

Socioeconomic Status and MOOC Enrollment: Enriching Demographic Information with External Datasets

John Hansen

Harvard University

john_hansen@mail.harvard.edu

Justin Reich

Harvard University

justin_reich@mail.harvard.edu

ABSTRACT

To minimize barriers to entry, massive open online course (MOOC) providers collect minimal demographic information about users. In isolation, this data is insufficient to address important questions about socioeconomic status (SES) and MOOC enrollment and performance. We demonstrate the use of third-party datasets to enrich demographic portraits of MOOC students and answer fundamental questions about SES and MOOC enrollment. We derive demographic information from registrants' geographic location by matching self-reported mailing addresses with data available from Esri at the census block group level and the American Community Survey at the zip code level. We then use these data to compare neighborhood income and levels of parental education for U.S. registrants in HarvardX courses and the U.S. population as a whole. Overall, HarvardX registrants tend to reside in more affluent neighborhoods. U.S. HarvardX registrants on average live in neighborhoods with median incomes approximately .45 standard deviations higher than the U.S. population. Parental education is also associated with a higher likelihood of MOOC enrollment. For instance, a seventeen year-old whose most educated parent has a bachelor's degree is more than five times as likely to register as a seventeen year-old whose most educated parent has a high school diploma.

Author Keywords

MOOCs; demographics; socioeconomic status; geographic analysis

ACM Classification Keywords

K.3.1 Distance Learning.

1. BACKGROUND AND CONTEXT

Advocates for Massive Open Online Courses (MOOC) have promoted their potential to make education “class-blind” [1] and “allow people who lack access to world-

class learning...an opportunity to make a better life for themselves and their families” [7]. In ideal circumstances for “class-blind” educational opportunities, we would hope to see society’s least affluent students equally or over-represented in the distribution of registrants. The earliest evidence of MOOC enrollment, however, suggests that most registrants in courses already have a Bachelor’s degree [4, 9], raising the question of whether MOOCs are “reinforcing the advantages of the ‘haves’ rather than ‘educating’ the ‘have-nots’” [5]. Understanding patterns of MOOC enrollment and engagement for students of different backgrounds is critical to understanding the role that MOOCs can play in ameliorating or exacerbating educational inequalities.

In order to understand and address educational inequalities, MOOC researchers need detailed demographic information about participating students. In particular, we are interested in demographic information that provides insight into registrants’ socioeconomic status (SES), which is broadly defined as “one’s access to financial, social, cultural, and human capital resources” [11]. A panel of experts assembled by the National Center for Education Statistics identifies a triad of measures that are generally agreed to be most useful in characterizing student SES: family income, parental educational attainment, and parental occupational status. This information is regularly available to higher education researchers, as colleges typically ask applicants to report their parents’ educational attainment and occupation, and applications for financial aid request detailed information on family income and assets.

These SES measures, however, are not available for MOOC registrants. In an effort to maximize enrollment, MOOC providers minimize barriers to entry for courses, so students can register with only a few mouse clicks. As a result, MOOC providers rely on brief, voluntary surveys to gather demographic data. For instance, to register for the edX site, registrants are only asked to provide their gender, year of birth, level of education, and address. These forms do not paint a detailed demographic picture of the registrant. Many courses then offer more detailed surveys for registered students, but course faculty and MOOC providers are rightly hesitant about soliciting personal information in the early days of a course. Investigating the impact of MOOCs on students from different backgrounds requires creative approaches to leveraging available data.

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

- ACM copyright: ACM holds the copyright on the work. This is the historical approach.
- License: The author(s) retain copyright, but ACM receives an exclusive publication license.
- Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in TimesNewRoman 8 point font. Please do not change or modify the size of this text box.

In this paper, we show how third-party census datasets can be used to enrich the limited SES data collected by a MOOC platform in order to provide a richer understanding of student background characteristics. We demonstrate three important uses of these third-party datasets: merging, comparing, and response bias testing. First, we use self-reported student addresses to identify student neighborhoods, and then use data about a neighborhood's median household income as a measure of SES, as discussed by the NCES panel [11]. Previous research has found that neighborhood affluence is related to college attendance for adolescents [3] and levels of peer group income and educational attainment for adults [10]. Second, we compare the distributions of median neighborhood income and parental education levels of MOOC registrants to the U.S. population as a whole. Third, we use the more complete address data—which becomes neighborhood data—from the edX site registration to check for SES-related response-bias in the parental education item from HarvardX course surveys.

We demonstrate the importance of these data-analytic strategies in an investigation of the relationship between SES and registration in nine MOOCs offered by HarvardX—an initiative by Harvard University to offer MOOCs through the edX platform—in the 2013-2014 academic year. To better understand how access to resources shapes MOOC enrollment, we offer comparisons of HarvardX registrants and the U.S. population with respect to neighborhood affluence and parental educational attainment. In doing so, we demonstrate the value of third-party datasets in estimating the extent to which new forms of online learning promote, or fail to promote, more equitable access to educational opportunities.

2. DATASET

Our primary dataset of MOOC enrollees includes registrants from nine HarvardX courses conducted during the 2013-2014 school year on the edX platform. We include all courses offered between September 2013 and June 2014, hosted on the edX platform that issued certificates to students who earned a sufficiently high grade. If a course was in a series of modules, we include only the first module in the series. Including US users, non-US users, and users whose country could not be identified, our dataset contains 238,572 unique registrants. We focus on the US population, whom we identify through a combination of self-reported country and IP geo-location. We drop 145,858 registrants reporting a country outside the US and registrants reporting no country who were not geo-located to the United States. To avoid issues of sparse data and possible misrepresentation of extremely young or old ages, we drop 12,275 registrants reporting an age younger than 13 or older than 69. This leaves us with 79,921 unique registrants.

In order to registering for a HarvardX course on the edX platform, users were invited to complete a site registration

on edX. This site registration includes a voluntary four-question survey asking for student gender, level of education, year of birth, and address. Address is collected as a single open-response text field. Of our 79,921 registrants—60,892 or 76% completed the entire site registration survey.

Upon enrolling in a HarvardX course, registrants were then asked to complete a “pre-course survey,” but no mechanism existed to enforce completion (students could register for the course without even starting the survey). Registrants were asked to complete the survey through in-course prompts and emails, but response rates were generally low as a proportion of all registrants, ranging from 14% to 45%. This survey included approximately 20 items, including a question asking students to self report their mother's and father's highest level of education, and response rates to this item varied from 13% to 34% across courses.

We use three additional sources of publicly available data in our analyses. The first source contains data from the American Community Survey (5-year estimates for 2008-2012) organized at the zip code level and available for download from the U.S. Census Bureau website. The U.S. Census Bureau delivers the American Community Survey (ACS) each year to several million American residents, and survey items include age, education, and income [16]. This dataset offers the simplest way for enriching our dataset with neighborhood-level demographic data. The second source also relies on the ACS 5-year estimates, but this dataset is organized at the person level and available for download from the Minnesota Population Center's IPUMS website [12]. We use this dataset to compare the U.S. and HarvardX distributions of parent education for adolescents. Lastly, we use a 2013 U.S. demographic dataset available from Esri, a for-profit provider of geographic information systems software and data. Esri's 2013 demographic dataset [6] provides estimates for median household income and age distributions within each census block group, and we used the ArcGIS software package to match registrants to census block group-level variables.

3. DATA ANALYTIC STRATEGY

Our analysis of the relationship between SES and MOOC enrollment proceeds in three parts. First, we demonstrate two ways of obtaining measures of median neighborhood income for HarvardX students, using only zip code and then using a full address. As an exploratory step, we compared the distributions of zip code median neighborhood for HarvardX registrants and the U.S. In subsequent analyses, we condition our findings on age, since both neighborhood income (as we will show) and educational attainment [8] are associated with age, and as shown in Figure 1, the distribution of ages among U.S. HarvardX registrants differs from the U.S. population.

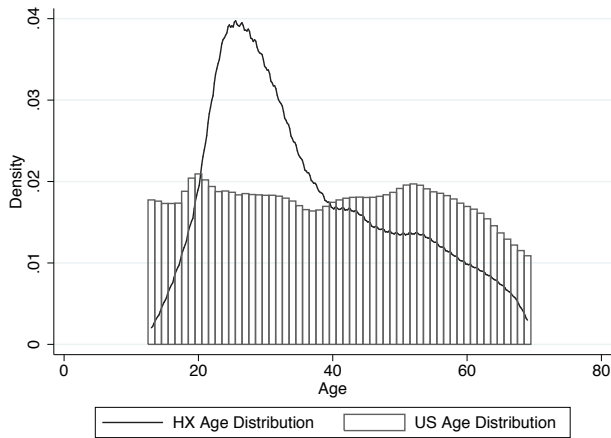


Figure 1: Distribution of US non-institutionalized residents and 2013-2014 HarvardX registrants by age.

Second, we obtain measures of parental education from pre-course surveys, and use our more complete address data to test for response bias. Again, as an exploratory step, we compare the distributions of parental education of HarvardX registrants and U.S. persons conditional on age.

Third, we append our HarvardX dataset to the ACS and Esri datasets and fit regression models to estimate the likelihood of a U.S. resident registering for a HarvardX course, conditioned on our two SES measures and age. From these models, we can evaluate the substantive effects of parent education and neighborhood income on MOOC enrollment. We also estimate SES differences by course, which confirms our general findings.

4. DERIVING NEIGHBORHOOD INCOME FROM ADDRESS

We demonstrate two approaches to parsing mailing addresses provided by HarvardX registrants and using this information to obtain further information about median neighborhood income. Our first approach was to parse self-reported mailing addresses from the edX site registration for a zip code and match that to demographic data available at the zip code level from the American Community Survey. The overall median zip code income estimate for the US is \$53,046 [14]. Since the ACS includes the count of occupied housing units by zip code, we computed the *mean* of median zip code household income in the US by taking a weighted average. The weighted mean of median zip code incomes in the US was \$56,532 (sd = 22,876). Estimates are in 2012 dollars. The mean of median zip code household income observed in the HarvardX sample of unique registrants was \$68,262. The dashed line in Figure 2 shows the zip code median household income for unique

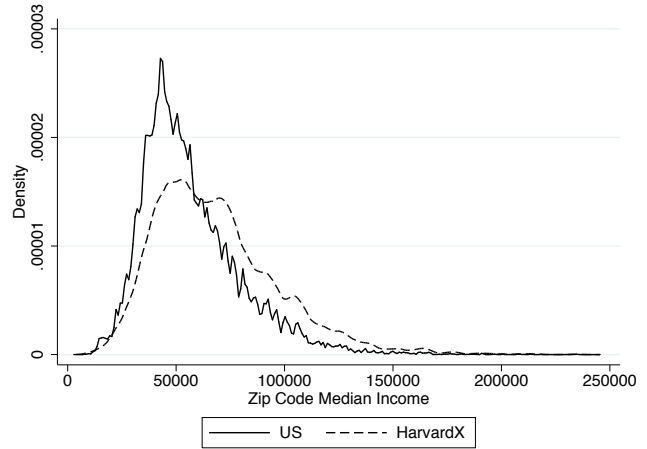


Figure 2: Weighted average distributions of median household income by zip code for US non-institutionalized residents and 2013-2014 HarvardX registrants.

HarvardX registrants reporting a valid zip code ($n=51,751$) tends to be higher than the general US population. The difference between unique HarvardX registrants and the US population is \$11,730, or approximately .51 standard deviations ($11,730/22,876$).

Zip codes have the advantage of being easy to parse from an open field mailing address, but there are drawbacks of using the ACS data aggregated at the zip code level. One issue is that zip code boundaries are drawn—and occasionally redrawn—by the US Postal Service in order to support efficient mail delivery [15], not the interests of demographically-minded researchers. Additionally, the zip code-level dataset we used here was not disaggregated by age; it only contained estimates of the number of households per zip code. Since age and median zip code income are associated—as shown in Figure 1—failing to take into account differences in the age distributions between HarvardX and the US could lead to biased estimates of the difference in neighborhood income for HarvardX registrants and the US population.

To address these concerns, in our second approach to deriving neighborhood income data, we use a demographic dataset available from Esri. In this dataset, the geographic unit of analysis is the census block group and population counts are disaggregated by age [6, 15]. We parsed the mailing address from the edX site registration survey into several fields, which allowed geocoding software to estimate latitude and longitude coordinates, and we performed a spatial join to match HarvardX registrants with demographic data for their census block group.

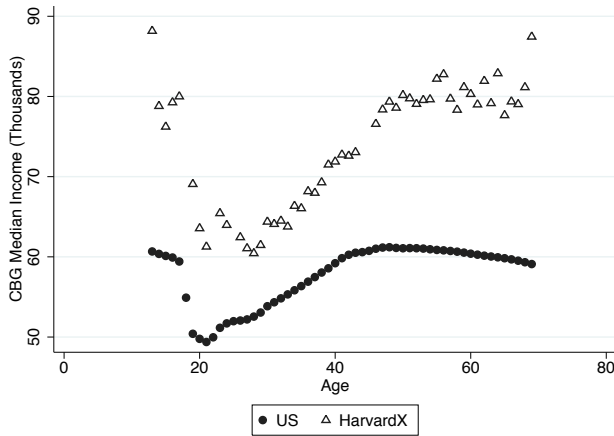


Figure 3: Average census block group median income by age for all U.S. non-institutionalized residents and 2013-2014 HarvardX registrants.

Figure 3 shows that HarvardX registrants reporting a parsable mailing address ($n=44,362$) tend to live in census block groups with higher median household incomes than individuals of the same age in the U.S. According to our dataset, where 2013 census block group median income is top-coded at \$200,001, the weighted mean in the US is \$57,642 ($sd = 30,535$). For unique HarvardX registrants, the comparable mean is \$70,646. Notice that we find the difference to be greater for younger and older students compared to students between 20 and 40. Using the same approach as we did with ACS zip code data, we compute a difference in neighborhood income of \$13,184 or .425 standard deviations. We can further refine these comparisons by simultaneously taking into account the differences in registration probability and neighborhood median income attributable to age. To estimate the average difference in neighborhood median household income between HarvardX registrants and the U.S. population *taking into account the cross-sectional association between age and neighborhood income*, we fit the following model:

$$med_hinc_{ij} = \beta(HX_i) + \delta_j + \epsilon_i,$$

where β is the estimated average difference in neighborhood median income between HarvardX registrants and members of the US population of the same age. δ_j estimates the average difference in neighborhood income attributable for each age from 13-69 (using a series of 57 dummy variables, one for each age), HX is a dummy variable equal to 1 for HarvardX registrants (and 0 otherwise), and ϵ is a random error term assumed to be normally distributed with mean 0. The coefficient of interest, β , can be thought of as the weighted average of differences between each pair of dots and triangles in Figure 3. Our estimate for β is \$14,009 ($p < .05$), with a 95% confidence interval of 13,727 to 14,291. This is approximately .46 standard deviations in terms of the unconditional standard deviation, though standard deviation of neighborhood income varies with age.

Taken together, these three methods—average difference in median zip code income, average difference in median census block income, and average difference in median census block income controlling for age—suggest that HarvardX registrants live in neighborhoods with median incomes almost one-half of a standard deviation higher than the general U.S. public. Our estimates speak to the general efficacy of deriving demographic data from self-reported addresses, and this approach can easily be extended to include other demographic data published at the “neighborhood” level. Considering the interest in identity verification in MOOCs, it’s plausible that address will remain an accessible and useful SES measure for researchers in the future, especially if trends in U.S. residential segregation by income [17] continue to rise.

5. COMPARING PARENTAL EDUCATION AND ADDRESSING MISSING DATA

To continue to develop our comparison of SES measures of HarvardX registrants and the U.S. population, we next examine the distributions of parental educational attainment. Here we focus primarily on the distribution of adolescent registrants for HarvardX. Recall that we obtain self-reported data of attainment levels from surveys disseminated after course registration, and we can compare these with U.S. characteristics obtained from the American Community Survey. We focus on adolescents in our HarvardX/U.S. comparisons for two reasons: first, this is an age group where parent education is clearly established in the literature as an important measure of SES [11], and second, the collection of ACS data on parent education becomes less reliable for individuals older than seventeen, since it is only gleaned from households where children live with their parents [13]. Following a common convention [11], we define parental educational attainment as the highest degree attained for one’s most educated parent. Figure 4 compares the parental education attainment levels for HarvardX registrants and the US population as

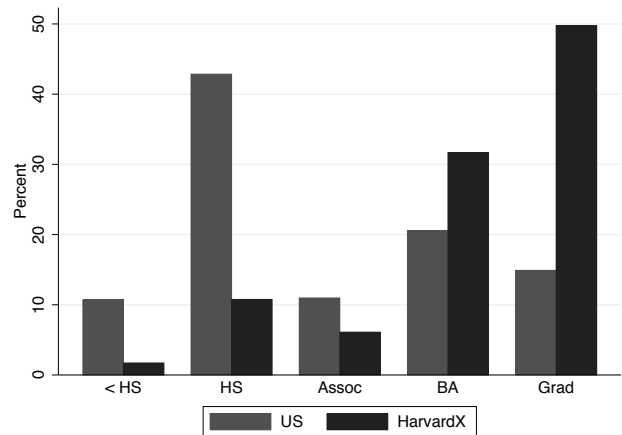


Figure 4: Parental educational attainment among 13-17 year olds, for all U.S. non-institutionalized residents and 2013-2014 HarvardX registrants.

computed from the ACS. Figure 4 shows that adolescent HarvardX registrants tend to report higher levels of parental educational attainment than typical adolescents in the US. While the modal value for the US is high school completion, the modal value for HarvardX registrants is a graduate degree.

One potential concern with these HarvardX data is response bias. Since only 34% of registrants provided their parent’s education, survey respondents and non-respondents systematically may differ. However, another asset of using these external data sources is that we can test for evidence of response bias. Since we have more complete data for HarvardX neighborhood median income than parental education, we can exploit the relationship between neighborhood household income and parent education to test the null hypothesis that, among HarvardX registrants reporting a parsable address, neighborhood median income is the same for individuals who did and did not respond to the parent education survey item. In order to take into account the relationship between neighborhood median income and age, we test this hypothesis by regressing census block group median income on dummy variables for age and a dummy variable equal to 1 for registrants not reporting parent education. In this case, we are interested in comparing one subset of HarvardX registrants to another subset of HarvardX registrants, rather than comparing HarvardX registrants to the U.S. population. We fit this model twice; first, we look at 13-17 year-old HarvardX registrants, and second we fit an analogous model on all HarvardX registrants for whom we have neighborhood income data. Specifically, first, we fit the following model on the 864 13-17 year-old HarvardX registrants for whom we have an estimate for the median household income in their neighborhood:

$$med_hinc_{ij} = \beta(missing_ped_i) + \delta_j + \epsilon_i,$$

where β is our coefficient of interest: the average difference in neighborhood median income between HarvardX registrants reporting parent education and HarvardX registrants not reporting parent education. δ_j is one of five dummy variables estimating the average difference in neighborhood income attributable to age in the population (one dummy for each age from 13 to 17); and ϵ is a random error term assumed to be normally distributed with mean 0. In our model, we fail to reject the null hypothesis that the coefficient β is statistically different from zero ($p > .05$). For a test with greater statistical power, we fit the same model on all 44,364 unique HX registrants of all ages with parsable mailing addresses, and again we fail to reject the null hypothesis ($p > .05$) that neighborhood income is different for individuals responding to the survey item about parental education. Failing to reject the null hypothesis does not prove that survey non-respondents are not demographically different from survey respondents with respect to SES or other important ways, but the lack of a within-age group difference in neighborhood income

suggests that SES levels for survey respondents are reasonable estimates for SES levels for HarvardX registrants not responding to the parent education survey item but reporting a parsable address.

6. HOW DOES SES PREDICT THE LIKELIHOOD OF MOOC REGISTRATION?

In the previous sections we used third-party datasets for three purposes: (1) to match address data with neighborhood income data to enrich our demographic portrait of HarvardX students, (2) to descriptively compare distributions of median neighborhood income and parental education between HarvardX registrants and the U.S. population, and (3) to evaluate response bias in our survey data. Here, we demonstrate their use in estimating the effect of SES measures on the likelihood of registering for a HarvardX course, conditional on student age.

In this section, we append our HarvardX dataset to one of two nationally representative datasets and fit regressions models to offer greater insight into the demographics of HarvardX registrants. We do this by literally appending the HarvardX dataset to either the ACS or Esri dataset, yielding a representative sample for the United States. Since variables for the ACS and Esri dataset focus on different units of observation (the person in the former and the census block group in the latter), we must evaluate the effect of median income and parental education separately. When we append the HarvardX dataset to the ACS dataset, we add approximately 75,000 observations to the probability weight-implied 232 million observations that represent the entire non-institutionalized U.S. population aged 13-69, making it only trivially less representative. Similarly, when we append the HarvardX dataset to the Esri dataset, we add the 75,000 observations to the approximately 232.5 million frequency weight-implied observations, again making it trivially less representative.

For parent education, we fit a model that can be expressed in the following mathematically equivalent ways:

$$Pr(reg_{ij} = 1 | X_{ij}) = (1 + \exp(-(\beta(ped_i) + \delta_j)))^{-1}$$

$$\text{logit}(pr(reg_{ij} = 1 | X_{ij})) = \beta(ped_i) + \delta_j,$$

where the coefficient of interest is β , the within-age difference in the probability—or the difference in the log-odds in the second equation—of HarvardX registration attributable to a difference in parent education. In Model 1, we estimate a separate β for each level of parental education, and in Model 2 we use an ordinal scale such that 1 = Less than HS completion, 2 = HS completion, 3 = Associate’s degree, 4 = Bachelor’s degree, and 5 = Graduate degree. In Models 1 and 2, δ_j is a dummy variable for each age in the model, accounting for the difference in registration likelihood attributable to age. The effect of β is nonlinear on a probability scale but constrained to be linear on a log-odds (or logit) scale.

Table 1: Logistic regression models of HarvardX registration likelihood for age 13-17 by parent education.

	(1) Parent Ed Dummies	(2) Linear Parent Ed
< HS	0 (.)	
HS	0.451 (0.413)	
Assoc	1.250** (0.438)	
Bach	2.271*** (0.392)	
Grad	3.054*** (0.387)	
Parent Ed		0.851*** (0.0499)
Observations	186320	186320
neg2ll	41239.8	41248.5
r2_p	0.0640	0.0638
df_m	8	5

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Coefficients for Age Dummy Variables Omitted

In Table 1, we present a taxonomy of logistic regression models estimating the likelihood of enrolling in a HarvardX course conditional on age and parental education. In Model 2, we include dummy variables for age and model the effect of parent education without constraining its functional form. We include dummy variables for the levels of educational attainment shown in Figure 3 (omitting “less than HS” in order to avoid collinearity). In Model 3, we assume a linear effect of parent education. Assuming a linear effect of parent education on the log-odds of registration, we estimate that on average, a one-unit increment on our parent education scale is associated with a .851 increase in the log of the odds of HarvardX registration.

The absolute difference on a probability scale depends on the baseline probability of the enrollee. The similarity of the results of Models 1 and 2, also shown in Figure 5, demonstrate that the relationship between registration and parent education is reasonably well approximated by our parent education scale, where we constrain each unit difference on the scale has the same effect on the relative odds of registration.

The difference in the probability of enrollment for two seventeen year-olds—where one’s most educated parent has a bachelor’s degree and the other’s most educated parent has a graduate degree—is approximately .0004. While this seems small on an absolute probability scale, the probability of registration for a random seventeen year-old whose most educated has a bachelor’s degree is only .0003. Therefore, we find that a one-unit change in parent

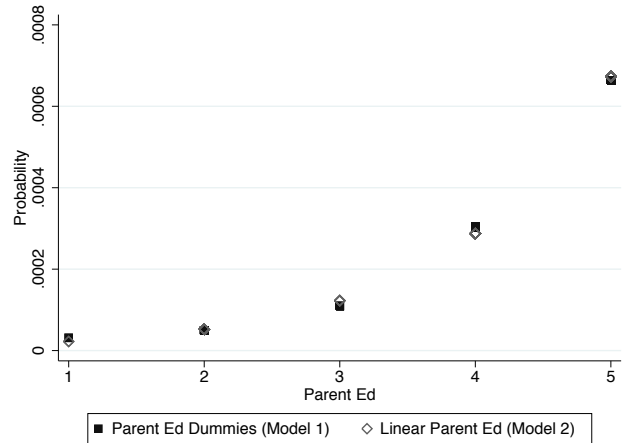


Figure 5: Probability of HarvardX registration by level of parental education for 17 year olds.

education is associated with an individual being more than twice as likely to register.

To interpret our model in an alternative but equivalent way, we estimate that for every 100,000 seventeen year-olds in the United States who do not have a parent who graduated from high school, 2 would register for HarvardX course in our dataset. Among 100,000 random seventeen year-olds whose most educated parent has a high school diploma, approximately 5 would register. For 100,000 whose parents have an associate’s, bachelor’s, and graduate degree, the respective estimates are approximately 12, 29, and 67. *If there were an equal number of seventeen year-olds whose most educated parent had a high school diploma and a bachelor’s degree*, we would estimate that more than five times as many registrants’ most educated parent would have a BA. However, HarvardX registrants with well-educated parents do not actually outnumber those with less-educated parents by such dramatic margins because, as shown in Figure 4, it is far more common in the US for parents of teenagers to have a high school degree compared to a postsecondary degree. Failing to take into account the underlying distribution of educational attainment in the US would understate how unrepresentative the parents of adolescent HarvardX registrants are compared to the US in terms of educational attainment.

In Table 2, we present a taxonomy of logistic regression models predicting HarvardX registration using the neighborhood median income measure from the Esri dataset. Unlike our parent education models, we don’t exclude older registrants. In Model 1, we constrain the predictive effect of neighborhood income to be the same for all age groups. In Model 2, we allow the effect of our predictor to vary across age groups, which we define as: 13-17, 18-24, 25-29, 30-39, 40-49, 50-59 and 60-69. In terms of log-odds, we estimate:

$$\text{logit}(\text{pr}(\text{reg}_{ijk} = 1 | X_{ijk})) = \beta_1(\text{med_hinc}_i) + \beta_k(\text{age_group}_k * \text{med_hinc}_i) + \delta_j,$$

Table 2: Logistic regression models of HarvardX registration likelihood by age and neighborhood income.

	(1) med_hinc	(2) Age Group Interactions
med_hinc	0.0123*** (0.000124)	0.0135*** (0.000769)
13.17 * med_hinc		0 (.)
18.24 * med_hinc		-0.0000824 (0.000840)
25.29 * med_hinc		-0.00265** (0.000845)
30.34 * med_hinc		-0.00266** (0.000852)
35.39 * med_hinc		-0.00287*** (0.000869)
40.49 * med_hinc		-0.00191* (0.000817)
50.59 * med_hinc		0.000287 (0.000820)
60.69 * med_hinc		0.00132 (0.000850)
Observations	232277133	232277133
neg2ll	829125.6	828981.4
r2_p	0.0228	0.0230
df_m	57	64

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Coefficients for Age Dummy Variables Omitted

where β_1 is the difference in the log-odds of registration attributable to neighborhood median income for the baseline age group (13-17), and β_k is the differential predictive effect of neighborhood compared to the baseline age group. δ_j is the estimated difference in the relative odds of registration attributable to age. Model 1 only includes a dummy variable for each age from 13-69, and the median household income in an individual's census block group. In Model 2, we allow the predictive effect of neighborhood median income to vary across age groups in order to account for heterogeneity in its effect across age groups.

For two seventeen year olds, a \$1,000 increment in neighborhood median income is associated with a .0135 increase in the log-odds of registration. The mean of census block group median income in the US is \$57,600, and standard deviation of neighborhood median income is \$30,000. If seventeen year-olds were uniformly distributed with respect to neighborhood median income, we would estimate that more than twice as many individuals residing in neighborhoods with \$117,000 median incomes would register compared to neighborhoods with median incomes around the national average. Once again, though, this does

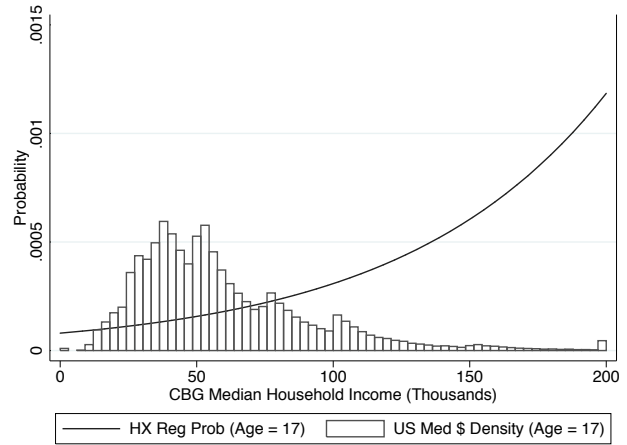


Figure 6: Fitted probability of HarvardX registration for prototypical 17 year-old by median neighborhood income.

not imply that HarvardX registrants from high-earning neighborhoods actually outnumber those from typical neighborhoods by these implied margins because, as shown in Figure 6, it is far more common in the US for an individual's neighborhood median income to be around \$57,000 compared to \$117,000.

Further, we can use the coefficients from the parent education and neighborhood income models to compare the effects of these two SES measures on a similar scale: their association with the likelihood of HarvardX registration for adolescents. Using the coefficients from the parent education and neighborhood income models, we estimate that a neighborhood median income difference of approximately \$63,000 (.851/.0135) is comparable to a one unit difference in parent education. We don't intend to argue for the relative importance of one predictor over the other. Instead, we present this to emphasize a virtue of the linearity of the log-odds scale for comparing predictive effects in the heterogeneous MOOC context.

Finally, in order to evaluate whether these aggregate findings are consistent at the course level, we model the association between HarvardX registration and neighborhood income separately for each of the nine courses in our dataset. The result will be an estimate of each course's "SES-neutrality" on a constant scale, including variability by age group. We could obtain substantively the same results with respect to the courses' "SES-neutrality" relative to one another from either of the following models:

$$med_hinc_{ijk} = \beta_1(HX_i) + \beta_k(age_group_k * HX_i) + \delta_j + \epsilon_i,$$

$$logit(pr(reg_{ijk} = 1 | X_{ijk})) = \beta_1(med_hinc_i) + \delta_j + \beta_k(age_group_k * med_hinc_i).$$

Table 3: Regression models predicting for nine HarvardX 2013-2014 courses the difference in neighborhood income for HarvardX registrants compared to U.S. residents, including variability by age. Coefficients are scaled in thousands of dollars.

	(1) Heroes	(2) Early Christ	(3) Clinical Trials	(4) Public Health	(5) Health Policy	(6) Genomics	(7) Science & Cooking	(8) China	(9) Global Health
hx_reg	19.92*** (3.125)	14.92*** (2.888)	22.59*** (3.911)	19.01*** (2.588)	18.95*** (3.452)	31.49*** (4.674)	19.70*** (1.562)	30.81*** (2.542)	23.88*** (4.567)
13.17 * hx_reg	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
18.24 * hx_reg	-6.601 (3.409)	-5.321 (3.090)	-5.238 (4.163)	-3.986 (2.769)	-0.393 (3.607)	-18.16*** (4.998)	-4.480** (1.687)	-15.63*** (2.790)	-6.520 (4.757)
25.29 * hx_reg	-12.06*** (3.414)	-9.807** (3.034)	-14.08*** (4.092)	-11.69*** (2.770)	-9.142* (3.588)	-24.81*** (4.873)	-8.942*** (1.642)	-20.25*** (2.788)	-13.60** (4.783)
30.34 * hx_reg	-13.80*** (3.491)	-8.377** (3.043)	-11.67** (4.101)	-12.61*** (2.832)	-9.625** (3.641)	-21.94*** (4.925)	-8.592*** (1.649)	-19.29*** (2.873)	-14.39** (4.874)
35.39 * hx_reg	-15.70*** (3.648)	-10.11** (3.086)	-8.487* (4.182)	-13.43*** (2.936)	-7.137 (3.727)	-18.27*** (5.026)	-6.308*** (1.684)	-17.83*** (3.001)	-15.07** (5.086)
40.49 * hx_reg	-10.73** (3.402)	-6.831* (2.987)	-3.881 (4.111)	-11.76*** (2.834)	-3.916 (3.628)	-10.39* (4.947)	-2.569 (1.635)	-14.86*** (2.807)	-8.839 (4.910)
50.59 * hx_reg	-4.353 (3.455)	-3.450 (2.982)	2.060 (4.238)	-7.043* (2.923)	-0.395 (3.696)	-6.150 (5.144)	3.046 (1.651)	-8.945** (2.874)	-6.392 (5.023)
60.69 * hx_reg	2.240 (3.606)	0.0637 (3.020)	1.185 (4.920)	-5.529 (3.356)	5.037 (3.892)	-6.804 (6.572)	3.666* (1.714)	-7.740** (2.908)	-1.654 (5.370)
Observations	232235693	232240911	232235955	232237352	232237346	232234831	232253657	232236799	232234896
r2	0.0158	0.0158	0.0158	0.0158	0.0158	0.0158	0.0159	0.0158	0.0158
df_m	64	64	64	64	64	64	64	64	64

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Coefficients for Age Dummy Variables Omitted

The first model estimates the average difference in neighborhood income between a course's registrants and the U.S., and the second model estimates the difference in the likelihood of course registration attributable to neighborhood income. To allow the association between HarvardX registration and neighborhood income to vary across age groups within a course, we include interactions for age categories. We choose to estimate the first because the coefficients will be estimated on a dollar scale, which is easier to interpret. We show the results in Table 3, and all coefficients are expressed in terms of thousands of dollars.

The top row of coefficients in Table 3 is each course's estimated β_1 , which is the average difference in neighborhood median income within a course between HarvardX registrants and the U.S. for the baseline age group (age 13-17). Each subsequent row contains a course-specific estimate of β_k , which is the average difference in neighborhood income between a given age group and the baseline age group for HarvardX registrants. δ_j accounts for differences in a course's average neighborhood income attributable to the distribution of registrant ages within a course, and ϵ is a random error term assumed to be normally distributed with mean 0. To give an example for the baseline age group, we estimate that Early Christianity's registrants tend to live in neighborhoods where on average

the median income is higher by \$14,920. While we observe heterogeneity across courses and within age groups, nothing from this analysis suggests that our aggregate findings are overly dependent on particular courses. Though our definition of age groups were not entirely arbitrary, estimates in the table above are primarily illustrative. These coefficients may be sensitive to alternative age groupings.

Estimates like these have advantages to descriptive statistics for understanding a course's demographic with respect to SES. If one's question is the extent to which a course's registrants tend to be drawn from more affluent neighborhoods, the table above offers sensible estimates on a scale that lends itself to straightforward cross-course comparisons and the possibility of differential effects by age group. Generalizing about which kinds of courses appeal to which kinds of students is premature given the small number of courses in this study, but a similar approach applied to a larger set of courses could be instructive. In this small number, we see no obvious differences between courses in the humanities, sciences or professions in regard to neighborhood income and course registration.

7. DISCUSSION

If edX and other MOOC platforms promoted a set of truly class-blind educational opportunities, we might expect at least two things to be true of the registered students. First, we would expect to find that the range of students registered for these courses reflected a wide diversity of peoples. Indeed, we do find that the 2013-2014 HarvardX courses included American students living in the nation's poorest neighborhoods as well as students living in the wealthiest neighborhoods, and included students whose parents earned advanced degrees as well as parents with no degree at all.

We might also expect, however, that in a class-blind set of educational opportunities, the distribution of registrants would reflect the distribution of the larger population, and in this dimension, the population of HarvardX registrants is disappointing. U.S. HarvardX registrants on average live in neighborhoods with median incomes approximately .45 standard deviations higher than the U.S. population. Parental education is also associated with a higher likelihood of MOOC enrollment. For instance, a seventeen year-old whose most educated parent has a bachelor's degree is more than five times as likely to register as a seventeen year-old whose most educated parent has a high school diploma. Our findings suggest that the HarvardX courses in our study disproportionately attract individuals already advantaged in terms of access to resources.

This analysis only examines one facet of the complex interplay between MOOCs, SES, and educational inequality. The digital divide is best understood as two divides, one of access and one of usage [2]. While we show here that students from advantaged backgrounds are more likely to enroll in MOOCs, this analysis does not shed light on whether students from low income neighborhood or from families with lower levels of education attainment have lower persistence, participation, or performance within MOOCs. We are currently pursuing these analyses, which are made possible by the more comprehensive demographic portrait available by linking internal MOOC surveys with external sources of data.

One of the challenges of conducting studies on how students from different background engage differently with MOOCs is that the desire to keep MOOC easily accessible often precludes asking probing demographic questions as a condition of entry into a course. For the foreseeable future, we expect that conducting rich inquiries about student background characteristics will require drawing on external datasets, both to expand the information that we have about each student and to better understand how MOOC students compare to populations as a whole. An important contribution of this paper is to demonstrate how external datasets can be used to expand demographic pictures of learners, provide a reference point of comparison between MOOC students and larger populations, and to characterize potential sources of response bias when MOOC data are

collected from multiple sources. While we demonstrate this strategies in the U.S. context, we hope that researchers around the world will explore the use of other national census datasets to extend this line of research into other countries.

There are important ethical issues to be considered in pursuing this research. While edX students voluntarily provide their address and other demographic information to us, and while the terms of service and public discourse around MOOCs make it clear that edX data is used for research purposes, this research demonstrates how the merging of multiple datasets makes it possible for researchers to glean deeper insights about an individual than a person might expect for a particular disclosure. When edX registrants share their address with edX, they may or may not understand that providing an address can provide a wealth of additional information about a person and his or her context. Further survey or interview research with MOOC participants might reveal whether MOOC registrants expect their data to be connected with other data sources as is common in a variety of marketing and political settings, or whether they feel that these kinds of research moves are a breach of trust. While we believe that our research strategies are necessary to better address critical questions about how students from different backgrounds engage with MOOCs, we recognize the importance of participating in a public conversation about how researchers use data from online learner.

8. REFERENCES

- [1] Agarwal, A. Online Universities: It's Time for Teachers to Join the Revolution. *The Observer*, (June 15 2013).
- [2] Attewell, P. Comment: The First and Second Digital Divides. *Sociology of Education*, 74, 3 (Jul. 2001), 252-259.
- [3] Bowen, W. G. and Chingos, M. M. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press, Princeton, NJ, 2009.
- [4] Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D. and Emanuel, E. *The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why?*, 2013. Retrieved October 12, 2014 from SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2350964
- [5] Emanuel, E. J. Online education: MOOCs taken by educated few. *Nature*, 503, 7476 (2013), 342-342.
- [6] Esri. *Esri Updated Demographics*. Esri, Redlands California, 2014.
- [7] Friedman, T. L. Come the Revolution. *New York Times*, (May 15 2012).
- [8] Goldin, C. D. *The race between education and technology*. Cambridge, Mass. : Belknap Press of Harvard University Press, 2008, Cambridge, Mass., 2008.
- [9] Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J. and Chuang, I. *HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013*. HarvardX & MITx Working Paper No. 1. , 2014. Retrieved August 12, 2014

from SSRN:
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263

[10] Kling, J. R., Liebman, J. B. and Katz, L. F. Experimental analysis of neighborhood effects. *Econometrica*, 75, 1 (2007), 83-119.

[11] National Center for Education Statistics. *Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress: A Theoretical Foundation*. National Center for Education Statistics, Washington, D.C., 2012. Retrieved October 15, 2014 from NCES: http://nces.ed.gov/nationsreportcard/pdf/researchcenter/socioeconomic_factors.pdf

[12] Ruggles, S., Alexander, J. T., Genadek, K., Schroeder, M. B. and Sobek, M. Integrated Public Use Microdata Series: Version 5.0. (2010).

[13] U.S. Census Bureau. American Community Survey, 2012. (2013). Retrieved October 12, 2014, from <http://www.census.gov/acs/www/Downloads/questionnaires/2012/Quest12.pdf>

[14] U.S. Census Bureau. American FactFinder: 2008-2012 American Community Survey Five-Year Estimates, Selected Economic Characteristics. . (2013). Retrieved August 18, 2014, from the US Census Bureau: factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_DP03&prodType=table

[15] U.S. Census Bureau. Geographic Terms and Concepts - ZIP Code Tabulation Areas. 2014, October 10 (2010). Retrieved October 12 from U.S. Census Bureau: <https://www.census.gov/geo/reference/zctas.html>

[16] U.S. Census Bureau. *Design and Methodology: American Community Survey*. U.S. Government Printing Office, Washington, D.C., 2009. Retrieved August 14, 2014 from U.S. Census Bureau: http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

[17] Watson, T. Inequality and the measurement of residential segregation by income in American neighborhoods. *Rev. Income Wealth*, 55, 3 (2009), 820-844.