Economics
Discussion Paper Series
EDP-1509

EVALUATING NON-COMPULSORY EDUCATIONAL INTERVENTIONS - THE CASE OF PEER ASSISTED STUDY GROUPS

Ralf Becker

Maggy Fostier

October 2015

Economics

School of Social Sciences

The University of Manchester

Manchester M13 9PL

Evaluating Non-compulsory Educational Interventions - The Case of Peer Assisted Study Groups

Ralf Becker[a] and Maggy Fostier[b]

[a] Economics, School of Social Sciences, The University of Manchester[1]

[b] Faculty of Life Sciences, The University of Manchester

September 2015

Abstract

Evaluating the effectiveness of any teaching innovation is notoriously difficult. This is mainly due to the importance of self-selection. In this paper we attempt to evaluate the effectiveness of a Peer Assisted Study Scheme (PASS) in which higher year students (PASS Leaders) lead study groups of lower year students. While the PASS Leaders are explicitly not teaching the lower year students there role is to facilitate the learning of the students in the groups. The issue of self-selection is a prominent obstacle in the evaluation of the scheme's impact on the course units grades of the participating students as it transpires, in our case, that better students tend to attend more PASS sessions.

In this paper we use the data on two different PASS schemes to isolate the causal effect of PASS attendance on course unit grades. A number of techniques proposed in the econometric literature to evaluate causal effects are applied. As it turns out their application is not without shortcomings which are discussed throughout the paper. Across a range of methods, the findings demonstrate that students judged to be regular PASS attendees gain in the region of 5 marks (on a 100 marks UK grade scale) which is the equivalent of a third of a standard deviation. This is at the lower end of the effect sizes reported in the literature. This can be explained by the fact that we have a richer set of conditioning variables which is used to control for the effect of self-selection than previous studies evaluating the effectiveness of PASS.

1. Introduction

The evaluation of educational interventions is notoriously difficult. Teachers and lecturers often introduce new methods or aspects to their educational offerings on the basis of their own beliefs of what would benefit the learning process of students. The subsequent analysis of whether the implemented intervention was successful is usually problematic due to a number of reasons. The most problematic being the issue of selection bias. Factors, often unobservable, that determine whether a student participates at all, or engages fully, with the new educational element may also be somewhat, or even strongly, related to the educational outcome. This is problematic as the educational outcomes of a course (usually some form of grades) is typically used to evaluate the effectiveness of the educational intervention.

The innovation considered in this paper is that of the Peer Assisted Study Sessions (PASS)[2]. In this scheme higher year students (PASS leaders) guide, in weekly sessions, small groups of lower year

[1] Corresponding author: Ralf Becker, ralf.becker@manchester.ac.uk.

students (typically two higher year students for a maximum of 15 lower year students) through the learning and revision process for a course unit (or degree year) that they studied in their previous year.

Participation in PASS schemes is typically voluntary. We can measure participation as attendance records are kept. Such an attendance variable can be used as an 'intervention variable' which we ultimately want to relate to the educational outcomes, measured by exam grades. There is a substantial literature that attempts to evaluate whether attendance in PASS sessions indeed improves exam grades. However, the main empirical difficulty faced by researchers in this area is that of selection bias. If generally weaker (stronger) students tend to utilise PASS more, then simple effects estimators may well under- (over-) estimate the true (causal) effect. As concluded in a recent review of this literature (Dawson *et al.*, 2014), there is little to no evidence for a positive effect of PASS attendance on grade outcomes that dealt with this issue of self-selection comprehensively.

The econometrics and policy evaluation literature has, of course, developed a whole toolbox of methodologies to tackle issues like these. It is the purpose of this paper to review the most prominent of these in the context of evaluating the effectiveness of PASS. The key to an effective analysis of this question is to either be able to control access to the PASS program (e.g. via a randomised control trial) or to be able to model the self-selection process using variables that are independent of the actual grade outcomes. Only if either of these can be implemented can we guard against the most difficult of self-selection phenomena, selection on unobservables. In such a situation we have to accept that there are variables that simultaneously determine PASS attendance and grade outcomes, without the researcher being able to observe or sufficiently proxy for these variables. Such variables could be intellectual ability, motivation and ambition.

We attempt to apply both strategies, and in the course of the investigation highlighting where the majority of the evaluation literature falls short. Our empirical results show that the application of these strategies are not straightforward and indeed, in this instance, may not deliver the hoped for results robust to selection on unobservables. We also apply methods that allow for self-selection that is due to variables that can be observed or proxied. In the PASS evaluation literature papers that do attempt to correct for self-selection tend to assume that using variables like prior academic achievement can sufficiently proxy for self-selection. In our view this is an assumption not typically justified. However, that fact that one of our PASS schemes is a second year PASS scheme allows us to exploit quasi-Panel features of our data, a feature that is not typically available to evaluations of 1st year PASS schemes. When using this feature we establish an effect size of PASS attendance that is at the lower end of the effect size that has been reported in the empirical literature.

The remainder of this paper is structured as follows. A description of the basic features of PASS schemes (Section 2) is followed by a brief review of the relevant evaluation literature (Section 3) and the experimental setup utilised in this paper (Section 4). A description of our two datasets (Section 5) precedes the application of the different estimations strategies (Sections 6 to 10) and their respective results. Section 11 concludes.

[2] In the U.S. these schemes are often called Supplementary Instruction (SI).

2. Peer Assisted Study Sessions - PASS

In this paper we analyse two PASS schemes implemented at The University of Manchester. A scheme implemented for 1st year Life Sciences students and one for 2nd year students of a course unit called Econometrics.

The main feature of a PASS scheme are weekly meetings between a group of students (typically between 5 and 15 participants) currently enrolled as 1st year Life Sciences students or Econometrics students with higher year students (typically two PASS Leaders) in the respective area which have successfully passed the year/course unit. It is the purpose of these meetings to provide a supportive learning environment in which the students amongst each other and with the help of the higher year students improve their understanding of the course unit that is being supported by the PASS scheme. In the case of the Econometrics PASS scheme the supported course unit is a 2nd year econometrics course unit. In the case of the Life Sciences PASS scheme the supported course unit is a 1st year course unit (Genes and Evolution), but in addition it is an explicit aim of that particular PASS scheme to support 1st year students beyond that particular unit.

PASS Leaders, who volunteer for that role and are not being paid, receive the equivalent of 1 to 1.5 days of training for their role. That training is mainly generic training in group facilitation. It is made quite explicit in all communications with Leaders and students/participants that the PASS Leaders are not meant to be teachers of the particular course unit. Attendance in the weekly PASS sessions is voluntary.

The activities that take place during the weekly meetings are ideally driven by the needs of the participants. In general they rely on peer instruction and learning between the participants with the PASS Leaders taking a facilitation role. In regular, weekly sometimes fortnightly meetings between the PASS Leaders, the course unit lecturer and a PASS support officer employed by the University, problems arising from PASS meetings (often issues about attendance but also issues arising from difficult group dynamics or common problems with course content) are being discussed.

3. PASS Evaluation Literature

Universities invest resources into PASS for the following three reasons: improve retention in difficult course units, improve grades in difficult course units and increase graduation percentages[3]. Potentially such results can be achieved via a number of routes such as improving the understanding of the academic material, improved student motivation and/or happiness, improved study skills. Consequently there is a wide ranging literature that attempts to establish whether positive effects of PASS on any of the above three final goals, or any of the intermediaries can be established.
A good review of this wide literature is provided in Dawson at al. (2014).

[3] See for instance the homepage of the Supplemental Instruction unit at The University of Kansas, http://www.umkc.edu/asm/si/overview.shtml. In addition to these potential benefits for participants, it is widely acknowledged that the PASS Leaders (the higher year students) potentially have a lot to gain as well; in particular skills that may increase their employability.

As this paper is only concerned with the potential effect on course unit grades we will concentrate on this aspect with a particular focus on whether the crucial problem of self-selection has been addressed and if so how.

The general findings of Dawson *at al.* (2014) are that there are positive effects which appear to be in the order of grade improvements between 0.2 and 0.6 standard deviations. While the papers reviewed in Dawson *at al.* (2014) often establish statistical significance, they do point out that there isn't one paper that fully addresses the issue of self-selection and the resulting potential bias in the estimated effect. While the authors remain quite positive in their outlook they do point out that these issues make a causal interpretation of these results at best tentative.

Where papers recognise the potential for bias arising from self-selection they tend to either attempt to create a setup that mimics a randomised controlled trial, or they attempt to control for self-selection by using variables that are suspected to be correlated with student's motivation and/or ability.

Papers in the former category are Hodges *et al.* (2001), Gattis (2002) and Lewis *at al.* (2005). Hodges *et al.* (2001) analyse the results of three groups of students, one for which attendance of the PASS/SI sessions were compulsory, one which they label the voluntary PASS group and a final group that was not exposed to PASS. However, the last group appears to be a self-selected group (i.e. those students who just did not use the opportunity to attend PASS sessions. While it remains somewhat unclear how students selected into the group with mandatory PASS it was apparent that there were also other aspects of the delivery that was different for this section. The approach of Gattis (2002) seemed to be an interesting one, as treatment and control groups were all selected from a group of students that expressed a wish to utilise PASS sessions. However, they do not apply a random process to allocate groups of students to one of these two[4]. A random allocation to PASS is also described in Lewis *at al.* (2005). They find, on the basis of this experiment no significant effect of PASS, but conjecture that the constraints of the randomisation may have turned their scheme ineffective. A fourth paper that attempts this approach is Fostier and Carey (2007). As it turns out the allocation in treatment and control groups was also sub-optimal. In Section 6 we return to this paper as its results will be revisited here.

A more widely used is the attempt to control for self-selection by including explanatory variables that potentially proxy for the variables that are thought to influence a student's probability to make use of any PASS sessions. Such variables are motivation (seen to be positively correlated with PASS attendance) and academic ability (Dawson *at al.* (2014) and Dancer *et al.* (2007) discuss a variety of papers which find positive, negative and insignificant relationships between academic ability and PASS attendance).

The most prominent approach in this vein is to use pre-University grade achievement as a proxy for academic ability. In most papers reviewed in Dawson *at al.* (2014) the inclusion of such a variable, however, does not significantly alter the findings of an essentially positive impact of PASS attendance on course unit grades. McCarthy (1997) points out that using pre-University measures of academic ability may not correlate very well with the factors that determine why students may do

[4] As this paper does not seem to be publically available the information is somewhat sketchy and taken from the Dawson *at al.* (2014) review paper.

well at University (and choose to attend PASS)[5]. However, as many PASS schemes are 1st year schemes, often no additional University-level information is available. Some papers attempt to circumnavigate this by using grades from other (concurrent) course units (e.g. McCarthy, 1997, and Dancer *et al.,* 2007). This is not a very convincing strategy as PASS schemes, are meant to have positive effects that go beyond those on a particular course unit. If that is the case then such grades may well also be dependent on PASS attendance.

When Dawson *at al.* (2014) reviewed the literature (ending about 2010) they concluded that the evidence in the literature which attempts to establish whether PASS attendance has a significantly positive impact on course unit grades was lacking in a number of aspects: no paper had tackled the issue of self-selection in a fully satisfactory manner; the measure of engagement with the Programme (often a binary variable) is crude; studies were often conducted on smallish numbers and the methodologies used were too sketchy to aggregate the results in a meta-analysis.

[5] One reason why this may be true is that often Universities, in particular very selective ones, recruit rather uniform (as measured by pre-University grade achievements) student bodies. The differences may well be aspects that are difficult to observe such as the ability to adjust quickly to a new environment and the ability to work independently.

4. The Basic Experimental Setup

Consider that we have a variable that measures the educational attainment of the ith student in the particular course under consideration. This could be a scale variable (grade), a categorical variable (grade classes like, first, upper second, lower second, etc) or a binary pass/fail variable[6], y_i. We also have a value for the intervention variable, x_i for the ith student. This is likely to be a binary variable (participated or not) or a count (hours of attendance) variable. Further we consider k additional variables, q_{1i} to q_{ki} that represent observable student characteristics, such as the degree programme they are enrolled in, which year of studies they are, whether they are a home or overseas student etc.

In this Section we will describe the basic experimental setup that will serve our analysis throughout the paper. At the core lies what has been called Rubin's causal model (Angrist and Pischke, 2009). While we will observe one outcome for the attainment variable for each student, y_i, we conceptualise that there are two potential outcomes, y_{0i}, the outcome in case the ith student does not receive the intervention (x_i=0), and y_{1i}, the outcome in case the ith student does receive the intervention (x_i=1)[7]. Depending on whether the student did receive the intervention we will observe only one of these:

$$y_i = \begin{cases} y_{0i} & if\ x_i = 0 \\ y_{1i} & if\ x_i = 1 \end{cases} \tag{1}$$

We will further recognise that these outcomes, other than on the received (or not received) intervention, also depend on a set of covariates, $q_i = (q_{1i}, \ldots, q_{ki})$:

$$y_i|q_i = \begin{cases} y_{0i}|q_i & if\ x_i = 0 \\ y_{1i}|q_i & if\ x_i = 1 \end{cases} \tag{2}$$

We recognise that these (conditional on q_i) outcomes are random variables with associated probability distributions. What we are interested in here is the following difference

$$E[y_{1i}|q_i - y_{0i}|q_i] = E[y_{1i}|q_i] - E[y_{0i}|q_i]. \tag{3}$$

This is the difference between the expectation of the outcome if a student had received the intervention (or was treated) and the expectation of the outcome in case a student did not receive the intervention (was not treated). These expectations are conditioned on the set of covariates collected in q_i. This term is commonly called the average treatment effect (ATE). At times we may also be interested in the average treatment effect only for those that were treated (x_i=1), ATT:

$$E[(y_{1i}|q_i, x_i = 1) - (y_{0i}|q_i, x_i = 1)] = E[y_{1i}|q_i, x_i = 1] - E[y_{0i}|q_i, x_i = 1] \tag{4}$$

or alternatively the average treatment effect only for those that were not treated (x_i=0), ATN:

$$E[(y_{1i}|q_i, x_i = 0) - (y_{0i}|q_i, x_i = 0)] = E[y_{1i}|q_i x_i = 0] - E[y_{0i}|q_i x_i = 0]. \tag{5}$$

[6] It should be noted that the intervention may be applied to an entire degree year in which case this variable will represent something like a grade point average.

[7] For the time being we assume that the intervention variable is a binary variable. This is for notational simplicity and this framework can be extended to the case of multi-level treatments.

It is important to understand that ATT, the effect had by those treated, and ATN, the effect those that were not treated could have had had they be treated, need not be equal. The empirical difficulty arises from the fact that, as per (11) or (2), for any individual we only observe either y_{1i} or y_{0i}.

From this specification it is clear that we are interested in modelling conditional expectations. In what follows we will sometimes specify a specific functional relationship for this conditional expectation and at other times we will leave this unspecified. When we specify a functional form for the conditional expectation function we do so in the tradition of linear educational production functions (Todd and Wolpin, 2003):

$$E[y_i|q_i, x_i] = \beta_0 + \beta_1 x_i + \gamma_1 q_{1i} + \cdots + \gamma_k q_{ki} \tag{6}$$

In such a model the coefficients γ_1 to γ_k represent the strength of the marginal effects of an increase in the value of the respective student characteristics q_{1i} to q_{ki} (holding everything else equal, ceteris paribus). β_0 is a constant term and β_1 represents the size of the marginal effect of an increase in the value of the intervention variable[8]. We will estimate the unknown parameters from a regression model

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 q_{1i} + \cdots + \gamma_k q_{ki} + \varepsilon_i \tag{7}$$

ε_i is an error term that reflects the fact that the functional form of the proposed model is (almost certainly) misspecified, that individual grades will not adhere exactly to any specified model and, importantly in this context, it will also capture the impact on grades of variables, that are not explicitly included into the model.

The value for the coefficient β_1 is of particular interest in this study, as it will indicate whether the intervention in question (here PASS) has a significant impact. If the intervention did have a positive impact we would expect this coefficient to be positive (assuming that more student engagement is reflected in higher values for x_i).

In order to obtain useful parameter estimates (in particular unbiased estimates) through OLS a number of assumptions need to be met. Without going into any detail, the crucial assumption here is the zero conditional mean assumption, formally $E(\varepsilon_i|x_i, q_{1i}, \dots, q_{ki})$ =0. It is easiest to illustrate the meaning of this assumption by explaining how it is likely to fail. As mentioned above the error terms ε_i will, amongst others, capture grade impacts of factors that have not been included into the model. These factors may be constructs such as learning competence, intelligence etc. Let's think about two students who are equal in terms of all their attributes captured by characteristics q_{1i} to q_{ki} (e.g. both are 2nd year home students, both are students on the same programme etc.). One of these students, Emilie, is a slouch and the other, Marie, is what we would call a well engaged student. For this reason alone Marie is likely to obtain a better grade than Emilie. But if it is now also true that Marie, because she is an engaged student, engages with the intervention and Emilie does not (as she is a slouch), then it will appear as if it was the engagement with the intervention that made the difference (whether it made a difference or not). The key to understanding is to realise that it is the correlation of the unobserved characteristic (captured by the error term and often called unobserved heterogeneity) with the intervention variable that causes the problem.

[8] We will discuss the precise interpretation using an example in Section 7.

What this example illustrates is that the zero conditional mean assumption is unlikely to hold in our context, as in the case of most educational offers it is virtually certain that the unobservable variables (captured by ε_i) are related to both the educational outcome variable y_i and the intervention variable x_i, as long as the value of the latter is partly determined by the student's initiative. In this case the intervention variable is said to be endogenous, or formally $E(\varepsilon_i|x_i, q_{1i}, \dots, q_{ki}) \neq 0$, the zero conditional mean assumption is breached. This is obvious when the programme or intervention is voluntary. Even if the program is compulsory, but the intervention variable measures some aspect of engagement like attendance, the intervention variable will be endogenous. In such a situation it is difficult to identify the real effect of the intervention (x_i) on the grade, as measured by β_1, which is our aim.

There are a number of ways in which this issue can be tackled. The cleanest and most comprehensive way to address this issue is to undertake a randomised controlled trial (Section 6). If the intervention is allocated to individuals using a random scheme, then, by definition, the intervention variable (if it is a binary variable) is uncorrelated to anything else and therefore also the unobserved characteristic that drives the outcome.

Alternatively, one would attempt to include variables that proxy for the unobserved heterogeneity into your set of explanatory variables $q_{1i}, \dots, q_{ki}$. The intention is to strip the error term of the variation that makes it correlated to the intervention variable x_i (Section 7).

Three additional approaches may be available. These approaches go beyond the OLS estimation framework of the model in equation (1). The first explicitly models the individual's decision to partake in the intervention and then finds individual specific control groups. This method is called matching and, as the inclusion of proxy variables, relies on identifying a set of observable variables that allows the participation decision to be modelled. It is described in Section 8.

In order to allow for unobservable heterogeneity different approaches have to be employed. First one could use an instrumental variables approach. This is used to essentially estimate the parameters in model (1). However, one requires an additional variable that influences the participation decision, but does not impact on the grade outcome. Such a variable is called an instrument and it is not always available. The methodology is discussed in Section 9.

Lastly, if assessment data are available pre- and post-intervention, then one can use a panel data approach[9]. The approach here is to investigate whether the intervention has any significant impact on the change in the grade information. Often one would hope that students that participate in the intervention show, on average, a greater improvement than their peers that did not participate. This approach will work if the unobserved heterogeneity is not time varying. In that case these factors would impact on grades in all years and hence would not influence the change in grades which now is the dependent variable. A more detailed discussion can be found in Section 10.

After describing the main features of the empirical datasets used, the following sections will outline how each of these methods can be applied in the context of our problem of evaluating the

[9] Panel data are cross-sectional datasets in which you have repeated observations for the same individual (e.g. grade data for Year 1, 2 and 3.

effectiveness of PASS. The dependent variable will be the grade outcome for a student in a course unit that is supported by a PASS scheme.

5. Data

In the course of this paper we will use two different data sets.

Dataset 1: Life Sciences

This dataset is used in this paper as it has a specific feature that potentially allows for the identification of a causal effect of PASS attendance on course unit grades.

In the academic year 2005/06 the Faculty of Life Sciences at the University of Manchester launched their PASS scheme. As they only recruited enough PASS leaders to support half their Year 1 Bioscience cohort, they introduced PASS in a staggered fashion, where half the cohort (approximately 230 students) was automatically enrolled in PASS in semester 1 and the other half in semester 2. This dataset, therefore, contains some element of random assignment that can potentially be used to identify the effect of PASS attendance on course unit grades.

PASS was launched in semester 1 in association with a challenging course unit, “Genes and Evolution” (labelled Genetics below). Practically, this meant that every two weeks, students studied in PASS the content of the unit and associated problem sets given to the entire cohort. On alternate weeks, students could discuss any units they wanted. Therefore PASS could have a potential grade impact on four different course units.

The data available for Semester 1 of the academic year 2005/06 are as follows. Exam grades for four different course units y_{ji} (j = Genetics, Molecular Genetics, Body-Systems or Biochemistry)[10]. We also have information on whether the student was allocated a semester 1 PASS group (p_i=1 if yes, 0 otherwise), how many of the weekly PASS sessions a student attended, pa_i, and finally a categorical variable that indicates which one of many programmes the students were enrolled into, pr_i[11] .

In Table 1 we display summary statistics for these variables.

	genetics	**molgen**	**bodysys**	**biochem**	**p**	**pa**
Mean	42.31	49.68	61.29	55.87	0.49	1.19
Median	41	63	63	57		0
St Dev	17.30	18.88	14.88	16.10		2.05
Min	2	0	18	8	0	0
Max	89	90	97	89	1	9

Table 1. Summary Statistics for 314 students of ECON20110 Econometrics

Clearly the genetics course is the most difficult one (with an average grade of 42, which in the British system is just above the pass mark of 40). About half the students were allocated a PASS group, but

[10] Grades are on a traditional British system, out of 100. Grades above 90 are extremely rare. The pass mark is 40.

[11] The most popular programmes are Biology (25% of students), followed by Biomedical Sciences (12%), Biochemistry (11%) and Life Sciences (7%). Altogether there were 19 different programs, the smallest with only three students.

average attendance in PASS sessions was low, just over one week (1.19) out of a maximum of 9 weeks. This number increases to about 2.5 when only considering those students that were actually allocated into a PASS group.

Dataset 2: Econometrics

This is the workhorse dataset in this paper. We have data on 324 students that were enrolled in the Econometrics course unit at The University of Manchester in the academic year 2013/14. This is a second year course unit compulsory for all students that specialise in Economics. It is taught across the entire academic year but our analysis will be restricted to Semester 1. After excluding exchange students and students who did not attempt the final exam we are left with 314 students.

The main advantage of this dataset, when compared to most datasets used to assess the effectiveness of PASS, is that it pertains to a 2nd year course unit. This vastly increases the amount of information on students available and therefore facilitates conditioning on better set of covariates.

In this course unit there are 4 pieces of assessment. Two pieces of coursework with a 5% weight each, a mid-term multiple-choice test worth 20% of the course unit grade and a final exam worth 70%. In addition to the grade information for these pieces of assessment (*cw1*, *cw2*, *mt* and *y* for the exam grade which will serve as our main variable of interest) we also have the following Year 1 grade information available: the exam grades of the Statistics pre-requisite course[12] (*stats*), as well as the Year 1 Microeconomics (*micro*), Macroeconomics (*macro*) and year average (*Y1gpa*) grades. From the summary statistics we learn that mean and median grades of the Semester 1 exam are lower than either the equivalent mid-term exam statistics or indeed the Year 1 grades in statistics, micro- or macroeconomics. However, all grade variables have similar standard deviations.

The intervention considered in this Table is attendance in PASS sessions. Our base measure of PASS attendance comes from attendance records kept by PASS Leaders. During a semester there were a maximum of 9 PASS sessions and *pa* records how many of these were attended. The variable *ph* is a binary variable that indicates which student had a high PASS attendance (has attended at least 4 sessions). As it turns out 35% of students fall in that category. On average students attend 2.59 PASS sessions. Apart from a large group of students that do not attend any PASS session (127 students) , we find that the distribution of *pa* is almost uniform across values 1 to 8 with only 3 students attending all nine sessions.

In Table 2 we can see some summary statistics for this dataset.

	Sem1Ex = Y	**Sem1MT**	**pa**	**ph**	**Y1_GPA**	**stats**	**micro**	**macro**
Mean	56.51	65.19	2.59	0.35	63.09	65.67	70.69	60.78
Median	58.00	67.00	1.00	0.00	63.00	66.67	72.50	60.00
St Dev	15.86	13.96	2.85	0.48	8.78	14.79	13.45	13.94
Min	5.00	22.00	0.00	0.00	37.08	24.24	32.50	26.67
Max	85.00	100.00	9.00	1.00	83.54	100.00	100.00	93.00

Table 2. Summary Statistics for 314 students of ECON20110 Econometrics

[12] Two different Statistics course units can serve as pre-requisites and we also know which of these two paths students come from.

In addition to the summary statistics we also report summary information on some of the available categorical variables (Table 3); whether they are a home or overseas student (*home* = 1 if home student), their gender (*fem* = 1 if female), their stated ethnicity and the degree programme they are enrolled in.

Gender	Freq	Programme	Freq	Ethnicity	Freq
Female	120	**BEconSc**	77	**White**	116
Male	194	**BA(Econ)**	197	**Chinese**	103
		IBFE	30	**Asian**	70
Year	Freq	**PPE**	9	**Other**	25
2	224	**Other**	1		
3	90				

Table 3. Frequency distributions for categorical variables.

Students on the Econometrics course unit come from a range of different degrees. The BEconSc is a single honours economics programme that only admits students with very high pre-University Mathematics qualifications. Other programmes have similar high overall entry criteria but do not require high Mathematics entry requirements. The biggest amongst the other degrees is the multi-disciplinary BA(Econ) degree in which students can specialise in different subject areas. 104 students in this cohort specialised in Economics, whereas 44 specialised in Economics plus another area (such as politics or sociology). PPE is a Politics, Philosophy and Economics degree and IBFE represents the International Business, Economics and Finance programme.

Of the 314 students in our sample 224 take this course unit in their 2nd year. Students who take it in their 3rd year are either students that entered university with significantly lower Mathematics qualifications and required two years of Mathematics and Statistics course units to qualify for the Econometrics course unit or belong to a subset of students (BA(Econ) students) whose regulations allow that this particular course unit can be taken in the third year.

If a student required two levels of statistics prerequisite course units, we use the grade of the second of these as the grade for the *stats* variable, noting that this course unit was taken in the student's second year.

6. Randomised control trials

One obvious approach to break the link between the intervention variable and the error term ε_i is to use a randomised control trial (RCT), whereby a randomly selected sub-group of students is exposed to the intervention (treatment) while the other sub-group isn't [13].

This approach presupposes that an intervention can be denied to a control group. As one would usually want to apply such evaluations to interventions which have unknown outcomes, such denial of treatment is usually ethically acceptable, as the wide application in the area of medicine clearly

[13] This approach is very common in medicine but less so in other areas in particular in the social sciences. However, it has recently received renewed support in the context of education (e.g. Hutchinson and Styles, 2010 or Torgensen and Torgensen, 2013 and in general through the work of the Education Endowment Foundation which funds educational projects that deliver evidence) and more generically (Haynes et al., 2012).

demonstrates. In the context of PASS the implementation of a randomised control trial is somewhat more difficult. The advantages of PASS for students attending are widely believed to go beyond a direct grade impact. Institutions that often have invested substantial amounts of resources into the provision of PASS tend to feel that withholding the full range of these benefits from students (even if the actual impact on grades is uncertain) cannot be justified. For this reason there is often little willingness to undertake randomised controlled trials.

In some instances, however, the circumstances do enable institutions to use RCTs. This was the case when the Faculty of Life Sciences at the University of Manchester launched their PASS scheme in 2005/06 (dataset 1: Life Sciences). As they only recruited enough PASS leaders to support half their Year 1 Bioscience cohort, they introduced PASS in a staggered fashion, where half the cohort (approximately 240 students) was automatically enrolled in PASS in semester 1 and the other half in semester 2. This setup thus imposed a RCT for the Semester 1 exam outcomes[14].

PASS was launched in semester 1 in association with a challenging course unit, "Genes and Evolution" (labelled Genetics below). Practically, this meant that every two weeks, students studied in PASS the content of the unit and associated problem sets given to the entire cohort. On alternate weeks, students could discuss any units they wanted. Therefore PASS could have a potential grade impact on four different course units.

As p_i was the result of a random allocation (this will be qualified below) enables an estimation of the following regression model.

$$\boldsymbol{y_{ji} = \beta_0 + \beta_1 p_i + \beta_2 q_i + \varepsilon_i} \tag{8}$$

where *j* = Genetics, Molecular Genetics, Body-Systems or Biochemistry and q_i includes dummy variables for study programmes[15]. For each *j* we regress Model (8) with and without the additional covariates q_i. The estimates for β_1 are shown in the upper panel (Full Sample) of Table 4.

[14] The work in this section is a re-examination of data which were first described and analysed in Fostier and Carey (2007). Their finding was that PASS attendance had a significant and positive effect on grades (amongst other, not easily measurable benefits described). However, no careful consideration of selection bias was made which is why these data are re-examined here.

[15] Students were enrolled on 20 different degree programmes, some of them with very small enrolment numbers. The categories used are: Anatomy, Biochemistry, Biology, Biomedical Sciences, Biotechnology and enterprise, Life Sciences and Zoology. E.g. $q_{BioChem,i} = 1$ if the *i*th student is a Biochemistry student and 0 otherwise.

Course Unit	Genetics		Molecular Genetics		Body-Systems		Biochemistry
Full Sample, n	384		452		377		451
Without q_i	4.9991	***	3.3839	*	2.2712		0.9252
	(1.7507)		(1.7724)		(1.5470)		(1.5194)
With q_i	0.1230		3.0871		1.5964		-1.9268
	(2.5541)		(2.5428)		(2.3673)		(2.1659)
Subsample, n	159		169		116		451
Without q_i	1.0793		5.3930		6.3975	*	0.9252
	(3.3936)		(3.5976)		(3.3001)		(1.5194)
With q_i	1.3632		5.7167		7.0669	**	-1.9268
	(3.3785)		(3.7485)		(3.4408)		(2.1656)

Table 4. Estimates for β_1 and their standard errors (using White standard errors to guard against residual heteroskedasticity) in parenthesis. n is the effective sample size which differs as not all students are enrolled in all four course units. * /**/* significant at 1/5/10% significance level (using White standard errors). Estimates are reported for two sample. Full Sample: all 459 Life Sciences students. Subsample: Subsample: students from selected degrees, see text.**

From the included programme dummies we tend to find that, on average, Biochemistry students do significantly better than students from other degree programmes for all course units but for Body-Systems[16].

The strongest result of Table 4 seems to be the result for the Genetics exam. This is plausible as the PASS scheme was designed to support the student's learning in this particular course unit. Taken for face value the estimated coefficient of 4.9991 seems to indicate that, on average, students that were enrolled for Semester 1 PASS achieved about 5 extra marks. The p-value for this coefficient estimate is very small (0.0045) indicating that it is very unlikely that such a result would have occurred by chance. This interpretation would be correct if the PASS dummy variable p_i was uncorrelated to unobserved individual factors that are relevant to the grade outcomes.

The results in Table 4 are illuminating in that this apparently obvious positive effect of being part of the Semester 1 PASS (for Genetics and marginally for Molecular Genetics) disappears when programme dummies are included. This suggests that the apparent random allocation to the Semester 1 PASS scheme turns out to be correlated with the student's degree programme. This is clearly visible from the data in Table 5.

[16] Biology students do significantly worse for Molecular Genetics and Biochemistry, whereas Biomedical Sciences students do worse for Genetics.

Course Unit	Biochemistry	Biology	Biomedical Sciences	Life Sciences	Other
All Students (n = 459)	10%	22%	22%	7%	39%
Sem 1 PASS $p_i = 1$ ($n_1 = 225$)	19%	36%	0%	11%	34%
Sem 2 PASS $p_i = 0$ ($n_0 = 234$)	0%	8%	44%	3%	45%
All Genetics Students (n = 384)	11%	24%	17%	8%	40%
Sem 1 PASS $p_i = 1$ ($n_1 = 224$)	19%	36%	0%	11%	34%
Sem 2 PASS $p_i = 0$ ($n_0 = 160$)	0%	8%	41%	3%	48%

Table 5. Programme percentage for all students and for students that took the Genetics exam. For these two groups also the degree distributions amongst those that were allocated to a Semester 1 PASS group ($p_i = 1$) and those that were not ($p_i = 0$).

Disproportionally many students from Biochemistry and Biology were allocated to the Semester 1 PASS programme. As the decision to enrol in a particular programmes may well be correlated to underlying unobserved factors, which in turn may be correlated to exam outcomes, it turns out that the identification strategy used here does not work, as the intervention variable, p_i, is correlated with unobserved heterogeneity, which, in turn is captured by programme information. When the programme dummy variables are included, the PASS dummy variable, p_i, is estimated to be insignificant for all four course units.

Looking at the degree distributions in Table 5 it is apparent that students from some degree programmes were not randomly allocated, e.g. Biochemistry students who always were allocated to Semester 1 PASS and Biomedical Sciences students who were never allocated to Semester 1 PASS. For this reason we repeat the analysis of Table 4 with a subsample from which we exclude all students from degree programmes in which there is no mix of students enrolled and not enrolled in PASS. Degree programmes included in this subsample are Biology and Life Sciences, but also the smaller Biotechnology and Enterprise, and Zoology degrees. Amongst the remaining 171 students only 34 students were not enrolled into PASS sessions.

The results (estimates for β_1 in Model (8)) for this subsample are presented in the lower panel of Table 4. We can clearly see that any positive effect on Genetics and Biochemistry exams of being enrolled in a PASS group disappears. The earlier Genetics result seems to be purely driven by the fact that Biochemistry students tend to achieve higher grades in the Genetics exam. For the Molecular Genetics and Body Systems there seem to be somewhat larger effects of PASS enrolment for the subsample. But it is worth keeping in mind that the proportion of students not enroled in PASS is fairly small (< 20%).

So far this has been merely a cautionary tale that highlighted that a randomised controlled trial can deliver misleading results if the randomisation is not done carefully.

Of course, using the dummy variable p_i as the intervention variable is unsatisfactory for one more reason. It is a very crude measure of a student's exposure to PASS. A student who was allocated to

Semester 1 PASS may well have decided to not attend[17]. Therefore the attendance variable, pa_i, carries more information. However, it is likely that this variable is correlated with unobserved student heterogeneity. This would be the case if good students were more likely to attend more PASS sessions. In such a situation, using pa_i as the intervention variable is likely to result in an overestimated effect of PASS.

This will be demonstrated in the following analysis. Model (8) could be amended by using different treatment variables x_i, either the number attended weekly sessions, pa_i, or a dummy variable that reflects whether a student attended at least four of the weekly PASS sessions, ph_i (=1 if attendance is at least 4 sessions, 0 otherwise). As just discussed, it is very likely that this will produce a model that will produce biased estimates for the impact of PASS attendance as both p_i and ph_i are almost certainly positively correlated with unobserved heterogeneity and therefore any estimated effect of engagement with PASS on the exam grade would be overestimated. One strategy to counter this is to include other variables that sufficiently proxy for this unobserved heterogeneity. This dataset contains, other than the programme the student is enrolled on, no additional information. One could argue that perhaps the average grade in all non-genetics course units ($\bar{y}_{non-gen,i}$) can be used as such a proxy, but we note that this is a sub-optimal strategy as this is information not available at the beginning of the year and therefore could potentially also be affected by PASS attendance[18].

$$y_{genetics,i} = \beta_0 + \beta_1 x_i + \beta_3 q_i + \beta_4 \bar{y}_{non-gen,i} + \varepsilon_i \tag{9}$$

	Sample **Full Sample** (9**a**)	(9**b**)	(9**c**)	(9**d**)	**Sub-Sample** (9**e**)	(9**f**)	(9**g**)	(9**h**)
n	384	384	384	384	159	159	159	159
pa$_i$	1.759*** (0.440)	0.4858 (0.366)			2.239*** (0.591)	0.7346 (0.487)		
ph$_i$			8.751*** (2.628)	2.0557 (2.142)			12.724*** (3.691)	3.8060 (2.998)
$\boldsymbol{q_i}$	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.	Incl.
$\bar{\boldsymbol{y}}_{\boldsymbol{non-gen,i}}$		Incl.		Incl.		Incl.		Incl.
R^2	0.1002	0.4382	0.0891	0.4368	0.1077	0.4311	0.0991	0.4290

Table 6. Parameter estimates of Model (9) and their standard errors (using White standard errors to guard against residual heteroskedasticity) in parenthesis. n is the effective sample size. Models that include $\bar{y}_{non-gen,i}$ only use observations for which grades for all course units were available. * /**/* significant at 1/5/10% significance level (using White standard errors). Full Sample: all students 459 Life Sciences students. Subsample: students from selected degrees, see text.**

[17] Indeed, about 18% of students who were allocated a semester 1 PASS group never attended. The average attendance, amongst the students that attended at least once was 3 weekly session (of a maximum of 9 weekly sessions). More details on attendances can be found in Fostier and Carey (2007).

[18] This is particularly true as it is an explicit purpose of the PASS scheme to support student's development as independent learners and help student's to develop a support network that reaches beyond the particular course unit supported by the PASS scheme. Further it is likely that there are significant synergies between the learning process for genetics and other course units, which implies that it is impossible to interpret $\bar{y}_{non-gen,i}$ as a proxy for unobserved heterogeneity only. A similar strategy had been proposed by McCarthy (1997) and was also implemented by Dancer *et al.* (2007).

In the (9a) column of Table 6 the genetics grade is explained by the pa_i variable and the programme dummy variables. At face value the result would indicate that an additional weekly PASS attendance would deliver an average grade increase of 1.759 marks. This seems statistically significant and indeed substantial as it would suggest that attending all nine PASS sessions would, on average, increase the course grade by about 16 marks, which is roughly equivalent to one standard deviation. However, the inclusion of the $\bar{y}_{non-gen,i}$ variable reduces the size of this coefficient significantly and removes all statistical significance. The inclusion of this variable also tremendously increases the R^2 of this regression, which is not really surprising as good students tend to be good in all course units.

A similar pattern of results can be observed when we use ph_i as the intervention variable (a face value effect of 8.751 marks for being a high PASS attender is reduced to 2.056 and becomes statistically insignificant[19]) and indeed when we restrict the analysis to the subsample described earlier. The inclusion of the variable that is meant to capture (imperfectly as argued above) the heterogeneity amongst students removes all statistical significance from the intervention variable.

None of the specifications discussed above delivers a watertight result. But what is obvious is that the engagement with PASS is correlated with unobserved individual factors that are also very likely to be correlated to exam outcomes. Therefore we conclude, at this stage, that there is little empirical evidence that higher PASS attendance improves grade outcomes. In the following analysis we turn to a different dataset (Dataset 2: Econometrics). This has the advantage that it is a second year PASS scheme and we can observe more student specific information that may be relevant for a student's decision to attend PASS or not.

7. The Use of Proxy Variables

In the previous section we argued that the presence of unobserved heterogeneity complicates the impact evaluation to a degree that makes it impossible to evaluate whether PASS attendance has a significant positive affect on a student's exam grades. This is, of course, a very common problem in the literature evaluating the effectiveness of educational innovations or programmes[20].

One way to tackle this problem is by including variables that proxy for this unobserved heterogeneity. Good examples for this approach are the work of Angrist and Lavy (2001) who attempt to evaluate the effectiveness of teacher training and the work by Andrietti (2014) who attempts to evaluate whether lecture attendance has any significant effect on student's educational outcomes as assessed by examination grades on a particular second year course unit. The latter is a problem that is very similar to that of evaluating PASS attendance as lecture attendance is, in most universities, voluntary, as is PASS attendance, and very likely to be highly correlated with other unobserved factors that are relevant for a student's performance. In other words, good students are more likely to attend lectures and obtain good grades (which they would have obtained even without attending lectures). It is therefore entirely predictable that there will be a positive

[19] An effect of 8.751 corresponds to an effect equivalent to half a standard deviation. An effect comparable to the size of effects commonly cited in the literature (see Dawson *et al.*, 2014).

[20] The following sections will heavily lean the books by Angrist and Pischke (2009, 2015) which provide excellent reference material and Blundell *et al.* (2005) who discuss different methodologies in the context of evaluating the effect of schooling on earnings.

correlation between lecture attendance and grade outcomes, a correlation that cannot be, by itself, mistaken for causation.

Andrietti (2014) combines lecture attendance and grade outcome data with additional, mainly administrative, data about students, such as age, gender, high-school grades, year 1 university grade point average, educational and family background and some survey data that are meant to proxy for a student's motivation. In contrast to the grade information for other course units, which was used as a proxy variable in the previous section, these variables can lay a better claim to proxy for unobserved heterogeneity as neither of Andrietti's proxies is potentially endogenous.

Indeed the inclusion of the proxies reduces (in size and statistical significance) the impact of lecture attendance although this remains marginally statistically significant. However, as discussed by Andrietti, it is unlikely that the proxies used, of which the year 1 grade point average seems to be the most powerful, do proxy the unobserved factors to an extend that removes the problem of potential correlation between the lecture attendance variable and the unobserved error term.

In the context of evaluating whether PASS attendance has any significant impact on exam grades we will pursue a similar strategy. What is needed is student specific information that relates to the general qualities of a student which will significantly affect both, the student's grade and make the student attend more PASS sessions, the confounding factors. When students come to University there is little such information available. This is particularly true for UK universities which have a large proportion of international students who come with a wide variety of different High School diplomas which are not easily comparable. For this reason we will focus on Dataset 2 which relates to a PASS scheme that is attached to a 2nd year course unit. A significant amount of information would have been accumulated on these students during their first year at university.

In addition to the exam information, y_i, our dependent variable, we have again information on the PASS attendance, ph_i (binary variable indicating high attendance) and pa_i (a count variable measuring the number of weeks a student attended PASS sessions). These will be used to measure the potential impact of PASS attendance. As PASS was made available to all students, this dataset has no equivalent to the p_i variable used earlier.

We will estimate the following model[21]

$$y_i = \delta_0 + \delta_1 ph_i + \delta_2 q1_i + \delta_3 q2_i + \varepsilon_i \quad (10)$$

where $q1_i$ and $q2_i$ are vectors of variables that are either directly relevant for explaining the variation in the econometrics exam grade ($q1_i$) or are used as proxy variables to capture unobserved heterogeneity amongst the students ($q2_i$). The variables that are included in $q1_i$ are the grade in the pre-requisite statistics course[22] for econometrics and a measure of the student's tutorial attendance in econometrics. For $q2_i$ we use the micro- and macroeconomics course unit grades, Year 1 grade point average, whether the student had an economics A-level[23], whether the student had a

[21] Here we use ph_i as the measure of PASS attendance, but in the analysis we also use pa_i.
[22] All grades used here are final exam grades only.
[23] A-level is the highest High-School qualification in a subject. For overseas students the admissions office judged whether a student had an A-level equivalent.

Mathematics A-level[24], what degree programme the student was enrolled in, whether the student was a home or overseas student, the ethnicity and whether the student took the econometrics course unit in year 2 or 3. None of these should have any direct influence on the econometrics exam grade (beyond the grade information available from the pre-requisite statistics course unit), but may serve to capture some variation in student attitude. To illustrate that self-selection is an issue that is relevant, it is useful to note that the Year 1 GPA is about 5 marks higher amongst those students that are deemed high PASS attenders ($ph_i = 1$) than amongst those that are not. This difference is highly statistically significant and a similar difference arises between the medians. There is, therefore, evidence that better students are more likely to attend PASS sessions.

We also know the students' gender. We do not include this variable as a proxy variable here as we will consider it as an instrumental variable in the following Section.

In Table 7 we report OLS parameter estimates and their p-values (in parenthesis). The p-values are calculated using t-statistics that use White standard errors to guard against the potential presence of heteroskedasticity[25].

Model (10)	**(10a)**	**(10b)**	**(10c)**	**(10d)**	**(10e)**	**(10f)**
N	314	314	303	303	303	303
ph_i	9.764 (0.000)		6.133 (0.000)		4.220 (0.001)	
pa_i		1.590 (0.000)		1.002 (0.001)		0.817 (0.001)
$q1_i$			Included	Included	Included	included
$q2_i$					Included	included
R^2	0.0862	0.0815	0.295	0.293	0.522	0.526

Table 7. Parameter estimates of Model (3) and their p-values (using White standard errors to guard against residual heteroskedasticity) in parenthesis. n is the effective sample size.

In models (10a) and (10b) (columns 1 and 2 in Table 7) no control variables are included. This is, of course, inappropriate, but it is done here to demonstrate what a naïve evaluation would produce. These results suggest that a student with high attendance could, on average, gain 9.8 marks, the equivalent of one degree class. Or alternatively (Model 10b) a student attending a maximum of 9 PASS sessions could gain about 14.3 marks compared to a student not attending any PASS session. In what follows we shall call this the "raw" effect.

The inclusion of covariates in $q1_i$ makes a significant difference to the estimated effect of PASS attendance. The effect size drops significantly (by about a third for both ph_i and pa_i) but remains statistically significant. Further, the R^2 of the regression model improves significantly. Amongst the two variables included in $q1_i$ it is the pre-requisite statistics grade that is accountable for most of this difference. The fact that the effect size is reduced is indicative for a positive correlation of PASS attendance and statistics exam grade (the correlation coefficient between statistics grade and pa_i is

[24] Students that enter University with only a GCSE qualification in Mathematics will only be able to take the Econometrics course unit in Year 2 after having gone through additional Mathematics and Statistics training.
[25] Tests for residual heteroskedasticity provide some evidence for residual variance that is larger for students with lower pre-requisite grades and lower PASS attendance (results not shown here).

0.19). This is indicative of the core problem at the heart of the analysis, i.e. the finding that good students tend to use PASS more than struggling students.

In order to proxy as much as possible of the unobserved student quality (which we suspect to be positively correlated to PASS attendance and to the econometrics grade) we add the variables in $q2_i$ to the regression model. The results are shown in columns (10e) and (10f) of Table 4. As the R^2 of these regressions shows another significant improvement it becomes obvious that these variables capture some important factors that influence the student's grades. The inclusion of these variables reduces the estimated PASS attendance parameters further; yet, they remain positive and significantly so. Their magnitude indicates that a student that attends all 9 PASS sessions can, everything else being equal, expect to gain about 9 marks compared to a student that does not attend any PASS session. In fact, the only variable in $q2_i$ that contributes to this significantly is the Year 1 grade point average[26].

Recall that the purpose of the inclusion of the variables in $q2_i$ was to remove any systematic factors that influenced both the student's decision to attend PASS and their exam grade from the regression's error term[27]. While the variables included in $q2_i$ cannot lay claim to eliminating all relevant heterogeneity from the regression residual, we are confident that they have captured a significant proportion. Before we discuss strategies that continue to tackle this particular issue, we shall investigate the origin of the potential positive effect of attending PASS sessions in somewhat more detail. In particular we are interested in whether the positive effect that seems to prevail is an effect that applies to all PASS groups or whether there are significant differences between groups. As we do have information on which group students attended, we can use this for an attempt to differentiate between the groups.

We shall use the estimated model (10f) as a starting point. Instead of including the PASS attendance variable directly we shall interact it with 18 group dummy variables, $dg_{i,j}$ which takes the value 1 if the *i*th student is a member of the *j*th PASS group and 0 otherwise. We then estimate the following model:

$$y_i = \delta_0 + \delta_{1,1}(pa_i * dg_{1,i}) + \dots + \delta_{1,18}(pa_i * dg_{18,i}) + \delta_2 q1_i + \delta_3 q2_i + \varepsilon_i \tag{11}$$

The estimate to $\delta_{1,j}$ will then indicate by how much, everything else being equal, a student, being a member of the *j*th PASS group, should expect one extra PASS session to increase the final exam grade[28]. In Table 8 we report the parameter estimates for $\delta_{1,j}$ for j = 1, …, 18.

[26] Qualitatively the same results as in columns (3e) and (3f) can be obtained when only including the Year 1 average instead of the entire list of $q2_i$ variables. Results are available on request.
[27] A good discussion of the underlying models applied to a similar problem is found in Angrist and Lavy (2001).
[28] Note that a student that was not allocated to any PASS group (26% of the cohort) would not be captured by any of the group dummy variables.

Model (11)	$R^2 = 0.5055$				
$\hat{\delta}_{1,1}$	0.6187	$\hat{\delta}_{1,7}$	-0.1270	$\hat{\delta}_{1,13}$	1.1469
$\hat{\delta}_{1,2}$	0.8395	$\hat{\delta}_{1,8}$	0.7037	$\hat{\delta}_{1,14}$	1.5962***
$\hat{\delta}_{1,3}$	1.9369***	$\hat{\delta}_{1,9}$	0.4469	$\hat{\delta}_{1,15}$	0.4804
$\hat{\delta}_{1,4}$	0.0513	$\hat{\delta}_{1,10}$	0.1781	$\hat{\delta}_{1,16}$	3.0316***
$\hat{\delta}_{1,5}$	0.2873	$\hat{\delta}_{1,11}$	0.4387	$\hat{\delta}_{1,17}$	1.3865**
$\hat{\delta}_{1,6}$	1.7218***	$\hat{\delta}_{1,12}$	1.4373***	$\hat{\delta}_{1,18}$	1.2715

Table 8. Parameter estimates of Model (4). * /**/* significant at 1/5/10% significance level (using White standard errors).**

It is apparent from Table 5 that the estimated effects vary quite dramatically from -0.1270 to 3.0316. And while a few individual groups seem to be associated with individually significant effects (in particular groups 3, 6, 12, 14, 16 and 17, the majority of groups have individually insignificant effects. But a note of caution is in order. The individual effects are estimated with a large degree of uncertainty as each group has between 12 and 15 members only. In fact a hypothesis test that tests the null hypothesis that all coefficients are of equal size cannot be rejected at any of the usual significance levels for this reason (F-test = 0.826 with p-value = 0.6625). While this last analysis illustrates that there could well be large variations in the effectiveness of individual PASS groups, the data are not really rich enough to establish such a result with confidence.

In this section we used a number of available variables to proxy for student heterogeneity. By allowing for such observed heterogeneity we reduced the *raw* effect of PASS participation by about 40-50%, depending on our measure of PASS participation. The information that was responsible for this reduction is mainly prior grade information (for a specific pre-requisite and the Year 1 average). It is likely that there is additional unobserved student heterogeneity that is relevant for the grade and also correlated with PASS attendance. For this reason, even the reduced effect of PASS attendance needs to be interpreted carefully as it is most likely still overestimated, assuming that the unobserved variation in students is positively correlated with both, PASS attendance and grade outcomes.

8. Matching Estimators

An alternative approach to allow for selection bias is to use matching estimators. Before describing details of these estimators it should be pointed out, that the approach taken here, while being different in some important details, achieves conceptually the same as the use of proxy variables described in the previous section. We control for selection bias that can be explained with the use of observed variables. We also say that the selection bias arises from observed heterogeneity. For the purpose of this paper we restrict the application of matching estimators to the binary treatment variable ph_i which indicates whether a student attended at least 4 of the 9 PASS sessions.

What distinguishes this approach from the use of the proxy variables is that matching estimators do not require the relation between the variables that capture the unobserved heterogeneity and the outcome variable to be linear.

The main idea behind matching is best explained by reference to the definition of the average treatment effect of the treated (ATT) in equation (4) which is replicated here

$$\delta_{ATT} = E[y_{1i}|q_i, ph_i = 1] - E[y_{0i}|q_i, ph_i = 1] \tag{12}$$

The first expectation, can be approximated by the sample grade average of those that received treatment ($ph_i = 1$) as for these $y_{1i} = y_i$, the observed grade. However, the second term is the counterfactual for the same group of students and, of course, y_{0i} is unobserved. What the matching estimator achieves is that it constructs a comparison group, from which to estimate $E[y_{0i}|q_i, ph_i = 1]$. The first crucial assumption here is that selection is independent of the potential outcomes, y_{0i} and y_{1i}, conditioned on covariates q_i. This is the same assumption that is used to justify the proxy variable approach in the previous section. Under this assumption we can replace $E[y_{0i}|q_i, ph_i = 1]$ with $E[y_{0i}|q_i, ph_i = 0]$ which in turn allows us to estimate this quantity from those observations that did not receive the intervention.

The issue a matching algorithm then has to tackle is to select a comparison group from all those observations with ($ph_i = 0$) that is comparable in terms of covariates, q_i, to those of the treated group[29]. A variety of different ways exist to assemble these comparison groups. The most intuitive way to do this is to select a comparison observations that match those in the treatment group in terms of their values for the conditioning variables in q_i. If the variables are continuous variables it will be difficult or impossible to find exact matches in which case one has to find observations that are close in some sense. There are different ways in which closeness can be defined. We implement an algorithm by Sekhon (2011) that attempts to find those comparison observations that match the empirical properties of the comparison group as close as possible to that of the treatment group. This estimator will be labelled *CovMatch* in the results table.

Perhaps the most common approach to deal with the multi-dimensional nature of closeness comparison is to use propensity scores to match. What is required here is a model to estimate the probability that an individual chooses treatment given the available covariate information, $P(ph_i = 1|q_i)$. Such models are standard in the econometric literature and the model used here is the logit model. Once estimated probabilities are available, then matching is achieved by matching the propensity scores of those in the treatment group to students in the control group. The resulting matching estimator is labelled *PropMatch* in the results table.

When applying either of the matching algorithms one needs to make decisions that are quite similar to the choice of explanatory variables in a regression framework. For the *CovMatch* algorithm one needs to decide which covariates ought to be matched on and for the *PropMatch* estimator one needs to decide which variables are to be used to model the propensity scores. Here we choose *Micro*, *Gender* and *Y1_GPA* as the variables to model the probability to be a high pass attender, $P(ph_i = 1|Micro_i, Gender_i, Y1\_GPA_i)$ which is then used as the matching variable in the *PropMatch* estimator[30]. The *CovMatch* estimator is chosen to minimise the empirical differences between the following variables: *Micro*, *Gender*, *Y1_GPA*, *Eth_cat* and *Year3*.

[29] This paper will not go into much technical detail. There is an extensive literature and thorough discussions can be found in Morgan and Winship (2007) and Imbens (2004). The matching algorithms used are those provided in the matching package written by J.S. Sekhon for the R Statistical software, Sekhon (2011).
[30] These are the variables that proved statistically significant in the Logit model.

The resulting estimators are shown in Table 9. In the ATT column we can see that the estimated effect of being a high PASS attender is an expected grade increase of 5.3 (*PropMatch* estimator) and 3.5 (*CovMatch* estimator) marks. The standard errors reveal that these effects are statistically significant at the 1% and 5% significance level respectively.

Matching Estimators	ATT			ATN			ATE		
	$\hat{\delta}$		$se(\hat{\delta})$	$\hat{\delta}$		$se(\hat{\delta})$	$\hat{\delta}$		$se(\hat{\delta})$
PropMatch	5.2576	***	(1.8200)	6.3268	***	(2.1346)	5.9557	***	(1.7836)
$\boldsymbol{CovMatch}$	3.4862	**	(1.7247)	5.0098	***	(1.7729)	4.4809	***	(1.5450)

Table 9. Matching estimators of the effect of being a high PASS attender. Standard errors are bootstrapped standard errors. * /**/* significant at 1/5/10% significance level.**

The description of the matching estimator was illustrated for the case of the average treatment effect of the treated (ATT), but the matching estimator can also be used to estimate the average treatment effect of the non-treated (ATN) and the average treatment effect (ATE). The differences arise from conditioning over different distributions of the covariate variables.

The results for ATN and ATE are also shown in Table 9 and we can see that for both matching estimators the estimated effect is somewhat larger for the non-treated students. In other words, the students that did not participate in PASS could have benefitted more from attending PASS than the students who were high PASS attenders. However, when compared to the estimated standard errors it also obvious that these differences are not statistically significant. The ATE estimators are located between the ATT and ATN estimators as they describe a mixture of both effects. In general we find the *CovMatch* estimators to be somewhat smaller than the *PropMatch* estimators.

It is also worth comparing these results to those obtained from the approach adopted in the previous section. In Model (10e) (see Table 7) we found the marginal effect of being a high PASS attender to be 4.22 grades. This value is clearly in the same ball-park as the matching estimators[31]. This is not really surprising when recalling that both approaches are designed to deal with selection on observables and indeed that a regression approach can be interpreted as a matching estimator with weight determined by the variances of the treatment probability (Angrist and Pischke, 2009, pp75-76). As a consequence of this interpretation it is also apparent that the regression estimate cannot be interpreted as either ATT or ATN or ATE.

When applying matching estimators it is customary to compare the empirical properties of the treatment and control groups to ensure that the matching indeed generated comparable treatment and control groups. This can be done in a variety of ways. Here we will look at mean values for covariates and where covariates are continuously distributed we also report p-values for Kolmogorov-Smirnov (KS) test statistics that test the null hypothesis that the respective covariate is equally distributed in the treatment and control groups. Results are shown in Table 10.

[31] When evaluating a slightly different peer assisted scheme Ward *et al.* (2013) also found that matching and regression estimators delivered very similar impact estimates.

Matching Estimators		Before matching		After matching PropMatch	CovMatch
covariate	Property	Treatment	Control	Control	Control
		$ph_i = 1$	$ph_i = 0$	$ph_i = 0$	$ph_i = 0$
Micro	mean	72.414	69.775	69.883	72.564
	KS (p-value)		0.099	0.071	0.388
Stats	mean	69.467	63.658	67.717	70.227
	KS (p-value)		0.004	0.136	0.757
Y1_GPA	mean	66.024	61.533	64.929	65.934
	KS (p-value)		0.000	0.447	0.789
Eth_cat(Chinese)	mean	0.3028	0.3415	0.3891	0.3211
Eth_cat(White)	mean	0.3945	0.3561	0.3249	0.4038
Eth_cat(Other)	mean	0.1101	0.0634	0.0757	0.0734
Year 3	mean	0.3395	0.2585	0.4121	0.3670
Gender(Male)	mean	0.5138	0.6732	0.4602	0.5138

Table 10. Properties of treatment and control groups. *Mean* is the mean value of the relevant covariate in the treatment and control group. For the discrete dummy variables, Eth_cat(Chinese), Eth_cat(White), Eth_cat(Black), Year 3 and Gender(male) the values indicate the relevant proportions. E.g. there are 51% males in the treatment group. Where covariates are continuous we also report the p-value of a Kolmogorov-Smirnov test that tests for differences in the distribution of the relevant covariate. The null hypothesis is that there is no difference and the p-values are bootstrap p-values.

First it is useful to look at the results in the "Before Matching" columns for treatment and control groups. Comparing the two columns we can see that in particular the mean values for the variables *Stats*, *Y1_GPA*, *Year 3* and *Gender(Male)* differ. This is not surprising as we are aware that better students are more likely to be high PASS attenders and therefore be in the treatment group ($ph_i = 1$). On average their Year 1 average is 4.5 grades and their prerequisite statistics grade 6 grades higher than that for students in the control group. This is, of course, just an indication for selection on observable variables, the issue matching estimators are to tackle. For the continuous variables (*Stats* and *Y1_GPA*) we can also see that the KS tests clearly reject the null hypothesis that these covariates are equally distributed in treatment and control groups. Interestingly we also note that the proportion of male students (which is 62% across the whole class) in the treatment group is significantly smaller than in the control group. Male students are significantly less likely to be high PASS attenders than female students.

In the "After Matching" columns we can see the empirical attributes of the matched control groups. In the column labelled "PropMatch" we can see the control group properties when using propensity scores in the matching algorithm. Most notably the KS test statistics now indicate that the distributions of grades (*Stats* and *Y1_GPA*) are not significantly different between treatment and control groups (p-values of 0.136 and 0.447 respectively). This is also reflected in smaller differences between the mean grades. We also see that the proportion of *gender(Males)* is more similar now (46% in matched control and 51% in treatment group). This is no surprise as the *gender* variable is a statistically significant variable in the propensity score model[32]. As the ethnicity variables do not contribute to the variation in propensity scores, this explains why the ethnicity group proportions in

[32] No details of this model are reported. The variables that determine variation in the propensity scores are the *gender* and *Y1_GPA* variables.

the propensity score matched control groups are not closer to those in the treatment group when compared to those in the unmatched control group.

The *CovMatch* algorithm produces a control group that is even closer in covariance properties to the treatment group. This is, of course, no surprise as that algorithm chooses the control group in a manner that minimises these differences (see Sekhon, 2011). In particular the grade variables and the gender proportions are now extremely close between the matched control and the treatment group. As we saw in Table 9, the effect estimators resulting from covariance matching were between 20 and 35% smaller than those from the propensity matching algorithm.

9. The Use of Instruments

We discussed in Section 6 that the cleanest way to disentangle any unobserved heterogeneity from the treatment/intervention variable is to assign the treatment randomly to a subgroup of students. If however, there is either a strong bias that a treatment will work or a treatment has already been offered widely such a strategy may be undesirable. This is certainly the case with PASS schemes. University administration and ethics committees may be unwilling to contemplate a randomised controlled trial as PASS schemes are widely believed to have a number of benefits (for students attending and indeed delivering the PASS sessions) which may go beyond the examinable grade impact.

In such cases an alternative, yet standard econometric technique to tackle the issue of endogeneity of the treatment variable, Instrumental Variable (IV) Estimation, may be available. What is required in this context is one (or several) instrumental variable(s). They need to have the following properties. *First*, they are irrelevant for the outcome (here exam grades) and hence can be excluded from your econometric model for exam grades. *Second*, they are uncorrelated to the regression error term, i.e. uncorrelated to the unobserved heterogeneity that causes the above problem of endogeneity. *Third*, it is correlated with the endogenous intervention variable, the measure of PASS attendance.

If such variable(s) were available, the parameters in model (10) could be estimated by IV[33]. The practical problem of IV is to identify suitable instruments. In the context of the current problem we require student specific information that explains why students attend PASS sessions without this information having any direct explanatory power for the exam grade.

Instruments tend to be very difficult to come by and often they arise from natural experiments. No such lucky circumstances are available in the context of this study. Sometimes a researcher can create useful instruments. Here we attempted to create an instrument using the following scheme. The course unit lecturer send an email to a randomly selected sub-group of students which attempted to encourage these students to attend PASS sessions. Conditional on such an email having no motivational effect beyond increasing the probability to go to PASS sessions[34] we can then

[33] Technical details of instrumental variables estimation can be found in any econometrics textbook, e.g. Wooldridge (2013). A more substantive and intuitive explanation can be found in Angrist and Pischke (2015).
[34] In other words, to be a useful instrument the email should not increase the general motivation of students to work harder/better for this course units. It should merely motivate them to go to PASS sessions.

use the dummy variable, enc_i (which takes a value of 1 if the *i*th student received that email and 0 otherwise) as an instrument.

As mentioned earlier we also have gender information which turned out to be correlated to PASS attendance. It was statistically significant in the logit model used to model propensity scores. From Table 10 we could also see that males were less likely to be high PASS attenders than females. Whether gender can be used as a potential instrument depends on whether we can justify the exclusion of a gender variable from the basic model (10) as an explanatory variable. While it turns out that gender does not appear to be a statistically significant addition to that model, it is important to note that this is not sufficient to conclude that the exclusion restriction for these variables is satisfied[35]. A paper by Ceci *et al.* (2014) reviews some of the relevant literature and comes to the conclusion that there is little evidence that there are significant performance differences (when using tests of mathematical skills) between the genders, but for males being overrepresented in the very right tail of the distribution. When interpreting the following results we will keep in mind that the use of the gender variable as instrument is at best controversial.

For the remainder of this section we will work with the two potential instruments, enc_i and gen_i, (= 1 if a student is male). It is important that the instruments (collectively) explain a significant amount of variation in the intervention variable. Only if that is the case, will the IV estimation method produce precise estimates of the intervention effect (δ_1 in Model (3)). To establish that this is the case we can look at the following (first stage) regression

$$pa_i = \gamma_0 + \gamma_1 q_i + \gamma_2 enc_i + \gamma_3 gen_i + v_i \tag{13}$$

where the variable vector q_i now contains those explanatory variables that, from the analysis in Section 7 have proved statistically significant, the grade in the prerequisite statistics course unit, the Year 1 GPA, a dummy variable indicating whether a student took the course unit in their 3rd year, a dummy indicating whether a student is an overseas student and a set of dummy variables indicating the Programme students are enrolled on. We are, however, mainly interested in the effect of the instruments enc_i and gen_i. While the gender dummy proves statistically significant at the 95% significance level, the encouragement email turns out to have no effect on students' attendance to PASS sessions[36]. When using a F-test to test the null hypothesis that $\gamma_2 = \gamma_3 = 0$ we obtain a F-statistic of 5.816 with a p-value of 0.003. While this may be considered statistically significant the fairly small F-test statistic is evidence of a very weak relationship. Even if we were to accept the role of gender as an instrument, the fact that we have weak instruments will lead us to not put too much store on the following IV estimation results.

Despite the weakness of the instruments we report the OLS and IV estimations for the following regression model

$$y_i = \delta_0 + \delta_1 pa_i + \delta_2 q_i + \varepsilon_i \tag{14}$$

[35] It is well known that the exclusion restriction cannot be tested empirically and has to be motivated through other arguments, e.g. the knowledge of random assignment.

[36] While it is no news to these lecturers that students tend to ignore the advice of lecturers, it nethertheless came as a relief to this particular lecturer (Becker) that similar experiments by other lecturers at The University of Manchester resulted in similar student (non) reactions.

in Table 11. In fact we will also report the results for the case in which we use ph_i as the intervention variable. The OLS versions of these models are close to models (10e) and (10f) and only differ in that statistically insignificant variables have been excluded.

Model (6)				
Method	OLS	IV	OLS	IV
Instruments		gen, enc		gen, enc
Const	-15.601 ***	-15.44 ***	-15.204 ***	-14.774 **
pa	0.793 ***	0.865		
ph			4.476 **	5.380
stats	0.118 *	0.118 *	0.117 *	0.116 *
Y1gpa	0.977 ***	0.972 ***	0.980 ***	0.969 ***
Y3stud	2.921 *	2.895 *	3.012 *	2.971 *
Overseas	-2.900 **	-2.883 *	-2.840 *	-2.790 *
Prog(BEconSc)	4.067 **	4.125 **	3.804 **	3.881 **
Prog(IBFE)	3.777	3.806	3.556	3.575
Prog(Other)	3.900	4.077	3.362	3.649
Prog(PPE)	-0.392	-0.336	-0.906	-0.887
R^2	0.495	0.495	0.493	0.493
Stage 1 F-test		5.816		4.334
(p-value)		(0.003)		(0.014)

Table 11. Parameter estimates of Model (4). * /**/* significant at 1/5/10% significance level (using White standard errors). The base category for the Programme membership is the BA(Econ).**

A number of things can be seen from this Table. Importantly, while the OLS estimations deliver statistically significant coefficients for the intervention variables (*pa* or *ph*), the IV estimations deliver coefficient estimates with very similar, but not statistically significant values. This is a result of the very weak instruments which were used in the IV estimation. The application of IV estimation has, in this case, not progressed our effort to isolate the effect of PASS attendance on the exam grade. As argued before, the significant estimates for (*pa* or *ph*) in the OLS estimations cannot necessarily be taken as evidence for a significant impact of PASS attendance on exam grades as we cannot be certain that the intervention variable is uncorrelated to the error term.

In the literature there are two papers we are aware of that attempt to use an instrumenting approach, Lewis *et al.* (2005) and Loviscek and Cloutier (1997). Their approach is different in the sense that they apply a two-stage Heckman estimator (see also Blundell, 2005), the validity of which, however, also crucially depends on the availability of an instrument. Both papers find apparently significant effects of PASS attendance, effects in the order of 1.5 times the standard deviation of grade outcomes. We are, however, not convinced that the authors did identify sufficiently strong instruments[37].

10. Using Panel features of the dataset

In the previous Section we demonstrated how IV estimators can potentially be used to estimate the causal effect of PASS attendance on exam grades even under the presence of unobserved

[37] Also see Bushway *et al.* (2007) for a discussion of problems that can arise in the application of a Heckman two-step correction. In a further indication that the result in Lewis *et al.* (2005) is inflated due to methodological issues is that a different methodology shows up virtually now significant effect (although on a slightly different dataset).

heterogeneity, or selection on unobservable variables. However, as the instruments available, were at best weak, and at worst invalid instruments (gender), IV estimation did not deliver any conclusive evidence.

There is one more feature in the econometrics data-set that can potentially deliver additional information. Most PASS schemes run in the first year of a student's degree. This PASS scheme , however, relates to a 2nd year course unit. This has the advantage that it delivers important conditioning information (*Stats*, *Micro*, *Y1_GPA*) from the student's first year that would not normally be available. This gives the dataset some Panel characteristics. We speak of Panel data when we have repeated observations for one variable for the same individual (just observed at different times). While this is not strictly true in that students do not take Econometrics in Year 1 and Year 2, we do have grade information (albeit pertaining to different course units) from Year 1 and 2[38].

Let's reconsider the basic regression model used to estimate the conditional expectation function:

$$y_{it} = \delta_0 + \delta_1 x_{it} + \delta_2 qf_{it} + \varepsilon_{it} \tag{15}$$

As before we label with y the outcome variable used to evaluate the intervention and x represents the intervention variable (either pa_{it} or ph_{it}). As this model stands we need to allow for the possibility that the intervention variable (PASS attendance) is correlated with the error term, resulting in biased estimates of δ_1 if we were to apply OLS to (15). Model (15), however, differs in the following aspects to that used previously: *First,* all variables now also have a time subscript *t* = Year 1, Year 2, which indicates that variables are observed in Years 1 and 2[39]. *Second*, the vector of covariate variables qf_{it} now only contains student characteristics such as their gender, programme, home or overseas and ethnic status. All of these are fixed, such that $qf_{i1} = qf_{i2}$.

The basic idea of using Panel data to isolate a causal effect of x on y is best illustrated by looking at the differenced version of model (15):

$$y_{i2} - y_{i1} = \delta_0 - \delta_0 + \delta_1(x_{i2} - x_{i1}) + \delta_2(qf_{i2} - qf_{i1}) + \varepsilon_{i2} - \varepsilon_{i1} \tag{16}$$

which results from subtracting model (15) for period *t* = 1, from model (15) for period *t* = 2. This will simplify significantly after noting that $qf_{i1} - qf_{i2} = 0$, $x_{i1} = 0$ as no student participated in any Year 1 PASS scheme[40] and defining $v_{i2} = \varepsilon_{i2} - \varepsilon_{i1}$:

$$y_{i2} - y_{i1} = \delta_1 x_{i2} + v_{i2}. \tag{17}$$

[38] The only other paper in this literature that states that it exploited some panel data feature is Dancer *et al.* (2007). However, the data description in that paper does not allow the conclusion that their data indeed do have a panel feature and therefore the results of fixed effects and random effects estimations are impossible to interpret.
[39] As discussed in the Data Section there are actually some students which take the econometrics course unit in Year 3 (and a subset of these which take the statistics prerequisite course in Year 2). For simplicity we abstract from this complication, but note that the *Year 3* dummy variable has been used to allow for these students.
[40] This is certainly true for the cohort of students considered here. A year later econometrics students would have also been able to attend a 1st Year PASS scheme.

We are now essentially modelling the improvement in grades as a function of PASS participation[41]. The coefficient δ_1 has a causal interpretation if all selection on unobservable variables was due to student characteristics that, while unobservable, were also constant through time. Further we need to assume that the effect of these unobservable variables was identical on y_{i1} and y_{i2} (see Angrist and Pischke, Chapter 5). Given these assumptions then this omitted effect will difference out in the new differenced error term $v_{i2} = \varepsilon_{i2} - \varepsilon_{i1}$. As this effect is thus removed from the error term this would also eliminate any correlation between the intervention variable x_{i2} and the error term v_{i2}, allowing parameter estimation by OLS.

Of course, while y_{i2} is the Semester 1 exam grade for the econometrics course unit students take in their 2nd year, there is no such exam in their 1st year, y_{i1}. This means that the pure Panel data approach is not applicable here. However, in the following analysis we will replace y_{i1} with either the *Year1_GPA* or the *stats* variable. This approach, which in a slightly different context has also been used by Andrietti and Velasco (2015), has two obvious shortcomings. *First*, we now treat variables (*Year1_GPA* and *stats*) that previously were used as conditioning variables as pre-intervention outcome variables. *Second*, we acknowledge that the difference between the definitions for y_{i1} and y_{i2} makes the above assumption that the unobserved variables have identical effects on y_{i1} and y_{i2} less credible. Despite these two concerns we believe that this analysis has marginal value in the context of the analysis in this paper as we offer this analysis as an additional robustness check amongst a set of techniques that all have some shortcomings.

Due to the different nature of y_{i1} and y_{i2} it is also likely that the effect of qf_{it} does not cancel out and hence Table 12 will also present results that include these variables.

[41] When estimating this model one would also include a constant term that allows for a trend the data. In fact it is an important identifying assumption that this trend is common across treatment and control groups.

Model (17)	Method	OLS				
Dependent Var	Intervention Var					
y_{i1}	x_{i2}	qf_{i2}	$\hat{\delta}_1$	$se(\hat{\delta}_1)$		R^2
$stats_{i1}$	pa_{i2}	no	0.6364	(0.2970)	**	0.01342
$Y1\_gpa_{i1}$	pa_{i2}	no	0.8909	(0.2412)	**	0.04533
$stats_{i1}$	pa_{i2}	yes	0.6308	(0.2849)	**	0.16860
$Y1\_gpa_{i1}$	pa_{i2}	yes	0.8473	(0.2374)	**	0.10350
$stats_{i1}$	ph_{i2}	no	3.9541	(1.7130)	***	0.01453
$Y1\_gpa_{i1}$	ph_{i2}	no	5.2725	(1.3134)	***	0.04452
$stats_{i1}$	ph_{i2}	yes	3.4095	(1.6392)	***	0.16660
$Y1\_gpa_{i1}$	ph_{i2}	Yes	4.7913	(1.2894)	***	0.09999

Table 12. OLS parameter estimates for δ_1 in Model (17). * /**/* significant at 1/5/10% significance level (using White standard errors). The base category for the Programme membership is the BA(Econ).**

Before interpreting the estimated coefficients in Table 11 it is useful to remember that the estimated marginal effects of pa_i and ph_i were estimated to be 0.817 and 4.220 respectively in the Models with all covariates (columns (10f) and (10e) in Table 6).

It is immediately apparent that the impact estimates from model (17) are fairly similar to those in the regression models with covariates. This is irrespective of the definition of y_{i1} in the dependent variable ($y_{i2} - y_{i1}$), the choice of intervention variable (x_{i2}) and whether extra covariates (qf_{it}) were included or not. All estimates are estimated to be statistically significant. A few interesting details arise from these results. The impact coefficient is smaller when using the first year statistics course result as the relevant Year 1 grade (for both definitions of the intervention variable). When not using any covariates (but for the intervention variable) we find higher, but still small, R^2 values when using the Year 1 GPA as the pre-treatment outcome variable. This is perhaps not surprising when we recognise that the 1st year statistics grade is more variable than the Year 1 GPA (see Table 2). When including the covariates (qf_{it}) this is reversed mainly due to the fact that overseas students tend to do significantly better in Year 1 statistics course units (but not overall in Year 1 and not in econometrics).

The additional benefit of this methodology, as compared to the regression and matching approach, is that, conditional on the above assumptions holding, we can allow for selection on unobservable variables, as long as these variables remain constant. Doing this has delivered impact estimates for PASS attendance that are very similar to those coming from the regression and matching approach. But it needs to be noted that any selection bias that comes from student heterogeneity that is captured by neither observed variables nor constant through time has not been sufficiently taken into account by our analysis.

11. Conclusion

It was the aim of this paper to establish whether attendance in voluntary PASS schemes has any significant and measurable effect on student's grade performance. Here we put particular emphasis on the possibility of self-selection bias and explicitly describe efforts to control for this.

One of the strategies to control for selection bias was to use covariates that have the potential to capture the reasons for self-selection. This is done, both by using regression and matching estimators. Both these estimators suggest that the effect of attending PASS sessions is potentially significant, statistically and indeed in substance. The measured advantage for being a regular PASS attendee is in the order of 5 marks and the average exam grade advantage of one additional PASS session per semester in the order of 0.8 marks. When comparing this to the exam grade standard deviation of 16, we obtain an effect size (of high PASS attendance) of a third of a standard deviation. The size of this effect is at the lower end of the effect sizes range that was described in the review paper by Dawson *et al.* (2014) who report effect sizes between 1/3 and 2/3 of a standard deviation.

The use of conditioning variables has significantly reduced the size of this effect. Without these conditioning strategies, the apparent positive grade effect of PASS attendance would be much closer to a value of around 0.6 standard deviations.

The effect size estimates obtained in this manner would have a causal interpretation if the self-selection was fully explained by the variables used in the above control strategy. We are confident that the controls used here are better controls than in most previous studies. This would mainly be due to the fact that we have pre-PASS attendance grade information obtained at University level (Year 1) as opposed to High-School level grade information used by almost all previous studies.

However, as we cannot exclude the possibility that some degree of self-selection arises from unobserved factors, we recommend that these results are interpreted with some degree of caution. We also outlined how one would use econometric techniques to arrive at reliable causal interpretations even if some selection was related to unobserved heterogeneity that also impacts the course unit grade directly. For instance we described an experimental setup that potentially could have delivered a valid instrumental variable, but were not successful in that endeavour.

Lastly, we revisited an earlier attempt at using a randomised control trial to eliminate any bias in our effect estimate. By revisiting that data-set we show that it was not carefully constructed to deliver any robust and positive results. We therefore recommend that institutions find ways to implement robust identification strategies (e.g. a randomised control trial). We acknowledge that this may not be a straightforward endeavour as a positive grade effect is merely one of many potential positive effects (see e.g. Dawson *et al.*, 2014, for a review of the range of effects). The fact that a large number of Universities have been implementing PASS schemes, sometimes over decades, illustrates that there is a deep belief that these schemes have benefits for students and PASS leaders, which is, of course, the reason why institutions will hesitate to implement randomised trials, thus preventing the accumulation of real robust evidence.

References

Andrietti, V. (2014), Does lecture attendance affect academic performance? Panel data evidence for introductory macroeconomics, International Review of Economics Education, 15, 1-16.

Andrietti, V. and Velasco, C. (2015) Lecture attendance, study time, and academic performance: A panel data study. *The Journal of Economic Education*, 26, 239-259, doi: 10.1080/00220485.2015.1040182.

Angrist, J.D. and Lavy, V. (2001), Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools, Journal of Labour Economics, 19, 343-369.

Angrist, J.D. and Pischke, J.-S. (2009) Mostly Harmless Econometrics, Princeton University Press.

Angrist, J.D. and Pischke, J.-S. (2015) Mastering Metrics: The Path from Cause to Effect, Princeton University Press.

Arulampalam W., Naylor R.A. and Smith J. (2012) Am I missing something? The effects of absence from class on student performance, Economics of Education Review, 31, 363-375.

Blundell R., Dearden, L. and Sianesi, B. (2005) Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey, *Journal of the Royal Statistical Society A*, 168, 473-512.

Bradlow E.T. (1998) Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design, Marketing Letters, 9, 383-391.

Bushway, S., Johnson, B.D. and Slocum, L.A. (2007) Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology, *Journal of Quantitative Criminology*, 23, 151-178.

Ceci, S.J., Ginther, D.K., Khan, S. and Williams, W.M. (2014) Women in Academic Sciences: A Changing Landscape, Psychoological Science in the Public Interest, 15, 75-141.

Dancer, D., Morrison, K. and Smith, M. (2007) Measuring the impact of a peer assisted learning program on students' academic performance in econometrics. In: M. Davies, J. Hirschberg, C. Johnston and J. Lye (Eds.), The quantitative analysis of teaching and learning in higher education: Forum proceedings (pp. 19–42). Melbourne, Victoria, Australia: University of Melbourne.

Dawson, P., van der Meer, J., Skalicky, J. and 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*, doi: 10.3102/0034654314540007.

Fayowski, V. and MacMillan, P. D. (2008) An evaluation of the supplemental instruction programme in a first year calculus course. *International Journal of Mathematical Education in Science and Technology*, 39, 843–855. doi:10.1080/00207390802054433

Fostier, M. and Carey, W. (2007) Exploration, experience and evaluation: Peer Assisted Study Scheme (PASS), sharing the experience of the University of Manchester, The Higher Education Academy, https://www.heacademy.ac.uk/resource/exploration-experience-and-evaluation-peer-assisted-study-scheme-pass-sharing-experience.

Gattis, K.W. (2002) Responding to self-selection bias in assessments of academic support programs: A motivational control study of supplemental instruction, *Learning Assistance Review*, 7, 26-36.

Haynes, L., Service, O., Goldacre, B. and Torgerson D. (2012) Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials, Cabinet Office, Behavioural Insights Team, Paper, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf.

Hirano K., Imbens G.W., Rubin, D.B. and Zhou, X.-H. (2000) Assessing the effect of an influenza vaccine in an encouragement design, Biostatistics, 1, 69-88.

Hodges, R., Dochen, C. and Joy, D. (2001) Increasing student's success: When supplemental instruction becomes mandatory, *Journal of College Reading and Learning*, 31, 143-156.

Hutchinson D and Styles B. (2010) A guide to running randomised controlled trials for educational researchers, National Foundation for Educational Research (NFER), http://www.nfer.ac.uk/nfer/publications/RCT01/RCT01.pdf.

Imbens, G.W. (2004) Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. Review of Economics and Statistics, 86, 4-29.

Loviscek, A.L. and Cloutier, N.R. (1997) Supplemental Instruction and the Enhancement of Student Performance in Economics Principles, *The American Economist*, 41, 70-76.

McCarthy, A., Smuts, B. and Cosser, M. (1997). Assessing the effectiveness of supplemental instruction: A critique and a case study. *Studies in Higher Education*, 22, 221–231. doi:10.1080/03075079712331381054

Morgan, S.L. and Winship, C. (2007) Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press.

Sekhon, J.S. (2011) Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R, Journal of Statistical Software, 42, 1-52.

Sole, Gisela; Rose, Andrew; Bennett, Tracey; Jaques, Katrina; Rippon, Zoe; and van der Meer, Jacques, (2012) A Student Experience of Peer Assisted Study Sessions in Physiotherapy, Journal of Peer Learning, 5(1). Available at:http://ro.uow.edu.au/ajpl/vol5/iss1/2.

Sovey A.J. and Green D.P. (2011) Instrumental Variables Estimation in Political Science: A Readers Guide.

Stock, W.A., Ward, K., Folsom, J., Borrenpohl, T., Mumford, S., Pershin, Z., Carriere, D. and Smart, H. (2013) Cheap and effective: The impact of student-led recitation classes on learning outcomes in introductory economics. *The journal of Economic Education*, 44, 1-16, doi: 10.1080/00220485.2013.740368.

Todd, P.E. and Wolpin, K.I. (2003) On the specification and estimation of the production function for cognitive achievement, The Economic Journal, 113, F3-F33.

Torgensen C.J. and Torgensen, D.J. (2013) Randomised trials in education: An introductory handbook, https://educationendowmentfoundation.org.uk/uploads/pdf/Randomised_trials_in_education-revised250713.pdf, accessed (17 July 2015).

Wooldridge, J.M. (2013) Introductory Econometrics: A Modern Approach, Cengage Learning.