

Gonzalez, Javier; Özak, Ömer

**Working Paper**

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

I4R Discussion Paper Series, No. 62

**Provided in Cooperation with:**  
The Institute for Replication (I4R)

*Suggested Citation:* Gonzalez, Javier; Özak, Ömer (2023) : Replication of Dickens (2022) "Understanding Ethnolinguistic Differences: The Roles of Geography and Trade", I4R Discussion Paper Series, No. 62, Institute for Replication (I4R), s.l.

This Version is available at:  
<https://hdl.handle.net/10419/276250>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



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**

## I4R DISCUSSION PAPER SERIES

I4R DP No. 62

# Replication of Dickens (2022) “Understanding Ethnolinguistic Differences: The Roles of Geography and Trade”

**Javier Gonzalez<sup>1</sup>, Ömer Özak<sup>2</sup>**

*<sup>1</sup>Southern Methodist University, Dallas/USA*

*<sup>2</sup>Southern Methodist University, Dallas/USA, IZA Bonn/Germany, Global Labour Organization (GLO) Germany*

SEPTEMBER 2023

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

### Editors

**Abel Brodeur**  
*University of Ottawa*

**Anna Dreber**  
*Stockholm School of Economics*

**Jörg Ankel-Peters**  
*RWI – Leibniz Institute for Economic Research*

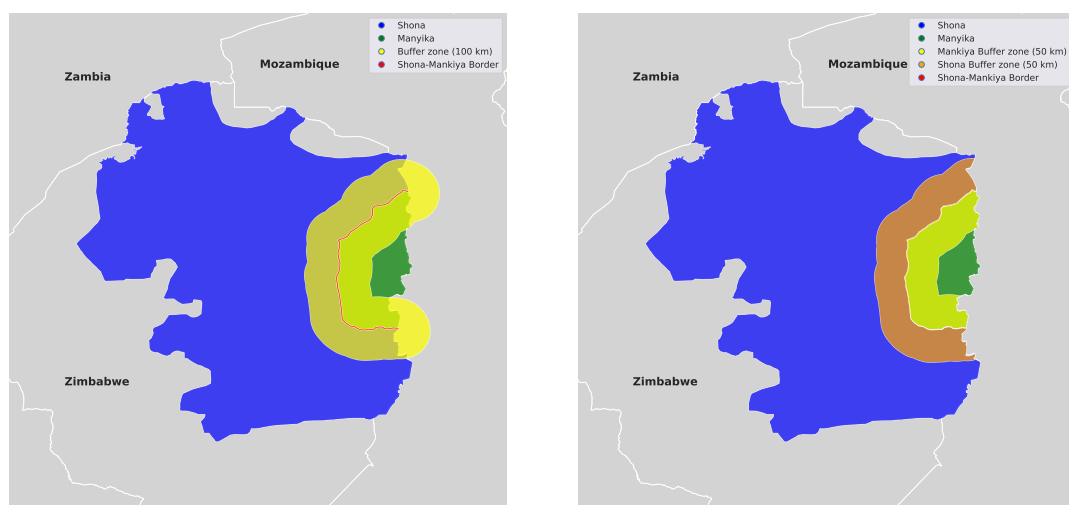


## 29 1 Introduction

30 [Dickens \(2022\)](#) studies the effect of potential gains from trade on linguistic distances.  
 31 He analyzes linguistic distances based on lexicographic distances ([Wichmann et al.](#)  
 32 [2016](#)) between spatially adjacent languages that were still spoken in world in 2009  
 33 ([Lewis 2009](#)).

34 [Dickens \(2022\)](#) constructed proxies of potential gains from trade between ethno-  
 35 linguistic groups using measures of the variation in agricultural productivity ([Galor](#)  
 36 [and Özak 2015, 2016](#)) in a 50km buffer around the borders where their homelands  
 37 met (see [Figure 1](#)).<sup>1</sup> He takes advantage of the spatial and temporal variation in  
 38 agricultural productivity to identify the effect of agricultural variation on linguis-  
 39 tic distances. Specifically, following [Galor and Özak \(2015, 2016\)](#), he exploits the  
 40 quasi-random changes in productivity generated in the course of the Columbian Ex-  
 41 change ([Crosby 2003](#)) as a natural experiment to identify their effect on linguistic  
 42 distances.

Figure 1: Buffer Zones



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

(b) Alternative Buffer Zones (Ethnologue v16)

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

<sup>1</sup>I.e., buffers with a diameter of 100km.

45 with a large set of geographical and spatial controls, language family and country  
46 fixed effects). He double clustered standard errors at the level of each language  
47 group to account for dependence between observations. His main results are pre-  
48 sented in his Table 1 (p.967). Specifically, in Column (1) he “ reports within-family  
49 estimates for pre-Columbian land productivity variation and the post-Columbian  
50 change in land productivity variation. Both estimates enter with the expected neg-  
51 ative sign and are statistically significant at the 1% level. This says that ethnic  
52 groups separated across high-variation regions are more similar in language than  
53 groups separated across low-variation regions. In particular, a one standard devia-  
54 tion increase in pre-1500 land productivity variation decreases linguistic distance by  
55 1.7 percentage points, while a standard deviation increase in productivity variation  
56 at the onset of the Columbian exchanges implies a 2.0 percentage point decrease  
57 in linguistic distance.” (p.967) Additionally, he shows that after accounting for all  
58 controls except country fixed effects (column 5) “[t]he coefficient on pre-1500 land  
59 productivity variation retains the expected sign, but loses significance at standard  
60 levels. Whereas the coefficient of interest—the effect of land productivity variation  
61 resulting from the Columbian Exchange in the post-1500 period—retains statistical  
62 significance with the expected negative sign, and is statistically equivalent to the  
63 unconditional estimate in column (1).” (p.968)

64 In the present paper, we investigate whether his empirical results are repro-  
65 ducible and replicable and further test the robustness to an alternative approach  
66 to measuring the potential gains from trade. Specifically, (1) we reconstructed his  
67 main independent variables from original sources ([Galor and Özak 2015, 2016](#), [Lewis  
68 2009](#)), (2) we reconstructed his main independent variables using an updated ver-  
69 sion of the ethnolinguistic map he employed to construct the buffers employed in  
70 his analysis, (3) using the same original and updated sources we constructed alter-  
71 native measures of potential inter-ethnic gains from trade. Using these various new  
72 measures, we replicated his estimation in column 5 of Table 1 for presentation here  
73 (although we provide a more complete replication in the repository with the new

74 data and code).<sup>2</sup>

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

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

## 86 2 Reproducibility

87 We downloaded the replication package provided by the author to the journal.<sup>4</sup>  
88 The code and data worked without any issues. We were able to reproduce the  
89 results of the paper using Stata-MP version 17. In order to improve reproducibil-  
90 ity and provide educational tools based on these analyses for individuals who may  
91 not have access to this paid software, especially individuals in developing coun-  
92 tries, we replicated these results using free tools. Specifically, we used the open  
93 source programming language Python and various statistical and graphical pack-  
94 ages based on it that are freely available. All our data and code are available at  
95 <https://osf.io/k3p7g/>. To help others replicate and learn from our replication anal-  
96 ysis, we have also created scripts to recreate our computational environment using  
97 [conda/mamba](#), [Deepnote](#), and [MyBinder](#).<sup>5</sup>

---

<sup>2</sup>The data and code are available at <https://osf.io/k3p7g/>.

<sup>3</sup>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.

<sup>4</sup>Downloaded from <https://doi.org/10.1093/ej/ueab065> on October 4, 2022.

<sup>5</sup>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)
$\Delta$ in land productivity variation (post-1500)	-0.114 (0.035) [0.001]	-0.113 (0.038) [0.003]	-0.094 (0.038) [0.013]	-0.117 (0.038) [0.002]	-0.111 (0.038) [0.003]	-0.078 (0.041) [0.056]
Land productivity variation (pre-1500)	-0.094 (0.033) [0.005]	-0.093 (0.034) [0.006]	-0.071 (0.040) [0.072]	-0.066 (0.033) [0.042]	-0.062 (0.039) [0.117]	-0.046 (0.043) [0.282]
$\Delta$ in land productivity (post-1500)		-0.003 (0.036) [0.932]	0.002 (0.035) [0.958]	-0.018 (0.036) [0.616]	-0.004 (0.035) [0.918]	0.027 (0.037) [0.453]
Land productivity (pre-1500)		-0.005 (0.024) [0.847]	-0.002 (0.024) [0.927]	0.005 (0.025) [0.846]	0.017 (0.025) [0.482]	0.037 (0.025) [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.

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



108 analysis.

### 109 **3 Replication**

#### 110 **3.1 Direct Replication using Original and Updated Data**

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

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

---

<sup>6</sup>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.

<sup>7</sup>Our replication code and data reconstructs the data using all versions of the CSI data.

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

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

149 Our reconstructed independent variables (CSL\_OJ, CSL\_change\_OJ, CSLSD\_OJ,  
150 CSLSD\_change\_OJ), based on the same and updated underlying raw data, are  
151 very highly correlated with the ones provided in the replication package (CSI,

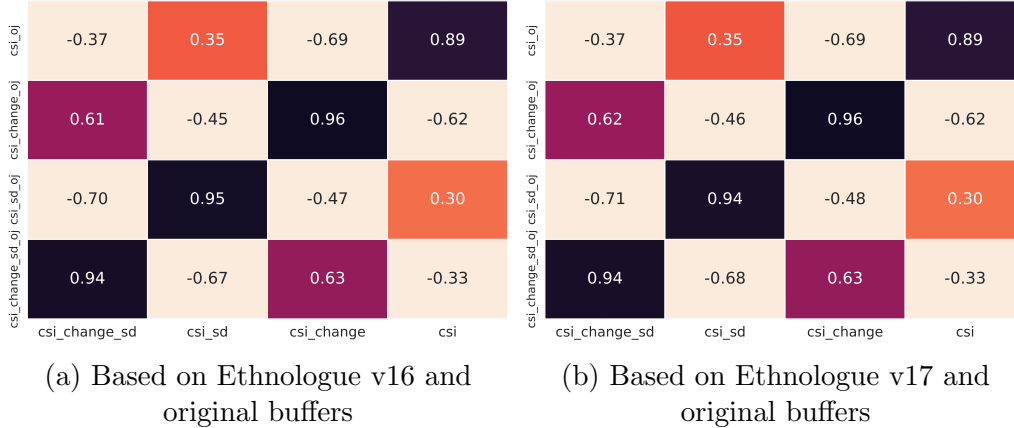
---

<sup>8</sup>Stata command: tab identifier if langIso1== langIso2.

<sup>9</sup>The language is Marwari which is spoken in India and Pakistan (see notebook for references).

<sup>10</sup>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



152 CSLchange, CSLSD, CSLSD\_change), as shown in Figure 2. Columns (1)-(5) in  
 153 Table 2 explore how the reconstructed data and samples affects his main result from  
 154 Table 1 column 5 in the paper. Specifically, column (1) replicates the original result,  
 155 while columns (2) and (3) show the result of constraining the original data to the  
 156 samples of language pairs we have in the reconstructed data with qualitatively simi-  
 157 lar results, albeit a bit stronger effects in the reduced sample of version 17. Columns  
 158 (4) and (5) replicate the analysis using our reconstructed measures and finds almost  
 159 identical results, although the estimated effects are always larger. Specifically, using  
 160 the original data and sample, a one standard deviation increase in the change in  
 161 land productivity variation (post-1500) decreased linguistic distances by 0.111 stan-  
 162 dard deviations (p-value = 0.003), while it decreases it by 0.115 standard deviations  
 163 (p-value = 0.001) in the reconstructed Ethnologue version 16 data, and by 0.153  
 164 standard deviations (p-value < 0.001) in the reconstructed Ethnologue version 17  
 165 data. Thus, there does not seem to exist statistically significant differences between  
 166 the replication and original analyses.

167 **4 Robustness - Inter-Ethnic vs Generalized Trade Potential**

168 Dickens’s (2022) “aim [...] is to shed light on an unexplored channel of cultural  
 169 change: inter-ethnic trade” (p.953). His main hypothesis, general theory, and model  
 170 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	v16	v17
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(8)	(9)
$\Delta$ in land productivity variation (post-1500)	-0.111 (0.038)	-0.111 (0.038)	-0.136 (0.040)	-0.140 (0.042)	-0.192 (0.042)	-0.192 (0.042)	-0.102 (0.039)	-0.127 (0.041)	-0.102 (0.039)	-0.127 (0.041)
Land productivity variation (pre-1500)	[0.003]	[0.003]	[0.001]	[0.001]	[0.000]	[0.000]	[0.009]	[0.002]	[0.009]	[0.002]
	-0.062 (0.039)	-0.062 (0.039)	-0.073 (0.042)	-0.064 (0.041)	-0.104 (0.042)	-0.104 (0.042)	-0.080 (0.043)	-0.094 (0.045)	-0.080 (0.043)	-0.094 (0.045)
$\Delta$ in land productivity (post-1500)	[0.117]	[0.115]	[0.078]	[0.121]	[0.014]	[0.014]	[0.061]	[0.034]	[0.061]	[0.034]
	-0.004 (0.035)	-0.003 (0.035)	0.016 (0.035)	0.007 (0.043)	0.038 (0.043)	0.038 (0.043)	-0.007 (0.035)	0.013 (0.035)	-0.007 (0.035)	0.013 (0.035)
Land productivity (pre-1500)	[0.918]	[0.922]	[0.637]	[0.873]	[0.375]	[0.375]	[0.849]	[0.698]	[0.849]	[0.698]
	0.017 (0.025)	0.018 (0.025)	0.016 (0.026)	-0.003 (0.032)	0.008 (0.033)	0.008 (0.033)	0.018 (0.025)	0.017 (0.026)	0.018 (0.025)	0.017 (0.026)
$\Delta$ in cross border $\Delta$ in land productivity (post-1500)	[0.482]	[0.478]	[0.531]	[0.924]	[0.799]	[0.799]	[0.460]	[0.513]	[0.460]	[0.513]
	-0.048 (0.025)	-0.048 (0.025)	-0.058 (0.025)	-0.048 (0.025)	-0.048 (0.025)	-0.048 (0.025)	-0.011 (0.025)	-0.011 (0.024)	-0.011 (0.025)	-0.011 (0.024)
Cross border $\Delta$ in land productivity (pre-1500)	[0.643]	[0.665]	[0.019]	[0.665]	[0.643]	[0.643]	[0.027]	[0.643]	[0.665]	[0.643]
	0.001 (0.022)	0.001 (0.022)	0.000 (0.022)	0.001 (0.022)	0.001 (0.022)	0.001 (0.022)	0.027 (0.023)	0.032 (0.023)	0.027 (0.023)	0.032 (0.023)
Geography controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language family FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted $R^2$	0.278	0.278	0.281	0.279	0.283	0.283	0.279	0.282	0.279	0.282
Observations	8402	8401	7552	8401	7552	7552	8401	8401	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.

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

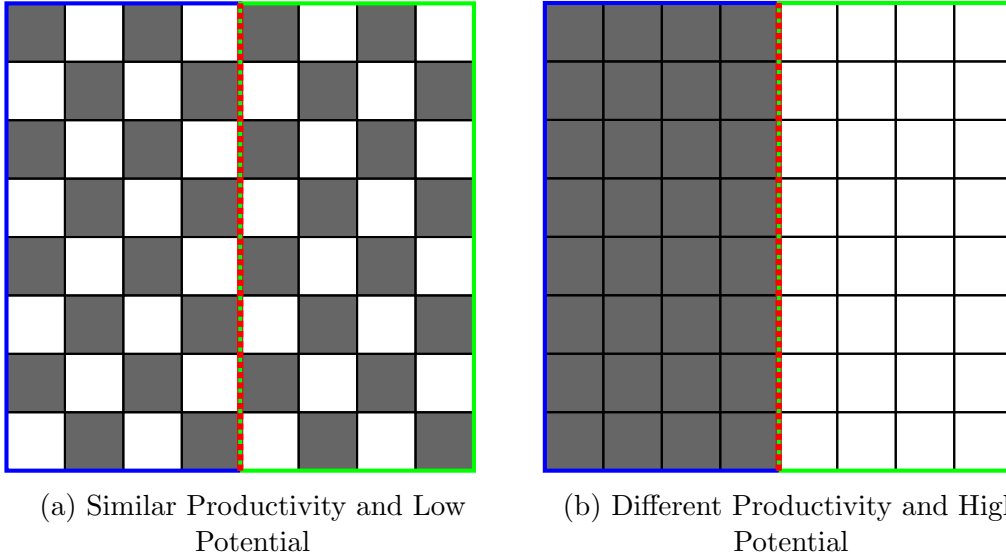
191 To understand this issue better, Figure 3 shows that the potential for *inter-ethnic*  
192 trade should depend (as in his model and theory) on the difference in productivity  
193 *between* ethnic groups around the border and not on the variation of productivity  
194 in the buffer as a whole, which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

---

<sup>11</sup>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.

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

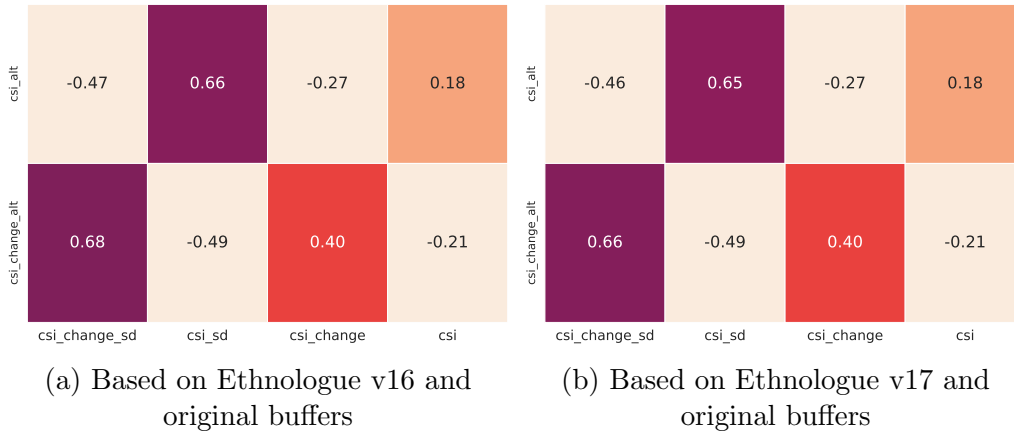
Figure 3: Productivity and Inter-Ethnic Trade Potential in Buffer around Border



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

216 0.058 (p-value = 0.019) using data from Ethnologue version 16 and 17 respectively.  
 217 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



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

<sup>12</sup>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.

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

## 239 **5 Conclusion**

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

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



261 [Özak 2015, 2016](#)).<sup>13</sup> Clearly, more research is needed to understand this important  
262 subject.

---

<sup>13</sup>It should not be difficult to use the code we provide in our repository to extend these analyses further.

263 **References**

- 264 Crosby, A. W.: 2003, *The Columbian exchange: biological and cultural consequences*  
265 *of 1492*, Vol. 2, Greenwood Publishing Group.
- 266 Depetris-Chauvin, E. and Özak, Ö.: 2020a, (de facto) historical ethnic borders and  
267 contemporary conflict in africa, *Technical report*, IZA Discussion Paper Series.
- 268 Depetris-Chauvin, E. and Özak, Ö.: 2020b, The origins of the division of labor in  
269 pre-industrial times, *Journal of Economic Growth* **25**(3), 297–340.
- 270 Dickens, A.: 2022, Understanding ethnolinguistic differences: The roles of geogra-  
271 phy and trade, *The Economic Journal* **132**(643), 953–980.
- 272 Fenske, J.: 2014, Ecology, trade, and states in pre-colonial africa, *Journal of the*  
273 *European Economic Association* **12**(3), 612–640.
- 274 Galor, O. and Özak, Ö.: 2015, Land productivity and economic development:  
275 Caloric suitability vs. agricultural suitability, *Agricultural Suitability (July 12,*  
276 *2015)* .
- 277 Galor, O. and Özak, Ö.: 2016, The agricultural origins of time preference, *American*  
278 *economic review* **106**(10), 3064–3103.
- 279 Lewis, M. P.: 2009, *Ethnologue: Languages of the world*, Vol. 16, SIL international  
280 Dallas, TX.
- 281 Lewis, M. P., Simons, G. F. and Fennig, C. D.: 2014, *Ethnologue: Languages of the*  
282 *world*, Vol. 17, SIL international Dallas, TX.
- 283 Wichmann, S., Holman, E. W. and Brown, C. H.: 2016, *The ASJP Database (ver-*  
284 *sion 16)*.

## 285 APPENDIX

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

	Dependent variable: Lexicostatistical linguistic distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
$\Delta$ in land productivity variation (post-1500)	-0.100 (0.030) [0.001]	-0.100 (0.034) [0.003]	-0.083 (0.034) [0.013]	-0.103 (0.034) [0.002]	-0.098 (0.033) [0.003]	-0.065 (0.034) [0.056]
Land productivity variation (pre-1500)	-0.061 (0.022) [0.005]	-0.061 (0.022) [0.006]	-0.046 (0.026) [0.072]	-0.043 (0.021) [0.042]	-0.040 (0.026) [0.117]	-0.029 (0.027) [0.282]
$\Delta$ in land productivity (post-1500)		-0.001 (0.010) [0.932]	0.001 (0.010) [0.958]	-0.005 (0.010) [0.616]	-0.001 (0.010) [0.918]	0.008 (0.010) [0.453]
Land productivity (pre-1500)		-0.001 (0.006) [0.847]	-0.001 (0.006) [0.927]	0.001 (0.006) [0.846]	0.004 (0.006) [0.482]	0.009 (0.006) [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	v16	v17
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(8)	(9)
$\Delta$ in land productivity variation (post-1500)	-0.098 (0.033) [0.003]	-0.098 (0.033) [0.003]	-0.117 (0.034) [0.001]	-0.115 (0.034) [0.001]	-0.153 (0.034) [0.000]	-0.090 (0.034) [0.009]	-0.110 (0.035) [0.002]	-0.090 (0.034) [0.009]	-0.090 (0.034) [0.009]	-0.110 (0.035) [0.002]
Land productivity variation (pre-1500)	-0.040 (0.026) [0.117]	-0.040 (0.026) [0.115]	-0.048 (0.027) [0.078]	-0.040 (0.026) [0.121]	-0.065 (0.026) [0.014]	-0.052 (0.028) [0.061]	-0.061 (0.029) [0.034]	-0.052 (0.028) [0.061]	-0.052 (0.028) [0.061]	-0.061 (0.029) [0.034]
$\Delta$ in land productivity (post-1500)	-0.001 (0.010) [0.918]	-0.001 (0.010) [0.922]	0.005 (0.010) [0.637]	0.002 (0.012) [0.873]	0.010 (0.012) [0.375]	-0.010 (0.009) [0.275]	-0.007 (0.009) [0.441]	-0.002 (0.010) [0.849]	-0.002 (0.010) [0.849]	0.004 (0.010) [0.698]
Land productivity (pre-1500)	0.004 (0.006) [0.482]	0.004 (0.006) [0.478]	0.004 (0.006) [0.531]	-0.001 (0.008) [0.924]	0.002 (0.008) [0.799]	0.002 (0.006) [0.750]	0.001 (0.006) [0.922]	0.004 (0.006) [0.460]	0.004 (0.006) [0.460]	0.004 (0.006) [0.513]
$\Delta$ in cross border $\Delta$ in land productivity (post-1500)						-0.056 (0.029) [0.050]	-0.066 (0.028) [0.019]	-0.012 (0.029) [0.665]	-0.012 (0.029) [0.665]	-0.013 (0.028) [0.643]
Cross border $\Delta$ in land productivity (pre-1500)						0.001 (0.018) [0.959]	0.000 (0.018) [0.995]	0.022 (0.019) [0.251]	0.022 (0.019) [0.251]	0.027 (0.019) [0.153]
Geography controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Spatial controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language family FEs	Yes	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	0.279	0.282
Observations	8402	8401	7552	8401	7552	8401	7552	8401	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.