al.'s (1966) cross-cultural project. PSE (point of subjective equality) is the percentage that segment a must be longer than b before subjects perceived the segments as equal in length. Children were sampled in the 5-to-11 age range.

BEHAVIORAL AND BRAIN SCIENCES (2010), Page 1 of 75
doi:10.1017/S0140525X0999152X

The weirdest people in the world?

Joseph Henrich
Department of Psychology and Department of Economics, University of British Columbia, Vancouver V6T 1Z4, Canada
joseph.henrich@gmail.com
http://www.psych.ubc.ca/~henrich/home.html

Steven J. Heine
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
heine@psych.ubc.ca

Ara Norenzayan
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
ara@psych.ubc.ca

Abstract: Behavioral scientists routinely publish broad claims about human psychology and behavior in the world's top journals based on samples drawn entirely from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. Researchers – often implicitly – assume that either there is little variation across human populations, or that these "standard subjects" are as representative of the species as any other population. Are these assumptions justified? Here, our review of the comparative database from across the behavioral sciences suggests both that there is substantial variability in experimental results across populations and that WEIRD subjects are particularly unusual compared with the rest of the species – frequent outliers. The domains reviewed include visual perception, fairness, cooperation, spatial reasoning, categorization and inferential induction, moral reasoning, reasoning styles, self-concepts and related motivations, and the heritability of IQ. The findings suggest that members of WEIRD societies, including young children, are among the least representative populations one could find for generalizing about humans. Many of these findings involve domains that are associated with fundamental aspects of psychology, motivation, and behavior – hence, there are no obvious *a priori* grounds for claiming that a particular behavioral phenomenon is universal based on sampling from a single subpopulation. Overall, these empirical patterns suggests that we need to be less cavalier in addressing questions of *human* nature on the basis of data drawn from this particularly thin, and rather unusual, slice of humanity. We close by proposing ways to structurally re-organize the behavioral sciences to best tackle these challenges.

Keywords: behavioral economics; cross-cultural research; cultural psychology; culture; evolutionary psychology; experiments; external validity; generalizability; human universals; population variability

WEIRD languages?

- Majid and Levinson (2010, 103) say that ”WEIRD languages have misled us, too”, arguing that linguists have ”projected assumptions based on English and familiar languages onto the rest”.
- It is certainly possible to agree with this statement. I think, however, that we may also be led a bit astray by the catchy acronym WEIRD in that the adjectives it encapsulates are not necessarily the most adequate for characterizing the biases that have influenced linguistics.

WEIRD languages have misled us, too

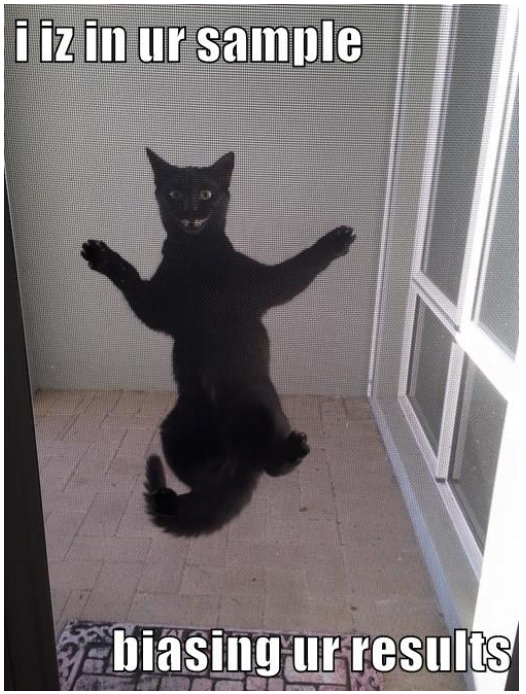
doi:10.1017/S0140525X1000018X

Asifa Majid and Stephen C. Levinson
Max Planck Institute for Psycholinguistics, Nijmegen 6500AH, The Netherlands.
asifa.majid@mpi.nl **http://www.mpi.nl/people/majid-asifa**
stephen.levinson@mpi.nl
http://www.mpi.nl/people/levinson-stephen

Abstract: The linguistic and cognitive sciences have severely underestimated the degree of linguistic diversity in the world. Part of the reason for this is that we have projected assumptions based on English and familiar languages onto the rest. We focus on some distortions this has introduced, especially in the study of semantics.

“Literate, Official, and with Lots of users”.

- It is true that Western (mainly European) languages have been in the focus for a long time; however, even after the Eurocentric bias has started to lose its grip on the choice of languages to be studied, there remains a bias that can be summed up in the acronym “LOL” for
- Literate
- Official, and with
- Lots of users



Question to be answered

How WEIRD is current typological research?

Bias in typology

- Even in typological works, the bias is visible.
- To a certain extent, it is probably unavoidable, given the restricted availability of information on smaller languages.
- The interesting question is how much the bias in the choice of languages influences the results.

An exclusive club

- It turns out that there is a very restricted set of LOL languages which are overrepresented in almost any sample.
- Let us look at that set!

Literate

- There are 128 language versions of Wikipedia with more than 10,000 articles

46	Norwegian (Nynorsk)	norsk nynorsk	**120,514**	279,078
47	Volapük	Volapük	**120,107**	248,920
48	Latin	Latina	**115,733**	217,741
49	Simple English	Simple English	**112,513**	364,978
50	Greek	Ελληνικά	**106,262**	317,825
51	Hindi	हिन्दी	**101,608**	514,952

10,000+ articles [edit]

№	Language	Wiki	Articles	Total
52	Azerbaijani	azərbaycanca	**95,323**	241,93
53	Thai	ไทย	**94,448**	548,75
54	Georgian	ქართული	**93,689**	261,55
55	Occitan	occitan	**88,326**	142,09
56	Belarusian	беларуская	**82,619**	202,42
57	Chechen	нохчийн	**82,513**	97,24
58	Macedonian	македонски	**81,783**	1,081,11
59	Malagasy	Malagasy	**79,283**	212,81
60	Newari	नेपाल भाषा	**72,361**	195,26
61	Urdu	اردو	**68,068**	314,98
62	Tatar	татарча/tatarça	**66,742**	151,46
63	Tamil	தமிழ்	**66,448**	204,23
64	Piedmontese	Piemontèis	**63,768**	94,16
65	Welsh	Cymraeg	**63,530**	149,00

Official

- According to Wikipedia, there are 191 languages which are ”official languages of sovereign countries”

Official languages of sovereign countries [edit]

A [edit]

Afar:

- Djibouti (with Arabic, French, Somali)

Afrikaans:

- South Africa (with English, Ndebele, Northern Sotho, Sotho, Swati, Tsonga, Tswana, Venda, Xhosa, Zulu)[1]

Aja-Gbe:

- Benin (a national language along with Anii, Bariba, Biali, Boko, Dendi, Fon-Gbe, Foodo, Fula, Gen-Gbe, Lukpa, Mbelime, Nateni, Tammari, Waama, Waci-Gbe, Yobe, Yom, Xwela-Gbe, Yoruba, the official languages is French)

Lots of users

- According to Ethnologue, there are 394 languages with at least one million speakers

Not the same thing as ”users” but it is hard to find reliable statistics about that. Cheat: Swahili added to the list

Table 2. Distribution of world languages by number of first-language speakers

Population range	Living languages			Number of speakers		
	Count	*Percent*	*Cumulative*	*Total*	*Percent*	*Cumulative*
100,000,000 to 999,999,999	8	0.1	0.1%	2,529,403,578	40.20547	40.20547%
10,000,000 to 99,999,999	82	1.2	1.3%	2,480,078,977	39.42144	79.62691%
1,000,000 to 9,999,999	304	4.3	5.5%	915,659,448	14.55462	94.18154%
100,000 to 999,999	943	13.3	18.8%	296,136,843	4.70717	98.88870%

The unique 57

- But only 57 languages fulfill all three criteria

Afrikaans	Danish	Halh Mongolian	Latvian	Russian	Tamil
Amharic	Dutch	Hebrew	Lithuanian	Serbian	Telugu
Armenian	Eastern Panjabi	Hindi	Malagasy	Sinhala	Thai
Belarusan	English	Hungarian	Mandarin Chinese	Slovak	Turkish
Bengali	Finnish	Indonesian	Nepali	Slovene	Ukrainian
Bulgarian	French	Italian	Norwegian	Swahili	Vietnamese
Burmese	Georgian	Japanese	Persian	Spanish	Yoruba
Catalan	German	Kazakh	Polish	Standard Arabic	
Croatian	Greek	Korean	Portuguese	Swedish	
Czech	Gujarati	Kyrgyz	Romanian	Tagalog	

The one per cent

- This is slightly less than one per cent of the world's languages

Possible reasons why LOL languages may bias a sample

- Areal bias: LOL languages overwhelmingly derive from the super-Saharan Old World
- ”Exotericness”: LOL languages are likely to be high-contact languages with many second language speakers – properties which have been claimed to be correlated to low morphological complexity and large phonological inventories
- Influence from written language: The extensive use of writing may influence syntax and phraseology
- Standardization: Conservative norms propagated in education may conserve obsolete features

Possible reasons why LOL languages may bias a sample

- Technology: The fact that LOL languages are spoken in technologically advanced societies may influence e.g. vocabulary (such as colour terms)
- Societal structure: Stratification of society influences e.g. pronouns and honorifics

LOL biases in typology

- Greenberg 1963: 14 LOL languages of 30 (47 per cent)
- Dahl 1985: 33 LOL languages of 64 (51 per cent)

LOL languages in WALS

- Number of LOL languages in WALS 100-sample: 24 (24 %)
- Number of LOL languages in WALS 200-sample: 27 (13.5 %)
- Percentage of WALS data coming from LOL languages: 7
- Maximal percentage of LOL data in a WALS map: 26.8

Why so many LOL languages?

- The large percentage of LOL language in the basic WALS samples is partly due to a conscious policy – the editors decided to add a few more “major languages of Eurasia” than would be motivated from the point of view of a sample that would be maximally free of genealogical and areal bias.
- However, this is not quite the whole story.
- The ”genealogically more balanced sample, with only one language per genus”, proposed on p. 6 in the book version of WALS, would remove four of the LOL languages in the 100 sample, leaving 20.

Bigger samples are better

- On the other hand, most maps in WALS contain large numbers of languages in addition to the basic samples, and these are much less obviously biased towards LOL languages.
- Example: Map 83A, Order of Object and Verb
  - 1519 languages
  - 53 LOL languages (3.4 %)
- Even so: more than 90 % of all LOL languages are in the sample, compared to about 22 % of all languages

Differences between total samples and LOL languages

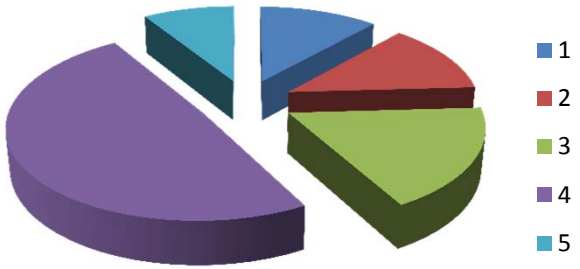
- Passive exists: 162 (43.4 %) vs. 28 (90 %)
- No antipassive: 146 (75.3 %) vs. 31 (100 %)
- No applicative constructions: 100 (54.6 %) vs. 26 (83.9 %)
- Both A and P arguments marked on verb: 193 (51.1 %) vs. 7 (22.6 %)
- Nonverbal encoding of predicative adjectives: 132 (34.2) vs. 29 (72.5 %)
- Relative pronouns on obliques: 13 (11.6 %) vs. 12 (46.2 %)
- Maximal number of basic colour categories: 11 (9.2 %) vs. 7 (87.5 %)
- No politeness distinction in pronouns: 136 (65.7 %) vs. 4 (10 %)
- Strongly suffixing: 406 (41.9 %) vs. 36 (76.0 %)

How much changes when removing the LOL languages?

feature	difference
45A Politeness Distinctions in Pronouns	13.34%
123A Relativization on Obliques	11.23%
76A Overlap between Situational and Epistemic Modal Marking	10.72%
62A Action Nominal Constructions	8.80%
75A Epistemic Possibility	8.40%
79A Suppletion According to Tense and Aspect	7.24%
122A Relativization on Subjects	6.62%
56A Conjunctions and Universal Quantifiers	5.55%
106A Reciprocal Constructions	5.44%
63A Noun Phrase Conjunction	4.97%

Relativization on obliques: removing the LOL languages

1 Relative pronoun	13	12	1
2 Non-reduction	14	1	13
3 Pronoun-retention	20	2	18
4 Gap	55	8	47
5 Not possible	10	3	7



with LOL languages

without LOL languages

Moral

- Literacy, political status, and number of speakers are factors that are not mentioned in the introduction to WALS; they are rarely seen as relevant for typological sampling.
- However, even if the effects may be restricted, it is worth keeping an eye on ”LOL biases” in linguistics.