Jul, 2025
GRINS DISCUSSION PAPER SERIES DP N° 03/2025 ISSN 3035-5567

GRINS

The consequences of promoting data literacy among graduate students

DP N° 03/2025

Authors:
Margherita Fort, Annalisa Loviglio, Susanna Tinti

Jul, 2025
GRINS DISCUSSION PAPER SERIES DP N° 03/2025 **ISSN 3035-5567**

The consequences of promoting data literacy among graduate students

Margherita Fort, Annalisa Loviglio, Susanna Tinti

KEYWORDS

data literacy | minor | tertiary education | human capital formation

JEL CODE

I20, J24

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Author(s): Margherita Fort, Annalisa Loviglio, Susanna Tinti. Title: The consequences of promoting data literacy among graduate students. Publication Date: 2025.

We study the impact of a program designed to enhance data literacy on graduate students' skills and academic outcomes in a large Italian university. The program (i.e. a minor) targets students who are expected to have weak quantitative competences and offers 120-hours training focused on improving the ability to interpret and process data, in addition to the regular courses of the master program in which students are enrolled (i.e. their major). The admission process to the minor is characterized by rationing, resolved by random assignment of available slots to applicants. Exploiting the resulting exogenous variation for identification, we find that the program largely improved digital literacy of participants with low pre-treatment levels of numeracy. Despite the additional effort required by the program, we can rule out any slowdown in the progress of the academic career in the major master program of participating students.
Keywords: data literacy, minor, tertiary education, human capital formation
JEL-Codes: I20, J24

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
GRINS FOUNDATION

GRINS FOUNDATION

Discussion Paper Series

The consequences of promoting data literacy among graduate students

Discussion paper n. 40/2025

Margherita Fort (a), Annalisa Loviglio (a), Susanna Tinti (a):

(a) University of Bologna

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
GRINS FOUNDATION

ISSN 3035-5567

The consequences of promoting data literacy among graduate students

Margherita Fort (a), Annalisa Loviglio (a), Susanna Tinti (a): (a) University of Bologna

Discussion Paper DP nº 40/2025

We study the impact of a program designed to enhance data literacy on graduate students' skills and academic outcomes in a large Italian university. The program (i.e. a minor) targets students who are expected to have weak quantitative competences and offers 120-hours training focused on improving the ability to interpret and process data, in addition to the regular courses of the master program in which students are enrolled (i.e. their major). The admission process to the minor is characterized by rationing, resolved by random assignment of available slots to applicants. Exploiting the resulting exogenous variation for identification, we find that the program largely improved digital literacy of participants with low pre-treatment levels of numeracy. Despite the additional effort required by the program, we can rule out any slowdown in the progress of the academic career in the major master program of participating students.

Keywords: data literacy, minor, tertiary education, human capital formation

JEL-Codes: I20, J24

The consequences of promoting data literacy among graduate students*

Margherita Fort[*], Annalisa Loviglio[†], Susanna Tinti[‡]

March 25, 2025

Abstract

We study the impact of a program designed to enhance data literacy on graduate students' skills and academic outcomes in a large Italian university. The program (i.e. a *minor*) targets students who are expected to have weak quantitative competences and offers 120-hours training focused on improving the ability to interpret and process data, in addition to the regular courses of the master program in which students are enrolled (i.e. their *major*). The admission process to the *minor* is characterized by rationing, resolved by random assignment of available slots to applicants. Exploiting the resulting exogenous variation for identification, we find that the program largely improved digital literacy of participants with low pre-treatment levels of numeracy. Despite the additional effort required by the program, we can rule out any slowdown in the progress of the academic career in the *major* master program of participating students.

JEL-Codes: I20, J24

Keywords: data literacy, minor, tertiary education, human capital formation

*In Alphabetic Order: *Conceptualization*: M.Fort, A. Loviglio; *Field work administration and data management*: M. Fort., A. Loviglio, S. Tinti; *Data analysis*: M. Fort, A. Loviglio, S. Tinti; *Draft preparation, review and editing*: M. Fort, A.Loviglio, S. Tinti; *Funding acquisition*: M.Fort, A. Loviglio. Funding from MIUR to the Department of Economics (Department of Excellence grant - 2018-2022) is gratefully acknowledged. We gratefully thank A. Saia who contributed to the design of the ad-hoc survey on data literacy and offered funds to support the data collection. S. Tinti gratefully acknowledges financial support by the EU - NextGenerationEU with funds made available by National Recovery and Resilience Plan (NRRP) Mission 4, Component 1, Investment 4.1 (MD 351/2022) – Public Administration (PhD fellowship 38-412-03-DOT22K2552-43 CUP J33C22001850002) . A. Loviglio and M. Fort gratefully acknowledge funding by the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP J33C22002910001). We are grateful to the board of the LEDA program and to the data warehouse of the University that hosted the *minor* program for their support. All authors have read and agreed to the published version of the manuscript. We thank participants of the Public Governance Management and Policy PhD Forum and the ESPANET Conference 2024, specifically L. Bonaccini, N. Montanari, G. Pignataro and L. Vergolini for useful comments and remarks on an earlier version of the project. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

1 Introduction

As the world has become more complex and data oriented (Carmi and Yates, 2020; Fontichiaro and Oehrli, 2016; Wolff et al., 2016), the need to make citizens able to understand, interpret and manage data has gained relevance among both policy makers and researchers. These competencies fall under the broad umbrella of *data literacy*. Data literacy can be defined as the ability to search for, read, understand, interpret and communicate data, extracting meaningful information to support decision-making in everyday life.[1] Understanding and managing data are fundamental for everyday tasks, while a lack of such skills expose individuals to various risks – personal, social, physical and financial – while also limiting their ability to be proactive citizens (Carmi and Yates, 2020). From interpreting news conveyed through graphs, info-graphics, and statistical summaries to performing basic financial tasks, data literacy plays a crucial role. It may be framed as a lever for social justice and equality (Elisa Raffaghelli, 2020), an increasingly essential requirement for the labor market (Chise et al., 2021; Fayer et al., 2017; Windisch, 2015), and a necessarily skill for navigating modern life – including, for instance, online banking, e-commerce and the management of bureaucratic processes (Hanushek et al., 2015).

Despite the widely acknowledge importance of data literacy, proficiency in this area remains notably low in Italy. While no standardized measures of data literacy currently exists, related skills, such as numeracy – as measured by the Programme for the International Assessment of Adult Competencies (PIAAC) – can serve as a useful proxy.[2] The

*University of Bologna, Dept. of Economics (Bologna, Italy), CEPR, CESifo, IZA.

†University of Bologna, Dept. of Economics (Bologna, Italy), IZA.

‡University of Bologna, Dept. of Economics (Bologna, Italy), Ministry of Economics and Finance.

[1]A range of alternative albeit similar definitions of data literacy can be found in the literature. In 2017, the European Commission introduced its own definition – alongside the concept of information literacy – describing it as the ability to search, read, and interpret data across various daily and academic communication contexts. The OECD, building on the definition proposed by Carlson et al. (2011), defines data literacy as "the ability to derive meaningful information from data, the ability to read, work with, analyze and argue with data, and understand what data mean, including how to read charts appropriately, draw correct conclusions from data, and recognize when data are being used in misleading or inappropriate ways". Wolff et al. (2016) analyze different definitions of data literacy and propose to describe data literacy as the ability to ask and answer real-world questions from large and small data sets through an inquiry process that considers ethical data use. This includes skills such as selecting, cleaning, analyzing, visualizing, critiquing, and interpreting data, as well as communicating data-driven narratives and incorporating data into a design process.

[2]According to the OECD, numeracy is defined as the ability to access, use, interpret, and communicate mathematical information and concepts to solve and manage problems in various contexts. This closely

PIAAC results indicate that Italy scores below the OECD average, emphasizing the urgent need for improvement in this domain (Rouet et al., 2021).

This study presents evidence on the effectiveness of an interdisciplinary course launched by a large Italian University to enhance data literacy skills among graduate students who may not otherwise receive such training within their standard curriculum. The program is offered at no cost to all master's students enrolled at the host University, and applications exceed available slots.

Participants have the opportunity to attend a *minor* consisting of four 30-hours courses focused on data collection, management, and interpretation. Since the program has no prerequisites, admission is determined by random assignment, which we exploit to identify its causal effects. Moreover, we leverage pre-treatment numeracy levels to study heterogeneous effects across the skill distribution. Our results suggest that most students may benefit from the program, with the largest improvements in data literacy competencies observed among those who began with lower numeracy skills. For instance, our estimates show that a participant with a numeracy level one standard deviation (s.d.) below the sample average achieves a data literacy score nearly 0.5 s.d. higher on an assessment administered 6 months after the end of classes, compared to an otherwise identical student not exposed to the program. Importantly, these gains do not come at the expense of the students' primary university career. In fact, the additional coursework does not slow down the completion of their *major* degree, as participants pass at least as many exams as their peers in the control group.

As highlighted by PIAAC results, non-STEM graduates have lower numeracy than STEM graduates, and may be at a disadvantage in the labor market.[3] Our findings suggest that university students who do not choose STEM majors can benefit from minor programs targeting transversal skills, such as data literacy. This additional training may help them better align their competencies with the demands of a rapidly evolving labor market (Ghodoosi et al., 2024), ultimately improving their employment prospects.

Finally, our study contributes to the broader debate about majors and minors choices, as well as admission policies in tertiary education (e.g., Bordon and Fu, 2015). In Italy,

aligns with the broader concept of data literacy (Kankaraš et al., 2016).

[3]According to Table A.2.6 (N) of the Annex available online at `https://stat.link/eb8dxq`, non-STEM graduates, on average, have numeracy levels nearly 19 points lower than STEM graduates (0.32 s.d.).

as in most European countries, students enroll in specialized programs, with relatively inflexible curricula. By contrast, in countries such as the U.S. and Canada, specialization is postponed and a greater flexibility in course selection is allowed. Our findings suggest that offering students the opportunity to complement their major with minor programs targeting diverse skill sets may be beneficial.

The paper is structured as follows: Section 2 describes the institutional setting and data, providing information about the program design and selection procedure, and it illustrates the data used in the paper. Section 3 explains the empirical strategy adopted and illustrates the main results. Section 4 concludes.

2 Institutional setting and data

During the 2021-2022 academic year, a large Italian University launched a series of educational programs targeting students enrolled in master's degrees. These programs offered blocks of four master's-level courses and allowed students to obtain formal qualifications in addition to their primary master's degree. In what follows, we refer to these programs as *minor*. The rational behind these programs is to introduce thematic courses to enrich students' primary university career (their *major*) with interdisciplinary competencies, which are valuable for both further studies and future employment. Designed to address relevant contemporary themes through an interdisciplinary approach, minor programs integrate conventional teaching with innovative methodologies, incorporating multimedia materials and group projects into lectures and seminars.

These programs were open to students already enrolled in a two-year master's degree at the host University and were offered free of charges.[4] All educational activities were conducted in person. In order to obtain the final certification, students were required to complete four courses and pass the corresponding exams.

Within this institutional framework, we focus on a specific *minor* titled *Learning from data* (LEDA), explicitly designed to promote *data literacy*. This program offers an interdisciplinary curriculum aimed at fostering a culture of knowledge centered on data comprehension and communication. The focal theme revolves around the concept

[4]Students enrolled in three-year bachelor's degrees were not eligible and could not apply. The University also offers a small number of 5 or 6-year degrees (mainly in the medical field), whose students were eligible provided that they were in their fourth year or beyond.

of data literacy, emphasizing the importance of interpreting data, extracting meaningful information, and develop basic data management skills as a valuable addition to students' academic profiles.

The program combines 4 courses offered by different University Departments (Computer Science and Engineering, Statistical Sciences, Management, and Economics) over approximately 10 months. Two courses began in mid-February 2022, while the remaining two started in September 2022.[5] Each course offers a field-specific perspective within the common goal of introducing students to data literacy and enhancing their analytical skills, particularly for those whose academic backgrounds includes minimal exposure to quantitative subjects. In line with this objective, only students enrolled in master's degrees outside the four contributing departments were eligible for LEDA.

Application and Take-up Application opened in December 2021. All students interested in LEDA were required to complete the "Education & Skills Online Assessment" as part of the application process. This questionnaire is designed to provide individual-level results linked to the OECD's *Survey of Adult Skills (PIAAC)* measures of literacy and numeracy.[6] Aggregate statistics on performance in the assessment were shared with LEDA's governance and instructors to document the heterogeneous background of the class, while students did not receive any feedback.

The program received 150 applications, of which 109 were valid, for 50 available seats.[7] All applicants were informed that, due to limited capacity, admission would be determined by random assignment, stratified by five areas: humanities, social sciences, technology, science, and medicine. 51% of valid applications came from students in humanities programs and 30% were from students in social sciences (Law, Sociology or Political Sciences), suggesting a much higher interest in the program among students pursuing non-quantitative majors.

Applicants offered a seat had a few days to enroll. In the event of withdrawal, the

[5]Specifically, the courses are: i) *Extracting, integrating and mining from complex sources* (Computer Science and engineering department); ii) *Describing phenomena and controlling uncertainty* (Statistical Sciences department); iii) *Managing data to support business activities* (Management Department); and iv) *Data to inform political and social choices* (Economics department).

[6]Further information is available at `https://www.oecd.org/en/about/programmes/piaac/education-and-skills-online-assessment.html`.

[7]Among the 150 applicants, 24 applicants were deemed ineligible for the program, while 17 failed to complete the entry test in due time and were thus excluded.

slot was offered to the next candidate on the waiting list within the same academic area. Overall, 62 students were offered admission, and 48 choose to enroll, while 47 were not selected.[8] Unfortunately, the only available participation measure is exam completion: as of June 2023, 23 out of 48 students had passed at least one exam.

Endline Assessment Information on students' data literacy competencies was collected through an assessment specifically designed for the LEDA minor. The questionnaire was administered in June 2023, roughly 6 months after the courses ended, to capture long-lasting effects of the program rather than temporary improvements and to allow students to have at least one opportunity to take the exam at the end of each class. Students were invited to participate to the endline survey and were offered a flat incentive of 20 euros.[9] A total of 85 students completed the questionnaire, including 48 of those who were offered a slot in LEDA (77%) and 37 of those who did not have the chance to enroll (79%).

The assessment aimed to evaluate students' proficiency in applying logical-mathematical reasoning to real-world problems, as well as their ability to understand and interpret tables and graphs. It was based on GRE and GMAT validated tests and included 15 questions to be solved in approximately 40 minutes. Our main outcome of interest is the total number of correct answers.

Progression in the main career We complemented application and survey data with administrative data on students' performance in their major. Specifically, we retrieved from the host University archives the number of exams passed between March 2022 (after the start of LEDA coursework) and April 2023.

Descriptive statistics Table A1 describes the sample of 109 applicants and one of the endline respondents used in the paper, grouped into Non-STEM (humanities and social sciences) vs STEM (technology, science, and medicine) disciplinary areas. We remind

[8]Two seats remained vacant due to late withdrawals, which prevented further enrollment from the waiting list.

[9]The incentive scheme included a small penalty of 5 euros for students who took more than ten days to complete it since invitation, however only 3 students received the reduced amount of 15 euros. Out of the 109 students with valid application, 4 students did not give consent to be contacted at the end of the LEDA program and were thus not invited to complete the endline survey (two of them were offered a slot to participate to the program and two were not offered a slot).

readers that, given eligibility criteria illustrate before, majors in Economics, Management, Statics and Computer Science and Engineering are excluded. The vast majority of applicants comes from Non-STEM fields and about 70% of applicants is female.[10] Almost 50% of non-STEM applicants and 30% of STEM-applicants are enrolled in the first year of their master's program. As expected, numeracy levels are lower for the Non-STEM group ($\approx$ 13 points, roughly 0.5 of a standard deviation of the corresponding score in this sample), while literacy levels are more similar ($\approx$ 6 points higher for Non-STEM fields applicants, roughly 0.15 of a standard deviation of the corresponding score).

3 Empirical strategy and findings

We aim to assess the causal effect of the LEDA minor program on students' data literacy competencies. Additionally, we investigate potential spillover effects that attending extra courses may have on students' primary university careers. Specifically, we examine whether the additional workload has any negative impact on their performance in their major.

We exploit the random variation introduced by the application process for identification: applicants were offered a slot to enroll in LEDA ("assigned") or not ("not assigned" or "controls") based on their position in a randomly sorted list. Panel a) of Table A2 compares pre-treatment characteristics between these two groups, showing no significant differences. This confirms that the two groups are well balanced in terms of gender, year of enrollment in their major, and baseline literacy and numeracy. Panel b) replicates the analysis for the subsample of students who completed the endline questionnaire, confirming that assigned and control students did not self-select differently when choosing to participate in the assessment.

Specification choices We first estimate the *intention-to-treat* (ITT) effects, that is, the effect of being offered a slot to enroll in LEDA on digital literacy and academic performance in the primary career. Specifically, we estimate the following model:

$$Y_i = \alpha A_i + \gamma N_i + \mu_i + \epsilon_i \tag{1}$$

[10]There is some indication there female applicants are more likely to answer endline survey. However, there is no evidence of non random attrition along any observed covariate.

where Y_i represents the outcome of interest for student i, A_i equals 1 if the student was offered a slot, N_i denotes the pre-treatment level of numeracy, μ_i is the area fixed effect, and ϵ is an error term. Furthermore, in our preferred specification, we allow for heterogeneous effects based on pre-treatment numeracy levels by introducing an interaction term:

$$Y_i = \alpha' A_i + \beta' A_i \times N_i + \gamma' N_i + \nu_i + \varepsilon_i, \tag{2}$$

Next, we estimate the effect of LEDA using a parametric *instrumental variable* (IV) approach, where our baseline estimates consider a binary indicator for actual program enrollment.[11] The indicator is interacted with the pre-treatment level of numeracy in the specification allowing for heterogeneous effects by numeracy. Not surprisingly, given the high take-up rate, the instrument is not-weak, as confirmed in Table A3.[12] Moreover, Table A4 shows that sufficient independent sources of variation are detected for the model allowing for heterogeneous effects.

One might expect larger effects of the program would correspond to intense participation rather than mere enrollment. Appendix B explores an alternative take-up definition, based on actual participation. Since data on attendance are not available, we use the more restrictive binary indicator of having taken at least one of the four LEDA exams by the beginning of June 2023, before the endline assessment.

Effects of assignment and participation in LEDA on data literacy Figure 1 summarizes the main findings of our analysis on data literacy. It displays the marginal effects of the LEDA program according to the baseline specification (dashed red line) and to our preferred specification, which allows for heterogeneity by pre-treatment numeracy levels (solid blue line). The left panel presents results for the ITT specifications, while the right panel shows results for the IV ones. Corresponding estimates can be found in Table A5 in the Appendix.

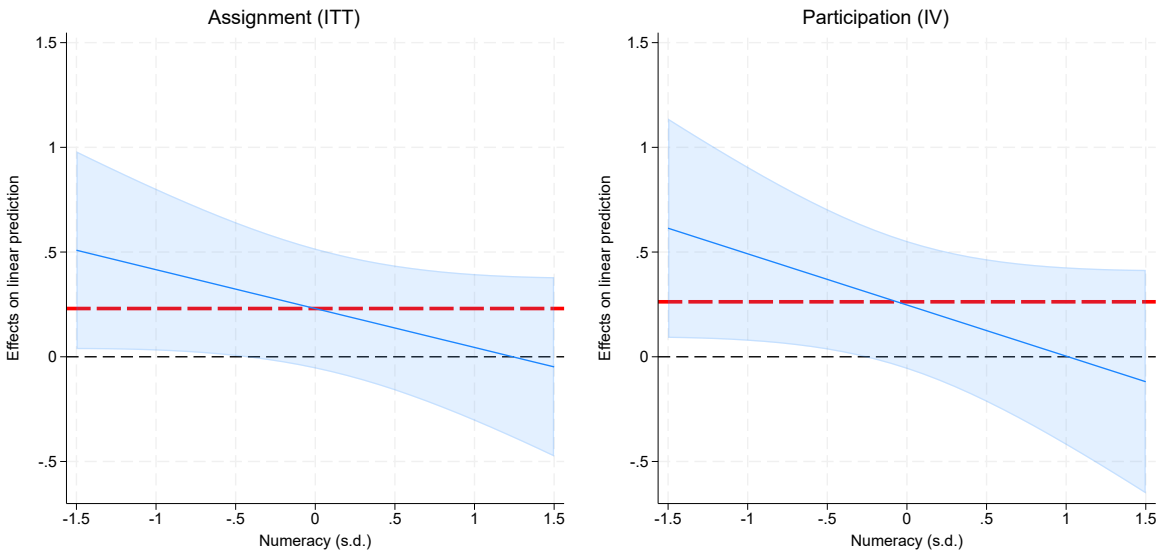
[11] We estimate the following model, using two-stage least squares:

$$Y_i = aT_i + cN_i + m_i + e_i \tag{3}$$
$$T_i = pA_i + qN_i + r_i + u_i, \tag{4}$$

where T_i equals 1 if the student enrolled in LEDA. The interaction term $A_i \times N_i$ is included as second instrument in the specification that accounts for heterogeneous effects by numeracy.

[12] The first-stage estimate of the probability of enrollment is 87.7%. The estimate is not sensitive to the omission of the control for pre-treatment levels of numeracy (see Table A3 in the Appendix).

Figure 1: Effect of the LEDA program on data literacy by baseline numeracy



Note: The figure displays estimates of the marginal effects of being assigned to LEDA (left panel) or enrolling in LEDA (right panel), based on the ITT (left panel) or IV (right panel) estimates or the model that allows for heterogeneity by pre-treatment levels of numeracy reported in the last two columns of Table A5 in the Appendix. Each estimate is complemented with a 90% confidence interval based on the asymptotic distribution of corresponding estimators and estimates of parameters and standard errors from the same table. The horizontal red dashed lines reports the marginal effects estimated through the models that do not allow for heterogeneity (first columns of Table A5).

Results suggest an average increase of data literacy by more than 0.2 standard deviations (0.23 s.d. in the ITT specification and 0.26 in the IV specification), though these estimates are not statistically significant.[13] However, there is evidence of heterogeneity along pre-treatment numeracy levels, with students with lower numeracy benefiting the most from the program. We detect statistically significant and large effects for individuals with below average numeracy: for instance, the ITT marginal effect of LEDA for students with numeracy 1 s.d. below average is 0.42.[14] Conversely, we find small and non-statistically significant effect for individuals with above average numeracy.

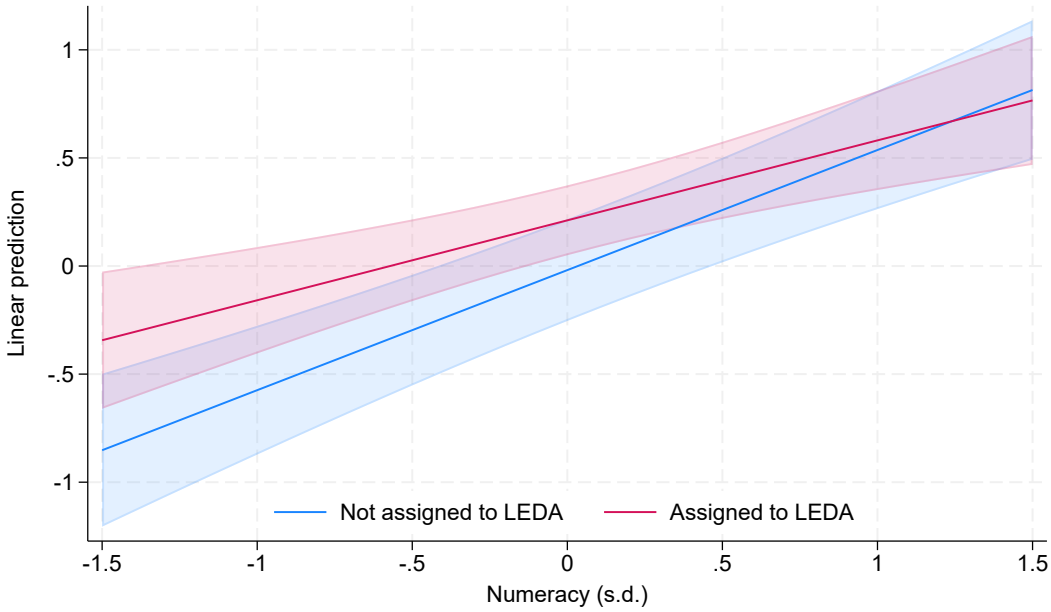
We find reassuring that the program benefited students with low pre-treatment levels of numeracy and, at the very least, did not harm those with high pre-treatment numeracy.

IV estimates further confirm positive and sizable effects on data literacy for individuals

[13] P-values are 0.19 for the ITT specification and 0.17 for the IV specification.

[14] The one-sided p-value for H_1 : marginal effect for individuals with pre-treatment levels of numeracy 1 s.d. below average > 0 is 0.037.

Figure 2: Predicted level of data literacy by baseline numeracy



Note: The figure shows predictions of expected values of data literacy based on the ITT parameter estimates of the model that allows for heterogeneity by pre-treatment levels of numeracy reported in Table A5 in the Appendix. Each estimate is complemented with a 90% confidence interval based on the asymptotic distribution of corresponding estimators and estimates of parameters and standard errors from the same table.

who start with lower numeracy levels (almost 0.5 s.d. for those with numeracy 1 s.d. below the mean) and negligible effects for those with above average levels of numeracy.

Figure 2 highlights a positive correlation between data literacy and pre-treatment numeracy among both controls students and those assigned to LEDA. However, this correlation is weaker for the latter, indicating that the program may help individuals compensate for initially low numeracy levels.

Overall, our results suggest that LEDA has a positive or null effects for applicants, and show a large gain for those with lower initial numeracy. We interpret these findings as evidence that the program acts as an equalizer, reducing gaps between individuals who are better equipped in terms of quantitative skills and those with weaker initial endowments.

Effects of assignment and participation in LEDA on primary career While positively impacting data literacy, the program could also generate negative spillovers effects on students' primary degree progress by slowing down their academic advancement.

To assess this possibility, we examine the effect of LEDA on the number of exams passed after the program began in winter 2022. As reported in Table A6 in the Appendix, we do not detect any statistically significant effect. If anything, LEDA appears to have a small positive average effect on students' primary career, with larger – but still insignificant – gains for those with low pre-treatment numeracy.[15]

4 Concluding Remarks

This paper studies the causal effects of the LEDA data literacy program, offered as a complement to students' primary degree programs at a large Italian University. Rationing of available slots is resolved through random assignment of applicants to the program, allowing us to exploit a clean identification strategy to assess causal effects. Notably, given this design, the exogenous source of variation is independent from students' pre-treatment numeracy levels, which were measured using a standardized OECD assessment at baseline. This allows us to explore heterogeneity in the program's effects based on students' initial competencies.

We find that students with lower baseline numeracy skills benefit significantly from the program, experiencing substantial improvements in data literacy as measured by an assessment administered six months after course completion. In contrast, we detect small or negligible effects for students with above-average numeracy skills. Importantly, LEDA does not slow down students' progresses in their primary career. If anything, students with lower numeracy slightly increase the number of exams taken, but the effect is not statistically significant.

Given the limited precision of our estimates, these results should be interpreted with some caution, but they suggest that offering training on data literacy to adults - such as master's students at University - can yield substantial benefits and has the potential to reduce inequalities in data literacy. Consistently, our findings suggest that interventions targeting individuals with expected low levels of data literacy may be effective in improving these essential transferable skills.[16]

[15] On average control students passed 3.6 exams, with a standard deviation of 2.9. Thus, the estimated effects in Table A6 for students with 1 s.d. below-average numeracy correspond to improvements of 0.11 s.d. (ITT) and 0.13 s.d. (IV).

[16] This paper uses data collected up to June 2023. While we initially planned to extend the analysis with additional waves of data collection, institutional constraints prevented this. First, the LEDA program was

Beyond the many non-economic reasons to invest in the skills of both youth and adult skill – including the view that a minimum level of literacy and numeracy is a civil right and a prerequisite for full participation in a modern democracy (Vignoles, 2016), the literature has shown a strong association between numeracy skills and wages. Specifically, Hanushek et al. (2015) report that in Italy a one-standard-deviation increase in numeracy is associated with a 13.2% increase in hourly wages among prime-age workers. While direct estimates of the wage returns to data literacy skills are not available, assuming similar returns to those of numeracy, our estimates suggest that a training program targeting data literacy could lead to a 6.5% wage increase for individuals starting with low quantitative skills.[17]

Such returns could be particular relevant for graduates from non-STEM fields, who typically have lower quantitative skills, potentially contributing to narrow the wage gap between STEM and non-STEM graduates. According to the main survey of Italian graduates, five years after graduation the wage gap between Humanities and STEM graduates exceeds 22%.[18] The previous figure suggests that data literacy training for Humanities graduates could reduce this gap by 5 percentage points, closing almost one fourth of it. Although we cannot directly measure the effects of this training on wages, our back-of-the-envelope calculations indicate that returns are likely to be economically relevant. As a consequence, the evidence presented in this paper may encourage further investments in data literacy training and research aimed at understanding the labor market impact of programs that develop transversal skills such as data literacy.

temporarily discontinued for one year. Second, when it was recently re-instated, the selection process changed, and rationing was no longer resolved through randomization. As a consequence, the new edition of the program does not allow us to implement the same identification approach to increase statistical power.

[17]This back-of-the-envelope calculation multiplies 13.2% by 0.49, the IV estimate for students with pre-treatment numeracy 1 s.d. below average (from Table A5).

[18]Authors' computation using AlmaLaurea national averages. Data available at `www.almalaurea.it/i-dati/le-nostre-indagini/condizione-occupazionale-laureati`

References

Bordon, P., Fu, C., 2015. College-major choice to college-then-major choice. The Review of Economic Studies 82, 1247–1288.

Carlson, J., Fosmire, M., Miller, C., Nelson, M.S., 2011. Determining data information literacy needs: A study of students and research faculty. Libraries and the Academy 11, 629–657.

Carmi, E., Yates, S.J., 2020. What do digital inclusion and data literacy mean today? Internet Policy Review 9.

Chise, D., Fort, M., Monfardini, C., 2021. On the intergenerational transmission of stem education among graduate students. The BE Journal of Economic Analysis & Policy 21, 115–145.

Elisa Raffaghelli, J., 2020. Is data literacy a catalyst of social justice? A response from nine data literacy initiatives in higher education. Education Sciences 10, 233.

Fayer, S., Lacey, A., Watson, A., 2017. Stem occupations: Past, present, and future. Spotlight on Statistics 1, 1–35.

Fontichiaro, K., Oehrli, J.A., 2016. Why data literacy matters. Knowledge quest 44, 21–27.

Ghodoosi, B., Torrisi-Steele, G., West, T., Heidari, M., 2024. Perceptions of data literacy and data literacy education. Journal of Librarianship and Information Science .

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.

Kankaraš, M., Montt, G., Paccagnella, M., Quintini, G., Thorn, W., 2016. Skills matter: Further results from the survey of adult skills. oecd skills studies. OECD Publishing .

Rouet, J.F., Britt, M.A., Gabrielsen, E., Kaakinen, J., Richter, T., Lennon, M., 2021. The assessment frameworks for cycle 2 of the programme for the international assessment of adult competencies: Piaac cycle 2 assessment framework: Literacy. OECD Publishing .

Vignoles, A., 2016. What is the economic value of literacy and numeracy? IZA World of Labor .

Windisch, H.C., 2015. Adults with low literacy and numeracy skills: A literature review on policy intervention .

Wolff, A., Gooch, D., Montaner, J.J.C., Rashid, U., Kortuem, G., 2016. Creating an understanding of data literacy for a data-driven society. The Journal of Community Informatics 12.

Appendix

A Tables

This section reports additional tables relevant to the paper.

Table A1 describes the sample of 109 applicants and one of the endline respondents used in the paper, grouped into Non-STEM (humanities and social sciences) vs STEM (technology, science, and medicine) disciplinary areas. We stress that because of the eligibility criteria for the LEDA program, the STEM applicants do not include students enrolled in degrees in Computer Science and Engineering and the Non-STEM applicants do not include students enrolled in degrees in Economics, Management and Statistics.

Table A2 compares the pre-treatment characteristics of students assigned to LEDA and those not assigned, on the sample of applicants and on the sub-sample of applicants who completed the endline survey.

Tables A3 and A4 report the first-stage estimates for specification without or with heterogeneous effects by numeracy, respectively. There is no indication of weak instruments issues, and sufficient independent sources of variation are detected for the models allowing for heterogeneous effects.

Tables A5 show the ITT and IV estimates of participating in LEDA on data literacy using enrollment as proxy for participation. Table A6 report the effect on students' primary university career. All tables include a baseline specification and a specification were we allow for heterogeneous effects based on pre-treatment numeracy levels.

Tables with corresponding first-stage, ITT and IV estimates using an alternative proxy for LEDA participation (i.e. exam taking behaviour) are presented in Section B.

Table A1: Descriptive Statistics

	LEDA Applicants		Endline respondents	
	Non-STEM fields	STEM fields	Non-STEM fields	STEM fields
Female	0.682	0.667	0.721	0.706
	(0.468)	(0.483)	(0.452)	(0.470)
Literacy	332.727	329.048	333.971	328.235
	(38.139)	(37.935)	(39.517)	(40.502)
Numeracy	312.273	325.500	312.500	325.882
	(25.085)	(23.278)	(25.060)	(24.253)
1° year Master	0.477	0.333	0.485	0.294
	(0.502)	(0.483)	(0.503)	(0.470)
Observations	88	21	68	17

Note: The table reports descriptive statistics on pre-treatment characteristics of eligible students for the sample that includes all LEDA applicants with valid application (left panel) and the sub-sample who also completed the endline survey (right panel), by field of study. We stress that because of the eligibility criteria for the LEDA program, the STEM applicants do not include students enrolled in degrees in Computer Science and Engineering and the Non-STEM applicants do not include students enrolled in degrees in Economics, Management and Statistics. *Literacy* and *Numeracy* represent the PIAAC literacy and numeracy scores, in levels.

Table A2: Balance of pre-treatment characteristics by assignment status

	Panel (a) LEDA Applicants			Panel (b) Endline respondents		
	Not-ass.	Assigned	Diff.	Not-ass.	Assigned	Diff.
Female	0.681	0.677	-0.021	0.703	0.729	0.006
	(0.471)	(0.471)	(0.832)	(0.463)	(0.449)	(0.955)
Literacy	330.213	333.387	3.564	332.162	333.333	1.357
	(38.813)	(37.545)	(0.652)	(40.766)	(38.994)	(0.882)
Numeracy	312.766	316.452	2.142	315.135	315.208	-1.952
	(27.560)	(23.195)	(0.665)	(28.735)	(22.690)	(0.728)
1° year Master	0.404	0.484	0.082	0.378	0.500	0.139
	(0.496)	(0.504)	(0.409)	(0.492)	(0.505)	(0.218)
Observations	47	62	109	37	48	85

Note: The table reports average characteristics of eligible students by assignment status and tests whether differences are statistically significant running independent OLS regressions. Consistently with the randomization design, all regressions include area fixed effects. Robust standard errors in parentheses. *Literacy* and *Numeracy* represent the PIAAC literacy and numeracy scores, in levels.
Legend: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. No significant differences detected.

Table A3: Effect of being offered a slot for the LEDA program on participation (First Stage)

	No controls	Pre-treatment controls
	(1)	(2)
Assigned to LEDA	0.878***	0.877***
	(0.051)	(0.052)
Numeracy		-0.007
		(0.036)
Area fe	Yes	Yes
F-test	297.205	287.164
Adj. R-Square	0.699	0.696
Observations	85	85

Note: The table reports estimate of first stage parameters assessing the causal effect of being offered a slot for LEDA on participation to the program using linear probability model and OLS estimator for alternative specifications that differ for the control variables included. Our preferred specification is the one presented in column (2). IV estimates corresponding to specification in column (2) are presente in Table A5 (data literacy) and Table A6 (students' careers). Binary indicator for participation into LEDA is enrollment in the program (vs not enrollment). Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.
Legend: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4: Effect of being offered a slot for the LEDA program on participation (First Stage) - Models allowing for heterogenous effects by pre-treatment numeracy levels

	Enrollment (P)	
	P	P · Numeracy
Assigned LEDA	0.877***	0.054
	(0.0512)	(0.0554)
Assigned LEDA · Numeracy	-0.046	0.713***
	(0.0738)	(0.1156)
Numeracy	-0.007	-0.001
	(0.0323)	(0.0132)
Area fe	Yes	Yes
F-test	164.53	28.77
SW F-test	321.48	44.91
Observations	85	85

Note: The table reports estimate of first stage parameters assessing the causal effect of being offered a slot for LEDA on participation to the program using linear probability model and OLS estimator in a model were heterogeneous effects are allowed (see Tables A5 for corresponding IV estimates on data literacy and Table A6 for students' careers). Binary indicator for participation into LEDA is enrollment in the program (vs not enrollment) and it is denoted with *P*. Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA. *P* is the participation indicator, either based on enrollment or on exams. Legend: SW for Sanderson-Windmeijer multivariate F test of excluded instruments. The SW complement the F-test on joint significance of the instrument as they are tests for under-identification and weak identification of individual endogenous regressors in models with multiple endogenous regressors and instruments. They check for "sufficient" independent source of variation, *partialling-out* linear projections of the remaining endogenous regressors. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: Effect of the assignment and of participation to the LEDA program on data literacy- ITT and IV results. Measure of participation based on: enrollment

	Model parameters' estimates			
	Baseline specification		Preferred specification	
	ITT	IV	ITT	IV
Assigned to LEDA	0.230		0.230	
	(0.173)		(0.172)	
Assigned to LEDA · Numeracy			-0.186	
			(0.140)	
LEDA Participant		0.262		0.247
		(0.187)		(0.186)
LEDA Participant · Numeracy				-0.244
				(0.226)
Numeracy	0.470***	0.472***	0.555***	0.552***
	(0.071)	(0.084)	(0.097)	(0.111)
Area fe	Yes	Yes	Yes	Yes
Observations	85	85	85	85
Assigned to LEDA (ITT)				
@**Low Numeracy** (-1sd) [m.e.]			**0.416**	
2-sided p-value (H_1 : m.effect $\neq 0$)			0.077	
@**High Numeracy** (+1sd) [m.e.]			**0.045**	
2-sided p-value (H_1 : m.effect $\neq 0$)			0.833	
LEDA **Participant** (IV)				
@**Low Numeracy** (-1sd) [m.e.]				**0.492**
2-sided p-value (H_1 : m.effect $\neq 0$)				0. 072
@**High Numeracy** (+1sd) [m.e.]				**0.003**
2-sided p-value (H_1 : m.effect $\neq 0$)				0.992

Note: The table reports results on independent regression assessing: i) the causal effect of being offered a slot for LEDA on data literacy (ITT); i) the causal effect of participation to LEDA on data literacy (IV). See Table A3 and Table A4 for corresponding first stage estimates. Binary indicator for participation into LEDA is enrollment in the program (vs not enrollment). The Table reports estimates from two different model specifications (baseline and preferred), that differ as the preferred specification allows for heterogeneous effects based on pre-treatment numeracy Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.

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

Table A6: Effect of the assignment and of participation to the LEDA program on primary University career - ITT and IV results. Measure of participation based on: enrollment

	Model parameters' estimates			
	Baseline specification		Preferred specification	
	ITT	IV	ITT	IV
Assigned to LEDA	0.165		0.165	
	(0.646)		(0.651)	
Assigned to LEDA · Numeracy			-0.166	
			(0.637)	
LEDA Participant		0.188		0.175
		(0.701)		(0.704)
LEDA Participant · Numeracy				-0.222
				(0.858)
Numeracy	0.398	0.399	0.474	0.472
	(0.327)	(0.315)	(0.427)	(0.421)
Area fe	Yes	Yes	Yes	Yes
Observations	85	85	85	85
Assigned to LEDA (ITT)				
@**Low Numeracy** (-1sd) [m.e.]			**0.332**	
2-sided p-value (H_1 : m.effect $\neq 0$)			0.690	
@**High Numeracy** (+1sd) [m.e.]			**-0.001**	
2-sided p-value (H_1 : m.effect $\neq 0$)			0.999	
LEDA **Participant** (IV)				
@**Low Numeracy** (-1sd) [m.e.]				**0.397**
2-sided p-value (H_1 : m.effect $\neq 0$)				0.702
@**High Numeracy** (+1sd) [m.e.]				**-0.047**
2-sided p-value (H_1 : m.effect $\neq 0$)				0.968

Note: The table reports results on independent regression assessing: i) the causal effect of being offered a slot for LEDA on regular University career (ITT); i) the causal effect of participation to LEDA on regular University career (see Table A3 and Table A4 for corresponding first stage estimates). Binary indicator for participation into LEDA is enrollment in the program (vs not enrollment). The Table reports estimates from two different model specifications (baseline and preferred), that differ as the preferred specification allows for heterogeneous effects based on pre-treatment numeracy Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.

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

B Alternative definition of the treatment

This section presents the results using an alternative definition of take-up based on actual participation. In the absence of attendance data, we employ a more restrictive measure, namely a binary indicator that takes the value 1 in case students completed at least one of the four LEDA exams by early June 2023 and zero otherwise. Even if the first stage decreases (see estimates in Table B1 versus estimates in Table A3 for the baseline model without heterogeneous effects and estimates in Table B2 versus those reported in Table A4 for the preferred model that allows heterogeneous effects), the instrument is not weak and we have sufficient independent sources of variation to identify parameters in the model with heterogeneous effects. As expected, the causal effect of actual participation (as proxied by taking at least one exam) has a larger effect than the mere enrollment leading to an increase in data literacy of over 0.5, nearly double the effect observed when considering the causal effect of enrollment. Additionally, heterogeneity by pre-treatment numeracy level is more pronounced: the causal effect of LEDA for students with low levels of numeracy is more pronounced, while the causal effect for students with high levels of numeracy is not statistically significant and close to zero in absolute value.[19] For completeness, Table B3 and Table B4 report both ITT and IV results, even if ITT estimates coincide with those previously reported in Table A5 and Table A6,

[19]We refer to students with low levels of numeracy as those with pre-treatment levels of numeracy 1sd below average and to students with hight levels of numeracy as those with pre-treatment levels of numeracy 1sd above average.

Table B1: Effect of being offered a slot for the LEDA program on participation (First Stage)

	No controls	Pre-treatment controls
	(1)	(2)
Assigned to LEDA	0.424***	0.425***
	(0.079)	(0.080)
Numeracy		0.005
		(0.040)
Area fe	Yes	Yes
F-test	28.655	28.224
Adj. R-Square	0.232	0.222
Observations	85	85

Note: The table reports estimate of first stage parameters assessing the causal effect of being offered a slot for LEDA on participation to the program using linear probability model and OLS estimator for alternative specifications that differ for the control variables included. Our preferred specification is the one presented in column (2). IV estimates corresponding to specification in column (2) are presented in Table B3 (data literacy) and Table B4 (students' careers). Binary indicator for participation into LEDA is enrollment in the program (vs not enrollment). Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.
Legend: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B2: Effect of being offered a slot for the LEDA program on participation (First Stage) - Models allowing for heterogeneous effects by pre-treatment numeracy levels

	At least one exam in the 1st exam session (P)	
	P	P · Numeracy
Assigned LEDA	0.425***	-0.053
	(0.0800)	(0.0670)
Assigned LEDA · Numeracy	-0.040	0.360***
	(0.0833)	(0.1094)
Numeracy	0.023	-0.011
	(0.0226)	(0.0167)
Area fe	Yes	Yes
F-test	14.14	6.41
SW F-test	27.33	10.40
Observations	85	85

Note: The table reports estimate of first stage parameters assessing the causal effect of being offered a slot for LEDA on participation to the program using linear probability model and OLS estimator in a model were heterogeneous effects are allowed (see TableB3 for corresponding IV estimates on data literacy and Table B4 for students' careers). Binary indicator for participation into LEDA is having taken at least one exam by the end of the first exam session (vs having taken none) and it is denoted with *P*. Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.
Legend: SW for Sanderson-Windmeijer multivariate F test of excluded instruments. on joint significance of the instrument as they are tests for under-identification and weak identification of individual endogenous regressors in models with multiple endogenous regressors and instruments. They check for "sufficient" independent source of variation, *partialling-out* linear projections of the remaining endogenous regressors. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B3: Effect of the assignment and of participation to the LEDA program on data literacy - ITT and IV results. Measure of participation based on: at least one exam taken by the end of the first exam session

	Model parameters' estimates			
	Baseline specification		Preferred specification	
	ITT	IV	ITT	IV
Assigned to LEDA	0.230		0.230	
	(0.173)		(0.172)	
Assigned to LEDA · Numeracy			-0.186	
			(0.140)	
LEDA Participant		0.542		0.485
		(0.399)		(0.421)
LEDA Participant · Numeracy				-0.462
				(0.491)
Numeracy	0.470***	0.467***	0.555***	0.539***
	(0.071)	(0.087)	(0.097)	(0.117)
Area fe	Yes	Yes	Yes	Yes
N	85	85	85	85
	Hypothesis testing (p-values)			
Assigned to LEDA (ITT)				
@**Low Numeracy** (-1sd) [m.e.]			**0.416**	
2-sided p-value (H_1: m.effect $\neq 0$)			0.077	
@**High Numeracy** (+1sd) [m.e.]			**0.045**	
2-sided p-value (H_1: m.effect $\neq 0$)			0.833	
LEDA **Participant** (IV)				
@**Low Numeracy** (-1sd) [m.e.]				**0.946**
2-sided p-value (H_1: m.effect $\neq 0$)				0.094
@**High Numeracy** (+1sd) [m.e.]				**0.023**
2-sided p-value (H_1: m.effect $\neq 0$)				0.974

Note: The table reports results on independent regression assessing: i) the causal effect of being offered a slot for LEDA on data literacy (ITT); ii) the causal effect of participation to LEDA on data literacy (IV). See Tables B1 and B2 for corresponding first stage estimates. Binary indicator for participation into LEDA is having taken at least one exam by the end of the first exam session (vs having taken none). The Table reports estimates from two different model specifications (baseline and preferred), that differ as the preferred specification allows for heterogeneous effects based on pre-treatment numeracy. Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.

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

Table B4: Effect of the assignment and of participation to the LEDA program on regular University career- ITT and IV results. Measure of participation based on: at least one exam taken by the end of the first exam session

	Model parameters' estimates			
	Baseline specification		Preferred specification	
	ITT	IV	ITT	IV
Assigned to LEDA	0.165		0.165	
	(0.646)		(0.651)	
Assigned to LEDA · Numeracy			-0.166	
			(0.637)	
LEDA Participant		0.389		0.337
		(1.437)		(1.475)
LEDA Participant · Numeracy				-0.424
				(1.717)
Numeracy	0.398	0.396	0.474	0.462
	(0.327)	(0.313)	(0.427)	(0.411)
Area fe	Yes	Yes	Yes	Yes
N	85	85	85	85
	Hypothesis testing (p-values)			
Assigned to LEDA (ITT)				
@**Low Numeracy** (-1sd) [m.e.]			**0.332**	
2-sided p-value (H_1: m.effect $\neq 0$)			0.690	
@**High Numeracy** (+1sd) [m.e.]			**-0.001**	
2-sided p-value (H_1: m.effect $\neq 0$)			0.999	
LEDA **Participant** (IV)				
@**Low Numeracy** (-1sd) [m.e.]				**0.761**
2-sided p-value (H_1: m.effect $\neq 0$)				0.701
@**High Numeracy** (+1sd) [m.e.]				**-0.088**
2-sided p-value (H_1: m.effect $\neq 0$)				0.972

Note: The table reports results on independent regression assessing: i) the causal effect of being offered a slot for LEDA on regular University career (ITT); ii) the causal effect of participation to LEDA on regular University career (IV). See Table B1 and Table B2 for corresponding first stage estimates. Binary indicator for participation into LEDA is having taken at least one exam by the end of the first exam session (vs having taken none). The Table reports estimates from two different model specifications (baseline and preferred), that differ as the preferred specification allows for heterogeneous effects based on pre-treatment numeracy. Consistently with the randomization design, all regression include area fixed effects. Robust standard errors in parentheses. *Numeracy* is the standardized numeracy score based on PIAAC test. The indicator is standardized with respect to the sample of endline respondents not assigned to LEDA.

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