This notebook contains my replication of the results from the following paper:

> Bleakley, Hoyt, and Chin, Aimee. 2010. Age at arrival, English proficiency, and social assimilat
> ion among US immigrants. American Economic Journal: Applied Economics, 2(1), 165-92.

Downloading and viewing this notebook:

-The best way to view this notebook is by downloading it and the repository it is located in from //GitHub//. Other viewing options like MyBinder or NBViewer may have issues with displaying images or coloring of certain parts (missing images can be viewed in the folder //files// on GitHub).

-The original paper, as well as the data and codes provided by the authors can be accessed here: https://www.aeaweb.org/articles?id=10.1257/app.2.1.165 (https://www.aeaweb.org/articles?id=10.1257/app.2.1.165)

In [1]:
```
#importing packages
import numpy as np
import pandas as pd
from statsmodels.formula.api import wls
from auxiliary import *
from auxiliary2 import *
#setting up dataframes
df_ind1 = pd.read_csv( 'data/p00use_mf_indiv1.csv')
df_ind2 = pd.read_csv( 'data/p00use_mf_indiv2.csv')
df_ind = df_ind1.append(df_ind2)
df_mat = pd.read_csv( 'data/p00use_mf_matched_with_spouse.csv')
```

Table of Contents

---

1. Introduction

Among numerous scholars, Beakley and Chin (2010) discuss the effect of English proficiency on social assimilation between immigrants in the US. While it is intuitive that immigrants with better English skills would be melt in the US society more thoroughly, competitive theories can be offered to explain the correlation between the two properties. A theory is that English skill lowers the difficulty, or, as economists put it, the cost of across group social interactions. As a result, interaction with the local increases when the immigrants have better English skills. On the other hand, it can also be argued that it is the people, who prefer the destination country's culture and society more, self-selected into immigrating into the US. And the higher English level of this group of people is another product of their endorsement towards the US culture. So the correlation is just driven by the cultural preference as a confounding factor, rather than causation.

Beakley and Chin (2010) show evidential support for the theory that English proficiency causes social integrations using 2000 census data. The empirical strategy employed is to use the age of arrival as an instrumental variable. The justification of this practice is that the age of arrival correlates with the ability to acquire a second language and at the same time, does not affect other endogenous variables. They also exploite immigrants from Anglophone countries as a control group for non-linguistic effects of age at arrival. The IV esitmation confirms the causal effect of English proficiency on social outcome. English proficiency raises the probabilities of being divorced, having a more educated, higher earning, or US-born spouse, having fewer children.

In this notebook, we will replicate this study of Beakley and Chin. After that, a discussion of the conceptual framework and robustness checks will be offered. Our analysis explores some conceptual and econometrics issues in the original study.

This notebook is structured as follows. In the next section, we present the theoretical background that raises the authors' research interest. Following that we will give a brief data description. In Section 3, we analyze the instrumental variable strategy employed and the specification of the estimations. The main highlights of this notebook are sections 4 to 7. In Section 4, we will discuss some problems of the conceptual framework of the authors. Section 5 presents our replication of the results in the benchmark. Section 6 consists of various robustness checks adn extensions based on the benchmark. in Section 7 we sum up this paper.

2. Theoretical Background and Data Description

2.1 Theoretical Background

Researchers are motivated to study the factors contributing to the social assimilation of immigrants. Previous studies, such as Gillian Stevens and Gray Swicegood (1987), Brian Duncan and Stephen J. Trejo (2007), Xin Meng and Robert G. Gregory (2005) have shown that English proficiency raises the probability of intermarriage. Duncan and Trejo (2007) particularly examine Mexican immigrants' marital outcomes. They investigate certain characteristics of Mexican Americans. It is revealed that Mexican Americans who are married to non-Mexicans "tend to speak better English, be more educated, be more likely to work, and earn more compared to Mexican Americans married to either Mexican immigrants or US-born Mexicans". Similar evidence to support English proficiency relationships with fertility outcomes are given by Ann Marie Sorenson et al. (1988), and Swicegood et al. (1988). Apart from marital and fertility outcomes, English proficiency also correlated with residential location outcomes, as Edward Funkhouser and Fernando A. Ramos (1993) and Maude Toussaint-Comeau and Sherrie L. W. Rhine (2004) demonstrated. The relationship between the proficiency of the dominant language in the destination country and the level of social assimilation brings significant policy implications. If knowing the mainstream language helps immigrants integrating into their new home, then policies help improving language skills can certainly enhance social harmony.

But the nature of that correlation is still debatable. The "casual hypothesis" claims that language proficiency lower the effort, or in an economics terminology - the "cost" of social interaction with the local residents. Under this partial equilibrium framework, lowering the cost of social interaction implies an increase in the consumption of it. Therefore the language proficiency causes social interaction. However, this hypothesis comes across challenges. The first one is that language proficiency is correlated with other variables that will affect the social outcome. For example, how well a person masters a second language is correlated with his ability such as IQ. It is possible that an immigrant is getting a well-paid simply because of his higher ability instead of language proficiency. Secondly, even if there is a direct relationship between language proficiency and social outcome, it still remains doubtful if it is the people have assimilated into the destination country, such as by having a spouse or a job in the destination country, can improve the English proficiency during their stay in the destination country.

2.2 Data Description

Our data is obtained from the 2000 US Census of Population and Housing. Measures including English level, a crucial variable for our computation of the instrumental variable, are self-reported by the respondent as a categorical variable. The census was conducted on household level. We can identify each respondent's and the cohabiting spouse's answer. Martial outcomes in social assimilation are therefore possible to be shown.

We subset the data to childhood immigrants currently aged from 25 to 55. A childhood immigrant is an immigrant who was under age 15 upon arrival in the US. A childhood immigrant typically did not choose to immigrate but just follow the parents' decision. Under our definition, a childhood immigrant spent at least 11 years and at most 55 years in the US. The minimum and maximum age of observations ensure enough years are given to expose to the English environment, at the same time avoid selection biases due to retired and deceased immigrants.

We further subset the data by their origins' English tie. A childhood immigrant who has a non-English-speaking country of birth belongs to the treatment group. A childhood immigrant whose countries of birth have English as 1) an official language and 2) predominant language, belongs to the control group

Our descriptive statistics can be shown as follows.

In [2]:
```
sumlist_ind = ["eng", "age", "female", "white" , "black" , "asianpi"  , "other",  "multi" , "hispdum" ,'
sumlist_mat = ["spouseeng", "marriednative" ,"couplesamebpld", "couplesameancestry1", "spouseage", "spou

#droppingh Null entry in english skills
df_ind['eng'].replace('  ', np.nan, inplace=True)
df_ind= df_ind.dropna(subset=['eng'])

df_ind["young9"] = np.where(df_ind.agearr<=9,1,0)

dfs_ind= df_ind[sumlist_ind]

dfs_ind.loc[(dfs_ind["nengdom"] ==1 ), 'English group'] = 'Treatment Group : Born in non-English-speakir
dfs_ind.loc[(dfs_ind["nengdom"] ==1 )&( dfs_ind["young9"] == 1), 'Age group'] = 'Arrived Age 0-9'
dfs_ind.loc[(dfs_ind["nengdom"] ==1 )&( dfs_ind["young9"] == 0), 'Age group'] = 'Arrived Age 10-14'

dfs_ind.loc[(dfs_ind["nengdom"] ==0 ), 'English group'] = 'Control Group: Born in English-speaking count
dfs_ind.loc[(dfs_ind["nengdom"] ==0 )&( dfs_ind["young9"] == 1), 'Age group'] = 'Arrived Age 0-9'
dfs_ind.loc[(dfs_ind["nengdom"] ==0 )&( dfs_ind["young9"] == 0), 'Age group'] = 'Arrived Age 10-14'

dfs_ind= dfs_ind.drop(['nengdom', 'young9'], axis=1)

#droppingh Null entry in english skills
df_mat['eng'].replace('  ', np.nan, inplace=True)
df_mat= df_mat.dropna(subset=['eng'])

df_mat["young9"] = df_mat["agearr"] <=9

# keeping useful variables
```

```
dfs_mat= df_mat[sumlist_mat]

dfs_mat.loc[(dfs_mat["nengdom"] ==1 ), 'English group'] = 'Treatment Group : Born in non-English-speakir
dfs_mat.loc[(dfs_mat["nengdom"] ==1 )&( dfs_mat["young9"] == 1), 'Age group'] = 'Arrived Age 0-9'
dfs_mat.loc[(dfs_mat["nengdom"] ==1 )&( dfs_mat["young9"] == 0), 'Age group'] = 'Arrived Age 10-14'

dfs_mat.loc[(dfs_mat["nengdom"] ==0 ), 'English group'] = 'Control Group: Born in English-speaking count
dfs_mat.loc[(dfs_mat["nengdom"] ==0 )&( dfs_mat["young9"] == 1), 'Age group'] = 'Arrived Age 0-9'
dfs_mat.loc[(dfs_mat["nengdom"] ==0 )&( dfs_mat["young9"] == 0), 'Age group'] = 'Arrived Age 10-14'
dfs_mat = dfs_mat.rename(columns={'nchild': 'nchild_spouse', 'haskid': 'haskid_spouse'})
dfs_mat= dfs_mat.drop(['nengdom', 'young9'], axis=1)

# the observations is to be weighted by perwt2
sumlist_ind.remove("nengdom")
sumlist_ind.remove("young9")
sumlist_ind.remove("perwt2")
sumlist_mat.remove("nengdom")
sumlist_mat.remove("young9")
sumlist_mat.remove("perwt2")


dfs_ind= dfs_ind.drop(['perwt2'], axis=1)
dfs_mat= dfs_mat.drop(['perwt2'], axis=1)
list_table1 = list(dfs_ind.columns)
list_table1.remove("English group")
list_table1.remove("Age group")
list_table2 = list(dfs_mat.columns)
list_table2.remove("English group")
list_table2.remove("Age group")


for i in list_table1:
    dfs_ind = dfs_ind.rename(columns={i: Dict_for_sumlist1[i]})
for i in list_table2:
    dfs_mat = dfs_mat.rename(columns={i: Dict_for_sumlist1[i]})
```

```
/Users/chingfungchow/opt/anaconda3/lib/python3.7/site-packages/pandas/core/indexing.py:844: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  self.obj[key] = _infer_fill_value(value)
/Users/chingfungchow/opt/anaconda3/lib/python3.7/site-packages/pandas/core/indexing.py:965: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  self.obj[item] = s
```

In [3]:
```
print("Table 1a: the mean and SD of variables for the control group and treatment group")
print("This table refers to column (4) and (1) in Table 1 of the benchmark" )


print("The mean of variables for control group and treatment group")
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(dfs_ind.groupby(["English group"]).mean(), dfs_mat.groupby(["English group"]).mean())

print("The standard deviation of variables for control group and treatment group")
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(dfs_ind.groupby(["English group"]).std(), dfs_mat.groupby(["English group"]).mean())
```

```
Table 1a: the mean and SD of variables for the control group and treatment group
This table refers to column (4) and (1) in Table 1 of the benchmark
The mean of variables for control group and treatment group
```

English group	English-speaking ability ordinal measure	Age	Female	White	Black	Asian/Pacific Islander	Other single race	Multiracial	Hispanic	Years of schooling	Is currently married with spouse present	Is currently divorced
Control Group: Born in English-speaking country	2.980587	38.763566	0.529178	0.690274	0.225087	0.027673	0.018989	0.037978	0.012157	14.474017	0.571562	0.12218
Treatment Group : Born in non-English-speaking country	2.714701	36.616986	0.501723	0.553140	0.028159	0.118414	0.247922	0.052365	0.521571	13.089313	0.605827	0.09717

English group	Spouse English-speaking ability ordinal measure	Spouse is US-born	Spouse has the same country of birth	Spouse has the same ancestry	Spouse age	Spouse years of schooling	Years of schooling	Spouse log(wages last year)	Spouse worked last year	Both worked last year	Number of children living in same household, only individuals married spouse present	Has a chi living sam household on individua marrie wi spous prese
Control Group: Born in English-speaking country	2.978607	0.812874	0.089315	0.243836	40.771411	14.528200	14.597188	10.362746	0.881479	0.779933	1.397906	0.71037
Treatment Group : Born in non-English-speaking country	2.587381	0.499035	0.390376	0.543241	38.176907	12.965574	13.103995	10.191646	0.825765	0.701292	1.736242	0.7970

```
The standard deviation of variables for control group and treatment group
```

English group	English-speaking ability ordinal measure	Age	Female	White	Black	Asian/Pacific Islander	Other single race	Multiracial	Hispanic	Years of schooling	Is currently married with spouse present	Is currently divorced

English group	English-speaking ability ordinal measure	Age	Female	White	Black	Asian/Pacific Islander	Other single race	Multiracial	Hispanic	Years of schooling	Is currently married with spouse present	Is currently divorced
Control Group: Born in English-speaking country	0.166150	8.440097	0.499158	0.462389	0.417648	0.164036	0.136488	0.191146	0.109591	2.469523	0.494862	0.327504
Treatment Group : Born in non-English-speaking country	0.624551	8.295871	0.499999	0.497170	0.165428	0.323098	0.431808	0.222762	0.499536	3.476692	0.488674	0.296201

English group	Spouse English-speaking ability ordinal measure	Spouse is US-born	Spouse has the same country of birth	Spouse has the same ancestry	Spouse age	Spouse years of schooling	Years of schooling	Spouse log(wages last year)	Spouse worked last year	Both worked last year	Number of children living in same household, only individuals married spouse present	Has a chi sam household on individua marrie wi spous prese
Control Group: Born in English-speaking country	2.978607	0.812874	0.089315	0.243836	40.771411	14.528200	14.597188	10.362746	0.881479	0.779933	1.397906	0.71037
Treatment Group : Born in non-English-speaking country	2.587381	0.499035	0.390376	0.543241	38.176907	12.965574	13.103995	10.191646	0.825765	0.701292	1.736242	0.79704

In [4]:
```
print("Table 1b: the mean and SD of variables for the age groups")
print("This table refers to column (5), (6), (2) and (3) in Table 1 of the benchmark" )


print("The mean of variables for age groups")
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(dfs_ind.groupby(["English group","Age group"]).mean(), dfs_mat.groupby(["English group","Age
print("The standard deviation of variables for age groups")
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(dfs_ind.groupby(["English group","Age group"]).std(), dfs_mat.groupby(["English group","Age
```

```
Table 1b: the mean and SD of variables for the age groups
This table refers to column (5), (6), (2) and (3) in Table 1 of the benchmark
The mean of variables for age groups
```

Our descriptive statistics, althought slightly deviate from the benchmarks,give support to the age-language relationship picture we drawn. Late arrivers in treatment group and control group demonstrate a significant difference in English level, while earlier arrivers' English level are similar.

3. Identification and Empirical Strategy

The challenges require advanced econometrics techniques beyond ordinary least square estimation. The authors choose to use the method of instrumental variable to tackle the issues. What stands out the benchmark compared to other literature is that the authors sensibly make use of the fact that children acquire foreign languages with ease. This strategy is supported by research in physiological studies where researchers identified the physiological feature in the brain which is known as the “critical period of language acquisition” (Eric H. Lenneberg 1967).

The social outcomes are assumed to be a function of age at arrival for non-Anglophone-origin immigrants. If a person coming from a non-English speaking country immigrates to the US at a younger age, one’s English level is going to be closer to the native speakers. It is further assumed that there is no direct effect of age of arrival on the social assimilation. Therefore age-of-arrival will be suitable IV to control for the ability effect.

Our instrumental variable is define as follows:

$$k_{ija} = max(0, a-9) * I(j \quad is \quad a \quad non-English-speaking \quad country) \tag{1}$$

where *a* is age at arrival, *I*(·) is the indicator function, and *j* is country of birth

$$ENG_{ija} = \alpha_1 + \pi_1 k_{ija} + \delta_{1a} + \gamma_{1j} + W'_{ija}\rho_1 + \epsilon_{1ija} \tag{2}$$

for individual *i* born in country *j* arriving in the US at age *a*. ENG_{ija} is a measure of English language skills, δ_{1a} is a set of age-at-arrival dummies, γ_{1j} is a set of country-of-birth dummies, and w_{ija} is a vector of exogenous explanatory variables (e.g., age, sex, race).

The OLS esitmation is given by

$$y_{ija} = \alpha_1 + \beta ENG_{ija} + \delta_a + \gamma_j + W'_{ija}\rho + \epsilon_{ija} \tag{3}$$

for individual *i* born in country *j* arriving in the US at age *a*. y_{ija} is the outcome , ENG_{ija} is a measure of English language skills (the endogenous regressor), δ_a is a set of age-at-arrival dummies, γ_j is a set of country-of-birth dummies, and w_{ija} is a vector of exogenous explanatory variables (e.g., age, sex, race).

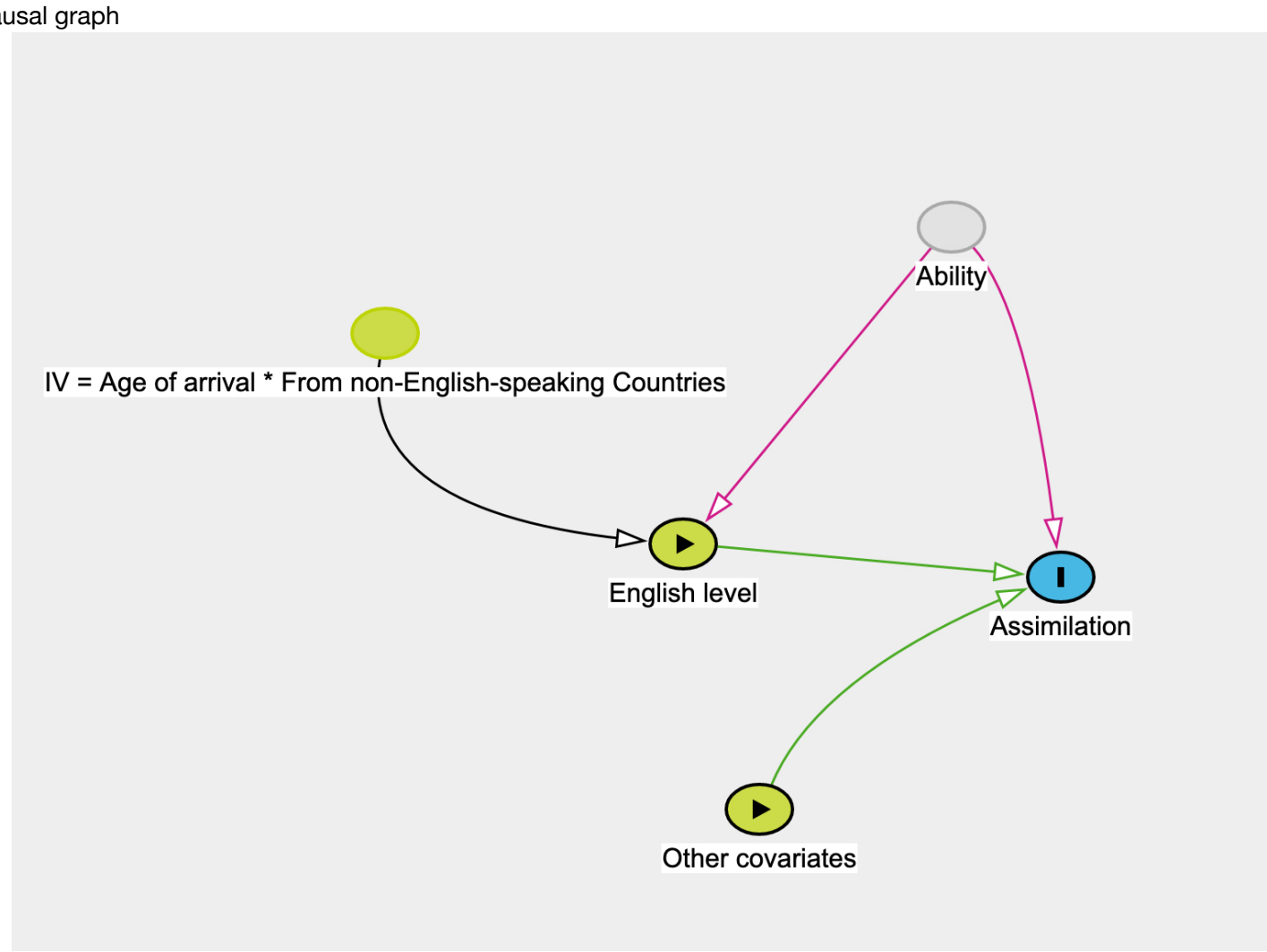
The OLS estimation of equation (3) will be biased because of endogeneity issue. The IV strategy we employed is to get the first stage esitmate of $\hat{ENG}_{ija}$ using equation (2). And then use $\hat{ENG}_{ija}$ instead of ENG_{ija} in the second stage estimate, i.e. equation (4)

$$y_{ija} = \alpha_1 + \beta^{2SLS} \hat{ENG}_{ija} + \delta_a + \gamma_j + W'_{ija}\rho + \epsilon_{ija} \tag{4}$$

The estimation of β^{2SLS} will be our target.

Our model can be visualized by a causal graph.

Figure 1. Causal graph



4. Conceptual issues in the benchmark

4.1 Social outcome and social assimilation

It seems to us that the authors are treating *social outcome* and *social assimilation* as a pair of synonymies when they are analyzing the statistics result.

It is not clear whether the variables measure the *wellness of the integration*. For example, why an immigrant having a child in the destination country implies that he or she better fitted into US society? Or why an immigrant and his or her spouse are both implies a higher level of social assimilation? We do not see any clear and strong intuition indicating the linkage between the proposed measures and our target concepts. Certainly, for some measures, a significantly low level indicates a potential integration problem. But we cannot tell if there is no problem by an increase in the level.

A thought experiment can elaborate our point. Suppose there is a group of immigrant K from a specific country. If the mean wage of K-people (when fixing other variables) is very low, then it might indicate that they on margins in the labor market. But what if the mean-wage of K-people is very high? It would also mean that they process some systematic properties that the locals do not. These properties also differentiate K-people systematically from the locals. The high mean wage does not make them more integrated into the destination society, although they are absolutely better off in terms of living standards.

An initial guess to solve the problem is to compare the difference of the mean of these variables between the immigrants and the US. If the deviation is small, then it means an immigrant is on average not having a higher or lower outcome from the locals.

However, this is not a solution neither. This issue is more complicated by the US liberal tradition. In principle, individuals are free to choose their own form of life. What does it mean if the immigrants are losing their “homeland features”? If k-people, after observing the undesirable outcome of keeping their “homeland features”, decide to abandon those. And they finally have an outcome having no difference in those outcomes than the locals. Are they really integrating in the US society in terms of the outcome? Or they are just integrating into the US on the surface, but they are not embracing the liberal spirit of the US, which more fundamental than those measure outcomes? To settle this debate, we have to go deep into a philosophical discussion of politics and culture. But that will not be our focus in this short article. We will just be cautious about the conceptual insufficiency of the result about certain outcomes.

4.2 Exogeneity of age at arrival

One assumption for using instrumental variable is that the instrument is as good as randomly assign and have no correlation with Y through other uncontrolled variables. The authors justify the exogeneity of age at arrival by stating that for childhood immigrants "age at arrival is not a choice variable since they did not time their own immigration, but merely come with their parents to the United State".

We remain skeptical concerning the exogeneity assumption. Maybe the children are not the ones *making the final call* on the immigrate decision. However, we can reasonably think that the parents, if being responsible, will take *the children's ability to assimilate into considerations* when they are deciding to immigrate or not. If the parents know that their children will have a very rough time living in a new country, they have a higher cost and less probability to immigrate. This is more easily found in East Asian nations (which are non-English-speaking) where the offspring's future is weighted heavily. In addition, for the potentially-non-assimilating children who have already entered secondary school and are unwilling to move with their parents, the parents can just send the children to boarding school or leave the children to relatives in the origins, if they have a lower cost to do so. So these non-assimilating children will not be included in the sample. This is more common in Asian or Middle-East countries (which are non-English-speaking) where family-collectivism is practiced. Siblings of the parents or grandparents have a larger moral responsibility to proxy the parents' role when the parents cannot take care of the children themselves.

These family characteristics, when being uncontrolled for, create self-selection bias where children that are expected to have higher social outcomes will be taking the treatment. This will lead to an overestimation of the treatment effect.

4.3 Cultural connotations in language

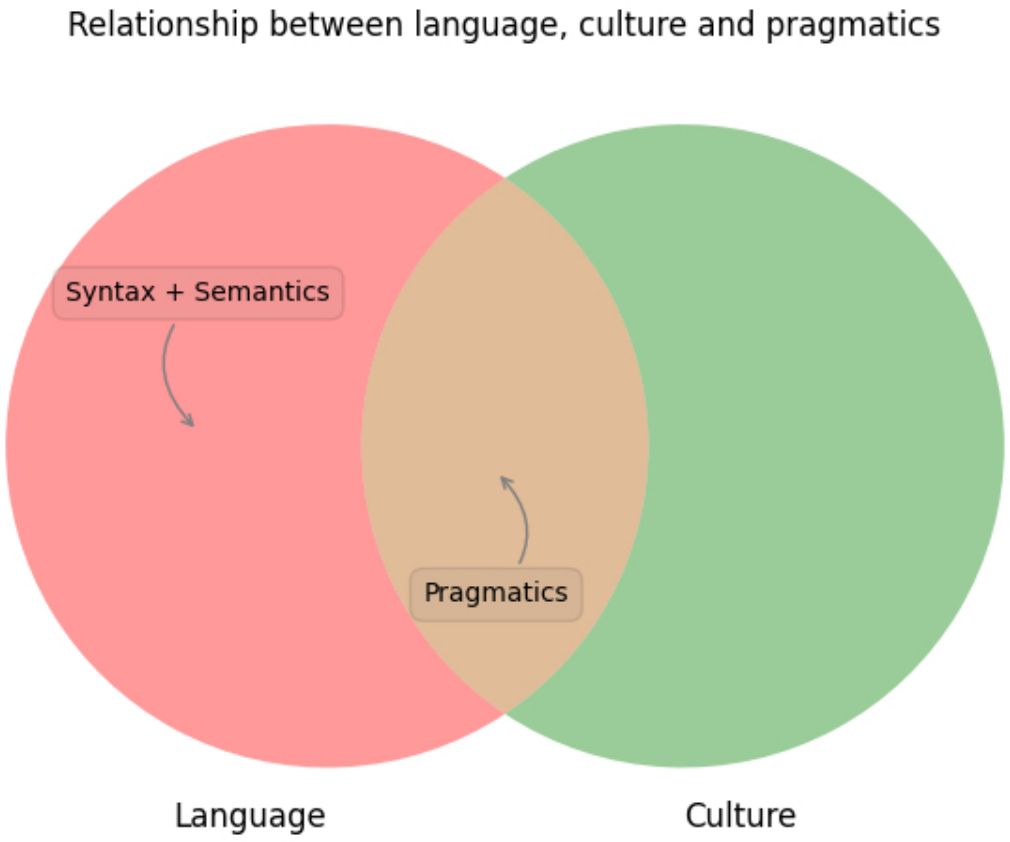
As the authors remark in the paper, controlling the effect of mastering English (language effect) cannot exclude the effect of “melting” into the US culture and institution. The authors’ solution is to incorporate immigrants from the English speaking countries as the control group, so as to control for the non-language effect. They explain their decision:

> *This is because, upon arrival in the United States, immigrants originating from English-speaking countries encounter everything that immigrants from non-English-speaking countries encounter except a new language. Thus, any difference in outcome between young and old arrivers from non-English-speaking countries that is over and above the difference from English-speaking countries can plausibly be attributed to language*. (Bleakley & Chin, 2010, p.169-170).

Singling out the language effect on immigrants from non-anglophone countries certainly has a significant methodological merit, as anglophone immigrants barely come across any enormous language barrier. But it seems to us doubtful whether the authors can really separate the language effect and the culture effect. Anglophone countries do not share the 100% same culture, however close they are.to each other. It is fairly possible that a British immigrant found the US metric system, jumbo-size Coca-Cola cups, and restaurant tipping manner very confusing. There can still be an influence of culture in the estimation of the parameters for the late-comers in treatment group.

We differ our understanding of the meaning of the parameters estimated from the authors’. Instead of claiming that the difference in outcome is completely attributed to language, we can defend our thesis better by arguing that we have to take into account the embedded cultural effect in language, by looking at the nature of language. As Morris (1949) pointed out, a (natural) language is composed of three parts: semantics, syntax and pragmatics; Semantics refers to “the relations of signs to the objects to which the signs are applicable” are; Syntax refers to “the formal relation of signs to one another”; Pragmatics is “the relations of signs to interpreters”. The authors’ mistake, as it appears to us, is to assume language skill level has only the syntaxial aspect. While anglophone immigrants have not many obstacles in the syntax of English in America, they might have problems in the pragmatics of American English. For example, when a British says “Ooh, isn’t it cold?”, what he is doing (with his words) is to invite the listener to a conversation , but an American will just encode it as merely a question about weather.

Figure 2. Relationship between language, culture and pragmatics



As it is shown by Wittgenstein (1967), “the speaking of language is part of an activity, or of a life-form”. Language is a social practice. When language users are interacting, the culture of the linguistic community is manifested. Culture is embedded in a language, through the pragmatics of it. Culture is instantiated in the language uses and culture-free language is not usable and empty. The language community, having their own culture, decide the rule of the “language game”, by which they determine who is a good speaker and who is not. In other words, to be a good language user, one has to know the culture of that linguistic community.

In our estimation, we could not separate the cultures in the pragmatics side of English in America. But a comprehensive understanding of languages implies that we do not have to so do. Rather, it accepts the cultural aspect being part of the estimator of language level. And it includes the understanding of culture when interpreting the estimation of parameters.

5. Replication

5.1 Main result

We first estimate equation (2') which is a modification of equation (2). In equation (2') the dependent variable is a social outcome rather than English proficiency.

$$Y_{ija} = \alpha_1 + \pi_1 k_{ija} + \delta_{1a} + \gamma_{1j} + W'_{ija}\rho_1 + \epsilon_{1ija} \tag{2'}$$

Equation (2') is a reduced-form equation that returns the effect of age at arrival on social outcomes for the treatment group in a regression setting. From Table 2, we can observe that age-at-arrival is significant to the majority of variables the authors are interested in, except for some fertility and all location outcomes. In the benchmark, the age-at-arrival effects are to be explained through language proficiency. If this hypothesis is right, further regressions, given by equation (4) should verify that.

In [5]:
```
df_ind['eng'].replace('  ', np.nan, inplace=True)
df_ind= df_ind.dropna(subset=['eng'])
df_ind["pwlinear"] = np.where(df_ind.agearr>=9, df_ind["agearr"]-9 ,0)
df_ind["pwlinear"].describe()
df_ind["idvar"] = df_ind["pwlinear"]*df_ind["nengdom"]
df_ind["idvar"].describe()
df_ind = (df_ind[df_ind.perwt2 !=0])

df_mat['eng'].replace('  ', np.nan, inplace=True)
df_mat= df_mat.dropna(subset=['eng'])
df_mat["young9"] = df_mat["agearr"] <=9
df_mat['eng'].replace('  ', np.nan, inplace=True)
df_mat= df_mat.dropna(subset=['eng'])
df_mat["pwlinear"] = np.where(df_mat.agearr>=9, df_mat["agearr"]-9 ,0)
df_mat["pwlinear"].describe()
df_mat["idvar"] = df_mat["pwlinear"]*df_mat["nengdom"]
df_mat["nchild_spouse"] = df_mat["nchild"]
df_mat["haskid_spouse"] = df_mat["haskid"]
df_ind["eng1"] = np.where(df_ind.eng>=1,1,0)
df_ind["eng2"] = np.where(df_ind.eng>=2,1,0)
df_ind["eng3"] = np.where(df_ind.eng>=3,1,0)
df_ind["tr9"] = df_ind["pwlinear"]*df_ind["nengdom"]
df_mat["tr9"] = df_mat["pwlinear"]*df_mat["nengdom"]
df_mat["nchild_spouse"] = df_mat["nchild"]
df_mat["haskid_spouse"] = df_mat["haskid"]
```

In [6]:
```
result_t2_ind = pd.DataFrame({"Dependent variable St": list_table2_ind})
for dep in list_table2_ind:
    get_table2_result(df_ind, dep, result_t2_ind)
result_t2_mat = pd.DataFrame({"Dependent variable St": list_table2_mat})
for dep in list_table2_mat:
    temp = get_table2_result(df_mat, dep, result_t2_mat)
result_table2 = result_t2_ind.append(result_t2_mat)


result_table2 =get_variable_full_name_table2(result_table2)
result_table2 = result_table2.sort_index(ascending=True)


get_asterisk(result_table2,"P Value", "Coef")
result_table2["Coefficient for identifying instrument variable"] = (result_table2["Parameter"].astype(st
```

```
eng1
eng2
eng3
eng
marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
spouseeng
marriednative
couplesamebpld
couplesameancestry1
spouseage
spouseyrssch
spouselnwage
spouseworkedly
bothworked
nchild_spouse
haskid_spouse
```

In [7]:
```
result_table2= result_table2[["Coefficient for identifying instrument variable","SE"]]
print("Table 2—Reduced-Form Effects")
print("This table refers to Table 2 in the benchmark")

display(result_table2)
```

```
Table 2—Reduced-Form Effects
This table refers to Table 2 in the benchmark
```

Panel	Dependent variable	Coefficient for identifying instrument variable	SE
Panel A. English proficiency measures	**1. Speaks English not well or better**	-0.006**	0.003
	2. Speaks English well or better	-0.029**	0.012
	3. Speaks English very well	-0.069***	0.013
	4. English-speaking ability ordinal measure	-0.104***	0.029
Panel B. Marital status	**1. Is currently married with spouse present**	0.011***	0.004
	2. Is currently divorced	-0.005***	0.002
	3. Has ever married	0.007***	0.003
Panel C. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	-0.086***	0.019
	2. Spouse is US-born	-0.034***	0.011
	3. Spouse has the same country of birth	0.037***	0.012
	4. Spouse has the same ancestry	0.019*	0.011
Panel D. Spouse's age and education	**1. Spouse age**	0.096***	0.035
	2. Spouse years of schooling	-0.249***	0.072
Panel E. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	-0.031***	0.012
	2. Spouse worked last year	-0.008***	0.003
	3. Both worked last year	-0.013**	0.005
Panel F. Fertility	**1. Number of children living in same household**	0.046***	0.014
	2. Has a child living in same household	0.008*	0.004
	3. Number of children living in same household, only individuals married spouse present	0.043***	0.014
	4. Has a child living in same household, only individuals married with spouse present	0.001	0.002
	5. Is a single parent	-0.002	0.003
	6. Is a never married, single parent	0.0	0.002
Panel G. Residential location	**1. Fraction of PUMA population from same country of birth**	0.001	0.001
	2. Fraction from same country of birth is above national mean for the country of birth	0.002	0.007
	3. Fraction of PUMA population with same primary ancestry	0.002	0.001
	4. Fraction with same ancestry is above national mean for the primary ancestry	0.002	0.006

The key results of the benchmark is replicated in table 3a, 3b and 3C, where we first look at the OLS and 2SLS results using equation (4) on the whole sample, male and female. English level, adjusted by the instrumental variable, affects most of the marital outcome. For example, the English skill level decreases the probability of being currently married. However, the effects on marital outcomes (that authors are interested in) do not tell a clear picture of the assimilation status of immigrants, which we explained earlier in the conceptual issues section.

In [8]:
```
#plot table 3
df_ind_t3 = get_1st_stage_result(df_ind)
df_mat_t3 = get_1st_stage_result(df_mat)

result_table3_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    get_ols_tsls_result(df_ind_t3, var, result_table3_ind)


result_table3_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    get_ols_tsls_result(df_mat_t3, var, result_table3_mat)


result_table3 = result_table3_ind.append(result_table3_mat)
result_table3 = get_variable_full_name_table3(result_table3)
result_table3 = result_table3.sort_index(ascending=True)

get_asteris_and_coef(result_table3,"OLS P Value", "OLS", "English effect(OLS)")
get_asteris_and_coef(result_table3,"2SLS P Value", "2SLS", "English effect(2SLS)")
```

```
marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
spouseeng
marriednative
couplesamebpld
couplesameancestry1
spouseage
spouseyrssch
spouselnwage
spouseworkedly
bothworked
nchild_spouse
haskid_spouse
```

In [9]:
```
print("Table 3a –Effect of English-Language Skills on Marriage Outcomes")
print("This table refers to the column (1) and (2) in Table (3),(4),(5) in the benchmark")

result_table3[["English effect(OLS)", "OLS SE" , "English effect(2SLS)", "2SLS SE", "2SLS Adj R sq"]]
```

```
Table 3a –Effect of English-Language Skills on Marriage Outcomes
This table refers to the column (1) and (2) in Table (3),(4),(5) in the benchmark
```

Out[9]:

Panel	Dependent variable	English effect(OLS)	OLS SE	English effect(2SLS)	2SLS SE	2SLS Adj R sq
Panel A. Marital status	**1. Is currently married with spouse present**	0.008	0.009	-0.108***	0.039	0.072147
	2. Is currently divorced	0.01***	0.003	0.052***	0.017	0.036220
	3. Has ever married	-0.002	0.007	-0.071***	0.024	0.153013
Panel B. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	0.512***	0.016	0.809***	0.180	0.244451
	2. Spouse is US-born	0.105***	0.009	0.322***	0.107	0.319407
	3. Spouse has the same country of birth	-0.12***	0.007	-0.351***	0.115	0.326626
	4. Spouse has the same ancestry	-0.076***	0.011	-0.18*	0.099	0.224261
Panel C. Spouse's age and education	**1. Spouse age**	-0.348***	0.035	-0.901***	0.334	0.727417
	2. Spouse years of schooling	1.353***	0.034	2.35***	0.679	0.265225
Panel D. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	0.169***	0.006	0.296***	0.111	0.193737
	2. Spouse worked last year	0.039***	0.003	0.078***	0.029	0.112868
	3. Both worked last year	0.09***	0.004	0.12**	0.048	0.033702
Panel E. Fertility	**1. Number of children living in same household**	-0.106***	0.021	-0.441***	0.136	0.150791
	2. Has a child living in same household	-0.005	0.004	-0.073**	0.037	0.120078
	3. Number of children living in same household, only individuals married spouse present	-0.177***	0.013	-0.41***	0.131	0.165262
	4. Has a child living in same household, only individuals married with spouse present	-0.019***	0.004	-0.007	0.020	0.101348
	5. Is a single parent	-0.001	0.004	0.021	0.026	0.061569
	6. Is a never married, single parent	-0.003	0.002	-0.002	0.017	0.034528
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	-0.01***	0.001	-0.007	0.007	0.448813
	2. Fraction from same country of birth is above national mean for the country of birth	-0.05***	0.007	-0.015	0.065	0.030742
	3. Fraction of PUMA population with same primary ancestry	-0.013***	0.002	-0.018	0.013	0.357922
	4. Fraction with same ancestry is above national mean for the primary ancestry	-0.041***	0.007	-0.022	0.057	0.024823

In [10]:
```
# plot table 3 female sample
df_ind_t3_female = df_ind[df_ind["female"] == 1]
df_mat_t3_female = df_mat[df_mat["female"] == 1]

df_ind_t3_female = get_1st_stage_result(df_ind_t3_female)
df_mat_t3_female = get_1st_stage_result(df_mat_t3_female)


result_table3_female_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_ols, temp_2SLS) = get_ols_tsls_result(df_ind_t3_female, var, result_table3_female_ind)


result_table3_female_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_ols, temp_2SLS) = get_ols_tsls_result(df_mat_t3_female, var, result_table3_female_mat)


result_table3_female = result_table3_female_ind.append(result_table3_female_mat)

result_table3_female = get_variable_full_name_table3(result_table3_female)
result_table3_female = result_table3_female.sort_index(ascending=True, sort_remaining =False)


get_asteris_and_coef(result_table3_female,"OLS P Value", "OLS", "English effect(OLS)")
get_asteris_and_coef(result_table3_female,"2SLS P Value", "2SLS", "English effect(2SLS)")
```

```
/Users/chingfungchow/Documents/GitHub/Microeconometrics-Paper/auxiliary.py:22: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  data['eng_hat'] = fit_st1.predict(data)
/Users/chingfungchow/opt/anaconda3/lib/python3.7/site-packages/pandas/core/generic.py:6746: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  self._update_inplace(new_data)

marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
spouseeng
marriednative
couplesamebpld
couplesameancestry1
spouseage
spouseyrssch
spouselnwage
spouseworkedly
bothworked
nchild_spouse
haskid_spouse
```

In [11]:
```
print("Table 3b –Effect of English-Language Skills on Marriage Outcomes, female sample")
print("This table refers to the column (3) and (4) in Table (3),(4),(5) in the benchmark")
result_table3_female[["English effect(OLS)", "OLS SE" , "English effect(2SLS)", "2SLS SE", "2SLS Adj R s
```

```
Table 3b –Effect of English-Language Skills on Marriage Outcomes, female sample
This table refers to the column (3) and (4) in Table (3),(4),(5) in the benchmark
```

Out[11]:

Panel	Dependent variable	English effect(OLS)	OLS SE	English effect(2SLS)	2SLS SE	2SLS Adj R sq
Panel A. Marital status	**1. Is currently married with spouse present**	-0.004	0.009	-0.076**	0.033	0.062049
	2. Is currently divorced	0.015***	0.003	0.064**	0.027	0.035673
	3. Has ever married	-0.001	0.005	-0.015	0.027	0.136230
Panel B. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	0.474***	0.008	0.615***	0.109	0.232672
	2. Spouse is US-born	0.101***	0.011	0.27**	0.118	0.340200
	3. Spouse has the same country of birth	-0.118***	0.009	-0.285**	0.127	0.338575
	4. Spouse has the same ancestry	-0.078***	0.013	-0.17	0.112	0.226091
Panel C. Spouse's age and education	**1. Spouse age**	-0.496***	0.055	-1.312***	0.461	0.707548
	2. Spouse years of schooling	1.367***	0.036	2.188***	0.592	0.264124
Panel D. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	0.166***	0.006	0.233**	0.114	0.118192
	2. Spouse worked last year	0.021***	0.005	0.022	0.019	0.027787
	3. Both worked last year	0.106***	0.006	0.12***	0.042	0.027167
Panel E. Fertility	**1. Number of children living in same household**	-0.162***	0.019	-0.472***	0.128	0.165340
	2. Has a child living in same household	-0.019***	0.003	-0.041	0.034	0.116814
	3. Number of children living in same household, only individuals married spouse present	-0.203***	0.017	-0.591***	0.140	0.193076
	4. Has a child living in same household, only individuals married with spouse present	-0.024***	0.004	-0.02	0.021	0.123443
	5. Is a single parent	-0.001	0.007	0.032	0.033	0.048728
	6. Is a never married, single parent	-0.005	0.004	-0.005	0.026	0.046236
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	-0.012***	0.002	-0.014***	0.004	0.464532
	2. Fraction from same country of birth is above national mean for the country of birth	-0.058***	0.006	-0.038	0.046	0.036516
	3. Fraction of PUMA population with same primary ancestry	-0.017***	0.002	-0.033***	0.006	0.369076
	4. Fraction with same ancestry is above national mean for the primary ancestry	-0.05***	0.007	-0.088**	0.041	0.030300

In [12]:
```
# plot table 3 male sample
df_ind_t3_male = df_ind[df_ind["female"] == 0]
df_mat_t3_male = df_mat[df_mat["female"] == 0]

df_ind_t3_male = get_1st_stage_result(df_ind_t3_male)
df_mat_t3_male = get_1st_stage_result(df_mat_t3_male)


result_table3_male_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_ols, temp_2SLS) = get_ols_tsls_result(df_ind_t3_male, var, result_table3_male_ind)


result_table3_male_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_ols, temp_2SLS) = get_ols_tsls_result(df_mat_t3_male, var, result_table3_male_mat)


result_table3_male = result_table3_male_ind.append(result_table3_male_mat)

result_table3_male = get_variable_full_name_table3(result_table3_male)
result_table3_male = result_table3_male.sort_index(ascending=True, sort_remaining =False)


get_asteris_and_coef(result_table3_male,"OLS P Value", "OLS", "English effect(OLS)")
get_asteris_and_coef(result_table3_male,"2SLS P Value", "2SLS", "English effect(2SLS)")
```

```
/Users/chingfungchow/Documents/GitHub/Microeconometrics-Paper/auxiliary.py:22: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  data['eng_hat'] = fit_st1.predict(data)
/Users/chingfungchow/opt/anaconda3/lib/python3.7/site-packages/pandas/core/generic.py:6746: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  self._update_inplace(new_data)

marriedpresent
divorced
evermarried
```

In [13]:
```
print("Table 3c —Effect of English-Language Skills on Marriage Outcomes, male sample")
print("This table refers to the column (5) and (6) in Table (3),(4),(5) in the benchmark")
result_table3_male[["English effect(OLS)", "OLS SE" , "English effect(2SLS)", "2SLS SE", "2SLS Adj R sq'
```

```
Table 3c —Effect of English-Language Skills on Marriage Outcomes, male sample
This table refers to the column (5) and (6) in Table (3),(4),(5) in the benchmark
```

Out[13]:

Panel	Dependent variable	English effect(OLS)	OLS SE	English effect(2SLS)	2SLS SE	2SLS Adj R sq
Panel A. Marital status	**1. Is currently married with spouse present**	0.019*	0.011	-0.14***	0.048	0.092846
	2. Is currently divorced	0.006*	0.004	0.038	0.023	0.031216
	3. Has ever married	-0.004	0.009	-0.133***	0.046	0.164015
Panel B. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	0.56***	0.027	1.028***	0.271	0.259198
	2. Spouse is US-born	0.109***	0.006	0.386***	0.110	0.310249
	3. Spouse has the same country of birth	-0.122***	0.006	-0.434***	0.104	0.326044
	4. Spouse has the same ancestry	-0.074***	0.009	-0.197**	0.089	0.229249
Panel C. Spouse's age and education	**1. Spouse age**	-0.147***	0.033	-0.382	0.648	0.727247
	2. Spouse years of schooling	1.343***	0.043	2.582***	0.848	0.272656
Panel D. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	0.172***	0.012	0.395**	0.184	0.055268
	2. Spouse worked last year	0.062***	0.004	0.141**	0.064	0.042286
	3. Both worked last year	0.072***	0.004	0.115*	0.067	0.041955
Panel E. Fertility	**1. Number of children living in same household**	-0.049*	0.026	-0.403**	0.165	0.133685
	2. Has a child living in same household	0.009	0.007	-0.113**	0.049	0.107167
	3. Number of children living in same household, only individuals married spouse present	-0.148***	0.013	-0.19	0.183	0.149393
	4. Has a child living in same household, only individuals married with spouse present	-0.014***	0.004	0.008	0.035	0.087784
	5. Is a single parent	-0.001	0.001	0.0	0.015	0.007007
	6. Is a never married, single parent	-0.002**	0.001	-0.008	0.008	0.011830
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	-0.008***	0.001	-0.0	0.011	0.433569
	2. Fraction from same country of birth is above national mean for the country of birth	-0.041***	0.009	0.007	0.091	0.026577
	3. Fraction of PUMA population with same primary ancestry	-0.009***	0.001	-0.002	0.021	0.347328
	4. Fraction with same ancestry is above national mean for the primary ancestry	-0.033***	0.008	0.056	0.084	0.020615

The coefficients for "Spouse English-speaking ability", "Spouse is US-born" and "Spouse has the same country of birth" are more direct indicators to an immigrant's status: His or her English level moves in the same direction with the spouse English ability and the probability of the spouse being US-born, and it moves in the opposite direction with the probability of marrying someone from the same country of birth. These results imply that English level helps an immigrant to depart from the social circle of his or her countrypeople and be able to build relationships, in one most intimate form, with the US locals. Similarly, English also improves the spouse labor outcome, such as wage and employment status. It is worth noticing that for female immigrants, higher English level statistically significantly decreases the age of their spouses, while we do not observe a significant effect for male respondents.

English level does not bring significant effects to fertility outcomes, except for the decrease in the number of children. It might be caused by the fact that the mean age in the treatment group (36.54) is slightly less than that in the control group (38.403). In addition, English skills bear no demonstrative improvement in ethnic enclaving or resident location outcomes. The technicality issue is that due to data privacy, the lowest level of geographic aggregation measure "public-use microdata area" (PUMA), which is an area containing at least 100,000 people. As a result, our definition for a neighborhood becomes too big, leading to the unsatisfactory theoretical ground for our estimation.

However, in the authors' analysis, the adjusted R-square of each estimation is not presented. As soon as we pay attention to the goodness-of-fit, we discover that equation (4) does a poor job in giving precise predictions. 2SLS estimations on an outcome such as "is currently married with spouse present", "is currently divorced", "has a child living in the same household", "is a single parent" and "spouse worked last year" have an adjusted R-squared less than 0.04, which as a sign that there is just too much residual. The poor goodness-of-fit of the outcome with currently married (or divorced) might be consistent with our previous analysis. The effect of English on an immigrant's decision on marriage is ambiguous, as there are reasonable theories to predict the opposite results.

5.2 Robustness checks in the benchmark

It is possible that in table 3a,3b and 3c there are uncontrol features for the treatment group (childhood immigrants from non-English-speaking countries) that will be included in the 2SLS estimator. The authors consider alternative hypotheses for robustness check, in order to eliminate the non-language effect in the estimations. The robustness checks done by the authors are summarized in table 4a. The results in alternative specifications, as shown in table 4b, are similar to the baseline 2SLS analysis.

Table 4a — Summary of robustness check in the benchmark: Feature, movitivation and solution

Feature	Motivation	Solution
Age-of-arrival × Place-of-birth development effect	English countries in general are better developed than in non-English countries. Immigrants from poorer non-Anglo countries might have received worse education before they arrive at the US. The age-of-arrival effect estimated might be mixed with the place-of-birth development effect and is therefore biased.	Control for an interaction term between the age-of-arrival and per capita GDP in the country of birth in 1980 (obtained from the Penn World Tables).
Age-of-arrival × Origin fertility rate effect	Age-at-arrival effect could depend on the fertility rate in the origin country. The fertility rate of the US is higher than in most developed countries, but lower than most developing countries. For immigrants from an English country (which is more likely to be developed), integrating in the US society means having a higher fertility rate. For immigrants from a non-English country (which is less likely to be developed), integrating in the US society means having a lower fertility rate. The age-of-arrival effect on fertility outcome estimated did not take the origin's fertility rate into account.	Control for an interaction term between the age-of-arrival and total fertility rate in the country of birth in 1982 (obtained from the World Development Indicators).
Culture effect	Immigrants from English speaking countries have experienced a culture and institution that are more similar to the one in the US compared to immigrants from non-English countries, regardless of age at arrival.	Restrict analysis to groups of countries that might be more similar to each other. Exclude Canadian and Mexican immigrants separately, as they might be seen as poor control for assimilation process for a representative immigrant

In [14]:
```
####################get baseline result (from table 3 )

result_table4_baseline = result_table3[["English effect(2SLS)", "2SLS SE"]]
# result_table4_baseline = result_table4_baseline.rename(columns={"English effect(2SLS)":"Base results (

####################get result with control in origin's GDP
df_ind_GDP = df_ind
df_mat_GDP = df_mat


df_ind_GDP['lngdp'].replace('  ', np.nan, inplace=True)
df_ind_GDP= df_ind_GDP.dropna(subset=['lngdp'])
df_mat_GDP['lngdp'].replace('  ', np.nan, inplace=True)
df_mat_GDP= df_mat_GDP.dropna(subset=['lngdp'])

df_ind_GDP["pwxgdp"]=(df_ind_GDP["pwlinear"]*df_ind_GDP["lngdp"])/100
df_mat_GDP["pwxgdp"]=(df_mat_GDP["pwlinear"]*df_mat_GDP["lngdp"])/100

df_ind_GDP = get_first_stage_result_table4(df_ind_GDP ,"pwxgdp" )
df_mat_GDP = get_first_stage_result_table4(df_mat_GDP, "pwxgdp")


result_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_2SLS) = get_tsls_result_table4(df_ind_GDP, var, result_ind,"pwxgdp")

result_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_2SLS) = get_tsls_result_table4(df_mat_GDP, var, result_mat,"pwxgdp")

result_table4_control_GDP = result_ind.append(result_mat)


result_table4_control_GDP = get_variable_full_name_table3(result_table4_control_GDP)
get_asteris_and_coef(result_table4_control_GDP,"2SLS P Value", "2SLS", "Control for origin GDP × age at
result_table4_control_GDP = result_table4_control_GDP[["Control for origin GDP × age at arrival (2)", "2

####################get result with control in origin's fertility
df_ind_tfr = df_ind
df_mat_tfr = df_mat

df_ind_tfr['tfr82'].replace('  ', np.nan, inplace=True)
df_ind_tfr= df_ind_tfr.dropna(subset=['tfr82'])

df_mat_tfr['tfr82'].replace('  ', np.nan, inplace=True)
df_mat_tfr= df_mat_tfr.dropna(subset=['tfr82'])

df_ind_tfr["pwxtfr"]=(df_ind_tfr["pwlinear"]*df_ind_tfr["tfr82"])
df_mat_tfr["pwxtfr"]=(df_mat_tfr["pwlinear"]*df_mat_tfr["tfr82"])

df_ind_tfr = get_first_stage_result_table4(df_ind_tfr ,"pwxtfr" )
df_mat_tfr = get_first_stage_result_table4(df_mat_tfr, "pwxtfr")


result_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_2SLS) = get_tsls_result_table4(df_ind_tfr, var, result_ind,"pwxtfr")

result_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_2SLS) = get_tsls_result_table4(df_mat_tfr, var, result_mat,"pwxtfr")

result_table4_control_fertility = result_ind.append(result_mat)
result_table4_control_fertility = get_variable_full_name_table3(result_table4_control_fertility)
get_asteris_and_coef(result_table4_control_fertility,"2SLS P Value", "2SLS", "Control for origin fertili
result_table4_control_fertility = result_table4_control_fertility[["Control for origin fertility × age a

####################get result dropping immigrants from Canada
df_ind_no_CA = df_ind

df_mat_no_CA = df_mat

df_ind_no_CA = (df_ind_no_CA[df_ind_no_CA.bpld !=15000])
df_mat_no_CA = (df_mat_no_CA[df_mat_no_CA.bpld !=15000])

df_ind_no_CA = get_first_stage_result_table4(df_ind_no_CA  )
df_mat_no_CA = get_first_stage_result_table4(df_mat_no_CA)


result_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_2SLS) = get_tsls_result_table4(df_ind_no_CA, var, result_ind)

result_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_2SLS) = get_tsls_result_table4(df_mat_no_CA, var, result_mat)

result_table4_no_CA = result_ind.append(result_mat)
result_table4_no_CA = get_variable_full_name_table3(result_table4_no_CA)
```

```
get_asteris_and_coef(result_table4_no_CA,"2SLS P Value", "2SLS", "Drop Canada (4)")
result_table4_no_CA = result_table4_no_CA[["Drop Canada (4)", "2SLS SE"]]


####################get result dropping immigrants from Mexico
df_ind_no_Mex = df_ind

df_mat_no_Mex = df_mat

df_ind_no_Mex = (df_ind_no_Mex[df_ind_no_Mex.bpld !=20000])
df_mat_no_Mex = (df_mat_no_Mex[df_mat_no_Mex.bpld !=20000])
df_ind_no_Mex = get_first_stage_result_table4(df_ind_no_Mex)
df_mat_no_Mex = get_first_stage_result_table4(df_mat_no_Mex)

result_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
for var in list_table3_ind:
    (temp_2SLS) = get_tsls_result_table4(df_ind_no_Mex, var, result_ind)

result_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
for var in list_table3_mat:
    (temp_2SLS) = get_tsls_result_table4(df_mat_no_Mex, var, result_mat)

result_table4_no_Mex = result_ind.append(result_mat)

result_table4_no_Mex = get_variable_full_name_table3(result_table4_no_Mex)
get_asteris_and_coef(result_table4_no_Mex,"2SLS P Value", "2SLS", "Drop Mexico (5)")
result_table4_no_Mex = result_table4_no_Mex[["Drop Mexico (5)", "2SLS SE"]]
result_table4_no_Mex
```

```
result_table4_baseline=result_table4_baseline.rename(columns={ "English effect(2SLS)":"Base results (1)"
result_table4_control_GDP=result_table4_control_GDP.rename(columns={"2SLS SE": "SE(2)"})
result_table4_control_fertility=result_table4_control_fertility.rename(columns={"2SLS SE": "SE(3)"})
result_table4_no_CA=result_table4_no_CA.rename(columns={"2SLS SE": "SE(4)"})
result_table4_no_Mex=result_table4_no_Mex.rename(columns={"2SLS SE": "SE(5)"})


result_table4 = pd.merge(result_table4_baseline, result_table4_control_GDP,
                         left_on=["Panel",'Dependent variable'], right_on = ["Panel",'Dependent variable'
result_table4 = pd.merge(result_table4, result_table4_control_fertility,
                         left_on=["Panel",'Dependent variable'], right_on = ["Panel",'Dependent variable']
result_table4 = pd.merge(result_table4, result_table4_no_CA,
                         left_on=["Panel",'Dependent variable'], right_on = ["Panel",'Dependent variable']
result_table4 = pd.merge(result_table4, result_table4_no_Mex,
                         left_on=["Panel",'Dependent variable'], right_on = ["Panel",'Dependent variable']

result_table4 = result_table4.sort_index(ascending=True, sort_remaining =False)
```

```
marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
```

In [15]:
```
print("Table 4b – 2SLS Effect of English Using Alternative Samples and Specifications")
print("This table refers to Table (6) in the benchmark")
with pd.option_context('display.max_rows', None, 'display.max_columns', None):
    display(result_table4)
```

Panel E. Fertility	**1. Number of children living in same household**	-0.441***	0.136	-0.399***	0.149	-0.349***	0.110	-0.414***	0.144	-0.476***	0.150
	2. Has a child living in same household	-0.073**	0.037	-0.041	0.028	-0.058	0.042	-0.058	0.043	-0.113*	0.064
	3. Number of children living in same household, only individuals married spouse present	-0.41***	0.131	-0.397***	0.134	-0.326***	0.116	-0.383***	0.136	-0.433***	0.143
	4. Has a child living in same household, only individuals married with spouse present	-0.007	0.020	0.004	0.018	-0.003	0.032	0.011	0.018	-0.008	0.046
	5. Is a single parent	0.021	0.026	0.034*	0.020	0.025	0.032	0.033	0.027	0.01	0.045
	6. Is a never married, single parent	-0.002	0.017	0.008	0.014	0.001	0.021	0.005	0.020	-0.009	0.031
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	-0.007	0.007	0.0	0.003	-0.014	0.011	-0.005	0.007	-0.027**	0.012
	2. Fraction from same										

6. Robustness checks and extensions

6.1 Durbin Wu Hausman test

We first examine the endogeneity of English proficiency, or in other words, whether employing IV add merits to our analysis. Although the benchmark authors test for endogeneity by comparing OLS and IV estimate, we can also test for endogeneity in a more formal way by running the Durbin–Wu–Hausman test on English proficiency.

As stated before, our equation (2) is:

$$ENG_{ija} = \alpha_1 + \pi_1 k_{ija} + \delta_{1a} + \gamma_{1j} + W'_{ija}\rho_1 + \epsilon_{1ija} \tag{2}$$

By equation (2), we get the estimation of ϵ_{1ija} ie. $\hat{\epsilon_{1ija}}$

Durbin–Wu–Hausman test result given by

$$y_{ija} = \alpha_1 + \beta ENG_{ija} + \omega\hat{\epsilon_{ija}} + \delta_a + \gamma_j + W'_{ija}\rho + \mu_{ija} \tag{5}$$

The estimation of ω will be our target.

In [16]:
```
WH Test

t_st1 = wls( "eng ~ idvar + C(agearr) + C(age) + C(bpld) + female + black + asianpi + other + multi + his
_ind['Epsilon']  = fit_st1.resid
_ind['eng_hat'] = fit_st1.predict(df_ind)


sult_DWH_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
r var in list_table3_ind:
 get_Durbin_Wu_Hausman_result(df_ind, var,result_DWH_ind )
t_st1 = wls( "eng ~ idvar + C(agearr) + C(age) + C(bpld) + female + black + asianpi + other + multi + his
_mat['Epsilon']  = fit_st1.resid
_mat['eng_hat'] = fit_st1.predict(df_mat)


sult_DWH_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
r var in list_table3_mat:
 get_Durbin_Wu_Hausman_result(df_mat, var,result_DWH_mat )
sult_DWH = result_DWH_ind.append(result_DWH_mat)
sult_DWH = get_variable_full_name_table3(result_DWH)
sult_DWH = result_DWH.sort_index(ascending=True, sort_remaining =False)

t_asteris_and_coef(result_DWH,"Epsilon P Value", "Epsilon", "Epsilon")

sult_DWH=result_DWH[["Epsilon","Epsilon SE"]]
```

```
marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
spouseeng
marriednative
couplesamebpld
couplesameancestry1
spouseage
spouseyrssch
spouselnwage
spouseworkedly
bothworked
nchild_spouse
haskid_spouse
```

In [17]:
```
print("Table 5 –Durbin Wu Hausman test result")
result_DWH
```

```
Table 5 –Durbin Wu Hausman test result
```

Out[17]:

Panel	Dependent variable	Epsilon	Epsilon SE
Panel A. Marital status	**1. Is currently married with spouse present**	0.117***	0.036
	2. Is currently divorced	-0.043***	0.016
	3. Has ever married	0.07***	0.023
Panel B. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	-0.303***	0.032
	2. Spouse is US-born	-0.22**	0.104
	3. Spouse has the same country of birth	0.234**	0.108
	4. Spouse has the same ancestry	0.106	0.104
Panel C. Spouse's age and education	**1. Spouse age**	0.558*	0.329
	2. Spouse years of schooling	-1.022***	0.348
Panel D. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	-0.131	0.086
	2. Spouse worked last year	-0.04*	0.024
	3. Both worked last year	-0.032	0.031
Panel E. Fertility	**1. Number of children living in same household**	0.339***	0.100
	2. Has a child living in same household	0.069*	0.036
	3. Number of children living in same household, only individuals married spouse present	0.235***	0.081
	4. Has a child living in same household, only individuals married with spouse present	-0.012	0.020
	5. Is a single parent	-0.022	0.024
	6. Is a never married, single parent	-0.001	0.017
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	-0.003	0.010
	2. Fraction from same country of birth is above national mean for the country of birth	-0.035	0.072
	3. Fraction of PUMA population with same primary ancestry	0.005	0.017
	4. Fraction with same ancestry is above national mean for the primary ancestry	-0.02	0.062

The test evaluate the endogeneity of English proficiency by verifying the economic and statistics significance of the coefficient for epsilon hat. If $\hat{\omega}$ is significant, it is implied that that endogenous regressors' effects on the estimates are significant, and an IV strategy is needed. Vice versa, if residual is not statistically significantly different from zero, so conclude that there is no endogeneity bias in the OLS estimates. Using DWH test, we discovered that with regard to quite a lot of outcomes, we have no statistical reasons to use English proficiency as the instrument. The exceptions are “Spouse English-speaking ability ordinal measure”, “Spouse years of schooling” , “Spouse age”, “Number of children living in same household”.

However, although the DWH test result is not satisfactory, we might still defend using this IV approach by arguing an "economics theory comes first" attitude. The statistical failure is a result caused by the imperfectness of the data on hand and possibly committing variable biases. As long as the economics reasoning is backed by a strong economics theory and intuition, we should feel confident employing the current IV.

6.2 Control for residence location

It is known that different states have different cultures and policies, and therefore different acceptance towards immigrants. Some states have an immigrant tradition such as Los Angeles and New York while some do not. It is possible that some locations of residence are more popular within non-English-speaking immigrants. For example, California has a lot of Asian immigrants because of the location and its immigrant heritage. Likewise, Mexican and South American immigrants are clustered in southern states such as Texas, New Mexico and Florida. States with a stronger immigrants heritage are easier for immigrants to assimilate. So the result might be biased. We therefore include a vector of dummy variables of state of residence and interaction between birthplace and state of residence into the model.

In [18]:
```
result_R2_ind = pd.DataFrame({"Dependent variable St": list_table3_ind})
fit_st1 = wls( "eng ~ idvar + C(agearr) + C(age) + C(bpld)+ C(statefip)*pwlinear + female + black + asia
df_ind['eng_hat'] = fit_st1.predict(df_ind)

for var in list_table3_ind:
    get_ols_tsls_result_Robust2(df_ind, var, result_R2_ind)

result_R2_mat = pd.DataFrame({"Dependent variable St": list_table3_mat})
fit_st1 = wls( "eng ~ idvar + C(agearr) + C(age) + C(bpld)+ C(statefip)*pwlinear + female + black + asia
df_mat['eng_hat'] = fit_st1.predict(df_mat)

for var in list_table3_mat:
    get_ols_tsls_result_Robust2(df_mat, var, result_R2_mat)


result_R2 = result_R2_ind.append(result_R2_mat)
result_R2 = get_variable_full_name_table3(result_R2)
result_R2 = result_R2.sort_index(ascending=True, sort_remaining =False)
get_asteris_and_coef(result_R2,"2SLS P Value", "2SLS", "English effect(2SLS)")

result_R2 = result_R2[["English effect(2SLS)", "2SLS SE"]]
```

```
marriedpresent
divorced
evermarried
nchild
haskid
singleparent
nevermarried_haskid
share_bpld_minusself
abovemean_bpld2
ancestpct_minusself
abovemean_ancestry2
spouseeng
marriednative
couplesamebpld
couplesameancestry1
spouseage
spouseyrssch
spouselnwage
spouseworkedly
bothworked
nchild_spouse
haskid_spouse
```

In [19]:
```
print("Table 6 — English effect with control on the residence location (state level)")
result_R2
```

```
Table 6 — English effect with control on the residence location (state level)
```

Out[19]:

Panel	Dependent variable	English effect(2SLS)	2SLS SE
Panel A. Marital status	**1. Is currently married with spouse present**	-0.096**	0.044
	2. Is currently divorced	0.05**	0.020
	3. Has ever married	-0.077***	0.028
Panel B. Spouse's nativity and ethnicity	**1. Spouse English-speaking ability ordinal measure**	0.852***	0.140
	2. Spouse is US-born	0.385***	0.117
	3. Spouse has the same country of birth	-0.403***	0.131
	4. Spouse has the same ancestry	-0.225**	0.114
Panel C. Spouse's age and education	**1. Spouse age**	-1.085***	0.421
	2. Spouse years of schooling	2.514***	0.604
Panel D. Spouse's labor market outcomes	**1. Spouse log(wages last year)**	0.321**	0.125
	2. Spouse worked last year	0.078**	0.032
	3. Both worked last year	0.113**	0.045
Panel E. Fertility	**1. Number of children living in same household**	-0.39***	0.120
	2. Has a child living in same household	-0.067	0.044
	3. Number of children living in same household, only individuals married spouse present	-0.371***	0.104
	4. Has a child living in same household, only individuals married with spouse present	0.007	0.027
	5. Is a single parent	0.007	0.025
	6. Is a never married, single parent	-0.005	0.018
Panel F. Residential location	**1. Fraction of PUMA population from same country of birth**	0.0	0.013
	2. Fraction from same country of birth is above national mean for the country of birth	-0.011	0.062
	3. Fraction of PUMA population with same primary ancestry	-0.001	0.023
	4. Fraction with same ancestry is above national mean for the primary ancestry	0.031	0.062

The result is displayed in table 6. We can observe similar outcomes from the baseline model presented in table 4_. One exception is the English effect on the variable "spouse has the same primary ancestry". In the baseline model, the effect is -0.181 and insignificant at 10% level. When we control for the interaction term, effect is -0.225 and significant at 5% level.

6.3 English effect on outcome gap between immigrants and spouses

We extend the authors' analysis to other outcomes that we are interested in. We use the baseline model to investigate the gap of age, years of schooling, and income between spouses. These three gaps can be seen as a proxy for age closeness, education closeness and economic closeness between a couple. Apart from the bassline model, we also add controls for the immigrants' level and their partners' level of the dependent variables and observe the effect.

In [20]:
```
df_mat_R3 = df_mat
df_mat_R3['age'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['age'])
df_mat_R3['spouseage'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['spouseage'])
df_mat_R3['yrssch'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['yrssch'])
df_mat_R3['spouseyrssch'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['spouseyrssch'])
df_mat_R3['lnwagely'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['lnwagely'])
df_mat_R3['spouselnwage'].replace('  ', np.nan, inplace=True)
df_mat_R3= df_mat_R3.dropna(subset=['spouselnwage'])


df_mat_R3["Wage_difference"] = abs(df_mat_R3["lnwagely"] - df_mat_R3["spouselnwage"])
df_mat_R3["Age_difference"] = abs(df_mat_R3["age"] - df_mat_R3["spouseage"])
df_mat_R3["Education_difference"] = abs(df_mat_R3["yrssch"] - df_mat_R3["spouseyrssch"])

list_mat_R3 = ["Wage_difference", "Age_difference", "Education_difference"    ]

result_mat_R3 = pd.DataFrame({"Dependent variable St": list_mat_R3})

get_tsls_result_table4(df_mat_R3, "Wage_difference", result_mat_R3,"lnwagely + spouselnwage" )
get_tsls_result_table4(df_mat_R3, "Education_difference", result_mat_R3, "yrssch +spouseyrssch ")
get_tsls_result_table4(df_mat_R3, "Age_difference", result_mat_R3,"spouseage" )

get_asterisk(result_mat_R3,"2SLS P Value", "2SLS")
result_mat_R3["2SLS Parameter (Full Sample)"] = (result_mat_R3["2SLS Parameter"].astype(str) + result_ma
result_mat_R3=result_mat_R3.rename(columns={ "Dependent variable St":"Dependent variable", "2SLS SE": "S
result_mat_R3 = result_mat_R3.set_index('Dependent variable')
result_mat_R3= result_mat_R3[["2SLS Parameter (Full Sample)", "SE (Full Sample)"]]


# subset female sample
df_mat_R3_female = df_mat_R3[df_mat_R3["female"] == 1]
result_mat_R3_female = pd.DataFrame({"Dependent variable St": list_mat_R3})

get_tsls_result_table4(df_mat_R3_female, "Wage_difference", result_mat_R3_female,"lnwagely + spouselnwag
get_tsls_result_table4(df_mat_R3_female, "Education_difference", result_mat_R3_female, "yrssch +spouseyr
get_tsls_result_table4(df_mat_R3_female, "Age_difference", result_mat_R3_female,"spouseage" )

get_asterisk(result_mat_R3_female,"2SLS P Value", "2SLS")
result_mat_R3_female["2SLS Parameter (Female Sample)"] = (result_mat_R3_female["2SLS Parameter"].astype(
result_mat_R3_female=result_mat_R3_female.rename(columns={ "Dependent variable St":"Dependent variable",
result_mat_R3_female = result_mat_R3_female.set_index('Dependent variable')
result_mat_R3_female= result_mat_R3_female[["2SLS Parameter (Female Sample)", "SE (Female Sample)"]]


# subset male sample
df_mat_R3_male = df_mat_R3[df_mat_R3["female"] == 0]
result_mat_R3_male = pd.DataFrame({"Dependent variable St": list_mat_R3})

get_tsls_result_table4(df_mat_R3_male, "Wage_difference", result_mat_R3_male,"lnwagely + spouselnwage" )
get_tsls_result_table4(df_mat_R3_male, "Education_difference", result_mat_R3_male, "yrssch +spouseyrssch
get_tsls_result_table4(df_mat_R3_male, "Age_difference", result_mat_R3_male,"spouseage" )

get_asterisk(result_mat_R3_male,"2SLS P Value", "2SLS")
result_mat_R3_male["2SLS Parameter (Male Sample)"] = (result_mat_R3_male["2SLS Parameter"].astype(str) +
result_mat_R3_male=result_mat_R3_male.rename(columns={ "Dependent variable St":"Dependent variable", "2S
result_mat_R3_male = result_mat_R3_male.set_index('Dependent variable')
result_mat_R3_male= result_mat_R3_male[["2SLS Parameter (Male Sample)", "SE (Male Sample)"]]


result_mat_R3 = pd.merge(result_mat_R3, result_mat_R3_female,
                         left_on=['Dependent variable'], right_on = ['Dependent variable'], how = 'left')
result_mat_R3 = pd.merge(result_mat_R3, result_mat_R3_male,
                         left_on=['Dependent variable'], right_on = ['Dependent variable'], how = 'left')
```

```
Wage_difference
Education_difference
Age_difference

/Users/chingfungchow/opt/anaconda3/lib/python3.7/site-packages/pandas/core/generic.py:6746: SettingWit
hCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexin
g.html#returning-a-view-versus-a-copy)
  self._update_inplace(new_data)

Wage_difference
Education_difference
Age_difference
Wage_difference
Education_difference
Age_difference
```

In [21]:
```
print("Table 7 — English effect on the outcome gap with spouse")
result_mat_R3
```

```
Table 7 — English effect on the outcome gap with spouse
```

Out[21]:

Dependent variable	2SLS Parameter (Full Sample)	SE (Full Sample)	2SLS Parameter (Female Sample)	SE (Female Sample)	2SLS Parameter (Male Sample)	SE (Male Sample)
Wage_difference	0.199**	0.083	0.094	0.062	0.213***	0.060
Age_difference	0.216	0.269	0.247	0.273	0.142	0.205
Education_difference	-0.696***	0.130	-0.359*	0.202	-1.057***	0.213

From the results we know that English level does not inflict significant effects on age and economic closeness. But it affects education closeness significantly. A simple explanation can be offered: English skills have a significant effect on communications and consequently exchanges of ideas. The cost of communication decreases if one's English level is better. As a result, immigrants will consume more intelligence closeness.

Suppose the three kinds of closeness are among the goods in deciding the utility of marriage (other goods will be, for example, the appearance of the partner). To consume education or intelligence closeness, one has to pay for the effort of communication. An increase in English level lowers the cost for communication, so the relative price of intelligence level closeness decrease. If intelligence closeness is not a Giffen good, then more of it will be consumed. However when it comes to age closeness or economic closeness, the increase of English level has no effect on the cost of these two goods. That explains why the English skill effects on the age gap and income gap are not significant. We can also observe that the magnitude of English effect is much stronger for male than for female immigrants. The differential can probably be explained by different utility curves of marriage between females and males.

7. Conclusion

The benchmark investigates the relationships between English proficiency and some social aspects of an immigrants in the US. The authors choose the age of arrival as the instrument to control for omitted variables bias and reverse causality problems. Although some doubts call for attention regarding endogeneity issues (section 6.1), nevertheless we still have theoretical reasons to retain the IV approach proposed by the benchmark. The IV results pass a lot of robustness checks (section 5.2 & 6.2). It is clearly demonstrated that English proficiency causes increases in certain outcomes, especially some marital and fertility outcomes (section 5.1 & 6.3). However, we have seen the conceptual pitfalls of the benchmark.

The outcomes might not be satisfactory measurements for the level of social assimilation, despite being social outcomes. Some outcomes do not imply the degree of integration. That significantly limits the policy implications of the paper. We agree with most theses with the benchmark but differ in conceptual issues. As the aim of this replication paper is not to explore theories in immigration and population economics, we leave the project of defining good measurements for social assimilations to future researchers.

8. References

- **Bleakley, Hoyt, and Chin, Aimee**. 2010. Age at arrival, English proficiency, and social assimilation among US immigrants. *American Economic Journal: Applied Economics*, 2(1), 165-92.
- **Duncan, Brian, and Stephen J. Trejo**. 2007. "Ethnic Identification, Intermarriage, and Unmeasured Progress by Mexican Americans." In *Mexican Immigration to the United States*, ed. George J. Borjas, 229–67. Chicago, IL: University of Chicago Press and National Bureau of Economic Research.
- **Funkhouser, Edward, and Fernando A. Ramos**. 1993. "The Choice of Migration Destination: Dominican and Cuban Immigrants to the Mainland United States and Puerto Rico." *International Migration Review*, 27(3): 537–56.
- **Lenneberg, Eric**. 1967. *Biological foundations of language*. New York (N.Y.): Wiley.
- **Morris, Charles.**. 1967. *Foundations of the theory of signs* (International encyclopedia of unified science 1/2). University of Chicago press.
- **Sorenson, Ann Marie**. 1988. "The Fertility and Language Characteristics of Mexican-American and Non-Hispanic Husbands and Wives." *Sociological Quarterly*, 29(1): 111–30.
- **Stevens, Gillian, and Gray Swicegood**. 1987. "The Linguistic Context of Ethnic Endogamy." *American Sociological Review*, 52(1): 73–82.
- **Swicegood, Gray, Frank D. Bean, Elizabeth Hervey Stephen, and Wolfgang Opitz**. 1988. "Language Usage and Fertility in Mexican-Origin Population of the United States." *Demography*, 25(1): 17–33.
- **Toussaint-Comeau, Maude, and Sherrie L. W. Rhine**. 2004. "Tenure Choice with Location Selection: The Case of Hispanic Neighborhoods in Chicago." *Contemporary Economic Policy*, 22(1): 95–110.
- **Wittgenstein, Ludwig**. 1968. *Philosophical investigations* (Repr. ed.). Oxford: Blackwell.