RESEARCH REPORT

On Whorfian socioeconomics

THOMAS B. PEPINSKY

Cornell University

Whorfian socioeconomics is an emerging interdisciplinary field of study that holds that linguistic structures explain differences in beliefs, values, and opinions across communities. This field, which draws on linguistic relativity but extends it radically, holds that linguistic features are a fundamental explanation for variation in human behavior. This essay provides a conceptual overview and methodological critique of Whorfian socioeconomics, with a particular emphasis on empirical studies that document a correlation between the presence or absence of a linguistic feature in a survey respondent's language and their responses to survey questions. Using the universe of linguistic features from the *World atlas of language structures online* and a wide array of responses from the World Values Survey, I show that such an approach produces highly statistically significant correlations in a majority of analyses, irrespective of the theoretical plausibility linking linguistic features to respondent beliefs and behavior. I show how two simple and well-understood statistical fixes can more accurately reflect uncertainty in these analyses, and use them to replicate two prominent findings in Whorfian socioeconomics. The essay concludes by reflecting on the common methodological challenges facing linguists and other social scientists interested in nonlinguistic effects of linguistic structures.

Keywords: linguistic relativity, Sapir-Whorf hypothesis, survey research, social sciences, observational studies

1. INTRODUCTION. An emerging field of inquiry on the social and economic correlates of linguistic structures has renewed interest across the social sciences in old debates in linguistics. This new field of inquiry, which I term 'Whorfian socioeconomics', attributes causal effect to language structure in explaining social, economic, and political outcomes around the world. That line of reasoning has a long pedigree in linguistics and anthropology, originating with the so-called 'Sapir-Whorf hypothesis'—today termed the 'linguistic relativity' thesis. Social scientists commonly hold, crudely, that linguistic relativity implies that 'language shapes thought'; a more precise statement is that crosslinguistic disparities affect nonlinguistic cognitive processes of the populations in question.[1] Drawing on the linguistic relativity thesis but radically extending its predictions, Whorfian socioeconomics argues that linguistic structure has a sociologically and economically meaningful causal effect on values, beliefs, and behaviors across human societies. This is particularly interesting because one's native language—unlike variables such as class or income—is exogenous (i.e. causally prior to) the beliefs that one holds. In this way, linguistic structure resembles variables such as geography, climate, and genetic diversity as a fundamental explanation of human behavior (see e.g. Ashraf & Galor 2018, Diamond 1997, Sachs 2001, Sachs & Warner 1995).

In this essay I provide a conceptual overview and methodological critique of Whorfian socioeconomics. The conceptual overview distinguishes the proximate cognitive predictions of linguistic relativity from the longer chain of association in Whorfian socioeconomics that links those cognitive tasks to values and behaviors in other domains. Whorfian socioeconomics assumes that linguistic relativity is a valid theoretical paradigm, which entails that critiques of linguistic relativity undermine Whorfian socioeco-

* Thanks to Keith Chen, Abby Cohn, Peter Enns, Sara Goodman, Amy Liu, and seminar participants at Aarhus University for comments on an earlier draft.

[1] I thank an anonymous referee for this formulation.

nomics, but the reverse is not true. My methodological critique, in turn, identifies a series of thorny empirical issues that make testing the causal claims of Whorfian socioeconomics particularly challenging. Although many of these issues are common to research on linguistic relativity, some are novel and stem from the different kinds of data used—and methodological commitments held—in other parts of the social sciences.

Drawing on these methodological critiques, I then subject a central empirical strategy of Whorfian socioeconomics to critical scrutiny. Much of the empirical evidence marshaled in favor of Whorfian socioeconomics comes from crossnational survey data, in which researchers document a correlation between the presence or absence of a linguistic feature in a survey respondent's language and their responses to survey questions. I show that such correlations are remarkably easy to find, even among highly implausible pairs of language features and beliefs. Using the universe of language features from the *World atlas of language structures online* (*WALS*; Dryer & Haspelmath 2013) and twenty-five 'values variables' and four behavioral outcomes from six waves of the World Values Survey (WVS; Inglehart et al. 2014) which include 225,362 survey respondents, I uncover highly statistically significant correlations between thousands of pairs of language features and individual survey responses, even when controlling flexibly for a rich set of respondent demographic characteristics and country and year differences. Statistically significant 'feature-value' correlations using *WALS* and WVS data are incredibly common.

The purpose of this exercise is not to uncover spurious correlations and provide implausible theories to explain them. Rather, it is to show that researchers who set out to find statistically significant correlations between linguistic features and beliefs and behaviors are almost certain to find them, because such tests are of limited probative value. Some large proportion of these feature-value correlations are certainly spurious. I then show that two well-known strategies for accounting for correlations among groups of respondents—here, clustering standard errors by country and survey year (Cameron et al. 2011) and multilevel modeling (Barr et al. 2013, Gelman & Hill 2007)—substantially increase our uncertainty about parameter estimates, guarding against the conclusion that linguistic features actually do predict beliefs and behaviors. My results cannot be used to overturn any particular claim about how language shapes preferences, beliefs, or values. But these findings do recommend caution in interpreting feature-value correlations across languages and respondents, even when taking special care to account for differences across and within national language communities.

In the next section, I briefly review the linguistic relativity thesis and distinguish it from Whorfian socioeconomics. I then outline the main empirical strategies employed in Whorfian socioeconomics, both explaining why nonlinguists working in the field of Whorfian socioeconomics find them attractive and outlining their attendant weaknesses. A simulation study, using *WALS* and WVS data, illustrates the impossibly high number of statistically significant feature-value correlations that exist in commonly used data sets. Next, I outline two simple statistical procedures that mitigate the risk of type I error in feature-value correlations, and apply them to two prominent publications in order to show how a more conservative statistical procedure affects their inferences.[2] The final section concludes with a brief discussion of how methodological approaches in linguistics and the psychological sciences differ from those in applied microeconom-

[2] All data and analysis code are available for download and analysis at Dataverse (https://doi.org/10.7910/DVN/TEAJNJ).

ics and related fields of the social sciences. A shared methodological orientation will strengthen the broader intellectual project of Whorfian socioeconomics.

2. From linguistic relativity to whorfian socioeconomics. Linguistic relativity is a broad term that captures a collection of theories about the relationship between language and thought. The core argument—that linguistic disparities affect nonlinguistic cognitive processes—encompasses a range of positions, from the hypothesis that language supplies the categories that speakers use to classify objects and concepts to the belief that one's language constrains their ability not just to express but even to conceptualize certain ideas. Today, the concept of linguistic relativity is most closely associated with Edward Sapir and Benjamin Lee Whorf and is sometimes termed the 'Sapir-Whorf hypothesis', but the first and most prominent proponent of linguistic relativity was Wilhelm von Humboldt, who advanced a 'strong' or 'hard' interpretation that does not enjoy much support today (see von Humboldt 1999 [1836]). The weakest interpretation of the linguistic relativity thesis is relatively uncontroversial, as any English speaker who has learned grammatical gender in Spanish or the distinction between голубой 'light blue' and синий 'dark blue' in Russian can attest. Russian supplies a lexical distinction between shades of blue that English does not possess, requiring native English speakers learning Russian to be conscious of a categorical distinction of which they would not otherwise be aware. Here, we see crosslinguistic variation in lexical categories that affect a nonlinguistic cognitive domain.

The literature that I term Whorfian socioeconomics invokes a strong version of the linguistic relativity thesis to explain variation in beliefs and behaviors across human communities (for a survey of this literature targeting an economics audience, see Mavisakalyan & Weber 2018). For example, if it is true that some languages grammatically encode the future tense, and if it is also true that this affects speakers' conceptualization of the future, then speakers of those languages might behave differently in future-oriented tasks like saving (Chen 2013, Liang et al. 2018, Mavisakalyan et al. 2018, Pérez & Tavits 2017). Speakers of languages with gendered pronouns or articles might be more likely to hold gendered beliefs about social roles (Hicks et al. 2015, Jakiela & Ozier 2018, Liu et al. 2018, Mavisakalyan 2015, Pérez & Tavits 2019, van der Velde et al. 2015). Speakers of pro-drop languages might be less individualistic than speakers of non-pro-drop languages (Kashima & Kashima 1998, Licht et al. 2007, Tabellini 2008). Speakers of languages with obligatory politeness distinctions in second-person pronouns might hold more hierarchical beliefs about politics and society (Davis & Abdurazokzoda 2016, Galor et al. 2016, Kashima & Kashima 1998). Speakers of languages that distinguish between the inclusive and exclusive 'we' might have more sociotropic preferences (Wieczorek 2013).

The distinction between the established literature on linguistic relativity and the emerging literature on Whorfian socioeconomics lies in the length of the proposed causal chain linking linguistic features to social and behavioral outcomes of interest. In the established tradition of linguistic relativity research in linguistics, psychology, and anthropology, outcomes of interest are primarily nonlinguistic cognitive processes or behaviors that are proximate to the specific linguistic features of interest (although Slobin 1996 proposes that 'thinking for speaking' matters through the actual production of speech). One illustrative example is Levinson's (2003) research on spatial frames of reference. Guugu Yimithirr, a Pama-Nyungan language of Hopevale village in far northern Queensland, uses only cardinal directions (north, south, east, and west) to communicate the spatial location or orientation of objects, making it the only known pure case of a language with

solely an 'absolute' frame of reference. This differentiates it from other languages that may communicate location or orientation with reference to relative and/or intrinsic features ('in front of', 'left hand', etc.). Levinson (2003) theorizes that linguistic competence in Guugu Yimithirr requires speakers to retain a detailed mental accounting of the cardinal directions at all times, and hypothesizes that this will affect various nonlinguistic cognitive domains: dead reckoning, solving maze puzzles, and so forth. The evidence in favor of these hypotheses is compelling: Guugu Yimithir speakers, for example, are remarkably adept at tracking how far they have traveled in various directions in order to calculate their cardinal orientation relative to a starting point. Yet note how proximate the outcome variables here are to the linguistic feature: linguistic spatial frames of reference affect the cognitive processes governing location and orientation, and in turn affect how speakers complete spatially oriented tasks.

Whorfian socioeconomics, by contrast, lengthens the causal chain beyond the proximate nonlinguistic task associated with a linguistic feature to values, beliefs, and behaviors that might plausibly follow from the presence, absence, or salience of that feature or as consequences of those nonlinguistic tasks. The examples of grammatical gender and obligatory politeness distinctions in second-person pronouns given above help to illustrate the difference between linguistic relativity and Whorfian socioeconomics. That grammatical gender applied to nouns encourages speakers to conceptualize masculine nouns with 'masculine' characteristics and feminine nouns with 'feminine' characteristics is a relatively proximate causal claim consistent with the linguistic relativity thesis (Boroditsky et al. 2003). That grammatical gender encourages gendered beliefs about social roles (Hicks et al. 2015, Jakiela & Ozier 2018, Liu et al. 2018, Mavisakalyan 2015, Pérez & Tavits 2019, van der Velde et al. 2015) requires an additional series of theoretical links from categories applied to specific nouns to categories applied to gendered social categories more generally and without context. Likewise, obligatory politeness distinctions in pronouns might encourage speakers to pay greater attention to relative status hierarchies in conversation; to claim that they encourage more hierarchical social formations more broadly or lead cultures to be less 'individualistic' (Davis & Abdurazokzoda 2016, Galor et al. 2016, Kashima & Kashima 1998) requires a more developed theory to extend the argument from conversational pragmatics to social and political structures.

The longer causal chain required for Whorfian socioeconomics is not itself an argument against this theoretical approach. Yet many linguists are skeptical of Whorfian socioeconomics, even if they are sympathetic to some version of linguistic relativity. In a recent review of crosslinguistic typological comparisons, Ladd et al. (2015:227) write:

> In general, however, attempts to link language structure with extralinguistic factors are almost intrinsically suspect: The Boasian tradition (e.g., Boas 1931) insists that there are no 'primitive' languages and emphasizes the suitability of all languages to their speakers' communicative needs; Whorfian ideas about the cognitive biases imposed by the native language are no longer widely credited (but see Carroll et al. 2004, Levinson 2012); Chomskyans assume the existence of universal (and probably innate) structural principles; and the Saussurean foundation of all modern linguistics means that linguists take for granted the arbitrariness of linguistic form almost from the first day of their first introduction to the subject.

Responding to Feldmann's (2019) argument linking pro-drop languages to lower rates of completing secondary and tertiary education, Kennedy (2018) notes that 'the original claims of Sapir and Whorf (cited in Feldmann's paper) were of a very different nature: that a language can always meet the needs of its speakers'. Such critiques notwithstanding, Whorfian socioeconomics is likely to remain attractive to nonlinguists for three simple reasons: languages vary; linguistic relativity seems intuitively plausible, espe-

cially in its weak form; and languages are not 'choice variables' (people cannot choose their native tongue as a consequence of the values they hold). This last condition is particularly important, because it suggests that unlike values, opinions, and behaviors, a survey respondent's native tongue is unlikely to be causally affected by the outcome one seeks to explain.

3. Methodological challenges to whorfian socioeconomics. Although one may endorse a weak version of linguistic relativity while rejecting the findings of Whorfian socioeconomics, the reverse is not true: the theoretical arguments that support Whorfian socioeconomics require linguistic relativity to be valid. Put otherwise, arguments against Whorfian socioeconomics—either theoretical or methodological—need not dismiss linguistic relativity. I assume, for the purposes of argument, that some form of the linguistic relativity thesis is theoretically coherent; were it not, then Whorfian socioeconomics could not be theoretically coherent either. However, I take no position on the theoretical viability of Whorfian socioeconomics in this essay save to reiterate that Whorfian socioeconomics requires linguistic relativity. Recent reviews of the empirical evidence in favor of linguistic relativity can be found in Niemeier & Dirven 2000 and Everett 2013, among other sources; for a critical review, see McWhorter 2014.[3] Pérez 2018 provides an overview of the possible connections between language and public opinion responses.

In this section, I turn to the evidence used in the Whorfian socioeconomics literature. To reiterate, this literature makes causal claims: linguistic features are not simply correlated with values, beliefs, and behaviors; they also have a causal effect on values, beliefs, and behaviors. The methodological strategies used in Whorfian socioeconomics literature are similar to those used throughout the social sciences: correlational analysis at the individual or country level, experiments, and so-called 'quasi-experiments' or 'natural' experiments. Although my focus is on individual-level correlational analyses, I first review other methodological strategies and the challenges associated with them.

3.1. Aggregate, experimental, and quasi-experimental evidence. Many recent contributions in Whorfian socioeconomics marshal evidence at the aggregate level, correlating national-level variables with characteristics of the major languages spoken in those countries (Davis & Abdurazokzoda 2016, Galor et al. 2016, Jakiela & Ozier 2018, Licht et al. 2007, Mavisakalyan 2015, Tabellini 2008, van der Velde et al. 2015). This approach is powerful in its simplicity: dominant languages—and in particular, the features of dominant languages—are unlikely to be endogenous to the social and behavioral outcomes of interest, which is a threat for most aggregate questions in the social sciences (economic development may cause or be caused by democracy, civil conflict may cause or be caused by ethnic heterogeneity, and so forth). But of course, the standard problem of omitted-variable bias—or 'confounding', in the language of contemporary causal analysis—remains serious. It is generally difficult to conclude that one has measured all of the relevant confounding variables that might explain a national-level correlation between dominant language features and social or economic outcomes.

For linguists, however, such concerns may be secondary to more fundamental concerns about the coarseness of aggregate correlations. Without data from individual respondents, such correlations may fall victim to the ecological fallacy, in which aggregate findings comparing groups do not hold for individuals within those groups. More trou-

[3] Note that if it could be shown that linguistic relativity is theoretically coherent but empirically invalid, then it would follow that the core premises of Whorfian socioeconomics are invalid as well.

blingly, such analyses do not provide empirical evidence of the causal pathways or mechanisms that link linguistic features to social and economic outcomes, which skeptical linguists may consider a fatal flaw for any such analysis. Yet although the methodological state of the art in applied economics and causal inference views mechanisms as valuable parts of holistic causal explanations, the dominant counterfactual model of causation (see Morgan & Winship 2015 for an introduction) holds that causal effects may be estimated and causal relations may be studied without reference to or even knowledge of the mechanisms that link cause and effect (for critical overviews of mechanisms in causal explanation, see e.g. Bullock et al. 2010, Gerring 2010, Morgan & Winship 2015:338–44). Notably, the index of a landmark methodological text for applied economics, *Mostly harmless econometrics* (Angrist & Pischke 2009), contains no entry for 'mechanism' or 'causal mechanism'. The implication for Whorfian socioeconomics is that in the understanding of authors working within this tradition, aggregate correlations, unsupported by a detailed empirical account of the mechanisms linking linguistic features to aggregate behaviors or social structures, may still be interpreted as causal relationships if other assumptions about the system of variables hold.

Experimental methodologies, by contrast, are better equipped to identify causal effects than are aggregate correlations and can be used to probe causal mechanisms as well. But for the very reason that language minimizes endogeneity problems in observational studies—language is not a choice variable—convincing experiments are nearly impossible for the questions of interest to Whorfian socioeconomics. It is not possible to randomly assign speakers to native languages, much less to speaking native languages in communities that differ in no other way than the languages they speak. As a result, experimental manipulations in Whorfian socioeconomics are customarily done only within bilingual populations. In studying the social consequences of grammatical gender, Liu et al. (2018), for example, use bilingual Hungarian and Romanian speakers in Transylvania, and Pérez and Tavits (2019) use bilingual Russian and Estonian speakers in Estonia. But these studies muddy the claims of Whorfian socioeconomics, because speakers are competent in both languages (so claims must be narrowly about the effects of the language being spoken in real time). They also limit the external validity of these claims (because how could effects identified from bilingual speakers apply to monolingual speakers?).

The problems run deeper, though, for experimental approaches in Whorfian socioeconomics. A proper experiment linking feature *F* to belief or behavior *Y* must be able to assign speakers to a language with feature *F* while holding everything else—both nonlinguistic and linguistic—constant in expectation. But there is no plausible way to randomly assign speakers to grow up speaking French with or without the *tu*/*vous* distinction, or Bahasa Indonesia without a clusivity distinction in the first-person plural. Linguistic features, in this way, are what are known as 'bundled treatments' when we compare languages. Interestingly, in both Liu et al. 2018 and Pérez & Tavits 2017 the language that is a member of the Indo-European language family is gendered and the language that is from the Uralic language family is nongendered. How can we distinguish the effect of grammatical gender from other differences between these two language families?[4] The best one can do is to compare a language with feature *F* and a language without that feature, on the assumption that no other linguistic feature (alone

[4] Hungarian and Estonian are both agglutinative languages, and Romanian and Russian are synthetic languages. Might one argue that explicit attention to gender in conjugations explains the observed difference between the language pairs?

or in interaction) explains any observed difference between respondents. Alternatively, one may compare words within the same language, as with Tavits and Pérez's (2019) study of the effects of masculine versus gender-neutral pronouns in Swedish on gendered attitudes. But this abandons the goal of cataloguing the crosslinguistic effects of linguistic structures.

In applied economics and related social scientific fields, one way to invoke the precision of experiments in observational studies that compare across populations is to look for quasi-experiments or natural experiments. These are cases where even though the researcher does not control the assignment of treatment versus control (as in a standard experiment), linguistic features vary 'as good as randomly' across subject populations due to quirks of history, policy, or nature. Although such research designs are now quite common in fields like economics and political science, I am aware of only one study in Whorfian socioeconomics that attempts this sort of approach (Galor et al. 2016). The reason for the paucity of quasi-experimental Whorfian socioeconomics is, once again, that language is not a choice variable, so history and policy do not present many circumstances where populations find themselves 'as good as randomly' assigned to speak one language versus another.

3.2. Individual-level observational evidence. Individual-level observational studies drawing on large crossnational surveys offer a helpful compromise that draws on the strengths of both experimental and aggregate observational studies. By replacing aggregate with individual-level data, they sidestep the ecological fallacy. By comparing large numbers of languages that differ across many features, it is possible to allay concerns about the indeterminacy of two-language experimental comparisons. Likewise, by comparing many groups of individuals—in most cases, comparing many respondents across many countries—crossnational surveys allay concerns about the nonlinguistic differences among particular communities found in comparative studies of small numbers of language communities. And by conditioning on differences across individuals within countries, as well as differences across countries, regression-based approaches using large crossnational surveys ('feature-value correlations') can begin to approximate the quasi-experimental ideal that a particular linguistic feature is 'as good as randomly' assigned, conditional on those observed differences. Many of the recent contributions to the Whorfian socioeconomics literature—among others, Chen 2013, Pérez & Tavits 2017, Feldmann 2019, Gay et al. 2018, Hicks et al. 2015, Liu et al. 2018, and Mavisakalyan et al. 2018—are based in whole or in part on individual-level survey data.

There are, nevertheless, challenges facing the kinds of individual-level observational studies used to support Whorfian socioeconomics. Some are inherent to any observational study and parallel the issues facing the aggregate correlations described above. Because there is no way to guarantee that the set of observed covariates at the individual or group level adjusts for all differences across individuals and groups, causal interpretations for feature-value correlations require the assumption that there are no omitted confounders. And feature-value correlations cannot shed light on causal mechanisms linking linguistic features to beliefs and behaviors. Instead, analysts rely on the theoretical claims of linguistic relativity to interpret these correlations as causal and to supply the mechanisms that link linguistic features to relatively distal attitudinal or behavioral outcomes.

Other critiques, however, are particular to this approach. One important challenge to using crossnational public opinion data to test general claims about human language is the representativeness of the survey data: if public opinion surveys capture speakers of a nonrepresentative sample of languages, then they will not reflect the distribution of lin-

guistic features across human populations. This is almost certainly the case, as speakers of indigenous languages of the Americas, Africa, Northern and Central Asia, and Australia are underrepresented in data sources such as the World Values Survey relative to speakers of Indo-European, Sino-Tibetan, Afro-Asiatic, and a handful of other language families. This may render results for certain linguistic features meaningless, because only a small fraction of the speakers of languages containing those features are analyzed.

A separate issue is whether crossnational survey data can accurately capture the beliefs, opinions, and values of all human populations. For speakers of indigenous languages of the Americas, Africa, Northern and Central Asia, and Australia, survey items that ask about saving, business ownership, left-right political self-positioning, and other topics may function poorly because they do not reflect sociologically relevant aspects of respondents' lives. To the extent that such languages have distinctive linguistic features, any correlation between those features and survey responses will be plagued with measurement error.

Another critique that has played a central role in the online commentary by linguists on websites such as *Language Log* (http://languagelog.ldc.upenn.edu/nll/) focuses on the nonindependence of related languages and the diachronic patterns of cultural and linguistic evolution that generate the observed synchronic distribution of language features and survey responses. For example, using a mixed-effects modeling approach, Roberts et al. (2015) account for the interrelatedness of languages and find that evidence presented in Chen 2013 weakens. To my knowledge, no follow-up studies have followed exactly their empirical strategy, although others have acknowledged these challenges and have sought to account for differences across language families and/or cultural areas using fixed effects (e.g. Chen et al. 2017). Roberts et al.'s (2015) contribution nevertheless demonstrates that the historical connectedness among languages may generate statistically significant feature-value correlations that are not due to the effects of language itself.

I advance a related but conceptually distinct criticism in the next section, showing that even without considering issues of the representativeness of the survey data of all human language communities or of linguistic relatedness and evolution, the structure of the WVS data—with individuals nested within countries across years—creates distinct methodological challenges that the Whorfian socioeconomics literature has yet to fully embrace.

4. A SIMULATION STUDY. The preceding discussion has reviewed the methodological challenges facing Whorfian socioeconomics and distinguished the intellectual project of Whorfian socioeconomics from that of linguistic relativity. In this section, I look more closely at the role of individual-level observational evidence from crossnational surveys. Feature-value correlations are insufficiently conservative tests in the form that they are frequently run. The implication is that they are overly likely to result in statistically significant results that appear to support the strong form of the linguistic relativity thesis when they are in fact spurious correlations.

4.1. DATA AND METHODS. To test whether linguistic features explain values and beliefs, authors working with individual-level observational data search for correlations between the presence or absence of a feature in a speaker's language and the speaker's responses in public opinion surveys, using multiple regression to adjust for potential confounding variables. This is the strategy I follow here. The basic regression appears in equation 1:

$$(1)\quad Y_{iclt} = \beta F_l + \delta X_{iclt} + \phi_c + \tau_t + \epsilon_{iclt},$$

in which Y_{iclt} represents one of twenty-five WVS survey responses from respondents i who speak language l in country c and year t, X_{iclt} is a set of demographic control variables,[5] ϕ_c and τ_t are indicator variables for country and year, and ϵ_{iclt} is an error term. F_l denotes a linguistic feature of interest, and the coefficient β estimates the relationship between a linguistic feature and a WVS survey response.

The WVS survey responses can be found Table 1. Each captures a belief or value with socioeconomic significance, from self-placement on a left-right political scale to beliefs about the importance of work.

	VARIABLE	WVS CODE
1	Important in life: family	A001
2	Important in life: friends	A002
3	Important in life: leisure time	A003
4	Important in life: politics	A004
5	Important in life: work	A005
6	Important in life: religion	A006
7	Interest in politics	E023
8	Most people can be trusted: agree or disagree	A165
9	How much freedom of choice and control	A173
10	When jobs are scarce, men should have more right to a job than women	C001
11	When jobs are scarce, employers should give priority to a nation's people over immigrants	C002
12	Approve of a woman as a single parent	D023
13	Men make better political leaders than women do	D059
14	Aims of country: first choice is a stable economy	E001
15	Most important: first choice is a strong economy	E005
16	Willingness to fight for country	E012
17	Self-positioning in political scale (left-right)	E033
18	Government should intervene to address income equality	E035
19	Private vs. state ownership of business	E036
20	Should the government or the people take more responsibility	E037
21	Importance for political system: having a strong leader	E114
22	Importance for political system: having a democratic political system	E117
23	How proud of nationality	G006
24	Religious person	F034
25	Feeling of happiness	A008

TABLE 1. WVS variables. The WVS code corresponds to the harmonized variable code in the aggregated WVS data file (Inglehart et al. 2014).

The dependent variables range from binary yes/no variables to ordinal scales from 1–10, but to ensure comparability across analyses I model each dependent variable using ORDINARY LEAST-SQUARES (OLS) regression. The elements of X include sex, age, employment status, marital status, highest level of education, self-assessed social class, and country-specific income decile. Each enters the regression equation as a series of indicators for each value of each variable except for age, which enters in linear and quadratic forms.

The linguistic features from *WALS*, F_l, are found in Table A1 in Appendix A. The original *WALS* indicators are nominal variables containing between two and eight different categories per feature, but for the purposes of highlighting, contrasts were recoded into a series of binary variables, usually indicating a 'presence/absence' or a 'most common versus others' contrast. Because these dichotomizations were made

[5] These include respondent age, age^2, and dummy variables capturing employment status, marital status, country-specific income level (ten categories), highest level of education, subjective social class, and sex.

without any particular theory in mind, it is possible that different codings of these variables would produce different results. A small number of *WALS* features do not vary at all among the languages of the respondents in the sample and were dropped from the analysis. In all, there are 138 usable linguistic feature variables from *WALS*.

WVS records the language each respondent reports speaking at home, which I matched to the linguistic features for that language found in *WALS*.[6] I was able to match 234 languages from WVS to their corresponding *WALS* languages (a list of languages and total number of respondents in the WVS data appears in Table A2 in Appendix A, along with the *WALS* language name that I used to match the WVS and *WALS* data). Of the languages that appear in WVS, 135 remain unmatched, but these amount to only 7.1% of all respondents. Using these data, I regressed each element of *Y* on each element of *F* and controls, for a total of 3,350 regressions, collecting the coefficients β and the *t*-statistics for each.

4.2. Results. If statistically significant feature-value correlations are common, then this entails either that all sorts of linguistic features explain all sorts of beliefs and values, or that at least some such correlations are spurious, an artifact of the statistical methods employed. Figure 1 displays the main results of this analysis with a density plot of the absolute value of the *t*-statistics from these regressions.

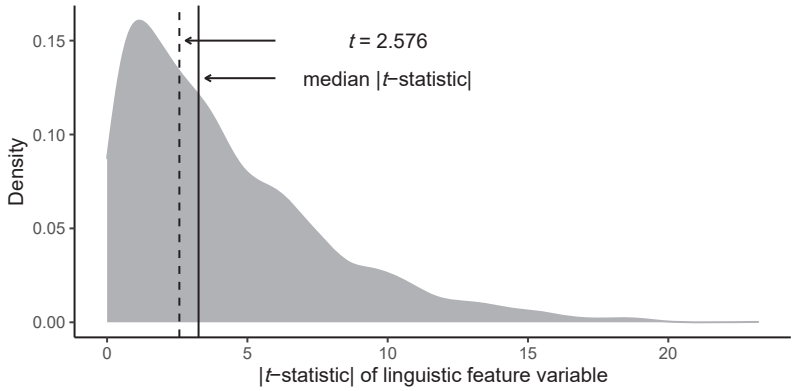


Figure 1. Basic results. The dashed line at $t = 2.576$ corresponds to $p < 0.01$ in a two-tailed test. The solid line is the overall median of all *t*-statistics.

The dashed reference line at $t = 2.576$ corresponds to a 99% significance level in a two-tailed test. Estimates to the right of that line would be considered highly statistically significant under the conventional null hypothesis significance-testing framework. The black reference line is the median of all the collected *t*-statistics. These results show that a substantial majority of feature-value regressions are highly statistically significant at conventional levels. Two-thirds of all results have a *t*-statistic of 1.96

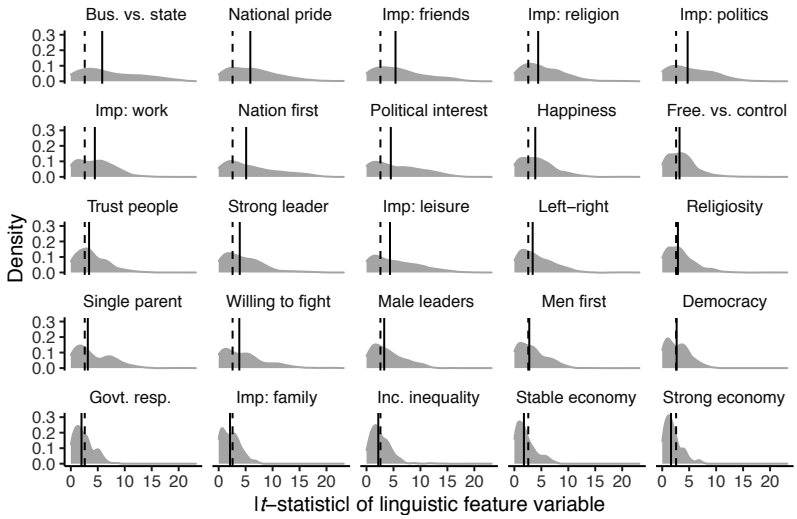
[6] WVS also includes a variable that captures the language used in the interview, but I use the language spoken at home on the hypothesis that this is the language in which the speaker is most 'naturally' competent. Nevertheless, to check that this decision did not influence my results, I replicated the analysis here using the language in which the interview was conducted as the match variable. Although individual results vary, the general pattern of highly statistically significant results remains the same. The mean *t*-statistic using the language spoken at home as the match variable is 4.19, and that using the language of the interview as the match variable is 3.80. A comparison of results is available in Figure A1 in Appendix A.

or greater (for $p < 0.05$) in absolute value, and one-third of all estimates feature *t*-statistics greater than 5.

To get a sense of the magnitude of these estimated effects, I take the absolute value of the estimated coefficients that are statistically significant at the $p < 0.01$ level and divide each by the standard deviation of its associated dependent variable. Because the features are all binary variables, this expresses the effect size for each coefficient as a change in the standard deviation of its dependent variable. Following Cohen's (1988) rules of thumb, 88% of the estimated effect sizes are substantively very small (< 0.2), with a median effect size of 0.094.

How to interpret the results? If we knew for certain that linguistic features had no effect on attitudes and beliefs, then a statistical test should uncover no evidence of a relationship between features and values. In many such tests, we should reject the null hypothesis of no association at the $p < 0.05$ level approximately 5% of the time, at the $p < 0.01$ level approximately 1% of the time, and so forth. That we reject the null hypothesis nearly 70% of the time in these regressions implies either that these associations are actually very common or that our statistical procedure is subject to type I error. Given the nature of the simulation exercise, exhaustively pairing all *WALS* features with a wide range of dependent variables, it is highly unlikely that each of these regressions captures a theoretically plausible causal link between a linguistic feature and a value or belief. Some proportion of the results from these analyses are certainly spurious, although it is possible that some of these estimated coefficients reflect true causal effects of linguistic features on beliefs.

Might these results be driven by one or two of the WVS variables that happen to be correlated with linguistic feature variables? In Figure 2, I check for this possibility by breaking down the results by each of the twenty-five dependent variables.



Hope: Send art with fonts embedded for figs 2-A2.

FIGURE 2. Results by dependent variable. The dashed line at $t = 2.576$ corresponds to $p < 0.01$ in a two-tailed test. The solid line is the median of the *t*-statistics for each dependent variable. Dependent variables are ordered from top-left to bottom-right by the proportion of results that are statistically significant at the $p < 0.01$ level.

Here we discover that there is, indeed, variation across WVS variables in terms of their tendency to correlate with linguistic features. In the case of the variable measuring 'the importance of work', fewer than 50% of all results are statistically significant at the $p < 0.01$ level. In other cases, such as beliefs in 'personal freedom versus control', the 'importance of religion', or 'political interest', statistically significant correlations are particularly common. But the main takeaway point from Fig. 2 is that even among those values and beliefs that are less likely to correlate with linguistic feature variables, highly statistically significant results remain quite common.

Finally, one might suspect that the above results are driven by the fact that the dependent variables all capture beliefs and values. Might a more objective behavioral outcome as dependent variable yield different results that are less likely to be spuriously correlated with linguistic features? To check, I chose four additional outcome variables that ask about actual respondent behavior from WVS and repeated the analyses above, which provided me with 536 more regression analyses. Table 2 describes these four new outcome variables, and Figure 3 presents the results.

	VARIABLE	WVS CODE
1	Participate in politics by signing a petition	E025
2	Frequency of attending religious services	F028
3	Respondent's number of children	X011
4	Family savings during the past year	X044

TABLE 2. WVS behavior variables. The WVS code corresponds to the harmonized variable code in the aggregated WVS data file (Inglehart et al. 2014).

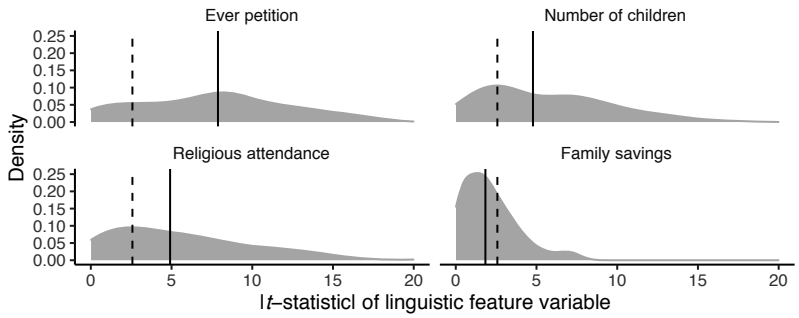


FIGURE 3. Behavioral results by dependent variable. The dashed line at $t = 2.576$ corresponds to $p < 0.01$ in a two-tailed test. The solid line is the median of the t-statistics for each dependent variable.

If anything, these results are even more discouraging. Linguistic features are consistently significant predictors of signing a petition, the number of children that a respondent has, the frequency with which they attend religious services, and saving money.

5. ALTERNATIVES. These results should be troubling for the Whorfian socioeconomics literature: either linguistic features are strong predictors of a great range of beliefs and behaviors, or some fraction of these results are spurious because such tests have limited probative value. Taken together, they should signal a word of caution to researchers encountering a statistically significant correlation between one particular linguistic feature and any particular value, belief, or behavior as captured in survey data. Given that standard regression approaches produce results that are almost certainly overconfident—and given that Whorfian socioeconomics ought to be tested empirically—how might researchers proceed? In what follows, I provide two suggestions.

Before I do so, however, let us consider a scenario in which a researcher has conducted exactly the analysis that I have shown above—examining all possible correlations between *WALS* features and WVS responses—and then selected a single statistically significant partial correlation (or some set of them) to publish as a result. Theories may emerge that correspond to the available statistically significant findings, a practice known as 'HARKing', or '**h**ypothesizing **a**fter the **r**esults are **k**nown' (Kerr 1998). The practice of publishing only those results that are statistically significant is sometimes described as '*p*-fishing' or '*p*-hacking'. In this scenario, standard critical values for statistical significance are no longer valid. This is because each test has some probability of rejecting even a true null hypothesis, so a collection of many such tests makes it almost certain that one or more will reject a null hypothesis.

Standard methods for adjusting for multiple comparisons are well understood by linguists (see e.g. Riazi 2016:22–23 on the Bonferroni correction) and may be applied in such a scenario. For example, if I wished to claim that speakers of languages with voicing gaps in plosive consonants are less likely to be willing to fight for their country, given that I have run 3,886 regressions in total in order to find that particular result, the Bonferroni-corrected critical *p*-value corresponding to 95% significance should be 0.05/3886 = 0.000013, a far stricter test.

However, adopting a Bonferroni correction presumes that scholars will report all of the tests they have conducted. But the same publication incentives that encourage HARKing also discourage faithful reporting of all possible tests that have been conducted. Concerns about HARKing, *p*-fishing, and related problems have prompted extensive discussion of how best to constrain researchers to avoid the publication of false positives, including radical transparency in methodological choices, preregistration of research designs, and so forth (see e.g. Humphreys et al. 2013, Ioannidis 2005, Simmons et al. 2011). Preregistration is increasingly common in the field of psycholinguistics, which shares many of psychology's concerns about reproducibility.

These discussions of how to address multiple comparisons or how to incentivize good research practices are ultimately issues about researcher discretion in analyzing and reporting results. The suggestions described below, by contrast, address the nature of the data in feature-value correlations. They should be applied regardless of how many analyses the researcher has conducted, and they apply equally whether or not the researcher has well-defined hypotheses before conducting a particular analysis. Finally, because they require no additional data beyond that which is already used in the analysis, using software routines that are already available in common statistical packages, they are nearly costless to implement as robustness tests.

5.1. Clustering standard errors. The baseline model specified in equation 1 includes fixed effects for both country and year, which absorb any baseline differences in responses that are particular to the country in which the respondent lives or the year in which the survey was taken. Nevertheless, it is possible that the error terms captured in ϵ_{iclt} are correlated among respondents, across either countries or years. As is well known, OLS regression produces unbiased estimates of regression coefficients in the presence of correlated errors (or any form of heteroscedasticity), but estimates of uncertainty may be too optimistic. A straightforward way to account for the clustered nature of the data is to estimate 'cluster-robust' standard errors, as described in Cameron et al. 2011, which relax the assumption that regression errors are homoscedastic.

I present formulas for calculating cluster-robust standard errors in Appendix B, but the intuition behind clustering standard errors is straightforward. The survey responses in a particular country, or in a particular year, are likely to be subject to idiosyncratic

shocks that are common to all respondents within countries or years. Such a possibility violates the assumption of homoscedasticity required in the calculation of 'classical' standard errors. Adjusting for country-specific differences through country fixed effects allows us to distinguish linguistic features from country effects, but this does not generally account for all possible sources of correlated errors within countries. The same conclusion holds for year fixed effects and correlated errors within years. Further intuition behind the distinction between fixed effects and clustered standard errors can be found in Cameron & Miller 2015:329–30.

Clustering could have a substantial impact on estimated standard errors if particular languages tend to be spoken within certain countries. This is obviously true: every WVS interview conducted in Hebrew[7] takes place in Israel, every WVS interview in Indonesian is from Indonesia, and 92% of all interviews in Hungarian were conducted in Hungary. Some published studies have adopted an approach of clustering standard errors by country. For example, Chen 2013 clusters results by country although only some regressions include continent (not country) fixed effects, and Pérez & Tavits 2017 clusters by country but does not include country fixed effects, while Pérez & Tavits 2019 includes country and year fixed effects and clusters by country only, and Chen et al. 2017 includes continent and year fixed effects and clusters by country only. Several recent studies (e.g. Gay et al. 2018, Liu et al. 2018, Mavisakalyan et al. 2018) do not cluster across any dimension, or cluster on a different dimension altogether (e.g. Liang et al. (2018), who cluster by firm).[8] Troublingly, Hicks et al. (2015:26) report that they do not cluster by country when including country fixed effects because 'doing both at the same level may produce unreliably smaller standard errors'. I am unaware of any work on Whorfian socioeconomics that clusters by country and survey year in studies using WVS/*WALS* data.

The question of whether one 'should' cluster standard errors seems at first glance to hinge on the assumption of homoscedasticity. The emerging literature on cluster-robust inference addresses different considerations, such as the number of groups (here, countries and years) within each cluster (see Cameron & Miller 2015:340–50) and the level at which the causal variable of interest varies (see Angrist & Pischke 2009:319–22). In the context of fixed-effects regressions such as the ones estimated here, Abadie et al. (2017) propose that if the analysis sample is not a random sample of the population of interest, then clustering is necessary so long as there is heterogeneity in the causal effects being estimated (which is almost certainly the case). Since neither the WVS countries nor the WVS survey years are a random sample from the population of humans speaking languages across time and space, which is the population of interest in analyses of feature-value correlations, it follows that clustering by country and year is necessary.

The other consideration—the number of clusters per dimension—warrants further care. In WVS data, there are more clusters across space (country) than there are across time (survey wave or year). When the number of clusters is small, the asymptotic results in Cameron & Miller 2015 do not apply, and clustered results may actually be too conservative. MacKinnon et al. (2017) suggest a bootstrap-based approach for data-

[7] Recorded for some reason as 'Jewish' in the WVS integrated data file.

[8] Gay et al. (2018) include fixed effects by country, country of origin, and country of origin by decade but do not cluster on any dimension. Mavisakalyan et al. (2018) include fixed effects by country and language family but do not cluster on any dimension, and they also estimate crossnational regressions for which clustering by country or year is not necessary. Liu et al. (2018) do not include fixed effects and do not cluster on any dimension, and they also include an experimental analysis for which clustering is also not necessary.

structures scenarios with multiway clustering and a small number of clusters on one dimension. See Cameron et al. 2008 for further details on what they term a 'wild cluster bootstrap' approach for cluster-robust inference with a small number of clusters.

5.2. Modeling language variation. A different approach to estimating the effects of linguistic features recognizes that languages differ from one another for reasons other than the linguistic feature variables themselves. Ideally, one would use language-specific fixed effects to adjust for any unobserved features particular to each language, but such an approach makes it impossible to estimate a separate coefficient for the linguistic feature. An alternative strategy that amounts to a compromise between ignoring differences across languages and accounting for them with language-specific fixed effects is to adopt a multilevel modeling approach (Gelman & Hill 2007:237–58). This approach makes a slight modification to equation 1.

$$(2)\quad Y_{iclt} = \beta F_l + \delta X_{iclt} + \phi_c + \tau_t + \lambda_l + \epsilon_{iclt}$$

In addition to country and year fixed effects ϕ_c and τ_t, equation 2 includes λ_l, a random intercept that is identified by assuming that its distribution follows $N(0, \sigma^2_\lambda)$, allowing σ^2_λ to be estimated from the data. This allows us to estimate β while accounting for differences across languages, and still also including country and year fixed effects as above. Unlike clustering, a multilevel modeling approach affects the estimate of β in addition to its standard error.

To my knowledge, no studies in the Whorfian socioeconomics tradition adopt a multilevel modeling approach that allows for random effects by language. Random-effects models are, however, standard in the analysis of experimental data in psycholinguistics (see Clark 1973 for an early methodological statement). Recent overviews of the use of multilevel models in psycholinguistics and related fields include Baayen et al. 2008 and Barr et al. 2013, and multilevel modeling approaches have recently been proposed for the fields of corpus linguistics (Gries 2015) and quantitative sociolinguistics (Johnson 2009).

Despite these methodological parallels, the logic behind using multilevel modeling for survey data differs in important ways from the logic in psycholinguistics and the analysis of experimental data with repeated subjects or word items. And the distinction between the terms 'fixed effects' and 'random effects' as employed in the Whorfian socioeconomics literature—whose methodological approach draws primarily from the field of applied microeconomics (see Angrist & Pischke 2009 for an overview)—can be misleading from the perspective of multilevel modeling in linguistics. Above, I used the terms 'country fixed effects' and 'year fixed effects' to describe a series of indicator variables that estimate separate intercepts for each country and year in order to account for any differences across countries and years. This differs from the use of the term 'fixed effects' to describe individual-level covariates in linguistic research, as used for example by Wieling et al. (2011:4): 'fixed-effect factors are factors with a small number of levels that exhaust all possible levels (e.g., the gender of a speaker is either male or female)'.

Fixed effects in the applied microeconomics sense absorb any differences across respondents at higher levels of aggregation at the cost of making it impossible to estimate the regression coefficients of aggregate-level covariates. However, because country- and year-level variables are not of direct interest in Whorfian socioeconomics, this cost is minimal. As noted above, the fact that linguistic features are of interest prevents the inclusion of language-specific fixed effects, so language-specific random effects in a multilevel modeling framework represent a feasible alternative.

5.3. Adjusted results. To illustrate how these two adjustments affect inferences about the statistical significance of linguistic feature variables, I collected the fifty re-

gression results with the highest *t*-statistics from all of the analyses produced so far (both values and behaviors) and reestimated each model using both methods of adjustment. The smallest of the *t*-statistics among these fifty regressions is 15.4, which corresponds to a *p*-value of 7.76×10^{-53}. For the clustered standard errors approach, I cluster by both country and survey year. For the multilevel modeling approach, I estimate equation 2. Figure 4 compares the results from the unadjusted regressions to the corresponding estimates using both adjustments.

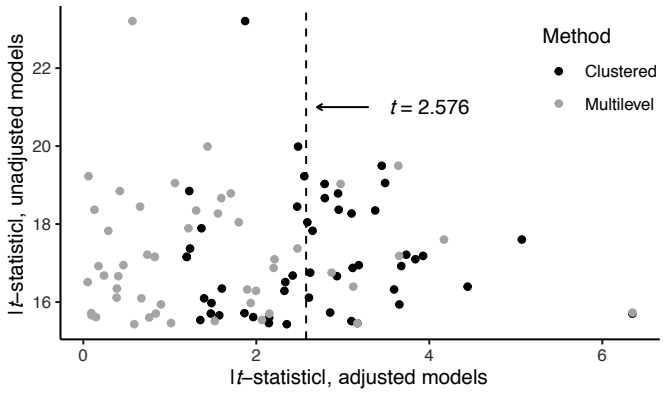


FIGURE 4. Adjusting the fifty most significant results. The dashed line at $t = 2.576$ corresponds to $p < 0.01$ in a two-tailed test.

A dramatic decrease in statistical significance relative to the baseline modeling approach is evident for both methods. The median *t*-statistic for the fifty most statistically significant unadjusted models is 16.96, but when adopting these adjustments for the same models, the median *t*-statistic drops to 2.62 for the clustered models and 1.04 for the multilevel models. This reduction in the statistical significance of the linguistic feature variables is consistent with the relatively more conservative nature of these statistical models, which account for either the grouped nature of the survey responses or the differences across languages aside from linguistic features themselves.

Unfortunately, we do not know the true effects of any language features on respondent beliefs or behaviors, so we cannot know how confident we should be in statistically significant feature-value correlations that survive these more conservative tests. In other words, we cannot know if these more conservative estimation strategies 'work'. However, the takeaway point is that either of these two approaches is preferable to a naive approach which ignores these other kinds of heterogeneity across survey respondents that happen to be correlated with particular linguistic features.

To further elucidate how these alternative methods fare in analyzing data of the form presented here, I present in Appendix C two additional simulation analyses. First, I demonstrate that these methods produce insignificant findings when effects are known to be zero. I do this by randomly generating binary linguistic features for the languages in the data set, and then showing that the standard approach, which does not model language-level variation, incorrectly produces statistically significant findings in the majority of analyses, whereas the multilevel modeling approach achieves the appropriate rejection rates, and the two-way clustering approach achieves rejection rates that are much closer to the appropriate ones. Second, I demonstrate that these methods do not overturn results for statistical associations that are theoretically plausible. When analyz-

ing the relationships between employment status and savings, social class and left-right political positioning, and household income and preferences for redistribution, coefficient estimates remain highly statistically significant even in the more conservative tests advocated here.

One final concern is what to infer from the finding that the inclusion of random effects by language overturns the main results for feature-value correlations. Although these results are strictly inconsistent with any Whorfian socioeconomics-derived hypothesis that linguistic features explain values, beliefs, and behaviors, does this not entail that something about language nevertheless matters? The answer is 'no': what these results actually demonstrate is that something correlated with language is associated with values, beliefs, and behaviors even when controlling for country and year differences as well as individual-level covariates. It could be language itself, or it could be something else that we identify when separating respondents by language. These methods do not allow us to resolve this ambiguity. If the goal is to determine whether language speakers differ because of the languages they speak, a more precise statistical approach is still necessary.

6. Two replications. The preceding discussion has proposed two simple statistical fixes to use when correlating linguistic features with WVS data, and it has shown how these fixes produce more conservative estimates of the relationships between linguistic features and survey responses. But would adopting such methods change the inferences we draw from any published results? In this section I replicate two prominent studies in Whorfian socioeconomics, each published in a top disciplinary journal in the social sciences, but fail to replicate their key findings when using the methodologies introduced above.

Chen 2013, published in the *American Economic Review*,[9] is perhaps the most prominent example of the new literature on Whorfian socioeconomics. It was selected as an editors' choice article in *Science* and was widely debated at *Language Log* and elsewhere. The article provides extensive evidence that speakers of languages that grammatically encode time have more future-oriented behaviors. The statistical methods employed in Chen's analyses of WVS data are common in fields such as labor economics and are quite conservative: his most conservative analyses use a series of interacted fixed effects for country, survey wave, age, sex, and other demographic characteristics to form precise comparisons among respondents who are identical across each of those background characteristics but who differ in whether they speak a language that encodes these distinctions.

Chen's (2013) findings have already been subject to critical scrutiny. In one notable example, Roberts et al. (2015) show that after accounting for the phylogenetic and geographical relatedness of languages using a multilevel modeling approach, the correlation Chen identified is no longer statistically significant in a number of tests. Roberts et al.'s use of random intercepts for language family and geographic area proves a principled account of the sources of heterogeneity across languages. The multilevel modeling approach I introduce differs in that it is agnostic about the sources of within-language correlation and leaves unmodeled any cross-language variation. Future research, of course, may combine our approaches by including random intercepts by language and language family.

[9] According to Google Scholar data from December 2019 (https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=bus_economics), the *American Economic Review* has the highest H-5 index of any journal in economics.

I was able to download Chen's replication materials and to reproduce his exact results. To extend his analysis and investigate how alternative modeling strategies might affect the inferences drawn from such regressions, I focus on his table 3 (Chen 2013:707), specifically on models 1 and 2. The outcome variable of interest is whether the respondent's family reported having saved in the past year (see Table 2 above). Unlike the simulation results presented above, which employ OLS regression, Chen uses logistic regression because his dependent variable of interest is binary. Because the arguments above about clustering and multilevel modeling apply equally to logistic regression, in order to ensure that my replication is as comparable as possible, I estimate logistic regression models as well.[10] His strategy of using interacted fixed effects differs from the simulation results I have shown thus far, but I follow this as well—with the exception that in the multilevel modeling approach, I replace his age × sex × country × wave × income × education fixed effects from model 2 with age × sex × income × education fixed effects and separate country and year fixed effects.[11] This change is conceptually quite minor and is not responsible for the different results I obtain. Finally, to ensure that differences in software implementation do not explain any of the differences in results that I find below, I follow the author and analyze his data using Stata, version 15 (StataCorp 2017).

The replication results appear in Table 3. Columns 1 and 5 correspond exactly to models 1 and 2 from Chen's (2013) table 3, although I report logistic regression coefficients rather than odds ratios. Columns 2 and 6 replace his country-clustered standard errors with two-way clustered standard errors by country and year in a logistic regression model. But because Chen's data contain only fifteen years, rather far from the number of clusters needed for the asymptotics in Cameron et al. 2011 to apply, they may be too conservative. So columns 3 and 7 implement the wild cluster bootstrap by year and the standard cluster adjustment by country via the methods implemented in Roodman et al. 2019. Finally, columns 4 and 8 show results from multilevel logistic regressions. Model 1 of Chen 2013 contains no country or year fixed effects, and these results remain significant with two-way clustering but not in the multilevel model. The results from his model 2, which introduces (interacted) country and year fixed effects, do not remain statistically significant at the $\alpha = 0.05$ level with two-way clustering (with or without the wild cluster bootstrap) or a multilevel model.

I next replicated the findings from Pérez & Tavits 2017, published in the *American Journal of Political Science*,[12] which also investigated the relationship between grammatical encoding of the future and prospective behavior. The authors employ a statistical approach similar to that of Chen (2013), although they choose not to estimate the more conservative statistical models Chen employs. My results below confirm that had they done so, following Chen's code beyond the first regression model, their statistical results would not have supported their hypotheses. I focus on table 2 (Pérez & Tavits 2017:724) and specifically on models 1 and 2, where the dependent variable is a respondent's belief in the importance of protecting the environment even if doing so were to have negative implications for the economy. As above, I follow the authors and estimate logistic regressions, again using Stata.

[10] Another useful feature of a logistic regression approach is that it drops any fixed-effects groups in which there is no variation in the dependent variable, as these can contribute no information about the effects of linguistic features. This underscores just how conservative Chen's methods are.

[11] A multilevel logistic regression model with the fully interacted fixed effects would require tens of thousands of dummy variables, too many for most software packages to estimate.

[12] According to Google Scholar data from December 2019 (https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=soc_politicalscience), the *American Journal of Political Science* has the highest H-5 index of any journal in political science.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Chen 2013	model 1	model 1	model 1	model 1	model 2	model 2	model 2	model 2
Future	−0.777**	−0.777**	−0.777**	−0.052	−0.331*	−0.331	−0.331	−0.227
	(0.151)	(0.193)	[a]	(0.192)	(0.158)	(0.193)	[a]	(0.120)
N	152,056	152,056	152,056	152,056	64,017	64,017	64,017	135,666
Extended controls	N	N	N	N	Y	Y	Y	Y
Country FEs	N	N	N	N	Y	Y	Y	Y
Year FEs	N	N	N	N	Y	Y	Y	Y
Country clusters	Y	Y	Y	N	Y	Y	Y	N
Year clusters	N	Y	Y	N	N	Y	Y	N
Wild cluster bootstrap	N	N	Y	N	N	N	Y	N
Bootstrap *p*-value[a]	—	—	< 0.001	—	—	—	0.094	—
Language REs	N	N	N	Y	N	N	N	Y

Table 3. Replication of Chen 2013, table 3. Logistic regression coefficients, with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$. [a] Standard errors are not calculated in the wild cluster bootstrap analysis; instead, inferences are based on the *p*-value calculated using Roodman et al. 2019.

Table 4 presents the results. Columns 1 and 7 correspond exactly to Pérez and Tavits's models 1 and 2. In columns 2 and 8, I first extend their models by including country and year fixed effects, with no adjustment of standard errors for clustering across countries and years. That the statistical significance of these results already drops precipitously is a worrying sign. In columns 3 and 9 I also add two-way clustered standard errors, further confirming this result. The number of years here is even smaller—only five—so columns 4 and 10 implement the wild cluster bootstrap for clustering by year, which still fails to reject the null hypothesis at the $\alpha = 0.05$ level. Columns 5 and 11 add random effects for language but remove the country or year fixed effects, whereas columns 6 and 12 add random effects as well as country and year fixed effects. The results are clear: unobserved country and year effects explain the findings in Pérez & Tavits 2017, and neither two-way clustering (with or without the wild cluster bootstrap) nor multilevel modeling can recover a significant partial correlation between grammatical future encoding and views on the environment.

It is important not to conclude too much from these results: judgments that correlations are significant or not rest on arbitrary thresholds of what constitutes statistical significance. Nevertheless, replicating two prominent publications on Whorfian socioeconomics reveals the importance of statistical modeling in linking linguistic features to survey responses. Although these statistical procedures are well understood, by linguists and by others, these findings confirm that failing to model variation across languages and survey respondents properly can produce overconfident results. Skeptics of what we can learn from feature-value correlations should nevertheless understand the methods upon which the statistical results rest; those that make no adjustments to account for the clustered nature of language data are bound to be anticonservative, and simple solutions may address them. To be clear, I have chosen to replicate these two publications because the authors have made their replication materials publicly available, and it is to Chen's (2013) and Pérez and Tavits's (2017) credit that they have done so. Making replication data publicly available makes testing these questions easier, but given the fragility of the results I have been able to replicate, researchers should consider those other results as provisional at best until their analysis data and code are available for public scrutiny.

7. Conclusion. This essay has presented a conceptual overview and methodological critique of Whorfian socioeconomics. Because language is not a choice variable, it is attractive to nonlinguists because it is unlikely to be endogenous; but this very fact also

	(1)	(2)	(3)	(4)	(5)	(6)
Pérez & Tavits 2017	model 1	model 1	model 1	model 1	model 1	model 1
Future	−0.309*	−0.185	−0.185	−0.185	−0.554**	−0.112
	(0.157)	(0.135)	(0.147)	[a]	(0.148)	(0.121)
N	64,666	64,666	64,666	64,666	64,666	64,666
Extended controls	N	N	N	N	N	N
Country FEs	N	Y	Y	Y	N	Y
Year FEs	N	Y	Y	Y	N	Y
Country clusters	Y	Y	Y	Y	N	N
Year clusters	N	N	Y	Y	N	N
Wild cluster bootstrap	N	N	N	Y	N	N
Bootstrap *p*-value[a]	—	—	—	0.095	—	—
Language REs	N	N	N	N	Y	Y
	(7)	(8)	(9)	(10)	(11)	(12)
Pérez & Tavits 2017	model 2	model 2	model 2	model 2	model 2	model 2
Future	−0.324*	−0.221	−0.221	−0.221	−0.519**	−0.091
	(0.155)	(0.129)	(0.137)	[a]	(0.150)	(0.126)
N	61,029	61,029	61,029	61,029	61,029	61,029
Extended controls	Y	Y	Y	Y	Y	Y
Country FEs	N	Y	Y	Y	N	Y
Year FEs	N	Y	Y	Y	N	Y
Country clusters	Y	Y	Y	Y	N	N
Year clusters	N	N	Y	Y	N	N
Wild cluster bootstrap	N	N	N	Y	N	N
Bootstrap *p*-value[a]	—	—	—	0.119	—	—
Language REs	N	N	N	N	Y	Y

Table 4. Replication of Pérez & Tavits 2017, table 2. Logistic regression coefficients, with standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$. [a] Standard errors are not calculated in the wild cluster bootstrap analysis; instead, inferences are based on the *p*-value calculated using Roodman et al. 2019.

creates a host of related methodological challenges. Credible experiments are difficult, but observational data are imperfect. Focusing on the widely publicized correlations between linguistic features and survey responses that lie at the heart of this emerging literature, and looking across all language features in *WALS* and a wide range of survey responses, I show that thousands of theoretically implausible yet highly statistically significant correlations between linguistic features and survey responses exist. Most of these correlations are certainly spurious. Fortunately, simple and well-understood statistical procedures can guard against the types of type I error that are so common in individual-level analyses of crossnational survey data. Two replication analyses demonstrate the fragility of prominent articles in both economics and political science. The empirical evidence in favor of the linguistic relativity thesis from the field of linguistics itself is compelling, but much narrower in scope.

This last point is important for evaluating how this manuscript fits into the broader body of research on Whorfian socioeconomics. My statistical critique of analyses of crossnational public opinion data addresses only one piece of evidence that links language to public opinion. As emphasized above, researchers have invoked a wide range of evidence, from aggregate correlations at the national level to lab-based studies that randomly assign tasks by language, to test the hypothesis that linguistic features have sociologically and economically meaningful causal effects on beliefs and behaviors. My findings do not imply that one should dismiss all statistical research linking language and public opinion. By addressing the inferential challenges associated with attributing causal meaning to variation across linguistic features in comparative statistical

research, however, my analysis highlights how different forms of statistical evidence fit together to support a broader research agenda.

It may be tempting to conclude that this essay amounts to an endorsement of correlational analysis of survey data just so long as Whorfian socioeconomists cluster their standard errors or estimate language-specific random effects. But these simple statistical fixes do not address other methodological objections to feature-value correlations, including problems of omitted-variable bias/unobserved confounding or linguistic and cultural relatedness. Roberts et al. (2015) offer a promising approach to address the latter; the former can only be addressed in the context of a specific Whorfian socioeconomics hypothesis. Nevertheless, some attempt to address the clustered nature of survey and language data is necessary in any case, even if other methodological concerns are being addressed as well. And an important feature of the methodological adjustments proposed here is that they require no data other than those that are already available in the data set. Proponents of the emerging field of Whorfian socioeconomics must be particularly attentive to the unobserved factors that might explain patterns of survey responses and that happen to correlate with the languages respondents speak even if they do not address other concerns about linguistic and cultural relatedness. Similarly, the statistical corrections discussed here do not address the issue of crossnational public opinion data being nonrepresentative relative to the full diversity of human languages. Internally and externally valid claims made from feature-value correlations rely on accurate linkages between languages and human communities.

Collectively, my simulation analyses show that both two-way clustered standard errors and multilevel modeling greatly outperform naive approaches to modeling feature-value correlations that do not take into account grouping structure of the data. But how should researchers choose between these two statistical alternatives? In the simulations presented in Appendix C, multilevel modeling clearly outperforms clustering, with the former achieving the appropriate nominal rejection rates and the latter still too overconfident (although only by a modest amount). The ability of multilevel models to explicitly account for crosslinguistic variation provides another attractive argument in their favor.

There are two caveats, however. The first is that the simulation results presented in Appendix C correspond to exactly the data-generating process for which multilevel models are the appropriate solution, which 'stacks the deck' in their favor. It is noteworthy, in this regard, how well two-way clustering performs even though it makes no effort to model language variation at all. The second is that multilevel models require assumptions—specifically, that the random intercepts by language are drawn from a distribution with mean zero and finite variance—that may not be attractive in all contexts. Instead of universally endorsing multilevel models over two-way clustering, then, I recommend that researchers weigh the benefits of explicitly modeling crosslinguistic variation versus the costs of the untestable assumptions multilevel models impose. I also emphasize that because these are purely statistical fixes, there is no reason not to explore the robustness of results to both approaches.

More broadly, the methodological discussions in this essay reveal a disconnect between standard approaches in linguistics and the psychological sciences, and in applied microeconomics and related fields of the social sciences. The latest innovations in the statistical analysis of clustered observational data should be of interest to linguists. So, too, should best practices for multilevel modeling as employed in linguistics and sociology be of interest to aspiring Whorfian socioeconomists. A shared methodological approach would enable more constructive empirical work in this emerging interdisciplinary field of inquiry, especially given many linguists' reservations about the intellectual project's ambitious agenda.

APPENDIX A: ADDITIONAL TABLES AND FIGURE

	VARIABLE	VALUE 1	VALUE 2	N_1	N_2
1	Alignment of verbal person marking	Accusative	Other	212	168
2	Expression of pronominal subjects	Other	Subject affixes on verb	274	437
3	Verbal person marking	No person marking	Other	82	296
4	Order of person markers on the verb	A and P do not or do not both occur on the verb	Other	187	192
5	Ditransitive constructions: the verb 'give'	Indirect object construction	Other	189	189
6	Reciprocal constructions	Distinct from reflexive	Other	99	76
7	Passive constructions	Absent	Present	211	162
8	Antipassive constructions	No antipassive	Other	146	48
9	Productivity of the antipassive construction	No antipassive	Other	146	40
10	Applicative constructions	Applicative	No applicative construction	83	100
11	Vowel nasalization	Contrast absent	Contrast present	180	64
12	Periphrastic causative constructions	Other	Purposive but no sequential	50	68
13	Nonperiphrastic causative constructions	Morphological but no compound	Other	254	56
14	Negative morphemes	Negative affix	Other	395	762
15	Symmetric and asymmetric standard negation	Other	Symmetric	183	114
16	Negative indefinite pronouns and predicate negation	Other	Predicate negation also present	36	170
17	Polar questions	Other	Question particle	370	585
18	Predicative possession	'have'	Other	63	177
19	Predicative adjectives	Nonverbal encoding	Other	132	254
20	Nominal and locational predication	Different	Identical	269	117
21	Zero copula for predicate nominals	Impossible	Possible	211	175
22	Comparative constructions	Other	Particle	145	22
23	Relativization on subjects	Other	Relative pronoun	154	12
24	Relativization on obliques	Other	Relative pronoun	99	13
25	X 'want' complement subjects	Other	Subject is left implicit	139	144
26	Purpose clauses	Deranked	Other	102	68
27	X 'when' clauses	Deranked	Other	51	123
28	Reason clauses	Balanced	Other	90	79
29	Utterance complement clauses	Balanced	Other	114	29
30	Hand and arm	Different	Identical	389	228
31	Numeral bases	Decimal	Other	125	71
32	Number of basic color categories	Eleven	Other	11	108
33	*M-T* pronouns	No *M-T* pronouns	Exist	30	200
34	*m* in first-person singular	*m* in first-person singular	No *m* in first-person singular	53	177
35	*m* in second-person singular	*m* in second-person singular	No *m* in second-person singular	78	152
36	Tea	Other	Words derived from Min Nan Chinese *te*	146	84
37	Tone	No tones	Other	307	220
38	Paralinguistic usages of clicks	Affective meanings	Other	71	72
39	Order of negative morpheme and verb	[V-Neg]	Other	202	1,122
40	Preverbal negative morphemes	NegV	Other	681	643
41	Postverbal negative morphemes	None	Other	711	613

(TABLE A1. *Continues*)

	VARIABLE	VALUE 1	VALUE 2	N_1	N_2
42	Fixed stress locations	No fixed stress	Other	220	282
43	Weight-sensitive stress	Fixed stress (no weight-sensitivity)	Other	281	219
44	Weight factors in weight-sensitive stress systems	Combined	Other	42	458
45	Rhythm types	Other	Trochaic	170	153
46	Absence of common consonants	All present	Other	503	64
47	Presence of uncommon consonants	'th' sounds	Other	40	527
48	Consonant inventories	Moderately large or large	Other	151	412
49	Fusion of selected inflectional formatives	Exclusively concatenative	Other	125	40
50	Exponence of selected inflectional formatives	Monoexponential case	Other	71	91
51	Exponence of tense-aspect-mood inflation	Monoexponential TAM	Other	127	33
52	Inflectional synthesis of the verb	0–1 category per word	Other	5	140
53	Locus of marking in the clause	Dependent marking	Other	63	173
54	Locus of marking in possessive noun phrases	Dependent marking	Other	98	138
55	Locus of marking: whole language typology	Dependent marking	Other	46	190
56	Zero-marking of A and P arguments	Non-zero-marking	Zero-marking	219	16
57	Prefixing vs. suffixing in inflectional morphology	Other	Strongly suffixing	563	406
58	Reduplication	No productive reduplication	Other	55	313
59	Case syncretism	Core and noncore	Other	22	176
60	Syncretism in verbal person-number marking	Other	Syncretic	138	60
61	Vowel-quality inventories	Large (7–14)	Other	184	380
62	Number of genders	None	Other	145	112
63	Sex-based and non-sex-based gender systems	Other	Sex-based	173	84
64	Systems of gender assignment	Other	Semantic and formal	198	59
65	Coding of nominal plurality	Other	Plural suffix	553	513
66	Occurrence of nominal plurality	Only human nouns, optional	Other	20	271
67	Plurality in independent personal pronouns	Other	Person-number stem	147	114
68	The associative plural	No associative plural	Other	37	199
69	Definite articles	Definite word distinct from demonstrative	Other	216	404
70	Indefinite articles	Indefinite word distinct from 'one'	Other	102	432
71	Inclusive/exclusive distinction in independent pronouns	Inclusive/exclusive	Other	63	137
72	Consonant-vowel ratio	Moderately high or high	Other	171	393
73	Inclusive/exclusive distinction in verbal inflection	'we' the same as 'I'	Other	12	188
74	Distance contrasts in demonstratives	No distance contrast	Other	7	227
75	Pronominal and adnominal demonstratives	Identical	Other	143	58

(TABLE A1. *Continues*)

	VARIABLE	VALUE 1	VALUE 2	N_1	N_2
76	Third-person pronouns and demonstratives	Other	Unrelated	125	100
77	Gender distinctions in independent personal pronouns	Third-person singular only	Other	61	317
78	Politeness distinctions in pronouns	Exists	No politeness distinction	71	136
79	Indefinite pronouns	Generic-noun-based	Other	85	241
80	Intensifiers and reflexive pronouns	Differentiated	Identical	74	94
81	Person marking on adpositions	No person marking	Other	209	169
82	Number of cases	No morphological case marking	Other	100	161
83	Voicing in plosives and fricatives	In both plosives and fricatives	Other	158	409
84	Asymmetrical case marking	Additive-quantitatively asymmetrical	Other	53	208
85	Position of case affixes	No case affixes or adpositional clitics	Other	379	652
86	Comitatives and instrumentals	Differentiation	Other	213	109
87	Ordinal numerals	First, second, three-th	Other	61	260
88	Distributive numerals	No distributive numerals	Other	62	189
89	Numeral classifiers	Absent	Exist	260	140
90	Conjunctions and universal	Formally different quantifiers	Other	40	76
91	Position of pronominal possessive affixes	Exist	No possessive affixes	642	260
92	Possessive classification	Exists	No possessive classification	118	125
93	Voicing and gaps in plosive systems	None missing in /p t k b d g/	Other	255	312
94	Genitives, adjectives, and relative clauses	Highly differentiated	Other	77	61
95	Adjectives without nouns	Other	Without marking	51	73
96	Action nominal constructions	Ergative-possessive	Other	21	147
97	Noun phrase conjunction	‘and’ identical to ‘with’	Other	103	131
98	Nominal and verbal conjunction	Differentiation	Other	125	176
99	Perfective/imperfective aspect	Grammatical marking	No grammatical marking	101	121
100	The past tense	Exists	No past tense	134	88
101	The future tense	Inflectional future exists	No inflectional future	110	112
102	The perfect	No perfect	Other	114	108
103	Position of tense-aspect affixes	Other	Tense-aspect suffixes	464	667
104	Uvular consonants	Exist	None	97	470
105	The morphological imperative	Exists	No second-person imperatives	425	122
106	The prohibitive	Normal imperative + normal negative	Other	113	382
107	Imperative hortative systems	Neither type of system	Some system	201	174
108	The optative	Inflectional optative absent	Inflectional optative present	271	48
109	Situational possibility	Other	Verbal constructions	76	158
110	Epistemic possibility	Other	Verbal constructions	175	65
111	Overlap between situational and epistemic modal marking	Other	Overlap for both possibility and necessity	171	36
112	Semantic distinctions of evidentiality	Exist	No grammatical evidentials	237	181
113	Coding of evidentiality	Modal morpheme	Other	7	411

(TABLE A1. *Continues*)

	VARIABLE	VALUE 1	VALUE 2	N_1	N_2
114	Suppletion according to tense and aspect	Other	Tense and aspect	169	24
115	Suppletion in imperatives and hortatives	Imperative	Other	29	164
116	Glottalized consonants	Exist	No glottalized consonants	158	409
117	Verbal number and suppletion	Exist	None	34	159
118	Order of subject, object, and verb	Other	SVO	889	488
119	Order of subject and verb	Other	SV	304	1,193
120	Order of object and verb	Other	VO	814	705
121	Order of object, oblique, and verb	Other	VOX	290	210
122	Order of adposition and noun phrase	Other	Postpositions	607	576
123	Order of genitive and noun	Noun-Genitive	Other	468	781
124	Order of adjective and noun	Adjective-Noun	Other	373	993
125	Order of demonstrative and noun	Demonstrative-Noun	Other	542	682
126	Order of numeral and noun	Numeral-Noun	Other	479	674
127	Lateral consonants	/l/, no obstruent laterals	Other	388	179
128	Order of relative clause and noun	Noun-Relative clause	Other	579	245
129	Order of degree word and adjective	Degree word-Adjective	Other	227	254
130	Position of polar question particles	Exist	No question particle	529	355
131	Position of interrogative phrases in content questions	Initial interrogative phrase	Other	264	638
132	Order of adverbial subordinator and clause	Initial subordinator word	Other	398	261
133	Relationship between the order of object and verb and the order of adposition and noun phrase	Other	VO and prepositions	686	456
134	Relationship between the order of object and verb and the order of relative clause and noun	Other	VO and NRel	463	416
135	Relationship between the order of object and verb and the order of adjective and noun	Other	VO and NAdj	860	456
136	Alignment of case marking of full noun phrases	Nominative-accusative (standard)	Other	46	144
137	Alignment of case marking of pronouns	Nominative-accusative (standard)	Other	61	111
138	The velar nasal	Exists	No velar nasal	234	235

TABLE A1. *WALS* variables. This table describes the variables created from the *World atlas of language structures online* (Dryer & Haspelmath 2013) and gives the number of languages in the database with each feature.

	LANGUAGE (WVS NAME)	FREQUENCY	LANGUAGE (*WALS* NAME)
1	Afar	2	Qafar
2	Afrikaans	2,677	Afrikaans
3	Albanian	2,529	Albanian
4	Amharic	670	Amharic

(TABLE A2. *Continues*)

	LANGUAGE (WVS NAME)	FREQUENCY	LANGUAGE (*WALS* NAME)
5	Arabic	30,453	Arabic (Modern Standard)
6	Armenian	3,124	Armenian (Eastern)
7	Assamese	64	Assamese
8	Ateso	71	Teso
9	AU: Arakanese	1	Arakanese (Marma)
10	Avarian	8	Avar
11	Avaric	9	Avar
12	Aymara	21	Aymara (Central)
13	Azari	967	Azari (Iranian)
14	Azerbaijani	2,646	Azerbaijani
15	Bajau	18	Bajau (Sama)
16	Balkarian	5	Karachay-Balkar
17	Baluchi	132	Baluchi
18	Bamanakan	7	Bambara
19	Bambara	1,131	Bambara
20	Barahvi	52	Brahui
21	Basque	42	Basque
22	Belarusian	446	Belorussian
23	Bemba	543	Bemba
24	Bengali	2,034	Bengali
25	Berber	476	Berber (Ayt Seghrouchen Middle Atlas)
26	Berto	3	Berta
27	Bosnian	1,407	Bosnian
28	Bulgarian	1,872	Bulgarian
29	Bussa	4	Busa
30	Cantonese	943	Cantonese
31	Catalan; Valencian	816	Catalan
32	Cebuano	309	Cebuano
33	Chewa	12	Chichewa
34	Chinese	4,438	Mandarin
35	Croatian	323	Serbian-Croatian
36	Czech	1,124	Czech
37	Dagbani	250	Dagbani
38	Dioula	191	Diola-Fogny
39	Dutch; Flemish	2,773	Dutch
40	DZ: Amazigh	180	Berber (Chaouia)
41	Efik	23	Efik
42	English	23,026	English
43	Estonian	1,639	Estonian
44	Ewe	160	Ewe
45	Filipino; Pilipino	533	Tagalog
46	Finnish	1,020	Finnish
47	Foulfoulde	83	Fulfulde (Maasina)
48	French	2,873	French
49	Ga	184	Gã
50	Gagauz (Turkish Orthodox)	41	Gagauz
51	Gallegan	179	Galician
52	Gamo	29	Gamo
53	Georgian	4,200	Georgian
54	German	4,604	German
55	GH: Bimoba	1	Bimoba
56	GH: Dagaaba	1	Dagaare
57	GH: Dagaare	3	Dagaare
58	GH: Dagaati	4	Dagaare
59	GH: Ewe	167	Ewe
60	GH: Mampruli	15	Mampruli
61	GH: Moor	3	Mooré
62	GH: Sisala	2	Sisaala

(TABLE A2. *Continues*)

	LANGUAGE (WVS NAME)	FREQUENCY	LANGUAGE (*WALS* NAME)
63	Gilaki	77	Gilaki
64	Greek	1,085	Greek (Modern)
65	Gujarati	188	Gujarati
66	Haitian; Haitian Creole	1,974	Haitian Creole
67	Hakka	18	Hakka
68	Harari	2	Harari
69	Hausa	1,452	Hausa
70	Hungarian	1,967	Hungarian
71	Iban	36	Iban
72	Ibibio	71	Ibibio
73	Igbo	1,192	Igbo
74	IN: Assamese	3	Assamese
75	IN: Awadhi	139	Awadhi
76	IN: Bengali	255	Bengali
77	IN: Bhili	3	Bhili
78	IN: Bhojpuri	173	Bhojpuri
79	IN: Dogri	1	Dogri
80	IN: Gondi	15	Gondi
81	IN: Gujarati	251	Gujarati
82	IN: Hindi	973	Hindi
83	IN: Kannada	144	Kannada
84	IN: Lakher	1	Mara
85	IN: Magadhi	119	Magahi
86	IN: Maithili	65	Maithili
87	IN: Malayalam	192	Malayalam
88	IN: Marathi	250	Marathi
89	IN: Nepali	6	Nepali
90	IN: Oriya	314	Oriya
91	IN: Tamil	17	Tamil
92	IN: Telugu	261	Telugu
93	IN: Urdu	126	Urdu
94	India: Hindi/Hindu	2,207	Hindi
95	India: Marathi	448	Marathi
96	India: Oriya	267	Oriya
97	India: Telegu	432	Telugu
98	Indonesian	751	Indonesian
99	Iranian	2	Persian
100	IT: Italian	630	Italian
101	Italian	258	Italian
102	Japanese	4,911	Japanese
103	Javanese	1,407	Javanese
104	Jewish	2	Hebrew (Modern)
105	Kabardian	35	Kabardian
106	Kadazan	12	Kadazan
107	Kalanga	7	Kalanga
108	Kalmyk	5	Kalmyk
109	Kankana-ay	2	Kankanay
110	Kannada	320	Kannada
111	Karakalpak	38	Karakalpak
112	Kashmiri	6	Kashmiri
113	Kazah	725	Kazakh
114	Kelabit	3	Kelabit
115	Khmer	2	Khmer
116	Kinyarwanda; Rwandese	2,982	Kinyarwanda
117	Kirghiz	1,646	Kirghiz
118	Komi	21	Komi-Permyak
119	Konkani	4	Konkani
120	Korean	5	Korean

(TABLE A2. *Continues*)

	LANGUAGE (WVS NAME)	FREQUENCY	LANGUAGE (*WALS* NAME)
121	Kurdish	82	Kurdish (Central)
122	Kurdish/Esid	1,006	Kurdish (Central)
123	Kurtce	11	Kurdish (Central)
124	Lampung	3	Lampung
125	Lao	1	Lao
126	Latvian	703	Latvian
127	Lazca	1	Laz
128	Lezgin; Lezghian	77	Lezgian
129	Lithuanian	881	Lithuanian
130	Lozi	148	Lozi
131	Luganda	280	Luganda
132	Lunda	9	Lunda
133	Luri/Lori	108	Luri
134	Luvalo	26	Luvale
135	Lwo	116	Luwo
136	LY: Tamazight	90	Berber (Siwa)
137	Macedonian	1,531	Macedonian
138	Malay	2,196	Malay
139	Malayalam	209	Malayalam
140	Malinke	49	Maninka
141	Maltese	3	Maltese
142	Mandarin	2,378	Mandarin
143	Maori	3	Maori
144	Melanau	5	Melanau
145	Moldavian	743	Moldavian
146	Mordovian	6	Mordvin (Erzya)
147	Moroccan	3	Arabic (Moroccan)
148	Ndebele	283	Ndebele (in South Africa)
149	Nepali	15	Nepali
150	NG: Efik	17	Efik
151	NG: Fulani	15	Fula (Nigerian)
152	NG: Gwari	2	Gwari
153	NG: Ibibio	6	Ibibio
154	NG: Idoma	6	Idoma
155	NG: Isoko	1	Isoko
156	NG: Kilba	1	Kilba
157	NG: Tiv	8	Tiv
158	NG: Urhobo	1	Urhobo
159	Northern Sotho, Pedi; Sepedi	836	Sotho (Northern)
160	Norwegian	988	Norwegian
161	Nsenga	31	Nsenga
162	Oromo	304	Oromo (Boraana)
163	Ossetian; Ossetic	2	Ossetic
164	Persian	1,730	Persian
165	Peul	86	Fula (Senegal)
166	PH: Aklanon	6	Aklanon
167	PH: Cagayan (mapun)	4	Phlong
168	PH: Chavacano	15	Chavacano
169	PH: Hiligaynon	127	Hiligaynon
170	PH: Iluko	45	Ilocano
171	PH: Kapampangan	11	Kapampangan
172	PH: Maguindanaon	9	Magindanao
173	PH: Manobo	2	Manobo (Western Bukidnon)
174	PH: Maranao	41	Maranao
175	PH: Pangasinense	8	Pangasinan
176	PH: Tausug	3	Tausug
177	PH: Tiduray	3	Piapoco
178	PH: Waray	69	Waray-Waray

(TABLE A2. *Continues*)

	LANGUAGE (WVS NAME)	FREQUENCY	LANGUAGE (*WALS* NAME)
179	Polish	1,978	Polish
180	Portuguese	2,682	Portuguese
181	Pushto	545	Pashto
182	Putonghua	24	Mandarin
183	Quechua	167	Quechan
184	Romanian	5,760	Romanian
185	Romany/Gypsi	109	Romani (Ajia Varvara)
186	Russian	15,170	Russian
187	Saho	4	Saho
188	Serbian	5,117	Serbian-Croatian
189	Serbo-Croatian	104	Serbian-Croatian
190	Shona	1,205	Shona
191	Sidama	151	Sidaama
192	Sindhi	356	Sindhi
193	Sinhala, Sinhalese	2	Sinhala
194	Slovak	1,011	Slovak
195	Slovenian	2,997	Slovene
196	Soddo	19	Soddo
197	Soninke	34	Soninke
198	Sotho, Southern, Sesotho	942	Sesotho
199	Spanish; Castilian	32,226	Spanish
200	Sundanese	510	Sundanese
201	Suomea	945	Finnish
202	Swahili; Kiswahili	978	Swahili
203	Swazi	128	Swati
204	Swedish	3,038	Swedish
205	Tadjic	121	Tajik
206	Tagalog	893	Tagalog
207	Taiwanese	1,243	Taiwanese
208	Talish	111	Talysh (Azerbaijan)
209	Tamil	813	Tamil
210	Tatar	144	Tatar
211	Thai: Central	1,201	Thai
212	Tigrinya	149	Tigré
213	Tiv	40	Tiv
214	Tokaleya	12	Tokelauan
215	Tsonga/Shangaan	473	Tsonga
216	Tswana	967	Tswana
217	Turkish	8,837	Turkish
218	Turkmen	16	Turkmen
219	TW: Hakka	37	Hakka
220	Twi	697	Akan
221	Twi (Akan)	941	Akan
222	Uigur	5	Uyghur
223	Ukrainian	2,828	Ukrainian
224	Urdu	1,012	Urdu
225	Uyghur	4	Uyghur
226	Uzbek	1,700	Uzbek
227	Venda	142	Venda
228	Vietnamese/Kiinaa	2,399	Vietnamese
229	Western Frisian	38	Frisian (Western)
230	Wolaita	36	Wolaytta
231	Xhosa	1,730	Xhosa
232	Yoruba	1,389	Yoruba
233	Zaza	29	Zazaki
234	Zulu	2,440	Zulu

TABLE A2. Languages and speakers. This table gives the number of speakers of each language in the World Values Survey, as defined by variable S016 'Language spoken at home'. The corresponding language in the *World atlas of language structures online* is also provided.

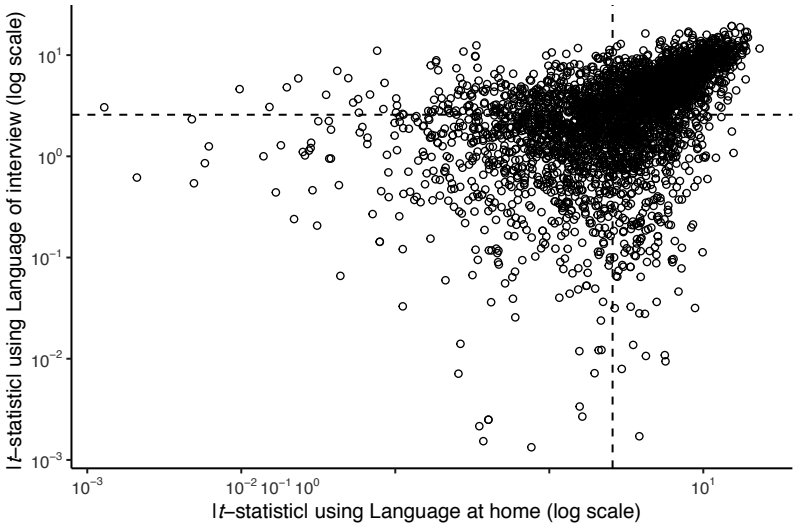


FIGURE A1. Comparing results using language at home and language of interview. The dashed lines at t = 2.576 correspond to $p < 0.01$ in a two-tailed test. The lower left quadrant includes all results where $p > 0.01$ using both methods. The upper right quadrant includes all results where $p < 0.01$ using both methods. The upper left and lower right quadrants contain results that are statistically significant at $p < 0.01$ in one but not the other.

APPENDIX B: CLUSTER-ROBUST STANDARD ERRORS

Consider a simplified version of equation 1 as $Y_{ic} = \beta X_{ic} + \epsilon_{ic}$, where i indexes survey respondents and c indexes countries. The variance of $\hat{\beta}$ can be expressed as in equation A1.

$$(A1) \quad V(\hat{\beta}) = \frac{\sum x_i^2 E[\epsilon_i]^2}{(\sum x_i^2)^2}$$

With homoscedastic errors, this reduces to $V(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$. Without that assumption, one must find an alternative method to estimate the numerator of equation A1. White (1980) showed that in large samples, this can be estimated using A2:

$$(A2) \quad \hat{V}(\hat{\beta}) = \frac{\sum x_i^2 e_i^2}{(\sum x_i^2)^2}$$

where e_i is simply the regression error obtained from $e_i = Y_i - \hat{\beta} X_i$.

Now consider a grouping of survey respondents i across countries c, and the assumption that errors are correlated within countries, but not across countries. This implies that $E(\epsilon_i, \epsilon_j) = 0$ for all observations i and j that are not in the same country, $E(\epsilon_i, \epsilon_j) \neq 0$ if they are. Cameron and Miller (2015:320–21) show that one can express the corresponding estimate of the variance of $\hat{\beta}$ using equation A3.

$$(A3) \quad \hat{V}(\hat{\beta}) = \frac{\sum\sum x_i x_j e_i e_j \times 1[i, j \text{ in the sme country}]}{(\sum x_i^2)^2}$$

This expression adjusts the estimate of variance of β to reflect the clustering of errors of an arbitrary form for respondents WITHIN countries.

Cameron et al. (2011:241) show that this logic extends to clustering across multiple dimensions. In the model $Y_{ict} = \beta X_{ict} + \epsilon_{ict}$, one may allow that errors are correlated for respondents within countries c and also for respondents within years t. One may construct the 'two-way clustered' estimate of the variance of $\hat{\beta}$ using equation A4:

$$(A4) \quad \hat{V}(\hat{\beta}) = \hat{V}^c(\hat{\beta}) + \hat{V}^t(\hat{\beta}) - \hat{V}^{c \cap t}(\hat{\beta}),$$

where each variance element on the right-hand side of the equation is estimated using the method in equation A3. They further establish the equivalence of this result for nonlinear models such as logistic regression (Cameron et al. 2011:242–43).

Appendix C: Comparisons with known null effects and known positive effects

In this section, I provide context for the simulation analysis in the main text, showing that the two methods I use correctly control false positive rates in feature-value correlations while not simultaneously increasing the rate of false negatives. I do this first by studying the performance of these methods when the true effect of a linguistic feature is known to be zero, and then by studying how these methods fare in the context of well-established sociodemographic correlations that are almost certainly present.

C1. Do more conservative tests correctly identify known null effects? Because we do not know the true proportion of feature-value correlations that actually are statistically insignificant, it is important to benchmark these methods' performance using data structures in which we know that the true effect of a language variable is zero. To do this, I adopt a simulation approach again, but this time rather than using actual features recorded in the *WALS* data, I randomly generate hypothetical linguistic features whose effect on all outcomes is known to be zero, and then evaluate the performance of these methods in estimating their effects.

Specifically, I take the WVS data set that was analyzed in the main text and extract the list of languages. I then randomly assign half of those languages to possess a linguistic feature or not, and merge that randomly generated variable back into the data set. Because I have randomly generated that linguistic feature myself, we know that it cannot have any relationship with any outcome variable. I then randomly select an outcome variable from the list of those analyzed and estimate the 'effect' of the randomly generated linguistic feature on the randomly selected dependent variable using the standard OLS regression approach, the two-way clustered standard errors approach, and the multilevel modeling approach. Repeating this process 250 times produces 750 simulated estimates of the effect of a variable whose true effect is known to be zero.

Figure A2 shows the distribution of the absolute value of the *t*-statistics for each of the three methods. The black density plot shows that the standard approach generates highly statistically significant results even when the true effect of the linguistic feature is known to be zero. Of the estimates, 61.4% are statistically significant at the $p < 0.01$ level, a result that is comparable to the findings in Figs. 2 and 3 in the main text. By contrast, the two gray density plots show that accounting for the clustered nature of the data produces estimates that are closer to what we would expect when the true effect of the randomly generated linguistic feature is known to be zero. In the two-way clustered standard errors approach, only 4.8% of estimates are statistically significant at the $p < 0.01$ level, and in the multilevel modeling approach only 0.8% of estimates are statistically significant at the $p < 0.01$ level.

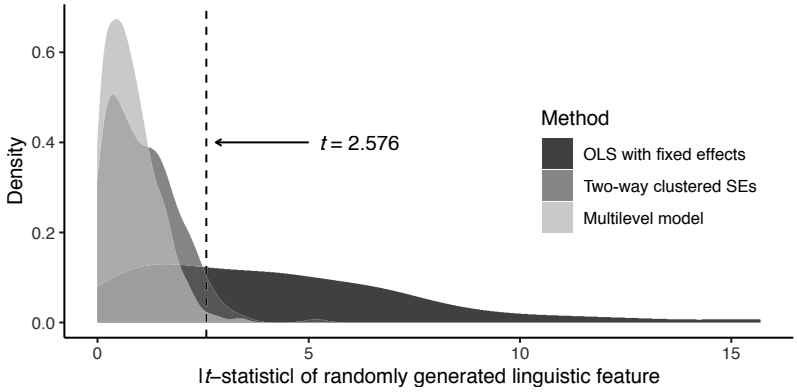


Figure A2. Comparing results for known null effects. The dashed line at $t = 2.576$ corresponds to $p < 0.01$ in a two-tailed test.

C2. Do more conservative tests overturn known positive associations? The other possibility is that the statistical approaches I have advocated here are too conservative, producing type II errors in which a test erroneously fails to reject a null hypothesis of no effect when such an effect does exist. I explore this using three intuitively plausible and empirically well-documented demographic associations: between employment status and household savings, between social class and self-positioning on a left-right ideological scale, and between income and preferences for income equality. For each of these analyses, I estimate the same statistical model as in the simulation analyses presented in the main text: this baseline statistical model

includes a range of demographic covariates as well as country and year fixed effects, but makes no effort to account for the clustered nature of the data.

First, I examine the relationship between employment status and household savings. In Table A3, I display the regression coefficients for seven occupation status choices, evaluated relative to the baseline category of 'employed full-time'. Column 1 contains the baseline OLS regression model, column 2 estimates two-way clustered standard errors, and column 3 estimates the multilevel model with random effects by language.

	OLS	TWO-WAY CLUSTERED *SE*S	MULTILEVEL MODEL
Housewife	−0.055 ***	−0.055 ***	−0.052 ***
	(0.008)	(0.012)	(0.008)
Other	−0.093 ***	−0.093 ***	−0.091 ***
	(0.016)	(0.026)	(0.016)
Part-time	−0.025 **	−0.025	−0.025 **
	(0.009)	(0.013)	(0.009)
Retired	−0.017	−0.017	−0.017
	(0.009)	(0.017)	(0.009)
Self-employed	0.011	0.011	0.010
	(0.008)	(0.012)	(0.008)
Student	−0.030 **	−0.030	−0.032 **
	(0.011)	(0.027)	(0.011)
Unemployed	−0.086 ***	−0.086 ***	−0.087 ***
	(0.009)	(0.022)	(0.009)
OBSERVATIONS	*121,764*		

TABLE A3. Employment and household savings. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

As anticipated, these adjustments generally have no substantial effects on the well-known finding that unemployed people have less savings than those employed full-time. They similarly have no effects on the coefficients for women who work inside the home, or for those whose employment status is listed as 'Other'. Clustering standard errors by both country and year DOES affect inferences about students, part-time employed, and retirees; the most substantial differences are for students. But this is appropriate: there is no well-established literature that documents a consistently lower household savings among students, controlling for family income (as these models do).

In Table A4 I examine the relationship between social class and self-positioning on a left-right ideological scale in which higher values correspond to more right-leaning individuals. Here the omitted category is 'lower class', so each comparison is made relative to those respondents with the lowest subjective class position. The three models are, as before, the baseline OLS regression, the two-way clustered standard errors model, and the multilevel model.

In this example, all substantive inferences about class and left-right ideology are unchanged when adopting the more conservative statistical approaches. In all models, the higher the respondent's subjective social class, the further to the right they report themselves to be.

	OLS	TWO-WAY CLUSTERED *SE*S	MULTILEVEL MODEL
Lower middle class	0.175 ***	0.175 ***	0.194 ***
	(0.028)	(0.051)	(0.028)
Upper class	0.510 ***	0.510 ***	0.531 ***
	(0.061)	(0.115)	(0.061)
Upper middle class	0.338 ***	0.338 ***	0.365 ***
	(0.031)	(0.097)	(0.031)
Working class	−0.014	−0.014	0.009
	(0.028)	(0.041)	(0.028)
OBSERVATIONS	*99,864*		

TABLE A4. Social class and left-right ideological positioning. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Finally, in Table A5 I study the relationship between income and preferences for income equality, where higher values correspond to LOWER support for income equality. Each independent variable is a dummy for a respondent's country-specific income level, with level 10 the highest income level and level 1 the lowest income level, serving as the omitted reference category. The three models are as before.

We see clear evidence once again that all substantive inferences about income and equality are unchanged when adopting the more conservative statistical approaches. In all models, the higher the respondent's income, the less they support income equality, and this difference is substantively larger and more statistically significant as incomes levels rise relative to the baseline.[13]

	OLS		TWO-WAY CLUSTERED *SE*s		MULTILEVEL MODEL	
Level 2	0.025		0.025		0.028	
	(0.036)		(0.055)		(0.036)	
Level 3	0.037		0.037		0.042	
	(0.035)		(0.078)		(0.035)	
Level 4	0.070		0.070		0.084	*
	(0.036)		(0.083)		(0.036)	
Level 5	0.166	***	0.166	*	0.179	***
	(0.035)		(0.081)		(0.035)	
Level 6	0.292	***	0.292	**	0.304	***
	(0.038)		(0.093)		(0.038)	
Level 7	0.418	***	0.418	***	0.429	***
	(0.040)		(0.097)		(0.040)	
Level 8	0.517	***	0.517	***	0.536	***
	(0.044)		(0.119)		(0.044)	
Level 9	0.609	***	0.609	***	0.632	***
	(0.054)		(0.110)		(0.054)	
Level 10	0.846	***	0.846	***	0.865	***
	(0.058)		(0.128)		(0.058)	
OBSERVATIONS	*125,268*					

TABLE A5. Income and preferences for income equality. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

REFERENCES

ABADIE, ALBERTO; SUSAN ATHEY; GUIDO W. IMBENS; and JEFFREY WOOLDRIDGE. 2017. When should you adjust standard errors for clustering? Working paper 24003. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w24003.

ANGRIST, JOSHUA D., and JÖRN-STEFFEN PISCHKE. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

ASHRAF, QUAMRUL H., and ODED GALOR. 2018. The macrogenoeconomics of comparative development. *Journal of Economic Literature* 56.1119–55. DOI: 10.1257/jel.20161314.

BAAYEN, R. HARALD; DOUGLAS J. DAVIDSON; and DOUGLAS M. BATES. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412. DOI: 10.1016/j.jml.2007.12.005.

BARR, DALE J.; ROGER LEVY; CHRISTOPH SCHEEPERS; and HARRY J. TILY. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.255–78. DOI: 10.1016/j.jml.2012.11.001.

BOAS, FRANZ. 1931. *The mind of primitive man*. New York: Macmillan.

BORODITSKY, LERA; LAUREN A. SCHMIDT; and WEBB PHILLIPS. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, ed. by Dedre Gentner and Susan Goldin-Meadow, 61–80. Cambridge, MA: MIT Press.

BULLOCK, JOHN G.; DONALD P. GREEN; and SHANG E. HA. 2010. Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology* 98.550–58. DOI: 10.1037/a0018933.

CAMERON, A. COLIN; JONAH B. GELBACH; and DOUGLAS L. MILLER. 2008. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics* 90.414–27. DOI: 10.1162/rest.90.3.414.

CAMERON, A. COLIN; JONAH B. GELBACH; and DOUGLAS L. MILLER. 2011. Robust inference with multiway clustering. *Journal of Business & Economic Statistics* 29.238–49. DOI: 10.1198/jbes.2010.07136.

[13] All results here generalize to cases with multiple explanatory variables.

Cameron, A. Colin, and Douglas L. Miller. 2015. A practitioner's guide to cluster-robust inference. *The Journal of Human Resources* 50.317–72. DOI: 10.3368/jhr.50.2.317.

Carroll, Mary; Christiane von Stutterheim; and Ralf Nuese. 2004. The *language and thought* debate: A psycholinguistic approach. *Multidisciplinary approaches to language production*, ed. by Thomas Pechmann and Christopher Habel, 183–218. Berlin: De Gruyter. DOI: 10.1515/9783110894028.183.

Chen, M. Keith. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review* 103.690–731. DOI: 10.1257/aer.103.2.690.

Chen, Shimin; Henrik Cronqvist; Serene Ni; and Frank Zhang. 2017. Languages and corporate savings behavior. *Journal of Corporate Finance* 46.320–41. DOI: 10.1016/j.jcorpfin.2017.07.009.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12.335–59. DOI: 10.1016/S0022-5371(73)80014-3.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.

Davis, Lewis S., and Farangis Abdurazokzoda. 2016. Language, culture and institutions: Evidence from a new linguistic dataset. *Journal of Comparative Economics* 44.541–61. DOI: 10.1016/j.jce.2015.10.015.

Diamond, Jared. 1997. *Guns, germs, and steel: The fates of human societies*. New York: W. W. Norton.

Dryer, Matthew S., and Martin Haspelmath (eds.) 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://wals.info.

Everett, Caleb. 2013. *Linguistic relativity: Evidence across languages and cognitive domains*. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110308143.

Feldmann, Horst. 2019. Do linguistic structures affect human capital? The case of pronoun drop. *Kyklos* 72.29–54. DOI: 10.1111/kykl.12190.

Galor, Oded; Ömer Özak; and Assaf Sarid. 2016. Geographical origins and economic consequences of language structures. CESifo working paper 6149. Munich: Center for Economic Studies & Ifo Institute. DOI: 10.2139/ssrn.2877619.

Gay, Victor; Daniel L. Hicks; Estefania Santacreu-Vasut; and Amir Shoham. 2018. Decomposing culture: An analysis of gender, language, and labor supply in the household. *Review of Economics of the Household* 16.879–909. DOI: 10.1007/s11150-017-9369-x.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gerring, John. 2010. Causal mechanisms: Yes, but … . *Comparative Political Studies* 43.1499–1526. DOI: 10.1177/0010414010376911.

Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10.95–125. DOI: 10.3366/cor.2015.0068.

Hicks, Daniel L.; Estefania Santacreu-Vasut; and Amir Shoham. 2015. Does mother tongue make for women's work? Linguistics, household labor, and gender identity. *Journal of Economic Behavior & Organization* 110.19–44. DOI: 10.1016/j.jebo.2014.11.010.

Humphreys, Macartan; Raul Sanchez de la Sierra; and Peter van der Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21.1–20. DOI: 10.1093/pan/mps021.

Inglehart, Ronald; Christian Haerpfer; Alejandro Moreno; Christian Welzel; Kseniya Kizilova; Jaime Diez-Medrano; Marta Lagos; Pippa Norris; Eduard Ponarin; and Bi Puranen (eds.) 2014. World Values Survey: All rounds—country-pooled datafile version. Madrid: JD Systems Institute. Online: https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp.

Ioannidis, John P. A. 2005. Why most published research findings are false. *PLOS Medicine* 2:e124. DOI: 10.1371/journal.pmed.0020124.

Jakiela, Pamela, and Owen Ozier. 2018. Gendered language. Policy research working paper 8464. Washington, DC: World Bank. DOI: 10.1596/1813-9450-8464.
Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3.359–83. DOI: 10.1111/j.1749-818X.2008.00108.x.
Kashima, Emiko S., and Yoshihisa Kashima. 1998. Culture and language: The case of cultural dimensions and personal pronoun use. *Journal of Cross-Cultural Psychology* 29.461–86. DOI: 10.1177/0022022198293005.
Kennedy, Bob. 2018. On pronoun typology and economic measures. *Language Log*, December 10, 2018. Online: http://languagelog.ldc.upenn.edu/nll/?p=40957.
Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2.196–217. DOI: 10.1207/s15327957pspr0203_4.
Ladd, D. Robert; Seán G. Roberts; and Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1.221–41. DOI: 10.1146/annurev-linguist-030514-124819.
Levinson, Stephen C. 2003. *Space in language and cognition: Explorations in cognitive diversity*. New York: Cambridge University Press.
Levinson, Stephen C. 2012. Introduction. *Language, thought, and reality: The selected writings of Benjamin Lee Whorf*, ed. by John B. Carroll, Stephen C. Levinson, and Penny Lee, vii–xxiii. Cambridge, MA: MIT Press.
Liang, Hao; Christopher Marquis; Luc Renneboog; and Sunny Li Sun. 2018. Future-time framing: The effect of language on corporate future orientation. *Organization Science* 29.989–1236. DOI: 10.1287/orsc.2018.1217.
Licht, Amir N.; Chanan Goldschmidt; and Shalom H. Schwartz. 2007. Culture rules: The foundations of the rule of law and other norms of governance. *Journal of Comparative Economics* 35.659–88. DOI: 10.1016/j.jce.2007.09.001.
Liu, Amy H.; Sarah Shair-Rosenfield; Lindsey R. Vance; and Zsombor Csata. 2018. Linguistic origins of gender equality and women's rights. *Gender and Society* 32.82–108. DOI: 10.1177/0891243217741428.
MacKinnon, James G.; Morton Ørregaard Nielsen; and Matthew D. Webb. 2017. Bootstrap and asymptotic inference with multiway clustering. Queen's Economics Department working paper 1386. Kingston, ON: Queen's University. Online: http://hdl.handle.net/10419/188898.
Mavisakalyan, Astghik. 2015. Gender in language and gender in employment. *Oxford Development Studies* 43.403–24. DOI: 10.1080/13600818.2015.1045857.
Mavisakalyan, Astghik; Yashar Tarverdi; and Clas Weber. 2018. Talking in the present, caring for the future: Language and environment. *Journal of Comparative Economics* 46.1370–87. DOI: 10.1016/j.jce.2018.01.003.
Mavisakalyan, Astghik, and Clas Weber. 2018. Linguistic structures and economic outcomes. *Journal of Economic Surveys* 32.916–39. DOI: 10.1111/joes.12247.
McWhorter, John H. 2014. *The language hoax: Why the world looks the same in any language*. New York: Oxford University Press.
Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and causal inference: Methods and principles for social research*. 2nd edn. New York: Cambridge University Press.
Niemeier, Susanne, and René Dirven (eds.) 2000. *Evidence for linguistic relativity*. Philadelphia: John Benjamins.
Pérez, Efrén O. 2018. The language-opinion connection. *The Oxford handbook of polling and survey methods*, ed. by Lonna Rae Atkeson and R. Michael Alvarez, 249–71. New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780190213299.013.18.
Pérez, Efrén O., and Margit Tavits. 2017. Language shapes people's time perspective and support for future-oriented policies. *American Journal of Political Science* 61.715–27. DOI: 10.1111/ajps.12290.
Pérez, Efrén O., and Margit Tavits. 2019. Language influences public attitudes toward gender equality. *The Journal of Politics* 81.81–93. DOI: 10.1086/700004.
Riazi, A. Mehdi. 2016. *The Routledge encyclopedia of research methods in applied linguistics*. New York: Routledge.

Roberts, Seán G.; James Winters; and Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. *PLOS One* 10:e0132145. DOI: 10.1371/journal.pone.0132145.

Roodman, David; Morten Ørregaard Nielsen; James G. MacKinnon; and Matthew D. Webb. 2019. Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal* 19.4–60. DOI: 10.1177/1536867X19830877.

Sachs, Jeffrey D. 2001. Tropical underdevelopment. Working paper 8119. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w8119.

Sachs, Jeffrey D., and Andrew M. Warner. 1995. Natural resource abundance and economic growth. Working paper 5398. Cambridge, MA: National Bureau of Economic Research. DOI: 10.3386/w5398.

Simmons, Joseph P.; Leif D. Nelson; and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22.1359–66. DOI: 10.1177/0956797611417632.

Slobin, Dan I. 1996. From 'thought and language' to 'thinking for speaking'. *Rethinking linguistic relativity*, ed. by John J. Gumperz and Stephen C. Levinson, 70–96. New York: Cambridge University Press.

StataCorp. 2017. Stata statistical software: Release 15. College Station, TX: StataCorp LLC.

Tabellini, Guido. 2008. Institutions and culture. *Journal of the European Economic Association* 6.255–94. DOI: 10.1162/JEEA.2008.6.2-3.255.

Tavits, Margit, and Efrén O. Pérez. 2019. Language influences mass opinion toward gender and LGBT equality. *Proceedings of the National Academy of Sciences* 116.16781–86. DOI: 10.1073/pnas.1908156116.

van der Velde, Lucas; Joanna Tyrowicza; and Joanna Siwinska. 2015. Language and (the estimates of) the gender wage gap. *Economics Letters* 136.165–70. DOI: 10.1016/j.econlet.2015.08.014.

von Humboldt, Wilhelm. 1999 [1836]. *On language: On the diversity of human language construction and its influence on the mental development of the human species*. Trans. by Peter Heath. Ed. by Michael Losonsky. Cambridge: Cambridge University Press.

White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48.817–38. DOI: 10.2307/1912934.

Wieczorek, Anna Ewa. 2013. *Clusivity: A new approach to association and dissociation in political discourse*. New York: Cambridge Scholars.

Wieling, Martijn; John Nerbonne; and R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLOS One* 6(9):e23613. DOI: 10.1371/journal.pone.0023613.

[pepinsky@cornell.edu]

[Received 27 March 2019;
revision invited 9 April 2019;
revision received 29 April 2019;
revision invited 25 August 2019;
revision received 12 December 2019;
revision invited 25 October 2020;
revision received 29 January 2021;
accepted pending revisions 22 May 2021;
revision received 1 June 2021;
accepted 19 June 2021]