# Democratizing Education? Examining Access and Usage Patterns in Massive Open Online Courses

# Title: Democratizing Education? Examining Access and Usage Patterns in Massive Open Online Courses

**Authors:** John D. Hansen*[1], Justin Reich[2]

**Affiliations:**

[1]Harvard University

[2]Massachusetts Institute of Technology

*Correspondence to: john_hansen@mail.harvard.edu

**Abstract**: Massive Open Online Courses (MOOCs) are often characterized as remedies to educational disparities related to social class. Using data from 68 MOOCs offered by Harvard and MIT between 2012 and 2014, we find that course participants from the United States tend to live in more affluent and better-educated neighborhoods than the average U.S. resident. Among those who do register for courses, students with greater socioeconomic resources are more likely to earn a certificate. Furthermore, these differences in MOOC access and completion are larger for adolescents and young adults, the traditional ages where people find on-ramps into STEM coursework and careers. Our findings raise concerns that MOOCs and similar approaches to online learning can exacerbate rather than reduce disparities in educational outcomes related to socioeconomic status.

**One Sentence Summary:** In 68 massive open online courses (MOOCs) offered by Harvard and MIT, students with greater socioeconomic resources enrolled and earned certificates at higher rates.

**Main Text:** For nearly a century, technologists have promised that new broadcast media will bridge resource gaps between students in more and less privileged environments. "With radio the underprivileged school becomes the privileged" was the promise in the 1930s*(1)*; in the 1960's boosters declared that television would "make available to these young people instruction of a higher order than they might otherwise receive"*(2)*. In the first years of the 2010s, technologists have heralded the possibility that massive open online courses (MOOCs) can "democratize education" *(3-5)*. Previous generations of broadcast and interactive technologies—film, radio, television, personal computers, Internet access, and Web 2.0 platforms—have yet to fulfill the promise of educational parity*(6)*, and these new claims from MOOC advocates warrant empirical study. In this study, we take advantage of the data collected from MOOC students about their demographics and course performance—generally unavailable in studies of broadcast technologies—to present a portrait of registration and completion patterns in 68 courses offered by Harvard and MIT on the edX platform.

Our analytical framework is guided by Attewell's argument that the "digital divide," the gap in education technology opportunities between students from different backgrounds, is best understood as two divides: one of access and one of usage*(7)*. More and less affluent students not only have different levels of basic access to emerging technologies; they use them for different purposes with different levels of support from mentors. Historically, digital divides of usage have compounded digital divides of access. Surveys from the National Assessment of

Educational Progress in 1996 and 2011 show that students from schools serving mostly affluent students were more likely to use computers for simulations or modeling; by contrast, students from schools serving low-income students were more likely to use computers for drill and practice exercises*(8, 9)*. Comparable patterns have been found across the sciences and other subject areas when comparing schools with similar computer-student ratios serving students from different backgrounds*(10)*. Attewell found evidence of similar patterns of computer usage at home, where the academic benefits of home computers were greater for children from affluent families*(11)*.

These patterns extend into the era of free Web tools as well. Reich and colleagues examined the use of freely-available wikis—platforms for collaborative Web publishing—in U.S., K-12 schools in the late '00s*(12)*. He found that free wikis were more likely to be created in affluent schools, and in these schools wikis were more likely to be used to support collaborative problem solving and new media literacy. In schools serving low-income students, wikis were more likely to be used for teacher-centered content delivery. This research suggests a potential paradoxical effect of free online learning resources: they can disproportionately benefit the affluent, who have the social, financial, and technological capital to take advantage of new innovations, including those that are free.

The earliest research on MOOCs hints at similar kinds of patterns. The majority of registrants in MOOC courses already have a college or graduate degree, and some studies find a positive, substantively modest correlation between a student's level of education and course completion*(13-16)*. We build upon these studies with a much richer demographic portrait of students across a wider range of courses.

Socioeconomic status (SES) denotes one's social and financial resources, and it is typically viewed through a combination of measures*(17)*. In this study, we use three indicators for SES: parental educational attainment, neighborhood median income, and neighborhood average educational attainment. When signing up for edX, students are asked to provide their mailing address, and for American MOOC registrants, we can use this address to identify each student's census block group, a "neighborhood" of approximately 1,500 people for which we have census data about median income and educational attainment*(18)*. While more direct measures of family income or wealth are preferred, these neighborhood-level measures have proven useful in other studies *(19)*. We are particularly interested in adolescents age 13-17 for several reasons. First, these are the years that have traditionally been critical for students finding an on-ramp into postsecondary STEM education and careers. Also, MOOC advocates have identified K-12 students as a promising target population for MOOCs*(20, 21)*, and universities and MOOC platforms are increasingly targeting this population with their offerings*(22)*,Pragmatically, these students likely live at home with their parents, and our three measures should identify an individual's SES with greatest fidelity in this age range.

In the 2012-2014 academic years, Harvard and MIT offered 68 free courses and modules on the edX learning management system, attracting 1,028,269 unique participants (individuals who enter the courseware of one or more courses)*(16)*. Our study examines 164,198 unique participants from the U.S. who report an age between 13 and 69 and provide a mailing address we can match to a census block group, which is 57% of U.S. participants in this age range (Table

S1). Since many participants registered for multiple courses, these students account for over 200,000 participant-course observations. We compare the demographic characteristics of American MOOC participants to the U.S. population to better understand the digital divide of access. This comparison can be understood as a case-control study*(23)*, with edX enrollees as cases and a synthetic set of 1-1 matched controls by geographic area, assuming that controls are unlikely to be enrolled in edX given the large population size. We then examine how measures of SES predict course completion to understand the digital divide of usage.

We first describe differences in neighborhood characteristics between HarvardX and MITx participants and the U.S. population as a whole. Figure 1 shows that for all ages from 13 to 69, MOOC participants lived in neighborhoods that are more affluent and have higher average levels of educational attainment. We find that, on average, MOOC participants resided in neighborhoods where median household income is $69,641 dollars, which is $11,998 dollars above the neighborhood national average of $57,643 (Table S2). If we restrict our comparison to individuals age 13 to 17, the difference is $23,181 (Table S2). We find large differences in neighborhood educational attainment across all age groups as well.

We conduct a variety of sensitivity analyses, presented in the supplementary materials, suggesting that this finding is robust and persists at the individual level (Fig. S4). Specifically, we find: the positive relationship between neighborhood SES and MOOC participation persisted across courses and within states, counties, and census tracts (Table S6); survey respondents appeared similar to non-respondents with respect to our measures of SES (Tables S7-S8); alternative demographic datasets and neighborhood identification approaches produce similar estimates; and participants also tend to live in more densely populated neighborhoods (Tables S9-S10), suggesting that MOOCs do not disproportionately serve the geographically isolated.

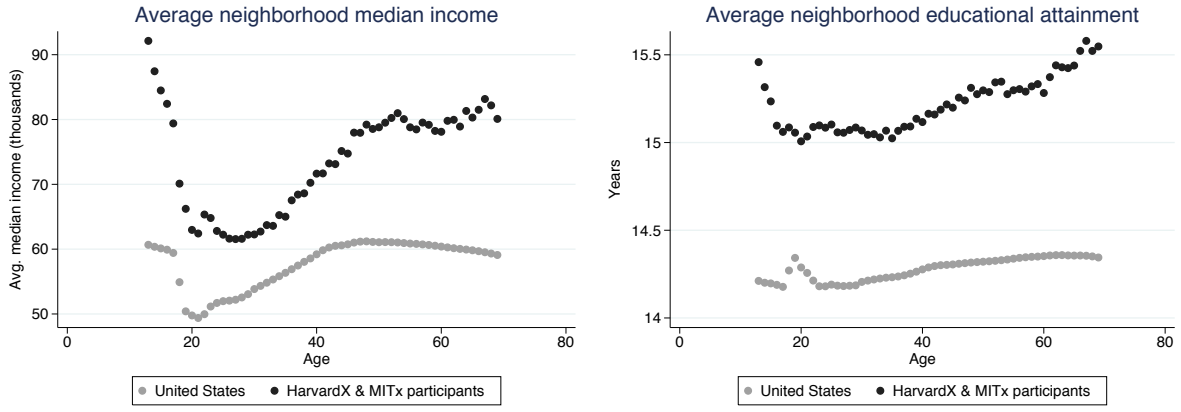


**Fig. 1.** Neighborhood income and educational attainment differences between MOOC participants and the U.S. population.

Predicting MOOC participation as a function of neighborhood SES makes these differences interpretable in terms of participation likelihood. Table 2 displays the results of logistic regression models, where the odds of participation are estimated in terms of a 1 standard deviation change in the predictor. Interpreting these results in dollars, we predict that an additional $20,000 in neighborhood median income increased the odds of participation by 27%. Each additional year of neighborhood-average educational attainment increased the odds of

participation by 69%. Among adolescents, the relationship between neighborhood SES and MOOC participation was even stronger.

| SES Variable | Age | Participation Odds ratio, +1 SD (standard error) | Certification Odds ratio, +1 SD (standard error) |
|---|---|---|---|
| Neighborhood Income SD = $30,536 | 13-69 | 1.44 (.003) | 1.06 (.014) |
| | 13-17 | 1.59 (.012) | 1.13 (.026) |
| Neighborhood Education SD = 1.27 years | 13-69 | 1.95 (.005) | 1.07 (.022) |
| | 13-17 | 2.09 (.024) | 1.32 (.049) |
| Parental Education SD = 2.92 years | 13-17 | 2.97 (.086) | 1.28 (.114) |

**Table 2.** Differences in MOOC participation and certification likelihood attributable to a one standard deviation increment in SES variables. An odds-ratio of 1 means equivalent odds. For age 13-69 regressions, the sample sizes are approximately 232 million for participation and 201k for certification. For age 13-17 regressions, the sample sizes are approximately 20.5 million for participation, 8,481 for neighborhood-SES certification models, and 2,112 for parental education certification models. See supplementary materials for model specification details..Robust standard errors clustered at the course level are used for certification models. All coefficients are statistically significant ($p < .01$).

Turning to the digital divide of usage, we found analogous patterns when examining the relationships between our measures and certificate attainment. Neighborhood- and individual-level SES measures were associated with higher rates of course completion, with larger magnitudes for younger participants. Examining the full age range of 13-69, we interpret the coefficients from Table 2 as modest in magnitude. Among the individuals who took the initiative to enroll and participate in a HarvardX course, neighborhood SES—like one's own educational attainment *(17)*—was a statistically significant but not substantively strong predictor of course completion on average. These relatively modest overall differences, however, mask important differences in attainment by SES for young people. For an adolescent participant whose most educated parent has a bachelor's degree, the odds of certification were approximately 1.75 times higher than an otherwise similar adolescent in the same course whose most educated parent has less than a bachelor's. Students from all backgrounds earn certificates in Harvard and MIT MOOCs, but especially among the young, high-SES students are more likely to earn a certificate.

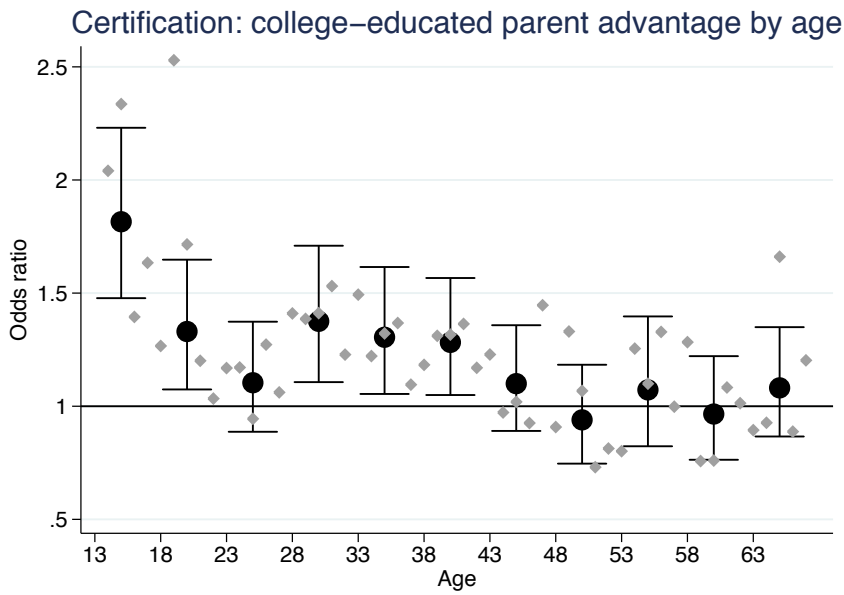


**Fig. 2.** Odds ratio of certificate-earning for participants with a college-educated parent compared to participants without one. Diamonds are estimated by means of a logistic regression model that include sets of binary indicators for age, course, enrollment mode, and the interaction of each age indicator with a binary indicator for college-educated parent. Circles with error bars are estimated in an analogous specification where age group indicators (13-17, 18-22, etc.) replace age indicators in the interaction. Error bars shows ±1 SE. Each point on the plot represents the multiplicative difference in the odds of certification among students of the same age whose parents have a bachelor's degree compared to those whose parents do not.

Overall, individuals living in high-SES neighborhoods in the United States were substantially more likely to participate in Harvard's and MIT's MOOCs, and, conditional on participation, high-SES students earned certificates at higher rates. These patterns were particularly strong among adolescents, precisely the age at which we hope students from low-income backgrounds can use education as a gateway to the middle class.

The rhetoric of "democratizing education" implies broad social benefits without precisely articulating how those benefits might be distributed. In Figure 3, we present two stylized representations of the effects of a technological innovation such as MOOCs on educational outcomes from students from different backgrounds. In the scenario we call "Closing Gaps," expanding access simultaneously benefits all students and ameliorates inequality. In the "Rising Tide, panel" all groups benefit from emerging technologies, but gaps in educational outcomes widen.

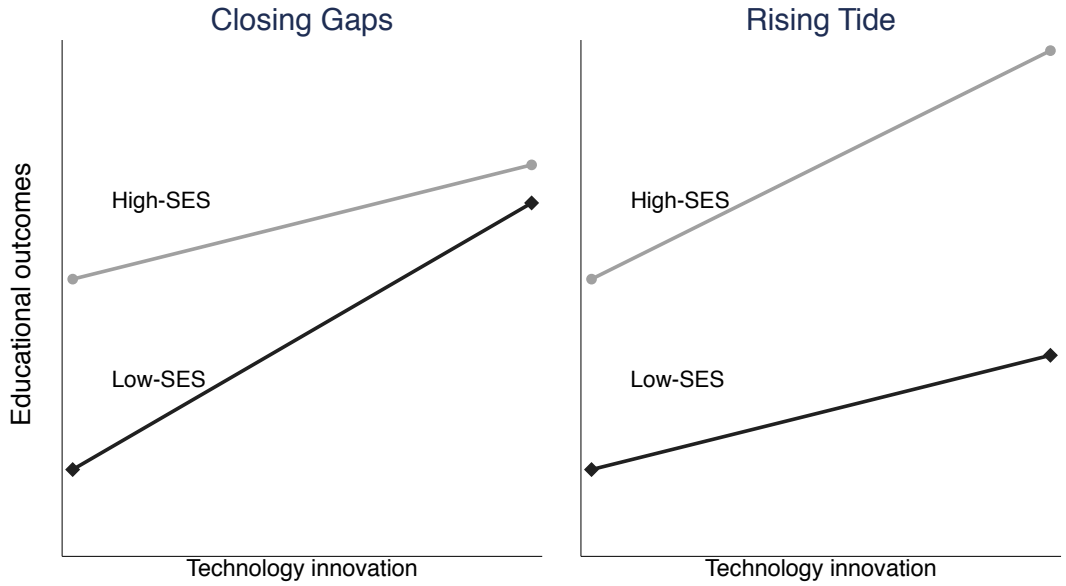


**Fig 3**: Two stylized representation of the hypothesized effects of a technological innovation on educational outcomes for students from high-SES and low-SES backgrounds.

Whether particular gaps will widen or close, for whom, and under what circumstances, are all questions worthy of further study as MOOCs and other new learning opportunities expand. The findings from this observational study appear more consistent with the "Rising Tides" than "Closing Gaps" scenario, but additional research will be necessary to identify causal effects on SES-education gaps. Despite early research that socially-advantaged children watched more Sesame Street and learned at least as much from watching *(24)*, later research found that it narrowed an SES-related gap in school readiness*(25)*.

MOOCs are one of many online learning opportunities, and our findings cannot be generalized to all open educational resources or education technologies. Nevertheless, our research on MOOCs—along with previous decades' research examining the access and usage patterns of emerging learning technologies—should provoke skepticism of lofty claims regarding democratization, level playing fields, and closing gaps that might accompany new genres of online learning, especially those targeted at younger learners. Freely-available learning technologies can offer broad social benefits, but educators and policymakers should not assume that the underserved or disadvantaged will be the chief beneficiaries. Closing gaps with digital learning resources requires targeting innovation towards the students most in need of additional support and opportunity.

**References**


1. L. Cuban, *Teachers and Machines: The Classroom Use of Technology Since 1920* (Teachers College Press, New York, 1986).

2. "*Teaching by Television,*" (Ford Foundation, , 1961).

3. R. Kanani, EdX CEO Anant Agarwal on the Future of Online Learning. *Forbes*., Jan 15, 2015 (2014).

4. D. Koller, MOOCs Can Be a Signifcant Factor in Opening Doors to Opportunity. *EdSurge*., Jan 15, 2015 (Dec. 31, 2013).

5. D. Faust, R. Reif, The Newest Revolution in Higher Ed. *Boston Globe*.(2013).

6. S. Reardon, in Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children, R. J. Murnane, G. Duncan, Eds. (Russell Sage Foundation Press, New York, 2011).

7. P. Attewell, Comment: The First and Second Digital Divides. *Sociology of Education*. **74**, 252-259 (2001).

8. H. Wenglinsky., "*Does it Compute? The Relationship Between Education Technology and Student Achievement in Mathematics,*" (Educational Testing Services, Princeton, NJ, 1998).

9. U. Boser., "*Are Schools Getting a Big Enough Bang for Their Education Technology Buck?*" (Center for American Progress, Washington, D.C., 2013).

10. M. Warschauer, M. Knobel, L. Stone, Technology and Equity in Schooling: Deconstructing the Digital Divide. *Educational Policy*. **18**, 562-588 (2004).

11. P. Attewell, J. Battle, Home Computers and School Performance. *Inf. Soc*. **15**(1999).

12. J. Reich, R. J. Murnane, J. B. Willett, The State of Wiki Usage in U.S. K–12 Schools. *Educational Researcher*. **41**, 7-15 (2012).

13. E. J. Emanuel, Online education: MOOCs taken by educated few. *Nature*. **503**, 342-342 (2013).

14. A. D. Ho *et al*., "*HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013,*" *Rep. No. HarvardX & MITx Working Paper No. 1*, 2014).

15. J. Reich, MOOC Completion and Retention in the Context of Student Intent. *EDUCAUSE Review Online*.(2014).

16. A. D. Ho *et al*., "*HarvardX and MITx: Two Years of Open Online Courses,*" *Rep. No. HarvardX Working Paper No. 10*, 2015).

17. 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).

18. J.D. Hansen, J. Reich, Socioeconomic Status and MOOC Enrollment: Enriching Demographic Information with External Datasets, (ACM, New York, NY, USA, 2015).

19. S. R. Sirin, Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*. **75**, 417-453 (2005).

20. C. E. Finn, MOOCs in Small Sizes Please. *Education Next*.(2012).

21. M. B. Horn, MOOCs for High School. *Education Next*. **14**, 82-83 (2014).

22. T. Lewin, Promising Full College Credit, Arizona State Offers Online Freshman Program. *New York Times*. **164**, A14-A14 (2015).

23. J. J. Schlesselman, P. D. Stolley, *Case Control Studies: Design, Conduct, Analysis* (Oxford University Press, New York, 1982).

24. T. D. Cook, "*Sesame Street*" *Revisited* (Russell Sage Foundation, New York, 1975).

25. M. S. Kearney, P. B. Levine, Early Childhood Education by MOOC: Lessons from Sesame Street. *NBER Working Papers*., 104 (2015).

26. Esri. *Updated Demographics*. (2014). Redlands, CA: Environmental Systems Research Institute http://doc.arcgis.com/en/esri-demographics/data/updated-demographics.htm

27. U.S. Census Bureau. American FactFinder: 2008-2012 American Community Survey Five-Year Estimates, Financial Characteristics. (2013). Retrieved August 18, 2014, from the US Census Bureau: http://factfinder2.census.gov/faces/nav/jsf/pages/download_center.xhtml

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

29. Esri. ArcGIS Desktop: Release 10.3. Redlands, CA: Environmental Systems Research Institute.

30. C.M. Hoxby, C. Avery, The Missing One-Offs: The Hidden Supply of High-Achieving, Low Income students. *NBER Working Papers*., 18586 (2012).

**Acknowledgments:** This work was funded in part by the Dean's Office of the Harvard Graduate School of Education. We are grateful to the HarvardX and MITx research communities for comments and support and to three anonymous reviewers for helpful feedback. Data on HarvardX and MITx students is available at from the Harvard Dataverse at http://dx.doi.org/10.7910/DVN/29779. These study files also include Stata code and log files for all analyses. Student level data is restricted to qualified researchers approved by the HarvardX research committee. Esri data is available for a fee from esri.com. The American Community Survey micro data is publicly available at ipums.org. The American Community survey zip code level data is available at http://factfinder2.census.gov/faces/nav/jsf/pages/download_center.xhtml

**Supplementary Materials:**
Materials and Methods
Figures S1-S5
Tables S1-S10
References (26-30)

# Supplementary Materials for

## Democratizing Education? Examining Access and Usage Patterns in Massive Open Online Courses

John D. Hansen,* Justin Reich

correspondence to: john_hansen@mail.harvard.edu

**This PDF file includes:**

Materials and Methods
Figs. S1 to S5
Tables S1 to S10

## Materials and Methods

Materials

*MOOC Participation Dataset*

This dataset includes observational and self-reported data for all HarvardX and MITx Massive Open Online Course (MOOC) participants in courses offered from 2012-2014 and hosted on the edX platform. This is the same dataset used for the joint HarvardX-MITx course report for 2012-2014 *(16)*. HarvardX and MITx are university initiatives supporting the offering of MOOCs on the edX platform. University faculty and staff create course content for HarvardX and MITx, and edX manages the platform, hosts the course content, and tracks user behavior in the courseware. From edX platform logs, we can determine whether a registrant enters the courseware (which we use to distinguish registrants from participants, including only the latter), and we can determine whether a participant earns a certificate.

Self-reported data comes from two sources: the edX site registration survey and course-specific pre-course surveys. Users submit the site registration survey when creating an edX account, and we use two key fields from the survey for our analyses: mailing address and year-of-birth. Since the site registration survey is completed only once, this information remains the same across courses for each individual. 95% of unique participants report a year of birth corresponding (approximately) to an age of 13-69 at the time of the course launch, and we discuss the quality of mailing address self-reports below.

Pre-course surveys were created by HarvardX and MITx researchers in tandem with course faculty. Upon enrolling in a course, registrants were asked to complete a survey, but no mechanism existed to enforce completion. Students could register for the course without taking the survey, and not all courses included one. Within our analytic sample, 51% completed the survey and another 17% submitted a partial response. The survey included approximately 20 items, including items about geographic location and an item asking students to self report their mother's and father's highest level of education. The only non-geographic pre-course survey item used in our analysis is parental education, and we have this item for approximately 32% of participants. Many items on pre-course surveys varied across courses in accordance with the interests of course faculty and researchers, and respondents were often assigned to one of several experimental versions of a course's survey in order to test the effects of survey length or the phrasing of a question. While each survey included a set of common items, systematic differences in survey composition randomly prevented some users from answering some questions, artificially deflating response rates for non-common items. For users who enrolled in multiple courses, survey responses for "course-invariant" user characteristics were imputed. For example, a user taking three courses who reported country of residence in only one pre-course survey was assumed to reside in the same country for all three courses.

*Esri 2013 Demographic Data*

Esri's 2013 demographic dataset draws primarily on the 2010 US Census and is annually updated. It includes a long list of variables available at various geographic levels, such as educational attainment among adults at the census block group level *(26)*. An important feature of the data is that the number of individuals in a census block group is disaggregated by age. That is, a separate variable exists for the number of individuals age 0, 1, 2, … 84. This allows us to reshape the dataset such that each unique age-block group combination is a separate row with

a frequency weight corresponding to the count of individuals of a given age living in the census block group.

*American Community Survey Data*

The US Census Bureau delivers the American Community Survey each year to several million American residents, and survey items include age, education, and income. We use data from the US Census Bureau's American Community Survey (ACS) in two forms. First, we use a version of the data available through the Minnesota Population Center's Integrated Public Use Microdata Series (IPUMS) website, where the ACS 5-year estimates of US population characteristics are organized at the person level. In Figure S4, we use this dataset to compare the distributions of parental education for individuals from 13 to 17 years of age. Second, we use a version of the data available through the US Census Bureau website *(27)*. In this case, the data are organized at the zip code level—or, more precisely, at the "zip code tabulation area" level *(28)*. 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, not the interests of demographic 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 neighborhood income are associated—as shown in Figure 1—failing to take into account differences in the age distributions between MOOC participants and the U.S. population would confound differences in age and neighborhood income.

*Geocoding and Final Sample Selection*

Our objective is unbiased estimates of neighborhood-level characteristics of HarvardX and MITx participants. Given our sample size—and the apparent similarity between individuals reporting a parsable address and those who do not (Table S8)—we prefer a relatively simple geocoding strategy that yields high-quality matches for 57% of participants compared to an approach that on average yields lower-quality matches for a higher proportion of participants. We present alternative approaches and demonstrate that they support our main findings.

First, the address parsing script takes the edX site registration mailing address field as an input, removes unwanted characters and phrases, and standardizes the format. For example, commas are removed, all characters are capitalized, and consecutive spaces are replaced with single spaces. Additional cleaning occurs intermittently (e.g. hyphens are not removed until later since some individuals report a nine-digit zip code containing a hyphen).

Before: 123 Pearl Street, San Francisco CA, 94123
After: 123 PEARL STREET SAN FRANCISCO CA 94123

Second, the script parses zip codes from the field, confirms that parsed zip codes match a real us zip code, and merges in state name (abbreviated) and city primarily affiliated with that zip code using a lookup table. Matched zip codes are removed and stored in a separate field.

Before: 123 PEARL STREET SAN FRANCISCO CA 94123
After: 123 PEARL STREET SAN FRANCISCO CA

Third, the script searches for the abbreviation or name of a US state at the end of the field. If the state corresponding to the zip code identified above is the same state found at the end of the address field, we consider them "matched." We remove the state abbreviation and proceed as if we have correctly identified the participant's correct zip code, city, and state. All other observations are dropped.

Before: 123 PEARL STREET SAN FRANCISCO CA
After: 123 PEARL STREET SAN FRANCISCO

Next, for matched observations, we remove the word immediately to the left of the location, which is typically where a city name should be, and the rest of the field is treated as the street address. Some cities are multiple words long, and in these cases only the last word is removed from the address field.

Before: 123 PEARL STREET SAN FRANCISCO
After: 123 PEARL STREET SAN

Including the "SAN" from "SAN FRANCISCO" is not problematic for the geocoding software we use. The final components are as follows, where the city and state fields are derived from the zip code:

Address: 123 PEARL STREET SAN
City: SAN FRANCISCO
State: CA
Zip Code: 94123

Of the 288,505 unique participants age 13-69 whom we believe to be US residents, 164,198 submitted an address that we could parse and match to a zip code using the criteria above (Table S1). We geocoded these addresses using ArcMap's USA Geocoding Service *(29)*. Using the software's default settings for determining match quality, all but six individuals were successfully matched to at least a zip code (the six individuals who were not matched were dropped from the sample). Overall, 88.5% of addresses were matched to a census block group, and another 11.5% were matched to a zip code. In the case of the zip code match, we estimate census block group characteristics using the characteristics of the census block group where the centroid of the 5-digit zip code falls.

## Methods

### *Main Findings*

The the difference in average neighborhood median income for MOOC participants and the U.S. population is approximately $12,000 (Table S2). We also estimate the average difference in neighborhood median income between MOOC participants and the US population adjusting for the cross-sectional relationship between age and neighborhood income. We make this adjustment because age is associated with neighborhood income and likelihood of MOOC participation (as shown in Fig. 1), and we want to avoid confounding a difference in neighborhood income with a difference in age distributions between MOOC participants and the general population. Including an indicator variable for each age, represented by the $\delta_j$ parameter below, allows us to identify

the average difference in neighborhood income between individuals of the same age. After appending a dataset representative of the U.S. population to our MOOC participant dataset—and keeping only one observation per participant, so individuals participating in multiple courses are weighted equally to individuals participating in only one course—we estimate the following model using ordinary least squares (OLS) regression:

$$BG\_INC_{ij} = \alpha + \beta\left(EDX_{ij}\right) + \delta_j + \epsilon_{ij}, \quad (1)$$

where $EDX_{ij}$ is a dichotomous variable equal to one for each unique participant and zero otherwise, $\alpha$ is the U.S. mean in models where age indicators are omitted, $\epsilon_{ij}$ is a stochastic error term, and $\beta$ is the parameter of interest: the average difference in neighborhood median household income between MOOC participants and individuals of the same age in the U.S. population. The models below restrict both groups to age 13-69. Table S2 shows that the baseline difference in means between these two groups is $11,998. Including $\delta_j$ yields an estimate of $13,508; this is essentially a weighted average of the differences between the U.S. and MOOC participant dots in the left panel of Figure. 1. A difference of $13,508 may seem smaller than one would expect based on a visual estimate from Figure 1, and this disparity is attributable to the high density of participants in their twenties, where the distance between the groups' dots is relatively small. The density of the age distribution of participants for age 13-69 is shown in Figure S1.

The statistical models for neighborhood education are analogous. Our measure of neighborhood educational attainment transforms the original variable from the Esri dataset, which is educational attainment in terms of highest degree attained for individuals of at least twenty-five years of age. We use the following convention to convert educational attainment from highest degree awarded to years:

Less than High School: 9 years
High School: 13 years
Associate's Degree: 15 years
Bachelor's Degree: 17 years
Graduate Degree: 19 years

We treat parental educational as a categorical variable in some cases (as shown in Fig. 2), but in other cases we transform it using the same convention above (as shown in Table 2). When fitting the model above for participants age 13-17, we restrict the sample to observations in this age range and omit the age indicators ($\delta_j$). We exclude 18 year-olds because we observe a sharp drop in average neighborhood income for 18 year-olds compared to 17 year-olds, which is presumably attributable to 18 year-olds moving out of their parents' house at a higher rate than 17 year-olds. We observe this within the U.S. and among MOOC participants, though the absolute difference is greater for participants. Restricting the comparison to 13-17 year-olds arguably provides a cleaner comparison of household SES levels during adolescence. Table S2 presents analogous estimates where individuals participating in multiple courses are counted

multiple times. All coefficients are higher in this specification, which is what one would expect if MOOC usage levels tended to rise with SES.

We estimate an analogous statistical model separately for each course as well. In this case, every participant enrolled in the course is included in the regression, unlike the samples above where a unique participant could only be counted once. Additionally, fitting separate regressions by subsample allows the $\delta_j$ parameters to vary across courses, which will reduces bias if the age distribution of participants varies by course. These results are available in the analysis log file (see "Acknowledgments" section for details), and show that our main findings about MOOC participation and neighborhood SES hold across courses and ages, with no notable exceptions.

In the "Participation" column of Table 1, we present parameter estimates from logistic regression models estimating the likelihood of MOOC participation as a function of SES. These coefficients answer the question: relatively how much more likely to participate is an individual with SES level *x* compared to an otherwise similar individual with SES level *x*?* Specifically, in the case of block group median income, we estimate:

$$\log\left(\frac{P(EDX_{ij}=1 \mid X_{ij})}{1-P(EDX_{ij}=1 \mid X_{ij})}\right) = \alpha + \beta(BG\_INC_{ij}) + \delta_j, \quad (2)$$

where the parameter of interest is $\beta$, the difference in the log-odds of participation attributable to neighborhood income among individuals of the same age. Exponentiating $\beta$, as presented in Table S3, displays the difference on an odds-ratio scale. As before, $\delta_j$ is a vector of dichotomous indicator variables for age and $\alpha$ is a trivial constant. Note that merging the MOOC participant dataset with an external dataset trivially biases downward our estimates of the likelihood of MOOC participation since MOOC participants are essentially double-counted (theoretically, they appear in the U.S. dataset as well as the MOOC dataset).

In Table 1, we present coefficients estimated from standardized SES variables. For observation *i,* we z-score (standardize) variable *x* by subtracting the U.S. population mean, $\mu_x$, and dividing by the population standard deviation, $\sigma_x$. We compute $\mu_x$ and $\sigma_x$ for the U.S. using the previously discussed frequency weights for the Esri neighborhood variables and the ACS-provided probability weights for parental education. There is far more within-age variability in our neighborhood SES variables than between-group variability, so computing within-age standardized dispersions from the age-conditional mean makes no substantive difference. We choose to standardize using the unconditional mean and standard deviation in order to maintain the ability to make statements about the relative difference in SES distributions by age on a constant dollar scale.

In the "Certification" column of Table 1, we present parameter estimates from logistic regression models estimating MOOC certification as a function of SES. These coefficients answer the question: relatively how much more likely to earn a certificate is an individual with SES level *x* compared to an otherwise similar individual with SES level *x*?* Specifically, in the case of block group median income, we estimate:

$$\log\left(\frac{P(CERT_{ijkl}=1 \mid X_{ijkl})}{1-P(CERT_{ijkl}=1 \mid X_{ijkl})}\right) = \alpha + \beta(BG\_INC_{ijkl}) + \delta_j + \omega_k + \lambda_l, \quad (3)$$

where the coefficient of interest is $\beta$, the estimated average difference in the log-odds of certification attributable to a one-thousand dollar difference in SES for same-aged participants taking the same course in the same mode. Exponentiating $\beta$, as presented in Table S4, presents the difference in likelihood on an odds-ratio scale. As before, $\delta_j$ is a vector of dichotomous indicator variables for age and $\alpha$ is a trivial constant. A vector of indicator variables for course is represented by $\omega_k$, and $\lambda_l$ is a vector of indicator variables for mode. The possible modes of course-taking are Honor, Verified, Audit, and Missing. Students in the Verified mode voluntarily pay between $25 and $250 to have their identity verified. (We also fit models without the $\lambda_l$ indicators and discuss the results below.)

Figure 2 compares by age the odds ratio of certificate-earning for participants with a college-educated parent compared to participants without one. It illustrates how the association between parental educational attainment and MOOC certification depends on the age of the participant. The figure overlays the results from two models. The first, whose results are shown by the diamonds, is:

$$\log\left(\frac{P(CERT_{ijkl}=1 \mid X_{ijkl})}{1-P(CERT_{ijkl}=1 \mid X_{ijkl})}\right) = \alpha + \delta_j + \omega_k + \lambda_l + \sum_{j=14}^{69} \beta_j (PBA_{ijkl} * \delta_j), \quad (4)$$

where the coefficients of interest are the vector $\beta_j$ $(\beta_{14}, \beta_{15}, \ldots, \beta_{69})$. As before, $\delta_k$ is a full set of indicators for age, $\omega_k$ is a full set of indicators for course, and $\lambda_l$ is a full set of indicators for mode. Each $\beta_j$ coefficient captures the estimated average difference in the log-odds of certification attributable to having a parent with at least a bachelor's degree for individuals of a given age (14-69) taking the same course in the same mode. We exclude thirteen year-olds from this model because only one thirteen year-old without a college-educated parent has completed a certificate (at least among the thirteen year-olds for whom we have data on parental education), which leads to an imprecise and anomalously high estimate. This specification is useful because it places no constraints on the relationship between age and the predictive effect of having a four-year college-educated parent, showing that our claim that parental education is more strongly related to certification for younger participants will be robust to model selection. The diamonds plotted in Figure 2 are exponentiated $\beta_j$ coefficients. The second model, which corresponds to the filled circles and error bars, is:

$$\log\left(\frac{P(CERT_{ijkl}=1 \mid X_{ijkl})}{1-P(CERT_{ijkl}=1 \mid X_{ijkl})}\right) = \alpha + \delta_j + \omega_k + \lambda_l + \sum_{g=1}^{11} \beta_g (PBA_{ijkl} * AG_g), \quad (5)$$

where the coefficients are analogous to equation (4), except participants are binned into eleven age groups in five-year increments, ranging from 13-17 to 63-68. Huber-White standard errors clustered at the course level are computed, and the error bars show one standard error above and below the point estimate.

*Main Findings: Alternative Specifications*

We also fit the main logistic regression models without controlling for mode (i.e. omitting $\lambda_l$). One could argue that these estimates are the most germane to our central claims, since controlling for mode will attenuate the relationship between SES and certification if higher-SES individuals are more likely to pursue an ID-verified certificate, and ID-verified participants complete certificates at a higher rate. This is borne out in the data, as shown in Table

S4, since the coefficient estimates are higher across the board in these specifications. The counter-argument is that ID-verified usage of MOOCs is sufficiently qualitatively different as to constitute an apples-to-oranges comparison. In other words, once we introduce ID-verification, a paid service, we are no longer comparing student usage of free courseware. We prefer the main specifications (with the $\lambda_l$), as these models compare individuals using the MOOC in a similar way, but, ultimately, omitting the $\lambda_l$ has little impact on the coefficients.

On page 4, we interpret coefficients from Table 2 in a way that suggests that the rate of MOOC participation is a monotonically increasing function of SES. Such a model not only entails that middle-SES individuals are more likely to participate than low-SES individuals, but also that high-SES individuals are more likely to participate than middle-SES individuals. This is a key point, substantively and statistically. Statistically, the logistic regression models from Table 2 constrain the relationship between neighborhood SES and the rate of MOOC participation to be a (positive or negative) monotonic function. Substantively, one might anticipate that low-SES individuals are unlikely to participate, but that little difference in participation rates exist between individuals from middle- and high-SES neighborhoods. To test this hypothesis, we fit a model with a less constrained functional form for the relationship between neighborhood SES and the likelihood of MOOC participation. We bin individuals by neighborhood SES and in the regression replace the continuous variables (i.e. *BG_EDU*) with a set of dichotomous indicator variables for the bins (e.g. *BIN_11* is equal to one for individuals for whom $10.5 \leq$ *CB_EDU* $< 11.5$ and zero otherwise). For *BG_INC,* we create bins from \$10,000 to \$210,000 by \$20,000 increments, assigning individuals to the nearest bin (Table S5). For *CB_EDU*, we bin from 9 to 19 by one year increments, though we use *BIN_11* as the reference category because few individuals—and no MOOC participants age 13-17—reside in neighborhoods for bins below the *BIN_11* level. Specifically, in the case of neighborhood education, we estimate:

$$\log\left(\frac{P(EDX_{ij}=1 \mid X_{ij})}{1-P(EDX_{ij}=1 \mid X_{ij})}\right) = \alpha + \delta_j + \sum_{k=9}^{19} \beta_k \left(BIN\_K_{ij}\right), \quad (6)$$

where the coefficients of interest are the vector $\beta_k$, as these coefficients characterize the MOOC participation rate among U.S. residents conditional on age and neighborhood SES. Table S5 presents estimates where the indicator for *BIN_11* is the omitted baseline group, and therefore each coefficient is the difference in the log-odds of participating in a MOOC for individuals in their own bin compared to individuals whose average neighborhood educational attainment is around 11 years. Unlike previous tables of results from logistic regressions, we present coefficient estimates on a logit scale (rather than an odds-ratio scale). The logit scale is desirable because the magnitudes of the differences between coefficients are on a linear scale. Therefore, if the difference between $\beta_{12}$ and $\beta_{13}$ is greater than the difference between $\beta_{13}$ and $\beta_{14}$ on a logit scale, one can more credibly claim that the former difference is larger than the latter. Overall, we find that the MOOC participation rate rises throughout the majority of the neighborhood education and neighborhood income distributions. Although it is not clear that the trend holds at the extremes, the trend does appear sufficiently strong to support the generalization that MOOC participation rate is an increasing function of neighborhood SES. Figures S2 and S3 display the results graphically, where one can interpret the plotted log-odds as local standardized estimates of the gap between the two distributions.

*Robustness of Findings*

Between-group relationships among variables do not necessarily persist within groups, so one could observe higher levels of neighborhood income for MOOC participants compared to the U.S. even though no such income difference exists at the household level. In terms of SES, our individual-level SES measure, parental educational attainment, allows us to address this concern directly for younger participants, where parental educational attainment is a well-established a component of SES *(19)*. For older participants, though, it is less clear whether parental education or neighborhood characteristics are better measures of an individual's SES. Evidence on one's own education has been discussed in previous work *(14, 16)*, and taking this evidence into account, we feel justified making general claims about individual SES and MOOCs since individual education, parental education, and neighborhood characteristics all point in the same direction with respect to participation and certification.

We report that "the positive relationship between neighborhood-level SES measures and MOOC participation persisted…within states, counties, and census tracts." Specifically, we regressed each standardized neighborhood SES variable on a set of binary age indicator variables ($\delta_j$), a set of binary indicators for one's state, country, or census tract ($\eta_k$), and a binary indicator equal to one for MOOC registrants and zero otherwise ($EDX_{ijk}$). The model is:

$$y_{ijk} = \alpha + \beta\left(EDX_{ijk}\right) + \delta_j + \eta_k + \epsilon_{ijk}, \tag{7}$$

where $y_{ijk}$ is the standardized neighborhood SES variable of participant $i$ of age $j$ living in geographic area $k$ (where the region is state, county, or census tract). The coefficient of interest is $\beta$, the difference in SES between edX participants and same-aged individuals living in the same state (or county or census tract). As shown below, $\beta$ remains positive and statistically significant in all specifications. To interpret the leftmost coefficient estimate of Table S6, we estimate that on average MOOC participants live in neighborhoods where neighborhood median income is $10,480 dollars higher than non-participants of the same age living in the same state.

The only household-level (rather than neighborhood-level) available to us is parental educational attainment. However, we only have this information for participants who submitted a pre-course survey. While we can identify neighborhood-level SES variables for more than half of participants, we can only observe parental education for 32% of them (33% of 13-17 year-olds). In this case, we use data from the American Community Survey to compare MOOC participants to the general US population. We restrict this comparison to 13-17 year-olds, since the ACS identifies parental educational for individuals residing with their parents. Moving away from home is more common beyond this age range, and a non-random pattern of missing data would likely yield biased estimates for older individuals.

Survey respondents may differ from non-respondents with respect to SES, so our sample of MOOC participants may not be representative of all MOOC participants. Since we have more complete data for neighborhood SES than parental education, we can exploit the positive relationship between neighborhood SES and parental education to test the null hypothesis that, among participants reporting a parsable address, neighborhood median income is the same for individuals who did and did not respond to the parental education survey item. In order to take into account the relationship between neighborhood median income and age, we test this hypothesis by regressing neighborhood SES variables on age dummies ($\delta_j$) and an indicator variable equal to 1 for registrants not reporting parental education. In this case, we compare one subset of MOOC participants to another subset of participants, rather than comparing MOOC

participants to the U.S. population. Specifically, we fit the following model:

$$BG\_INC_{ij} = \alpha + \beta\left(MISSING\_PED_{ij}\right) + \delta_j + \epsilon_{ij}, \quad (8)$$

where $\beta$ is our coefficient of interest: the average difference in neighborhood median income between participants reporting parental education and participants not reporting parental education. We fit this model for the entire population of participants and restricted to participants age 13-17 (in both cases, we retain the age dummies). While we find statistically significant evidence that individuals reporting parental education live in lower SES neighborhoods on average (Table S7), the differences (approximately $1,085 and .056 years of education) are substantively small. Overall, we find much higher levels of parental educational for 13-17 year-old MOOC participants compared to the U.S. (Fig. S4, Table S3), and we find that 13-17 year-olds reporting parental education on a survey live in neighborhoods with similar income and educational attainment levels as 13-17 year-olds not reporting parental education on a survey. This suggests that SES—not only neighborhood SES—levels are higher for young MOOC participants compared to the U.S.

One hypothetical scenario in which our approach to geocoding participants by mailing address could introduce bias is if high-SES individuals are somehow easier to geocode than low-SES individuals. In this case, we would have more success identifying neighborhoods for high-SES participants than low-SES participants, which would explain the difference in our neighborhood SES estimates. To address this concern, we use data available from the American Community Survey, which estimates median income data by zip code. Parsing a mailing address for a zip code is straightforward, so concerns about methods of parsing and matching are less of a concern. The 2012 median zip code income estimate for the US according to the ACS is $53,046 *(27)*. 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). The mean of median zip code household income observed in the MOOC participant sample of unique registrants is $68,119. The difference between unique participants and the U.S. population is $11,587, or .51 standard deviations. Reassuringly, this approach yields very similar estimates for differences in neighborhood median income as we find using census block groups.

A possible concern is that low-SES students are less likely to divulge *any* geographic information about themselves. Thus far, we have shown that our main findings are robust to choice of dataset, definition of neighborhood, and choice of self-reported item. Since it is theoretically possible that low-SES participants are less likely to divulge any geographic information at all about themselves—be that parent education, zip code, or mailing address—we compare address-matched and non-matched participants using IP address data, which is purely observational and available for 99% of participants. While this may seem like the most appealing geolocation method to use overall, a weakness of IP address data is its lower reliability and precision compared to address-based geolocation. Furthermore, even if IP address were perfectly accurate, it would problematic that if a student connects from school or work, his or her IP-identified location would differ from his or her home address. Nevertheless, the IP address geolocation offers an opportunity to test the generalizability of our estimates. We find that higher SES participants appear less likely to self-report address information we can link to a neighborhood. Using the coordinates identified by each participants' most commonly occurring

IP address, we match participants to the same Esri dataset as before and extract the same neighborhood variables, which we call *BG_INC** and *BG_EDU**. Next, we create an indicator variable, *ADD_MATCH* equal to 1 for participants whose location we were able to identify through mailing addresses parsing and zero otherwise. Since variation in neighborhood SES is related to age and likelihood of submitting a parsable mailing address, we include a full set of indicator variables for age and restrict our sample to individuals reporting an age between 13 and 69. The following regression compares the IP address-estimated, neighborhood SES levels of same-aged participants who did and did not report a parsable mailing address. In the case of *BG_INC**, we estimate the following equation:

$$BG\_INC_{ij} = \alpha + \beta\left(MISSING\_PED_{ij}\right) + \delta_j + \epsilon_{ij}, \quad (9)$$

where the coefficient of interest is $\beta$, the average difference in neighborhood median income when we use IP address to identify neighborhoods. The $\delta_j$ constrains the coefficient to estimate within-age differences in neighborhood median income. Additionally, we fit models including a full set of indicator variables for U.S. states where we exclude individuals who are geo-located to a different state by IP address compared to mailing address. On average we estimate that neighborhood SES levels appear to be lower for participants reporting a parsable mailing address (Table S8). This suggests that our main findings may, on average, slightly underestimate the true neighborhood SES difference between the typical MOOC participant and the U.S. population. For adolescents, Table S8 shows that our main findings may overestimate neighborhood SES for the typical adolescent MOOC participant, but these estimated differences are relatively small, sensitive to choice of neighborhood SES variable, and sensitive to model specification. Overall, major neighborhood SES differences between the address-matched and non-matched participants seem unlikely.

Finally, one could imagine that participants are less socioeconomically disadvantaged than they are geographically disadvantaged with respect to access to high-quality educational opportunities. There are considerable differences in college-going behavior between low-income, high-achieving students from large metropolitan areas compared to otherwise similar students attending schools in smaller districts with lower population densities *(30)*. To test the hypothesis that MOOCs are especially attractive to students living in less densely population areas, using the main analytic sample, we regress the natural logarithm of census block group population density on a full set of indicator variables for age and an indicator variable equal to 1 for MOOC participants whose mailing address was matched:

$$LN\_PD_{ijk} = \alpha + \beta\left(EDX_{ijk}\right) + \delta_j + \eta_k + \epsilon_{ijk}, \quad (10)$$

where the coefficient of interest is $\beta$, which can be interpreted as the approximate percentage difference in neighborhood population density for MOOC participants compared to the U.S. population. As before, $\delta_j$ constrains the coefficient to estimate within-age differences in neighborhood population density. We include separate estimates for age 13-17 and age 13-69, and we fit models with and without state indicators ($\eta_k$). Table S9 shows that estimates range from 14% to 60%, and all are positive and statistically different from zero, which means that MOOC participants tend to live in more densely populated neighborhoods than the typical U.S. resident. Figure S5 plots the cumulative distribution functions of population density for the U.S.

and MOOC participants age 13-17. The gap between the red and blue lines throughout the first five deciles shows that the teenagers living in the less densely populated half of neighborhoods are underrepresented in MOOCs.

To further examine the relationships among population density, neighborhood SES characteristics, and MOOC participation, we fit a set of logistic regression models to estimate the likelihood of participation as a function of multiple variables. In these models, we estimate the association between MOOC participation and neighborhood population density controlling for neighborhood measures of SES. Where neighborhood median income is controlled, for example, we estimate the following equation:

$$\log\left(\frac{P(EDX_{ijk}=1 \mid X_{ijk})}{1-P(EDX_{ijk}=1 \mid X_{ijk})}\right) = \alpha + \beta(LN\_PD_{ijk}) + \gamma(BG\_INC_{ijk}) + \delta_j + \eta_k, \quad (11)$$

where the coefficient of interest is $\beta$, the average difference in the log-odds of MOOC participation attributable to a one-unit increment in the natural logarithm of neighborhood population density, controlling for neighborhood SES. As before, $\delta_j$ constrains the coefficient to estimate within-age differences in neighborhood population density. We include separate estimates for age 13-17 and age 13-69, and we fit models with and without state indicators ($\eta_k$), which account for between-state differences in the likelihood of MOOC participation. Table S10 shows the results. In all specifications, $\beta$ is positive and statistically different from zero ($p <$ .001). On average, individuals living in more densely populated neighborhoods are more likely to participate in MOOCs. Our findings contradict the possibility that MOOCs are disproportionately serving the geographically underserved by attracting students living in less dense populated areas.

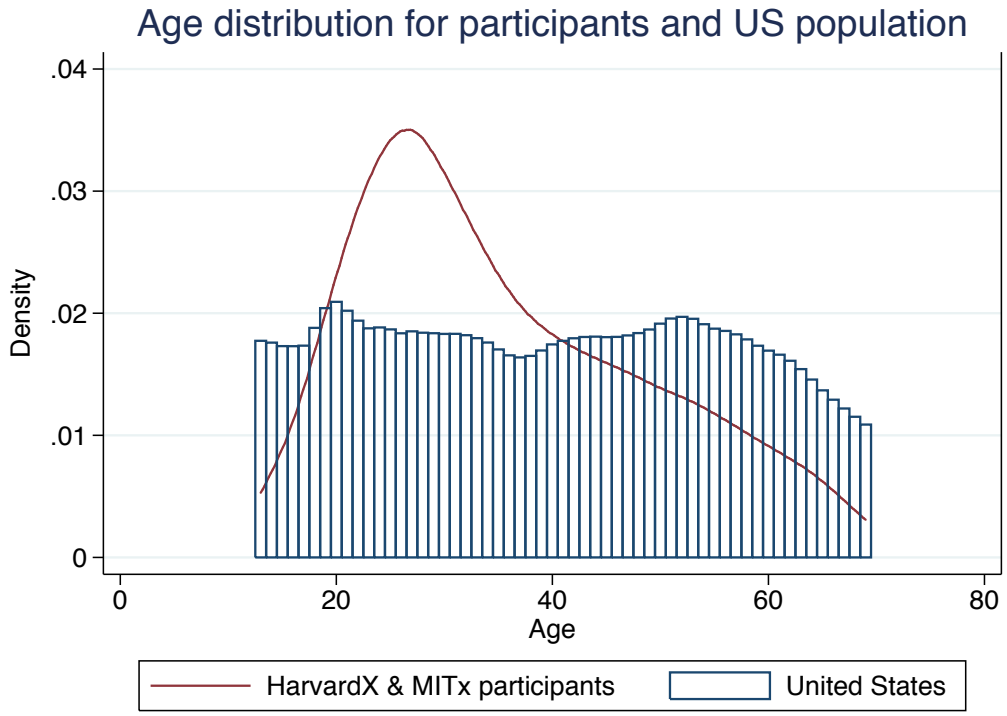


**Fig. S1.**

Age distributions for MOOC participants and the U.S. The red line is a smoothed kernel density function.

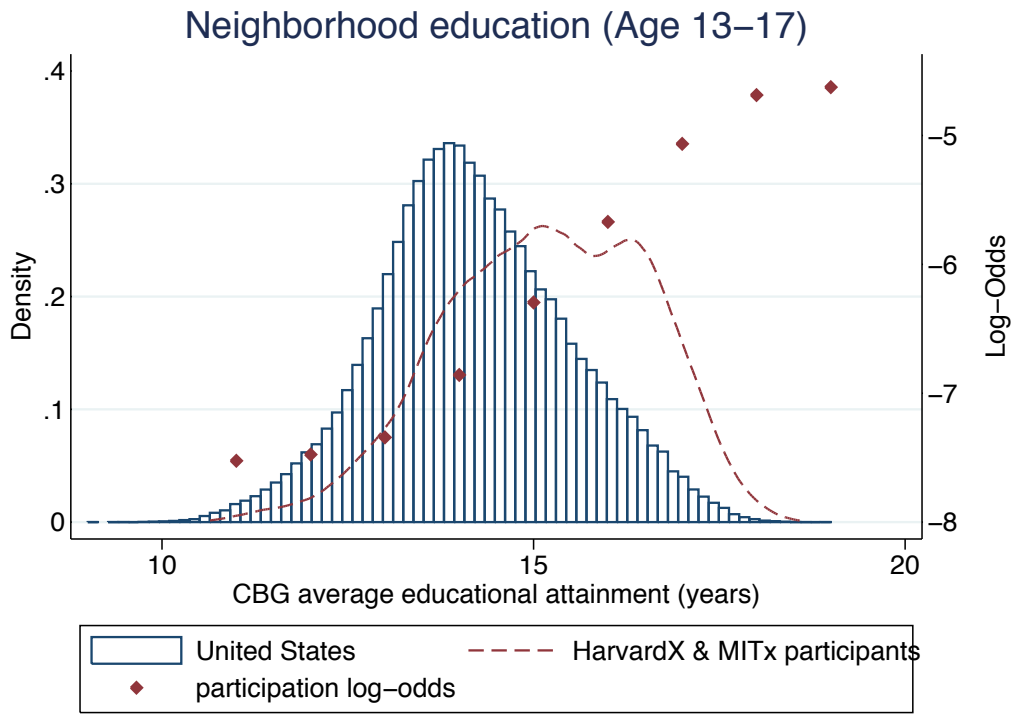


**Fig. S2**

Neighborhood educational attainment distribution (scaled in years) for MOOC participants and the U.S. and log-odds of MOOC participation by attainment bin.

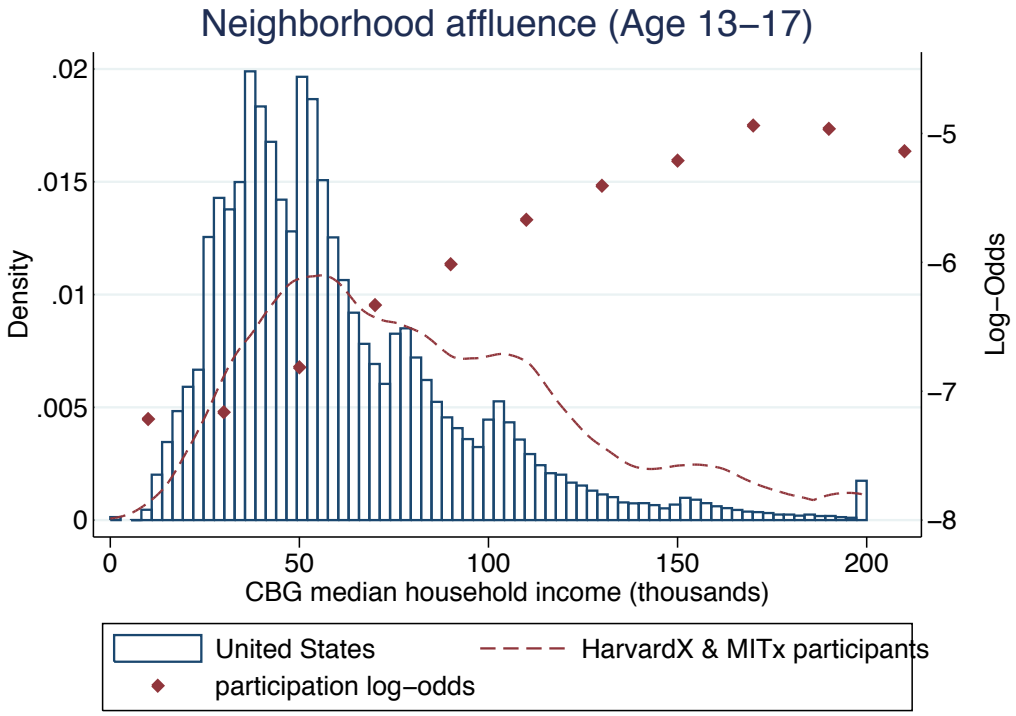


**Fig. S3**

Neighborhood median household income distribution for MOOC participants and the U.S. and log-odds of MOOC participation by income bin.

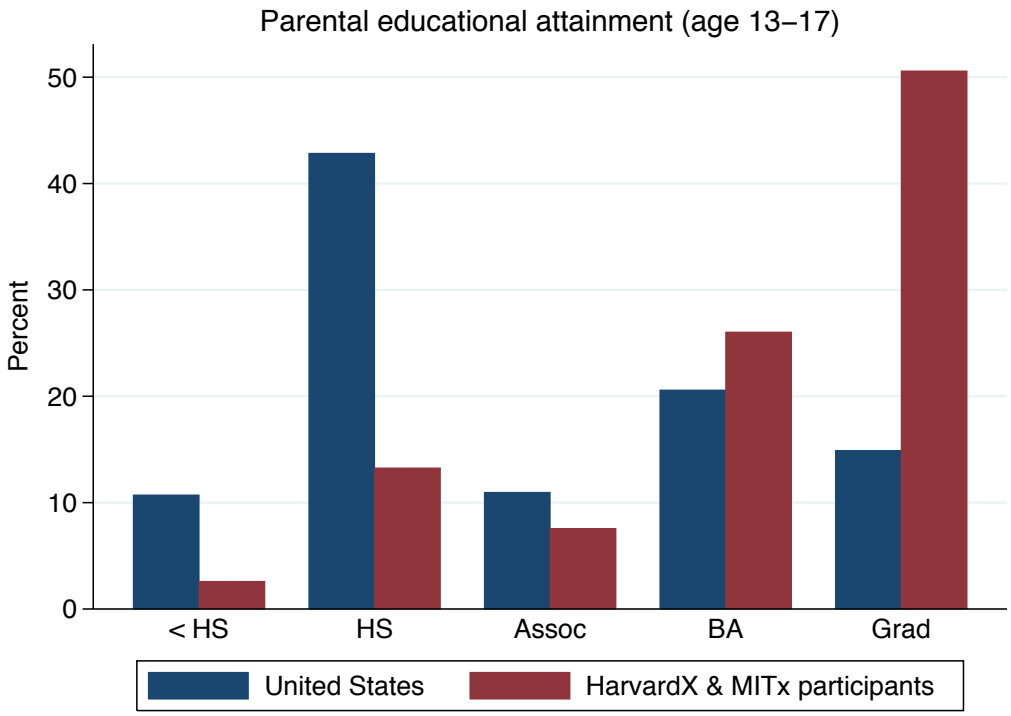


**Fig. S4**

Parental educational attainment for 13-17 year-old MOOC participants compared to the U.S. Estimates for the U.S. obtained from the American Community Survey (2013, 5-year estimates).

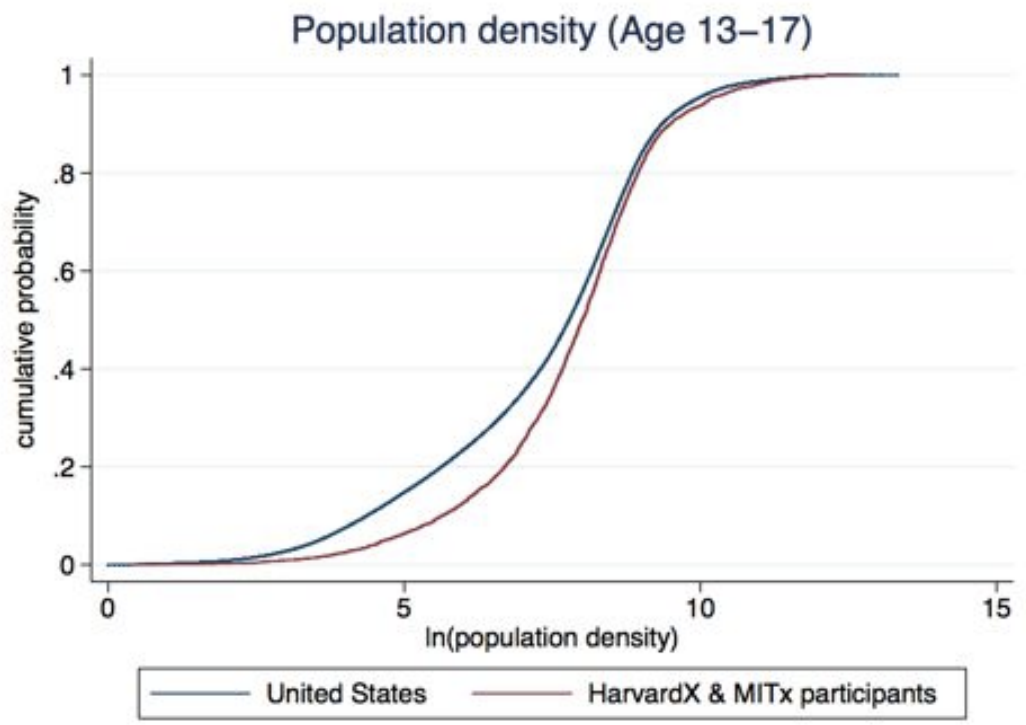


**Fig. S5**

Neighborhood population density for 13-17 MOOC participants compared to the U.S. Cumulative probability distribution functions shown.

| | All participants | | Unique participants | | | Unique & matched | |
|---|---|---|---|---|---|---|---|
| | Total | Cert. | Count | Med. age | Par. ed. | BG inc. | BG edu. |
| All | 1,478,703 | 6.5% | 912,781 | 26 | 15.95 | - | - |
| U.S. | 441,375 | 5.5% | 288,505 | 30 | 16.16 | - | - |
| Address matched | 248,514 | 6.6% | 164,198 | 32 | 16.06 | $69,641 | 15.15 |

**Table S1.**

MOOC participants age 13-69, BG Inc is the mean of census block group median income. BG Ed is the mean years of education for individuals age 25 and older in a census block group. The mean of parental education is reported in years of education, which is estimated from self-reports of degree attainment. The identification of U.S. in the second row uses a combination of IP address geolocation and self-reports.

Participation by Age and SES: Unique Participants

| | Neighborhood income | | | Neighborhood educational attainment | | |
|---|---|---|---|---|---|---|
| Unique Part. | Unadjusted | Within Age | Age 13-17 | Unadjusted | Within Age | Age 13-17 |
| *EDX* | 11998.5*** | 13507.9*** | 23180.9*** | 0.874*** | 0.898*** | 0.973*** |
| | (75.40) | (74.80) | (372.5) | (0.003) | (0.003) | (0.015) |
| Obs. | 232396969 | 232396969 | 20277134 | 232393719 | 232393719 | 20277121 |

Participation by Age and SES: Participant-weighted

| | Unadjusted | Within Age | Age 13-17 | Unadjusted | Within Age | Age 13-17 |
|---|---|---|---|---|---|---|
| *EDX* | 12177.8*** | 13611.5*** | 23278.8*** | 0.888*** | 0.911*** | 0.992*** |
| | (61.30) | (60.82) | (301.1) | (0.003) | (0.003) | (0.012) |
| Obs. | 232481285 | 232481285 | 20281177 | 232478013 | 232478013 | 20281164 |

**Table S2.**

Ordinary Least Square (OLS) regression comparisons of neighborhood SES variables for MOOC participants compared to the U.S. population. In the neighborhood income model where we do not adjust for age, the intercept is $57,643, the U.S. mean. For age 13-17, the intercept is $60,103. Estimates use unstandardized SES variables and include only unique participants in the top panel. Standard errors in parentheses.

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

Participation by Unstandardized SES variables

| | Neighborhood BG Median Inc. | | Neighborhood BG Education | | Parental Educ. |
|---|---|---|---|---|---|
| | All Ages | 13-17 | All Ages | 13-17 | 13-17 |
| *BG_INC* | 1.012*** (.0001) | 1.015*** (.0003) | - | - | - |
| *BG_EDU* | - | - | 1.694*** (.0032) | 1.786*** (.0158) | - |
| *PAR_ED* | - | - | - | - | 1.452*** (.0145) |
| Obs. | 232396969 | 20277134 | 232393719 | 20277121 | 20823120 |

Participation by Standardized SES variables

| | Neighborhood BG Median Inc. | | Neighborhood BG Education | | Parental Educ. |
|---|---|---|---|---|---|
| | All Ages | 13-17 | All Ages | 13-17 | 13-17 |
| *BG_INC* | 1.444*** (.0029) | 1.595*** (.0122) | - | - | - |
| *BG_EDU* | - | - | 1.951*** (.0046) | 2.087*** (.0235) | - |
| *PAR_ED* | - | - | - | - | 2.975*** (.0866) |
| Obs. | 232396969 | 20277134 | 232393719 | 20277121 | 20823120 |

**Table S3.**

Participation likelihood as a function of SES variables. Exponentiated coefficients (i.e. odds ratios) are reported. *BG_INC* coefficient estimates are scaled in thousands of dollars (i.e. a one-thousand dollar difference in neighborhood income is associated with 1.012 times the odds of participation). Age dummies are included in all specifications. Note that the ACS dataset includes probability (i.e. sampling) weights, while the Esri demographic data is reshaped to include frequency weights. We present the number of observations implies by the ACS sampling weights. Standard errors in parentheses.

$^{*}$ $p < 0.05$, $^{**}$ $p < 0.01$, $^{***}$ $p < 0.001$

Certification by Unstandardized SES variables

| | Neighborhood BG Median Inc. | | Neighborhood BG Education | | Parental Educ. |
|---|---|---|---|---|---|
| | All Ages | 13-17 | All Ages | 13-17 | 13-17 |
| *BG_INC* | 1.002*** (.0004) | 1.004*** (.0008) | - | - | - |
| *BG_EDU* | - | - | 1.055** (.0172) | 1.242*** (0.0365) | - |
| *PAR_ED* | - | - | - | - | 1.088** (0.0333) |
| Observations | 201225 | 8481 | 201146 | 8481 | 2112 |

Certification by Standardized SES variables

| | Neighborhood BG Median Inc. | | Neighborhood BG Education | | Parental Educ. |
|---|---|---|---|---|---|
| | All Ages | 13-17 | All Ages | 13-17 | 13-17 |
| *STD_BG_INC* | 1.059*** (0.0142) | 1.130*** (0.0264) | - | - | - |
| *STD_BG_EDU* | - | - | 1.070** (0.0222) | 1.317*** (0.0491) | - |
| *STD_PAR_ED* | - | - | - | - | 1.277** (0.1125) |
| Observations | 201225 | 8481 | 201146 | 8481 | 2112 |

Certification by Standardized SES variables ("mode" indicators omitted)

| | Neighborhood BG Median Inc. | | Neighborhood BG Education | | Parental Educ. |
|---|---|---|---|---|---|
| | All Ages | 13-17 | All Ages | 13-17 | 13-17 |
| *STD_BG_INC* | 1.063*** (0.0141) | 1.145*** (0.0247) | - | - | - |
| *STD_BG_EDU* | - | - | 1.078*** (0.0221) | 1.344*** (0.0517) | - |
| *STD_PAR_ED* | - | - | - | - | 1.313** (0.120) |
| Observations | 201225 | 8481 | 201146 | 8481 | 2112 |

**Table S4.**

Certification by SES. Neighborhood block group median income is in thousands of dollars. Exponentiated coefficients (i.e. odds ratios) and robust standard errors clustered at the course level are reported. Age, course, and mode dummies are included in all specifications. Stars indicate coefficients statistically significantly different from 1.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Participation by neighborhood SES (binned)

| | *CB_EDU* | | | *CB_INC* | |
|---|---|---|---|---|---|
| | All Ages | Age 13-69 | | All Ages | Age 13-69 |
| *BIN_9* | 1.978*** | 0 | *BIN_10* | 0 | 0 |
| | (0.356) | (.) | | (.) | (.) |
| *BIN_10* | -0.563* | 0 | *BIN_30* | -0.139*** | 0.056 |
| | (0.233) | (.) | | (0.0146) | (0.097) |
| *BIN_11* | 0 | 0 | *BIN_50* | 0.118*** | 0.406*** |
| | (.) | (.) | | (0.0143) | (0.095) |
| *BIN_12* | 0.219*** | 0.054 | *BIN_70* | 0.449*** | 0.889*** |
| | (0.0449) | (0.173) | | (0.0147) | (0.095) |
| *BIN_13* | 0.463*** | 0.184 | *BIN_90* | 0.675*** | 1.204*** |
| | (0.045) | (0.161) | | (0.0153) | (0.096) |
| *BIN_14* | 0.900*** | 0.671*** | *BIN_110* | 0.935*** | 1.557*** |
| | (0.042) | (0.158) | | (0.0155) | (0.096) |
| *BIN_15* | 1.419*** | 1.229*** | *BIN_130* | 1.140*** | 1.817*** |
| | (0.042) | (0.158) | | (0.0185) | (0.102) |
| *BIN_16* | 1.955*** | 1.861*** | *BIN_150* | 1.287*** | 2.006*** |
| | (0.042) | (0.158) | | (0.0205) | (0.105) |
| *BIN_17* | 2.496*** | 2.470*** | *BIN_170* | 1.482*** | 2.285*** |
| | (0.042) | (0.159) | | (0.0243) | (0.111) |
| *BIN_18* | 3.026*** | 2.838*** | *BIN_190* | 1.540*** | 2.275*** |
| | (0.044) | (0.174) | | (0.0341) | (0.132) |
| *BIN_19* | 2.337*** | 2.895** | *BIN_210* | 1.749*** | 2.083*** |
| | (0.126) | (1.014) | | (0.0282) | (0.129) |
| Observations | 232393719 | 20259024 | | 232396969 | 20277134 |

**Table S5.**

Logistic regression models of participation by neighborhood SES where bins are constructed from levels of neighborhood educational attainment (or median income) and treated as unordered categories in regressions. Coefficients are reported in logits. Standard errors in parentheses. $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Within state, census tract, and county differences

| | State | | County | | Census Tract | |
|---|---|---|---|---|---|---|
| | *BG_INC* | *BG_EDU* | *BG_INC* | *BG_EDU* | *BG_INC* | *BG_EDU* |
| *PARTICIPANT* | 10.48*** | 0.803*** | 6.75*** | 0.560*** | 0.800*** | 0.0428*** |
| | (.0873) | (0.0033) | (0.0806) | (0.0030) | (0.0399) | (0.0013) |
| ≈Observations | 232.4 mil | 232.4 mil | 232.4 mil | 232.4 mil | 232.4 mil | 232.4 mil |

**Table S6.**

Neighborhood SES differences between MOOC participants and U.S. population estimated within states, census tracts, and counties. *BG_INC* coefficients are scaled in thousands of dollars. Age dummies are included in all specifications. In *BG_INC* models, the exact number of observations are 232,396,969, 232,396,963, and 232,396,963. In the *BG_EDU* models, the exact number of observations is 232,393,719 in all specifications. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Neighborhood SES by missing survey data

| | (1)<br>All | (2)<br>Age 13-17 | (3)<br>All | (4)<br>Age 13-17 |
|---|---|---|---|---|
| *MISSING_PED* | 1.085*** | 1.142 | 0.0558*** | -0.0033 |
| | (0.195) | (1.020) | (0.0072) | (0.0329) |
| Observations | 164198 | 7625 | 164123 | 7625 |

**Table S7.**

Neighborhood SES differences between MOOC participants reporting parental education and MOOC participants not reporting parental educational attainment. The dependent variable in models (1) and (2) is block group median income, and the dependent variable in models (3) and (4) is block group educational attainment. Age dummies are included in all specifications. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Mailing address-matched and non-matched participants (all ages)

| | *BG_INC** | *BG_EDU** | *BG_INC** | *BG_EDU** | *BG_INC** | *BG_EDU** |
|---|---|---|---|---|---|---|
| *ADD_MATCH* | -433.6*** | -0.077*** | -1034.7*** | -0.0916*** | -1049.1*** | -0.0868*** |
| | (151.1) | (0.0054) | (154.9) | (0.0055) | (152.9) | (0.0056) |
| Age indicators | No | No | Yes | Yes | Yes | Yes |
| State indicators | No | No | No | No | Yes | Yes |
| Obs. | 282457 | 278198 | 282457 | 278198 | 257596 | 257050 |

SES differences between address-matched and non-matched participants (age 13-17)

| | *BG_INC** | *BG_EDU** | *BG_INC** | *BG_EDU** | *BG_INC** | *BG_EDU** |
|---|---|---|---|---|---|---|
| *ADD_MATCH* | 1414.8* | 0.0248 | 1264.7* | 0.0213 | 1012.3 | 0.0084 |
| | (613.2) | (0.0205) | (612.5) | (0.0205) | (586.2) | (0.0203) |
| Age indicators | No | No | Yes | Yes | Yes | Yes |
| State indicators | No | No | No | No | Yes | Yes |
| Obs. | 19243 | 19047 | 19243 | 19047 | 18553 | 18512 |

**Table S8.**

Neighborhood SES differences between MOOC participants reporting parsable and MOOC participants not reporting a parsable mailing address. Neighborhoods are identified using modal IP address, including for participants for whom a parsable mailing address is available. Only individuals located to the same state by both IP address and mailing address are included (49% of all U.S. participants reporting an age between 13 and 69). Standard errors in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

MOOC participant neighborhood population density compared to US

| | Age 13-69 | Age 13-17 | Age 13-69 | Age 13-17 |
|---|---|---|---|---|
| *EDX* | 0.606*** | 0.413*** | 0.379*** | 0.140*** |
| | (0.0049) | (0.0225) | (0.0038) | (0.0171) |
| State indicators | No | No | Yes | Yes |
| Observations | 232396969 | 20277134 | 232396969 | 20277134 |

**Table S9.**

Neighborhood population density differences between MOOC participants and U.S. population**.** The dependent variable is the natural logarithm of population density. Age dummies are included in all specifications. Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Population density

| | Ages 13-69 | | | | Age 13-17 | | | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| *LN_PD* | 1.213 | 1.204 | 1.142 | 1.112 | 1.151 | 1.117 | 1.121 | 1.068 |
| | (0.0019) | (0.0022) | (0.002) | (0.002) | (0.0085) | (0.0089) | (0.009) | (0.008) |
| *BG_INC* | 1.0121 | 1.0113 | | | 1.015 | 1.0141 | | |
| | (0.0001) | (0.0001) | | | (0.0002) | (0.0003) | | |
| *BG_EDU* | | | 1.623 | 1.572 | | | 1.754 | 1.707 |
| | | | (0.003) | (0.003) | | | (0.016) | (0.015) |
| State indic. | No | Yes | No | Yes | No | Yes | No | Yes |
| Obs. | 232mil | 232mil | 232mil | 232mil | 20mil | 20mil | 20mil | 20mil |

**Table S10.**

Logistic regression models of MOOC participation by population density, controlling for neighborhood SES characteristics. Exponentiated coefficients (i.e. odds ratios) are reported. Age dummies are included in all specifications. In specifications 1-2, the exact number of observations is 232,396,969. In specifications 3-4, the exact number of observations is 232,393,719. In specifications 5-6, the exact number of observations is 20,277,134. In specifications 7-8, the exact number of observations is 20,277,121. Standard errors are in parentheses. All coefficients are statistically significant ($p < 0.001$).