INSTITUTE for REPLICATION

No. 62
DISCUSSION PAPER SERIES

Replication of Dickens (2022) "Understanding Ethnolinguistic Differences: The Roles of Geography and Trade"

Javier Gonzalez

Ömer Özak

This paper received a response:

Dickens, Andrew. 2023. Response to Gonzalez and Özak's (2023) Replication Report. I4R Discussion Paper Series No. 63. Institute for Replication.

September 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 62

Replication of Dickens (2022) "Understanding Ethnolinguistic Differences: The Roles of Geography and Trade"

Javier Gonzalez[1], Ömer Özak[2]

[1]*Southern Methodist University, Dallas/USA*
[2]*Southern Methodist University, Dallas/USA, IZA Bonn/Germany, Global Labour Organization (GLO) Germany*

SEPTEMBER 2023

ISSN: 2752-1931

Replication of Dickens (2022) "Understanding Ethnolinguistic Differences: The Roles of Geography and Trade"*

Javier Gonzalez† and Ömer Özak‡

June 27, 2023

Abstract

Dickens (2022) studies the role of trade on long-run inter-ethnic linguistic differences. He establishes that neighboring ethnolinguistic groups have smaller (lexicostatistical) linguistic distances when there is a larger agricultural productivity variation between them. Specifically, he establishes that pre-1500 land productivity variation (CSI_SD) and its change due to Columbian Exchange in the post-1500 (CSI_SD_CHANGE) era decrease linguistic distances between groups. In what can be considered his main specification, which includes geographical controls, spatial controls, and language family fixed effects (Table 1 column 5), he estimates that a one standard deviation increase in the change in land productivity variation (post-1500) decreases linguistic distances by 0.11 standard deviations (p-value < 0.01) and a one standard deviation increase in land productivity variation (pre-1500) decreases linguistic distances by 0.06 standard deviations (p-value $= 0.12$). We conduct a direct replication of the paper by (i) reconstructing the main independent variables using the same original sources and following the procedures explained in the original study, (ii) using an updated version of the linguistic map (Ethnologue v17 instead of v16), and (iii) constructing alternative measures of inter-ethnic potential gains from trade. Our results basically confirm the sign, magnitude, and statistical significance of the point estimates in the original study.

Keywords: Replication, Linguistic Distances, Trade

JEL codes: F10, O10, Z10

*This paper was written as part of the "Reproducibility and Replicability in Economics" study. Authors do not have any conflicts of interest.

†Department of Economics, Southern Methodist University. PO Box 0496, Dallas, TX 75275-0496. E-mail: gonzalezjj@smu.edu.

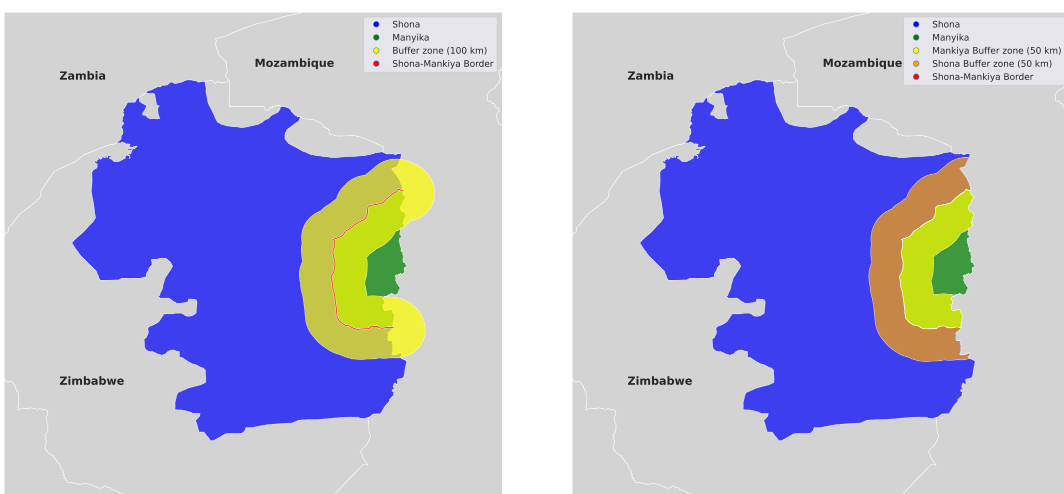
‡*Corresponding author.* Department of Economics, Southern Methodist University, IZA, and GLO. PO Box 0496, Dallas, TX 75275-0496. E-mail: ozak@smu.edu. Tel: (214) 768-2755. Fax: (214) 768-1821. ORCID 0000-0001-6421-2801.

1 Introduction

Dickens (2022) studies the effect of potential gains from trade on linguistic distances. He analyzes linguistic distances based on lexicographic distances (Wichmann et al. 2016) between spatially adjacent languages that were still spoken in world in 2009 (Lewis 2009).

Dickens (2022) constructed proxies of potential gains from trade between ethno-linguistic groups using measures of the variation in agricultural productivity (Galor and Özak 2015, 2016) in a 50km buffer around the borders where their homelands met (see Figure 1).[1] He takes advantage of the spatial and temporal variation in agricultural productivity to identify the effect of agricultural variation on linguistic distances. Specifically, following Galor and Özak (2015, 2016), he exploits the quasi-random changes in productivity generated in the course of the Columbian Exchange (Crosby 2003) as a natural experiment to identify their effect on linguistic distances.

Figure 1: Buffer Zones



(a) Original Buffer Zones (Replication, Ethnologue v16)

(b) Alternative Buffer Zones (Ethnologue v16)

Dickens (2022) estimates this effect using a cross section of all spatially adjacent language pairs and ordinary least squares regressions (in his most comprehensive

[1] I.e., buffers with a diameter of 100km.

with a large set of geographical and spatial controls, language family and country fixed effects). He double clustered standard errors at the level of each language group to account for dependence between observations. His main results are presented in his Table 1 (p.967). Specifically, in Column (1) he " reports within-family estimates for pre-Columbian land productivity variation and the post-Columbian change in land productivity variation. Both estimates enter with the expected negative sign and are statistically significant at the 1% level. This says that ethnic groups separated across high-variation regions are more similar in language than groups separated across low-variation regions. In particular, a one standard deviation increase in pre-1500 land productivity variation decreases linguistic distance by 1.7 percentage points, while a standard deviation increase in productivity variation at the onset of the Columbian exchanges implies a 2.0 percentage point decrease in linguistic distance." (p.967) Additionally, he shows that after accounting for all controls except country fixed effects (column 5) "[t]he coefficient on pre-1500 land productivity variation retains the expected sign, but loses significance at standard levels. Whereas the coefficient of interest—the effect of land productivity variation resulting from the Columbian Exchange in the post-1500 period—retains statistical significance with the expected negative sign, and is statistically equivalent to the unconditional estimate in column (1)." (p.968)

In the present paper, we investigate whether his empirical results are reproducible and replicable and further test the robustness to an alternative approach to measuring the potential gains from trade. Specifically, (1) we reconstructed his main independent variables from original sources (Galor and Özak 2015, 2016, Lewis 2009), (2) we reconstructed his main independent variables using an updated version of the ethnolinguistic map he employed to construct the buffers employed in his analysis, (3) using the same original and updated sources we constructed alternative measures of potential inter-ethnic gains from trade. Using these various new measures, we replicated his estimation in column 5 of Table 1 for presentation here (although we provide a more complete replication in the repository with the new data and code).[2]

Both us and, independently, the Institute for Replication were successfully able to reproduce the original study using the replication package provided by the author on the journal's website. We did not find any differences between the published results and the ones we obtained using his data and code.[3]

Our reconstructed independent variables, based on the same and updated underlying raw data, are very highly correlated with the ones provided in the replication package and generate almost identical results. Thus, we provide a successful direct replication of his work. As a final step, we explore further the role of inter-ethnic trade by constructing alternative measures of potential gains from inter-ethnic trade using the same underlying data. These alternative measures do not have an effect on linguistic distances between ethnic groups.

2 Reproducibility

We downloaded the replication package provided by the author to the journal.[4] The code and data worked without any issues. We were able to reproduce the results of the paper using Stata-MP version 17. In order to improve reproducibility and provide educational tools based on these analyses for individuals who may not have access to this paid software, especially individuals in developing countries, we replicated these results using free tools. Specifically, we used the open source programming language Python and various statistical and graphical packages based on it that are freely available. All our data and code are available at https://osf.io/k3p7g/. To help others replicate and learn from our replication analysis, we have also created scripts to recreate our computational environment using conda/mamba, Deepnote, and MyBinder.[5]

[2] The data and code are available at https://osf.io/k3p7g/.

[3] The replication code and data is available from the website of the Economic Journal (https://doi.org/10.1093/ej/ueab065). The code reproduces all tables and most figures (except maps), although it does not save the output to a usable file.

[4] Downloaded from https://doi.org/10.1093/ej/ueab065 on October 4, 2022.

[5] Our reproduction uses Jupyter notebooks and performs the same analyses in Python and Stata. To execute cells in the notebooks that use Stata, this software needs to be installed on the computer.

Table 1: Reproduction of Table 1 in Dickens (2022)
Standardized betas

	Dependent variable: Lexicostatistical linguistic distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.114	-0.113	-0.094	-0.117	-0.111	-0.078
	(0.035)	(0.038)	(0.038)	(0.038)	(0.038)	(0.041)
	[0.001]	[0.003]	[0.013]	[0.002]	[0.003]	[0.056]
Land productivity variation (pre-1500)	-0.094	-0.093	-0.071	-0.066	-0.062	-0.046
	(0.033)	(0.034)	(0.040)	(0.033)	(0.039)	(0.043)
	[0.005]	[0.006]	[0.072]	[0.042]	[0.117]	[0.282]
Δ in land productivity (post-1500)		-0.003	0.002	-0.018	-0.004	0.027
		(0.036)	(0.035)	(0.036)	(0.035)	(0.037)
		[0.932]	[0.958]	[0.616]	[0.918]	[0.453]
Land productivity (pre-1500)		-0.005	-0.002	0.005	0.017	0.037
		(0.024)	(0.024)	(0.025)	(0.025)	(0.025)
		[0.847]	[0.927]	[0.846]	[0.482]	[0.135]
Geography controls	No	No	Yes	No	Yes	Yes
Spatial controls	No	No	No	Yes	Yes	Yes
Language family FEs	Yes	Yes	Yes	Yes	Yes	Yes
Country FEs	No	No	No	No	No	Yes
Adjusted R^2	0.248	0.248	0.259	0.264	0.278	0.369
Observations	8402	8402	8402	8402	8402	7291

Notes: Unit of observation: border buffer zone (100 km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair‘s lexicostatistical linguistic distance. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, mean precipitation and its standard deviation, and the prevalence of malaria. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, logged distance between group centroids, the absolute difference in latitude and longitude, logged land area and logged population. Standard errors are double clustered at the level of each language group and are reported in parentheses. p-values are reported in square brackets.

Table 1 reproduces the results of Table 1 in the original paper and shows the estimates using standardized beta coefficients to simplify interpretability of the results (Table A1 in the Appendix reproduces the original Table 1 without this transformation). The results in column (5) of Table 1 suggest that after accounting for geographical and spatial controls, and language family fixed effects, a one standard deviation increase in the change in land productivity variation (post-1500) decreases linguistic distances by 0.11 standard deviations (p-value = 0.003) and a one standard deviation increase in land productivity variation (pre-1500) decreases linguistic distances by 0.06 standard deviations (p-value = 0.117). To simplify the presentation, we will focus on this same specification in the rest of the replication analysis.

3 Replication

3.1 Direct Replication using Original and Updated Data

We also replicated the construction of the main independent variables used in his analysis. We assumed the original study employed the langa.shp file from the World Language Mapping System, which provides polygons for the homelands of languages spoken across the world in 2009 (version 16) and 2014 (version17) based on the Ethnologue (Lewis 2009, Lewis et al. 2014). While the paper explains clearly most steps involved in the construction of the data, it did not clarify the geographical projection used in the analysis, which provided us with one degree of freedom. We followed the most straightforward (though perhaps not the most correct) strategy and used the original projections of the raw data in the construction.[6] Additionally, Galor and Özak (2015, 2016) provide two distinct Caloric Suitability Indices for the pre-1500 period, but Dickens (2022) does not clarify which one he used. Thus, we have a second degree of freedom, and we use their pre-1500 measure that includes Asian crops in Africa for our main replication analyses.[7]

Once we had these data sources and had made choices regarding the projections to be used, we followed a similar strategy to him to construct buffers around the borders. In the construction of borders and buffers, (i) we excluded pairs in which the border reflected the same language in two countries (e.g., the language Abron is spoken both in Ghana and Côte d'Ivoire, Spanish in many neighboring Latin American countries, and similar for Arabic in the Middle East and North Africa). Clearly, including these would have been problematic and bias the analysis. In the

[6] Both the linguistic maps for Ethnologue (Lewis 2009, Lewis et al. 2014) and the Caloric Suitability Index (Galor and Özak 2015, 2016) are provided in the WGS84, i.e., EPSG:4326, geographic coordinate system. These are not the best projections to use in a global analysis given the underlying data and question. Using a projected coordinate system like a cylindrical equal area projection (ESRI:54034) would have been preferable. Yet, previous experience suggests that using either of these generates highly correlated measures so we do not consider this a major issue. Moreover, it provides us with an alternative robustness analysis, which clearly could be improved by selecting different projections in the construction.

[7] Our replication code and data reconstructs the data using all versions of the CSI data.

original replication file, we found 111 pairs of languages that were the same language across two different countries.[8] We also identified one language that had the same name (NAME2 in the original data), was spoken in two countries, but had different ISO codes. We decided to treat it as a unique language and exclude it from the data construction.[9] None of these language pairs were included in the (original) empirical analyses since they were excluded due to being assigned missing linguistic distances. (ii) We excluded pairs in which the homeland of one language was (fully) contained inside the other, since this may reflect recent changes in the location of speakers and may bias the results.[10]

We reconstructed the data using both the version 16 and 17 of the Ethnologue. We found that between version 16 and 17, Ethnologue dropped many languages, which may have become extinct, or by mistake (e.g., version 17 does not have a polygon for Spanish in Paraguay). Additionally some new languages appeared in the data set (not clear whether these are languages that did not have a polygon in version 16 or which became more vibrant). Specifically, version 17 has 414 languages not contained in version 16, while version 16 has 537 languages not contained in version 17. These differences explain the main differences in the samples for replication.

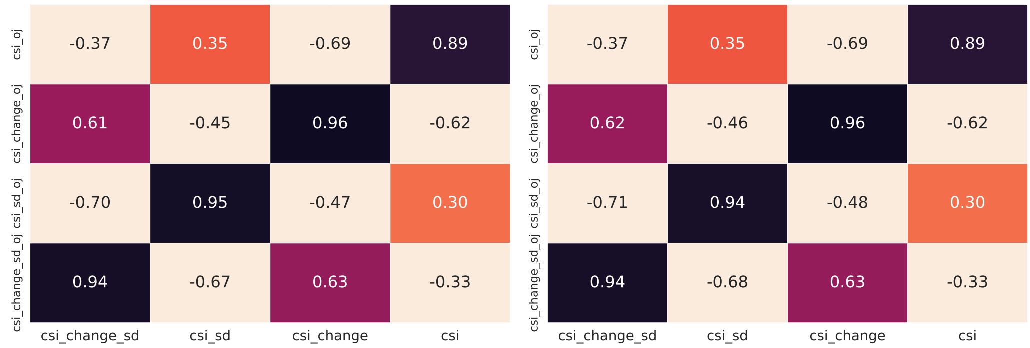
Our reconstructed independent variables (CSI_OJ, CSI_change_OJ, CSI_SD_OJ, CSI_SD_change_OJ), based on the same and updated underlying raw data, are very highly correlated with the ones provided in the replication package (CSI,

[8] Stata command: tab identifier if langIso1== langIso2.

[9] The language is Marwari which is spoken in India and Pakistan (see notebook for references).

[10] Our sample using version 16 of the Ethnologue map differed from the original study by only 1 observation, which is the pair BZX-MLI-FFM-MLI (Bozo, Hainyaxo and Fulfulde, Maasina in Mali), where one is strictly contained in the other. The replication data contains a variable called overlap, which is used in Table A12 in the original paper to explore the robustness to excluding languages that overlap. We assumed these would include the ones we excluded due to being one contained in the other, but this variable identifies many more language pairs as overlapping. We find that 2848 of the language pairs with linguistic distance data that are identified as overlapping according to the original study are not identified as overlapping in our procedure. These include cases like KHK-MNG-CMN-CHN, AAC-PNG-DBY-PNG, and AAL-CMR-KOT-CMR, which the Ethnologue map does not show as overlapping. Without the code used to generate the original data we could not determine how these overlaps were identified and why our results differed in this aspect. We did not explore this issue further since it had no impact on the main regression we analyze here.

Figure 2: Correlation Between Original and Reconstructed Independent Variables



(a) Based on Ethnologue v16 and original buffers

(b) Based on Ethnologue v17 and original buffers

CSI_change, CSI_SD, CSI_SD_change), as shown in Figure 2. Columns (1)-(5) in Table 2 explore how the reconstructed data and samples affects his main result from Table 1 column 5 in the paper. Specifically, column (1) replicates the original result, while columns (2) and (3) show the result of constraining the original data to the samples of language pairs we have in the reconstructed data with qualitatively similar results, albeit a bit stronger effects in the reduced sample of version 17. Columns (4) and (5) replicate the analysis using our reconstructed measures and finds almost identical results, although the estimated effects are always larger. Specifically, using the original data and sample, a one standard deviation increase in the change in land productivity variation (post-1500) decreased linguistic distances by 0.111 standard deviations (p-value $= 0.003$), while it decreases it by 0.115 standard deviations (p-value $= 0.001$) in the reconstructed Ethnologue version 16 data, and by 0.153 standard deviations (p-value < 0.001) in the reconstructed Ethnologue version 17 data. Thus, there does not seem to exist statistically significant differences between the replication and original analyses.

4 Robustness - Inter-Ethnic vs Generalized Trade Potential

Dickens's (2022) "aim [..] is to shed light on an unexplored channel of cultural change: inter-ethnic trade" (p.953). His main hypothesis, general theory, and model are based on the idea that "the gains from trade are a function of land productivity

Table 2: Replication and Extension of Main Result

	Dependent variable: Lexicostatistical linguistic distance $\in (0,1)$								
	Replication					Robustness			
	Original Data			Reconstructed Data		Alternative Measure			
	Original Sample	v16	v17	v16	v17	v16	v17	v16	v17
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ in land productivity variation (post-1500)	-0.111	-0.111	-0.136	-0.140	-0.192			-0.102	-0.127
	(0.038)	(0.038)	(0.040)	(0.042)	(0.042)			(0.039)	(0.041)
	[0.003]	[0.003]	[0.001]	[0.001]	[0.000]			[0.009]	[0.002]
Land productivity variation (pre-1500)	-0.062	-0.062	-0.073	-0.064	-0.104			-0.080	-0.094
	(0.039)	(0.039)	(0.042)	(0.041)	(0.042)			(0.043)	(0.045)
	[0.117]	[0.115]	[0.078]	[0.121]	[0.014]			[0.061]	[0.034]
Δ in land productivity (post-1500)	-0.004	-0.003	0.016	0.007	0.038	-0.034	-0.024	-0.007	0.013
	(0.035)	(0.035)	(0.035)	(0.043)	(0.043)	(0.031)	(0.031)	(0.035)	(0.035)
	[0.918]	[0.922]	[0.637]	[0.873]	[0.375]	[0.275]	[0.441]	[0.849]	[0.698]
Land productivity (pre-1500)	0.017	0.018	0.016	-0.003	0.008	0.008	0.002	0.018	0.017
	(0.025)	(0.025)	(0.026)	(0.032)	(0.033)	(0.024)	(0.025)	(0.025)	(0.026)
	[0.482]	[0.478]	[0.531]	[0.924]	[0.799]	[0.750]	[0.922]	[0.460]	[0.513]
Δ in cross border Δ in land productivity (post-1500)						-0.048	-0.058	-0.011	-0.011
						(0.025)	(0.025)	(0.025)	(0.024)
						[0.050]	[0.019]	[0.665]	[0.643]
Cross border Δ in land productivity (pre-1500)						0.001	0.000	0.027	0.032
						(0.022)	(0.022)	(0.023)	(0.023)
						[0.959]	[0.995]	[0.251]	[0.153]
Geography controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language family FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.278	0.278	0.281	0.279	0.283	0.278	0.280	0.279	0.282
Observations	8402	8401	7552	8401	7552	8401	7552	8401	7552

Notes: Unit of observation: border buffer zone (100 km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair‘s lexicostatistical linguistic distance. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, mean precipitation and its standard deviation, and the prevalence of malaria. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, logged distance between group centroids, the absolute difference in latitude and longitude, logged land area and logged population. Standard errors are double clustered at the level of each language group and are reported in parentheses. p-values are reported in square brackets.

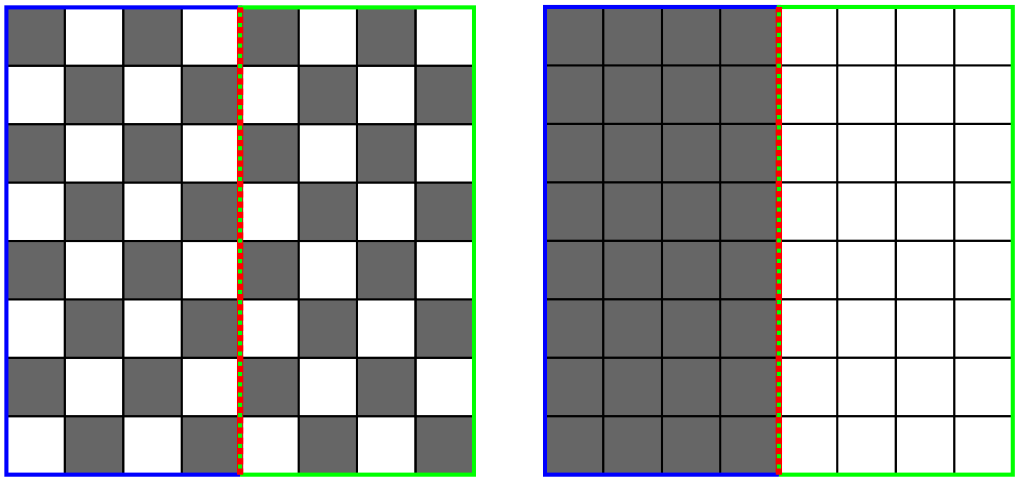
variation *between* groups" (p.959, emphasis added). Thus, the idea is that linguistic distances should decrease due to *different* groups coming together to exchange *different* types of goods they specialize in. Yet, neither his measure of inter-ethnic trade potential, variation in agricultural productivity, nor the evidence he presents based on it, *directly* provide support for this *inter-ethnic* aspect of trade, since they fail to capture cross-ethnic economic specialization. In particular, the ethno-linguistic group level analyses, which show that larger variation in land productivity is associated with more trade and less conflict (Section 4 of the paper, especially Table 4 and Figure 9), is similar and inline with previous research that has established the positive effect of diversity (ecological, geographical, and population) within an ethno-linguistic group (or its homeland) on the emergence of economic specialization, trade, and states (Depetris-Chauvin and Özak 2020b, Fenske 2014), all of which can directly reduce linguistic distances across languages in a region. Thus, his border level analyses may reflect the negative effect of *generalized* trade on linguistic distances even in cases where there is no cross-ethnic specialization. Specifically, it may very well be the case that linguistic distances decrease because people from *different* ethnic groups interact when trading *similar* goods, i.e, these interactions are not driven by ethnic level specialization. While this would have no impact on the empirical analyses themselves, as agricultural variation would decrease linguistic distances, it would affect their interpretation.

To understand this issue better, Figure 3 shows that the potential for *inter-ethnic* trade should depend (as in his model and theory) on the difference in productivity *between* ethnic groups around the border and not on the variation of productivity in the buffer as a whole, which should reflect incentives for generalized trade.[11] In particular, these figures depict two artificial buffers around an ethnic border, which is assumed to have the same mean and variation (i.e., the standard deviation) in productivity. Clearly, inter-ethnic trade potential can be expected to be higher in

[11] Equivalently, if we one were to do the analysis using the complete homelands, it should be driven by the difference in productivity between the homelands and not by the variations across both homelands.

Figure 3b than in Figure 3a since it provides stronger incentives for cross-ethnic economic specialization.

Figure 3: Productivity and Inter-Ethnic Trade Potential in Buffer around Border



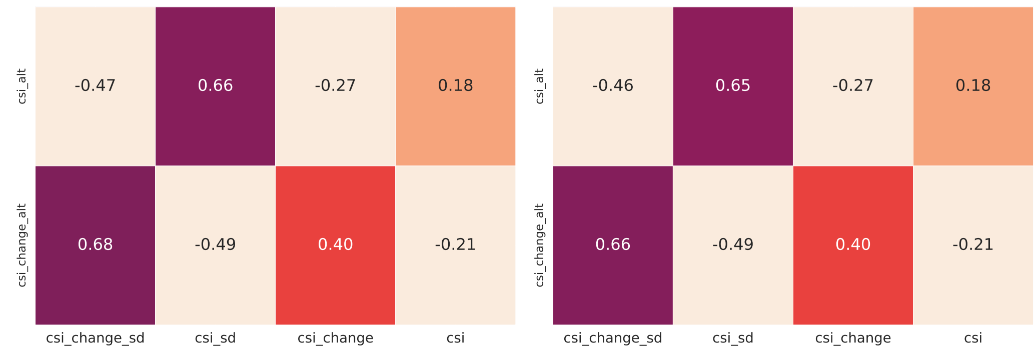
(a) Similar Productivity and Low Potential

(b) Different Productivity and High Potential

To explore this issue further, as a final step, we construct measures of the difference in productivity pre-1500 and its change post-1500 between the two sides of each buffer (as depicted in Figure 1b) and replicate the main specification using these. Specifically, for each buffer, we compute the absolute difference in productivity pre-1500 between both sides of the border, i.e., for each part of the buffer that belong to a specific homeland, and the change in this difference post-1500. Clearly, a significant effect of these differences would be prima facie evidence for the role of inter-ethnic trade. Columns (6) and (7) in Table 2 show the results of this analysis, in which we replace the original proxies of inter-ethnic trade (agricultural variation and its change) with our new measures. Specifically, the results show the effect of differences in productivity across the border pre-1500 and its change post-1500 on linguistic distances after accounting for all the same controls as in the main specification we have been focusing on in this replication (i.e., Column (5) in Table 1 of the original study). The estimates suggest that a one standard deviation increase in the changes post-1500 in the differences in agricultural productivity across the two sides of the borders decreases linguistic distances by 0.048 (p-value = 0.050) and by

0.058 (p-value = 0.019) using data from Ethnologue version 16 and 17 respectively. This is strong direct evidence that inter-ethnic trade does affect linguistic distances.

Figure 4: Correlation Between Original and Alternative Measures of Potential Gains from Inter-Ethnic Trade



(a) Based on Ethnologue v16 and original buffers

(b) Based on Ethnologue v17 and original buffers

Still, the results may reflect the forces of generalized trade. In particular, as Figure 4 shows, the correlation between our alternative measures of inter-ethnic trade potential (CSI_alt, CSI_change_alt) and the original measures of agricultural productivity variation both in the pre- and post-1500 periods is quite high with correlations above 0.65.[12] Clearly, given these high correlations, disentangling the two forces may be difficult. While not a perfect solution, as a first step, in columns (8) and (9) of Table 2 we include both sets of variables jointly. In this horse race, the effect of our measures of inter-ethnic trade potential fall between 77 and 80% and become statistically insignificant with p-values over 0.6. On the other hand, the original measures, which reflect generalized trade, remain basically unchanged and remain statistically significant at the 1% level. Thus, the results fail to provide strong evidence in favor of the inter-ethnic aspect of trade and may reflect the role of generalized trade instead. Nonetheless, these results need to be taken with a grain of salt, since there are many issues that may still bias the analysis. In particular, the location of these contemporary linguistic borders may not reflect historical ethnic borders (Depetris-Chauvin and Özak 2020a), so that our measures

[12]This could be expected due to the way the data is constructed and the relation between the standard deviation across the whole buffer and difference in the mean across both sides of the border.

may reflect poorly the historical conditions that may have generated inter-ethnic specialization. Moreover, our measures may not fully capture the incentives to specialize, e.g., across crops. Additionally, it is important to remember, that even if endowments are similar, inter-ethnic specialization and trade may still have evolved and driven linguistic convergence. Clearly, more research is needed.

5 Conclusion

Dickens (2022) presents novel and important evidence to understand the cultural convergence between ethno-linguistic groups and the role trade and geography play in it. Following his methodology, we have successfully replicated his central result establishing the negative effect of agricultural variation on linguistic distances. We have shown the results are robust to changes in the construction method and the underlying linguistic maps finding that the empirical analyses remained qualitatively unchanged, although the estimated effects tended to be larger.

As an extension of his work, we have delved deeper into the role of inter-ethnic as opposed to generalized trade in this result. Since both types of trade should increase contact between individuals, they should decrease linguistic distances due to increased contact among individuals. We show that the original measures used in the study failed to properly distinguish between these two forces and construct novel measures of potential gains from inter-ethnic trade at ethno-linguistic borders. We show that our novel measures generate qualitatively similar although quantitatively smaller effects. Also, we find that these results are not robust to accounting for the original measures. This may reflect the fact that it is generalized trade and not inter-ethnic trade that drives linguistic convergence, or various shortcomings of the new measures. Fully understanding the role of inter-ethnic trade may require different strategies like building more sophisticated measures that account for potential mismeasurement issues (Depetris-Chauvin and Özak 2020a) or proper crop specialization across ethnic groups using more data from the CSI project (Galor and

Özak 2015, 2016).[13] Clearly, more research is needed to understand this important subject.

[13] It should not be difficult to use the code we provide in our repository to extend these analyses further.

References

Crosby, A. W.: 2003, *The Columbian exchange: biological and cultural consequences of 1492*, Vol. 2, Greenwood Publishing Group.

Depetris-Chauvin, E. and Özak, Ö.: 2020a, (de facto) historical ethnic borders and contemporary conflict in africa, *Technical report*, IZA Discussion Paper Series.

Depetris-Chauvin, E. and Özak, Ö.: 2020b, The origins of the division of labor in pre-industrial times, *Journal of Economic Growth* **25**(3), 297–340.

Dickens, A.: 2022, Understanding ethnolinguistic differences: The roles of geography and trade, *The Economic Journal* **132**(643), 953–980.

Fenske, J.: 2014, Ecology, trade, and states in pre-colonial africa, *Journal of the European Economic Association* **12**(3), 612–640.

Galor, O. and Özak, Ö.: 2015, Land productivity and economic development: Caloric suitability vs. agricultural suitability, *Agricultural Suitability (July 12, 2015)* .

Galor, O. and Özak, Ö.: 2016, The agricultural origins of time preference, *American economic review* **106**(10), 3064–3103.

Lewis, M. P.: 2009, *Ethnologue: Languages of the world*, Vol. 16, SIL international Dallas, TX.

Lewis, M. P., Simons, G. F. and Fennig, C. D.: 2014, *Ethnologue: Languages of the world*, Vol. 17, SIL international Dallas, TX.

Wichmann, S., Holman, E. W. and Brown, C. H.: 2016, *The ASJP Database (version 16)*.

APPENDIX

Table A1: Reproduction of Table 1 in Dickens (2022)

	Dependent variable: Lexicostatistical linguistic distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.100	-0.100	-0.083	-0.103	-0.098	-0.065
	(0.030)	(0.034)	(0.034)	(0.034)	(0.033)	(0.034)
	[0.001]	[0.003]	[0.013]	[0.002]	[0.003]	[0.056]
Land productivity variation (pre-1500)	-0.061	-0.061	-0.046	-0.043	-0.040	-0.029
	(0.022)	(0.022)	(0.026)	(0.021)	(0.026)	(0.027)
	[0.005]	[0.006]	[0.072]	[0.042]	[0.117]	[0.282]
Δ in land productivity (post-1500)		-0.001	0.001	-0.005	-0.001	0.008
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
		[0.932]	[0.958]	[0.616]	[0.918]	[0.453]
Land productivity (pre-1500)		-0.001	-0.001	0.001	0.004	0.009
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
		[0.847]	[0.927]	[0.846]	[0.482]	[0.135]
Geography controls	No	No	Yes	No	Yes	Yes
Spatial controls	No	No	No	Yes	Yes	Yes
Language family FEs	Yes	Yes	Yes	Yes	Yes	Yes
Country FEs	No	No	No	No	No	Yes
Adjusted R^2	0.248	0.248	0.259	0.264	0.278	0.369
Observations	8402	8402	8402	8402	8402	7291

Notes: Unit of observation: border buffer zone (100 km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair‘s lexicostatistical linguistic distance. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, mean precipitation and its standard deviation, and the prevalence of malaria. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, logged distance between group centroids, the absolute difference in latitude and longitude, logged land area and logged population. Standard errors are double clustered at the level of each language group and are reported in parentheses. p-values are reported in square brackets.

Table A2: Replication and Extension of Main Result
Non-Standardized Coefficients

	Dependent variable: Lexicostatistical linguistic distance $\in (0,1)$								
	Replication					Robustness			
	Original Data			Reconstructed Data		Alternative Measure			
	Original Sample	v16	v17	v16	v17	v16	v17	v16	v17
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ in land productivity variation (post-1500)	-0.098	-0.098	-0.117	-0.115	-0.153			-0.090	-0.110
	(0.033)	(0.033)	(0.034)	(0.034)	(0.034)			(0.034)	(0.035)
	[0.003]	[0.003]	[0.001]	[0.001]	[0.000]			[0.009]	[0.002]
Land productivity variation (pre-1500)	-0.040	-0.040	-0.048	-0.040	-0.065			-0.052	-0.061
	(0.026)	(0.026)	(0.027)	(0.026)	(0.026)			(0.028)	(0.029)
	[0.117]	[0.115]	[0.078]	[0.121]	[0.014]			[0.061]	[0.034]
Δ in land productivity (post-1500)	-0.001	-0.001	0.005	0.002	0.010	-0.010	-0.007	-0.002	0.004
	(0.010)	(0.010)	(0.010)	(0.012)	(0.012)	(0.009)	(0.009)	(0.010)	(0.010)
	[0.918]	[0.922]	[0.637]	[0.873]	[0.375]	[0.275]	[0.441]	[0.849]	[0.698]
Land productivity (pre-1500)	0.004	0.004	0.004	-0.001	0.002	0.002	0.001	0.004	0.004
	(0.006)	(0.006)	(0.006)	(0.008)	(0.008)	(0.006)	(0.006)	(0.006)	(0.006)
	[0.482]	[0.478]	[0.531]	[0.924]	[0.799]	[0.750]	[0.922]	[0.460]	[0.513]
Δ in cross border Δ in land productivity (post-1500)						-0.056	-0.066	-0.012	-0.013
						(0.029)	(0.028)	(0.029)	(0.028)
						[0.050]	[0.019]	[0.665]	[0.643]
Cross border Δ in land productivity (pre-1500)						0.001	0.000	0.022	0.027
						(0.018)	(0.018)	(0.019)	(0.019)
						[0.959]	[0.995]	[0.251]	[0.153]
Geography controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language family FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.278	0.278	0.281	0.279	0.283	0.278	0.280	0.279	0.282
Observations	8402	8401	7552	8401	7552	8401	7552	8401	7552

Notes: Unit of observation: border buffer zone (100 km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair‘s lexicostatistical linguistic distance. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, mean precipitation and its standard deviation, and the prevalence of malaria. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, logged distance between group centroids, the absolute difference in latitude and longitude, logged land area and logged population. Standard errors are double clustered at the level of each language group and are reported in parentheses. p-values are reported in square brackets.