Can MOOC Programs Improve Student Employment Prospects?*

Aboozar Hadavand[1]
Ira Gooding[1]
Jeffrey Leek[1]

[1]Johns Hopkins Bloomberg School of Public Health

October 4, 2018

Abstract

Massive open online courses (MOOCs) exploded into popular consciousness in the early 2010s. Huge enrollment numbers set expectations that MOOCs might be a major disruption to the educational landscape. However, there is still uncertainty about the new types of credentials awarded upon completion of MOOCs. We surveyed close to 9,000 learners of the largest data science MOOC program to assess the economic impact of completing these programs on employment prospects. We find that completing the program that costs less than $500 led to, on average, an increase in salary of $8,230 and an increase in the likelihood of job mobility of 30 percentage points. This high return on investment suggests that MOOCs can have real economic benefits for participants.

Keywords: MOOCs; Online Education; Return to Education; Data Science; Propensity Score Matching

JEL Codes: I26; J62; J24

1 Introduction

Massive Open Online Courses (MOOCs) are recent developments in education. They take the form of online courses that are provided for free or at low cost and are open with no admission process. Growth in MOOC participation was so rapid that it was even faster than user growth on social media platforms such as Facebook.[1] Proponents argue that MOOCs are cheaper and more scalable than traditional college courses and can serve a wider variety of students with different needs and learning habits (Haber, 2014).

However, the enthusiasm surrounding MOOCs has been tempered by early results on MOOC

*Correspondence to: jtleek@jhu.edu. Research reported in this paper was supported by National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under award number 1R01GM115440-01A1. We would like to acknowledge the Coursera Research team and the Johns Hopkins Data Science Lab for support in completing this project. We are grateful for the feedback on earlier drafts provided by Leonardo Collado-Torres, Lucy DAgostino McGowan, Shannon Ellis, Stephanie Hicks, Leah Jager, Kai Kammers, Sarah McClymont, Elizabeth Stuart, Margaret Taub, and Sarah Thomas.

[1]The Year of the MOOC, The New York Times, November 2, 2012, https://nyti.ms/2NRGiEO.

completion rates and learner populations (Drezner, 2013). Critics of MOOCs have highlighted high attrition rates among learners (Jordan, 2015) and the fact that MOOCs tend to serve those who already have a college degree or above (Christensen et al., 2013) as causes for skepticism. Moreover, it is unclear whether the non-degree credentials typically offered after completion of a MOOCs can serve as an alternative to a university degree or lead to tangible economic or career benefits for learners.

Research on MOOCs is still at an early stage. Although there is an increasing number of studies on learners behavior and educational outcomes on MOOC platforms, to date, there is no study that quantifies the economic benefits of MOOCs to learners. We found only two qualitative studies on how MOOCs improve job prospects of participants. The first study found that 72% of survey respondents reported career benefits (Zhenghao et al., 2015). The findings are based on a survey of 52,000 individuals who had completed a Coursera MOOC. A second study looking at the impact of MOOCs on employability surveyed 441 learners who indicated they were motivated to take MOOCs for financial or employment-related reasons (Dillahunt et al., 2016). Both of these papers make no attempt to quantify the benefits.

We focus on studying and quantifying two potential economic benefits of taking a MOOC: increased earnings and job mobility. Our analysis is based on comparing completers and non-completers of the Johns Hopkins University (JHU) Data Science Specialization (DSS) on the MOOC platform Coursera. Launched in January 2012, Coursera is the most popular provider of MOOCs in the world with more than 25 million learners. Coursera partners with close to 150 universities across 29 countries to produce more than 2,000 MOOCs covering subjects ranging from computer science to philosophy. A typical course on Coursera includes recorded video lectures, graded assignments, quizzes, and discussion forums. Upon successful completion of certain courses or bundle of courses, students have the option to obtain a certificate from Coursera and the partner university.

Johns Hopkins University (JHU) Data Science Specialization (DSS) contains a set of 9 courses and a capstone project. A description of each course is provided in Table 1. The courses mainly constitute video lectures, peer-assessed programming assignments, quizzes, and forums. The last class included a final project that was peer assessed as well, i.e., each submission was randomly assigned to a few other students who graded the project using a fixed rubric. Data science is a heavily applied field that has grown out of traditional statistics. It is defined as a science that "seeks actionable and consistent pattern for predictive uses" (Dhar, 2013) and, therefore, is different from the existing practice of data analysis in various disciplines, which focuses only on explaining data sets.

The initial JHU data science courses on Coursera were created by three Department of Biostatistics faculty members in 2012. These courses reflected, to a large extent, material that was developed for and used in traditional courses at JHU. The offering of these courses, first as a combination of three courses and later as a combination of ten courses, coincided with Coursera's plan to offer Specializations. This collaboration led to the birth of the DSS on Coursera in 2014. There have since been more than 4 million enrollments for the entire program (Kross et al., 2017) and 13,119 learners have completed the entire Specialization.

Table 1: List of courses in Johns Hopkins University Data Science Specialization on Coursera.

Course	Description
The Data Scientists Toolbox	Version control and the most common tools used by data scientists
R programming	Fundamentals of how to program in R
Getting and cleaning data	How to obtain data from databases, the web, and other sources, then clean it up
Exploratory data analysis	Plotting and other initials explorations of a data set
Reproducible research	Basics of R markdown, literate programming, and the principles of reproducible research
Statistical inference	Probability and statistical inference
Regression models	Linear models and their application to real data
Practical machine learning	Basics of machine learning and predicting in R
Developing data products	Technologies for building data products like R packages, swirl courses, and Shiny apps
Data Science Capstone	Allowing students to create a usable/public data product

2 Data

Our study relies on two unique sources of data. First, we use data provided by Coursera, which tracks learner behavior on the Coursera website. Coursera provides user data for specific courses or programs only to the instructors and institutions involved in the creation of those courses. Coursera collects information on learners in five major categories: course content, demographics, learner progress, learner outcome, and forum participation.

Our second source of data is a survey that we sent to learners of DSS. Approximately 9,000 DSS learners responded to the survey. The survey includes questions on job market experience before and after completing the program, educational background, and other demographic variables. More specifically, we ask questions about employment status, occupation, income, the number of job changes since graduation from the program, and the probability of changing jobs in the future. Appendix C includes our survey questionnaire. More than three-quarter of learners described their pre-enrollment level of expertise in data science as a beginner. A majority of our respondents (64.1%) were employed before the DSS and this ratio increased (to 69.3%) following program participation. The share of people who used data science skills in their job increased by ten percentage points after taking DSS courses.

The survey was sent to all of the learners who interacted with the most popular course in the Specialization, R Programming. Interaction is defined as clicking on the course, reading the course description, pre-registration, or enrollment. As shown in Figure 1, a total of 297,372 learners have interacted with the R Programming course since 2015 out of which 57,993 enrolled in the course, 55,360 completed at least one course-item (a lesson, quiz, or exercise), and 29,465 completed the entire course. We only include respondents from the top five and the top ten countries (countries with the highest shares of learners) in our analysis (due to the sudden drop in the share of respondents in the rest of the countries), which leaves us with 3,513 observations.

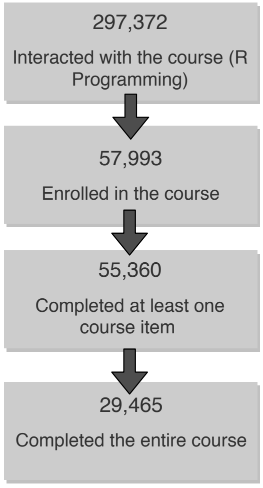

Figure 1: The number of learners with different levels of interaction with the course R Programming

The top ten countries in the sample are the United States, India, United Kingdom, Brazil, Canada, Germany, China, Netherlands, Spain, and Australia.

The survey respondents were representative of the learners in DSS as shown in Table 2 and Figure A1 in Appendix A. The table shows that in terms of genders, employment groups, and countries the survey sample closely resembles the broader population. For instance, among our survey respondents the share of female learners is 31%, the share residing in the United States is 35%, the share who are employed full time is 64%, and the share with a masters degree or above is 65%, while among general DSS learners on Coursera these shares are 29%, 34%, 65%, and 61%, respectively.

The main difference between the groups is when we look at education groups. More specifically, there are slightly more people with a doctoral degree in the survey sample than in the broader population (17 vs. 11 percent). Demographic data on Coursera is optional and usually filled upon signing up on the platform, therefore, there are a lot of missing values, and it is likely that some of the information is outdated. Due to the limitation of Coursera's demographic data, we could only compare the two groups in the dimensions mentioned above. We were only able to study those who interacted with the course since 2015 since Coursera changed their data collection standards in 2015, which made it impossible for us to link pre- and post-2015 data.

The survey link was sent through Coursera's announcement dashboard. Using the dashboard allowed us to link our survey respondents to their learner data through a hashed user ID that we embedded in the email. This allowed us to exploit both the survey data and learners data. Coursera has provided partner institutions and instructors a dashboard for downloading data for a single course or all courses associated with the institution. These data called, Research Data Exports, are sets of 100 CSV files for each course. One of the advantages of the data is the existence of a single hashed user ID for each learner. The hashed user ID is consistent for learners across all courses offered by an individual institution and allows for connecting learner grades and progress across courses. Table B1 in Appendix B explains what kind of variables are collected in each of these categories.

Table 2: Comparison of the share of learners in gender, country, employment, and education groups for both the survey respondents and Coursera DSS learners. Demographic data on Coursera are only available in these four dimensions.

	Survey Respondents	Coursera Learners
Gender	Female: 30.9% Male: 68.8% Other: 0.3%	Female: 29.0% Male: 71.0%
Country	United States: 35% India: 11% United Kingdom: 4% Brazil: 4% Canada: 4% Germany: 3% Netherlands: 3% China: 3% Spain: 2% France: 2% Other: 29%	United States: 34% India: 12% United Kingdom: 4% Brazil: 3% Canada: 3% Germany: 3% Netherlands: 2% China: 2% Spain: 3% France: 2% Other: 32%
Employment status	Full time: 64% Part time: 9% Unemployed: 27%	Full time: 65% Part time: 6% Unemployed (Looking for work and not looking for work): 20% Other: 9%
Highest Education	Less than high school degree: 0% High school degree or equivalent (e.g., GED): 2% Associate degree: 1% Some college but no degree: 3% Bachelor's degree: 29% Master's degree: 44% Professional degree: 4% Doctoral degree: 17%	Less than high school degree: 0% High school degree or equivalent (e.g., GED): 2% Associate degree: 1% Some college but no degree: 5% Bachelor's degree: 31% Master's degree: 46% Professional degree: 4% Doctoral degree: 11%

3 Binary and Non-binary Treatment Propensity Score Matching

The main challenge in program evaluation studies using observational data is endogeneity (or self-selection) issues. In the case of MOOC program evaluation, we cannot observe inherent differences between those who complete the program and those who do not. Individuals who complete a training program are by definition different from those who do not complete. These differences, if they influence the response, may discredit causal comparisons of outcomes by treatment status, even if one controls for the observed covariates. In the absence of experimental designs, there are various ways we could employ to analyze the effectiveness of the program using observational data. As a result, standard methods such as ordinary least squares, which are based on covariate adjustment are not sufficient (Imbens, 2000).

A method for adjusting for pre-treatment variables has been proposed by Rosenbaum and Rubin (1983) that is based on calculating the conditional probability of receiving the treatment, also known as a Propensity Score (PS), given pre-treatment variables. Propensity-score based methods are advantageous to standard regression methods since they (a) work as dimension reduction tools since they reduce multiple covariates into one propensity score, which is an important feature when there are a large number of pretreatment covariates (Rosenbaum and Rubin, 1984), (b) make it clear when it is and it is not possible to separate the effect of the treatment from the effect of other covariates (Stuart, 2010), (c) do not require a formal model of the response (outcome) variable, which reduces potential biases caused by misspecification such as multicolinearity (Ho et al., 2007), and (d) do not extrapolate beyond observed data (Stuart, 2010). Propensity-score based methods are becoming more popular in program evaluation analyses using observational data. Various studies use propensity-score based methods to examine the effect of educational or job training programs (Heckman et al., 1997; Dehejia and Wahba, 1999; Black and Smith, 2004; Lechner, 1995; Blundell et al., 2005).

The PS weighting method scans the observational data and weights observations to create balanced covariate distributions between treated and control groups. Therefore, unlike ordinary least squares that attributes the same weight to each observation, in PS weighting, observations are weighted by the probability of treatment - in our case, the probability that a given learner would complete the program. Adjusting for this conditional probability, also known as a propensity score (PS), removes all biases that result from imbalances in pre-treatment variables for the treated and control groups. We calculate this probability through a logit model using the course completion variable on the left-hand side.

Emphasis has to be made about what our estimates will mean. The weights are calculated using the following formula:

$$\begin{cases} 1/PS & \text{if } T = 1 \\ 1/(1 - PS) & \text{if } T = 0 \end{cases} \tag{1}$$

where PS is the propensity score and T is the treatment indicator. Therefore, each observation, depending on whether it is in the treatment ($T = 1$) or control group ($T = 0$), receives a different weight. Given this method, our estimates reflect the average treatment effect (ATE). ATE reflects the mean impact of treatment if everyone in the population received the treatment and, therefore, shows the difference in the population means of the two potential outcomes, completing beyond the threshold or not. The threshold divides the sample into control and treatment units. For instance, if we set the treatment threshold at 50 percent of the courses (completing at least 5 out of 10 courses), those who complete less than 50% of the courses will be in the control group and those who complete equal to or more than 50% of the classes will be in the treatment group.

We used the following variables in the regression used to calculate the weights: age, gender, race (white vs. non-white), country of residence, highest level of education, college major, binary variable indicating whether their motivation for enrollment was career-driven, pre-treatment employment status, binary variable indicating whether JHU DSS was the only data science training they received, and categorical variable indicating whether their major is related to data science. Figure A2 in Appendix A illustrates the distribution of the propensity score in the treated and control groups using the original and the weighted sample. It is generally desired to have an overlap of the distributions for the treated and the control groups. Figure 2 shows

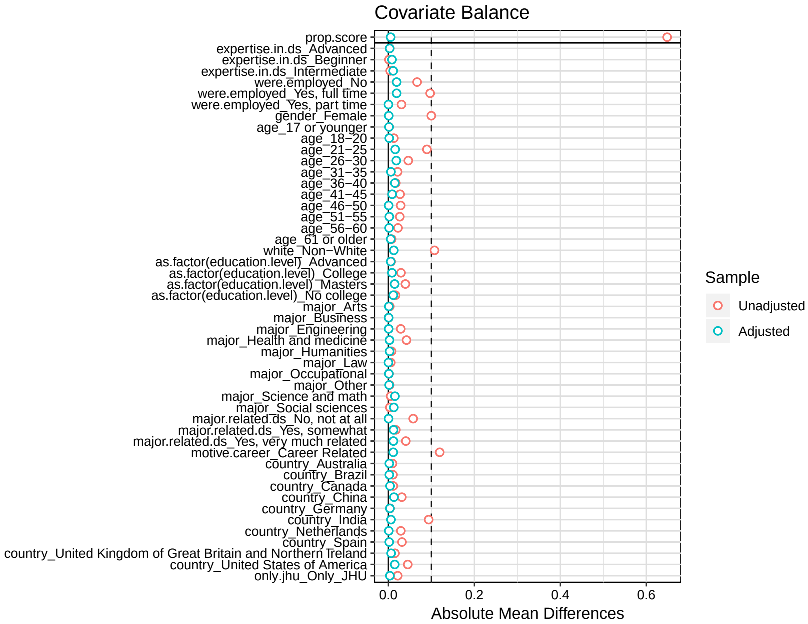

Figure 2: Balance of the propensity score and the covariates between the control and the treatment groups for both the unadjusted (unweighted) and the adjusted (weighted) samples. The binary treatment cutoff point is at 50%, i.e., those who passed fewer than five courses are in the control group and those who passed equal to or more than five courses are in the treatment group. The sample includes the top ten countries with the highest shares of respondents to our survey including China and India.

the balance of covariates (the absolute mean difference between control and treatment groups) for the unadjusted (unweighted) and the adjusted (weighted) samples. Achieving a balance of the covariate is the very purpose of weighting, allowing us to make a better causal inference. As Figure 2 shows, we achieve more balanced covariates after propensity score weighting.

Figure 3 shows the distribution of post-completion income using the resampled weighted observations. In other words, we create a pseudo-population using the propensity score weights in order to find the distribution of the new sample. The figure shows the incomes for the top ten countries based on the share of respondents in our sample for three different groups: those who did not complete any of the courses, those who completed between one to five courses, and those who completed more than five courses.

4 Empirical Results

To better understand the effect of the program on incomes, we perform a regression analysis using the PS-weighted sample. We then regress post-completion income on demographic covariates, education, and country of residence. Table 3 reports the regression results when we regress income on the binary variable indicating whether the number of courses a student passes falls above or below the threshold. The threshold is set at three different levels: 20%, 50%, and 80% of all courses. Panel (A) shows the result for when we only look at the top five countries with the highest shares of respondents for both when we include and exclude China and India.

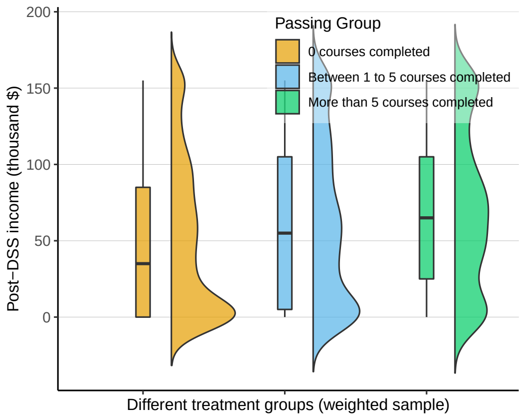

Figure 3: The propensity-score weighted and resampled distribution of income in dollars for the top ten countries in the sample for three different groups: those who did not pass any courses, those who passed between 1-5 courses, and those who passed more than five courses.

We exclude China and India due to the fact that income per capita in those countries differ significantly from the rest of the countries in our sample and it might skew our estimates. Panel (B) similarly shows the result for the top 10 countries.

Table 3: Regression of post-DSS income on completion rate using the propensity-score weighted sample.

Panel (A): Top 5 countries with the highest shares of respondents

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Threshold	20%	50%	80%	20%	50%	80%
India/China	Incl.	Incl.	Incl.	Excl.	Excl.	Excl.
Effect	6222.4***	4820.4**	2996.0	7820.0***	6565.3**	3570.2
t-statistic	3.466	2.123	0.912	3.673	2.471	0.936
Adjusted R^2	0.35	0.34	0.32	0.25	0.24	0.22

Panel (B): Top 10 countries with the highest shares of respondents

	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Threshold	20%	50%	80%	20%	50%	80%
India/China	Incl.	Incl.	Incl.	Excl.	Excl.	Excl.
Effect	5412.0***	4840.2**	2885.1	6644.8***	6245.8***	2790.8
t-statistic	3.367	2.317	0.948	3.489	2.663	0.829
Adjusted R^2	0.35	0.34	0.31	0.26	0.24	0.22

Note: Effects are in dollars. Threshold shows the cutoff point at which treatment and control groups are separated. For instance, threshold 20% means that those who passed fewer than two courses are counted as the control group and those who passed equal to or more than two courses are counted as the treatment group. Panel (A) shows the results for the top five countries and Panel (B) shows the results for the top ten countries with the highest respondents. Other covariates include gender, race, age, education, and whether the learner has taken any other training in data science besides DSS. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

If we set the treatment threshold at 50 percent of the courses (completing at least 5 out of 10 courses), the average increase in income due to participation in the program is $4,840 for the top ten countries in the sample. Depending on the number of countries considered and the treatment threshold, the effect can range from $2,790 to $7,820. All values are in U.S. dollars. Figure A3 in Appendix A shows a scatter plot of residuals on the y axis and fitted values (estimated responses) on the X-axis for Model (8). As the figure represents, the correlation between the residuals and the fitted values is not significant [correlation = -0.01, P = 0.316].

To better analyze different levels of course completion on income, we exploit a recent generalization of binary treatment propensity score models. Following the approach in Hirano and Imbens (2004), we estimate propensity scores for each course completion level (zero percent to 100 percent).

While in the binary treatment case the propensity score is estimated using logistic regression, in the multiple treatment approaches, where treatments correspond to ordered levels of treatment, as it is in our case, a parametric model is fitted to the observed outcomes. A generalization of the of the binary treatment propensity score has been proposed by Hirano and Imbens (2004) that they call generalized propensity score (GPS). The GPS is defined as the treatment assignment density calculated at a particular treatment value and set of covariates. The model is based on the conditional distribution of treatment T_i given covariates X_i. If the GPS is evaluated at the realized X_i and any specific treatment level, it becomes a random variable that can be used to balance covariates. Therefore, the GPS can be defined as

$$r(t,x) = f(T_i = t|X_i = x) \tag{2}$$

In practice, the GPS is calculated in two stages. In the first stage, we fit a model to estimate $f(T_i|X_i)$. In the second stage, we use the model in the first stage to obtain an estimate of $E(Y_i(t))$. We use the R package causaldrf[2] to perform the continuous treatment analysis.

After calculating the propensity score, we regress post-completion income on the number of courses passed as a continuous variable, the propensity score, and other covariates including education level, gender, age, race, a categorical variable for college major, and country of residence. The result of the regression are reported in Panel (A) in Table 4. If we consider the top five countries in the sample, we find that for every passed course income increases by \$117 $\times$ 10 = \$1,170 if we include China and India and \$111 $\times$ 10 if we exclude India and China. The effect decreases if we look at the top ten countries in the sample with the highest share of respondents. The effect is equal to \$82 $\times$ 10 if we include India and China and \$73 $\times$ 10 if we exclude them.

Figure 4 shows the percentage increase in income for every additional course in the Specialization as well as the 95 percent confidence interval. As it can be seen from the plot, most of the increase in income happens in the first quarter of the Specialization.[3]

According to Coursera, course completion is defined as completing all graded assignments for the courses in the Specialization. In fact, if the learner does not finish all the assignments, they will not receive a certificate from JHU. We test whether there is any sheepskin effect, also referred to as a signaling effect. The sheepskin effect states that employees with a credential will have higher earnings than those with the same training but no credential (Hungerford and

[2]Galagate, D., Schafer, J. (2015). Package causaldrf: Tools for Estimating Causal Dose Response Functions, Version 0.3, https://cran.r-project.org/web/packages/causaldrf.

[3]Although each learner has the freedom to take the courses in the order they wish, we found through clustering that most learners take two courses in the first 25 percent of their progress toward the Specialization certificate. These two courses are Data Scientists Toolbox and R Programming.

Table 4: Regression of post-DSS income on completion rate using the propensity-score weighted sample.

Panel (A): Binary course completion

	Model 1	Model 2	Model 3	Model 4
No. of countries	5	5	10	10
India/China	Incl.	Excl.	Incl.	Excl.
Effect	117.02**	111.29*	82.39*	73.36
t-statistic	2.295	1.936	1.937	1.541
Adjusted R^2	0.35	0.25	0.35	0.26

Panel (B): Completion defined as the percentage of course items completed

	Model 1	Model 2	Model 3	Model 4
No. of countries	5	5	10	10
India/China	Incl.	Excl.	Incl.	Excl.
Effect	115.16**	106.18*	85.44**	69.22
t-statistic	2.385	1.943	2.097	1.513
Adjusted R^2	0.34	0.25	0.35	0.25

Note: Effects are in dollars and show the increase in earnings for each additional course in the Specialization. Panel (A) shows the results for when the completion rate is calculated based on completing a course and Panel (B) shows the results for when the completion rate is calculated as the share of course items passed. Other covariates in both regressions include gender, race, age, education, and whether the learner has taken any other training in data science besides DSS. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

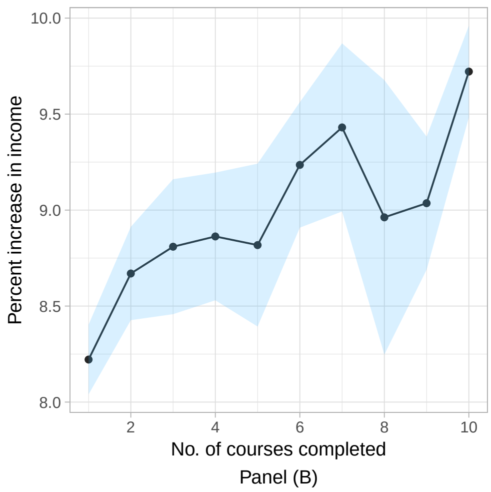

Figure 4: The percentage increase in income for every additional course in the Specialization as well as the 95 percent confidence interval for the top ten countries in the sample.

Solon, 1987). In our Coursera dataset, for each learner in each course, we can see how many items of the course they have passed. We use this to compute a non-binary course completion rate that captures the percentage of items within a course a learner has completed. Aggregating this across all courses creates a new continuous outcome that is the fraction of the Specialization completed by a student. Figure A4 in Appendix A shows the distribution of the binary and non-binary completion rates. The two distributions are slightly different but are highly correlated [correlation = 0.98, $P < 1 \times 10^{-3}$].

Using the same method by Hirano and Imbens (2004), we run a regression similar to the one above with the same set of covariates. In all of the regressions presented above, we include a binary variable indicating whether the learner has taken other data science courses to be able to isolate the effect of the DSS. We find the coefficient on the share of courses passed as \$85, reflecting the increase in income for every percentage point progress throughout the Specialization. This means that for every 10 percent progress (roughly equivalent to one course), income increases by \$850, which is similar to when we only looked at binary course completion, which suggests that there is a limited signaling effect from completing the courses. We use the same covariates as the previous regression. The results are shown in Panel (B) of Table 4. The effects are similar to the ones reported in Panel (A). This may be due to similar reasons: the two scores are similar for most individuals in our sample and/or, if they are different, there is not much of a signaling effect from receiving a certificate.

We further consider whether completing MOOCs has any impact on job mobility in addition to the effect on earnings. Even if taking MOOCs does not lead to an increase in income, it could potentially lead to a change in career path or promotions. We asked learners whether they had changed jobs or gotten promoted since completion of DSS. We then fit a model to evaluate the association between DSS completion and the probability of changing jobs using the continuous treatment variable in Hirano and Imbens (2004). The results of the regression are presented in Table 5. As shown in the table, for every additional course completed, the likelihood that the learner changes job increases by 2.8-3.5% depending on the number of countries included in the model. Therefore, on average, completion of the entire DSS increases this likelihood by 28–35 percentage points.

Table 5: Regression of whether the learner has changed jobs since the start of taking DSS on completion rate using the propensity-score weighted sample

	Model 1	Model 2	Model 3	Model 4
No. of countries	5	5	10	10
India/China	Incl.	Excl.	Incl.	Excl.
Effect	3.1***	2.8***	3.5***	3.3***
t-statistic	4.608	3.941	6.137	5.519
Adjusted R^2	0.06	0.06	0.06	0.06

Note: Effects show the increase in the likelihood that a learner has changed job since the start of taking the DSS for each additional course in the Specialization. Other covariates in both regressions include gender, race, age, education, and whether the learner has taken any other training in data science besides DSS. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5 Conclusion

We have studied the effectiveness of one of the most extensive and most popular data science programs offered online. Our analysis is the first attempt to measure the economic benefits of MOOC programs. The results point to positive and significant benefits of taking DSS MOOCs. College credentials are still highly regarded and employers are still not confident that MOOC credentials greatly vouch skills.[4] However, on the employees side, we find that a significant share of our respondents mention acquiring MOOC certificates on their LinkedIn profiles (69%), on their resumes (66%), during a job interview (41%), to their colleagues (40%), and to their managers (36%).

Similar to most studies on return to education, our analysis suffers from biases caused by unobserved characteristics, such as innate abilities (Cawley et al., 1998) that were impossible for us to measure in a survey. Second, DSS does not represent a typical MOOC. The high demand for data scientists makes the market for data scientists unique.[5] Despite this, we show in Table 2 that those who enrolled in DSS replicate an average learner on Coursera or similar platforms regarding demographics (outside of gender) and educational backgrounds. Third, learners were asked to report their income in the year following the completion of the courses from DSS. Thus, this only represents an estimate of the short-term earning and job mobility benefit of MOOC completion.

Despite having a low completion rate similar to other MOOCs, the number of completers of the DSS program (13,119) outweighs the number of masters degrees in statistics and biostatistics conferred in a given year in the United States (2,486) (Kross et al., 2017). This high graduation rate reflects the accessibility of MOOCs given that they are provided at a meager cost. The certificate program, on average, costs less than \$500.[6] In comparison, we estimate that the cost of a masters program in data science or analytics averages \$53,300 for a total of two years. Figure 5 compares DSS to college degrees and other certificates in terms of costs and benefits of each program for men and women. Given costs, DSS ranks among the highest in terms of the return on investment.

In Figure 5, the estimates for the cost of and the return to associate degree programs in the U.S. are from Jepsen et al. (2014). The estimates for the return to undergraduate and postgraduate degrees are from authors calculation using Current Population Survey (Annual Social and Economic Supplement) data from the Bureau of Labor Statistics. The estimates for college costs are from Abel and Deitz (2014). Average tuition costs of colleges and universities at bachelors and masters levels are authors calculations based on National Center on Education Statistics data.

In addition to the high return on investment, DSS is available to anybody around the world.

[4]The Economist, Special Report, Established education providers v. new contenders, January 12, 2017, https://econ.st/2Nh3RWv.

[5]A 2016 report by McKinsey Company projects a 12% annual increase in demand for data scientists in the United States. The age of analytics: competing in a data-driven world, McKinsey Global Institute, December 2016, https://mck.co/2JKQKif.

[6]Our data indicate that about 22% of the learners in the DSS received financial aid from Coursera that is in the form of a fee waiver.

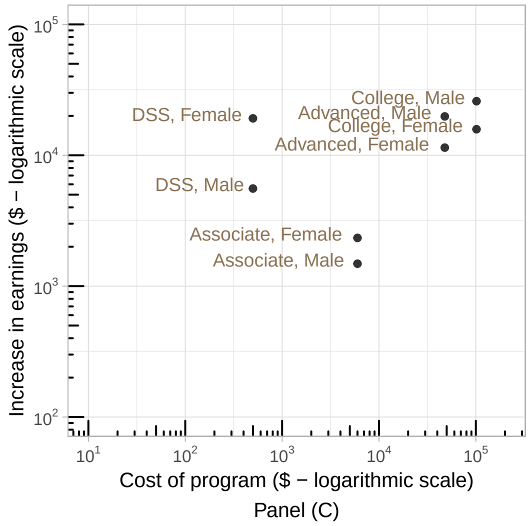

Figure 5: Comparison of the DSS to other degrees and certificates in terms of costs and benefits of each program for men and women.

In fact, one out of every four learners in the DSS program is from a lower-middle income or low-income country. MOOCs may not have disrupted traditional college programs, but our results indicate they may be a positive and economically viable supplement for people without the access or means to enroll in traditional programs.

References

Abel, Jaison and Richard Deitz. 2014. "Do the benefits of college still outweigh the costs?" Federal Reserve Bank of New York, Current Issues, Volume 20, Number 3.

Black, Dan A and Jeffrey A Smith. 2004. "How robust is the evidence on the effects of college quality? Evidence from matching." *Journal of econometrics* 121:99–124.

Blundell, Richard, Lorraine Dearden, and Barbara Sianesi. 2005. "Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168:473–512.

Cawley, John, James Heckman, and Edward Vytlacil. 1998. "Understanding the role of cognitive ability in accounting for the recent rise in the economic return to education." Technical report, National Bureau of Economic Research.

Christensen, Gayle, Andrew Steinmetz, Brandon Alcorn, Amy Bennett, Deirdre Woods, and Ezekiel Emanuel. 2013. "The MOOC phenomenon: Who takes massive open online courses and why?" Available at ssrn: http://dx.doi.org/10.2139/ssrn.2350964.

Dehejia, Rajeev H and Sadek Wahba. 1999. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association* 94:1053–1062.

Dhar, Vasant. 2013. "Data science and prediction." *Communications of the ACM* 56:64–73.

Dillahunt, Tawanna R, Sandy Ng, Michelle Fiesta, and Zengguang Wang. 2016. "Do massive open online course platforms support employability?" In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 233–244. ACM.

Drezner, Daniel. 2013. "Twilight of the MOOCs?" *Foreign Policy*, July 2013. Available: https://foreignpolicy.com/2013/07/22/twilight-of-the-moocs/ [Last accessed: 01 October 2018].

Haber, Jonathan. 2014. *MOOCs*. Cambridge, MA & London, MIT Press.

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *The review of economic studies* 64:605–654.

Hirano, Keisuke and Guido W Imbens. 2004. "The propensity score with continuous treatments." *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164:73–84.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15:199–236.

Hungerford, Thomas and Gary Solon. 1987. "Sheepskin effects in the returns to education." *The review of economics and statistics* pp. 175–177.

Imbens, Guido W. 2000. "The role of the propensity score in estimating dose-response functions." *Biometrika* 87:706–710.

Jepsen, Christopher, Kenneth Troske, and Paul Coomes. 2014. "The labor-market returns to community college degrees, diplomas, and certificates." *Journal of Labor Economics* 32:95–121.

Jordan, Katy. 2015. "Massive open online course completion rates revisited: Assessment, length and attrition." *The International Review of Research in Open and Distributed Learning* 16.

Kross, Sean, Roger D Peng, Brian S Caffo, Ira Gooding, and Jeffrey T Leek. 2017. "The democratization of data science education." PeerJ Preprints.

Lechner, Michael. 1995. "Effects of continuous off-the-job training in East Germany after unification." ZEW Discussion Papers.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70:41–55.

Rosenbaum, Paul R and Donald B Rubin. 1984. "Reducing bias in observational studies using subclassification on the propensity score." *Journal of the American statistical Association* 79:516–524.

Stuart, Elizabeth A. 2010. "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics* 25:1.

Zhenghao, Chen, Brandon Alcorn, Gayle Christensen, Nicholas Eriksson, Daphne Koller, and Ezekiel Emanuel. 2015. "Whos benefiting from MOOCs, and why." *Harvard Business Review* 25.

Statement of Conflict of Interest

Dr. Leek receives financial compensation through the Johns Hopkins Tech Transfer Program from revenue generated by the Johns Hopkins Data Science Specialization.

A Appendix A: Additional Figures

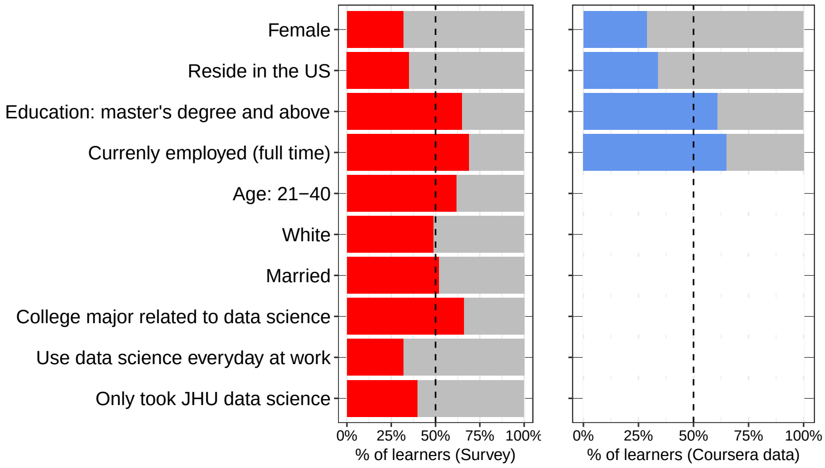

Figure A1: Summary of the share of survey respondents (left) and the share of Coursera learners (right) in various demographic and educational group. Demographic data on Coursera is only available in gender, geography, education, and employment status domains.

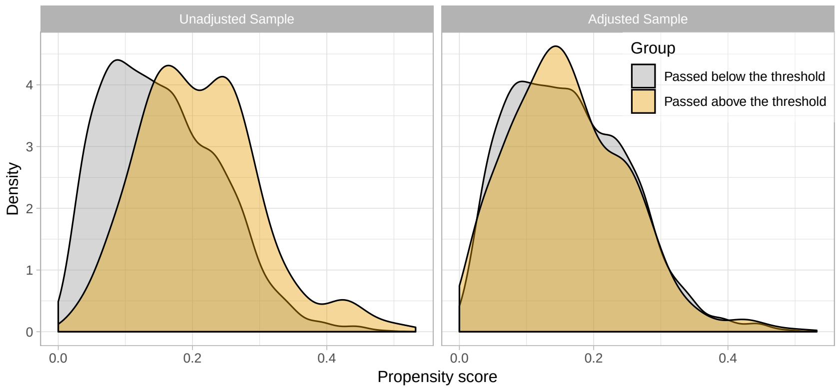

Figure A2: Density functions of the propensity score for the treatment and control groups for both the unadjusted (unweighted) and the adjusted (weighted) samples. The binary treatment cutoff point is at 50%, i.e., those who passed fewer than five courses are in the control group and those who passed equal to or more than five courses are in the treatment group. The sample includes the top ten countries with the highest shares of respondents to our survey including China and India.

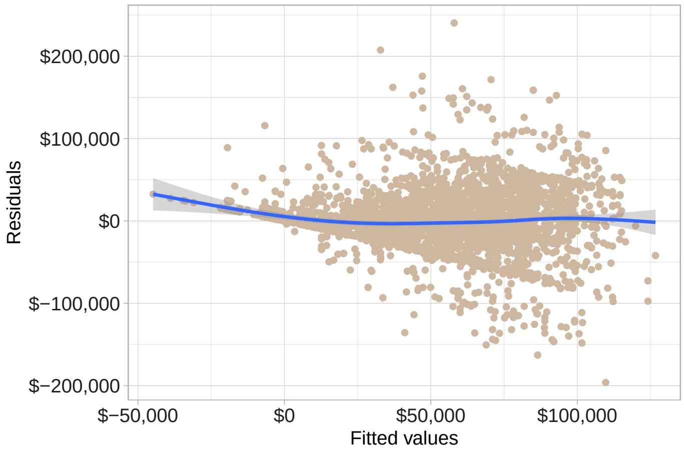

Figure A3: The plot of the residuals versus the fitted values for the weighted sample when we regress post-DSS income on the completion rate and other covariates. The binary treatment cutoff point is at 50%, i.e., those who passed fewer than five courses are in the control group and those who passed equal to or more than five courses are in the treatment group. The sample includes the top ten countries with the highest shares of respondents to our survey including China and India.

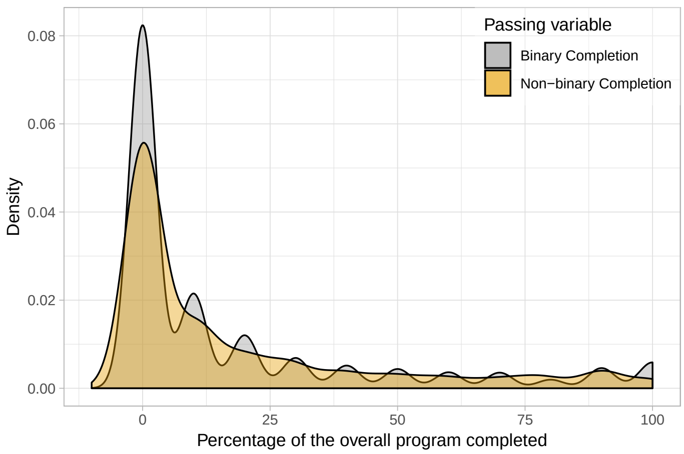

Figure A4: The distribution of the binary and non-binary completion rates among the survey respondents.

B Appendix B: Additional Tables

Table B1: Types of Coursera research data.

Data Type	Description
Assessment submission data	Assessment submissions of quizzes, peer review, and programming assignments by learners.
Course grade data	Contains the highest grade achieved by each learner on each required assessment as well as the time stamp of the learners highest-scoring submission. This table also includes each learners overall grade in the course.
Course progress data	Contains data documenting the time stamp for when the learner interacted with each piece of course content and the time stamps for when items were opened, completed, reopened, reattempted, etc.
Demographic data	Contains the following information for all enrolled learners: general geographical data (based on IP address), browser language preference, and information for learners who completed their learner profile responses or participated in Courseras platform-wide demographic survey (including age, gender, education level, and employment status).
Discussion data	Contains forum activity data such as posts, responses, upvotes/downvotes, flags, and questions and answers associated with course content items.

C Appendix C: Survey Questionnaire

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

* The purpose of this survey is to understand job market experiences of Johns Hopkins University (JHU) Data Science Specialization Students on Coursera before and after taking the Specialization courses. You are being invited to participate in this survey because Coursera Inc. records show that you have taken or started at least one of the courses of the Specialization. Your participation will greatly help our understanding of the effectiveness of the Specialization. The survey includes demographic and economic questions but no personally identifiable information about you is being collected and our research team will ensure that your answers will remain confidential. The results of this research study may be published in aggregate but your identity will not be revealed. Note that nonparticipation in the study will not affect the grade you have received (or will receive) on the Specialization courses. There are no foreseeable risks or discomforts as a result of participating in this study. Do you agree to participate in this study?

- ◯ Yes
- ◯ No

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

The following questions are about your job and experience during the months **before starting** the JHU Data Science Specialization. Press "OK" to continue.

How would you describe your **level of expertise** in data science **prior** to starting the JHU Data Science Specialization?

- ◯ Beginner
- ◯ Intermediate
- ◯ Advanced

Were you employed before starting the JHU Data Science Specialization?

- ◯ No
- ◯ Yes, part time
- ◯ Yes, full time

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

Are you **currently employed**?

- No
- Yes, part time
- Yes, full time

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

Is the current job you hold **the same job** you had just prior to taking the JHU Data Science Specialization?

- Yes, exactly
- Yes, but I've been promoted or received a raise
- No

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

Are you either **self-employed** or a **freelancer**?

- Yes
- No

Which of the following best describes your**current occupation**?

- ◯ Management Occupations
- ◯ Business and Financial Operations Occupations
- ◯ Computer and Mathematical Occupations
- ◯ Architecture and Engineering Occupations
- ◯ Life, Physical, and Social Science Occupations
- ◯ Community and Social Service Occupations
- ◯ Legal Occupations
- ◯ Education, Training, and Library Occupations
- ◯ Arts, Design, Entertainment, Sports, and Media Occupations
- ◯ Healthcare Practitioners and Technical Occupations
- ◯ Healthcare Support Occupations
- ◯ Protective Service Occupations
- ◯ Food Preparation and Serving Related Occupations
- ◯ Building and Grounds Cleaning and Maintenance Occupations
- ◯ Personal Care and Service Occupations
- ◯ Sales and Related Occupations
- ◯ Office and Administrative Support Occupations
- ◯ Farming, Fishing, and Forestry Occupations
- ◯ Construction and Extraction Occupations
- ◯ Installation, Maintenance, and Repair Occupations
- ◯ Production Occupations
- ◯ Transportation and Materials Moving Occupations
- ◯ Other (please specify)

How many times have you **changed your job** since obtaining the certificate from JHU?

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5+

How frequently do you **need data science skills** in your **current** job?

- ◯ Every day
- ◯ From time to time
- ◯ Never

What is your **annual income** from your **current job**? [please convert your income to **US dollars**]

How many years have you been at your **current job**?

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5+

Are you currently **looking for jobs**?

- Yes, continuously
- Yes, from time to time
- No

On a scale of 1-5 what is the likelihood that you **find/switch to a new job** in the **next 12 months**?

1 5

On a scale of 1-5 what is the likelihood that you **find/switch to a new job** in the **next 2 years**?

1 5

On a scale of 1 to 5, how much do you think the JHU Data Science Specialization **improved your skills in data science**?

Not at all Very much

On a scale of 1 to 5, how much do you think the JHU Data Science Specialization courses **helped** you in your career?

Not at all Very much

What **other courses or programs in data science** have you taken/attended? [select all that apply]

- JHU data science specialization was the only data science program I participated in
- Other online MOOCS (by other universities)
- Data science bootcamps
- Other (please specify)
- Online master's/micromaster's programs
- On campus classes
- Youtube videos/Other online resources

Finally, would you like to tell us a little more about**yourself** and **your motivation** for taking JHU Data Science Specialization courses?

- Yes
- No

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Survey on **Johns Hopkins University** Data Science Specialization on **Coursera**

What is your **gender**?

- Female
- Male
- Other

How old are you?

- 17 or younger
- 18-20
- 21-25
- 26-30
- 31-35
- 36-40
- 41-45
- 46-50
- 51-55
- 56-60
- 61 or older

What's your **marital status**?

- Single (never married)
- Living with partner
- Married
- Separated
- Widowed
- Divorced

In what **country** do you **currently live**?

What is your **ethnicity**? [Please select all that apply]

- [ ] American Indian, Native Hawaiian, or Alaskan Native
- [ ] Asian or Pacific Islander
- [ ] Black or African American
- [ ] Carribean
- [ ] East Asian
- [ ] Other (please specify)
- [ ] Hispanic or Latino
- [ ] Middle Eastern
- [ ] North African
- [ ] South Asian
- [ ] White / Caucasian

What is the **highest level** of **education** you have **completed**?

- Less than high school degree
- High school degree or equivalent (e.g., GED)
- Some college but no degree
- Associate degree
- Bachelor degree
- Master degree
- Professional degree
- Doctoral degree

Which major category best describes the**main theme of your academic studies**?

- Arts
- Humanities
- Social sciences
- Business
- Health and medicine
- Science and math
- Engineering
- Occupational
- Law
- Other

Are any of the **academic degrees** you've pursued or are currently pursuing (associate, masters, or advanced) **related to data science**?

- No, not at all
- Yes, somewhat
- Yes, very much related

Select the items that describe **your motivation for enrolling** in JHU Data Science Specialization? [You can select more than one choice]

- ☐ General interest in data science
- ☐ Relevant to job
- ☐ Relevant to school or degree program
- ☐ For personal growth and achievement
- ☐ For career change
- ☐ To meet new people
- ☐ To experience an online course
- ☐ To earn a certificate
- ☐ To improve my English skills
- ☐ Courses are offered by prestigious university/faculty

If you received a JHU Data Science Specialization **certificate**, in what contexts have you **mentioned** it? [select all that apply]

- ☐ During a job interview
- ☐ On my Linkedin profile
- ☐ On my resume
- ☐ To my colleagues
- ☐ To my managers/supervisors