

Evaluating the impact of student learning support programs in an open-access, low-tuition higher education institution

Antoine Platteau

(TIMES2, Université libre de Bruxelles)

Philippe Emplit

(OPERA-Photonics, Université libre de Bruxelles)

Dorothee Baillet

(CRSE, Université libre de Bruxelles)

Catherine Dehon

(ECARES, Université libre de Bruxelles)

Abstract

Faced with high failure rates in higher education, many institutions have introduced learning support programs to enhance student success. This study evaluates the causal impact of a peer tutoring program in the low-tuition, open-access context of French-speaking Belgian universities, using data from the Université libre de Bruxelles. Applying propensity score matching, results show that tutoring improves first-year students' grades, by 1 to 2.5 points out of 20 and enables one third of participants to pass courses they would otherwise have failed. These findings complete existing evidence from selective systems and highlight peer tutoring as a cost-effective way to promote success.

JEL codes : I21, I23, C31, C35

Keywords : Student Learning Support, Peer Tutoring, Academic Performance, Belgian Higher Education, Impact Evaluation, Propensity Score Matching

1. Introduction

Higher education systems differ markedly in how they regulate access, finance studies, and select students, with important consequences for student outcomes. Over the past two decades, global university enrolment has more than doubled, reaching 254 million students worldwide in 2024 (UNESCO Institute for Statistics, 2024). Yet this expansion has taken place within highly contrasted institutional frameworks. Drawing on a comparative perspective, Lambert (2023) distinguishes three broad types of higher education systems: Anglo-Saxon, Nordic, and Continental European. Anglo-Saxon systems typically combine high tuition fees with strict entry selection, while Nordic countries maintain selective admission policies alongside low or zero tuition fees. Despite these differences in funding, both models rely on strong selection at entry, resulting in relatively high graduation rates (around 80%) and low dropout rates (approximately 12%) (OECD, 2021).

In contrast, Continental European systems, including French-speaking Belgian universities under the Fédération Wallonie-Bruxelles (FWB), adopt an open-access approach. In the FWB, students holding an upper secondary education certificate (CESS) are eligible to enrol in most higher education faculties. Around 70% of graduates enter higher education in the year following completion of secondary school, a proportion that rises to 76% within three years (Dujardin et al., 2023). Tuition fees are comparatively low (835€ per year) as education is funded largely by public investment, contributing to these high entry rates. However, this inclusive approach is also associated with substantially poorer student outcomes. On-time graduation rates for first-year undergraduate students barely reach 21%, with dropout rates exceeding 30%. Even three years beyond the expected programme duration, only about half of students graduate, a performance well below that observed in Nordic and Anglo-Saxon systems (Lambert, 2023; OECD, 2021).

These disparities reflect structural differences in student selection and preparedness. In a system like the FWB, where open-access policies lead to a highly heterogeneous student population, implementing measures to reduce dropout and failure rates is a particularly ambitious challenge. Many students enter higher education underprepared due to persistent inequalities in compulsory education. This situation is further exacerbated by high rates of grade repetition in secondary school, with nearly one in two students in the final year having repeated at least one grade. In this context, academic support programs are not merely complementary but play a central role in promoting student success (FWB, Fédération Wallonie-Bruxelles, 2023; Lambert, 2020).

To address the challenge of success in higher education, universities worldwide have introduced academic support interventions, often involving peer tutoring (Stigmar, 2016). These programs are often implemented at the national or institutional level and commonly target high-risk courses (Dawson et al., 2014; Topping, 1996). One well-known example is Supplemental Instruction (SI), widely adopted in Anglo-Saxon universities, which has been proven to improve academic outcomes, including higher grades and lower dropout rates (Bowman et al., 2023; Guarcello et al., 2017). However, most published studies stem from selective, tuition-based higher education systems, limiting generalizability of these findings to more inclusive, publicly funded systems such as that of the FWB. Moreover, several studies examining comparable support programs rely primarily on descriptive approaches, not allowing for strong causal conclusions.

In this respect, the FWB constitutes a particularly relevant case for extending the international literature. Characterized by open-access policies and high dropout rates, the system has placed student support at the core of higher education policy. As part of the Bologna process in the early

2000s, legislation required all universities to allocate specific funding to learning support activities through the *Aides à la Réussite* program, aimed at reducing initial disparities, facilitating the transition into higher education, and improving student success (FWB, Fédération Wallonie-Bruxelles, 2004). The 2013 ‘Paysage’ Decree further expanded the list of recommended support measures, and since 2022 institutions have been required to provide additional support to students who obtain fewer than half of the credits in their first year.

The aim of these student learning support programs, targeted at first-year undergraduate students, is twofold: strengthening students’ understanding of core theoretical concepts and encouraging reflection on study strategies in order to foster practices better aligned with academic expectations. In the FWB, universities are free to choose which support activities to implement from the list provided by the decree. The Université libre de Bruxelles (ULB) has implemented a wide range of learning support activities, among which peer tutoring constitutes a central component. Each semester, approximately ten tutoring sessions are organised for selected courses, and students may attend freely according to their needs. Sessions focus on course content and are led by advanced undergraduate or master’s students who previously passed the course with a grade of at least 16 out of 20 and received prior training (Salmon et al., 2009)¹.

This study evaluates the causal impact of peer tutoring on the academic performance of first-year undergraduate students at the ULB. Using propensity score matching as a complement to ordinary least squares (OLS) regressions, we approximate the causal impact of peer tutoring conditional on observed characteristics. The analysis focuses on three first-year courses within the Faculty of Psychological and Educational Sciences that implemented peer tutoring activities.

While the effectiveness of peer tutoring is well documented in selective, tuition-based systems, evidence from open-access contexts remains limited. This study provides updated evidence within the specific case of the FWB, offering new insights into how tutoring performs in a publicly funded, non-selective system. The findings aim to inform future research and policy evaluation on student support programs in the inclusive higher education setting.

This paper is structured as follows. First, we review the literature on the assessment of student learning support programs and their impact on academic outcomes. Next, we detail our methodological framework and analyse data from ULB. Finally, we discuss the findings, implications, and limitations, offering recommendations for future research.

2. Literature review

The literature review is structured into three main sections. The first presents an overview of student learning support programs implemented across the world and the ways in which their effectiveness has been evaluated. The second examines impact evaluation approaches and highlights the importance of addressing selection bias inherent in simple descriptive analyses. The third discusses participation rates in such programs.

2.1. Assessment of student learning support programs

To support students facing academic difficulties, higher education institutions around the world have implemented various tutoring programs. These programs take different names across countries: *Supplemental Instruction* (SI) in the United States, *Peer-Assisted Study Sessions*

¹ Baillet 2015, not published.

(PASS) in Australia, and *Peer-Assisted Learning* in the United Kingdom. In Continental Europe, similar initiatives exist, such as the *Aides à la Réussite* program in Belgium's FFW and comparable initiatives in France and Germany. In Nordic countries, peer tutoring is less standardized at national level but more often driven by universities, through models like *Supplemental Instruction* in Sweden or equivalent systems in Norway (Malm et al., 2012; Topping, 1996).

Since the late 20th century, numerous studies have linked participation in peer tutoring and other support programs to better academic outcomes. Descriptive and correlational studies generally report higher grades among participants compared to non-participants (Dancer et al., 2007; Dawson et al., 2014; Lei et al., 2018; Malm et al., 2012; Peterfreund et al., 2008). However, these designs cannot establish causality, since students who attend tutoring may differ from those who do not. For instance, Mason (2018) argues that participants often show higher motivation and persistence, introducing selection bias (Bruffaerts et al., 2011; Endrizzi, 2010). Therefore, identifying the causal impact of tutoring on grades requires methods that correct for this bias (Rokusek et al., 2022).

2.2. Impact evaluation and methodological approaches

In this regard, experimental and quasi-experimental approaches have become essential to assessing the true effect of tutoring. Randomized controlled trials (RCTs) are often considered the gold standard for impact evaluation. Studies using RCTs in higher education, conducted in countries such as Spain, Germany, and the Netherlands, have generally shown that peer tutoring improves GPAs, reduces failure rates, and fosters more effective study habits (Arco-Tirado et al., 2020; Dekker et al., 2023; Hardt et al., 2023).

When solely observational data are available and voluntary participation into treatment does not mimic random assignment, as would be the case in an RCT, quasi-experimental methods become necessary. These approaches aim to approximate the conditions of a randomized experiment, allowing to compute the treatment effect. Among them, matching methods, such as propensity score matching, compare participants who received the intervention with similar non-participants, using observable characteristics to estimate the program's impact while reducing selection bias (Imbens & Wooldridge, 2009; Rosenbaum & Rubin, 1983). This approach increases comparability thereby enabling to isolate the program's impact (Gertler et al., 2016). Guarcello et al. (2017) used propensity score matching to study the impact of SI in improving success rates for an introductory psychology course at a large US university. Their study found that attending even one SI session increased the odds of passing the exam by 2.2 times ($p = 0.006$) while attending two or more sessions increased the odds by 2.8 times ($p = 0.03$). Bowman et al. (2023) similarly found, in US Midwestern universities, that SI improved grades (0.10 to 0.19 on a four-point scale) and reduced failure rates (4 to 6 percentage points), with greater improvements for students attending five or more sessions. Rivera (2022) reached similar positive effects of SI on student success.

2.3. Participation challenges and research limitations

Despite these encouraging results, participation rates in tutoring programs remain low regardless of the educational setting. Research consistently shows that students who would benefit most, often those least academically prepared, are also the least likely to seek help (Anfuso et al., 2022; Etter et al., 2000). This self-selection limits the reach and impact of support initiatives. Moreover, most empirical studies originate from selective and tuition-based higher education systems, making it difficult to generalize findings to publicly funded, open-access environments.

In Continental Europe, and particularly in Belgium's FWB, evidence on the effectiveness of tutoring remains scarce. Although peer support activities such as the *Aides à la Réussite* program have been widely implemented following policy changes, their evaluation has been largely descriptive and rarely causal. As a result, it remains unclear whether the positive effects reported internationally also hold in systems characterized by low tuition fees and open-access.

2.4. Present study

This study addresses this research gap by providing updated causal evidence on the effectiveness of peer tutoring in an open-access, publicly funded higher education system. Using propensity score matching, it estimates the impact of participation in peer tutoring on academic performance among first-year undergraduate students at the ULB. By focusing on a context that differs from those commonly studied, open-access, low fees, and heterogeneity in student preparedness, it contributes to the international discussion on whether and how tutoring programs can improve outcomes in inclusive higher education systems. Beyond its empirical contribution, this study addresses a timely and highly debated issue in the FWB, where recent policy changes in the higher education decree have raised concerns among students and institutions, while placing peer tutoring at the center of student support strategies. As such, it offers a foundational reference for future, more comprehensive analyses of the effects of these reforms and of student support initiatives on academic success.

3. Data

3.1. Sources of data

The data originates from the ULB and focuses on one device from a broader student learning support program, implemented by the Faculty of Psychological and Educational Sciences. This study focuses on peer tutoring organized for first-year undergraduate students, where the tutoring sessions are led by an upper-year student, as explained in the introduction. During each session, tutors passed out a sheet to establish an attendance list. Depending on the courses, between 7 and 10 tutoring sessions were dispensed. The organisation of these tutoring sessions is left to the tutor's choice, the details of which are often discussed with the lecturer. There is often a short theoretical reminder on a specific chapter of the course, a review of the prerequisites, followed by a question-and-answer session. Here, we focus on three courses that are part of the same first year bachelor degree, for which tutoring is organized, namely Neuropsychology, Statistics and Cytology. Evaluation in these three courses are similar, i.e. multiple-choice questions.

A first dataset, encompassing grades and tutoring attendance over four academic years (2013, 2014, 2015 and 2016), is provided by the faculty's Learning Support Services. Personal data such as gender, educational background, socio-economic situation (SES) and high school background are provided by the institutional databases of the university and are merged with the first database.

3.2. The population

We focus on first-year undergraduate students enrolled at the Université libre de Bruxelles (ULB) in the Faculty of Psychological and Educational Sciences who were entering higher education for the first time. Over the four academic years considered, 1,227 new students enrolled in the faculty, of whom 82.7% held a Belgian upper secondary school diploma. Information on high school curriculum was missing for approximately twenty students, who were therefore excluded from the analysis. In addition, although all remaining students were enrolled in the three courses considered, taking the exam was not compulsory. As grades are only available for students who

took the exam, the analysis was restricted to those who took at least one exam in the first or second session. This results in a final analytical sample of 879 students who took at least one of the exams in the three selected courses. The number of observations per course differs slightly as students may choose to take the exam for some courses but not others: 822 for Neuropsychology, 664 for Cytology, and 547 for Statistics.

3.3. Dependent variable

The dependent variable is the student's course grade, ranging from 0 to 20. Focusing on individual course grades allows us to assess the isolated impact of tutoring on each specific course without direct cross-course effects. At the ULB, students may only sit exams during official examination sessions: January, June, and August for Neuropsychology, and June and August for the other two courses. After the June session, a grade of 10 or higher validates the course and finalizes the mark. Grades below 10 do not validate the course, in which case students are allowed to retake the exam in the August session. When multiple exam attempts are available, we keep the highest grade obtained, regardless of the session in which it was achieved. Accordingly, the analysis focuses on the final course grade rather than solely on a binary pass-fail outcome.

3.4. Treatment variable

Our variable of interest is whether a student participated in at least one tutoring session (1 if yes, else, 0). While some studies explore different levels of engagement through dose-response analysis (Bowman et al., 2023), we focus on a binary distinction between students who attended at least one session and those who never participated. This dichotomous approach allows us to assess whether minimal exposure to tutoring is associated with improved outcomes.

3.5. Covariates

In propensity score matching, the objective is to compare students with similar achievement profiles. To do so, the propensity score model includes covariates that are related to both the treatment assignment and the academic outcome. These covariates are selected based on established theories and prior research on factors influencing university success (Patrick et al., 2011). In the present study, these variables fall into three categories: personal characteristics, socio-economic status, and high school background.

Among individual characteristics, gender has been shown to be a strong and consistent predictor of academic performance. A large body of research shows that women tend to outperform men academically due to better engagement, organization, and adherence to academic norms. Therefore, gender is included as a key control variable in the analysis (Borg, 2015; Jansen & Bruinsma, 2005).

Socio-economic background constitutes a second important dimension. As household SES data is unavailable, scholarship eligibility is used as a proxy. Prior studies indicate that students receiving scholarships, or those with jobs, face disadvantages compared to those who are financially supported by their families (Bruffaerts et al., 2011). Socio-economic disparities also operate at the school level: students from higher SES schools tend to perform better academically, reflecting differences in available resources, such as class size and resources availability (Li & Dockery, 2015). In the FWB, a 1998 decree promotes 'differentiated supervision', or 'positive discrimination', allocating additional resources to lower-SES secondary schools, identified by an index ranging from 1 to 5 out of 20 (FWB, Fédération Wallonie-Bruxelles, 2009)

Prior academic trajectory in secondary education is also closely linked to success in higher education. Students who took Latin, Greek, or extensive mathematics courses (≥ 6 hours) are

more likely to succeed in their first year at university (Arias Ortiz & Dehon, 2008). In French-speaking Belgium, secondary education diplomas are categorized as general, technical, artistic, and professional. For our analysis, we group these into two categories: General and Other. This distinction highlights the difference between students prepared for higher education and those ready to enter the labour market. Morlaix & Suchaut (2012) showed significant performance differences between students with professional/technical baccalaureates and those with general ones. In 2015, one in three students in Belgium had repeated a grade in high school (OECD, 2016). Evidence suggests that grade repetition adversely affects psychosocial outcomes, dropout risks, and academic performance (Galand et al., 2019). In our data, we assume students graduating from high school at 18 did not repeat a grade.

Finally, the analysis accounts for institutional changes over time. Our database spans multiple academic years under different decrees, including the 2014 'Paysage' Decree, which shifted from a year-based to a credit-based system. This new system has been shown to extend study durations (Fédération Wallonie Bruxelles, 2013; Dehon & Leboutteiller, 2025). A categorical variable is used to account for year-on-year changes and give more granularity to the results.

4. Descriptive evidence

This section provides a descriptive overview of the data, assesses the representativeness of the sample relative to the population, and identifies the characteristics associated with participation in the tutoring program.

Table 1 provides a descriptive overview of the sample. As expected from the literature, only a small proportion of students (14%) attended at least one tutoring session, with Neuropsychology showing the highest participation rate. Although low, this figure is consistent with previous findings (Etter et al., 2000).

Table 1: Descriptive breakdown of the sample, distributed between the three courses (n,%).

		Neuropsychology		Cytology		Statistics	
		n	%	n	%	n	%
Tutoring attendance	Not tutored	681	82.8%	579	87.2%	482	88.1%
	Tutored	141	17.2%	85	12.8%	65	11.9%
Academic year	2013	178	21.7%	143	21.5%	124	22.7%
	2014	214	26.0%	170	25.6%	152	27.8%
	2015	222	27.0%	178	26.8%	142	26.0%
	2016	208	25.3%	173	26.1%	129	23.6%
Gender	Male	221	26.9%	161	24.2%	127	23.2%
	Female	601	73.1%	503	75.8%	420	76.8%
Latin/Greek during HS	No	688	83.7%	543	81.8%	444	81.2%
	Yes	134	16.3%	121	18.2%	103	18.8%
Strong math profile during HS	No	719	87.5%	567	85.4%	456	83.4%
	Yes	103	12.5%	97	14.6%	91	16.6%
Benefits from a scholarship	No	570	69.3%	452	68.1%	385	70.4%
	Yes	252	30.7%	212	31.9%	162	29.6%
Grade repetition during HS	No	401	48.8%	356	53.6%	324	59.2%
	Yes	421	51.2%	308	46.4%	223	40.8%
HS positive discrimination	No	624	75.9%	511	77.0%	435	79.5%
	Yes	198	24.1%	153	23.0%	112	20.5%
Type of high school degree	General	628	76.4%	532	80.1%	460	84.1%
	Other	194	23.6%	132	19.9%	87	15.9%

In Psychological and Educational Sciences at the ULB, approximately three quarters of the students are female, which is not surprising given that psychology studies attract mainly women: 70-75% of students are female (Fowler et al., 2018). About 17% have a Latin/Greek background, and 15% have an advanced background in mathematics. The limited presence of students with advanced mathematics training is coherent, since those with stronger quantitative skills usually opt for STEM or economics programs. 30% of students receive a scholarship, aligning with the FWB figures (Paume et al., 2021). Around 22% of students come from a positively discriminated high school, reflecting a low SES. Nearly half have repeated a grade in high school, which is high but consistent with the fact that nearly half of the students in their final year of high school (47.8%) have repeated at least one year during their secondary education (FWB, Fédération Wallonie-Bruxelles, 2023). About 20% have professionally oriented degrees, which is more than the general figures from the FWB, where around 10% of such graduates attend university (ARES, 2016). Finally, the number of students is stable across the academic years, showing no major fluctuation in enrolment. Overall, the sample appears reasonably representative of the faculty's student population in the humanities.

Table 2: Difference in characteristics between the two groups, by course.

Variables	Neuropsychology			Cytology			Statistics		
	Treated (%)	Control (%)	Difference (pp)	Treated (%)	Control (%)	Difference (pp)	Treated (%)	Control (%)	Difference (pp)
Female	87.9	70.0	17.9	90.8	74.9	15.9	89.4	73.7	15.7
Latin/Greek during HS	18.4	15.9	2.5	18.5	18.9	-0.4	24.7	17.3	7.4
Mathematics during HS (>5h)	13.5	12.3	1.2	16.9	16.6	0.3	15.3	14.5	0.8
Scholarship	39.0	28.9	10.1	36.9	28.6	8.3	44.7	30.1	14.6
Grade repetition during HS	41.8	53.2	-11.4	38.5	41.1	-2.6	42.4	47.0	-4.6
HS positive discrimination	33.3	22.2	11.1	27.7	19.5	8.2	31.8	21.8	10.0
General type of CESS	83.0	75.0	8.0	83.1	84.2	-1.1	82.4	79.8	2.6

Note. Values are percentages. Differences are expressed in percentage points (treated minus control).

Values are in bold when the absolute difference is higher than 5%.

Table 2 reports differences in the distribution of student characteristics between tutored and non-tutored students, providing a descriptive assessment of potential imbalances between the two groups. The results indicate that the profiles of participants and non-participants differ substantially. This initial comparison highlights important patterns in participation and helps assess whether the tutoring program effectively reaches the student it is intended to support. For instance, in Neuropsychology, tutored students are more often female and more likely to hold a general secondary education degree, both characteristics positively associated with academic success. At the same time, they are more likely to come from disadvantaged backgrounds (as indicated by scholarship eligibility or positive discrimination status) and less likely to have repeated at least one year during secondary education, which are negatively correlated with performance.

This pattern suggests that while the tutoring program succeeds in attracting students with greater socioeconomic disadvantage, it appears less effective at reaching students with weaker prior academic outcomes, such as those who repeated a grade or did not come from general secondary education, despite these students being a key target population for such support programs. This observation ties in with Etter et al. (2000)'s idea that those who need help are more reluctant to seek it. Applying the same analysis to the two other courses leads to similar conclusions.

5. Methodological approach

The contribution of tutoring to students' final grades can first be assessed using OLS regressions, which model the relationship between course performance and a set of observed covariates. However, the validity of OLS relies on strong identifying assumptions, in particular the functional form and the exogeneity of tutoring participation conditional on observed characteristics. In a context where participation in tutoring is voluntary and likely subject to selection, this assumption may be violated, limiting the causal interpretation and reliability of OLS estimates. In addition, OLS estimates can be affected by extrapolation when treated and untreated students differ substantially in their characteristics, leading to extrapolations outside the region where data overlap.

To approximate causal interpretation, we use a matching analysis. Unlike OLS regressions, matching methods do not require strong assumption on the functional form between the outcome of interests and the covariates (Angrist & Pischke, 2009; Black, 2015). Instead, under the

assumptions outlined below, matching corrects for selection on observables by constructing a control group of untreated individuals with characteristics similar to those of treated students. By ensuring that treatment effects are estimated only within the region of common support, where treated and untreated students are comparable, matching provides a more robust framework for approximating causal effects from observational data (Imbens & Wooldridge, 2009; Stuart, 2010). This strategy reinforces causal validity by relying on a design-based approach consistent with the potential outcomes framework.

The treatment effect for an individual i is the difference between the value of the outcome while receiving the treatment and the value of the outcome had they not been treated:

$$T_i = Y_i(1) - Y_i(0) \quad (1)$$

where T_i is the causal treatment effect for individual i , $Y_i(D_i)$ are the potential outcomes in the two treatment situations (D_i being the treatment indicator, equal to 1 if individual i is assigned to the treatment, 0 if not). The issue is that only one of the two counterfactual treatment situations is observed for each individual. As we want to estimate the average treatment effect on the treated (ATT), defined as follow:

$$T_{ATT} = E[T|D = 1] = E[Y(1) - Y(0)|D = 1] \quad (2)$$

matching methods allow to impute missing observations for counterfactual outcomes, $E[Y(0)|D = 1]$.

Two assumptions are needed to use matching. The first is the conditional independence or unconfoundedness assumption. It states that given a set of observable covariates, treatment assignment is independent of the potential outcomes (Imbens & Wooldridge, 2009; Stuart, 2010)

$$Y(1); Y(0) \perp D | X. \quad (3)$$

The second assumption of common support refers to the overlap in propensity score distribution between treatment and control group. If there is no sufficient overlap, then it becomes impossible to find matches and no analysis can be run.

These assumptions imply that within subpopulations where the covariate values are the same, the assignment to treatment can be interpreted as if a completely randomized experiment had been conducted, even though the actual probabilities of assignment are unknown (Imbens & Rubin, 2015).

Propensity score matching can be used to simplify finding matches in high-dimensional data. The propensity score represents the conditional probability of enrolling in the program given the covariates (Rosenbaum & Rubin, 1983) and is defined as follows:

$$p_i(X_i) \equiv \Pr(D_i = 1|X_i) \quad (4)$$

where $p_i(X_i)$ is the propensity score for the individual i . Then, units are matched based on this score rather than matching on the covariate vector. The propensity score is either estimated through a logit or probit model, given covariates. After defining the matching parameters, only matched control units are used to compute the treatment effect (Gertler et al., 2016).

6. Results

This section first presents the regression results, followed by the findings from the propensity score matching analysis and the sensitivity analysis. Taken together, these approaches provide a more comprehensive assessment of the impact of tutoring on students' grades.

6.1. Regression

Table 3 compares regressions outcomes for tutored and non-tutored students across the three courses, with the dependent variable being the course grade, out of 20. Multiple linear regressions allow us to explore the structure of the data and assess whether the relationships between covariates and academic performance align with theoretical expectations.

Table 3: Multiple regressions outcome with the dependent variable being the course grade, out of 20.

Variables	Modalities	Neuropsychology		Cytology		Statistics	
		Estimates	Std. Error	Estimates	Std. Error	Estimates	Std. Error
Intercept		7.81***	0.49	8.78***	0.41	8.48***	0.64
Tutored	Untutored
	Tutored	2.42***	0.41	0.76**	0.36	1.29**	0.59
Year	2013
	2014	0.95**	0.42	-0.59	0.34	-1.37**	0.53
	2015	0.57	0.42	1.00***	0.35	-1.77***	0.54
	2016	0.80*	0.43	1.34***	0.35	-0.62	0.55
Gender	Male
	Female	1.03***	0.34	0.45	0.29	0.82*	0.46
Latin/Greek during HS	No
	Yes	1.51***	0.41	0.68**	0.33	2.45***	0.50
Mathematics in HS (> 5h)	No
	Yes	0.95**	0.45	0.38	0.35	1.48***	0.53
Scholarship	No
	Yes	0.01	0.32	-0.17	0.27	-1.12***	0.43
Grade repetition during HS	No
	At least 1	-1.32***	0.35	-1.01***	0.28	-1.14**	0.44
HS positive discrimination	No
	Yes	-1.37***	0.36	-0.59**	0.30	-0.32	0.49
CESS type	General
	Other	-1.89***	0.38	-1.42***	0.32	-1.80***	0.55
Observations		822		664		547	
R^2 / R^2 adjusted		0.23 / 0.22		0.18 / 0.17		0.20 / 0.18	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note. The '.' means that the modality is the reference for the variable.

The tutoring variable is positively associated with student grades in all three courses and statistically significant for Neuropsychology ($p < 0.01$), where tutored students score on average 2.42 points higher than their peers. A similar conclusion is reached for the two other courses, for which the coefficient is positive and significant ($p < 0.01$). This effect is both substantial and consistent with previous findings on the benefits of peer tutoring (Borg, 2015; Carvalho, 2016).

The other variables align with existing studies. Students with a Greek or Latin high school background, as well as those with strong math background, tend to perform better. Conversely, students who repeated a year or hold technical/professional diplomas achieve lower grades, confirming prior studies (Arias Ortiz & Dehon, 2008; Galand et al., 2019; Morlaix & Suchaut, 2012). Gender effects are also observed, particularly in Neuropsychology, where women score on average 1.03 points higher than men, a difference significant at the 1% level. Regarding socioeconomic background, receiving a scholarship is negatively associated with performance in the Statistics course ($p < 0.01$). Similarly, having attended a high school under a positive discrimination scheme is negatively associated with grades in Neuropsychology and Cytology ($p < 0.01$ and $p < 0.05$, respectively).

While these regression results provide useful insights and show a positive association between tutoring and student performance, they do not allow for causal interpretation because of the selection bias. Since participation in the tutoring program is voluntary, the study does not rely on randomized assignment as in an RCT. To improve the results of the comparison between tutored and non-tutored students, we complete the regression analysis with a propensity score matching approach. Although this method does not fully solve the issue of unobserved confounders, it allows us to compare students with similar observable characteristics and helps provide results that are one step closer to causality.

6.2. Propensity score matching

The application of propensity score matching rests on two key requirements. First, the unconfoundedness assumption must hold (equation (3)) which, in theory, requires that conditioning on all variables that influence both, the treatment assignment and the outcome in the analysis. When this assumption holds, treatment status is independent of potential outcomes, allowing the estimated effect to be interpreted causally rather than as the result of confounding. In practice, however, unconfoundedness cannot be tested directly. It can only be approximated by conditioning on a rich set of observed covariates that are plausibly related to both participation in tutoring and academic performance. Accordingly, the propensity score is estimated using individual characteristics, socio-economic indicators, and high school background variables.

To assess the robustness of our results to potential violations of the unconfoundedness assumption, we rely on Rosenbaum bounds sensitivity analysis (Rosenbaum, 2002). Rather than testing whether all relevant variables have been included, this approach evaluates how sensitive the results are to potential hidden bias arising from unmeasured confounders. It quantifies the strength an unmeasured factor would need to have to substantially affect treatment assignment and thus call the causal conclusion into question. Put simply, it tests how large a hidden bias must be for the estimated effect to lose statistical significance. If a small hidden bias makes the results insignificant, then results cannot be considered robust (Olitsky, 2014; Zhao et al., 2025). The findings from this sensitivity analysis using Rosenbaum bounds are presented in the corresponding subsection.

A second requirement concerns covariate balance between treated and control units conditional on the propensity score,

$$D \perp X | p(X), \tag{5}$$

ensuring that the matching process is performed correctly. In that case, observations with the same propensity score should have the same distribution of characteristics, regardless of

treatment status (Becker & Ichino, 2002). However, if propensity score is mis-specified, imbalances in baseline characteristics between treatment groups can persist. Thus, it is crucial to report balance diagnostics after matching to ensure the process is accurate and effective (Zhang et al., 2019).

Supposing that assignment to treatment is unconfounded, equation (3) holds, then assignment to treatment is unconfounded given the propensity score,

$$Y(1); Y(0) \perp D | p(X). \quad (6)$$

So, if we ensure that we control for as many covariates as possible such that (3) – and thus (6) – holds, that there is sufficient overlap propensity score distribution between the groups and that the matching provides balanced covariates, then we can be confident in the treatment effect provided by the analysis (Imbens & Rubin, 2015).

The detailed methodology and results are presented below for the Neuropsychology course, which has the highest proportion of tutored students and offers the most balanced comparison between treated and untreated groups. This course is used as an illustrative case to describe the matching procedure in detail and avoid unnecessary repetition. Results for the two other courses are presented in a more concise manner to assess the consistency of the estimated effects across contexts and to inform the discussion of their potential generalizability.

The second hypothesis of common support was confirmed by examining the propensity score distribution for the three courses, showing a high proportion of overlap between groups. Figure 1 is an example for Neuropsychology.

Figure 1: Distribution of propensity score for the Neuropsychology course.



After matching, balance of baseline characteristics between control and treatment groups has been assessed by examining the standardized mean difference (SMD). It allows to evaluate the quality of the matches. According to Stuart et al. (2013), an imbalance exists for a given variable

when the SMD exceeds 0.1. In Table 4, the Neuropsychology course shows that, after logit KNN1² matching with replacement, the absolute SMD is 0.1 or lower for all variables, indicating comparable groups for computing the treatment effect. Similar results are observed for the other two courses.

Table 4: Covariate balance comparison for the Neuropsychology course before and after logit KNN1 matching with replacement.

Variables	Before matching			After matching		
	Means Treated	Means Control	Absolute std. mean difference	Means Treated	Means Control	Absolute std. mean difference
2013	0.348	0.189	0.332	0.348	0.397	0.104
2014	0.440	0.223	0.436	0.440	0.404	0.071
2015	0.106	0.304	0.641	0.106	0.099	0.023
2016	0.106	0.283	0.574	0.106	0.099	0.023
Female	0.879	0.700	0.550	0.879	0.865	0.044
Latin/Greek during HS	0.184	0.159	0.067	0.184	0.199	0.037
Mathematics during HS (> 5h)	0.135	0.123	0.033	0.135	0.099	0.104
Scholarship	0.390	0.289	0.207	0.390	0.362	0.058
Grade repetition during HS	0.418	0.532	0.229	0.418	0.418	0.000
HS positive discrimination	0.333	0.222	0.237	0.333	0.326	0.015
General type of CESS	0.830	0.75	0.211	0.830	0.837	0.019
Observations	141	681				
Matched				141	94	
Unmatched					587	

The figures at the bottom of the table show the number of matched individuals. Out of 822 total individuals, 141 attended at least one tutoring session. With ‘KNN1 with replacement’ matching, each treated individual found a match, reusing several control individuals. Consequently, 587 control individuals were not matched. The ATT is then calculated on the 141 matched pairs.

To ensure consistency of the results across models, different propensity score matching specification were run. Using the MatchIt package on R, the following match were run: KNN1 without replacement, KNN1 with replacement, KNN2 with replacement, KNN3 with replacement, caliper matching ($r = 0.01$ and $r = 0.05$)³.

6.2.1. Matching results

Table 5 shows the ATT following the different matching algorithms. In very few cases, the absolute SMD for one or two variables ranges between 0.10 and 0.15, but the ATT is still reported. The Neuropsychology course shows the most consistent results, with tutored students scoring 2.19 to 2.57 points higher out of 20, significantly at the 1% level. Similar trends are observed in the

² KNN1 stands for ‘K-Nearest Neighbours’ with $K = 1$, meaning that each treated individual is matched with the closest control group member based on the propensity score. ‘With replacement’ allows a control individual to be matched to multiple treated individuals, ensuring optimal matches.

³ In caliper matching, $r = 0.01$ means the difference in propensity scores between matched treatment and control units should not exceed 0.01. Treatment units without a suitable control within this range are discarded from the matched sample.

Cytology and Statistics courses, with grade increases of 1.00 to 1.67 and 1.16 to 1.78, respectively, when significant.

Table 5: Propensity score matching: results of the ATT using various algorithms.

Model	Specification	Neuropsychology		Cytology		Statistics	
		ATT	T-stat	ATT	T-stat	ATT	T-stat
Logit	KNN 1 no repl.	2.19***	4.84	1.44**	2.78	1.78**	2.51
	KNN 1 repl.	2.49***	3.94	1.53***	2.95	1.73*	1.88
	KNN 2 repl.	2.41***	4.04	1.44***	3.16	1.33*	1.74
	KNN 3 repl.	2.47***	4.57	1.67***	3.76	1.47**	2.13
	Caliper 0.01	2.26***	4.78	1.30**	2.46	1.56**	2.16
	Caliper 0.05	2.35***	5.09	1.42***	2.94	1.61**	2.26
Probit	KNN 1 no repl.	2.20***	4.60	1.06*	1.96	1.16*	1.66
	KNN 1 repl.	2.56***	3.45	1.19**	2.03	0.98	1.21
	KNN 2 repl.	2.38***	4.02	1.32**	2.80	0.93	1.31
	KNN 3 repl.	2.57***	4.80	1.54***	3.54	1.35**	2.33
	Caliper 0.01	2.36***	4.79	0.79	1.43	1.41*	1.93
	Caliper 0.05	2.35***	4.93	1.00*	1.90	1.26*	1.76

* $p < 0.1$. ** $p < 0.05$. *** $p < 0.01$

We further investigate potential heterogeneity in the treatment effect across selected student characteristics. To ensure sufficient statistical power, we focus on two variables for which the post-matching distributions leave more than 30 treated and control observations in each subgroup. Tables 6 and 7 report the results for the Neuropsychology course only, based on logit KNN1 matching with replacement, distinguishing students by grade repetition status and scholarship status, respectively⁴.

Table 6: ATT by grade repetition status for Neuropsychology.

Grade repetition status	ATT	Std. Error	95% CI
No repetition	2.93***	0.82	[1.33 ; 4.54]
Repetition	1.87*	1.01	[-0.11 ; 3.84]

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note. Estimates report ATTs obtained after propensity score matching using logit KNN1 matching with replacement.

⁴ For the other courses and covariates, the distribution of observations after matching does not leave sufficiently large subgroups (below 30 observations), which limits statistical reliability. These results are therefore not reported.

Table 7: ATT by scholarship status for Neuropsychology.

Scholarship status	ATT	Std. Error	95% CI
No scholarship	2.52***	0,85	[0.86 ; 4.19]
Scholarship	2.43**	1,04	[0.40 ; 4.47]

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note. Estimates report ATTs obtained after propensity score matching using logit KNN1 matching with replacement.

Table 6 suggests a larger estimated treatment effect for students who did not repeat a grade during high school, with an ATT of 2.93 compared to 1.87 for students who did, corresponding to a difference of 1.06 points out of 20. Table 7 examines heterogeneity by socioeconomic status, proxied by scholarship eligibility, and shows very similar treatment effects for scholarship participants and non-participants (2.43 versus 2.52), with a negligible difference of 0.09 points.

However, formal Wald tests do not provide evidence of statistically significant heterogeneity in treatment effects in either case ($p = 0.14$ for grade repetition status and $p = 0.86$ for scholarship status). As a result, we do not find statistically significant evidence for heterogeneous treatment effects along these dimensions.

6.2.2. Implications for pass-fail outcomes

Because the primary objective of the peer-tutoring program is to help students who might otherwise struggle to succeed, it is informative to translate these estimated grade improvements into implied pass or fail outcomes. To this end, let's fix an ATT, for instance, the KNN1 logit estimator with replacement, for Neuropsychology, which reaches 2.49⁵.

To illustrate the practical importance of this effect, we apply the estimated gain to construct counterfactual grades by subtracting 2.49 points from students' actual exam scores⁶. This interpretative exercise indicates that 45 of the 141 tutored students (31.9%) in Neuropsychology appear to have passed thanks to the program. Applying the same procedure to Cytology and Statistics yields similar findings: 30 out of 85 students (35.3%) and 17 out of 65 students (26.2%), respectively, seem to have passed because of tutoring. This calculation is intended to provide policy-relevant intuition about the magnitude of the effect rather than to represent a separate causal estimate.

⁵ Note that this is an *average* treatment effect, meaning that its magnitude may vary across students depending on their observed characteristics.

⁶ For instance, student who earned 11/20 would be expected to score 8.51/20 without tutoring, meaning the program helped them pass.

Table 8: Breakdown of results (pass or fail) for (non-)participants in peer tutoring for each course, based on the logit KNN1 matching with replacement ATT.

Courses	Neuropsychology	Cytology	Statistics
n tutored	141	85	65
% tutored*	17.1%	12.8%	11.9%
% of students who passed the course thanks to peer tutoring	31.9%	35.3%	26.2%
% of students who would have passed anyway	34.8%	23.5%	23.1%
% of students who would have failed anyway	33.3%	41.2%	50.7%

*The percentage represents the rate of participation in tutoring activities for the course.

From the last rows of Table 8, we can summarize outcomes among tutored students: on average, 31.1% seem to have passed thanks to tutoring, 27.1% did not need it to pass, and 41.7% would have failed even with support. This highlights both meaningful effectiveness for a sizable share of students and the presence of barriers that tutoring alone does not fully overcome, especially in Statistics. These results suggest that peer tutoring can substantially benefit students.

6.2.3. Sensitivity analysis

In addition to implementing various matching specifications to assess the robustness of our results, we conducted a Rosenbaum bounds sensitivity analysis. This method evaluates whether our findings could be explained by an unobserved confounder. Specifically, it examines how strong the influence of an unobserved variable on the odds of treatment would need to be for the estimated treatment effect to lose statistical significance. Put simply, it asks: *how large would hidden bias have to be to overturn the results?* (Rosenbaum, 2002).

The analysis was performed using the Wilcoxon signed-rank-based procedure implemented in the `rbounds` package in R. Table 9 reports the upper-bound p-values from the Rosenbaum sensitivity test under increasing levels of hypothetical unobserved bias, denoted by Γ . When $\Gamma = 1$, treatment assignment is assumed to depend only on observed covariates. As Γ increases, we allow for progressively greater influence of an unobserved factor on treatment assignment. For example, if $\Gamma = 2$, and if two individuals are identical on observed characteristics, then one might be twice as likely to receive the treatment due to unmeasured factors. Treatment effects that remain statistically significant at higher values of Γ are considered robust, while effects that lose significance at low levels of Γ are more likely to be sensitive to hidden bias (Olitsky, 2014; Rosenbaum, 2002).

Table 9: Sensitivity analysis: critical p-values from Rosenbaum bounds following the logit KNN1 matching with replacement for each course.

Γ	Neuropsychology	Cytology	Statistics
1	<0.001	0.003	0.004
1.1	<0.001	0.009	0.009
1.2	<0.001	0.020	0.019
1.3	<0.001	0.038	0.034
1.35	<0.001	0.051	0.044
1.4	<0.001	0.067	0.056
1.5	0.001	0.105	0.085
1.6	0.004	0.154	0.121
1.7	0.008	0.211	0.163
1.8	0.016	0.274	0.210
1.9	0.029	0.341	0.262
2	0.048	0.410	0.315
2.05	0.060	0.444	0.343
2.1	0.074	0.478	0.370

Note. Upper bound significance levels are reported here. The Rosenbaum bounds were computed using the rbounds package in R (Keele, 2025). Boldfaced values indicate where the effect is no longer significant at the 5% level.

In Table 9, bold values indicate the point at which the treatment effect becomes insignificant at the 5% level. The effect is considered robust as long as the upper-bound p-value remains below this threshold. Results show that the tutoring effect in Neuropsychology is the most robust to unobserved confounding as it remains significant up to $\Gamma = 2$ ($p = 0.048$). This implies that even if an unobserved factor doubled a student’s likelihood of receiving tutoring relative to another student with identical observed characteristics, the estimated effect would still hold. By contrast, the effect becomes insignificant at $\Gamma = 1.35$ in Cytology ($p = 0.051$) and at $\Gamma = 1.40$ in Statistics ($p = 0.056$), indicating moderate sensitivity to unobserved confounding. Overall, the Neuropsychology results appear highly robust, whereas the effects in Cytology and Statistics appear to be more sensitive to potential hidden bias (Olitsky, 2014; Rosenbaum, 2002)

7. Discussion and limitations

The present study evaluates the impact of a peer tutoring program on student success. While the literature on tutoring is already extensive, the context we examine differs substantially from most existing research. We focus on the FWB, in Belgium, where open-access policies and low tuition fees make higher education accessible to a large and diverse student population. The limited studies that focus on this setting rely primarily on descriptive analyses that do not account for selection bias. Using a sample collected from a French-speaking Belgian university, we apply propensity score matching to approximate the causal effect of tutoring on course success. This contribution is key in a changing political context, where peer tutoring programs are being placed at the center of higher education reforms. Our analysis therefore provides an evidence base for ongoing and future policy evaluations and decisions in this context.

Our findings align with the existing literature in showing that peer tutoring improves academic performance. The magnitude of the effect varies across courses, but ranges from approximately

1 to 2.5 points out of 20. We assess the robustness of our results by using different matching specifications and running sensitivity tests for hidden bias. Moreover, comparing descriptive tests, OLS and matching estimates confirms the presence of selection into tutoring and supports the use of matching as a more reliable estimator of the treatment effect in this setting.

Regression and matching estimates lead to highly consistent conclusions. In Neuropsychology, for instance, the effect is 2.42 points using regressions and 2.49 with KNN1 logit matching with replacement, with similarly close patterns in Cytology (0.76 vs 1.53) and Statistics (1.29 vs 1.73). Although participation in tutoring is voluntary and thus subject to self-selection, descriptive evidence indicates that tutored students tend to exhibit slightly weaker academic and socioeconomic profiles, consistent with negative selection into treatment. In this context, the alignment between regression and matching results together with the Rosenbaum sensitivity analysis, suggests that any remaining unobserved bias would need to be substantial (for Neuropsychology) or moderate (for the two other courses) to overturn our findings. While raw mean differences already pointed toward a positive association, the causal framework strengthens this result by adjusting for selection and demonstrates again that simple comparisons alone are insufficient to assess program effectiveness.

Our results also show heterogeneity in effect magnitude across courses, with the greatest impact in Neuropsychology. Although we did not investigate this difference further, several course-specific explanations may be considered: variation of skills involved, differences in tutor quality, and distinct session formats. These factors underscore the importance of course context and caution against assuming generalization of those results to other across disciplines. Regarding treatment effect heterogeneity, we find no statistically robust evidence of differences along the two dimensions considered, i.e. grade repetition status and scholarship status. This absence of significance likely reflects limited sample sizes within subgroups rather than the absence of underlying heterogeneity. Analyses based on larger samples of treated students would be better suited to determine whether such differences exist but remain undetected in the present analysis.

Having established the effectiveness of peer tutoring in the unique context of French-speaking Belgian higher education, policymakers now have a solid base to support and expand these programs. Tutoring, especially peer tutoring represents a low-cost intervention capable of generating positive academic gains in a context where student preparedness is heterogeneous and public resources are constrained. A key challenge, however, is encouraging participation among students, as low attendance remains a bottleneck for the development of the program's full potential. Institutions may benefit from actively encouraging engagement through targeted invitations, inclusion in the curriculum, and proactive early-semester invitations to increase engagement. Overall, peer tutoring represents a promising, scalable lever to reduce inequalities upon entry into higher education, provided participation barriers are addressed and implementation is thoughtfully designed.

7.1. Limitations

Several limitations should be noted. First, even though the propensity score matching allows us to approximate the causal effect in the absence of randomized assignment, conditional on observed characteristics, we cannot fully assert causality. While we control for key variables, additional data on student motivation, and cognitive abilities could further strengthen the analysis, and their omission may affect the unconfoundedness assumption. To address this concern, we rely on Rosenbaum bounds sensitivity analysis to evaluate the robustness of our estimates to potential hidden bias arising from unobserved confounders. Although this approach

cannot establish the absence of such bias, the results indicate that the estimated effects remain robust to moderate levels of unobserved confounding. Future studies should consider complementary methodologies, like the instrumental variables (Paloyo et al., 2016) and the difference-in-differences (DiD) approaches, to further isolate causal impacts. Although these methods come with their own challenges, such as the difficulty of identifying valid instruments (Becker & Ichino, 2002) or the fact that the context may not lend itself well to a DiD approach, they remain valuable tools for causal inference when used appropriately.

Second, potential spillover effects among students may bias results. In educational settings, knowledge transfer between tutored and non-tutored students could decrease the measured impact of tutoring. These spillovers represent an unintended but beneficial outcome of the program, complicating the estimation of its direct effects.

Finally, this study focuses on three courses within a single faculty at the ULB. While this scope allowed for a detailed and context-specific analysis, generalizing these findings to other faculties or universities should be done cautiously. Although the FWB presents distinct features, such as low tuition fees and open-access, that make the results not directly transferable to other academic contexts, our methodology remains applicable to those settings. Other studies have used comparable methods in different institutional settings, and this work contributes to that growing literature. While the results themselves may not be directly transferable, the methodological can be applied to other institutional contexts.

7.2. Implications for future research

Future research could build on this study by investigating the long-term effects of tutoring on student retention and graduation rates. Using a larger sample would also allow for an analysis of heterogeneous treatment effects across different student profiles. In our data, estimates suggest that students who did not repeat a grade may benefit more from tutoring, but this difference is not statistically robust. Should future evidence establish meaningful heterogeneity, an important implication would be that tutoring may insufficiently reach or support students who enter higher education with prior academic difficulties, such as those with a history of grade repetition. In that case, refining targeting strategies and strengthening communication about the program's potential benefits could help increase participation and effectiveness among the students most at risk. In contrast to Paloyo et al. (2016), who devised a randomized encouragement design using monetary incentives, future studies could consider alternative mechanisms, such as mandatory attendance, to determine how compulsion affects the impact of tutoring on academic outcomes. Nevertheless, such approaches may risk diminishing the spontaneous and self-driven engagement that characterizes voluntary participation.

Given the structural challenges faced by the FWB system, including limited funding and high dropout rates, improving success rates through student learning support programs is crucial. Encouraging struggling students to seek help and engage with tutoring remains a key priority, and this requires coordinated efforts between policymakers and universities. Better integration of tutoring within academic programs, targeted outreach, and early semester interventions could help ensure that support reaches those who need it most. Ultimately, Peer tutoring should be viewed as one tool among others to address student difficulties, rather than as a comprehensive solution, but in an open-access system with constrained resources, it stands out as a scalable strategy to reduce inequalities and support student success. If the goal is not only to open the doors of higher education, but to give all students the same chances of success, then investing in peer support programs is essential.

8. References

- Anfuso, C., Awong-Taylor, J., Curry Savage, J., Johnson, C., Leader, T., Pinzon, K., Shepler, B., & Achat-Mendes, C. (2022). Investigating the impact of peer supplemental instruction on underprepared and historically underserved students in introductory STEM courses. *International Journal of STEM Education*, 9(1), 1–17. <https://doi.org/10.1186/s40594-022-00372-w>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. <https://doi.org/10.2307/j.ctvc4j72>
- Arco-Tirado, J. L., Fernandez-Martin, F. D., & Hervas-Torres, M. (2020). Evidence-based peer-tutoring program to improve students' performance at the university. *Studies in Higher Education*, 45(11), 2190–2202.
- ARES. (2016). *Indicateurs de l'enseignement supérieur*. Statistiques. <https://www.ares-ac.be/fr/statistiques/indicateurs>
- Arias Ortiz, E., & Dehon, C. (2008). What are the Factors of Success at University? A Case Study in Belgium. *CESifo Economic Studies*, 54(2), 121–148. <https://doi.org/10.1093/cesifo/ifn012>
- Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2(4), 358–377. <https://doi.org/10.1177/1536867X0200200403>
- Black, D. (2015). Matching as a regression estimator. *IZA World of Labor*. <https://doi.org/10.15185/izawol.186>
- Borg, E. (2015). Classroom behaviour and academic achievement: How classroom behaviour categories relate to gender and academic performance. *British Journal of Sociology of Education*, 36(8), 1127–1148. <https://doi.org/10.1080/01425692.2014.916601>
- Bowman, N. A., Preschel, S., & Martinez, D. (2023). Does supplemental instruction improve grades and retention? A propensity score analysis approach. *The Journal of Experimental Education*, 91(2), 205–229. <https://doi.org/10.1080/00220973.2021.1891010>
- Bruffaerts, C., Dehon, C., & Guisset, B. (2011). Can schooling and socio-economic level be a millstone to a student's academic success? In *ECARES Working paper 2011-016*.
- Carvalho, R. G. G. (2016). Gender differences in academic achievement: The mediating role of personality. *Personality and Individual Differences*, 94, 54–58. <https://doi.org/10.1016/j.paid.2016.01.011>
- Dancer, D., Morrison, K., & Smith, M. (2007). Measuring the impact of a peer assisted learning program on students' academic performance in econometrics. *The Quantitative Analysis of Teaching and Learning in Higher Education: Forum Proceedings*, 19–42. <https://doi.org/10.1080/03075079.2014.916671>
- Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: A systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research*, 84(4), 609–639. <https://doi.org/10.3102/0034654314540007>
- Décret définissant le paysage de l'enseignement supérieur et l'organisation académique des études, Pub. L. No. 39681, Moniteur Belge 99347 (2013). https://www.galilex.cfwb.be/document/pdf/39681_004.pdf
- Décret définissant l'enseignement supérieur, favorisant son intégration dans l'espace européen de l'enseignement supérieur et refinançant les universités, Pub. L. No. 28769, Moniteur Belge 45239 (2004). https://www.galilex.cfwb.be/document/pdf/28769_003.pdf

- Décret organisant un encadrement différencié au sein des établissements scolaires de la Communauté française afin d'assurer à chaque élève des chances égales d'émancipation sociale dans un environnement pédagogique de qualité, Pub. L. No. 34295, Moniteur Belge 47476 (2009). https://www.galilex.cfwb.be/document/pdf/34295_024.pdf
- Dehon, C., & Leboutellier, L. (2025). A comparison between two systems of university education: Years of study versus credit accumulation. *Education Economics*, 33(2), 258–276. <https://doi.org/10.1080/09645292.2023.2301333>
- Dekker, I., Luberti, M., & Stam, J. (2023). Effects of supplemental instruction on grades, mental well-being, and belonging: A field experiment. *Learning and Instruction*, 87, 101805. <https://doi.org/10.1016/j.learninstruc.2023.101805>
- Dujardin, C., Meulewaeter, C., & Tulumoglu, H. (2023). *De l'enseignement secondaire vers l'enseignement supérieur en Fédération Wallonie-Bruxelles: Analyse des transitions à partir du Cadastre des Parcours Educatifs et Post-Educatifs* (p. 60) [Analyse quantitative]. Fédération Wallonie-Bruxelles. https://statistiques.cfwb.be/fileadmin/sites/ccfwb/uploads/documents/Zoom_Parcours_educatif_version_finale.pdf
- Endrizzi, L. (2010). Réussir l'entrée dans l'enseignement supérieur. *Dossier d'actualité de La VST*, 59, 1–23.
- Etter, E. R., Burmeister, S. L., & Elder, R. J. (2000). Improving student performance and retention via supplemental instruction. *Journal of Accounting Education*, 18(4), 355–368. [https://doi.org/10.1016/S0748-5751\(01\)00006-9](https://doi.org/10.1016/S0748-5751(01)00006-9)
- Fowler, G., Cope, C., Michalski, D., Christidis, P., Lin, L., & Conroy, J. (2018). Women outnumber men in psychology graduate programs. *Monitor on Psychology*, 49(11), 255–278.
- FWB, Fédération Wallonie-Bruxelles. (2023). *Retard scolaire*. Chiffres Clés. <https://statistiques.cfwb.be/enseignement/fondamental-et-secondaire/retard-scolaire/>
- Galand, B., Lafontaine, D., Baye, A., Dachet, D., & Monseur, C. (2019). Le redoublement est inefficace, socialement injuste, et favorise le décrochage scolaire. *Cahiers Des Sciences de l'Education*, 38. <https://hdl.handle.net/2268/234067>
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact Evaluation in Practice*. World Bank Publications.
- Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing Student Success: Assessing Supplemental Instruction Through Coarsened Exact Matching. *Technology, Knowledge and Learning*, 22(3), 335–352. <https://doi.org/10.1007/s10758-017-9317-0>
- Hardt, D., Nagler, M., & Rincke, J. (2023). Tutoring in (online) higher education: Experimental evidence. *Economics of Education Review*, 92, 102350. <https://doi.org/10.1016/j.econedurev.2022.102350>
- Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (1st Edition). Cambridge University Press.
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86. <https://doi.org/10.1257/jel.47.1.5>
- Jansen, E. P., & Bruinsma, M. (2005). Explaining achievement in higher education. *Educational Research and Evaluation*, 11(3), 235–252. <https://doi.org/10.1080/13803610500101173>

- Keele, L. J. (2025, July 23). *Perform Rosenbaum Bounds Sensitivity Tests for Matched and Unmatched Data*. Cran. <https://cran.r-project.org/web/packages/rbounds/rbounds.pdf>
- Lambert, J.-P. (2020). L'enseignement supérieur peut-il être à la fois excellent et démocratique? Une analyse comparée des systèmes. *Center for Research in Economics*, 5. <https://cerec.be/wp-content/uploads/2020/10/lenseignement-supecc81rieur-peut-il-ecc82tre-acc80-la-fois-excellent-et-decc81mocratique-.pdf>
- Lambert, J.-P. (2023). Les politiques de l'éducation et leurs effets sociétaux. *La Thérésienne*, 1. <https://popups.uliege.be/2593-4228/index.php?id=1682>
- Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal*, 46(3), 517–528. <https://doi.org/10.2224/sbp.7054>
- Li, I. W., & Dockery, Am. (2015). Does school socio-economic status influence university outcomes? *Australian Journal of Labour Economics*, 18(1), 75–94. <http://hdl.handle.net/20.500.11937/19422>
- Malm, J., Bryngfors, L., & Mörner, L.-L. (2012). Supplemental instruction for improving first year results in engineering studies. *Studies in Higher Education*, 37(6), 655–666. <https://doi.org/10.1080/03075079.2010.535610>
- Mason, H. D. (2018). Grit and academic performance among first-year university students: A brief report. *Journal of Psychology in Africa*, 28(1), 66–68. <https://doi.org/10.1080/14330237.2017.1409478>
- Morlaix, S., & Suchaut, B. (2012). *Analyse de la réussite en première année universitaire: Effets des facteurs sociaux, scolaires et cognitifs* (p. 34). Institut de Recherche sur l'Education, Sociologie et Economie de l'Education.
- OECD. (2016). PISA 2015 Results (Volume II): Policies and Practices for Successful Schools. *OECD Publishing, II*, 468. <https://doi.org/10.1787/9789264267510-en>
- OECD. (2021). *Education at a Glance 2021: OECD Indicators*. OECD Publishing. <https://doi.org/https://doi.org/10.1787/b35a14e5-en>
- Olitsky, N. H. (2014). How do academic achievement and gender affect the earnings of STEM majors? A propensity score matching approach. *Research in Higher Education*, 55(3), 245–271.
- Paloyo, A. R., Rogan, S., & Siminski, P. (2016). The effect of supplemental instruction on academic performance: An encouragement design experiment. *Economics of Education Review*, 55, 57–69. <https://doi.org/10.1016/j.econedurev.2016.08.005>
- Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiology and Drug Safety*, 20(6), 551–559. <https://doi.org/10.1002/pds.2098>
- Paume, J., Cauwe, J., & Gires, J. (2021). *Enquête sur les ressources économiques des étudiant-e-s* [Rapport analytique]. Université libre de Bruxelles, Observatoire de la Vie Etudiante.
- Peterfreund, A. R., Rath, K. A., Xenos, S. P., & Bayliss, F. (2008). The impact of supplemental instruction on students in STEM courses: Results from San Francisco State University. *Journal of College Student Retention: Research, Theory & Practice*, 9(4), 487–503.
- Rivera, P. E. (2022). *Generalized Propensity Score Methods for Assessing the Impact of Supplemental Instruction Attendance Frequency* [Master thesis]. San Diego State University.

- Rokusek, B., Moore, E., Waples, C., & Steele, J. (2022). Impact of Supplemental Instruction Frequency and Format on Exam Performance in Anatomy and Physiology. *HAPS Educator*, 26(2), 5–13. <https://doi.org/10.21692/haps.2022.013>
- Rosenbaum, P. R. (2002). Observational studies. In *Observational studies* (pp. 1–17). Springer New York. https://doi.org/10.1007/978-1-4757-3692-2_1
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>
- Salmon, D., Baillet, D., Boulvain, M., Cobut, B., Coupremagne, M., Duchâteau, D., Lanotte, A.-F., Dubois, P., Goemaere, S., Noël, B., Houart, M., & Slosse, P. (2009). Construction d'un outil d'évaluation de la qualité des actions d'accompagnement pédagogique. Synthèse d'échanges et d'analyse de pratiques professionnelles en Communauté française de Belgique. *Revue internationale de pédagogie de l'enseignement supérieur*, 25(2). <https://doi.org/10.4000/ripes.252>
- Stigmar, M. (2016). Peer-to-peer Teaching in Higher Education: A Critical Literature Review. *Mentoring & Tutoring: Partnership in Learning*, 24(2), 124–136. <https://doi.org/10.1080/13611267.2016.1178963>
- Stuart, E. 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), 1. <https://doi.org/10.1214/09-STS313>
- Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8), S84–S90. <https://doi.org/10.1016/j.jclinepi.2013.01.013>
- Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, 32(3), 321–345. <http://dx.doi.org/10.1007/BF00138870>
- UNESCO Institute for Statistics. (2024). *Enrolment by level of education*. Other Policy Relevant Indicators. <https://data.uis.unesco.org/>
- Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, 7(1). <https://doi.org/10.21037/atm.2018.12.10>
- Zhao, T., Perez-Felkner, L., & Hu, S. (2025). The impact of merit aid on STEM major choices: A propensity score approach. *Educational Evaluation and Policy Analysis*, 47(3), 939–959. <https://doi.org/10.3102/01623737241254842>