The Effect of Assimilation on the Earnings of South Hayward Immigrants

Andriy Gostik

California State University, Hayward

Arayik Gyulumyan

California State University, Hayward

Maisha House

California State University, Hayward

Effendy Juraimin

California State University, Hayward

Abstract

This paper examines the effect of foreign birth and length of residency in the United States on the earnings of foreign-born in South Hayward residents. Field survey of South Hayward residents was conducted in 2002 to collect the data for analysis. And OLS is the primary method for analysis. The result suggests that foreign-born of South Hayward initially has 52 percent lower annual earnings than that of the native-born. As they stay in the U.S., their earnings begin to catch-up with those of comparable native-born. The catch-up for the foreign-born will take place in 29 years. The convergence of the earning differential also occurs to South Hayward foreign-Latino and foreign-non Latino. But the convergence for the foreign-non Latino will occur at much slower rate, about 16 years longer, than that of the foreign-Latino. The slower assimilation rate of the foreign-non Latino may be due to their higher ethnic enclave effects and taste based discrimination. Confirmation to these possible explanations requires future research in these topics for foreign-non Latino in South Hayward.

I. Introduction

According to South Hayward Community Information Book (2001), 77 percent of South Hayward residents are ethnic minorities, and 41 percent of the ethnic minorities are Latino or Hispanic descent. Moreover 36 percent of the total residents are non-English speakers. About one third of the residents ages 25 and over do not have a high school degree. The report also indicates that lower income households in South Hayward have been increasing steadily from 40 percent in 1980 to over half in 2000. See Appendix B for Demographics and Social Profile of South Hayward.

The statistics seem to suggest that significant portion of the South Hayward residents are most likely foreign born, and these demographics coincide with rising low-income income households in the community. If the foreign-born are mostly low-income earners, then we would like to learn the *effect of foreign birth and length of residency in the United States on the earnings of foreign-born in South Hayward residents*. Specifically, we would like to know whether their earnings are rising as they assimilate in the country and the community, whether they are rising faster than those of native born, and whether the *economic assimilation rate* is different among specified characteristics of the foreign born.

We will discuss literatures in assimilation of foreign-born to examine theories that support and develop our hypothesis in the next section. In section III, we describe the research method that enables us to collect relevant evidence to answer the research questions. Section IV introduces empirical model, details statistical approach to test the hypothesis, and observes possible complications of the approach. Section V describes the data, variables, and sample used in the analysis. Section VI discusses findings of our empirical estimation. Section VII summarizes the results and their policy implications.

II. Theoretical Background

A pioneering work by Chiswick (1978) indicates that at the time of entry into the United States, the annual earnings of immigrant men are 15 percent lower than the earnings of native men with the same schooling, age, number of weeks worked, and other demographic characteristics. How can we explain this difference in earning? One possibility is that knowledge and skills are not perfectly mobile across countries (*International transferability of human capital*). How relevant is the training acquired in the country of origin to U.S. labor markets? For example, a Mexican immigrant lawyer may have the same number of schooling but may not be able to practice in the United States because of different legal codes between both countries. In other words the *international transferability of human capital* is imperfect. We can now identify our first testable hypothesis. (1) South Hayward immigrants lack U.S. labor market skills when they first arrive. Thus, ceteris paribus, *initially their earnings are significantly lower than that of native-born persons*.

Once the immigrants stay in the United States, they adapts to U.S. specific labor market requirements by acquiring knowledge and experience. Thus the earning gaps are narrower as they stay longer in the country. Their earnings begin to grow at faster rates than the earnings of comparable natives. Chiswick (1978) finds that foreign-born will reach earning parity in fourteen years after arrival, and earn ten percent more than comparable natives after thirty years. There are two explanations to this finding.

First, immigrants are “more able and more highly motivated “ than natives. This theory implies that non-random selection of immigrants occur that lead to productive people migrating out of their countries into the United States (*Self-selection in migration*). In the words of Borjas (1990), “Presumably, this sorting occurs because only persons with exceptional ability, drive, and motivation would pack up everything they own, leave family and friends behind, and move to foreign country to start life anew."

Second, immigrants are more likely to self-finance their human capital investment. They are regularly "job hopping" as a way to gain information in U.S. job market[1] (Chiswick, 1978b), which discourages employers to finance job-specific investment. Becker (1964) discovers that for the same total investment in training, the smaller proportion of firm-specific training financed by the employer, experience-earning profiles are steeper. Therefore the immigrants with higher proportion of self-financed training have *steeper experience-earning profiles*. Our second hypothesis then is as follow. (2) As the South Hayward immigrants accumulate U.S. specific human capital, they gain these skills and labor market assimilation takes place, meaning that *their earnings begin to catch up to the earnings of comparable natives overtime since years of migration.*

The Chiswick's (1978) assimilation theory basically draws inferences of earnings adjustment over time from a single snapshot of the immigrant population at that time. But the theory does not consider the fact that the newly arrived immigrants may be different from those when Chiswick conducted his study. Specifically, Borjas (1985) argues that separate immigrant waves in different periods may have different productivity level ("*cohort effect*"). Thus what appears to be a catch-up in a cross-sectional data of immigrants may not occur in separate immigrant cohorts. This implies that we need to extend our hypothesis for the immigrants to different cohorts. The problem is what criteria should we use to identify the cohort.

Recent literature by Borjas (2000) suggests that the existence of a large ethnic enclave lowers the return on investment in the country specific human capital and lowers the assimilation rate (*ethnic enclave effect*). For example, Latino immigrants can work in community where employers speak their language, thus have less incentive to assimilate in the United States. Since South Hayward has large Latino community, we select Latino and non-Latino as our cohorts, and subjects for our last hypothesis. (3) Among the immigrants of the South Hayward, *Latino or Hispanics have lower assimilation rates than those of non-Latino or Hispanics*, ceteris paribus.

III. Method

A. Survey Method

A survey method was used to extract data for analysis. The South Hayward Community was divided into eight surveying tracts that were consistent with tracts used by U.S. Census data collectors in the year 2000 (see Appendix A Census Tract Map). The tract boundaries were based on major streets or railways as well as a large enough sample of residents in the section. All residents were randomly selected and surveys were only administered to randomly selected residences assigned to field interviewers. Stratified random sampling was used, which requires dividing the population into sub groups and using simple random sampling on each strata or sub group (Cassuto, 2002).

Surveys were conducted between the dates of April 20, 2002 to May 20, 2002. Each interviewer was required to compete survey training before he or she was allowed to survey residents in the field. Two hundred forty-six completed surveys were collected. The interviewers conducted surveys from morning to evening however approximately 61% of completed surveys were collected between the times of 4:00PM and 10:00PM.

B. Survey Design

The survey instrument was designed by the HIRE Center to address the research question of skills that allow workers to obtain, retain, and advance from entry-level jobs in the Bay Area labor market. The research question flows from the research design, which creates a research

[1] Youths just entering labor market, similar to recent immigrants, also tend to have high quit rates.

structure prior to data collection and analysis. The survey instrument was created with the idea of obtaining evidence that enables the researcher to answer the research question as unambiguously as possible (de Vaus, 2001).

C. Research Design

The research design team that created the instrument was specifically concerned with what skills employers hiring for entry level positions that require high school diplomas want and what skills people with high school diplomas have in one labor market. This resulted in the creation of two surveys, one for employers and one for residents. The team also considered the population to be surveyed and South Hayward seemed to have a population most representative of the types of respondents needed.

The design is considered non-experimental in that it is not subject to the control of the researcher. Gujarati explains that this type of data can create special problems for the researcher in that it causes difficulty in finding cause and effect relationships (Gujarati, 1995). For example our analysis looks at the rate foreign born persons assimilate into the United States Labor Market. We may find several relationships between the variables we have chosen that cause assimilation rates for some foreign born people to be faster than that of other foreign born people. If this happens we will take special care in analyzing cause and effect relationships as it relates to the groups involved due to the fact our data sets and survey instruments are non-experimental in nature. Additionally, we may consider adding more background factor variables from the data set we use to help accurately identify these types of relationships.

As mentioned our research question looks at the *effect of foreign birth and length of residency in the United States on the earnings of foreign-born in South Hayward residents*. Specifically, whether their earnings are rising as they assimilate in the country and the community, whether they are rising faster than those of native born, and whether the *economic assimilation rate* is different among specified characteristics of the foreign born.

The research question for our analysis coincides with the research design and research question used above, thus the survey instrument used will also allow us to extract evidence that addresses our research question unambiguously, even though the survey instrument was not designed specifically for our research. We believe our research design is valid and provides a structure to accurately test our research question.

D. The Strengths and Weaknesses of Cross-sectional Research Design

According to Nan Maxwell, Executive Director at the HIRE Center at California State University Hayward, research design can be the single most important stage of research and analysis. Two criteria are used to evaluate research design; internal validity and external validity. Maxwell also admits you cannot have external validity without internal validity (Maxwell, 2002).

Cross-sectional research design is used for the purposes of this paper. Cross-sectional designs are most widely used for social research because they enable the researcher to obtain results quickly, as well as can be ideal for descriptive analysis. Additionally, cross-sectional designs can be considered more cost effective than comparable experimental and longitudinal designs (de Vaus, 2001). We now explore our research design in terms of internal and external validity as it relates to cross-sectional research design.

In terms of internal validity, or design that does not enable the author to choose unambiguously one explanation of results over another, cross-sectional research designs present a number of problems that the authors of this paper acknowledge. Primarily, cross-sectional designs are problematic in establishing cause and affect relationships between variables because they lack the time element. This in turn affects the level of meaning, or the inferences that can be made about the relationships between dependent and independent variables as it relates to the survey population.

The authors of this paper attempt to ameliorate this problem by using the appropriate statistical controls. Following de Vaus (2001), we are making our comparison groups as similar as possible by statistically removing differences between groups after data have been collected. In the case of our analysis, we will attempt to create controls that ensure our US-born group is similar to our non-US born group and our Latino and Non-Latino non-US born groups are both similar to each other, as well as the US-born group, after data collection before we apply our empirical analysis.

Additional internal validity is minimized by the survey method used to collect data. Problems that cause biases like Testing or the effect of taking a test on a respondent, and Instrumentation or the ways questions are phrased by Interviewers were minimized by the excellent survey instrument used and requiring Interviewers to complete training before entering the field. Problems such as Selection Bias at Entrance or basing analysis on part of the survey population (Selectivity Bias) were minimized by using stratified random sampling.

In the case of external validity, or the extent to which results from a study can be generalized to beyond a particular study, cross-sectional designs are considered to be strong in accurately describing the survey population (deVaus, 2001). Probability sampling methods were employed during the data collection process. Additionally, the authors have addressed the opportunities for sampling bias and have reviewed the appropriate corrections to the problem should it arise (see Section V Data and Sample).

In addition, problems that cause biases like Interactive effect of testing or changes in responses to questions based on the way they are phrased by the Interviewer and Selection Interactions or systematic exclusion of parts of the survey population are minimized by the survey instrument design and methods.

IV. Estimation

A. Basic Model

To obtain a preliminary insight into the nature of the issue, we compute *descriptive statistics* of annual earnings, gender, high school graduate, college graduate, labor market experience, hours work per week, marital status, and years since migration for all sample, native-born, foreign-born, foreign-Latino, and foreign-non Latino. To test the hypotheses we employ *multiple linear regression* method. The regression is evaluated by the means of *Ordinary Least Squares (OLS)* unless violations of some of OLS assumptions are detected. We specify the equation using the human capital earnings function. In this respect, current research draws heavily on seminal works of Chiswick (1978) and Mincer (1974).

But the specification of the model to be discussed in this paper is slightly different from those proposed before to account for the peculiarities of the study. The key idea is that individual's annual earnings are determined by years of schooling and years of labor market experience. Following by the hypotheses, a dichotomous variable signifying if the individual is foreign-born or not has to enter the model explicitly. Pooled data comprising information on both foreign-born and U.S.-born individuals is to be used for the analysis. The final regression equation to test the hypotheses is specified in the following form:

$$\ln E_i = \ln E_0 + c_0\, FOR_i + c_1 HSG_i + c_2 COLG_i + c_3 T_i + c_4 T_i^2 + c_5 (FOR_i)(YSM_i) + c_6 FEM_i + u_i \quad (1)$$

where

- ln E is natural logarithm of annual earnings of individual i;
- ln E_0 is the intercept, a logarithm of hypothetical annual earnings of a person with no education and work experience;
- FOR is a dummy variable equal to unity for foreign-born and to zero otherwise.
- HSG is a dummy variable equal to unity if respondents are graduated from high school and equal to zero otherwise;
- COLG is a dummy variable with value of one for those who are graduated from college, and with value of zero otherwise.
- T is years of labor market experience proxied as age – years spent for education - 6;
- YSM stands for years since migration, applicable only to foreign-born;
- FEM is a dummy variable equal to unity for female and to zero otherwise.

Because education and job market experience are factors that are believed to contribute to higher earnings, all slope coefficients in the regression equation (1) are expected to be positive, except for c_4 and c_6, as explained below. Coefficient c_0, which is a part of the intercept in the equation for the foreign-born is anticipated to be negative, because we expect that initial earnings of foreign-born residents upon their arrival in the U.S. are lower than initial earnings of natives, ceteris paribus. We select female gender as our control variable by looking at its descriptive statistics and explanatory contribution to the regression. The coefficient to this variable, c_6, is expected to be negative as a body of earlier research suggests. Female gender has the best explanatory power to the regression in comparison to other control variables, such as marital status and hours work per week. Other than that, we resist the temptation to add more control variables into the model, such as those characterizing various skills the surveyed residents possess, keeping in mind precepts of the *Occam's razor*.

B. Derivation of the Model

There are several points regarding this specification that need some explanation. Given that the model is designed not for time series data, dependent variable in the form of natural logarithm needs to be explained. The theory behind this was elaborated by Mincer (1974). In the simplest form, equation (1) represents so-called *schooling model*, where earnings is a function of only years of schooling. The assumption is that schooling results in a postponement of earnings and, therefore, in reduction of a present value of earnings flow:

$$PV_s = E_s \sum_{t=s+1}^{n} \left(\frac{1}{1+r} \right)^t , \quad (2)$$

where

- s – years of schooling;
- E_s – annual yearnings of an individual with s years of schooling;
- n – length of working life plus years of schooling;
- r – discount rate.

To make the formula more convenient, continuous discounting process is further assumed yielding

$$PV_s = E_s \int_s^{m+s} e^{-rt} dt = \frac{E_s (e^{-rs} - e^{-r(m+s)})}{r} = \frac{E_s e^{-rs}(1 - e^{-rm})}{r} , \quad (3a)$$

where

- e is a base of natural logarithms;
- m = n-s is a fixed span of earning life.

Similarly, for an individual with zero years of schooling (s = 0) the expression becomes

$$\text{PV}_0 = \text{E}_0 \int_0^m e^{-rt} dt = \frac{E_0(1-e^{-rm})}{r} \tag{3b}$$

Since PVs = PV_0 in equilibrium, equating expressions (3a) and (3b) the following result is obtained:

$$\frac{E_s}{E_0} = \frac{1}{e^{-rs}} = e^{rs}\,. \tag{4}$$

Taking logarithms of (4) yields ln E_s = ln E_0 + rs, which is exactly the schooling model proposed by Mincer.

In our study, however, we tried to take a bit more realistic approach by measuring the impact of education on earnings through two dummy variables, HSG and COLG. Thereby we aim to address the issue of discontinuity of schooling variable, because such important thresholds as graduating from high school and college should be associated with earnings as compared to the earnings of residents who spent a comparable number of years for schooling, but didn't graduate or receive a degree. We focus our attention only on these two schooling variables because we consider them to have the most effect on earnings and also because the data on them are sufficient to do a statistical analysis. We further expand this simple equation by adding labor market experience as another key variable. Since surveys didn't contain a direct question about the years of work experience, we use a proxy to measure it. Given the hypotheses we set forward to test, we decided that the most appropriate proxy would be defined as age minus years a person spent for schooling minus six (childhood before school). Years spent for education are obtained in a manner described in the Data section of this research. The most serious drawback of such definition of years of labor market experience is that years when a person was unemployed are ignored. But as our data shows, most working people in the sample we use indicated that they had a long-term work experience, which makes our approach more justified. Another possible proxy defined as years a person has been working for his or her current employer does not serve our purpose to measure a catch-up between earnings of the foreign-born and U.S.-born residents.

Since the relationship between earnings and experience is deemed to be *nonlinear* (earnings rise at a declining rate with an increase in of years of experience), a quadratic term is included in the regression equation with coefficient b_4 expected to be negative (Mincer, 1974,p. 84). This expands the core equation to be:

$$\ln E_i = \ln E_0 + b_1 HSG_i + b_2\, COLG_i + b_3 T_i + b_4\, T_i^2 + u_i \tag{5}$$

However, to single out the factor of the labor market experience in the U.S. for those South Hayward residents who were born outside the U.S., variable of work experience (T) is decomposed into two new variables – YSM (years since migration), which is a proxy of years of labor market experience after migration[2], and YPM for years of work experience prior to migration. Therefore, the basic equation for foreign-born individuals becomes:

$$\ln E_i = \ln E_0' + b_1'\, HSG_i + b_2 COLG_i + b_3' YPM_i + b_4'\, YPM_i^2 + b_5' YSM_i + b_6'\, YSM_i^2 + u_i \tag{6}$$

Following Chiswick (1978), we make a substitution YPM = T –YSM in the equation (6) that results in below equation:

$$\ln E_i = \ln E_0' + b_1'\, HSG_i + b_2' COLG_i + (b_5' - b_3') YSM_i + (b_4' + b_6')\, YSM_i^2 + b_3' T_i + b_4'\, T_i^2 - 2b_4 YSM_i T_i + u_i. \tag{7}$$

Like Chiswick (1978), we omit the insignificant term YSM*T to simplify (7). Also, in our regression YSM-squared turned to be insignificant. But the results improved dramatically when we excluded it. Therefore, our final equation does not include a squared term for YSM.

In order to combine equations (5) and (7) we need additional assumptions about regression coefficients. Namely, we assume that $b_1 = b_1'$, $b_2 = b_2'$, $b_3 = b_3'$ and $b_4 = b_4'$. Thereby, variables T from equation (5) and from equation (7) will be combined. We also introduce a dummy variable FOR to finally obtain equation (1). FOR enters equation (1) for two purposes. First, interacting with other variables it distinguishes variables pertaining only to foreign-born persons. Second, it also appears as an independent term to control for a perceived difference in intercepts for equations (5) and (7).

C. Hypotheses Testing

To test the hypotheses suggested at the beginning of the research, the regression equation can be used the following way. First, it is necessary to determine if the goodness of fit of the overall model is sufficient to make any conclusions about the relationship between variables. F-test is a proper indicator for this. If p-value associated with F-statistic is less then 0.05 (required significance level), then the model can be used for further investigation.

To test the hypothesis 1 coefficients $\ln E_0$ and c_0 have to be significant. That is, p-values associated with these coefficients have to be less than 0.05. The reasoning is as follows. Setting YSM = 0 reduces equation (1) for foreign-born residents (FOR = 1) to

$$\ln E_i = \ln E_0 + c_0 + c_1 HSG_i + c_2 COLG_i + c_3 T_i + c_4 T_i^{2\,+} c_6 FEM_i \tag{8}$$

Hypothesis 1 is supported if c_0 is negative, because this mean that all other things being equal people from the community who just arrived at the U.S. from abroad would have lower initial earnings.

To test hypothesis 2, equation (1) is differentiated with respect to the dummy variable FOR. The resulting expression is nothing but a percent difference in earnings between the native and foreign-born residents:

$$\frac{\partial \ln E}{\partial \ln FB} = c_0 + c_5(YSM_i) \tag{9}$$

[2] We assume that since migration individuals only worked and didn't spend time on formal schooling.

Significance of the coefficients is required for legitimize conclusions. Expression (9) is expected to be negative for YSM = 0, as explained before, and then it is to rise with increasing values of YSM, as determined by the coefficients. Performing simple simulations by plugging in expression (9) increasing magnitudes of YSM gives an idea of how fast the gap between earnings of natives and the foreign-born is shrinking (if it is at all). Such simulations enable to test if catch-up exists.

Testing of the hypothesis 3 is analogous. Instead of using the data for natives and all foreign-born residents, first the regression is run for the subset of natives and Hispanic/Latino foreign-born residents and then the operation is repeated for the pooled data of natives and non-Hispanic foreign-born. Provided both regressions are significant as judged by F-statistics, hypothesis 3 is supported if expression (9) for the pooled regression of natives and non-Hispanic residents increases faster than expression (9) for the pooled regression of natives and Hispanic/Latinos does for increasing values of YSM, because this would mean that earning of non-Hispanic foreign-born residents approach those of the native faster as the years since migration increases.

D. Possible Complications with the Model

The proposed model is based on the assumption that since migration, foreign-born South Hayward residents have only been working. The idea follows Chiswick's work in 1978, where years of labor market experience were replaced by years since migration. He argues that such a substitution doesn't affect conclusions in any substantive way and so do we. However, should there be many enough cases in the underlying data when *foreign-born individuals spent a lot of time studying or not working after their immigration in the U.S.*, the conclusions drawn from the model would be biased. Nevertheless, given the limitations of data availability from the surveys, this research will proceed with the initial model. Overall, major complications are to be expected from insufficient or unreliable data. There is not much that can be done in this situation, except for collecting new and better data and trying the model again.

Other possible complications have to do with statistical characteristics of the data. Since the data used is cross-sectional, the problem of *heteroscedasticity* (non-constant variance of error terms) might arise. This would result in unreliable estimation of regression coefficients because confidence intervals would appear narrower than in reality. The problem can be detected by examining plots of residuals against dependent variable. Solving the problem requires the use of Weighted Least Squares method. On one hand the dependent variable in equation (1) is in logarithmic form, which lessen likelihood of heteroscedasticity. On the other hand, the nature of the problem, where the variance of earnings may increase as education and experience rise, inherits the property of heteroscedasticity.

Another possible complication is *misspecification* of regression equation. If the relationship between log of earnings and its factors has a different form from equation (1), the conclusions may be misleading. That is why it is important to base a regression equation on an appropriate theory. To derive some insight about the specification of the model, dependent variable should be plotted against independent ones and the pattern of the plot investigated. In particular, non-linear (quadratic) pattern of logarithm of earnings with respect to years of work experience acquired should be verified.

V. Data and Sample

According to Gujarati (1995), data should be analyzed on the basis of three categories. One should analyze data on the basis of data type, data source, and data reliability. In this section we analyze the data used for the study on the basis of the categories mentioned above.

A. Data Type

Three types of data are usually available for empirical analysis and they include time, series, cross-sectional, and pooled. Time series data is collected at regular time intervals, cross-sectional data is collected at the same point in time, and pooled data contains elements of both time series and cross sectional data. The data used for this study is cross-sectional data. Cross-sectional data has its own peculiarity in heterogeneity (Gujarati 1995). When applying statistical analysis to heterogeneous units the size or scale effect takes place. We acknowledge this problem with cross-sectional data and have accounted for it in our model.

B. Data Source

The data source used for this analysis is secondary and involves the use of door-to-door surveyors administering a survey instrument created by the HIRE Center at California State University Hayward.

The survey instrument was designed to address the question of skills that allow workers to obtain, retain, and advance from entry-level jobs in the Bay Area labor market. The research design team that created the instrument was specifically concerned with what skills employers hiring for entry level positions that require high school diplomas want and what skills people with high school diplomas have in one labor market. This resulted in the creation of two surveys, one for employers and one for residents. The team also considered the population to be surveyed and South Hayward seemed to have a population most representative of the types of respondents needed (see Appendix B for South Hayward Demographic Statistics). South Hayward is described as the area bounded by Harder Street, Mission Street, Industrial Street, and Interstate 880. Appendix B reveals that 77 percent of South Hayward residents are ethnic minorities, and 41 percent of the ethnic minorities are of Latino or Hispanic descent. Thirty six percent of the total residents are non-English speakers and the percentage of lower income households in South Hayward is increasing steadily, from 40 percent in 1980 to over half in 2000. These statistics reveal the strength of our data source as it relates to our research topic as mentioned in the introduction and mentioned briefly above. We can assume from the statistics that many South Hayward residents are born outside of the United States and this gives us a good opportunity to test assimilation on a population most representative of the types of respondents needed for our analysis.

The data used for our analysis comes from a Resident Survey, which consists of a series of questions that focus on identifying specific skills, training, and experience associated with employment. Additionally, the survey includes questions on background factors that might contribute to a person's ability to obtain and retain employment (HIRE Center, 2002). Also see Appendix C for Sample Survey.

Table 1: Cross-References Our Variables with the Survey Questions

Variable	Variable Description	Survey Question(s)
Ln E	Natural Log of Annual Earnings	Q15L: How much do you earn per hour x 1920
HSG	High School Graduate (Dummy Variable Yes =1, 0 otherwise)	Q42: Highest level of education completed
COLG	Bachelor's Degree Holder (Dummy Variable Yes =1, 0 otherwise)	Q42: Highest level of education completed
T	Labor market experience, proxied as age – years in school – 6	Q38: Age
		Q42: Highest level of education completed
YSM	Years since migrating to the U.S.	Q43A: Years have been living in the U.S.
FOR	Dummy variable. Unity for foreign birth, zero for native-born	Q43: Country were you born
FEM	Female (Dummy Variable Yes =1, 0 otherwise)	QS3: Respondent's Gender
Latino/Hispanics, and non-Latino/Hispanics	Latino/Hispanics and non-Latino/Hispanics	Q39: Are you Latino or Hispanic Descent?

The survey instrument has given us the opportunity to pull a variety of data for our analysis. We do not use all data extracted from available questions on the survey instrument for analysis. This is because we are only interested in the questions that allow us to analyze our topic the effect of foreign birth and length of residency in the United States on the earnings of foreign-born South Hayward residents (Table 1). For this reason we focus our analysis on the response to the several questions available on the Resident Survey for which we have constructed the data from each response into independent and dependent variables for the purposes of our analysis. You were introduced to the independent and dependent variables and why we chose them in the Estimation Section.

D. Data Reliability

Data reliability refers to the accuracy and quality of the data used for statistical analysis. This relates to the methods used for data collection. In terms of the data used for this study there are several opportunities for the collection of non-reliable, poor quality data. As mentioned a survey method is used, which can be biased in nature, partial response to the survey questionnaire can cause selectivity bias or analysis based on a population not representative of the population surveyed, and the data is collected using door-to-door surveyors who can possibly record errors by commission or omission.

Another way selectivity bias can occur is during the survey process and methods. If all interviewers systematically survey residents between the times of 6:00PM and 10:00PM Monday through Friday, our data set will be biased toward residents who are home and willing to take the survey during the time slot mentioned. Our data set may be systematically overlooking foreign-born South Hayward residents, who are either not home at the time or unable to participate in a survey during the time slot. As it relates to our data collection, the interviewers administered the bulk of the completed between 4:00PM and 10:00PM. Sixty One percent of the completed surveys were completed during this time.

It is important to mention the opportunity for selectivity bias in the data collected because the survey and data collection methods were not specifically designed for our analysis.

James Heckman (1979) discusses sample selection bias as the bias that results from using non-randomly selected samples to estimate behavioral relationships as an ordinary specification error or omitted variables bias.

> "Sample selection bias may arise in practice for two reasons. First, there may be self-selection by the individuals or data units being investigated. Second, sample selection decisions by analysts or data processors operate in much the same fashion as self selection (Heckman, 1979)."

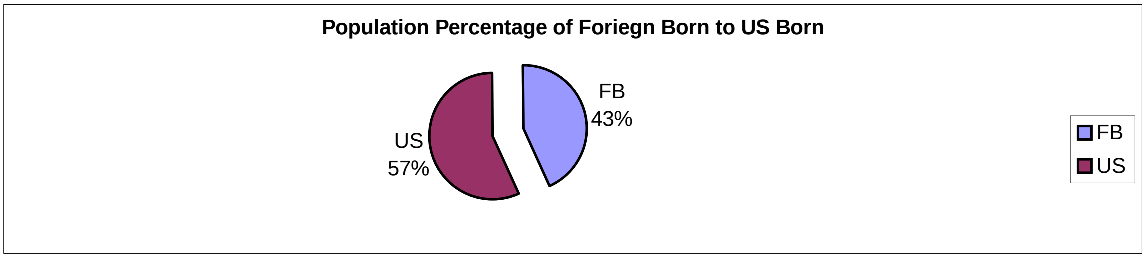
We consider the data and sample are randomly selected, but acknowledge the case where data units self select as suggested by Heckman. For example, our sample data set was reduced by 49% due to the amount of respondents that did not respond to one of our key variable questions.

This study takes into account data type, data source, and data reliability. We have tried to create a model that minimizes some of the negative affects of the data used for analysis. The authors of this paper contend that the data set used for analysis is quite significant and most of the weaknesses found are weaknesses that are found in many other data sets.

E. Population and Sample Statistics

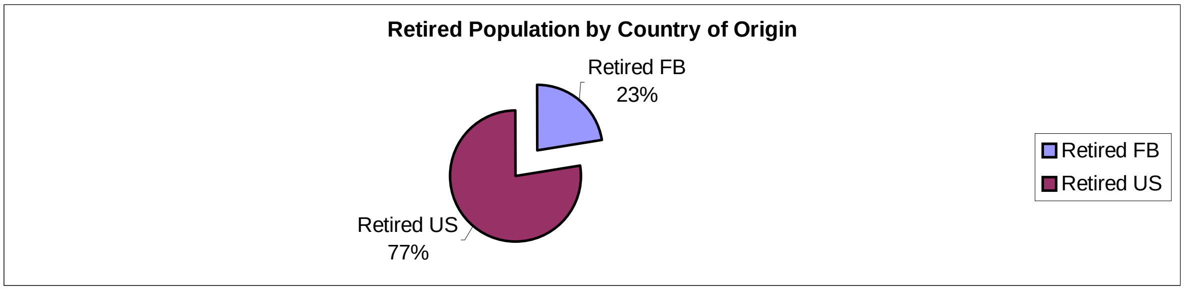
As previously mentioned, the population size is bigger than the data used for analysis purposes. We define population as the 246 completed surveys. There were two hundred and forty six completed surveys for which 57% of the population respondents were born in the United States US Born and 43% were born in another country, Foreign Born as revealed by Figure 1.

Figure 1: Population Percentage of Foreign-Born to US-Born



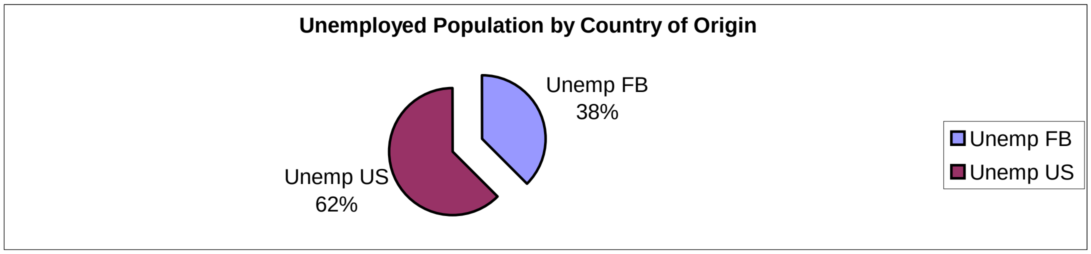
Approximately 16% of the population is retired. About 23% of the respondents were Foreign-Born and 77% were native-born. Observe Figure 2.

Figure 2: Retired Population by Country of Origin



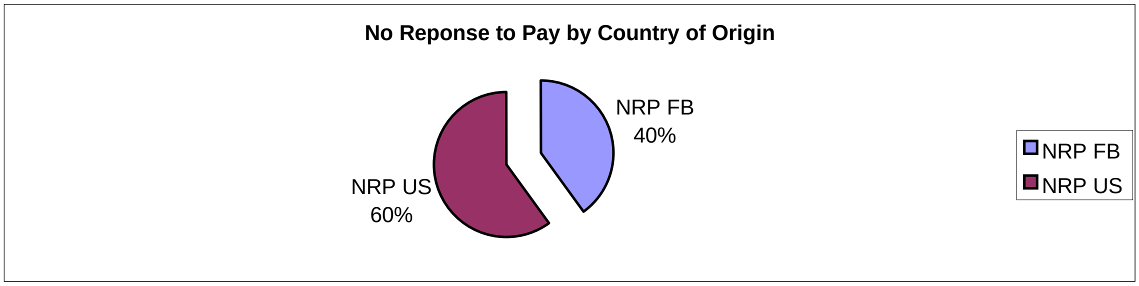
Nearly 23% of the population respondents were unemployed excluding retired respondents and the breakdown for Foreign Born/US Born is revealed in Figure 3.

Figure 3: Unemployed Population by Country of Origin



Finally 49% of the respondents including retired respondents did not respond to Q15L: “How much do you ear per hour multiplied by fifteen weeks?” as displayed below in Figure 4.

Figure 4: No Response to Pay by Country of Origin



As mentioned 50% of our survey population was reduced and this is how our sample differs from the survey population. Our sample size is comprised of 123 observations. The omission of 123 observations from our analysis might cause some to question whether our sample is random. Again, we acknowledge that there will be some omission of observations because the research design and instrument were not made specifically for our research. We feel the data set used is representative of a population needed to unambiguously test our hypothesis for several reasons.

Table 2 displays the percentage of Foreign Born and US Born respondents based on 123 observations. Notice that the US Born respondent percentage is 55%, which is approximately 2% less than the population percentage of 57%. The same difference occurs for the Foreign Born respondent percentage, which is 45% for the sample as compared to 43% in the population. For the most part the sample data represent the population data.

Table 2: Percentage of Foreign-Born and U.S.-Born Respondents

Population Total			**Sample Total**		
USA	140	57%	USA	67	55%
Foreign	106	43%	Foreign	56	45%
Total	*246*		*Total*	*123*	

A glaring weakness of our data is the reduction in observations used for analysis because of the lack of response to Q15L. We acknowledge this weakness in our data set and suggest revising this question on the survey instrument or re-training interviewers to accurate extract data for this particular question in order to increase the response rate to Q15L.

VI. Empirical Results

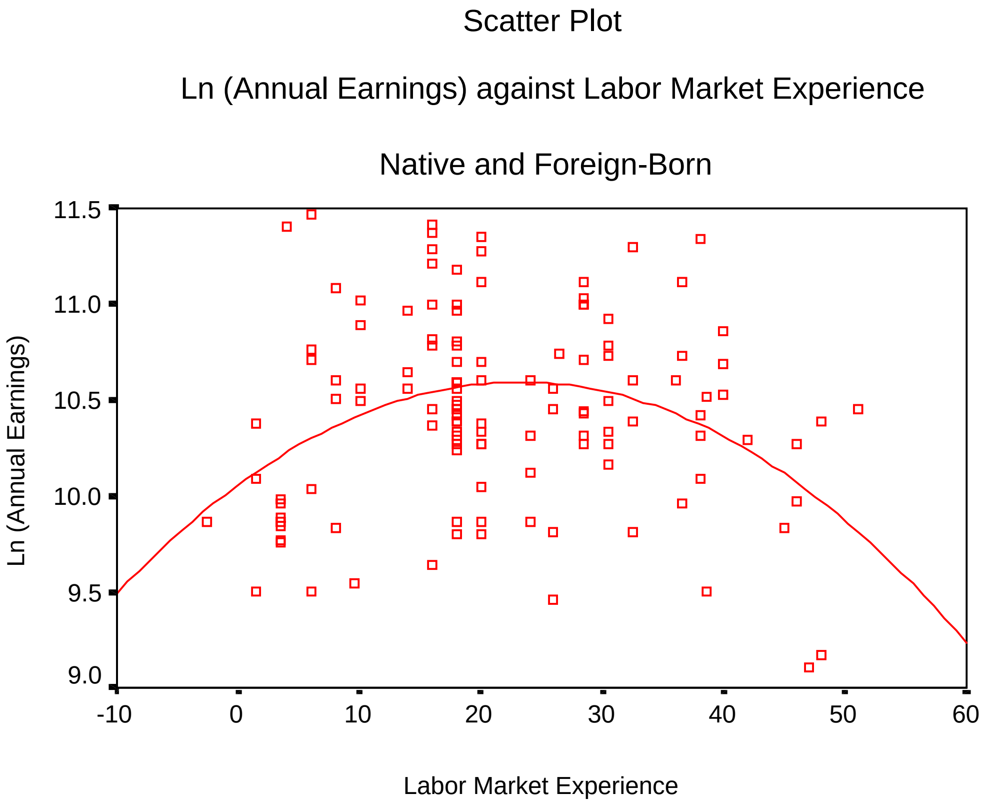
The simple observation of descriptive statistics for our data sample may supply initial information of the validity of our hypotheses. Table 3 shows descriptive statistics for overall sample, native-born, foreign-born, foreign-Latino, and foreign-non Latino. Mean of annual earnings for the all sample is $38,938 whereas a native-born earns $41,623 on average. Both numbers clearly outperform all foreign-born, including and outperform an immigrant by about at least $3,000. Earning of foreign-Latino is slightly higher than that of foreign-non Latino. In comparison to foreign-born, native-born has higher percentage of female and high school graduate, higher labor market experience, higher hours work per-week, and higher percentage of not married - spouse present. Also note that only one third of Foreign-Latino are graduated from high school whereas Native-born and Foreign-non Latino have nearly 90 percent.

Table 3: Descriptive Statistics

	All			Native Born			Foreign Born			Foreign Latino			Foreign Non-Latino		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Annual Earnings ($)	126	38,938.03	19,319.25	68	41,623.58	19,383.46	58	35,789.46	18,925.36	19	36,477.98	22,815.27	39	35,454.03	17,039.11
Ln (Annual Earnings)	126	10.45	0.50	68	10.53	0.46	58	10.35	0.54	19	10.33	0.61	39	10.36	0.50
Female (%)	244	48%	50%	139	52%	50%	104	44%	50%	41	54%	50%	62	37%	49%
High School Graduate (%)	245	79%	41%	140	88%	33%	104	67%	47%	41	37%	49%	62	87%	34%
College Graduate (%)	245	21%	41%	140	15%	36%	104	28%	45%	41	15%	36%	62	37%	49%
Labor Market Experience	244	24.31	14.32	139	25.12	15.01	104	23.30	13.39	41	23.74	13.04	62	23.22	13.71
Hours Work per Week	147	40.44	8.89	82	40.98	8.67	64	40.24	8.34	21	39.69	12.35	43	40.51	5.61
Not Married, spouse present (%)	245	36%	48%	140	44%	50%	104	26%	44%	41	24%	43%	62	27%	45%
Foreign born (%)	245	43%	50%	140	0%	0%	105	100%	0%	41	100%	0%	63	100%	0%
Latino (%)	244	32%	47%	139	27%	45%	104	39%	49%	41	100%	0%	63	0%	0%
Years Since Migration	104	16.61	10.18	1	1.00	.	103	16.76	10.11	40	16.29	9.28	62	16.95	10.72

To validate the quadratic relationship between logarithm of annual earnings and labor market experience, we plot a scatter graph in Figure 5. Introducing fitted-line into the plots, quadratic form seems to be the most appropriate with the highest R^2. Therefore our assumption of quadratic labor market experience in equation (7) is acceptable for our purpose.

Figure 5: Ln (Annual Earnings) Against Labor Experience for Native and Foreign-Born



To test our first hypothesis we perform regression analysis using OLS method. First we run a regression following the equation (1), where we observe how the education level and labor market experience affect logarithm of annual earnings, for a pooled native and foreign-born. Then we test the regression result for heteroscedasticity by plotting the residuals against the predicted values of logarithm of earnings (Figure 5). The scatter plot indicates no systematic data pattern, thus no presence of heteroscedasticity.

Figure 6: Est. Residuals against Ln (Annual Earnings) for Native and Foreign-Born

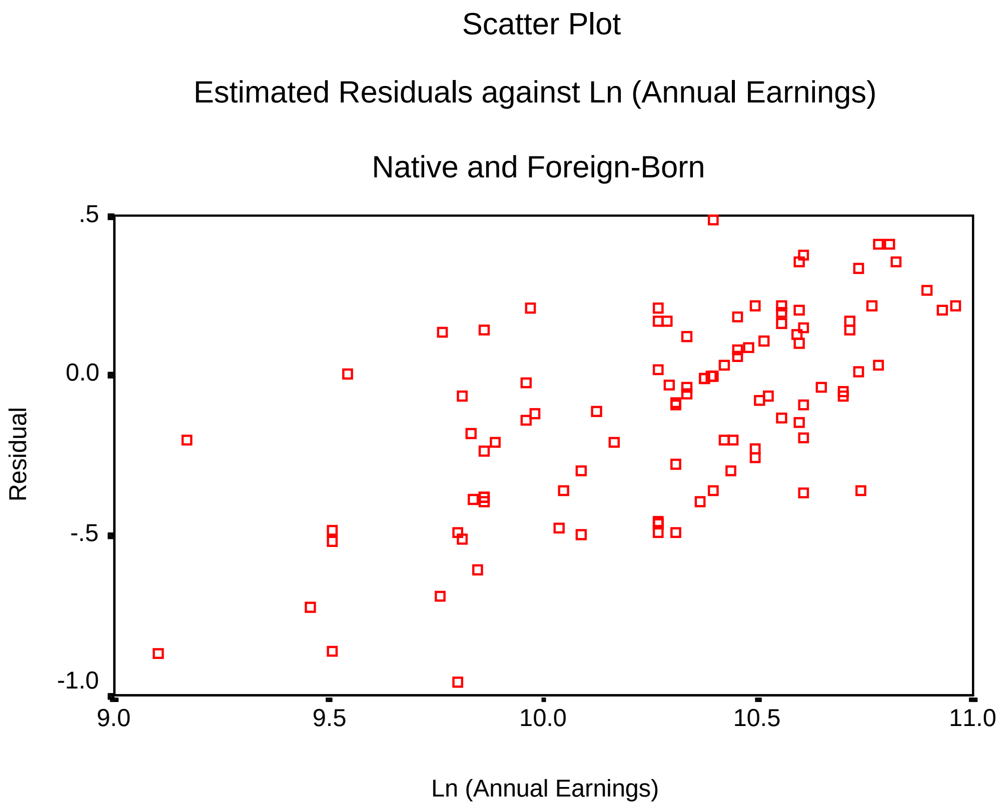


Table 4 shows results summary for all regressions. Details of regression outputs are in Appendix D, E, and F. Examining regression (1) for the first hypothesis, the adjusted R square is 0.39 with F-ratio of 12.213. The high F-ratio indicates that the regression is significant for further testing. All coefficients in the regression (1) are significant except for HSG (0.071 significant level). Signs of the coefficients follow our prior assumptions as well. FOR coefficient is particularly our sole interest in testing the first hypothesis. It means that *initial earnings of the foreign-born upon their arrival in the United States are about 52.3 percent lower than that of the native*. Thus the result validates our first hypothesis.

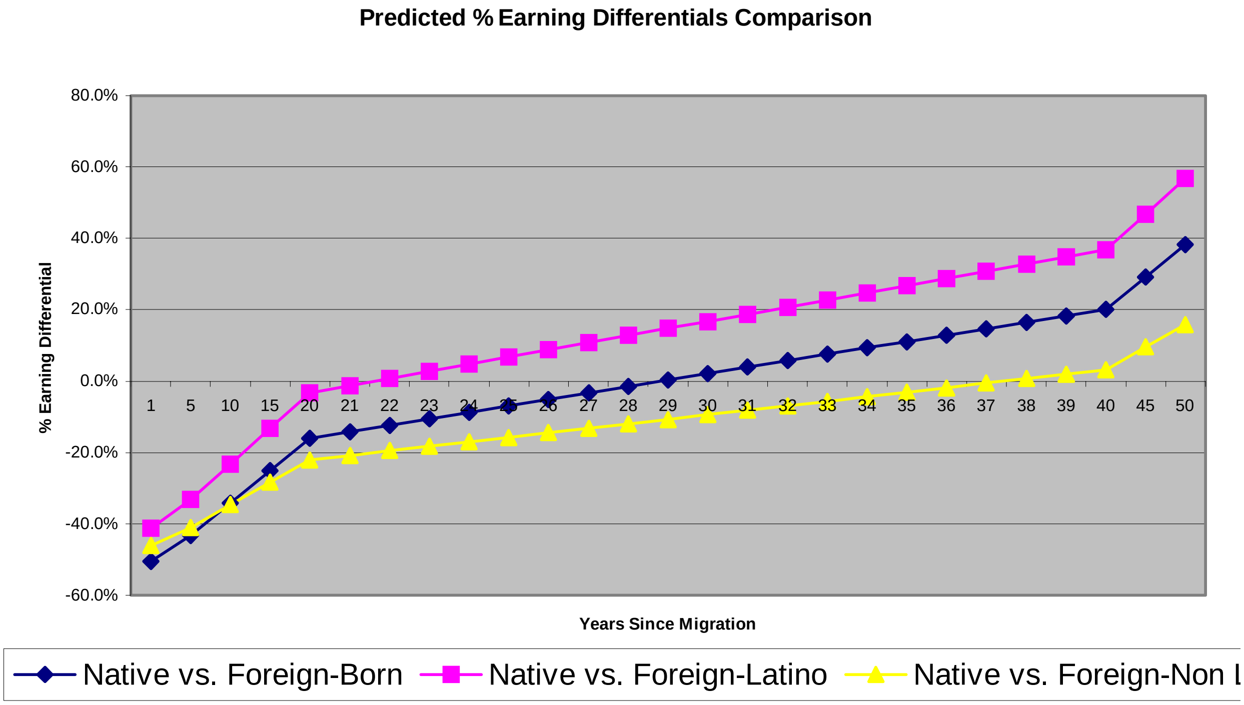
Table 4: Regression Analysis of Earnings for Native and Foreign-Born (Latino and Non-Latino)

	Native and Foreign-Born	**Native and Foreign-Born Latino**	**Native and Foreign-Born Non-Latino**
	(1)	(2)	(3)
CONSTANT	10.10622	9.97630	10.11411
	(0.000)	(0.000)	(0.000)
FOR	-0.52272	-0.43170	-0.47305
	(0.000)	(0.027)	(0.001)
HSG	0.22120	0.40551	0.17618
	(0.071)	(0.009)	(0.294)
COLG	0.33972	0.26051	0.38244
	(0.000)	(0.016)	(0.000)
T	0.03817	0.03160	0.03760
	(0.000)	(0.007)	(0.001)
T2	-0.00083	-0.00057	-0.00080
	(0.000)	(0.021)	(0.001)
FOR_YSM	0.01811	0.01998	0.01263
	(0.003)	(0.034)	(0.096)
FEM	-0.35383	-0.43796	-0.31299
	(0.000)	(0.000)	(0.000)
Observations (N)	123	85	104
Adjusted R Square	0.39	0.442	0.343
F-Ratio	12.213	10.618	8.741

Note: parenthesis indicates significant level

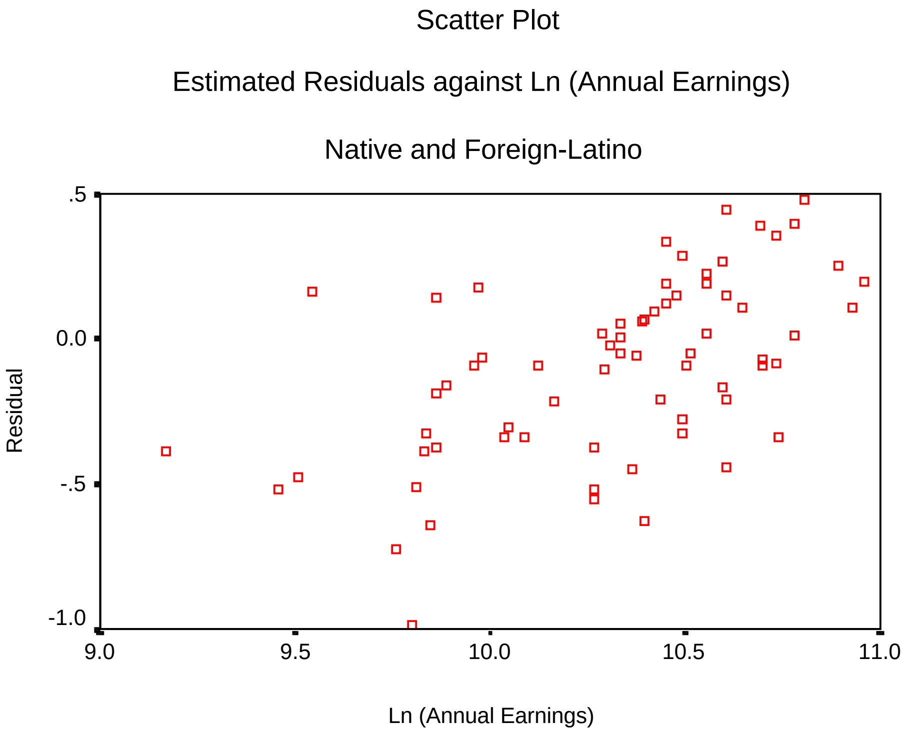
We are now ready to test the second hypothesis. We simulate the progression of earning differential over the years since migration using equation (9). Running the simulation with years since migration up to 50 years, we have a list of the predicted percentage of earning differential in Appendix G. The negative earning differential between native and foreign-born in column (1) decreases as the years since migration increases. Therefore the list shows that *catch-up does occur for foreign-born but it will take almost 29 years* (Figure 7). Thus it legitimizes our second hypothesis.

Figure 7: Equalization of income for all immigrants, Latino, and non-Latino immigrants



To test the third hypothesis, we run two additional regressions based on the equation (1). The first regression is for the pool data of native and foreign-born Latino. The second one is for native and foreign-born non-Latinos. Similar to regression (1), we test for heteroscedasticity by plotting residuals against logarithm of annual earnings for regression (2) and (3). Figure 8 shows that the scatter plots have no systematic pattern, thus no indication of heteroscedasticity for regression (2). We have same pattern in Figure 9 as well for regression (3) that indicates no presence of heteroscedasticity.

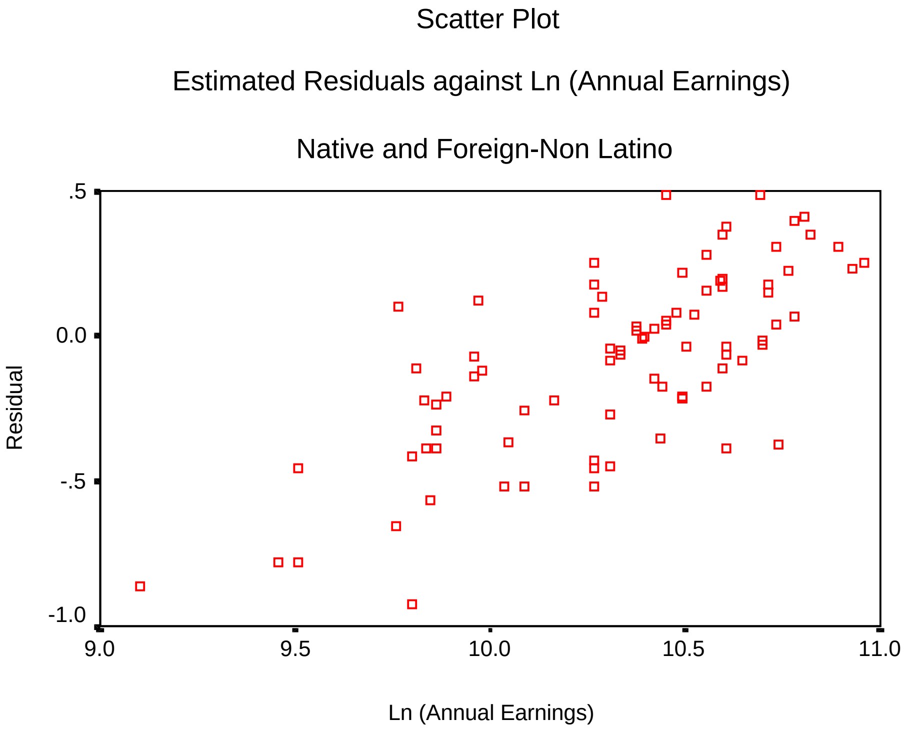
Figure 8: Est. Residuals against Ln (Annual Earnings) for Native and Foreign-Latino


Scatter Plot

Estimated Residuals against Ln (Annual Earnings)

Native and Foreign-Latino

Figure 9: Est. Residuals against Ln (Annual Earnings) for Native and Foreign-Non Latino


Scatter Plot

Estimated Residuals against Ln (Annual Earnings)

Native and Foreign-Non Latino

Column (2) and 3 in Table 4 show the regression results for the first and the second regressions respectively. Regression (2) has 0.442 adjusted R square and 10.618 F-ratio, whereas regression (3) has 0.343 adjusted R square and 8.741 F-ratio. Both regressions are significant with high F-ratios, which validate the relationship of our variables. Signs of the coefficients are in line with our prior assumptions as well. All coefficients in regression (2) are significant, but regression (3) has insignificant HSG and FOR_YSM although FOR_YSM is significant at 10 percent confidence level. Following similar approach to test the second hypothesis, we employ equation (9) using relevant coefficients in regression (2) and (3) to test hypothesis (3). Column (2) and (3) in Appendix G shows that negative earning differentials between native and foreign-Latino, and native and foreign-non Latino decrease as the years since migration increase. However the earning differential for native and foreign-non Latino decrease at a much slower pace. Whereas the *foreign-Latino needs 22 years to catch-up, it takes 38 years for the foreign-non Latino to achieve similar state* (Figure 7). Therefore the opposite holds for our last hypothesis.

VII. Conclusion

Our analysis suggests that foreign-born of South Hayward initially has 52 percent lower annual earnings than that of the native-born. This finding is in par with the assimilation theory that claims the immigrants' lack of local labor market skills upon their arrival in the United States explains their lower initial income in comparison to native-born employees.

We also find that as South Hayward immigrants stay in the U.S., their earnings begin to catch-up with those of comparable native-born. Our analysis estimates that the catch-up will take place in 29 years.

The convergence of the earning differential also happens to South Hayward foreign-Latino and foreign-non Latino. But the convergence for the foreign-non Latino will occur at much slower rate, about 16 years longer, than that of the foreign-Latino. This finding seems to contradict with the theory that cohorts with large ethnic enclave, such as foreign-Latino, have less incentive to invest in human capital, thus lowering their assimilation rates. Since we aggregate foreign-born non-Latino ethnic groups into one category, we have no information to determine if they have higher ethnic enclave effects than that of foreign-born Latino. Slower assimilation for foreign-non Latino also may be due to taste based discrimination (Becker, 1971) as Lalonde and Topel (1997) explain, “If there is discrimination, the wage gap would not be closed completely between natives and assimilated immigrants.” But we will leave this speculative explanation to future research.

References

Altonji, J. G. and R. M. Blank. “Race and Gender in the Labor Market.” Chapter 48 in Ashenfelter, O. and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3c, North-Holland, 1999.

Alameda County Public Health Department. *South Hayward Community Information Book 2001*. August 2001.

Becker, Gary S. *Human Capital*. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1964.

--------------------.*The Economics of Discrimination*. Second edition, Chicago & London: The University of Chicago Press, 1971.

Borjas, George J. “Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants.” *Journal of Labor Economics*, 3(4), Oct. 1985, pp. 463-89.

---------------------. “Assimilation and the Earnings of Immigrants.” *Friends or Strangers: The Impact of Immigrants on the U.S. Economy*. Chapter 6, New York: Basic Books, Inc., 1990, pp. 97-114.

---------------------. "Ethnic Enclaves and Immigrant Assimilation." *Paper presented at the Swedish Economic Council's Conference on: The Assimilation of Immigrants in the Labour Market*, Stockholm, March 13, 2000.

Brescia-Pena, Ande, and Debbie Jones-Ohel. *Resident Survey Handbook: A Guide to Conducting Surveys of Households in South Hayward Community Outreach Partnership Center (COPC).* The HIRE Center, California State University, Hayward, April 2002.

Casutto, Alex. “Sampling Lecture”, California State University, Hayward, Spring 2002

Chiswick, Barry R. “The Effect of Americanization on the Earnings of Foreign-born Men.” *The Journal of Political Economy*, Vol. 86, Issue 5 (Oct., 1978), 897-921.

----------------------. “A Longitudinal Analysis of the Occupational Mobility of Immigrants.” In *Proceedings of the 30th Annual Winter Meeting, Industrial Relations Research Association,* edited by Barbara Dennis, Wis., 1978 (b).

De Vaus, David. *Research Design in Social Research*. Sage Publications, 2001.

Gujarati, Damodar. “Regression on Dummy Variables.” *Basic Econometrics*, 3rd ed., Chapter 15, McGraw Hill, 1995, pp. 499-539.

Hanoch, G. “A Multivariate Model of Labor Supply: Methodology for Estimation,” *Rand Corporation Paper R-1980*, September, 1976.

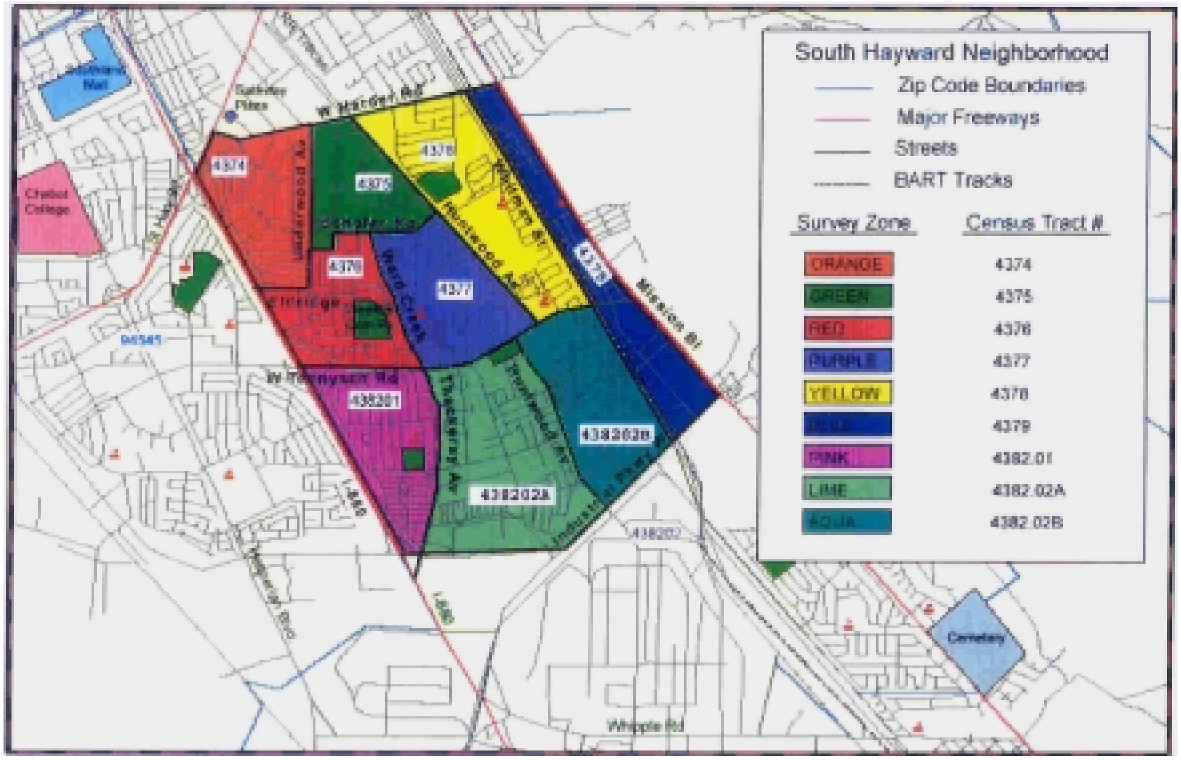
Heckman, James J. “Sample Selection Bias as a Specification Error.” *Econometrica*, Volume 47, Issue 1, January 1979, pp. 153-162.

LaLonde, Robert J., and Robert H. Topel. “The Economic Impact of International Migration and the Economic Performance of Migrants.” Chapter 14 in Rosenzweig and Stark (eds.), *Handbook in Population and Family Economics*, Elsevier Science B.V., 1997.

Maxwell, Nan. “Research Design Lecture”, California State University, Hayward, Spring 2002

Mincer, Jacob. *Schooling, Experience, and Earnings*. New York: Nat. Bur. Econ. Res., 1974.

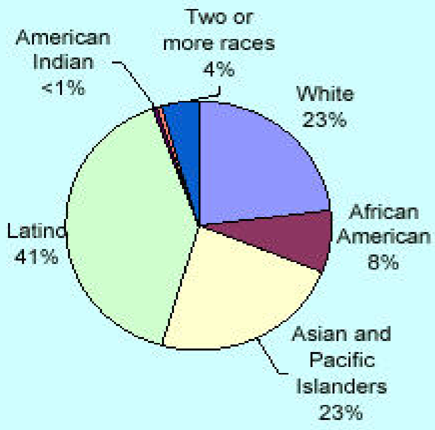
Appendix A: South Hayward Census Tracts Map



Source: Resident Survey Handbook, HIRE Center (April, 2002)

Appendix B: Demographics and Social Profile of South Hayward

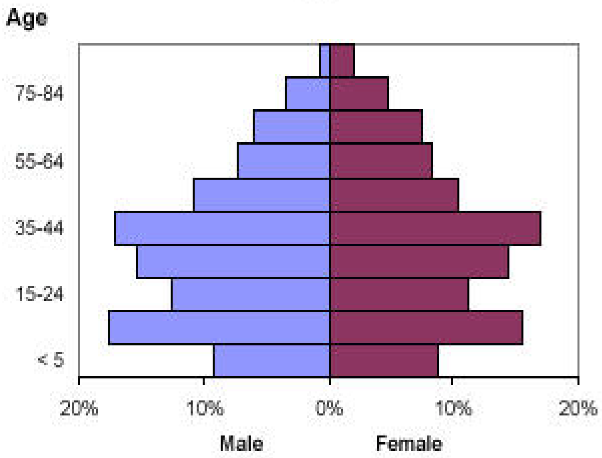
South Hayward
Racial and Ethnic Composition, 2000



Total Population = 37,639

Source: Census, 2000

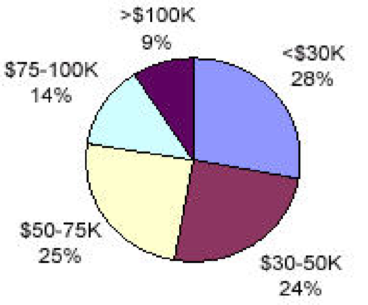
South Hayward Age Distribution by Sex, 1999



Total = 15,897 Total = 16,710

Source: Claritas, 1999

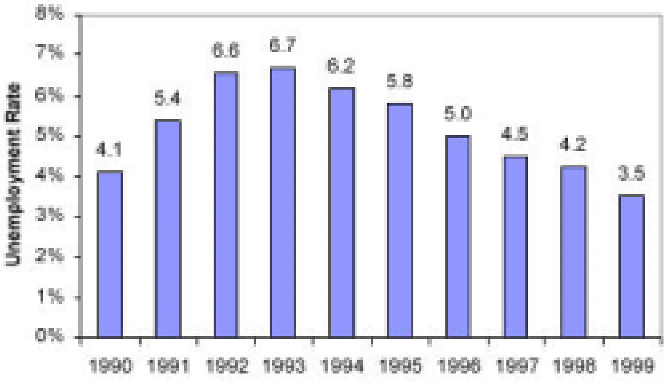
South Hayward Household Income Distribution, 1999



Total Number of Households = 10,605

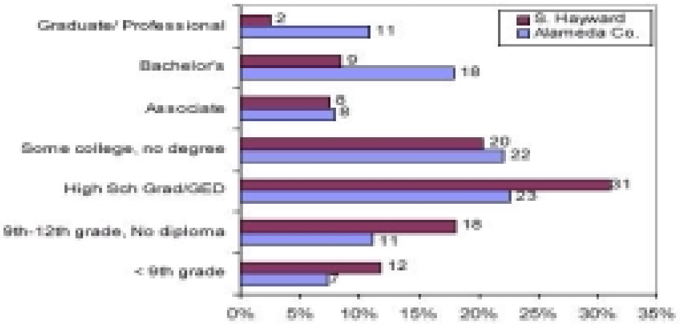
Source: Claritas, 1999

Average Annual Unemployment Rate
Hayward, 1990-1999



Source: California Dept. of Finance, 2000

South Hayward Educational Attainment, 1999



Total Adults Ages 25+ = 20,422

Source: Claritas, 1999

Appendix C: Survey Instrument

Appendix D: Regression Output for *Native and Foreign-Born*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.651	.424	.390	.3937

a Predictors: (Constant), FEM, COLG, T2, FOR, HSG, FOR_YSM, T

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.251	7	1.893	12.213	.000
	Residual	17.979	116	.155		
	Total	31.230	123			

a Predictors: (Constant), FEM, COLG, T2, FOR, HSG, FOR_YSM, T
b Dependent Variable: LN_AE

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	10.106	.168		60.177	.000
	FOR	-.523	.121	-.519	-4.309	.000
	HSG	.221	.121	.159	1.825	.071
	COLG	.340	.085	.302	3.999	.000
	T	3.817E-02	.010	.917	3.691	.000
	T2	-8.272E-04	.000	-.981	-3.949	.000
	FOR_YSM	1.811E-02	.006	.369	2.995	.003
	FEM	-.354	.073	-.349	-4.846	.000

a Dependent Variable: LN_AE

FOR – dummy variable for foreign-born if unity, zero for native-born

HSG – dummy variable for high school graduate if unity, zero for otherwise

COLG – dummy variable for college graduate if unity, zero for otherwise

T – total Labor Market experience

T2 – squared of total Labor Market experience

FOR_YSM – interaction between FOR dummy variable and years since migration

FEM – dummy variable for female if unity, zero for male

Appendix E: Regression Output for *Native and Foreign-Latino*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.699	.488	.442	.3733

a Predictors: (Constant), FEM, T2, COLG, FOR, HSG, FOR_YSM, T

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.359	7	1.480	10.618	.000
	Residual	10.871	78	.139		
	Total	21.231	85			

a Predictors: (Constant), FEM, T2, COLG, FOR, HSG, FOR_YSM, T
b Dependent Variable: LN_AE

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	9.976	.199		50.062	.000
	FOR	-.432	.192	-.360	-2.250	.027
	HSG	.406	.151	.301	2.693	.009
	COLG	.261	.106	.213	2.469	.016
	T	3.160E-02	.011	.762	2.763	.007
	T2	-5.660E-04	.000	-.653	-2.353	.021
	FOR_YSM	1.998E-02	.009	.336	2.154	.034
	FEM	-.438	.085	-.438	-5.141	.000

a Dependent Variable: LN_AE

FOR – dummy variable for foreign-born if unity, zero for native-born

HSG – dummy variable for high school graduate if unity, zero for otherwise

COLG – dummy variable for college graduate if unity, zero for otherwise

T – total Labor Market experience

T2 – squared of total Labor Market experience

FOR_YSM – interaction between FOR dummy variable and years since migration

FEM – dummy variable for female if unity, zero for male

Appendix F: Regression Output for *Native and Foreign-Non Latino*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.622	.387	.343	.3904

a Predictors: (Constant), FEM, COLG, T, FOR, HSG, FOR_YSM, T2

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	9.328	7	1.333	8.741	.000
	Residual	14.788	97	.152		
	Total	24.116	104			

a Predictors: (Constant), FEM, COLG, T, FOR, HSG, FOR_YSM, T2
b Dependent Variable: LN_AE

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		**B**	**Std. Error**	**Beta**		
1	**(Constant)**	10.114	.198		51.061	.000
	FOR	-.473	.142	-.474	-3.323	.001
	HSG	.176	.167	.103	1.055	.294
	COLG	.382	.090	.360	4.263	.000
	T	3.760E-02	.011	.951	3.276	.001
	T2	-7.959E-04	.000	-.998	-3.323	.001
	FOR_YSM	1.263E-02	.008	.246	1.680	.096
	FEM	-.313	.078	-.325	-4.019	.000

a Dependent Variable: LN_AE

FOR – dummy variable for foreign-born if unity, zero for native-born

HSG – dummy variable for high school graduate if unity, zero for otherwise

COLG – dummy variable for college graduate if unity, zero for otherwise

T – total Labor Market experience

T2 – squared of total Labor Market experience

FOR_YSM – interaction between FOR dummy variable and years since migration

FEM – dummy variable for female if unity, zero for male

Appendix G: Comparison of Predicted % Earning Differentials

YSM	Predicted % Earning Differential		
	Native vs. Foreign-Born (1)	**Native vs. Foreign-Latino (2)**	**Native vs. Foreign-Non Latino (3)**
1	-50.5%	-41.2%	-46.0%
5	-43.2%	-33.2%	-41.0%
10	-34.2%	-23.2%	-34.7%
15	-25.1%	-13.2%	-28.4%
20	-16.1%	-3.2%	-22.1%
21	-14.2%	-1.2%	-20.8%
22	-12.4%	**0.8%**	-19.5%
23	-10.6%	2.8%	-18.3%
24	-8.8%	4.8%	-17.0%
25	-7.0%	6.8%	-15.7%
26	-5.2%	8.8%	-14.5%
27	-3.4%	10.8%	-13.2%
28	-1.6%	12.8%	-12.0%
29	**0.2%**	14.8%	-10.7%
30	2.1%	16.8%	-9.4%
31	3.9%	18.8%	-8.2%
32	5.7%	20.8%	-6.9%
33	7.5%	22.8%	-5.6%
34	9.3%	24.7%	-4.4%
35	11.1%	26.7%	-3.1%
36	12.9%	28.7%	-1.9%
37	14.7%	30.7%	-0.6%
38	16.5%	32.7%	**0.7%**
39	18.3%	34.7%	1.9%
40	20.2%	36.7%	3.2%
45	29.2%	46.7%	9.5%
50	38.3%	56.7%	15.8%