Evaluation and Program Planning 68 (2018) 81–89

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

A counterfactual impact evaluation of a bilingual program on students' grade point average at a spanish university

J.L. Arco-Tirado[a,*], F. Fernández-Martín[a], A.M. Ramos-García[b], L. Littvay[c], J. Villoria[b], J.A. Naranjo[d]

[a] Department of Developmental and Educational Psychology, University of Granada, Spain
[b] Department of Didactis of Languge and Literature, University of Granada, Spain
[c] Department of Political Science, Central European University, Hungary
[d] Department of Didactics of Experimental Sciences, University of Granada, Spain

ARTICLE INFO

Keywords:
Bilingual program
Counterfactual impact evaluation
Causal inference
Matching
English as medium of instruction
Grade point average

ABSTRACT

This observational study intends to estimate the causal effects of an English as a Medium of Instruction (EMI) program (as predictor) on students Grade Point Average (GPA) (as outcome) at a particular University in Spain by using a Counterfactual Impact Evaluation (CIE). The need to address the crucial question of causal inferences in EMI programs to produce credible evidences of successful interventions contrasts, however, with the absence of experimental or quasi-experimental research and evaluation designs in the field. CIE approach is emerging as a methodologically viable solution to bridge that gap. The program evaluated here consisted in delivering an EMI program in a Primary Education Teacher Training Degree group. After achieving balance on the observed covariates and recreating a situation that would have been expected in a randomized experiment, three matching approaches such as genetic matching, nearest neighbor matching and Coarsened Exact Matching were used to analyze observational data from a total of 1288 undergraduate students, including both treatment and control group. Results show unfavorable effects of the bilingual group treatment condition. Potential interpretations and recommendations are provided in order to strengthen future causal evidences of bilingual education programs' effectiveness in Higher Education.

1. Introduction

The development and practice of plurilingual education is one of the priorities of the Council of Europe (De Wit, Hunter, Howard, & Egron-Polak, 2015) and the implementation of effective plurilingual education models is an on-going empirical process facing significant challenges at the scientific, institutional and policy levels. In an environment of increased dominance of English as the language of communication in research and education, and its use as a global lingua franca, there is a need to stimulate bilingual and plurilingual learning and programs at all educational levels including Higher Education (HE) in non-Anglophone countries (Bradford 2012; De Wit et al., 2015; Doiz, Lasagabaster, & Sierra, 2013). Along this line, Higher Education Institutions (HEIs) are feeling the pressure to offer students opportunities for developing comprehensive bilingual, biliteracy, and cross-cultural skills in their discipline of study (Bradford, 2012; Ramos-García, 2013; Dafouz & Smit, 2016; Doiz et al., 2013 p. 217).

English-taught, English-medium instruction, bilingual degree programs, bilingual or plurilingual learning or bilingual Massive Open Online Courses (MOOCs) are just a few examples reflecting how HEIs are responding to such internationalization, globalization and marketization forces. Interestingly, authors like Dafouz and Camacho-Miñano (2016) point to the need to analyze carefully potential conflicts between national differences in terms of language policies, implementation strategies or teaching traditions and that "Englishized" background. Furthermore, other authors like Dor (2004); Kirkpatrick (2011) warn against the inimical effects of the increasing role(s) English is playing in HEIs on local language and scholarship written in the local language in both Europe and Asia. This is the case for countries such as South Korea (Kim, Son, & Sohn, 2009), China (Hu, Li, & Lei, 2014; Johnstone, 2010) and Spain (Aguilar & Rodríguez, 2012; Dafouz & Guerrini, 2009; Dafouz, Núñez, and Sancho, 2007; Dafouz, Núñez, Sancho, and Foran, 2007; Doiz, Lasagabaster, & Sierra, 2011; Fernández-Costales & González-Riaño, 2015; Fernández-Viciana & Fernández-Costales, 2017; Ramos-García, 2013).

Additionally, in this certain rush to internationalize, there may be

* Corresponding author at: Department of Developmental and Educational Psychology, University of Granada, Campus de Cartuja, s/n, Granada, 18071, Spain.
E-mail address: jlarco@ugr.es (J.L. Arco-Tirado).

https://doi.org/10.1016/j.evalprogplan.2018.02.013
Received 29 September 2017; Received in revised form 15 February 2018; Accepted 18 February 2018
Available online 21 February 2018

variability in the quality of student experience for an international student (Dearden, 2014) but also for national and local students, which threats mobility and quality two core elements of the Bologna Declaration (European Ministers in charge of Higher Education, 1999). Interestingly, to protect quality and effectiveness of the European Higher Education Area, the Bologna process established the evaluation plans and mechanisms necessary for the renewal of the accreditation of the bachelor (monolingual) degrees (see Ministerio de Educación y Ciencia, 2007), but not for bilingual degrees, which were not contemplated at that time as another actual short-term possibility, at least in our country (Arco & Fernández, 2016; Ramos-García, Arco-Tirado, Fernandez-Martín, & Villoria-Prieto, 2016). Coincidentally, 2010 was both the deadline set for the Bologna process in Europe and the departing moments of several bilingual programs in Spain like for example the one we report here.

In this context, as a part of such accreditation renewal process, coordination, monitoring and evaluation activities yield preliminary positive evaluation results when comparing monolingual and bilingual groups with very little percentage point differences in the four years following-up period on key indicators and benchmarks (e.g., performance rate, success rate, GPA).

However, although apparently both intervention programs were working effectively in the case of more radical innovations such as EMI provisions it was necessary the application of more complex research designs and statistical techniques conducive to filter high-quality evidence of the EMI programs net impact effects. Following Slavin (2008) the need to establish a causal link between interventions and results based on high-quality evaluation strategies and techniques is essential for generating reliable evidence of what works. In this context, it is surprising, however, that the significant expansion of these programs worldwide in tertiary education has not been accompanied yet by large scale governmental efforts to measure the scientific quality of the good practices, promising practices, evidence-based practices, practice-based evidence and/or any other type of EMI practice or program to inform future evidence-based plurilingual higher education policies. This is particularly important in this case of EMI programs due to the apparently contradictory abundance of net impacts results on key students academic outcomes. In this regard, while many studies show that there is a cost for the students' GPA associated to this modality of delivering the curriculum (Byun, Chu, Kim, Park, Kim, & Jung, 2011), other studies show the benefits for students linked to this programs including a transition period (Airey, 2009; Del Campo, Cancer, Pascual-Ezama, & Urquía-Grande, 2015; Klaassen, 2001), while others show no effects on significant academic variables for students (Dafouz, Camacho-Miñano, & Urquía, 2014; Hellekjaer, 2008).

From the statistical decision theory perspective, the validity of such diverse statistical conclusions depends on the probability of obtaining Type I error (concluding that a treatment has an effect when it does not) or Type II error (failing to detect that a treatment has an effect when the true treatment effect is nonzero) when making the statistical inferences. So research efforts should be aimed at, primarily, increasing Statistical Power, that is, avoiding Type II error, a major threat to the statistical conclusion validity of educational research studies (Shadish, Cook, & Campbell, 2002).

1.1. The counterfactual impact evaluation (CIE) approach

Randomized Control Trial is the ideal way to study the net effects of educational programs or reforms, although these programs and reforms rarely adopt ex-ante evaluation designs. That is the case of the evaluation studies of the EMI programs mentioned above and also the case of the bilingual program analysed here. In all these cases regular ex-post comparisons are inadequate as students who chose an EMI program might be very different from those who opt for monolingual Degree programs. So, a highly convincing approach is needed, one which devotes far more attention to methods accounting for potential (ex-ante) differences between treatment group members and potential controls that are likely to affect the decision to participate (selection bias) and the results (before-after bias) obtained (European Commission, 2013). In this regard, CIEs-comparison of results to estimates of what would have occurred otherwise, provide the statistical technique necessary to counteract these potential sources of bias.

According to Holland (1986) the counterfactual approach conceives of two potential results when determining the effect of our intervention program on students. The first result is the student academic performance subsequent to having taken part in the bilingual-EMI group. This is the observed result for the student who receives the intervention. The second potential result is this student's performance had they not taken part in the bilingual education program, all else (measured covariates) being equal. In these circumstances this second result is referred to as the counterfactual result. In reality we do not and cannot observe counterfactuals results for individuals exposed to an intervention, because observing both outcomes for the same individual at the same time is not possible (Caliendo & Kopeinig, 2008; Gordon, 2015). What is done instead by using the matching approach is to estimate counterfactual results from selected individuals in the control group, assuming that potential unobserved confounding variables will not bias the selection of controls from the large group of nonparticipants available, who must be similar to the participants in all relevant pre-treatment variables (European Commission, 2013). Conventional matching using covariates can work well; however, as the number of covariates increases, it becomes difficult to find good matches for subjects in the treatment group (Olmos & Govindasamy, 2015). For these cases, in which conditioning on all relevant covariates is limited, the use of so-called balancing scores (i.e., functions of the relevant observed covariates like the propensity score) have been offered as a solution (Caliendo & Kopeinig, 2008). Some of the benefits associated with the use of this statistical technique (i.e., propensity scores) according to Olmos and Govindasamy (2015) are: (a) Creating adequate counterfactuals when random assignment is infeasible or unethical; (b) The development and use of propensity scores reduces the number of covariates needed to control for external variables (thus reducing its dimensionality) and increasing the chances of a match for every individual in the treatment group; (c) The development of a propensity score is associated with the selection model, not with the outcomes model, therefore the adjustments are independent of the outcome.

Noted in Thoemmes and Kim (2011), the propensity score is a conditional probability which expresses how likely a participant is to be assigned or to select the treatment condition given certain observed baseline characteristics. In a propensity score analysis this conditional probability is used to condition observed data, for example, through matching or stratification on the propensity score. The aim of conditioning on the propensity score is to achieve balance on the observed covariates and recreate a situation that would have been expected in a randomized experiment. Since the proliferation of propensity matching approaches in the literature, methodologists suggested additional matching methods to achieve appropriate balance between the quasi-experimental treatment and control groups (Diamond & Sekhon, 2015; Iacus, King, & Porro, 2012).

Among the wide range of approaches to mimic randomization in CIE to build a credible control group (without the use of randomization) from existing non-participants groups and to estimate causal effects (Gordon, 2015; Hahs-Vaughn & Onwuegbuzie, 2006) matching methods are experiencing a tremendous increase of interest in many scientific areas including the social sciences (Thoemmes & Kim, 2011). In our case three matching approaches have been compared: genetic matching (Diamond & Sekhon, 2015), nearest neighbor matching on a propensity score (Caliendo & Kopeinig, 2008; Hahs-Vaughn & Onwuegbuzie, 2006; Harder, Stuart, & Anthony, 2010) and Coarsened Exact Matching (CEM) (Iacus et al., 2012). The reason to run different matching methods has to do with identifying which one reaches a better balance on the covariates for the treatment and control groups before

calculating the treatment effect. Once both groups are fully comparable on their covariates, the impact of the bilingual-EMI education program for the student is simply the difference between the observed results on the treatment group and the counterfactual results estimated from the control group. Out of this comparison results, we provide a causal *description* rather than a causal *explanation* (European Commission, 2013) of the EMI program model implemented at the University of Granada, Spain.

So, once completed the process of the renewal of the accreditation of the university bachelor degree programs in Primary Education Teacher Training, affecting both modalities (i.e., monolingual and bilingual), as mentioned earlier in this section, and following Dearden (2014); Hu et al. (2014) recommendation of addressing potential gaps between policy rhetoric around high quality education standards in plurilingual education programs in HE and ground-level reality in the implementation of the EMI programs, our main goal for this study is to find evidence of the effectiveness of the EMI program implemented at the undergraduate level at a research university on a non-English speaking country, and its effects on the academic performance of the students. Our working hypothesis, based on our preliminary following-up data, is that there will not be a difference on performance average between Treatment and Control group, after controlling for confounding factors.

2. Method

2.1. Sample

The sample was drawn from an official observational (Cochran, 1965) dataset managed by the Academic Organization Office of the University of Granada. The complete original dataset consisted of N = 1288 students (mean age M = 19.75 years old, SD = 3.51, 57.69% female) registered for the academic cohorts of 2011/2015 and 2012/2016 of the Primary Education Teacher Training Degree. The institution reported bilingual education modality for N = 132 students adding both cohorts (10.25%), which leaves an N = 1156 (89.75%) for monolingual education modality. The reason to limit the dataset to these two cohorts has to do with the fact they were the only two cohorts having completed the four-year career at the time of implementing this study.

2.2. Instruments

2.2.1. Language group

Number of languages spoken during instruction time (lecturing) across courses served as our key independent variable. It was measured binarily, with 1 indicating “bilingual group” and 0 indicating “monolingual group”. Thus, language group was conceptualized as a binary treatment variable, with bilingual group as the treatment condition and monolingual group as the control condition.

2.2.2. Grade point average

Academic achievement at Graduation served as our outcome and reflected the main indicator of success in terms of academic success including employability.

2.3. Covariates

We used a number of covariates as selection variables, regression control variables, or both, as informed by previous empirical work (Neuville et al., 2007; Tinto, 1997). These variables were: entry-exam score, father’s job, mother’s job, father’s studies, mother’s studies, via of admission, entry year, end year, dropout, and gender. Gender was coded binarily, with 1 indicating *male* and 0 indicating “female”. Age was a continuous variable measured in years. Father and mother job situation was a nominal variable with 10 categories ranging from 1(=Unpaid workers) 2(=Non qualified workers), 3(=Qualified workers in agriculture and fishery), 4(=Qualified workers in industry, construction, and miner), 5(=Qualified workers on services, hotels, sales), 6(=Qualified in the Army), 7(=Administrative auxiliary), 8(=Technicians, support professionals of medium level), 9(=Technicians and professionals of high level with or without university studies), and 10(=Directors and Managers of public administration enterprises). Father’s and/or mother’s education were ordinal variables, with 7 categories ranging from 1(=Illiterate) 2(=No education), 3(=Primary education), 4(=Secondary education including medium vocational training), 5(=Post secondary education including higher vocational training), 6(=Higher education less than 4 years or similar), and 7(=Higher education more than 4 years). University entry route was another nominal variable with 3 categories 1(=University entry exam) 2(=Vocational Training level 2 or level 3), 3 (=University entry exam for older than 25). Entry exam score was a continuous variable normalized into a scale from 5 to 10 points. No special requirements were imposed on EMI students so comparability among groups is warrantee. The administrative decision of creating a separate EMI group entailed that both groups (treated and control were receiving different ‘treatment'), this situation coincidently aligns with what a theoretical “Hierarquical design” recommends in order to minimize potential problems of contamination between treatments, because only one treatment is present in the same cluster (e.g., in the same classroom). In other words, this type of design helps to alleviate contamination because the whole cluster (e.g., the classroom) receives the treatment (Hedges & Rhoads, 2010). All covariates were used in the matching procedure and to generate propensity scores in a logistic regression model with language group as the binary outcome. All variables were used as control variables in regression analyses. Nominal variables were binarized, with one reference category and ordinal was treated as continuous. No other important variables informed by previous empirical related work involving academic performance in higher education were missing from our analyses.

2.4. Procedure

The CIE process implemented comprises four major stages, with several sub-stages: (1) exploring the viability of the study; (2) determining observational covariates; (3) balancing the treatment and control group using matching methods; and (4) calculating the treatment effects.

Stage 1. Exploring the viability of the study. The viability of the study was explored by making some strategic decisions in terms of measuring the EMI program impact on the student's academic performance as evidenced by GPA. In this regard, we confirmed that the EMI program met the basic requirements of a counterfactual approach, i.e., the treatment delivered was clearly distinguishable from other interventions and participants in the intervention were exposed to broadly the same package of measures. The conditions above lead us to assume a coherent causal mechanism underpinning the intervention. Additionally, it was checked that the type of data available required to conduct this CIE were available from administrative sources. After that, cohorts of treated and non-treated units who were the focus of the evaluation were identified and applied the mechanisms to collect data from the administrative system cohorts. Subsequently, treatment and control group data on the selected outcome as well as covariates selected for this evaluation were extracted from official students’ record kept at University. At a later phase, students’ personal, family and academic data were anonymized for evaluation purposes granting confidentiality this way. Additionally, a definition of treatment was elaborated, which consisted of being registered at the bilingual group of the Degree in Primary Education Teacher Training at the University of Granada. Out of a total of 240 ECTS in this Degree, 156 ECTS (65%) were delivered through the EMI model. The definition for the control group consisted of being registered at the monolingual group of the

Degree in Primary Education Teacher Training at the University. The final sample of treatment and equivalent control group on which statistical analysis were applied was conformed based first on estimating the propensity score, and second on balancing both groups using matching methods as described earlier. For this Spanish-taught group none of the credit hours were delivered through the EMI model. So, all members from both groups were subject to selection processes based on choice motivated by potentially unobserved factors (Card, Ibarraran, & Villa, 2011).

Stage 2. Determining observational covariates. First, collected data (covariates) from both treated individuals and a sample of similar non-treated students. When selecting the covariates Shadish et al. (2006) recommend using a rich set of covariates in order to make credible the strongly ignorable treatment assignment assumption, which basically means that selection into treatment is not based on unobserved factors, or in other words, that the selection process can be characterized by the observable data (European Commission, 2013). Consequently, to yield unbiased causal effect estimates for the matching procedure, we considered all the variables that could potentially influence the student's participation in the treatment group according to previous research (Cham & West, 2016; Guill, Lüdtke, & Köller, 2017; Steiner, Cook, Shadish, & Clark, 2010; Ward & Johnson, 2008). The included covariates: (a) concerned the time before the treatment was assigned; (b) were considered to be stable over time in order to ensure that the covariates would not be affected by the treatment itself (Caliendo & Kopeinig, 2008; Kretschman, Vock, & Lüdtke, 2014); and (c) did not exclude or collapse any categorical covariate (Ho, Imai, King, & Stuart, 2007). Hence, variables such as entry-exam score, father's job, mother's job, father's studies, mother's studies, entry route, entry year, cohort, born in/out of Spain, birth year, dropout and gender were included. The nominal variables father and mother's job initially coded into 11 categories; the variables father and mother' education were considered continuous; the ordinal variable via of admission was coded into 3 categories becoming two dummy variables, the variable gender was coded into 2 categories becoming a dummy variable. The variables entry-exam score and entry year were treated as continuous variables. Afterwards, a logistic regression model with treatment and control as control and covariates to balance scores conditioning on all relevant covariates were used (Jensen, Shafer, Guo, & Larson, 2017). After collecting covariates, the propensity score was estimated using a logistic regression model in which all variables were included (non-parsimonious). These scores are later used to assess (and, in the case of nearest neighbor matching, to achieve) balance between the treatment and control groups on the included covariates.

Stage 3. Balancing the treatment and control group using matching methods. After estimation of the propensity scores the data was conditioned using the three matching procedures: nearest neighbor matching using propensity score, genetic matching and CEM algorithm. The later two are known to be the most efficient in producing treatment/control balance (Iacus et al., 2012). The plausibility of these approaches rests on the assumption, among others, that selection into treatment can be fully characterized by the observable data, which means that there are no unobserved differences between treatment and control groups that are related to results and the decision to participate in the intervention. The credibility of this assumption is enhanced by the incorporation of a rich range of variables into the matching estimation as well as the selection of variables based on prior knowledge and theory (European Commission, 2013; Shadish et al., 2006). With regard to these probability values, we then checked for the so-called overlap assumption (Guill et al., 2017), through visual balance checks (e.g. in this case clearly showed why means would be insufficient for checking the balance) as well as means comparison for both genetic and nearest neighbor matching balance results.

Stage 4. Calculating the treatment effects. After matching the matching performance of the methods applied was examined by considering the absolute standardized differences in the means between the

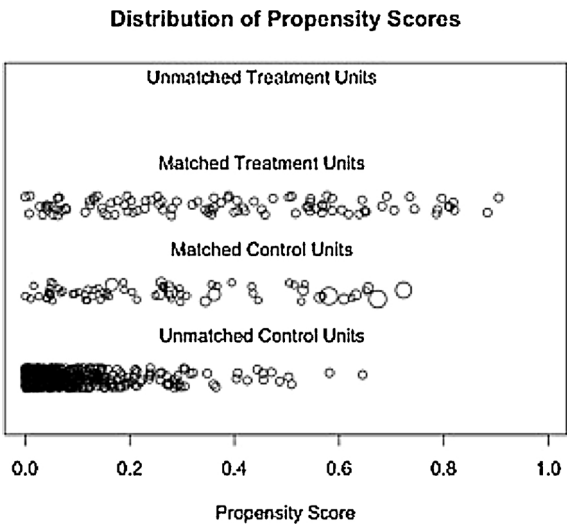

Fig. 1. Genetic Matching distribution of Propensity Scores by treatment status.

treatment and comparison group regarding the propensity scores and the number of covariates given. Therefore, genetic matching was the final choice: the matching procedure that leads to the smallest maximum bias in the covariates (Kretschman et al., 2014; Stuart, 2010). After matching, we re-ran the same regression models with post-matching sample to check for treatment effects again.

2.5. Statistical analysis

Analysis was done in R (version 3.4.0). For matching the MatchIt package (3.0.1) was used.

3. Results

Our results are presented in three parts. First, the results of the row comparisons of treatment and control on the selected covariates are shown. Secondly, the extent of the success of matching algorithms in removing differences in the covariates' distributions is offered. Finally, the treatment effects on students' academic performance are presented.

Figs. 1 and 2 compare the distribution of propensity scores between matched treatment units, matched control units and unmatched control units after genetic matching and nearest neighbor matching respectively. It can be seen that there is a good degree of overlap in the unmatched control units and, on one hand, a much better distribution of

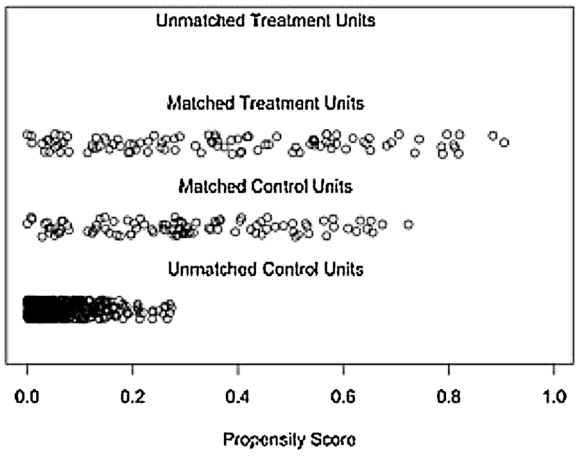

Fig. 2. Nearest Neighbour Matching distribution of Propensity Scores by treatment status.

Table 1
Logistic regression model with T/C as control and covariates.

	Estimate Std	Error	Z Value	Pr(> |z|)
(Intercept)	−160,808	180,864	−0,889	0,374
Cohort	0,266	0,264	1008	0,313
women	0,947	0,324	2924	0,003**
birthyr	0,074	0,091	0,82	0,412
bornoutsp	0,800	1288	0,621	0,534
as.factor(fatsocnom)2	0,386	0,814	0,475	0,635
as.factor(fatsocnom)3	0,429	1118	0,384	0,701
as.factor(fatsocnom)4	−0,308	0,992	−0,31	0,756
as.factor(fatsocnom)5	0,041	1264	0,033	0,974
as.factor(fatsocnom)6	1454	1,067	1,363	0,173
as.factor(fatsocnom)7	0,592	0,876	0,676	0,499
as.factor(fatsocnom)8	1,134	1,766	0,642	0,521
as.factor(fatsocnom)9	0,296	0,873	0,339	0,735
as.factor(fatsocnom)10	0,038	1586	0,024	0.981
as.factor(motsocnom)2	0,542	0,629	0,861	0,389
as.factor(motsocnom)3	−1,411	1,414	−0,998	0,318
as.factor(motsocnom)4	1,311	1,291	1,015	0,310
as.factor(motsocnom)5	0,693	1,025	0,676	0,499
as.factor(motsocnom)6	2,018	2,449	0,824	0,410
as.factor(motsocnom)7	−0,205	0,407	−0,505	0,614
as.factor(motsocnom)8	1,267	1,399	0,905	0,365
as.factor(motsocnom)9	0,536	0,407	1318	0,187
as.factor(motsocnom)10	0,436	1430	0,305	0,761
fatstud	0,210	0,125	1684	0,092
motstud	0,261	0,123	2133	0,033*
as.factor(entryroutenom)2	−2,922	1108	−2638	0,008**
as.factor(entryroutenom)3	−11,721	733,456	−0,016	0,987
entryscore	0,979	0,150	6,54	0,001***
dropout	NA	NA	NA	NA

Significance codes: ***p < .001. **p < .01. *p < .05.

individuals for any propensity scores in both the matched treatment and the matched controls; on the other hand, a slight outperformance of the treatment group over the control cohort. Both findings are important, because the essential principle of propensity score analysis is that if two individuals are found, one in each treatment or control condition, we can imagine that those two individuals were 'randomly' assigned to each group in the sense of either allocation being equally likely.

The β-values obtained from the regression model applied to raw data of treatment and control group changed from unmatched (β = .12492, t = 2.676, p < .00754**) (R^2 = .004765, F(1, 1286) = 7.162, p = .007541), to matched (β = −0089275, t = −2.245, p < .024944*) (R^2 = .3693, F(31, 1097) = 22.31, p < .2e-16, t = −2.245).

As it can be seen in Table 1, the logistic regression model with t/c as control and covariates shows that, for the statistically significant variables, *Entry-exam score* has the lowest *p*-value, thus suggesting a strong association of the students' entry-exam score with the probability of selecting the bilingual education program, followed by other variables such as *Women (sex)*, *Mother studies* (mother studies) and *Entry route 2*. The negative coefficient for the last predictor suggests that all other variables being equal, students accessing the university through this route (i.e., University entry exam for older than 25) are prone to ending up in the Control condition.

Table 2 shows how many treatment and control cases were eliminated from the sample depending on the matching method used. Improvement on the balance on the covariates comparing Treatment and Control underlies these cases attrition. CEM discarded too many cases from the treatment group to remain a viable option from a power perspective. The other two methods Genetic Matching (GM) and Nearest Neighbor Matching (NNM) yield better results in terms of size and power sample.

Table 3 presents the balance for unmatched, propensity matched and genetic matched covariates means between Treated and Controls cases. Genetic matching means comparisons yield the following values

Table 2
Unmatched and matched sample sizes resulting from each matching method.

	Genetic Matching (GM)		Nearest Neighborhood Matching (NNM)		Coarsened Exact Matching (CEM)	
	Control	Treated	Control	Treated	Control	Treated
All	729	103	729	103	729	103
Matched	74	103	103	103	52	21
Unmatched	655	0	626	0	677	82
Discarded	0	0	0	0	0	0

(β = −0.15058, t = −2.669, p < .00832**), with the multiple (genetic) regression model produced a significant regression equation R^2 = .03362, F(1, 175) = 7.123, p = .008324). While balance was weaker, for sensitivity analysis the nearest neighbor means comparisons are provided, yielding similar results (β = −0.10505, t = −2.027, p < .044*), with the multiple (nearest) neighbor regression model producing a significant regression equation R^2 = .01493, F(1 204) = 4.107, p = .04401).

Finally, with genetic matching the multiple regression model with all predictors produced a (β = −0.1669430, t = −3.659, p < .000349***) and a significant regression equation R^2 = .406, F (26, 150) = 5.626, p < .0001). The multiple (nearest) neighbor regression model with all predictors produce a (β = −0.12751, t = −2.976, p < .00333**) and a significant regression equation R^2 = .3576, F(27, 178) = 5.227, p < .0001).

4. Discussion

In a time of strain on public funds it is critical that academics and policy makers understand the effects of the educational interventions (European Commission, 2013). In this regard, our results provide evidence that there is a cost in academic performance for students taking the bilingual program. This means that two students who were statistically equivalent on all covariates before starting the Degree, except for their choice of enrolling the bilingual group, the one in the bilingual intervention group had a higher likelihood of completing their Degree with a lower GPA. The unexpected initial positive effect of the program on the EMI language group, when comparing raw data, could be attributed to the program effect; however, when controlling for covariates the results where the exact opposite. The reason for this change lies on the matching estimator accounting for the self-selection based on observables, which allows the true (negative) program effect to surface. In this vein, if the self-selection observed from the observables goes in the direction of selecting the better students into treatment, it must be expected to have the same kind of selection based on unobservables. For example, one of the potential confounding factors underlying these net impacts results could be families with higher cultural and social capital persuading their kids to register on the bilingual group, while for others these opportunities are rejected or ignored because of the lack of support from their families and/or avoidance of risk taking. As Hernández-Nanclares and Jiménez-Muñoz (2017) suggest, another potential confounding factor could be the gap between the English level command students develop in high-school and the requirements at HE. Still another potential confounder could be the students' motivation level resulting from the decision of being accepted or excluded from the language group, including the selection mechanism. If that were the case, controlling for the unobservable characteristics of the self-selected sample would probably yield an even more negative treatment effect. In this regard, it can be concluded that even with the limitation of not controlling for selection on unobservables this study provides some evidence that there is a negative effect.

Alternative accounts to this negative effect like differences on learning contents attributed to different curricula, or other confounding

Table 3
Balance for unmatched, Propensity matched and Genetic matched covariates between Treated and Control.

	Unmatched				Nearest Neighbor (Propensity) matched				Genetic matched			
	Means Treated	Means Control	SD Control	Mean Diff	Means Treated	Means Control	SD Control	Mean Diff	Means Treated	Means Control	SD Control	Mean Diff
Distance	0,370	0,089	0,121	0,281	0,370	0,299	0,184	0,071	0,370	0,334	0,228	0,036
Cohort	1573	1495	0,500	0,078	1573	1563	0,498	0,010	1573	1485	0,503	0,087
Women	0,835	0,632	0,485	0,203	0,835	0,854	0,354	−0,019	0,835	0,854	0,355	−0,019
Birth year	1993,061	1992,129	2432	0,939	1993,068	1992,690	2505	0,379	1993,068	1993,058	2131	0,010
Born out Spain	0,010	0,008	0,090	0,001	0,010	0,010	0,098	0	0,010	0,010	0,099	0
Father ocupation1	0,019	0,049	0,217	−0,03	0,019	0,010	0,098	0,0010	0,019	0,019	0,139	0
Father ocupation2	0,388	0,595	0,491	−0,207	0,388	0,437	0,498	−0,048	0,388	0,408	0,495	−0,019
Father ocupation3	0,029	0,029	0,167	0,003	0,029	0,010	0,098	0,0194	0,029	0,010	0,099	0,019
Father ocupation4	0,039	0,101	0,309	−0,068	0,039	0,039	0,194	0	0,039	0,039	0,194	0
Father ocupation5	0,019	0,021	0,142	−0,001	0,019	0,029	0,169	−0,001	0,019	0,010	0,099	0,010
Father ocupation6	0,048	0,012	0,110	0,036	0,048	0,048	0,216	0	0,048	0,010	0,099	0,039
Father ocupation7	0,126	0,075	0,264	0,051	0,126	0,126	0,334	0.000	0,126	0,116	0,323	0,010
Father ocupation8	0,010	0,003	0,052	0,007	0,010	0,010	0,098	0	0,010	0	0	0,010
Father ocupation9	0,3107	0,1029	0,304	0,208	0,311	0,282	0,452	0,029	0,311	0,379	0,488	−0,068
Father ocupation10	0,0097	0,0055	0,074	0,004	0,010	0,010	0,098	0	0,010	0,010	0,099	0
Mother ocupation2	0,0388	0,0274	0,163	0,011	0,039	0,039	0,194	0.000	0,039	0,039	0,194	0
Mother ocupation3	0,0097	0,0151	0,122	−0,005	0,010	0,010	0,098	0.000	0,010	0,010	0,099	0
Mother ocupation4	0,0097	0,0055	0,074	0,004	0,010	0,019	0,139	−0,010	0,010	0,010	0,099	0
Mother ocupation5	0,0194	0,0261	0,159	−0,007	0,019	0,029	0,169	−0,010	0,019	0,010	0,099	0,010
Mother ocupation6	0,0097	0,0014	0,037	0,008	0,010	0	0	0,010	0,010	0,010	0,099	0
Mother ocupation7	0,1262	0,1166	0,321	0,010	0,126	0,126	0,334	0	0,126	0,126	0,334	0
Mother ocupation8	0,0194	0,0014	0,037	0,018	0,019	0,010	0,098	0,010	0,019	0,019	0,139	0

potential factors like teacher English instruction competency (Dearden, 2014), students' language command (Coetzee-Van Rooy, 2010; Kim et al., 2009) and/or teacher practices and/or thoroughness differences, as Bradford (2012) notices, are initially discarded since all groups share the same curricula in the first case, and a balanced distribution of teachers related factors among all participants is assumed for the second and third factors. In this regard, it must be claimed that the negative difference on GPA may be attributed to the effect of the program, since differences in the covariates' distributions have been removed, as Slavin (2008) suggests by: (a) matching both groups on those key covariates that specialized literature recommends; and (b) the statistical techniques utilized have proven useful to evaluate treatment effects when using observational data.

These results, although to some extent limited, hopefully will contribute to mitigate the international scenario of scarcity of data on the success or impact of EMI programs, as Bradford (2012) claimed. Drawing from Hu et al. (2014); Dearden (2014), as we mentioned in an earlier section, one way to raise the quality of education standards in plurilingual education programs would be through developing consensus in HE around the idea of implementing effective plurilingual education policies. In case there are not enough high-quality evidences to support those policies, then the priority would be to create better conditions to develop more adequate datasets in a searchable format for example. These changes, instead, could trigger the process of building further evidence, particularly if interdisciplinary research teams are promoted and collaboration among universities developing similar experiences work. If eventually these conditions are met for long enough, then it is the opportunity for brokerage activities to mediate and facilitate the use of those evidence-based practices and programs by policy-makers.

Another implication for the future of EMI programs that our results entails that if significant improvement on the key factors affecting the quality of these programs is intended, additional outcomes have to be included in the regression models. Rather than solely looking into one outcome such as GPA, other important outcomes linked to this modality of plurilingual education need to be monitored as well. For example, learning a student language proficiency and trade-off may be worth being observed. However, the lack of provision mentioned above on one hand and, on the other hand, the objections posed by the university records office to provide academic records from specific courses (which involves confidentiality concerns) along with the cost in time of preparing those specific data (i.e. scores on English courses) from as many as twenty groups including both cohorts, prevented this research from including academic performance data on that key specific outcome variable.

Future student achievement studies should include as key covariate students' command of English as a baseline, when they start their studies at the University, and continue its measurement throughout the whole period of studies (Kim et al., 2009) including its impact on cognitive development (Coetzee-Van Rooy, 2010). This baseline measure for instance would be much more reliable than those obtained in postsecondary education, although the cost of gathering freshmen data at the writing, speaking and listening level is simply not affordable at this point. Alternativelly, other courses, which are part of this Degree curriculum belonging to the English department, could have been a proxy variable in order to measure the impact of these EMI programs on students' second language command in the future. Another example of research deficit on EMI programs impact and processes is how different students with diverse second language command "develop complex, cognitively demanding content knowledge and skills" (Johnstone, 2010, p. 123). Attention to how these plurilingual education-training models impact on in-service teacher perceptions (Pérez, 2014) and other non-cognitive (soft) skills are also in high demand due to the key role of those competencies in the labour market (Angel, Cabrales, & Carro 2016; De Wit et al., 2015).

5. Limitations

Designing observational studies (Cochran, 1965) to approximate randomized trials is challenging. In the previous sections, we have described and illustrated the use of regression adjustment and propensity scores for the analysis of observational data. In this regard, it is important to note that, while randomized trials allow balance over known and unknown covariates, observational data analysis only allows balance over known covariates, which is an unavoidable limitation. Also, if we thought of a hypothetical randomized experiment in our study that led to the observed dataset, unquestionably an important outcome variable is missing in our study, which is the student's performance on the second language. In this regard, future studies should include this outcome, as Hernández-Nanclares and Jiménez-Muñoz (2017); Jiménez-Muñoz (2014) suggest, so that in case of persistence of unfavorable outcome for treatment students on GPA, potential gains on this competing outcome can be demonstrated. Another controversial issue is the ratios of sample sizes needed to obtain well-matched samples (Rubin, 2008). In this regard, our matching ratio before balance (1156:132) and after balance on the covariates comparing control and treatment (729:103) are above the general recommendation of 4:1 (control to treatment) recommended by (Linden & Samuels, 2013). According to these authors, this ratio elicits, a priory, the lowest bias and allows investigators to maximize the number of controls matched to each treated individual to increase the likelihood that a sufficient sample size will remain after attrition. Alternatively, as Austin (2010) points out, increasing the number of untreated subjects matched to each treated subject tends to increase the bias in the estimated treatment effect and, conversely, increasing the number of untreated subjects matched to each treated subject decreases the sampling variability of the estimated treatment effect. Thus, Austin (2010) recommends matching either 1 or 2 untreated subjects to each treated subject when using propensity-score matching since this ratio minimizes the mean squared error. Consequently, although the distribution of propensity scores between control and treated we have reached is not entirely satisfactory (1:1) (i.e., 103:103) it still lies within the recommendable interval.

Finally, further work is needed to describe and justify the "approximating randomized assignment mechanism" (Rubin, 2008, p. 816) or the admission mechanism involved in this type of studies and its potential consequences on (motivational) covariates distribution. In this regard, future studies based on a discontinuity regression design selecting a certain number of students distributed above and below the entry cut-off grade could provide a research design quite close to randomization.

In sum, our results provide a baseline for evidence on the effects of an undergraduate EMI program implemented at a public research university in Spain. Our negative effects results also align with those found by Angel et al. (2016) although in their case referred to primary education students. In any case, both studies reinforce the need to look at the effect on these important outcomes when evaluating bilingual programs across the whole educational system in Spain and other EMI countries. After all, the implementation of high-quality bilingual education programs and policies depends on the type of evidence available.

Funding

This work was supported by the Junta de Andalucía-funded Proyecto de Excelencia: "Análisis y Garantía de Calidad de la Educación Superior Plurilingüe en la Educación Superior de Andalucía [Junta de Andalusia-funded Project of Excelence: Analysis and Warrantee of the Quality of Plurilingual Higher Education in Andalucia] (AGCEPESA; Grant Agreement No. P12-SEJ − 1588).

References

Aguilar, M., & Rodríguez, R. (2012). Lecturer and student perceptions on CLIL at a Spanish university. *International Journal of Bilingual Education and Bilingualism, 15*(2), 183–197. http://dx.doi.org/10.1080/13670050.2011.615906.

Airey, J. (2009). *Science, language and literacy. Case studies of learning in Swedish university physics (Doctoral dissertation)*. Uppsala, Sweden: Faculty of Science and Technology Retrieved on June 2, 2017 from: http://www.diva-portal.org/smash/get/diva2:173193/FULLTEXT01.pdf.

Angel, B., Cabrales, A., & Carro, J. M. (2016). Evaluating a bilingual education program in Spain: The impact beyond foreign language learning. *Economic Inquiry, 54*(2), 1202–1223. http://dx.doi.org/10.1111/ecin.12305.

Arco, J. L., & Fernández, F. D. (2016). Skills learning through a bilingual mentors program in higher education. *International Journal of Bilingual Education and Bilingualism,* 1–11. http://dx.doi.org/10.1080/13670050.2016.1228601.

Austin, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology, 172*(9), 1092–1097. http://dx.doi.org/10.1093/aje/kwq224.

Bradford, A. (2012). English-medium degree programs in Japanese universities: Learning from the European experience. *Asian Education and Development Studies, 2*(3), 225–240. http://dx.doi.org/10.1108/AEDS-06-2012-0016.

Byun, K., Chu, H., Kim, M., Park, I., Kim, S., & Jung, J. (2011). English-medium teaching in Korean higher education: Policy debates and reality. *Higher Education, 62*(4), 431–449. http://dx.doi.org/10.1007/s10734-010-9397-4.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72. http://dx.doi.org/10.1111/j.1467-6419.2007.00527.x.

Card, D., Ibarraran, P., & Villa, J. M. (2011). *Building in an evaluation component for active labour market programs: A practitioner's guide (Discussion Paper No. 6085)*. Bonn, Germany: IZA Retrieved on June 20, 2017 from http://ftp.iza.org/dp6085.pdf.

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods, 21*(3), 427–445. http://dx.doi.org/10.1037/met0000076.

Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, 128*, 134–155. http://dx.doi.org/10.2307/2344179.

Coetzee-Van Rooy, S. (2010). Complex, multilingualism and academic success in South African higher education. *Southern African Linguistics and Applied Language Studies, 28*(4), 309–321. http://dx.doi.org/10.2989/16073614.2010.548021.

Dafouz, E., & Camacho-Miñano, M. M. (2016). Exploring the impact of English-medium instruction on university students academic achievement: The case of accounting. *English for Specific Purposes, 44*, 57–67. http://dx.doi.org/10.1016/j.esp.2016.06.001.

Dafouz, E., & Guerrini, M. (2009). *CLIL across education levels: Experiences from primary, secondary and tertiary contexts.* Madrid, Spain: Richmond Publishing.

Dafouz, E., & Smit, U. (2016). Towards a dynamic conceptual framework for English-medium education in multilingual university settings. *Applied Linguistics, 37*(3), 397–515. http://dx.doi.org/10.1093/applin/amu034.

Dafouz, E., Camacho-Miñano, M. M., & Urquia, E. (2014). Surely they can't do as well: A comparison of business students' academic performance in English-medium and Spanish-as-first-language-medium programmes. *Language and Education, 28*(3), 223–236. http://dx.doi.org/10.1080/09500782.2013.808661.

Dafouz, E., Núñez, B., Sancho, C., & Foran, D. (2007). Integrating CLIL at the tertiary level: Teachers' and students' reactions. In D. Marsh, & D. Wolff (Eds.). *Diverse contexts-converging goals. CLIL in europe* (pp. 91–101). Frankfurt Germany: Peter Lang. http://dx.doi.org/10.1093/applin/amu010.

Dafouz, E., Núñez, B., & Sancho, C. (2007). Analysing stance in a CLIL university context: Non-native speaker use of personal pronouns and modal verbs. *International Journal of Bilingual Education and Bilingualism, 10*(5), 647–662. http://dx.doi.org/10.2167/beb464.0.

De Wit, H., Hunter, F., Howard, L., & Egron-Polak, E. (2015). *Internationalisation of higher education.* Brussels, Belgium: European Unionhttp://dx.doi.org/10.2861/444393.

Dearden, J. (2014). *English as a medium of instruction − a growing global phenomenon.* Oxford, UK: British Council and Oxford University Department of Education.

Del Campo, C., Cancer, A., Pascual-Ezama, D., & Urquía-Grande, E. (2015). EMI vs. Non-EMI: Preliminary analysis of the academic output within the INTE-R-LICA project. *Procedia –Social and Behavioral Sciences, 212*, 74–79. http://dx.doi.org/10.1016/j.sbspro.2015.11.301.

Diamond, A., & Sekhon, J. S. (2015). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics, 95*(3), 932–945. http://dx.doi.org/10.1162/REST_a_00318.

Doiz, A., Lasagabaster, D., & Sierra, J. M. (2011). Internationalisation, multilingualism and english-medium instruction. *World Englishes, 30*(3), 345–359. http://dx.doi.org/10.1111/j.1467-971X.2011.01718.x.

Doiz, A., Lasagabaster, D., & Sierra, J. M. (2013). Future challenges for English-mMedium instruction at the tertiary level. In A. Doiz, D. Lasagabaster, & J. M. Sierra (Eds.). *English-medium instruction at universities: Global challenges* (pp. 213–221). Bristol, UK: Multilingual Matters.

Dor, D. (2004). From englishization to imposed multilingualism: Globalization, the internet, and the political economy of the linguistic code. *Public Culture, 16*(1), 97–118 Retrieved from: https://muse.jhu.edu/article/54374/pdf.

European Commission (2013). *Design and commissioning of counterfactual impact evaluations.* Luxembourg: Publications Office of the European Unionhttp://dx.doi.org/10.2767/94454.

European Ministers in charge of Higher Education (1999). *The bologna 1999 declaration.* Retrieved from https://www.eurashe.eu/library/bologna_1999_bologna-declaration-pdf/.

Fernández-Costales, A., & González-Riaño, X. A. (2015). Teacher satisfaction concerning the implementation of bilingual programmes' in a Spanish University. *Porta Linguarum, 23*, 93–108.

Fernández-Viciana, A., & Fernández-Costales, A. (2017). El pensamiento de los futuros maestros de inglés en Educación Primaria: Creencias sobre su autoeficacia docente [Thinking of future English teachers in Primary Education: Believes on their teaching self-efficacy]. *Bellaterra Journal of Teaching & Learning Language & Literature, 10*(1), 42–60. http://dx.doi.org/10.5565/rev/jtl3.684.

Gordon, R. A. (2015). *Regression analysis for the social sciences.* New York, NY: Routledge. Taylor & Francis Group.

Guill, K., Lüdtke, O., & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: Evidence from a propensity score matching study. *Learning & Instruction, 47*, 43–52. http://dx.doi.org/10.1016/j.learninstruc.2016.10.001.

Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education, 75*(1), 31–65. http://dx.doi.org/10.3200/JEXE.75.1.31-65.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 5*(3), 234–249. http://dx.doi.org/10.1037/a0019623.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research.* IES National Center for Special Education Research Retrieved from https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf.

Hellekjaer, G. O. (2008). A case for improved reading instruction for academic English reading proficiency. *Acta Didactica Norge, 2*(1), 1–17. http://dx.doi.org/10.5617/adno.1022.

Hernández-Nanclares, N., & Jiménez-Muñoz, A. (2017). English as a medium of instruction: Evidence for language and content targets in bilingual education in economics. *International Journal of Bilingual Education and Bilingualism, 20*(7), 883–896. http://dx.doi.org/10.1080/13670050.2015.1125847.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199–236. http://dx.doi.org/10.1093/pan/mpl013.

Holland, P. (1986). Statistics and causal inference. *Journal of American Statistical Association, 81*, 945–970.

Hu, G., Li, L., & Lei, J. (2014). English-medium instruction at a Chinese University: Rhetoric and reality. *Language Policy, 13*(1), 21–40. http://dx.doi.org/10.1007/s10993-013-9298-3.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. http://dx.doi.org/10.1093/pan/mpr013.

Jensen, T. M., Shafer, K., Guo, S., & Larson, J. H. (2017). Differences in relationships stability between individuals in first and second marriages: A propensity score analysis. *Journal of Family Issues, 38*(3), 406–432. http://dx.doi.org/10.1177/0192513x15604344.

Jiménez-Muñoz, A. J. (2014). Measuring the impact of CLIL on language skills: A CEFR-based approach for Higher Education. *Language Value, 6*(1), 28–50. http://dx.doi.org/10.6035/LanguageV.2014.6.4.

Learning trough English: Policies, challenges and prospects. In R. Johnstone (Ed.). *Insights from east asia.* Kuala Lumpur, Malaysia: British Council.

Kim, A., Son, Y. D., & Sohn, Y. (2009). Conjoint analysis of enhanced English Medium Instruction for college students. *Expert Systems Applications, 36*, 10197–10203. http://dx.doi.org/10.1016/j.eswa.2009.01.080.

Kirkpatrick, A. (2011). *Internationalization or englishization: Medium of instruction in today's universities.* Hong Kong: Centre for Governance and Citizenship Working Paper Series 2011/003 Retrieved from: http://repository.lib.ied.edu.hk/pubdata/ir/link/pub/AK%20CGC%20occasional%20paper%20final_final_——%20Prof%20%20Kirkpatrick.pdf.

Klaassen, R. (2001). *The international university curriculum: Challenges in English-medium engineering education (Doctoral dissertation)*. Delft, Netherlands: Delft University of Technology Retrieved on June 2, 2017 from https://repository.tudelft.nl/islandora/object/uuid:dea78484-b8c2-40d0-9677-6a508878e3d9/datastream/OBJ/download.

Kretschman, J., Vock, M., & Lüdtke, O. (2014). Acceleration in elementary school: Using propensity score matching to estimate the effects on academic achievement. *Journal of Educational Psychology, 106*(4), 1080–1095. http://dx.doi.org/10.1037/a0036631.

Linden, A., & Samuels, S. J. (2013). Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice, 19*(5), 968–975.

Ministerio de Educación y Ciencia (2007). *Real Decreto 1393/2007, de 29 de octubre por el que se establece la ordenación de las enseñanzas universitarias oficiales* Retrieved from http://www.aneca.es/eng/Evaluation-Activities/VERIFICA/Bachelor-s-Degree-and-Master-s-Degree/Regulations.

Neuville, S., Frenay, M., Schmitz, J., Boudrenghien, G., Noël, B., & Wert, V. (2007). Tinto's Theoretical perspective and expectancy-value paradigm: A confrontation to explain freshmen's academic achievement. *Psychologica Belgica, 47*(1/2), 31–50. http://dx.doi.org/10.5334/pb-47-1-31.

Olmos, A., & Govindasamy, P. (2015). Propensity scores: A practical introduction using R. *Journal of Multidisciplinary Evaluation, 11*(25), 68–88.

Pérez, M. L. (2014). Teacher training needs for bilingual education, in-service teacher perceptions. *International Journal of Bilingual Education and Bilingualism, 19*(3), 266–295. http://dx.doi.org/10.1080/13670050.2014.980778.

Ramos-García, A. M. (2013). Higher education bilingual programmes in Spain. *Porta Linguarum, 19*, 101–111 Retrieved from http://www.ugr.es/~portalin/articulos/

PL_numero19/7%20A%20M%20Ramos.pdf.
Ramos-García, A., Arco-Tirado, J. L., Fernandez-Martín, F. D., & Villoria-Prieto, J. (2016). Towards a bilingual education model in higher education in a non- anglophone country: The case of the Bilingual Group in Primary Education Teacher Training Degree at University of Granada (Spain). *Paper presented at the annual meeting for Bilingual Education: II Congreso Internacional Sobre Educación Bilingüe* 15-18 Noviembre.
Rubin, D. B. (2008). For objective causal inference design trumps analysis. *The Annals of Applied Statistics, 2*(3), 808–840. http://dx.doi.org/10.1214/08-AOAS187.
Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.
Shadish, W., Luellen, J., & Clark, M. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin, & P. E. McKnight (Eds.). *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington DC: American Psychological Association. http://dx.doi.org/10.1016/j.jclinepi.2004.10.016.
Slavin, R. E. (2008). Evidence-based reform in education: Which evidence counts? *Educational Researcher, 37*(1), 47–50. http://dx.doi.org/10.3102/0013189x08315082.
Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*, 250–267. http://dx.doi.org/10.1037/a0018719.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1–21. http://dx.doi.org/10.1214/09-STS313.
Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the Social Sciences. *Multivariate Behavioural Research, 46*(1), 90–118. http://dx.doi.org/10.1080/00273171.2011.540475.
Tinto, V. (1997). Classrooms as communities. *Journal of Higher Education, 68*, 599–623.
Ward, A., & Johnson, P. J. (2008). Addressing confounding errors when using non-experimental, observational data to make causal claims. *Syntheses, 163*, 419–432. http://dx.doi.org/10.1007/s11229-007-9292-4.

Prof. Jose L. Arco-Tirado, Ph.D., has eighteen years of teaching and research experience. Before entering the University of Granada he worked at the Provincial Center for Drug addiction and the Andalusian School of Public Health. Currently he teaches two courses on a bilingual group in Primary Education Teacher Training. Dr. Arco-Tirado focuses his research interests on different topics like plurilingual education, service-learning, self-regulation and public program evaluation. He has published several books and articles in national and internationally indexed and recognized journals and editorials. He is part of the research project founded by the Regional government Junta de Andalucía on Quality of Higher Education Plurilingual programs.

Prof. Francisco D. Fernández, PhD., works at the Department of Developmental and Educational Psychology, University of Granada (Spain). For almost 13 years I have been teaching in Higher Education at the same time implementing several innovation and research projects aimed at improving quality of Education across the Educational System. I have research experience on national and international projects. I have published several articles, books, book´s chapters and papers on nationally and internationally recognized journals, publishers, congress and conferences. He is part of the research project founded by the Regional government Junta de Andalucía on Quality of Higher Education Plurilingual programs.

Prof. Ana M. Ramos belongs to the Dept. of Didactics of Language and Literature at the University of Granada. She has ten years of teaching and research experience on foreign languages teaching. She is a member of the scientific committee of Portal Linguarum an international and interuniversity journal of foreign language didactics, indexed in the Arts & Humanities Citation Index, SCOPUS, Latindex, The Linguist, MLA, the ISOC, DIALNET and UCUA. She has published several books and scientific articles in both national and international journals. She is part of the research project founded by the Regional government Junta de Andalucía on Quality of Higher Education Plurilingual programs.

Prof. Levente Littvay researches survey and quantitative methodology, twin and family studies (as the co-director of the Hungarian Twin Registry), and the psychology of radicalism and populism. Received around a half million in grants for his research and is an award-winning teacher of graduate courses in applied statistics, electoral politics, voting behavior, political psychology, American politics. He is an academic co-convener of ECPR's Methods Schools and an Associate Editor of Twin Research and Human Genetics.

Prof. Javier Villoria belongs to the Dept. of Didactics of Language and Literature at the University of Granada. He has fifteen years of teaching and research experience on several topics in the field. Member of the scientific committee of Portal Linguarum an international and interuniversity journal of foreign language didactics, indexed in the Arts & Humanities Citation Index, SCOPUS, Latindex, The Linguist, MLA, the ISOC, DIALNET and UCUA. He has several publications of books and articles in both national and international journals. Currently he is the Dean of the Faculty of Education at UGR. He is part of the research project founded by the Regional government Junta de Andalucía on Quality of Higher Education Plurilingual programs.

Prof. Jose A. Naranjo belongs to the Dept. of Didactics Experimental Sciences at the University of Granada. He has over 25 years of teaching and research experience in the field. He has either directed or collaborates on national and international research projects. He has published several books and articles in both national and international journals. He is part of the research project founded by the Regional government Junta de Andalucía on Quality of Higher Education Plurilingual programs. Currently he is the vice-rector of students and employability at the UGR.