OXFORD

Ethnolinguistic diversity and the spread of communicable diseases: a cross-country study on the COVID-19 pandemic

Cong Wang[1],[*], ID **and Jimin Wang[2]**

[1]Department of Economics, Macquarie University, North Ryde, NSW 2109, Australia and
[2]Accounting and Finance, University of Western Australia, Perth, WA, Australia

*Corresponding author. E-mail: cong.wang@mq.edu.au

Summary

Motivated by the varying effectiveness of government intervention policies to contain the COVID-19 pandemic, and the potential positive relationship between ethnolinguistic diversity and social distance, this paper aims to provide empirical evidence on the relationship between ethnolinguistic diversity and the spread of COVID-19. In particular, using global data from 113 developed and developing countries during the early stages of the pandemic (from 31 December 2019 to 8 July 2020), we have found a significant negative effect of ethnolinguistic diversity on the spread of the virus. The result is robust to alternative measures of ethnolinguistic diversity and estimator that addresses endogeneity. Moreover, we also show that the impact of ethnolinguistic diversity on the spread of COVID-19 differs in economies characterized by different levels of democracy, policy stringency on addressing COVID-19 and health expenditure.

Keywords: ethnolinguistic diversity, COVID-19 pandemic, communicable diseases, healthcare infrastructure, public policy

INTRODUCTION

In response to the COVID-19 pandemic, governments around the world have put in place various measures addressing social distancing including domestic and international travel bans, stay-at-home policy and complete lockdown of movements. These measures have been common in both developing and developed countries with only differences in stringency so that more severely affected countries chose more strict lockdown measures while less impacted countries choose to enforce them only partially.

Despite the various efforts from governments around the world, we still witness a large disparity across countries in terms of the effectiveness of controlling the spread of the disease. Further studies on determinants of the spread of the disease other than mitigation policies by governments are certainly needed. The current literature has increasingly turned its attention toward the potential association between ethnicity and COVID-19. Some focus on the outcome of COVID-19 [see e.g. (Graselli *et al.*, 2020; Pareek *et al.*, 2020)], and others discuss the general association between the two variables [see e.g. (Pan *et al.*, 2020; Vaduganathan *et al.*, 2020)]. So far, to the best of our knowledge, none has examined the potential causal relationship between ethnolinguistic diversity and COVID-19, despite the fact that theories on the impact of behaviours of humans belonging to different ethnic and linguistic groups on the spread of communicable diseases are already in place.

It is long established in the literature that ethnolinguistic diversity can impact social capital, an important capital that determines social gatherings and social connections, which then is an important determinant of social distancing outcomes and subsequent communicable disease prevention. The hypothesis that ethnic and linguistic diversity negatively affects social capital is rooted in three strands of theoretical arguments. The first one, as introduced by McPherson *et al.* (McPherson *et al.*, 2001) is known as the 'heterogeneity averse' argument. In this view, people like to socialize with peers who have similar backgrounds in ethnicity

and language. This is also known as the 'homophily' principle in structuring network ties and relationships and is widely understood to be present in many areas, such as in marriage, advice and support among, general friendship among others. This line of reasoning suggests that social connections and social capital are easier to foster in environments that exhibit homogenous characteristics such as ethnicity and language [see e.g. (Uslaner, 2001; Hooghe, 2007)]. Hooghe (Hooghe, 2007) specifically contends that one key aspect of social capital that can be fostered by closer ethnic and linguistic background is trust, since the main objective of fostering trust in a relationship is to minimize risk. The risk of betrayal is considered low in relationships with 'like-minded' people (Dasgupta, 1988) who share the same ethnicity and language. Therefore, according to this line of reasoning, low ethnolinguistic diversity tends to foster social capital, enhance social connections and gatherings, and thereby reduce social distance.

The second line of reasoning concerning the negative relationship between ethnolinguistic diversity and social capital focus on the 'group conflict theory' [see e.g. (Sheriff, 1966)]. According to this view, diverse ethnic and linguistic backgrounds in a concentrated area lead to the emphasis on within-group loyalty and hostility toward the outsiders, since the diverse background gives pressure on the within-group members to unite against a common threat from the out-group, due to mistrust between groups with different ethnic and linguistic backgrounds. The underlying reasoning concerning group conflicts also focuses on the context of scarce economic resources such as employment, so that in times of economic hardship, prejudice toward the out-group people will increase. Sides and Citrin (Sides and Citrin, 2007) also find that this group conflict can happen for fighting over cultural resources as well. For example, Sniderman and Hagendoorn (Sniderman and Hagendoorn, 2007) point out that attitudes toward immigrants are more derived from cultural and psychological factors than do from demographic and economic situations. Therefore, according to this view, low ethnolinguistic diversity generates fewer between-group conflicts, increasing social connections and capital, and thereby reducing social distance.

The third argument on the negative association between ethnolinguistic diversity and social capital is presented in classic sociology. It has long been argued since the inception of sociology that rapid social change can undermine social solidarity. The evolution of ethnic and linguistic diversity in a society can be viewed as a type of social change, and the underlying social solidarity exhibits social capital. Kesler and Bloemraad (Kesler and Bloemraad, 2010) therefore contend that an increase in ethnic and linguistic as well as cultural diversity erode social norms on certain issues, and that the 'normative consensus' on some issues can be broken. This then results from group conflicts as well as generating an increasing level of uncertainty about the then prevailing social norms, and doubts on how the current social norms might be replaced by something else. In the more recent development, Putnam (Putnam, 2007) put forward the 'hunkering down' theory which combines the previous three strands of theoretical arguments, but he argues that the long-term impact of a diverse society with multi-cultural backgrounds is likely to be positive, since people may create an 'overarching identity' such that a shared social norm may form and 'hunker-down' reactions may not be triggered.

The above theories on the negative relationship between ethnolinguistic diversity and social capital give important implications on the current COVID-19 pandemic. In a society where ethnolinguistic diversity is high, it is likely that social distance is high as well (due to lower social capital), at least for such distance between ethnic and linguistic groups. This then may be easier for the government in such societies to enact underlying social distancing rules and thereby more effectively control the spread of the disease. On the contrary, in societies where both ethnic and linguistic diversity are low, social distancing rules may be harder to enforce, as people tend to socialize more with each other within group. Consequently, government mitigation policies on COVID-19 that aim to achieve greater social distance may be less effective. Moreover, one other possible channel through which ethnolinguistic diversity may affect infectious disease outcomes is that ethnic diversity can often lead to a larger talent pool (including increased human capital in healthcare industries). Therefore, healthcare quality in such settings may be improved and thus causing much better infectious disease treatment outcomes and further leading to a lower spread of the disease [see e.g. (Fafchamps, 2000; Paulus *et al.*, 2008; Lee, 2015; Nathan, 2015; Churchill *et al.*, 2017)]. We, therefore, theorize a negative relationship between ethnolinguistic diversity and the spread of communicable diseases such as COVID-19.

The current empirical evidence on the above-proposed hypothesis is very lacking. There are some studies concerning the association between ethnic background and outcomes of COVID-19 mainly in the context of a systematic review. For example, Van Bavel *et al.* (Van Bavel *et al.*, 2020) in an overview article, contend that the pandemic is a massive global crisis that requires the use of social and behaviour sciences to 'help align human behaviour with the recommendations of epidemiologists and public health experts'. They especially focus on review articles on

the association between social and cultural influences and the spread of communicable diseases, and identify that tailored public policies that address the pandemic in societies with different ethnic and linguistic backgrounds are needed. Pan *et al.* (Pan *et al.*, 2020) also conduct a systematic review on the relationship between ethnicity and COVID-19, and find that of the 207 articles they search from various health research databases such as EMBASE, MEDLINE, Cochrane library and PROSPERO, 'five reported ethnicity, two reported no association between ethnicity and mortality'. The inclusive results certainly promote further investigation into this matter. Eaton and Kalichman (Eaton and Kalichman, 2020) use four decades of HIV pandemic data to inform current policies on containing the spread of COVID-19. They find that social and behaviour approaches are very important to address disease transmission in HIV cases. They note that 'multiple levels of intervention including intrapersonal, interpersonal, community and social factors' are necessary in the fight against COVID-19. Bruns *et al.* (Bruns *et al.*, 2020) point out that healthcare workers need to 'be aware of the facts of COVID-19, cultural implications, and potential for stigmatization of populations affected' by the disease.

For the very few empirical studies that examine the association between culture (arising from ethnic backgrounds) and COVID-19, Li *et al.* (Li *et al.*, 2020) use corporate conference call data from the USA and show that corporations with a strong corporate culture tend to do better in the pandemic. Renzaho (Renzaho, 2020) examines 'demographic, economic, political, health and socio-cultural differentials in COVID-19 morbidity and mortality' in Sub-Saharan Africa, and contends that 'the health system need to be strengthened through extending the health workforce by mobilising and engaging the diaspora' in the region. Deopa and Forunato (Deopa and Forunato, 2020) study how cultural differences can explain the spread of COVID-19 using data on mobility in Switzerland and find that 'mobility declined after the outbreak but significantly less in the German-speaking region. Within the Swiss region, higher generalized trust in others is strongly associated with lower reductions in individual mobility'. Huynh (Huynh, 2020) examines the role of culture in practicing social distancing in a global study and finds that 'countries with higher "uncertainty avoidance index" predict the lower proportion of people gathering in public such as retain and recreation, grocery and pharmacy, parks, transit stations and work places'. Shaw *et al.* (Shaw *et al.*, 2020) find that 'strong community solidarity and community behaviour' in East Asian countries such as China, Japan and South Korea are potential drivers to the successful mitigation of the pandemic, and that the homogeneity nature of

these societies might be a source of this strong community solidarity. Rathore *et al.* (Rathore *et al.*, 2020) examine the potential impact of social class diversity (in the context of the caste system) on the spread of COVID-19 in India. Their finding suggests a positive association between social class (caste) diversity and the spread of COVID-19. Whilst this seemingly contradicts the above theories and evidence that suggests a negative relationship between ethnic diversity and spread of COVID-19, we do not think this is surprising. After all, social class diversity cannot be equivalent to ethnic diversity. The unique caste system in India forms very strong within-group cooperation in the case of a major crisis like COVID-19. This however is not necessarily true in the context of ethnicity, where people in the same ethnic group are more likely to keep a low social distance compared with inter-ethnic relations, hence contributing to the spread of communicable diseases.

To test the above theory on the potential causal relationship between ethnolinguistic diversity and the spread of COVID-19 through social distance, we use global data (113 developed and developing countries) on COVID-19 cases and death (from 31 December 2019 to 8 July 2020), as well as ethnolinguistic diversity as obtained from Ethnologue to explore the potential relationship between the two variables, while controlling for macroeconomic conditions such as initial health conditions (diabetes and smoking prevalence), healthcare infrastructure (number of hospital beds) and GDP per capita. Our results indicate that high ethnic and linguistic diversity can reduce the cumulative cases of COVID-19 infections (i.e. slows its spread), therefore public policies in addressing the spread of COVID-19 may also need to take into account the existing conditions of ethnolinguistic diversity. Moreover, we show that our results are robust to alternative measures of ethnolinguistic diversity and estimator that addresses endogeneity and that the effect differs across economies endowed with different levels of democracy, policy stringency and health expenditure.

The contributions of this paper are therefore fourfold: (i) it provides empirical evidence on the relationship between ethnolinguistic diversity and the spread of communicable diseases using global data on the current COVID-19 pandemic. (ii) We attempt to investigate the potential causal relationship between ethnolinguistic diversity and COVID-19, by taking into account the potential endogeneity of ethnolinguistic diversity in its relationship with communicable diseases. We adopt various models that take care of endogeneity, such as two-stage least squares (2SLS), using physical instruments. (iii) We examine this relationship in different economies characterized by different levels of health expenditure, development of democracy

and policy stringency, effectively examining the moderation effects of these variables on the relationship between ethnolinguistic and COVID-19. (iv) We control the effects of initial health conditions (diabetes and smoking prevalence), income (GDP per capita) as well as health infrastructure (number of hospital beds) on the spread of COVID-19, providing further empirical evidence on the effectiveness of these potential determinants of containing COVID-19.

The remainder of this paper is structured as follows. The section 'Empirics' introduces data and identification strategies. The section 'Robustness to alternative measures of ethnolinguistic diversity and COVID-19' discusses the core results and performs robustness checks. The section 'Are the effects of ethnic diversity variated among different economies?' examines the relationship between ethnolinguistic diversity and COVID-19 in different economies characterized by healthcare resources, public policy effectiveness and democracy and the final section concludes.

EMPIRICS

Data

We collect unbalanced panel data on coronavirus pandemic (COVID-19), ethnolinguistic diversity and control variables on economic, health status and health infrastructure for 137 developing and developed countries (113 effective ones in the core sample) from 31 December 2019 to 8 July 2020. Our selection of days is mainly driven by data availability for COVID-19. Detailed data sources can be found in Supplementary Appendix A3. Supplementary Table A1 shows relevant summary statistics of the main variables used in the regressions.

Measures of COVID-19

The dependent variable (COVID-19 pandemic) is measured in two different ways: total accumulated number of cases per million population per day and total accumulated number of deaths per million population per day. The data are obtained from 'Our World in Data', and these measures of coronavirus spread are broadly adopted in the literature [see e.g. (Azarpazhooh *et al.*, 2020; Cheng *et al.*, 2020; Elsayed and Khan, 2020; Roser *et al.*, 2020; Stock, 2020)]. We use the total number of cases per million in the core regression and the total number of deaths per million as the alternative measure of coronavirus diseases. Some missing data are interpolated.

Ethnolinguistic diversity

We follow the study by Michalopoulos (Michalopoulos, 2012) [see also (Wang and Steiner, 2015)] to measure ethnic diversity in two different ways: log number of

languages (ln(number of languages)) and ethnolinguistic fractionalization (ELF). We use a log number of languages in our core regression. ELF and ELF at different aggregation levels (ELF 4, 8, 12 and 15) are utilized as alternative measures of ethnic diversity in the robustness check. The data are obtained from Michalopoulos (Michalopoulos, 2012).

Given that the number of languages in each geographic unit, e.g. each country is a proxy of overall ethnolinguistic diversity, the information on the location of these linguistic groups could reflect ethnic diversity (Michalopoulos, 2012). Due to the high skewness of the number of languages, we adopt a natural log of languages in this study following Michalopoulos (Michalopoulos, 2012). The index of ELF is widely adopted as the measure of ethnic diversity and the data are obtained from Michalopoulos (Michalopoulos, 2012). The ELF data at different aggregation levels are obtained from Desmet *et al.* (Desmet *et al.*, 2012).

The relationships between COVID-19 infections (total number of cases and deaths per million) and ethnolinguistic diversity can be preliminarily visualized through Supplementary Figures S1–S3. We take the average values of COVID-19 infections and ethnolinguistic diversity by each country to produce these graphs. The graphs generally indicate that ethnolinguistic diversity seems to have a negative association with COVID-19 infections. Whether this relationship is causal needs to be examined in a more advanced econometric model that treats endogeneity, which we discuss in the identification strategy section.

Control variables

Goenka et al. (Goenka *et al.*, 2014) find that the relationship between infectious diseases and economic growth is a two-way interaction: the incidence of disease influences supplies of productive labour, while health investment is also associated with recuperation from the disease. Furthermore, a number of studies show that cumulative detrimental health of infectious diseases hampers economic growth and exerts significant impacts on the poor (Sachs, 2002). Bonds et al. (Bonds *et al.*, 2010) and Tol *et al.* (Tol *et al.*, 2007) find that economic growth can reduce the burden of infectious diseases, given that economy with a high growth rate tend to have a healthier and more productive labour force and less vulnerable population that are prone to disease. Therefore, we adopt the annual GDP growth rate in each country as the control variable.

Hospital beds per thousand population are one of the most significant features of hospital service, and the scarcity of hospital beds can heavily impact the total number of cases treated (Feldstein, 1964). Nap *et al.* (Nap *et al.*, 2007) find that the peak of pandemic influenza requires a high intensity of hospital

bed deployment. Matheny *et al.* (Matheny *et al.*, 2007) contend that pandemic influenza could cause 90 million people infected and 9.9 million hospitalized in the USA, but hospital beds capacity is only 1 million, so a large scarcity of beds would swiftly overwhelm hospital capacity and lead to more infections. Therefore, we also consider hospital beds per thousand population as a control variable.

We also use smoking prevalence and diabetes prevalence as control variables in this study. Lawlor *et al.* (Lawlor *et al.*, 2003) and West (West, 2017) contend that smoking is one of the most prominent causes of morbidity and premature mortality. According to epidemiology studies, obesity pandemic in Canada is largely due to smoking and its subsequent sedentary behaviour (Spanier *et al.*, 2006). Another example is that smoking and diabetes lead to the pandemic of cardiovascular diseases in China [see (Smith and Zheng, 2010)]. Wei *et al.* (Wei *et al.*, 2000) and Treviño *et al.* (Treviño *et al.*, 2008) find that diabetes can reduce people's energy sufficiency and physical fitness. Therefore, both smoking and diabetes prevalence indicate national fitness levels to some extent, and show be used as control variables.

Identification strategy and instrumental variable

When we investigate the relationship between ethnolinguistic diversity and the COVID-19 pandemic, it is highly possible that we encounter endogeneity problems. This is largely because ethnolinguistic diversity and COVID-19 could be potentially impacted by a third variable such as education. Castles (Castles, 2009) shows that education often imposes a dominant culture, ignoring diversity and minority cultures. Therefore, the assimilation of education works like 'melting pot' that eliminates cultural differences and increases the sense of group belonging [see (Sizemore, 1978)]. In addition, it is largely agreed in the literature that schooling is utilized as the vehicle of intervention for infectious diseases (Halloran *et al.*, 1989; Zaidi *et al.*, 2004). Richardus and Kunst (Richardus and Kunst, 2001) find that the differences in infectious diseases spread between black and while people are explained by the different income and education levels. Therefore, education could exert influence on both ethnolinguistic diversity and COVID-19 pandemics. Furthermore, measurement errors could also cause issues concerning endogeneity.

We adopt the 2SLS estimator in the core specifications to deal with the endogeneity issue [see (Kelejian, 1971; Cumby *et al.*, 1983; Wang, 2013; Madsen *et al.*, 2017; Wang and Lu, 2020; Naveed and Wang, 2022)]. Current literature generally agree that ethnic diversity is not exogenous, the formation of ethnic diversity is

determined by several exogenous factors including climate, elevation, latitude, distance from the sea, distance from eastern Africa and so on [see (Cashdan, 2001; Michalopoulos, 2012)]. we follow Michalopoulos's (Michalopoulos, 2012) idea on the origins of ethnolinguistic diversity to instrument ethnolinguistic diversity by latitude, the migration distance from eastern Africa, and the distance from the sea. Detailed data definition can be found in Appendix A1. All three instrumental variables are likely to satisfy the exclusion restriction requirement, given that these variables seem to directly affect ethnolinguistic diversity, but not infectious diseases such as the COVID-19 pandemic.

Furthermore, we adopt a number of alternative measures of both ethnolinguistic diversity and the COVID-19 pandemic. Primarily, the total number of deaths per million is utilized as an alternative measure for the COVID-19 pandemic. We also use the average ELF adopted by Michalopoulos (Michalopoulos, 2012) as an alternative measure of ethnolinguistic diversity. Moreover, we follow Desmet *et al.* (Desmet *et al.*, 2012) to adopt ELF at different aggregation levels (ELF4, 8, 12 and 15) as further alternative measures of ethnolinguistic diversity.

Ordinary least square and 2SLS results in the core specifications

Columns (1) and (2) in Table 1 provide results of ordinary least square (OLS) regressions with and without control variables. The results of OLS regression show that there is a negative effect of ethnolinguistic diversity on the total number of cases of COVID-19. Columns (3) and (4) in Table 1 show results from the 2SLS estimator with and without control variables. The 2SLS results are consistent with the results from the OLS regressions. The signs and significance of all three IVs (Abslatclip, Migdistclip and Seadistclip) are consistent with findings from Michalopoulos (Michalopoulos, 2012). We report both the non-standardized and standardized coefficients so that the coefficients can be easily interpreted. For example, Column (4) in Table 1 suggests that a 1 SD increase in ethnolinguistic diversity results in a 0.641 SD decrease in COVID-19 cases. Countries used in the core regressions as well as the correlation table are shown in Appendices A2 and A4, respectively).

Furthermore, the negative effect of the log number of languages in the 2SLS regression seems to be much stronger than that of OLS regression. The differing sizes of coefficients from the OLS and 2SLS regressions could be a result of measurement errors on the endogenous variable: ethnolinguistic diversity, which is another source of endogeneity. This further justifies our use of 2SLS to deal with endogeneity. The *p*-values from the Hansen test of

Table 1: Ethno-diversity and COVID-19 (OLS/2SLS).

	(1)	**(2)**	**(3)**	**(4)**
	OLS	OLS	2SLS	2SLS
Panel A: second stage regression: total cases PM				
Ln(languages)	−114.0630***	−89.3210***	−406.5595***	−579.3395***
Ln(languages) beta	−0.1316***	−0.0988***	−0.4692***	−0.6410***
	(−21.1656)	(−11.7187)	(−4.4507)	(−3.3710)
Hospital beds		−3.6374		−104.0902*
Hospital beds beta		−0.0067		−0.1930*
		(−0.5990)		(−1.7174)
GDP growth		−68.4599***		−6.5776
GDP growth beta		−0.1163***		−0.0112
		(−16.1480)		(−0.2118)
Diabetes		19.5563***		45.3359*
Diabetes beta		0.0485***		0.1124*
		(6.9795)		(1.9041)
Smoking		6.5055***		−1.5323
Smoking beta		0.0407***		−0.0096
		(3.7228)		(−0.0950)
Test of over-identification (*p*-value)			0.4752	0.5665
F statistics/Wald statistics (*p*-value)	447.98	184.78	19.81	19.23
	(0.000)	(0.000)	(0.000)	(0.002)
N	19 761	16 884	19 761	16 884
R^2	0.0173	0.0344	0.0173	0.0159
Panel B: first stage information				
Instrumented			Ln(languages)	Ln(languages)
Abslatclip			−0.0425***	−0.0341***
			(−72.98)	(−40.24)
Migdistclip			−0.0153***	−0.0069***
			(−10.18)	(−3.64)
Seadistclip			0.736***	0.836***
			(23.78)	(22.80)
Wald statistics (*p*-value)			6381	10 186
			(0.000)	(0.000)

Notes. The regressions are estimated using the panel 2SLS, with and without controls, heteroscedasticity consistent standard errors and random effect. The standardized coefficients of all explanatory variables are reported in the table, *z*-values are in parentheses. *N* is the number of observations. We report the overall R^2 for the second stage regressions. We report the over-identification test, the null hypothesis is that the instruments are not over-identified. See Table 1 of Supplementary Material for more information. Significance at the 10, 5 and 1% levels is indicated by * and ***.

over-identification show that the first stage equations are not over-identified, hence, the validity of the instruments is confirmed. Given that our control variables are mostly time invariant background variables, we assume that the individual-specific effect is random and not correlated with the control variables, and therefore, adopt the random effect model (Schmidheiny and Basel, 2011). The Wald chi-square statistics indicate the joint significance of our core ethnolinguistic diversity variables and the control variables.

In general, the negative effect of ethnic diversity on total cases of COVID-19 per million is highly consistent with existing findings on the relationship between ethnic diversity and infectious diseases in the current literature. For example, Van Der Laan Bouma-Doff (Van Der Laan Bouma-Doff, 2007) shows that identical ethnicity has the preference of co-ethnic neighbours, which increases the residential segregation between different ethnicities. This trend of residential segregation could be intensified by inter-ethnic prejudice. Acevedo-Garcia (Acevedo-Garcia, 2000) supports the view that

residential segregation indeed has a direct decreasing impact on the transmission of infectious diseases such as tuberculosis. Moreover, the signs and significance of control variables in our results are also consistent with the findings in the current literature. In particular, hospital beds per thousand population and annual GDP growth have negative impacts on the spread of infectious diseases [see e.g. (Feldstein, 1964; Goenka *et al.*, 2014)]. Diabetes and smoking prevalence have positive impacts on the spread of infectious diseases [see e.g. (Lawlor *et al.*, 2003; West, 2017)].

ROBUSTNESS TO ALTERNATIVE MEASURES OF ETHNOLINGUISTIC DIVERSITY AND COVID-19

Thus far, we have tried to adopt only one particular measure of ethnic diversity (ln(languages)) to explain the COVID-19 spread. We also have used only the 2SLS estimator as the core model that deals with endogeneity. In fact, various types of measures of COVID-19 disease and ethnic diversity can be used to confirm that our main results are not driven by the specific measures of ethnolinguistic diversity and COVID-19 adopted. There are also the concern that the specific instrumental variables and control variables that we use may be driving the results. To address the above concerns, we implement four types of robustness checks, First, we adopt alternative measures of ethnolinguistic diversity [ELF, and ELF and different aggregation levels (4, 8, 12 and 15)] and an alternative measure of COVID-19 (death per million) to ensure that the core results are not driven by the specific measures adopted in the core regressions. Second, we adopt alternative control variables on top of the existing ones (Note that, in the additional control variable model, GDP per capita levels replaces GDP per capita growth rates, since the two variables are highly correlated.) [i.e. GDP per capita in levels, population age above 70 (number of elders), education, total number of tests per thousand, level of COVID awareness, life expectancy and extreme poverty] to ensure that the core results are not driven by the specific control variables we use. These variables are commonly identified in the literature as determinants of spread of infectious diseases [see e.g. (Feldstein, 1964; Lawlor *et al.*, 2003; Goenka *et al.*, 2014; West, 2017)]. The definitions and data sources of these additional controls can be found in Appendices A1 and A3, respectively. Third, we utilize alternative instruments to ensure the core results are not driven by the specific instrumental variables adopted. Specifically, we use the indigenous population as defined by population composition in 1500 AD, land quality, population density in 1500 AD and mean temperature as additional instruments

for ethnolinguistic diversity. These variables are contended by Michalopoulos (Michalopoulos, 2012) to be determinants of ethnolinguistic diversity to some extent, although their significance is weaker than latitude, migration distance and distance from the sea, which is why we use the latter in the core 2SLS regressions, and use the former for robustness checks only. Like instruments used in the core regressions, these new instruments are highly likely to not have any direct impact on a new infectious disease such as COVID-19, since they are exogenous factors dating back centuries so that the exclusion restriction requirement is likely to hold for these new instruments. For variable definitions and data sources of additional instrumental variables, please see Appendices A1 and A3, respectively. Finally, we also adopt clustered standard errors to account for the situation that infectious diseases spread tend to form clusters even at the country level, so that adjacent countries (e.g. USA and Canada, European nations) may be co-affected. This is to ensure that our core results are not biased by using non-clustered standard errors despite the fact that clustering is present in our sample.

Columns (1) to (6) in Supplementary Table A3 show results of robustness check from alternative measures of COVID-19 and ethnolinguistic diversity. Column (1) in Supplementary Table A3 shows that ethnolinguistic diversity has a negative effect on total cases of COVID-19 using average ELF as a measure of ethnolinguistic diversity. Column (2) in Supplementary Table A3 presents the result of an alternative measure of COVID-19 using COVID death per million, which also confirms the negative impact of ethnolinguistic diversity. Columns (3)–(6) in Supplementary Table A3 provide the results of ELF at different aggregation levels (i.e. 4, 8, 12 and 15). We find that the signs and significance of coefficients for ELF at all aggregation levels are consistent with other measures of ethnolinguistic diversity (ln(languages) and ELF).

Columns (1) and (2) in Supplementary Table A5 show results from using clustered standard errors without and with control variables, respectively. We follow Lee *et al.* (Lee *et al.*, 2021) to include two-way clustered standard errors at both the time and country dimensions. It is clear from these results that the negative impact of ethnolinguistic diversity on COVID-19 cases is confirmed regardless if we use two-way clustered standard errors or not, and that this significant effect is present when using two-way clustered standard errors regardless if we add control variables or not. By taking into account the clustering effect of infectious diseases spread at both the time and country dimensions, we are able to show consistent results with that in the core regression, further suggesting that our core results are not driven by the use of non-clustered standard errors.

Column (3) in Supplementary Table A5 shows the result from adding additional control variables [i.e. GDP per capita, number of elders (population aged 70 and above), education, COVID-19 test per capita, level of COVID-19 awareness, life expectancy and extreme poverty]. It is clear from this result that the negative effect of ethnolinguistic diversity on COVID-19 cases is once again confirmed here. One standard deviation increase in ethnolinguistic diversity results in a 0.764 SD decrease in COVID-19 cases. Both the original and additional control variables show expected signs where significant, and the results from the first stage confirm the validity of instruments by showing both expected signs and significance of the instruments and the passed over-identification test. The significant *F*-test result suggests the joint significance of additional controls and ethnolinguistic diversity. This result confirms that our core results are not driven by the use of specific control variables.

Columns (1) and (2) in Supplementary Table A6 show results from using alternative instruments (i.e. indigenous population as defined by population composition in 1500 AD; land quality; population density in 1500 AD and mean temperature) for ethnolinguistic diversity with and without controls, respectively. It is clear from these results that the negative impact of ethnolinguistic diversity on COVID-19 cases is once again confirmed here regardless if control variables are added or not. The new instruments are all significant in the first stage regressions, and the Hansen over-identification tests all pass for these instruments, further confirming the validity of the new instruments. These results suggest that our core results are not driven by the use of specific instrumental variables.

ARE THE EFFECTS OF ETHNIC DIVERSITY VARIATED AMONG DIFFERENT ECONOMIES?

In this section, we show how the impact of ethnolinguistic diversity on COVID-19 may be different across countries with different characteristics, such as degree of democracy, health expenditure and stringency of government policy on COVID-19. The government stringency index on COVID-19 is provided by 'Our World in Data'. Health expenditure data are collected from WDI. The democracy index is obtained from the Polity 5 project (more details on data sources can be found in Supplementary Appendix A3). By closely examining results from Columns (1) and (2) in Supplementary Table A4, it is quite evident that the negative effect of ethnolinguistic diversity on COVID-19 is twice as strong in countries with high government stringency than in countries with low government stringency. This result is consistent with findings from at least Hale

et al. (Hale *et al.*, 2020) and Fang *et al.* (Fang *et al.*, 2020), which show that government stringency and intervention could effectively reduce the total number of COVID-19 cases.

Columns (3) and (4) in Supplementary Table A4 present the results from high and low health expenditure countries. The negative impact of ethnolinguistic diversity can only be found in the high health expenditure group. This result is consistent with the findings from Nap *et al.* (Nap *et al.*, 2007), which postulate that health expenditure can exert a great impact on the spread of infectious diseases.

Columns (5) and (6) in Supplementary Table A4 show results from more and less democratic countries. In particular, we find that the impact of ethnolinguistic diversity on COVID-19 is stronger in more democratic countries. This result is not surprising and also consistent with findings in the current literature regarding democracy and infectious diseases. For example, Chang (Chang, 2020) shows that democracy plays an active role in providing government funds to public health organizations and allocating public health resources. Besley and Kudamatsu (Besley and Kudamatsu, 2006) also support the view that democracy has a positive association with individual health and life expectancy, and that health policy intervention is superior in democratic countries.

Finally, we also examine specific government policies in response to the COVID-19 pandemic as additional channels through which the relationship between ethnolinguistic diversity and COVID-19 spread might take hold. Specifically, we look at lockdown-related policies (i.e. stay-at-home and workplace closure), debt relief and income support. The data are obtained through the Our World in Data COVID-19 policy response. Detailed variable definitions and source can be found in the table notes of Supplementary Table A7. We split the sample using these variables by whether a country has such policies (Yes groups, > 0) or does not (No groups, = 0) at a particular date.

It is clear from Columns (1) to (8) in Supplementary Table A7 that the Yes groups all exhibit higher impacts of ethnolinguistic diversity on the spread of COVID-19 compared with the No groups where significant. In fact, the No groups are only significant in two out of the four cases (i.e. stay-at-home policy and debt relief), and the impacts are considerably smaller compared with the Yes group. This suggests that the Yes groups are important in the relationship between ethnolinguistic diversity and the spread of COVID-19 and that the four policy moderator variables that we choose are valid channels through which this relationship may take place. Furthermore, the control variables and instrumental variables have consistent signs where significant, and the Wald/*F* stats

indicate joint significance. [Note that in Columns (1) and (3), the signs on the instrument: migration distance are wrong. These may testify to the weak models exhibited by the No groups, especially considering that the core variable in Column (3) is insignificant. Moreover, the Wald stats in Columns (5) and (7) are insignificant, further give proof that the No groups regression models are not reliable (the core variable in Columns (7) are also insignificant). These findings further suggest that the moderators (stay-at-home policy, workplace closure, debt relief and income support) we choose are important in the relationship between ethnolinguistic diversity and the spread of COVID-19.] In all regressions, the over-identification tests pass, given further proof on the validity of the instruments.

The above results suggest that lockdown-related government policies (stay-at-home, workplace closure), as well as government financial supports (i.e. debt relief, income supports), are very important channels through which the impact of ethnolinguistic diversity on the spread of COVID-19 may be significantly enhanced and promoted. These results not only provide potential policy implications, but also suggest interesting theoretical underpinnings. On the one hand, in countries with higher ethnic and linguistic diversity, the government may be more pro-active in their policy making, due to simply the fact they have to pay more attention to potential inter-ethnic and inter-linguistic group dynamics. This enables them to enact more stringent lockdown-related policies and lower the spread of the virus. On the other hand, in countries where the government is less constraint by financial resources and able to provide higher financial support, inter-ethnic and inter-linguistic diversity and distance may be enhanced as people can afford to keep further social distance from one another on top of the existing social distance created by ethnolinguistic diversity, further enhancing the negative impact of ethnolinguistic diversity on the spread of COVID-19.

CONCLUSION

To conclude, this paper has investigated the potential causal relationship between ethnolinguistic diversity and the spread of COVID-19, which is motivated by theoretical and empirical arguments that assert a significant negative relationship between infectious diseases and ethnic segregation or social distance [see (Acevedo-Garcia, 2000)]. Particularly, we adopt various measures of COVID-19 and ethnolinguistic diversity to examine this relationship. This paper has also shed light on the potential moderation effects of government stringency, health expenditure and democracy on the relationship between ethnolinguistic diversity and COVID-19. By using the 2SLS estimator that addresses endogeneity, we have found evidence that supports the following:

First, we identify a significant negative effect of ethnolinguistic diversity on COVID-19 after controlling for health infrastructure, health status and macroeconomic conditions. Our core results are consistent with the findings of at least Deopa and Forunato (Deopa and Forunato, 2020), Huynh (Huynh, 2020), Li *et al.* (Li *et al.*, 2020), Renzaho (Renzaho, 2020) and Shaw *et al.* (Shaw *et al.*, 2020), among others. Second, this result is robust to different measures of ethnolinguistic diversity and COVID-19; alternative control variables; clustered standard errors and alternative instruments.

Third, we find that the effect of ethnolinguistic diversity varies in different countries with different levels of government COVID-19 control policy stringency, health expenditures and democracy. In particular, countries with more stringent government policies on COVID-19 tend to benefit more from the negative impact of ethnolinguistic diversity on COVID-19. Countries with larger health expenditures tend to exhibit a significant impact of ethnolinguistic diversity on COVID-19, whilst countries with smaller health expenditures do not. Countries that are more democratic also exhibit a larger negative relationship between ethnolinguistic diversity and COVID-19 compared with countries that are less democratic.

The above results have significant implications for policymakers and relevant stakeholders who try to contain infectious diseases such as COVID-19. On the one hand, this paper has shown that higher ethnolinguistic diversity gives a better environment for policymakers to contain the spread of infectious diseases as low social capital ensures higher social distance and less social interactions. For policymakers in a country with high ethnolinguistic diversity, they may be able to take advantage of this situation and implement less stringent non-pharmaceutical interventions such as lockdowns and moving restrictions, as people tend to be socially distant in the first place. Whereas policymakers located in low ethnolinguistic diversity countries might have to consider strict measures to prevent mass gatherings and high social face-to-face interactions. On the other hand, policymakers have to be careful about the differing magnitude of this effect in countries endowed with different characteristics such as democracy and health expenditure. The section 'Are the effects of ethnic diversity variated among different economies?' certainly shows that replying on ethnolinguistic diversity alone to contain the spread of infectious diseases is certainly not possible, since for example, the effect becomes insignificant in countries endowed with low health expenditure.

Finally, results from the section 'Are the effects of ethnic diversity variated among different economies?'

also show that government COVID-19 policies such as staying at home order, workplace closure, debt relief and income support; as well as health expenditure and degree of democracy are possible moderators through which the negative impact of ethnolinguistic diversity on COVID-19 can be enhanced. Social capital as a channel through which diversity can negatively impact COVID-19 is therefore very plausible. As illustrated in the introduction, societies endowed with higher ethnolinguistic diversity tend to have lower social capital, which results in higher social distance. Under this condition, it is easier for government to enact COVID-19 mitigation policies and make disease control more effective, thereby reducing the number of cases. That said, we do acknowledge the lack of exact channel identification in this paper due to data paucity. This is a natural limitation of this paper and could be addressed by future research.

Supplementary Material

Supplementary material is available at *Health Promotion International* online.

REFERENCES

Acevedo-Garcia, D. (2000) Residential segregation and the epidemiology of infectious diseases. *Social Science and Medicine*, **51**, 1143–1161.

Azarpazhooh, M. R., Morovatdar, N., Avan, A., Phan, T. G., Divani, A. A., Yassi, N. *et al.* (2020) Covid-19 pandemic and burden of non-communicable diseases: an ecological study on data of 185 countries. *Journal of Stroke and Cerebrovascular Diseases*, **29**, 105089.

Besley, T. and Kudamatsu, M. (2006) Health and democracy. *American Economic Review*, **96**, 313–318.

Bonds, M. H., Keenan, D. C., Rohani, P. and Sachs, J. D. (2010) Poverty trap formed by the ecology of infectious diseases. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1185–1192.

Bruns, D. P., Kraguljac, N. V. and Brus, T. R. (2020) COVID-19: facts, cultural considerations, and risk of stigmatization. *Journal of Transcultural Nursing*, **31**, 326–332.

Cashdan, E. (2001) Ethnic diversity and its environmental determinants: effects of climate, pathogens, and habitat diversity. *American Anthropologist*, **103**, 968–991.

Castles, S. (2009) *World Population Movements, Diversity, and Education*. In *The Routledge International Companion to Multicultural Education*. Routledge, pp. 49–61.

Chang, W. Y. (2020) Democracy, natural resources, and infectious diseases: the case of malaria, 1990–2016. *Studies in Comparative International Development*, **55**, 354–380.

Cheng, V. C., Wong, S. C., Chuang, V. W., So, S. Y., Chen, J. H., Sridhar, S. *et al.* (2020) The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2. *Journal of Infection*, **81**, 107–114.

Churchill, S. A., Ocloo, J. E. and Siawor-Robertson, D. (2017) Ethnic diversity and health outcomes. *Social Indicators Research*, **134**, 1077–1112.

Cumby, R. E., Huizinga, J. and Obstfeld, M. (1983) Two-step two-stage least squares estimation in models with rational expectations. *Journal of Econometrics*, **21**, 333–355.

Dasgupta, P. (1988) Trust as a commodity. In Gambetta, D. (ed), *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell, Oxford, pp. 49–72.

Deopa, N. and Forunato, P. (2020) *Coronagraben. Culture and Social Distancing in Times of COVID-19, SSRN Working Paper, No. 3635287.*

Desmet, K., Ortuño-Ortín, I. and Wacziarg, R. (2012) The political economy of linguistic cleavages. *Journal of Development Economics*, **97**, 322–338.

Eaton, L. A. and Kalichman, S. C. (2020) Social and behavioral health responses to COVID-19: lessons learned from four decades of an HIV pandemic. *Journal of Behavioral Medicine*, **43**, 341–345.

Elsayed, Y. and Khan, N. A. (2020) Immunity-boosting spices and the novel coronavirus. *ACS Chemical Neuroscience*, **11**, 1696–1698.

Fafchamps, M. (2000) Ethnicity and credit in African manufacturing. *Journal of Development Economics*, **61**, 205–235.

Fang, Y., Nie, Y. and Penny, M. (2020) Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: a data-driven analysis. *Journal of Medical Virology*, **92**, 645–659.

Feldstein, M. S. (1964) Effects of differences in hospital bed scarcity on type of use. *British Medical Journal*, **2**, 561–564.

Goenka, A., Liu, L. and Nguyen, M. H. (2014) Infectious diseases and economic growth. *Journal of Mathematical Economics*, **50**, 34–53.

Hale, T., Petherick, A., Phillips, T. and Webster, S (2020) *Variation in Government Responses to COVID-19, Blavatnik School of Government Working Paper, 31.*

Halloran, M. E., Bundy, P. and Pollitt, E. (1989) Infectious disease and the UNESCO basic education initiative. *Parasitology Today*, **5**, 359–362.

Hooghe, M. (2007) Social capital and diversity generalized trust, social cohesion and regimes of diversity. *Canadian Journal of Political Science*, **40**, 709–732.

Huynh, D. (2020) Does culture matter social distancing under the COVID-19 pandemic? *Safety Science*, 104872.

Kelejian, H. H. (1971) Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, **66**, 373–374.

Kesler, C. and Bloemraad, I. (2010) Does Immigration Erode Social Capital? The conditional effects of immigration-generated diversity on trust, membership, and participation across 19 countries, 1981–2000. *Canadian Journal of Political Science*, **43**, 319–347.

Lawlor, D. A., Frankel, S., Shaw, M., Ebrahim, S. and Smith, G. D. (2003) Smoking and ill health: does lay epidemiology explain the failure of smoking cessation programs among deprived populations? *American Journal of Public Health*, **93**, 266–270.

Lee, N. (2015) Migrant and ethnic diversity, cities and innovation: firm effects or city effects? *Journal of Economic Geography,* **15**, 769–796.

Lee, J. N., Mahmud, M., Morduch, J., Ravindran, S. and Shonchoy, A. S. (2021) Migration, externalities, and the diffusion of COVID-19 in South Asia. *Journal of Public Economics*, **193**, 104312.

Li, K., Liu, X., Mai, F. and Zhang, T. (2020) *The Role of Corporate Culture in Bad Times: Evidence from the COVID-19 Pandemic, SSRN Working Paper No. 3632395*.

Madsen, J. B., Wang, C. and Steiner, B. (2017) The creation of effective states in the OECD since 1870: the role of inequality. *European Journal of Political Economy*, **50**, 106–121.

Matheny, J., Toner, E. and Waldhorn, R. (2007) Financial effects of an influenza pandemic on US hospitals. *Journal of Health Care Finance*, **34**, 58.

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001) Birds of a feather: homophily in social networks. *Annual Review of Sociology*, **27**, 415–444.

Michalopoulos, S. (2012) The origins of ethnolinguistic diversity. *American Economic Review*, **102**, 1508–1539.

Nap, R. E., Andriessen, M. P., Meessen, N. E. and Van der Werf, T. S. (2007) Pandemic influenza and hospital resources. *Emerging Infectious Diseases*, **13**, 1714–1719.

Nathan, M. (2015) Same difference? Minority ethnic inventors, diversity and innovation in the UK. *Journal of Economic Geography,* **15**, 129–168.

Naveed, A. and Wang, C. (2022) Innovation and labour productivity growth moderated by structural change: analysis in a global perspective. *Technovation*, 102554. doi:10.1016/j. technovation.2022.102554

Pan, D., Sze, S., Minhas, J. S., Bangash, M. N., Pareek, N., Divall, P. *et al*. (2020) The impact of ethnicity on clinical outcomes in COVID-19: a systematic review. *EClinicalMedicine*, 100404.

Pareek, M., Bangash, M. N., Pareek, N., Pan, D., Sze, S., Minhas, J. S. *et al*. (2020) Ethnicity and COVID-19: an urgent public health research priority. *Lancet*, **395**, 1421–1422.

Paulus, R. A., Davis, K. and Steele, G. D. (2008) Continuous innovation in health care: implications of the Geisinger experience. *Health Affairs*, **27**, 1235–1245.

Putnam, R. D. (2007) *E. pluribus* Unum: diversity and community in the twenty-first century. *Scandinavian Political Studies*, **30**, 137–174.

Rathore, U., Das, U. and Sarkhel, P. (2020) Birds of a feather flock together? Diversity and spread of COVID-19 cases in India, arXiv preprint arXiv:2011.05839.

Renzaho, N. (2020) The need for the right socio-economic and cultural fit in the COVID-19 response in Sub-Saharan Africa: examining demographic, economic political, health and socio-cultural differentials in COVID-19 morbidity and mortality. *International Journal of Environmental Research and Public Health*, **17**, 3445.

Richardus, J. H. and Kunst, A. E. (2001) Black–White differences in infectious disease mortality in the United States. *American Journal of Public Health*, **91**, 1251–1253.

Roser, M., Ritchie, H., Ortiz-Ospina, E. and Hasell, J. (2020) Coronavirus disease (COVID-19)—statistics and research. *Our World in Data*, **4**.

Sachs, J. D. (2002) Macroeconomics and health: investing in health for economic development. *Revista Panamericana de Salud Pública*, **12**, 143–144.

Schmidheiny, K. and Basel, U. (2011) Panel data: fixed and random effects. *Short Guides to Microeconometrics*, **7**, 2–7.

Shaw, R., Kim, Y. K. and Hua, J. (2020) Governance, technology and citizen behavior in pandemic: lessons from COVID-19 in East Asia. *Progress in Disaster Science*, **6**, 100090.

Sheriff, M. (1966) *In Common Predicament: Social Psychology of Inter-group Conflict and Cooperation*. Houghton Mifflin, Boston.

Sides, J. M. and Citrin, J. (2007) European opinion about immigration: the role of identities, interests, and information. *British Journal of Political Science*, **37**, 477–504.

Sizemore, B. A. (1978) The melting pot revisited: assimilation, pluralism and integration in American society. *The Review of Education/Pedagogy/Cultural Studies*, **4**, 379–387.

Smith, Jr S. C. and Zheng, Z. J. (2010) The impending cardiovascular pandemic in China. *Circulation: Cardiovascular Quality and Outcomes*, **3**, 226–227.

Sniderman, P. and Hagendoorn, L. (2007) *When Ways of Life Collide. Multiculturalism and Its Discontents in the Netherlands*. Princeton University Press, Princeton, NJ.

Spanier, P. A., Marshall, S. J. and Faulkner, G. E. (2006) Tackling the obesity pandemic. *Canadian Journal of Public Health*, **97**, 255–257.

Stock, J. H. (2020) *Data Gaps and the Policy Response to the Novel Coronavirus (No. w26902)*. National Bureau of Economic Research.

Tol, R. S., Ebi, K. L. and Yohe, G. W. (2007) Infectious disease, development, and climate change: a scenario analysis. *Environment and Development Economics*, **12**, 687–706.

Treviño, R. P., Fogt, D. L., Wyatt, T. J., Leal-Vasquez, L., Sosa, E. and Woods, C. (2008) Diabetes risk, low fitness, and energy insufficiency levels among children from poor families. *Journal of the American Dietetic Association*, **108**, 1846–1853.

Uslaner, E. M. (2001) *The Moral Foundations of Trust*. Cambridge University Press, New York.

Vaduganathan, M., Vardeny, O., Michel, T., McMurray, J. J., Pfeffer, M. A. and Solomon, S. D. (2020) Renin–angiotensin–aldosterone system inhibitors in patients with COVID-19. *New England Journal of Medicine*, **382**, 1653–1659.

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M. *et al*. (2020) Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, **4**, 460–471.

Van Der Laan Bouma-Doff, W. E. N. D. A. (2007) Involuntary isolation: ethnic preferences and residential segregation. *Journal of Urban Affairs*, **29**, 289–309.

Wang, C. (2013) Can institutions explain cross country differences in innovative activity? *Journal of Macroeconomics*, **37**, 128–145.

Wang, C. and Lu, Y. (2020) Can economic structural change and transition explain cross-country differences in innovative activity? *Technological Forecasting and Social Change*, **159**, 120194.

Wang, C. and Steiner, B. (2015) Can ethno-linguistic diversity explain cross-country differences in social capital? A global perspective. *Economic Record*, **91**, 338–366.

Wei, M., Gibbons, L. W., Kampert, J. B., Nichaman, M. Z. and Blair, S. N. (2000) Low cardiorespiratory fitness and physical inactivity as predictors of mortality in men with type 2 diabetes. *Annals of Internal Medicine*, **132**, 605–611.

West, R. (2017) Tobacco smoking: health impact, prevalence, correlates and interventions. *Psychology and Health*, **32**, 1018–1036.

Zaidi, A. K., Awasthi, S. and deSilva, H. J. (2004) Burden of infectious diseases in South Asia. *BMJ*, **328**, 811–815.