Online Appendix: Unified Language, Labor and Ideology

Yang You

Last Updated: Jan. 2018

A. Survey Question Selection

This appendix describes the four survey sources used in the paper and explicitly lists the survey questions used in the paper. In this section, we also discuss the differences of definitions for the similar variables. For example, Putonghua (language) proficiency appears in CLDS, CGSS and WVS survey but the three surveys elicit and define the proficiency in different ways. WVS and ABS also share some similar questions in ideology and politics with nuance difference.

A1. China Labor Dynamics Survey 2012 (CLDS)

Language Proficiency Measure (Table 3): Question 9 in Interviewer Self-reported Questionnaire.

A2. China General Social Survey 2012 (CGSS)

Language Proficiency Measure (Table 3): Question A50 (Putonghua Speaking Proficiency)

A3. World Value Survey 1990, 1995, 2001, 2007, 2012 (WVS)

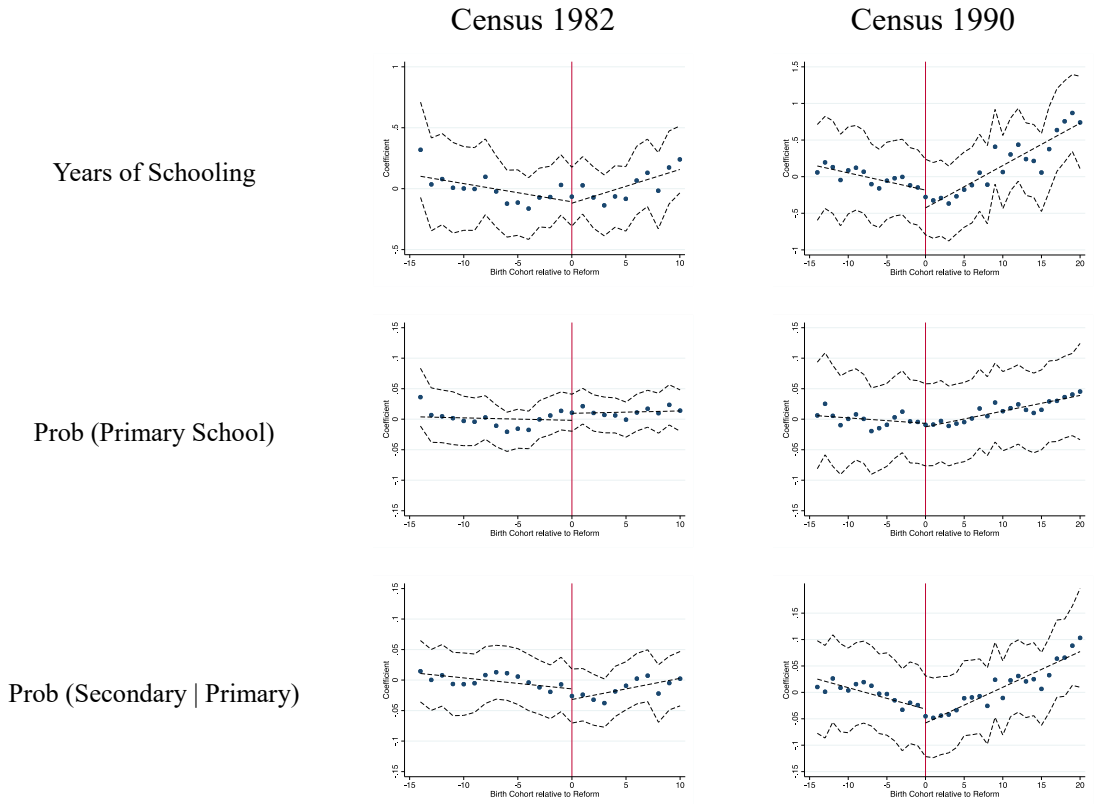
Position in Paper	Question Content	Variable Name	v1990	v1995	v2001	v2007	v2012
Table 3 Column (7)	What language do you normally speak at home?	followpoli	NA	NA	v217	NA	NA
Table 8 Panel A	How often do you follow politics in the news on television or on the radio or in the daily papers?	followpoli	NA	NA	v217	NA	NA
Table 9 Q1	How proud are you to be [Nationality]?	proud	v322	v205	v216	v209	v211
Table 9 Q2	I see myself as part of the Chinese nation.	belongcountry	NA	NA	NA	v211	v214

Table 9 Q3	I see myself as part of my local community.	belonglocal	NA	NA	NA	v212	v213
Table 9 Q4	To which of these geographical groups would you say you belong first of all?. And the next? And which do you belong to least of all?	belongfirst	v320	v203	v214	NA	NA
Table 11 Panel A Q1	How much confidence you have in them: Political Parties	trustparty	v285	NA	v155	v139	v116
Table 11 Panel A Q3	Willingness Signing a petition	petition	NA	NA	NA	v96	v85
Table 11 Panel B Q1	Who should be responsible for public welfare: individual or government	govwelfare	NA	v127	v143	v118	v98
Table 11 Panel B Q2	Private Ownership or Public Ownership	govbus	NA	v126	v142	v117	v97

A4. Asian Barometer Survey Wave 1, 2, 3 (ABS)

Position in Paper	Question Content	Variable Name	Wave1	Wave2	Wave3
Table 8 Panel A	Q4: How often do you follow news about politics?	followpolitic	q057	q057	q44
Table 8 Panel B	Main Source of Info	television	NA	qii51_1	NA
Table 8 Panel B	Main Source of Info	newspaper	NA	qii51_2	NA
Table 8 Panel B	Main Source of Info	radio	NA	qii51_3	NA
Table 8 Panel B	Main Source of Info	internet	NA	qii51_4	NA
Table 8 Panel B	Main Source of Info	cell	NA	qii51_5	NA
Table 9 Q5	Q2: Does (answer in Q156) do more good or harm to the region? (NEW)	goodchina	NA	NA	q157
Table 10 Panel A	Q1:Where would you place our country under the present government? (RATING BOARD)	democurrent	q100	q100	q91
Table 10 Panel B	Q2:To what extent would you want our country to be	demodesire	q101	q101	q93

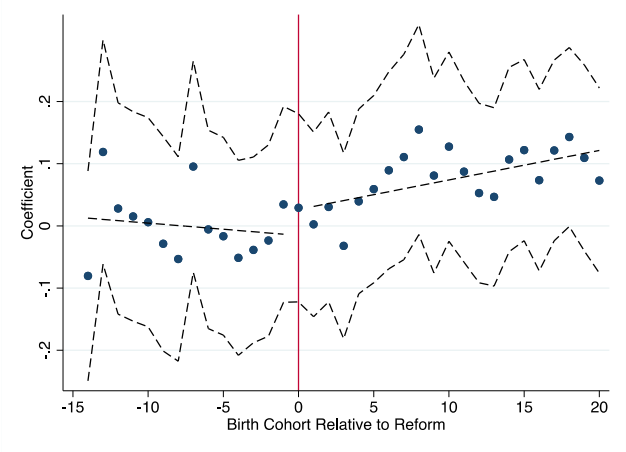
	democratic now? (RATING BOARD)				
Table 10 Panel B	Q3: Here is a similar scale of 1 to 10 measuring the extent to which people think democracy is suitable for our country. I	demosuit	q103	q103	q94
Table 10 Panel B	Q7: Which of the following statements comes closest to your own opinion?:Democracy is always preferable to any other kind of government	demopref	q117	q117	q124
Table 10 Panel C	Q8:Democracy is capable of solving the problems of our society	demotrust	q118	q118	q125
Table 10 Panel C	Q9: If you had to choose between democracy and economic development, which would you say is more important?	demodev	q119	q119	q126
Table 10 Panel C	Q11: Do you agree or disagree with the following statement: “Democracy may have its problems, but it is still the best form of government.”	demobest	NA	NA	q128
Table 11 Q2	Q1: Which country has the most influence in Asia? (NEW)	influcountry	NA	NA	q156
Table 11 Q5	Q1: The government should maintain ownership of major state-owned enterprises.	govbus	q140	q140	NA



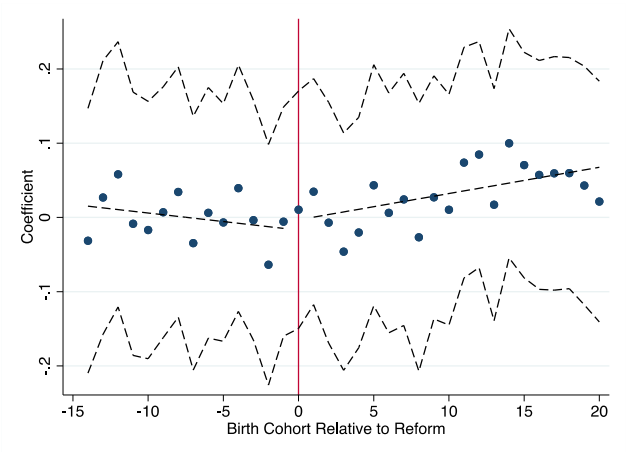
Appendix Fig 1: The left panel plots $\beta_{2,t}$ with Census 1982 data and the right panel plots $\beta_{2,t}$ with Census 1990. The sample includes 15 pre-reform birth cohorts and 10 post-reform birth cohorts in the left panel and 20 post-reform birth cohorts in the right panel. Birth cohort -15 is set as the base. $y_{i,j,t}$ in top row is years of schooling, $y_{i,j,t}$ in middle row is the probability of formal primary education and $y_{i,j,t}$ in bottom row is the conditional probability for secondary education. The dashed straight lines are fitted lines in pre- and post-treatment birth cohorts. 90% confidence intervals are plotted. Standard errors used for confidence intervals are clustered at the county level.

$$y_{i,j,t} = \sum_t \beta_{1,t} Post_{i,j,t} + \sum_t \beta_{2,t} Post_{i,j,t} * Distance_j + \alpha_j + \zeta_{prov,t} + \varepsilon_{i,j,t}$$

Panel A: Census 2000-Male Subsample



Panel B: Census 2000-Female Subsample



Appendix Fig 2: This figure plots the birth cohort specific coefficients $\beta_{2,t}$ with non-agricultural participation as the dependent variable ($y_{i,j,t}$) for 15 pre-reform and 20 post-reform birth cohorts. Birth cohort -15 is set as the base. Panel A and Panel B report estimates with male and female subsamples of Census 2000. The dashed straight lines are fitted lines in pre- and post-treatment birth cohorts. 90% confidence intervals are plotted. Standard errors used for confidence intervals are clustered at the county level.

$$y_{i,j,t} = \sum_t \beta_{1,t} Post_{i,j,t} + \sum_t \beta_{2,t} Post_{i,j,t} * Distance_j + \zeta_{prov,t} + \alpha_j + \varepsilon_{i,j,t}$$

Appendix Table 1: Language Effect on Education Attainment

Panel A: Census 1982 Full Sample				
Number of Post-Reform Cohorts	**5**	**10**	**15**	**20**
	$y_{i,j,t}$: Years of Education			
$Post_{i,j,t} * Distance_j$	-0.224	-0.227	-0.199	-0.103
	(0.238)	(0.249)	(0.252)	(0.240)
Obs.	600,877	732,397	828,401	970,027
	$y_{i,j,t}$: Primary School Enrollment			
$Post_{i,j,t} * Distance_j$	-0.005	-0.012	-0.013	-0.012
	(0.018)	(0.020)	(0.021)	(0.023)
Obs.	600,877	732,397	828,401	970,027
	$y_{i,j,t}$: Porb (Middle School | Primary School)			
$Post_{i,j,t} * Distance_j$	-0.029	-0.018	-0.016	-0.007
	(0.032)	(0.037)	(0.040)	(0.040)
Obs.	416,756	521,345	603,415	727,681

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Panel B: Census 1990 Full Sample				
Number of Post-Reform Cohorts	**5**	**10**	**15**	**20**
	$y_{i,j,t}$: Years of Education			
$Post_{i,j,t} * Distance_j$	-0.311**	-0.227	-0.122	-0.002
	(0.150)	(0.152)	(0.155)	(0.153)
Obs.	1,749,311	2,133,087	2,437,948	2,951,260
	$y_{i,j,t}$: Primary School Enrollment			
$Post_{i,j,t} * Distance_j$	-0.006	-0.003	0.004	0.008
	(0.015)	(0.016)	(0.017)	(0.019)
Obs.	1,749,311	2,133,087	2,437,948	2,951,260
	$y_{i,j,t}$: Porb (Middle School | Primary School)			
$Post_{i,j,t} * Distance_j$	-0.042*	-0.027	-0.018	-0.005
	(0.023)	(0.024)	(0.025)	(0.026)
Obs.	1,332,533	1,665,056	1,941,447	2,418,669

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Notes: This table reports the regression results with education as the dependent variables ($y_{i,j,t}$). α_j is the county fixed effect, $\zeta_{prov,t}$ is the province-cohort fixed effect. $Post_{j,t}$ is the post-treatment dummy for county j in birth cohort t. $Distance_j$ is the linguistic distance between local dialect in county j and Putonghua. Data sample includes 15 pre-reform birth cohorts. All regressions include county and province-birth cohort fixed effects. Robust standard errors clustered by survey county are reported in parenthesis.

$$y_{i,j,t} = \beta_1 Post_{i,j,t} + \beta_2 Post_{i,j,t} * Distance_j + \zeta_{prov,t} + \alpha_j + \varepsilon_{i,j,t}$$

Appendix Table 2: School Dropout Rate

Number of Post-Reform Cohorts	**5**	**10**	**15**	**20**
	$y_{i,j,t}$: Years of Education			
$Post_{i,j,t} * Distance_j$	-0.008	-0.006	-0.010	-0.015
	(0.021)	(0.017)	(0.016)	(0.015)
Obs	132,822	165,657	193,050	238,637
	$y_{i,j,t}$: Primary School Enrollment			
$Post_{i,j,t} * Distance_j$	-0.004	-0.005	-0.007	-0.012
	(0.019)	(0.015)	(0.015)	(0.014)
Obs	132,822	165,657	193,050	238,637
	$y_{i,j,t}$: Porb (Middle School | Primary School)			
$Post_{i,j,t} * Distance_j$	0.0002	0.0001	-0.006	-0.008
	(0.017)	(0.014)	(0.013)	(0.012)
Obs	53,869	73,765	93,049	126,431
	$y_{i,j,t}$: Porb (High School | Middle School)			
$Post_{i,j,t} * Distance_j$	-0.005	0.0002	0.0009	-0.003
	(0.022)	(0.017)	(0.015)	(0.015)
Obs	15,774	23,070	31,029	41,863

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Appendix Table 3: Non-agricultural Sector Employment-Census 1982 and 1990

	Full Population Sample			Male Subsample			Female Subsample		
Number of Post-Reform Cohorts	1982	1990	Obs	1982	1990	Obs	1982	1990	Obs
5	0.012	0.012	126,887	0.018	0.018	69,640	0.002	0.002	57,247
	(0.025)	(0.025)		(0.032)	(0.032)		(0.035)	(0.035)	
10	0.032	0.031	163,828	0.053*	0.051*	89,151	0.005	0.006	74,677
	(0.024)	(0.024)		(0.029)	(0.029)		(0.032)	(0.032)	
15	0.052**	0.052**	223,466	0.066**	0.065**	120,447	0.035	0.039	103,019
	(0.024)	(0.025)		(0.028)	(0.029)		(0.032)	(0.032)	
20	0.062**	0.065**	283,650	0.079***	0.078***	152,432	0.042	0.052	131,218
	(0.026)	(0.027)		(0.029)	(0.030)		(0.034)	(0.035)	

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Notes: This table reports the regression results with non-agricultural job participation as the dependent variables ($y_{i,j,t}$) using the subsample of rural residents. α_j is the county fixed effect, $\zeta_{\text{prov,t}}$ is the province-cohort fixed effect. $Post_{j,t}$ is the post-treatment dummy for county j in birth cohort t. $Distance_j$ is the linguistic distance between local dialect in county j and Putonghua. Geo_j is the geographical distance from county j to Beijing. Data sample include 15 pre-reform birth cohorts. All regressions include county and province-birth cohort fixed effects. Robust standard errors clustered by survey county are reported in parenthesis. Columns (1), (3) and (5) report the estimates with full sample, male subsample and female subsample of Census 1982. Columns (2), (4) and (6) report the estimates with full sample, male subsample and female subsample of Census 1990

$$\text{y}_{\text{i,j,t}} = \beta_1 \text{Post}_{\text{i,j,t}} + \beta_2 \text{Post}_{\text{i,j,t}} * \text{Distance}_\text{j} + \beta_3 \text{Post}_{\text{i,j,t}} * \text{Geo}_\text{j} + \zeta_{\text{prov,t}} + \alpha_\text{j} + \varepsilon_{\text{i,j,t}}$$

Appendix Table 4: Sectorial Decomposition of Non-Agricultural Participation Census 1982 and 1990

	Non-Agricultural Sector	Gov. Officials	Admin. Staff	Tech. Specialists	Service Workers	Factory Worker
	Panel A: Census 1982 + OLS Model					
$Post_{i,j,t} * Distance_j$	0.047***	-0.004	0.001	0.005	0.003	0.049**
	(0.015)	(0.003)	(0.002)	(0.006)	(0.003)	(0.015)
Obs	2,333,783	1,883,256	1,884,004	2,000,594	1,892,798	2,137,467
	Panel B: Census 1982 + Logit Model					
$Post_{i,j,t} * Distance_j$	4.937***	3.439***	1.154***	1.906***	3.817***	6.473***
	(0.230)	(0.382)	(0.206)	(0.121)	(0.242)	(0.291)
$Post_{i,j,t}$	-2.058***	-1.510***	-0.589***	-0.800***	-1.722***	-2.711***
	(0.108)	(0.162)	(0.105)	(0.0534)	(0.122)	(0.143)
Obs	2,333,783	1,883,256	1,884,004	2,000,594	1,892,798	2,137,467
	Panel C: Census 1990 +OLS Model					
$Post_{i,j,t} * Distance_j$	0.049*	0.005*	0.002	0.022***	0.011**	0.049
	(0.029)	(0.003)	(0.002)	(0.008)	(0.005)	(0.033)
Obs	3,262,322	2,798,586	2,791,006	2,882,239	2,819,034	3,098,377

	Panel D: Census 1990 + Logit Model					
$Post_{i,j,t} * Distance_j$	7.065***	7.099***	4.108***	4.178***	4.940***	8.412***
	(0.488)	(0.629)	(0.679)	(0.394)	(0.552)	(0.568)
$Post_{i,j,t}$	-2.753***	-2.364***	-1.600***	-1.595***	-2.079***	-3.327***
	(0.218)	(0.248)	(0.297)	(0. 169)	(0.230)	(0.265)
Obs	3,262,322	2,798,586	2,791,006	2,882,239	2,819,034	3,098,377

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Notes: We replicate Table 6 with Census 1982 and 1990. We classify the non-agricultural occupations with the one-digit Chinese occupation classification code (GB-19XX), which is slightly different with GB/T6565-1999, into the same five categories: Government officials (classification code: 2), technology specialists (classification code: 0/1/4), Administrative staff (classification code: 3), Service workers (classification code: 5) and Factory workers (classification code: 7/8/9). We estimate the specification (1) (Table 5 Column 1) by each occupation category. All regressions include 15 pre-reform birth cohorts and 20 post-reform birth cohorts for 1990 and 12 post-reform cohorts for 1982. In Panel A, we report the OLS estimators for β_2 with clustered robust standard errors. In Panel B, we estimate the specification using the Logit model without county fixed effects and report the MLE estimators for β_1 and β_2 with clustered standard errors at the county level.

Appendix Table 5: Full Dynamics of Migration

Panel A: Full Population Sample								
Number of Post-Reform Cohorts	5	10	15	20	5	10	15	20
	Rural + Urban Residents				Rural Residents			
All Type of Migration	0.023	0.030	0.029	0.037*	0.038**	0.040***	0.043**	0.045**
	(0.0195)	(0.019)	(0.020)	(0.022)	(0.016)	(0.015)	(0.017)	(0.020)
Migration within Province	-0.028	-0.028*	-0.030*	-0.032**	0.002	-0.007	-0.007	-0.016
	(0.018)	(0.017)	(0.016)	(0.016)	(0.015)	(0.013)	(0.014)	(0.015)
Across Province	0.051***	0.058***	0.059***	0.069***	0.036***	0.047***	0.050***	0.061***
	(0.009)	(0.011)	(0.013)	(0.017)	(0.007)	(0.008)	(0.010)	(0.014)
Across Language Area	0.053***	0.059***	0.057***	0.065***	0.033***	0.043***	0.047***	0.054***
	(0.009)	(0.011)	(0.012)	(0.015)	(0.006)	(0.007)	(0.009)	(0.012)
Geo Distance to Beijing	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	184,208	237,426	319,046	402,035	140,632	179,675	242,634	306,589

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Panel B: Male Subsample								
Number of Post-Reform Cohorts	5	10	15	20	5	10	15	20
	Rural + Urban				Rural			
All Type of Migration	0.0026	0.009	0.008	0.016	0.017	0.017	0.017	0.019
	(0.023)	(0.021)	(0.022)	(0.024)	(0.019)	(0.016)	(0.019)	(0.022)
Migration within Province	-0.041**	-0.034*	-0.038**	-0.041**	-0.001	-0.010	-0.016	-0.027*
	(0.021)	(0.018)	(0.017)	(0.017)	(0.016)	(0.013)	(0.014)	(0.016)
Across Province	0.044***	0.043***	0.046***	0.057***	0.018*	0.027**	0.033***	0.046***
	(0.012)	(0.013)	(0.014)	(0.019)	(0.010)	(0.010)	(0.013)	(0.016)
Across Language Area	0.044***	0.043***	0.043***	0.050***	0.014	0.020**	0.027**	0.035**
	(0.011)	(0.012)	(0.013)	(0.017)	(0.009)	(0.009)	(0.011)	(0.014)
Geo Distance to Beijing	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	95,584	122,829	164,382	206,512	71,285	91,029	122,724	155,151

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Panel C: Female Subsample								
Number of Post-Reform Cohorts	5	10	15	20	5	10	15	20
	Rural + Urban				Rural			
All Type of Migration	0.043*	0.051**	0.049**	0.059**	0.059**	0.062***	0.067***	0.071***
	(0.025)	(0.023)	(0.022)	(0.024)	(0.024)	(0.021)	(0.020)	(0.023)
Migration within Province	-0.015	-0.022	-0.023	-0.024	0.008	-0.005	-0.003	-0.007
	(0.024)	(0.021)	(0.020)	(0.020)	(0.023)	(0.019)	(0.018)	(0.019)
Across Province	0.057***	0.073***	0.073***	0.083***	0.052***	0.068***	0.069***	0.078***
	(0.010)	(0.011)	(0.013)	(0.016)	(0.008)	(0.009)	(0.010)	(0.013)
Across Language Area	0.060***	0.076***	0.072***	0.080***	0.052***	0.068***	0.068***	0.075***
	(0.009)	(0.011)	(0.013)	(0.015)	(0.007)	(0.008)	(0.009)	(0.012)
Geo Distance to Beijing	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	88,624	114,597	154,664	195,523	69,347	88,646	119,910	151,438

Level of Significance * p<0.10 ** p<0.05 *** p<0.01

Notes: This table shows full dynamics of the migration patterns in year 2000. Data sample include 15 pre-reform birth cohorts. All regressions include birth cohort, county, province-birth cohort fixed effects and geographical distance control. Robust standard errors clustered by birth county are reported in parenthesis.

$$\text{Migration}_{i,j,t} = \beta_1 \text{Post}_{i,j,t} + \beta_2 \text{Post}_{i,j,t} * \text{Distance}_j + \beta_3 \text{Post}_{i,j,t} * \text{Geo}_j + \zeta_{prov,t}\alpha_j + \varepsilon_{i,j,t}$$

A Brief History of Chinese Language Unification

Both political and economic factors drive the rising demand for language unification in China over the last century. On the political side, the government demands a more united country and develops stronger state capacity in local areas, rather than delegating governance to the existing local powers. The central government cannot appoint officials to areas where the official cannot speak the local language. The communication barrier keeps the political system segregated. Also, language diversity makes it more difficult for the government to inform the public or promote ideologies since it is too costly to cover all language groups. On the economic side, language diversity restricts economic development because people may fail to trade or invest if they cannot bargain or negotiate deals in the same language.

Among all candidates for the unified language, Mandarin is the top choice since it is the most widely spoken language in China. The Republic of China (ROC) made the first attempt to unify the language across the country. Beijing Mandarin was first legislated to be the official language in China in 1932 (ROC Year 21) after some modifications in grammar, and was also called the "National Language" (Guo Yu)[1]. However, the language unification was not a high priority policy for ROC government given World War II and the Chinese Civil War with the Communist Party. The first large scale language unification movement officially started in the ninth year after the establishment of People's Republic of China (PRC) in 1958. The special office for the Pinyin education reform started its mission in December 1954[2] and completed the reform proposal and Pinyin textbook within three years.

[1] People's Republic of China (PRC) renames "National Language" (Guo Yu) to be called Putonghua in mainland China after 1949. The Guo Yu is still the official language in Taiwan. Thus, there is no communication barrier between mainland China and Taiwan.

[2] Youguang Zhou, a Japan-trained economist and linguist who passed at the age of 112, was invited as the leader of the special Pinyin Reform office. Zhou proposed the two promotion standards for Putonghua (Beijing Mandarin based), 1. Putonghua should be designated as the only language on campus. 2. Putonghua should be the only language for communication in public areas. These two standards are the golden rules in the promotion of Putonghua.

A Brief Discussion on Research Limitations

There are three main limitations worth notice in this study. First, we cannot analyze the welfare of language unification since all the outcomes are measured in the world where language unification happens. We do not observe the counterfactual world without Putonghua; thus welfare analysis is almost impossible. Second, the working hypothesis is that the linguistically distant area is more treated in the language unification movement. Although we do observe a larger language proficiency improvement, the working hypothesis is still hard to access given the ambiguous metrics of the linguistic distance. Moreover, the model can also be misidentified if the treatment magnitude is not linear in the linguistic metric. Third, our estimation can be downward biased because we neglect the spillover effect across birth cohorts. The post-reform birth cohorts may teach Putonghua to the pre-reform cohorts, thus people under the old regime may pick up Putonghua as well. The spillover leads to underestimation of the language effects.