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