Report - Causal Inference Exam

Michele Cucinella - 1140859

April 2025

1 Paper Summary

This paper aims to identify the causal effect of offering a training program in data processing and interpretation to students with low level of quantitative skills (usually enrolled in masters like law, political science, medicine ecc). Given the growing importance of data literacy in the labor market, young students usually lack of the fundamental skills of data analysis and interpretation, with consequences on the employability and remunerations of just-graduated students. The paper analyzes the effect of an experimental design consisting in a training program called "LEDA" offered to a sample of units exploiting Instrumental Variable identification strategy (IV) with heterogeneous effects. In order to participate, the students (only if enrolled in non-quantitative courses) had to send the application to LEDA's instructors, complete a Numeracy evaluation test based on OECD's PIAAC measures of data literacy, and then wait for a random selection of the winners of available seats, in order to create treated and control groups. The courses, lasting approximately 10 months, were offered to 62 of the 109 valid applicants with similar characteristics between treated and control groups (a balance check was performed before the treatment) and ended with a final examination, 6 months after the conclusion of the program. Furthermore, data from applicants' primary university career were collected to verify any kind of spillover effect. The results of the ITT and IV, having secured the relevance of the instrument (assignement to LEDA), showed positive but insignificant effect of treatment (enrollment to LEDA) on data literacy, which turns significant when we include an interaction between the program participation and numeracy. In fact, for students with numeracy 1 s.d. below the average level, the effect of program participation is an increase of 0.42 standard deviations in the outcome Y and small but insignificant for high data-experienced students, with no impact on primary university career of those enrolled. This means that the LEDA could act as an equalizer of quantitative skills among students with different data analysis background, with practically no slowdowns on the performances in their major degree program. This paper also stresses the importance of such results, highlighting the possible consequences of rise in expected remuneration of those who acquire data interpretation and manipulation competences in the labor market, predicting roughly a 6.5% increase in wages of program participants and a reduction of 5% of wage gap between STEM and non-STEM students.

1.1 Identification strategy

The identification strategy used in this study is the Instrumental Variable (IV) strategy. IV strategy aims to identify the Average Treatment Effect (in homogeneous effect framework) $\beta_{ATE} = E[Y_i(1) - Y_i(0)]$ in the model

$$Y_i = \alpha + \beta X_i + \epsilon$$

solving the endogeneity problem, caused by the correlation between the X and the composite error term ϵ (due to for example unobserved characteristics M inside the $\epsilon = \phi M + \pi$), using an instrumental variable Z correlated with the endogenous variable (Relevance assumption: $\gamma \neq 0$).

$$X_i = \alpha_d + \gamma Z_i + \mu$$

In addition, we need that $E[Z, \epsilon] = 0$ (Exclusion restriction assumption) and $E[Z, \mu] = 0$. The specific IV procedure used in this paper is the 2SLS (Two Stage Least Square), in which we extract the fitted values $\hat{X}$ from the First Stage equation and then regress the outcome of interest on these values, in order to exploit only the exogenous variation of the X variable:

$$\text{FIRST STAGE} : \hat{X} = \hat{\alpha_d} + \hat{\gamma} Z_i$$

$$\text{SECOND STAGE} : Y_i = \alpha + \beta \hat{X} + \epsilon$$

The resulting β coefficient, after algebraic manipulation, is the ratio between $COV(Y_i, \hat{X})$ and $VAR(\hat{X})$. In the particular framework of the paper, that is called "lottery design" and is similar to a Randomized Controll Trial with one-sided non compliance (meaning that only among treated or control group there could be non-compliance), the instrument is a random assignment (Z) to the LEDA program of a group of interested applicants and the treatment (the X) is participation to the program (Enrollment P). The author analyses an heterogeneous treatment effect framework, meaning that the treatment effect changes at different level of covariates and compliance classes (Imbens & Angrist 1994). In the Heterogeneous effect framework the author assumes

- *SUTVA*: Stable Unit Treatment Value Assumption of no spillover or general equilibrium effects.
- *Indipendence*: the instrument is assigned randomly regardless of potential outcomes $Y_i(1)$, $Y_i(0)$ and potential treatments $D_i(1)$ and $D_i(0)$. The balance check performed in the paper after the assignment and at the time of the final evaluation controls for possible self-selection of groups.
- *Exclusion restriction*: the Instrument affects the outcome only through the treatment, which is true given that assignment to LEDA is a random process
- *Relevance of the instrument*: the instrument is correlated with the treatment variable as table A3 and A4 show.
- *Monotonicity*: the instrument affect the treatment status always in the same direction ($D_i(1) - D_i(0) \geq 0$). This assumption is automatically guaranteed by the design of the experiment.

The Heterogeneous treatment framework usually divides the population in: Compliers, Alwaystakers, Nevertakers and Defiers. Since it is not possible to participate to LEDA if not selected in the random assignment, the experiment excludes the possibility to encounter Alwaystakers and Defiers, and we are left with two remaining groups, so Monotonicity is satisfied. When there are heterogeneous effects, we are only able to identify the average treatment effect on those whose treatment status change in response to the assignment (either Compliers or Defiers). Therefore, in our case, we are searching for the local average treatment effect (LATE) for Compliers. In particular, we look for:

$$\delta_{LATE} = \frac{E[Y_i(1, D(1)) - Y_i(0, D_i(0)]}{E[D_i(1) - D_i(0)]} = E[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1]$$

The proportion of Compliers over the entire population is identified in the First Stage regression, where we see the correlation between treatment and instrument assignment. Given the results of the A3/A4 Tables, Compliers represent roughly 88% of the population with respect to Nevertakers. Given that there are no Alwaystakers, the LATE estimated is actually the average treatment effect on the treated (the "Bloom Result", which generalizes the effects to the treated/non-treated population whenever we don't have Alwaystakers/Nevertakers).

The heterogeneity referred to the variation in treatment effect across different levels of predictors is related to the change in "Numeracy", meaning that the effect of the program is different depending on the pre-treatment degree of students' data literacy. Therefore, the author specified also a model introducing an interaction between A_i dummy (assignment to the Program) and N_i, which refers to the Numeracy level of the observation. For this reason, we are actually working in an environment with multiple endogenous variable and multiple first stage equations. According to Angrist and Imbens (1995), it is possible to calculate the average causal effect by making a weighted average of all the X's specific LATE, weighting proportionally to the average conditional variance of the population frst-stage fitted values, at each value of the X.

Thus, the author exploits the instrumental variable strategy consisting in the random assignment to account for possible unobserved characteristics correlated with participation to the program (M) and heterogeneity in response to the assignment (different compliance classes and changes for different level of covariates). The 2SLS estimator is used to compute the (multiple) local average treatment effects on compliers.

2 Paper Strengths

Experimental Design: The possibility of the author to exploit the result of a real experiment, consisting in a program offered to a sample of students observing their pre-treatment characteristics and assessing the evolution of their quantitative skills across time is a considerable advantage of the paper. Experimental designs are usually expensive and relying on non-experimental design could sometimes bring to biased results given wrong specification or identification strategy. In fact, the experimental design gives the possibility to apply randomization in treatment assignment, which is needed to avoid selection biases or guarantee independence assumption. In addition, the author can exploit information on students' characteristic to perform balance checks and data collected by the PIAAC starting test in order to identify different effects by numeracy level and use the final examination to keep track of data literacy evolution both for treated and control group.

Powerfulness of Identification Strategy: Another advantage of this study is the correct identification of the actual causal effect of program assignment/participation on students' data literacy. With the useful data that the experimental design guarantees, we can exactly isolate the improvements in data literacy after assignment to LEDA controlling for pre-treatment existing knowledge in quantitative analysis. With the heterogeneous effect framework, the author takes into account how much a group of students with a specific level of numeracy reacts to the assignment and the size of the LATE effect in that specific group. Thus, at least among the small sample of applicants, we have a clear idea of the causal effect of being assigned to LEDA on data literacy.

Importance of Data literacy: The importance of the claimed results could be sustained by the evidence cited by the author of an effective increase in wages given higher data literacy, as well as satisfying a civil duty and allowing for more informed decision-making of individuals. Therefore, showing the effectiveness of programs such as LEDA could be helpful to inform Universities on how to widen their curricular offer in order to adapt to the evolution of the labor market. However, as described in the weaknesses part, the link between attending the program and wages is probably too arbitrary considering that further research is needed to support this causal effect.

3 Paper Weaknesses

Limited External validity: The instrumental variable assumptions, given the randomization applied to valid applicants and the specific design of the experiment, are respected. However, this is completely true only if we consider the program from the assignment to treatment stage onward. The application to LEDA process first collected those interested in the program and then performed the randomization in order to select the treated and the controls, but only after the units self-selected in the pre-treatment group, generating a distortion that we could name "Pre-Selection-Bias". The 150 applicants could have been particularly interested in handling data (and therefore could have already developed some data analysis skills) even if actually enrolled in masters of other fields like humanities and social sciences. These characteristics could have distorted the identification of the actual average treatment effect of attending to the LEDA program of a random individual selected from the population, showing instead the **local** average effect on those who already were susceptible to demonstrate improvements in data literacy. The random assignment carried out in the second stage (after collecting the applicants) *does* solve the identification problem given by the self-selection bias, since we have an assignment rule that doesn't depend on the "Education & Skills Online Assessment" or other characteristics (given that there is voluntary decision to enroll after being assigned to treatment, we have to look at IV estimates given by the Wald ratio in order to get rid of self-selection *among units with* $Z_i = 1$ *that are already interested/proficient*), but the validity of the analysis is limited to this group of voluntary applicants, which might have higher data skills performances (In some sense, the initial PIAAC test divided the sample in low and high level of Numeracy groups, yet among those that probably had on average a higher PIAAC score test than any other randomly chosen student). Therefore, the analysis of the effect on data literacy of the assignment/participation to the program was, in my opinion, correctly identified but only in the second stage, because of that "Pre-Selection Bias" effect in the initial stage of collecting applicants that limited the external validity of the results. The improvement in data literacy could

also be higher due unobserved ability, if we assume that people interested in these topics are usually talented students.

Small sample size: The experiment is constrained by the limited sample size of the program participation offerings and, consequently, by the precision and the generalizability of the estimates. 109 applications is a considerably small sample on which we could hardly make strong conclusions (considering that this group is divided further into treated and control groups, with 48 enrolled and 47 not selected). This pitfall is likely to undermine, therefore, internal and external validity of the results.

Progression in primary career: The same discussion made on "Pre-Selection-Bias" effect on data literacy outcome could be valid also on the analysis of the influence of participating to the program on major academic performance. Students selected for LEDA program were, as already explained, among those who voluntarily applied to the selection process. Therefore, nothing prevents us from guessing that these applicants are already proficient students that would have performed well in their primary career anyway, independently from attending to LEDA program or not. Even if the comparison is made with respect to the control group, we are not aware of how different is the group of 109 valid applicants compared to the average student in terms of academic performance. Again, the analysis could have considered a control group randomly selected from the population (performing a balance check of students characteristics relatively to the selected applicants) in order to avoid, in the identification of the effect of LEDA on major grades, the "interest in data literacy"/"ability" selection-bias (when these two elements are correlated with higher average academic performance).

Scarce measure of participation: In the baseline and preferred specification, the only measure of course participation is "enrollment" (P) in the program, which is, however, an uninformative measure of compliance to the assignment and conveys no clear information on participants' actual attendance to LEDA courses. Students formally enrolled could have decided to stop following the courses, maybe particularly those already experienced and interested in quantitative analysis, yet taking the final examination test (probably incentivized by the 20 euro flat incentive. This could be the reason why students with high pre-treatment numeracy ends up with not significant increase in data literacy outcomes). When we use the specification included in Appendix B, where treatment measure is substituted with completion of at least one exam by June 2023, we have a clearer measure of attendance, even if still with some limitations (as said, they could have decided to give only the exams given their pre-existing experience in data analysis, or they might have stopped following the courses after the first exam). The fact that the first stage results of this alternative participation measure are substantially lower (it's a huge jump from 0.87 to 0.42) with respect to just "enrollment" support the idea of this pattern of actual fewer applicants that really attended the courses after selection. This lower participation rate, in turn, could be another signal of pre-existing knowledge of data analysis by students that applied to LEDA, that made following the courses useless. Thus, not having completely reliable data on actual participation, the causal relation between LEDA and increased data literacy slightly weakens.

The back-of-envelope calculation: Even if it could be an interesting consideration, it seems a bit too much ambitious to state that the effect of the program has a specific effect of 6.3% increase in wages given the result of Hanusheck et al (2015) and a reduction in wage gap between STEM and non-STEM graduates by 5% points. The first result is obtained by assuming that the numbers of Hanusheck regarding the effect of numeracy on wages are the same if we substitute the measure of quantitative skills with data literacy (assessed with a test constructed for the sake of this specific experiment). The latter is another too vague statement that lacks of clarity and rigorous calculation.

4 Suggested Improvements

Solution to Pre-Selection-Bias: In order to avoid the pre-selection of students among those particularly interested in data analysis and evaluate more precisely the effect of the program on the average non-quantitative students, the author should have sent the offer to apply to a group of randomly chosen units enrolled in those master fields. After that, carry out the same randomization performed in the second stage of the actual LEDA to split the sample in treated and control groups (performing a balance check

of students characteristics among the two clusters), thus allowing for the causal identification of program's effects on data literacy and primary academic performance. Otherwise, it should have simply selected the control group among the population of non quantitative students' population, instead of picking from the sample of voluntary applicants to LEDA. This suggestion holds both for the identification of the effect on data literacy and on primary university career of the applicants.

Increase sample size: As already mentioned in the weakness part, the experiment should have collected more students and, maybe, should have been carried out in more than only one University, because results could not be representative of the larger population of non quantitative master students (for example, the University in which LEDA took place could be a top-quality campus, so the final examination scores could have been different if taken by students from other less prestigious universities)

Data on actual attendance: As explained in the Weaknesses part, the LEDA program could have represented a convenient choice for students already experienced in data analysis (maybe because they had previously attended to other quantitative classes) to enhance their curriculum by showing the participation to extra-curricular courses. Even if we control for the numeracy level at the start of the program, the majority of the few available seats could have been occupied by these students, of whom we could suspect very scarce actual participation to the LEDA. Thus, the estimates of our LATEs could be biased because we are assuming that some people attended the courses while actually they have not (the "P" variable fails to detect real participation of individuals to the program). Instead of just "enrollment" (P) or, as described in Appendix B, the number of students that completed at least one exam, it could be more informative to track the actual class attendance of students to LEDA courses. For example, the final test evaluation could have asked to the students how many courses they have attended or what was the percentage of LEDA program they have completed. This could represent a clearer image of real course participation, detecting the relevance of the instrument and isolating the causal relation between attending the program and increased data literacy.

The back-of-envelope calculation: Instead of the calculation of percentage changes in wage or wage gap reduction between STEM and non-STEM graduates, mixing the results coming from different studies, the author should have performed a more rigorous analysis, exploiting the same measures of data literacy to make comparisons (checking if the final test structure perfectly matches with the "numeracy" measure used by Hanusheck and, if not, search for paper on wage effects where data literacy assessment exactly overlaps with the one used in this framework) or, if possible, modify the experiment including a long-term survey of wage gap between individuals that attended LEDA and those not selected. This could give a clear image of the effect of the program on remuneration.

Essential Heterogeneity: There are cases in which, instead of self-selection in levels, meaning that units are treated depending on the value of the covariates X, there could be self-selection in gain $\beta = Y(1) - Y(0)$. This type of context is called essential heterogeneity and happens when people, depending on the value of the expected gain from the treatment, redistribute themselves among treated and controls such that $COV(\beta, D) \neq 0$. Given the argument of this paper, namely that it possible that might have applied to the LEDA particular students whose expected data literacy gain after treatment was higher (maybe due to more interests in such topics), we could find ourselves in a setting of essential heterogeneities, in other words, with heterogeneity in response to the treatment depending on perceived gain. According to Heckman et al. (2006), there are particular tests for essential heterogeneity that could have been used to verify wether, in this case, the author should have increased the complexity of model specification to account for these kind of heterogeneity. In particular, given that in absence of essential heterogeneity the outcomes Y are linear in $P(Z)$ (the propensity score, namely the probability to choose the treatment given that you received the assignment Z), it could have been useful to test this linearity in the conditional expectation of Y given $P(Z)$ in the sample of students that received the offer. However, this test that warrants additional complexity could be limited by the small sample size available, both for the calculation of the propensity score and for the linearity test, since we have only the data for 62 students, of which 48 decided to enroll.

5 Conclusion

This paper succeed to identify the causal effect of participation to the LEDA program to a sample of students randomly selected from a group of applicants. The result are reasonably internally valid, given the correct IV specification, though the small sample size and scarce measure of actual treatment could give imprecise results. If the research question is limited at analyzing the effect of course participation to a small fraction of volunteers, the paper succeeds to bring empirical evidence of the effect of this program.

References

(Angrist & Imbens 1995) Angrist, J. D., & Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. Journal of the American Statistical Association, 90(430), 431–442. https://doi.org/10.2307/2291054

(Imbens & Angrist 1994) Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica, 62(2), 467–475. https://doi.org/10.2307/2951620

(Hanushek et al. 2015) Hanushek, E.A., Schwerdt, G., Wiederhold, S., Woessmann, L., (2015). Returns to skills around the world: Evidence from PIAAC. European Economic Review 73, 103–130.

(Heckman et al. 2006) Heckman, J. J., Urzua, S., Vytlacil, E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. The Review of Economics and Statistics, 88(3), 389–432. http://www.jstor.org/stable/40043006