RUHR

ECONOMIC PAPERS

Julia Bredtmann
Carsten J. Crede
Sebastian Otten

Methods for Evaluating Educational Programs – Does Writing Center Participation Affect Student Achievement?

RUHR UNIVERSITÄT BOCHUM RUB RWI #275

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #275

Responsible Editor: Thomas K. Bauer

ISSN 1864-4872 (online) – ISBN 978-3-86788-320-7

Ruhr Economic Papers #275

Julia Bredtmann, Carsten J. Crede, and Sebastian Otten

Methods for Evaluating Educational Programs – Does Writing Center Participation Affect Student Achievement?

RUHR UNIVERSITÄT BOCHUM RUB

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

ISSN 1864-4872 (online)
ISBN 978-3-86788-320-7

Julia Bredtmann, Carsten J. Crede, and Sebastian Otten[1]

Methods for Evaluating Educational Programs – Does Writing Center Participation Affect Student Achievement?

Abstract

This paper evaluates the effectiveness of the introduction of a Writing Center at a university. The center has the purpose to provide subject-specific courses that aim to improve students' abilities of scientific writing. In order to deal with presumed self-perceptional biases of students in feedback surveys, we use different quantitative evaluation methods and compare the results to corresponding qualitative student surveys. Based on this evaluation, we present and discuss the validity of the approaches to evaluate educational programs. Although almost all students reported the writing courses to be helpful, we find no significant effect of course participation on students' grades. We attribute the difference in the results between quantitative methods and qualitative surveys to the inappropriateness of student course evaluations for assessing the effectiveness of educational measures.

JEL Classification: I20, I21, C81

Keywords: Performance evaluation; educational programs; student evaluation; empirical methods

August 2011

1 *Julia Bredtmann, RUB and RWI; Carsten J. Crede, RUB; Sebastian Otten, RUB and RWI.* *– The authors are grateful to Thomas K. Bauer and Christoph Ehlert for helpful comments and suggestions and Viktoria Frei for excellent research assistance. We also thank Anika Limburg and the Department of Economics at Ruhr University Bochum for providing the data. All remaining errors are our own.* *– All correspondence to Julia Bredtmann, Ruhr-Universität Bochum, 44780 Bochum, Germany, E-Mail: julia.bredtmann@rwi-essen.de.*

1. Introduction

In recent years, there has been a growing trend in universities to employ students' course evaluations to improve teaching quality and perform cost-benefit analysis. These evaluations are often carried out using qualitative feedback surveys of students – which after decades of debate – are still a controversial issue. Against this background, the following analysis of the effectiveness of a new Writing Center for students that was designed to improve their writing ability refrains from solely relying on students surveys to evaluate educational measures. To avoid the problems associated with students feedback surveys, different identification strategies are employed to measure the causal effect of WrC participation on students' achievement using the written examination grades of both participants and non-participants as outcome measure. Based on a perfect experiment as benchmark for the evaluation of the writing courses, the different strategies and necessary assumptions to establish comparability of the groups in limited observational data are presented. In addition to the cross-sectional comparison applicable for cross-sectional data, we employ the before-after comparison as well as the pooled Difference-in-Differences estimator for repeated cross-section data and the Fixed-Effects Difference-in-Differences estimator for panel data analysis. The results of these quantitative methods are then compared to those of the qualitative feedback surveys to examine the validity of student evaluations.

Studies of Remedios and Lieberman (2008) and Centra (2003) as well as older literature surveys indicate that although some factors unrelated to teaching influence students' evaluations, the evaluation remains reliable, valid and mostly unaffected of bias (see, e.g., Marsh, 1987; Marsh and Roche, 2000; Cashin, 1995). However, more recent studies find various biases undermining the validity of student course evaluations. Francis (2011) and McNatt (2010) discover that pre-course attitudes of students have an important influence on subsequent course evaluation. Findings of Davies et al. (2007) indicate that next to various course factors partly unrelated to the actual teaching quality, evaluations are also affected by the cultural background of the students. In contrast to findings of Centra (2003), McPherson et al. (2009) conclude that course instructors can improve their course ratings by raising the students' grade expectations. The contradictory results of the validity of student feedback surveys are casting doubts on the use of this qualitative approach for institutional evaluation of educational measures. In the case

of Writing Center evaluation, it is clear that student surveys measure the degree of satisfaction with their course and indicate their impression of its effectiveness. However, Walker and Palmer (2009) found that many students lack ability to assess their own level of understanding of course material. This perceived learning effect might be heavily influenced by the performance of the instructor (Jaarsma et al., 2008; Lo, 2010). The findings undermine the informative value of such results. In other environments, e.g., lectures, these qualitative surveys are even more problematic. Besides the various problems listed earlier, it is not clear what they measure specifically. Students might evaluate to what extend the course prepares them for tasks in working life and trains independent thinking. However, they also might just indicate the course's success in teaching a certain amount of content in a limited period.

In the following section, we describe the data used for our empirical analysis. In section 3, a description of the different identification strategies used to evaluate the effectiveness of the WrC is presented and the individual advantages and disadvantages as well as the requirements concerning the data are described. In addition to this, information about the implementation of the different approaches is provided. The empirical results are presented in section 4. Section 5 concludes and section 6 briefly outlines how similar evaluation studies should be carried out in future analyses.

2. Data

2.1. Qualitative data

The Writing Center (WrC) was established in August 2009 at the Faculty of Business Administration and Economics at Ruhr University Bochum, Germany. The fee-free and voluntary service was introduced to improve the ability of students to write scientific papers and help with specific problems they are facing while creating written seminar assignments as well as Diploma, Bachelor and Master theses. The WrC does not provide any feedback with regards to content or proofreading. It rather informs about scientific language style and supports students in realizing a successful writing process. Students can either attend face-to-face counseling for the discussion of individual problems during their writing assignments or workshops comprised of small groups dealing with different aspects of scientific writing. The student feedback surveys provided by WrC visitors, which the faculty so far relied on, indicated very good responses of the students towards the service. After

workshop attendance, 103 students provided feedback by stating their overall impression of the course. They could choose five different answers on a qualitative survey sheet ranging from *insufficient* to *excellent*. Roughly 45% of all visitors reported the workshop to be *excellent*, 47% answered *good* and 8% *satisfactory*, whereas no one replied with *unsatisfactory* or *insufficient*.

2.2. Quantitative data

The empirical analysis is based on administrative student records provided by the Examination Department of the faculty, which contain over 188,000 entries of all examinations conducted at the faculty since 1994. It does not only feature extensive information about the exams such as grade, nature of the exam, and respective semester of entry, it also comprises socioeconomic information about the corresponding student, including gender, age, and citizenship. The entries are assigned to students with a uniquely identifying student identification number (ID) that every student receives after enrollment at the university. In addition to these administrative data, a survey was carried out to identify students having visited the Writing Center. In order to collect these information, the Writing Center kept record of all students using their services (by asking them to provide their student ID), their dates of visit and the type of service they received (individual consultancy or workshop training). Using the student ID of each student, we were able to merge the two data sets. Since the Writing Center was introduced for the purpose of students to overcome challenges in written seminar assignments and Diploma, Bachelor and Master theses, we choose the grades received for these tasks ($WExG$) as our outcome measure.[1] Subsequently, all estimation coefficients measure the absolute effect on the test score, which ranges from 1.0 to 5.0.[2]

The following explanatory variables are included in the analysis. An indicator variable *Female* controls for gender-based study performance differences. For the purpose of controlling for German citizenship, we include the dummy *German* as nationals are expected to have better grades than exchange or foreign students. However, this variable has limited explanatory

[1] Throughout the text, the expression written examinations is used to refer to these examinations.

[2] Grades are sub-divided into 12 grade points, where 1.0 is best and 4.0 is the minimum pass grade. Therefore, a negative coefficient implies a positive influence on the written examination grade.

power, as it does not contain information on a possible migration background of persons with German nationality. Thus, it can rather be interpreted as an indicator for German being the first language. Furthermore, we include a variable for the students' age *(Age)* and a dummy variable flagging observations of students that are in semesters exceeding the prescribed period of study *(LtStud)*. Higher age might imply more experience and therefore a better performance in the course of study, whereas students with an above-average number of semesters are expected to be primarily those showing a lower ability to master their study. To separate Bachelor students from Diploma and Master students, we include the dummy (*Bachelor*) to mark Bachelor students.[3] The variable *AvGrade* represents the average grade that a student has received in all his end-of-term examinations in the course of studies, i.e., it consolidates all exam grades. It was calculated using the panel character of the database under exclusion of all entries after the introduction of the WrC and the written examination grades (theses grades), which serve as our outcome measure. It acts as a proxy variable for unobserved motivation and cognitive abilities, which are supposed to have a negative effect on the written examinations – the more distinct they are, the better is the grade. In addition, we use *Thesis* to flag entries of Diploma/Bachelor/Master theses to take into account that theses usually tend to be better graded than seminar assignments due to the higher experience of students. If not included, the effect of Writing Center participation on students' performance would be overestimated since the relative share of final examinations is higher among those having attended the Writing Center.

As the data only cover two semesters (half-year terms in Germany) after the introduction of the Writing Center (summer term 2010 and winter term 2010/2011), we restrict the data to observations of two semesters before (summer term 2009 and winter term 2009/2010) and the two available semesters after the introduction of the WrC. This mostly prevents estimates to be biased by changing teaching environments such as changing curricula, new professors or grade inflation.

Due to the fact that the Writing Center is a new institution and a relatively small number of students participated in the WrC yet, we decided not

[3]As Diploma and Master students are to a large extent similar in their characteristics as well as in their study environment and there are only few Master students in the data, we do not separately control for these two degree programs.

to differentiate between the effects of individual consultancy and workshop training. Therefore, whenever using information on whether a student has visited the WrC or not, we only differentiate between participation in either these two programs and no participation at all and assume that the effects of these programs are similar to a large degree.

After excluding observations with missing information on at least one of the variables used in the empirical analysis we end up with a sample of 2,414 observations with 229 entries of participants and 2,185 of non-participants. Thus, 9% of all observations represent students having received Writing Center support. Comparing the observations of Writing Center visitors with those of non-visitors shows that the two groups differ significantly in grades, gender and nationality, indicating strong self-selection in WrC participation (see Table 1). Most striking is the highly above-average participation of women in the program, which might be explained by a more pessimistic or critical self-assessment of writing abilities of female students or by female students being more motivated and diligent than their male felllows. Another significant difference is the above-average participation of German students, which could bias the WrC effect as students with superior language skills are more likely to perform better in the course of their studies and particularly in written exams. In addition, the problem of self-selection is amplified by the fact that WrC visitors have a significantly better average grade, suggesting that better students tend to use the service more often. This fact has strong implications for the empirical analysis. If visitors constitute a non-representative sample of all students with higher motivation and/or cognitive talent, the evaluation of the WrC effectiveness by comparing average scores of the written examination grades of the two groups would lead to biased estimates. Hence, although visitors have significantly better examination grades than non-visitors (2.68 vs. 2.84), this does not allow any conclusions regarding the causal impact of the WrC.

3. Methods

The main problem of evaluation of policies is the differentiation of correlation and causality. Ideally, the causal effect of an intervention on participants would be identified using a randomized controlled trial (RCT). Random assignment of a large number of persons to either a treatment group or a control group equally distributes all relevant observable and unobserv-

Table 1: Descriptive Statistics

	All Mean/StdDev	Participants Mean/StdDev	Non-Participants Mean/StdDev	Diff. in Means t-value
Grade of Written Examination	2.82	2.68	2.84	1.98^{**}
	(1.16)	(1.07)	(1.16)	
Female	0.41	0.56	0.40	-4.72^{***}
	(0.49)	(0.50)	(0.49)	
German	0.89	0.95	0.88	-3.02^{***}
	(0.32)	(0.22)	(0.32)	
Age	24.60	24.81	24.58	-1.38
	(2.39)	(2.41)	(2.39)	
Bachelor	0.40	0.26	0.41	4.61^{***}
	(0.49)	(0.44)	(0.49)	
Long-term student	0.23	0.27	0.23	-1.35
	(0.42)	(0.44)	(0.42)	
Average Grade	3.05	2.95	3.06	2.40^{**}
	(0.67)	(0.56)	(0.68)	
Thesis	0.14	0.18	0.13	-2.21^{**}
	(0.34)	(0.39)	(0.34)	
Writing Center Participant (WrCP)	0.09	–	–	–
	(0.29)			
Introduction Writing Center (IntWrC)	0.53	0.46	0.53	2.07^{**}
	(0.50)	(0.50)	(0.50)	
Observations	2414	229	2185	2414

Notes: – Significant at: ***1% level; **5% level; *10% level. – Calculations based on all observations, not person. – Grade of Written Examination is measured on a scale of 1.0 to 5.0, 1.0 is best. – Female, German, Bachelor, Long-time student, Thesis, WrCP and IntWrC are dummy variables, WrCP marks WrC participants and IntWrC observations after the introduction of the WrC.

able characteristics across both groups and facilitates their comparability. In such a situation, differences in mean values of the outcomes of both treatment group and control group after program participation can be attributed to the effect of program participation on participants (*treatment effect on the treated*). However, due to possible ethical, political or financial problems and limitations, respectively, randomized controlled experiments are not always feasible (Kluve et al., 2007). In case of the WrC, the best way to evaluate its effectiveness would have been to carry out a RCT in advance to its introduction by randomly assigning students to either of the two groups and comparing their average written examination grades afterwards. This arbitrary selection of participants was not a feasible solution, as it would have forced some students to take part against their intention while preventing others from desired participation. Instead, the service was established on a voluntary basis in order to let students decide themselves whether to attend counseling or not.

In such situations observational studies have to be carried out, in which the data is not derived in a manner guaranteeing a random assignment to the treatment similar to that of the RCT. Most important, the equal distribution of students with respect to their characteristics into the treatment group and control group is highly unlikely, as most likely students with certain charac-

teristics or the believe that they will profit from WrC visitation will decide to use the service. Therefore, in our case it is not sufficient to compare average written examination grades of WrC visitors with those of non-visitors to evaluate the impact of WrC participation on students' achievement, since participants significantly differ from non-participants with respect to important individual characteristics. In particular, students with higher average scores tend to use the WrC services more often than their fellow students with lower grades. If these students also perform better in written examinations, the corresponding higher mean value does not imply any causal effect of WrC participation on students' grades.

In order to measure the causal impact in observational data, one would like to observe the value of the outcome variable of a participant in case he would not have participated. This unobservable state is referred to as *counterfactual situation* and can never be observed. The objective of any observational study is to approximate this counterfactual situation from the available data. This is only possible through some restrictive *identification assumptions*. The quantitative approaches outlined in this paper rely on different identification assumptions, which cannot be tested. The comparison to the perfect experiment, however, provides some indication on the severeness of the identification assumption.

In this study, we define the following sub-groups. All students ever having visited the WrC constitute the treatment group, whereas all other students are assigned to the control group. Further, we distinguish between the time before the WrC was introduced (before treatment) and after its introduction (after treatment). All models applied examine the average treatment effect of WrC participation on its visitors, the so-called *average effect of treatment on the treated*. To isolate the effects of the WrC program on participants, we employ four different empirical methods applicable to observational data and examine to what extent they produce valid results in this special case. In doing so, we outline their data requirements and identification assumptions.

3.1. Cross-sectional comparison

Following the idea of the cross-sectional comparison (CSC), the written examination grades of WrC visitors are compared with those of students not having received any advisory (Cameron and Trivedi, 2005). In Figure 1, participants are marked with a continuous horizontal line and non-participants with a dashed line. The time before the WrC introduction is referred to as t_{-1}, the point of introduction as t_0 and the time after introduction as t_1. The

difference γ (A-B) between these groups is the average treatment effect on the treated.

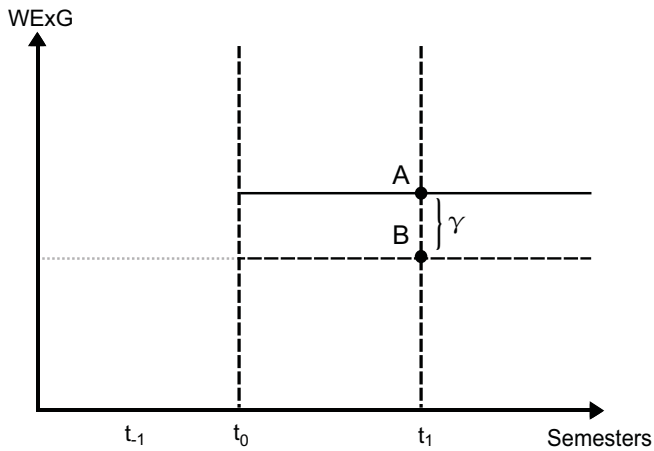

Figure 1: Cross-Sectional Comparison
Source: Bauer et al. (2009).

The counterfactual situation – how would a WrC visitor have performed without participation – is identified by assuming that if the treated had not participated, they would have had the same average grade as the control group in t_1, i.e., the two groups did not differ in advance to the introduction of the WrC in t_0.

The advantage of this approach is the fact that it does not require panel data but can be applied to simple cross-sectional data. Its weakness is its strong identification assumption that in most cases may not be legitimate. The causal effect of WrC participation on students' grades estimated with this method is considered to be biased; Table 1 documents strong self-selection of WrC visitors into participation, contradicting the validity of the identification assumption. Moreover, important unobserved characteristics may affect both the students' written examination grade as well as the decision to participate in the WrC. For example, one might argue that higher motivated students tend to use the WrC to a higher extend and have better written examination grades – irrespective of WrC participation. If being correlated with any of the

included variables, this unknown variable hidden in the error term potentially leads to biased estimates of all coefficients, which is especially problematic for the treatment effect as it prevents the correct identification of γ. Although we use the students' average grade as a proxy variable for their motivation and cognitive talent to capture at least a part of the unobserved features of a student, the results are still at risk of being biased.

3.2. Before-after comparison

For the application of the before-after comparison (BAC), repeated cross-section data are required, i.e., observations in at least two points of time – one prior to the WrC introduction (t_{-1}) and one afterwards (t_1) – are needed. To identify the causal effect of WrC participation, the written examination grades of WrC participants before WrC attendance are compared with those after the visits. As can be seen in Figure 2, a change in grades between these two points of time is considered to be the treatment effect δ (Wooldridge, 2010).

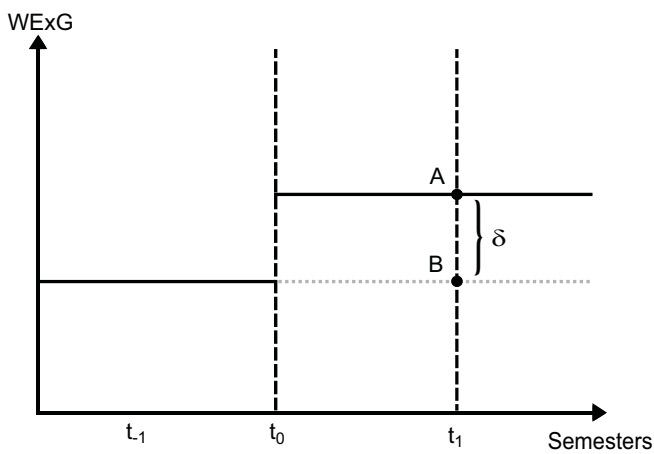

Figure 2: Before-After Comparison
Source: Bauer et al. (2009).

The BAC assumes that without engaging in the WrC, the students' written examination grades would have remained unchanged (dotted grey line). In contrast to the cross-sectional comparison, the before-after comparison features the same persons in the treatment group (after WrC visit) and in the

control group (before the WrC visit). Therefore, time-constant unobserved heterogeneity cannot bias the estimations. However, note that if participants select themselves into treatment, the estimated treatment effect is not valid for all students, since treatment could affect participants and non-participants differently. Furthermore, it is susceptible to unobserved time-variant factors affecting the written examination grade. Grade inflation, new professors or changed types of examinations as well as unobserved learning effects of students caused by gains in experience might influence the written examination grades between the two points of time used for comparison. This may result in biased estimates of the treatment effect, as all these influences would be attributed to the causal effect of WrC participation. However, as we limited our data to two years (four semesters), time-variant unobserved factors affecting our dependent variable could be a minor problem.

3.3. Pooled Difference-in-Differences estimator

The Difference-in-Differences (DiD) estimator combines the ideas of the cross-sectional and the before-after comparison and makes use of both longitudinal and cross-sectional information in the data (Schlotter et al., 2011; Wooldridge, 2010). As can be seen in Figure 3, in a first step the difference in grades between participants (continuous line) and non-participants (dashed line) both before (C-D) and after (A-B) WrC visitation and introduction, respectively are calculated. Then, the difference between these points of time is being compared. This double difference is the treatment effect, the effect of WrC participation on students' grades. In case panel data on participants and non-participants do not exist and individuals cannot clearly be identified both before and after the treatment, the pooled DiD estimator (PDiD) can be used to estimate the effect by comparing the mean values of the two groups.

The underlying identification assumption of the pooled DiD approach is that without the introduction of the WrC, the difference in grades between the two groups would have stayed constant over time. Hence, in contrast to the before-after comparison, the pooled DiD estimator allows for time-variant influences on the written examination grade affecting both groups equally, such as grade inflation. Moreover, the pooled DiD approach controls for systematic differences between treatment and control group, i.e., differences in students' mean grades in the absence of treatment. However, the validity of the pooled DiD approach presupposes exogenous treatment, i.e.,

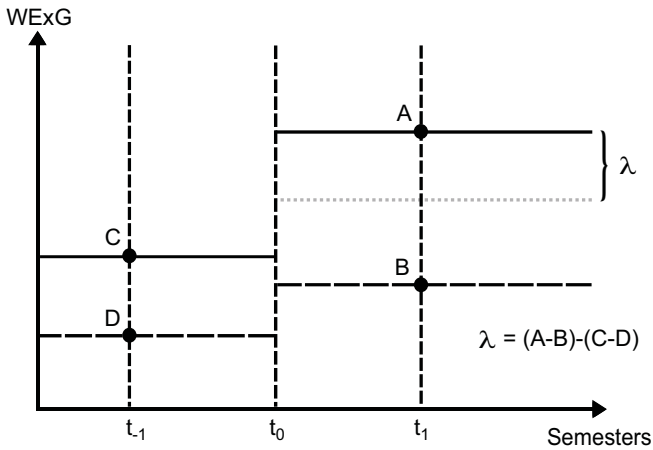

Figure 3: The Difference-in-Differences Estimator
Source: Bauer et al. (2009).

participation in the program may not be voluntarily chosen by the participants. This is the case in a so-called *natural experiment*, in which due to an exogenuous event such as new laws or programs a "random" distribution into participation and non-participation is guaranteed. Assume, for example, that the WrC has been introduced for the purpose of specifically helping foreign students in overcoming writing difficulties. While for them participation in the WrC is compulsory, German students are not allowed to utilize this service. In this case, we have an exogenous treatment and differences in written examination grades between participants (foreign students) and non-participants (German students) already existing before treatment can be "differenced out" with the pooled DiD estimator. In our case, however, participation in the program is the result of individual decision making. Hence, differences between participants and non-participants are not exogenous from treatment, but arise from self-selection of participants into participation.[4] As

[4]Thus, in our case the pooled DiD approach is not suitable for eliminating unobserved differences between treatment and control group. However, as in other contexts educational programs are designed to allow for exogenous treatment and the use of the pooled DiD estimator (see ,e.g., Dynarski, 2003; Meghir and Palme, 2005; Hanushek and Wößmann, 2006), we include it for the sake of completeness.

mentioned before, it is likely that higher motivated students tend to use the WrC to a higher extend and have better written examination grades – irrespective of WrC participation. Thus, unobserved heterogeneity (due to omitted variables) and time-variant factors affecting treatment and control group differently still bias the estimations in the pooled DiD approach.

3.4. Fixed-Effects Difference-in-Differences estimator

The Fixed-Effects DiD estimator (FEDiD) uses the panel character of the data set. Whereas the pooled DiD carries out pooled OLS regressions (thus, ignoring that a single student is featured several times in the data set), the FEDiD estimator considers the change of each person in the course of time. It is especially designed to avoid biased estimators due to unobserved time-invariant heterogeneity (Wooldridge, 2010; Schlotter et al., 2011). In our case, one might assume the individuals' unobserved cognitive talent not fully captured by *AvGrade* to be correlated with both the decision to participate in the WrC and students' grades. The FEDiD estimator can be implemented by including a dummy for each person in the regression capturing the individual unobserved effects.

The special advantage of this approach is its ability to yield unbiased coefficients in the presence of unobserved fixed heterogeneity across units of observation. However, as the pooled DiD-estimator, it is sensitive to time-variant heterogeneity. Another downside is the need of strong requirements for the data (panel data with uniquely identified persons).

3.5. Implementation

The implementation of all estimation strategies requires a careful design of additional control variables. Firstly, the dummy variable *WrCP* is introduced to separate treatment- and control group. It equals 1 for all observations of participants, no matter at which time they actually have visited the WrC.

Secondly, a variable is needed that marks the introduction of the WrC and the time of visit of participants, respectively. We will refer to it as *IntWrC*. The time-related comparison between the two groups is being complicated by the fact that participants can be exposed to WrC support either in the summer term 2009 or winter term 2009/2010. For non-participants, a distinction between the two semesters cannot be made since time of participation cannot be observed. Therefore, for non-participants, *IntWrC* marks entries

after the introduction of the WrC, whereas for participants this dummy flags entries after actual voluntary visits of the WrC.[5]

Thirdly, an interaction variable – *WrCP*IntWrC* – is needed that contains the product of *WrCP* and *IntWrC*. Thus, only observations of Writing Center visitors after their WrC participation are flagged with this dummy.

The cross-sectional comparison is implemented by including the dummy variable *WrCP* and restricting the observations to the time after WrC introduction/participation (i.e. *IntWrC = 1*), resulting in the following model

$$\begin{aligned} WExG_i = \beta_0 + \beta_1 Female_i + \beta_2 German_i + \beta_3 Age_i + \beta_4 Bachelor_i \\ + \beta_5 LtStud_i + \beta_6 AvGrade_i + \beta_7 Thesis_i + \gamma WrCP_i + \varepsilon_i. \end{aligned} \quad (1)$$

The estimated coefficient of *WrCP* ($\hat{\gamma}$) displays the causal effect of WrC participation on the visiting students, in case the identification assumption is valid.

The before-after comparison is implemented by introducing *IntWrC* to the regression and restricting the estimation to participants (i.e. *WrCP = 1*):

$$\begin{aligned} WExG_{it} = \beta_0 + \beta_1 Female_i + \beta_2 German_i + \beta_3 Age_{it} + \beta_4 Bachelor_{it} \\ + \beta_5 LtStud_{it} + \beta_6 AvGrade_i + \beta_7 Thesis_{it} + \delta IntWrC_{it} + \varepsilon_{it}. \end{aligned} \quad (2)$$

Hence, only WrC visitors are considered for calculation and the estimated coefficient of *IntWrC* ($\hat{\delta}$) shows the treatment effect.

The implementation of the pooled DiD estimator requires all three additional variables in the regression, but no limitations of the sample:

$$\begin{aligned} WExG_{it} = \beta_0 + \beta_1 Female_i + \beta_2 German_i + \beta_3 Age_{it} + \beta_4 Bachelor_{it} \\ + \beta_5 LtStud_{it} + \beta_6 AvGrade_i + \beta_7 Thesis_{it} + \beta_8 WrCP_i \\ + \beta_9 IntWrC_{it} + \lambda(WrCP_i * IntWrC_{it}) + \varepsilon_{it}. \end{aligned} \quad (3)$$

Contrary to the cross-sectional comparison, the coefficient of *WrCP* does not indicate the treatment effect but controls for general heterogeneity between

[5] This has to be done to prevent WrC participants to be featured in the control group in advance to the treatment.

participants and non-participants, i.e., for differences in WExG between the two groups that are independent of WrC participation. The coefficient of *IntWrC* does not show the treatment effect either (as in the before-after comparison), but displays possible generic shifts in grades of non-participants. In this approach, the treatment effect is measured by the estimated coefficient of the interaction variable $WrCP{*}IntWrC$ ($\hat{\lambda}$).

Due to the fact that a person is covered several times in the data set, the standard errors are correlated with each other. To prevent biased standard errors caused by this serial correlation, we estimate cluster-robust standard errors on the individual level for the three approaches presented above.

The Fixed-Effects DiD estimator assumes the presence of unobserved time-invariant factors, e.g. the individual's cognitive talent, α_i which has constant effects for each student in all periods and is correlated with both the individual's grades and the decision to participate in the WrC. In order to prevent these individual time-constant factors to bias the estimations being hidden in the error term and n dummy variables – one for each individual in our data – are included in the regression:

$$\begin{aligned} WExG_{it} = {} & \beta_1 Age_{it} + \beta_2 Thesis_{it} + \beta_3 IntWrC_{it} + \beta_4 LtStud_{it} \\ & + \omega(WrCP_i{*}IntWrC_{it}) + \sum_{i=1}^{n} \alpha_i D_i + \varepsilon_{it}. \end{aligned} \tag{4}$$

Unfortunately, coefficients for all time-invariant variables cannot be estimated with this method, as these variables do not vary in the course of time for a student. However, since we are mainly interested in the causal effect of WrC participation, which varies over time, this problem is of minor relevance in our context. The crucial assumption of the Fixed-Effects DiD estimator is that the unobserved factors do not have a varying effect over time. In our example, it has to be assumed that the students' cognitive talent is fixed and does not vary over the course of study. If this assumption is violated, the FEDiD estimator is inconsistent. As in the pooled DiD-estimator, the treatment effect is represented by the estimated coefficient of the interaction variable $WrCP{*}IntWrC$ ($\hat{\omega}$). To take heteroscedasticity into account, robust standard errors are calculated for the FEDiD estimations.

4. Results

Table 2 shows the estimation results of the four models presented above. The cross-sectional comparison as well as the before-after comparison indicate no significant effect of WrC participation on student performance. In contrast, the DiD approach finds a significant positive coefficient for the treatment effect, indicating a significant worsening of written examination grades after WrC visitation. The coefficient of $WrCP$ in the DiD estimation is insignificant indicating that prior to the introduction of the WrC, participants do not differ from non-participants in their ability to write dissertations and theses. The coefficient of $IntWrC$ is significantly negative, showing that non-participants got better written examination grades after the introduction of the WrC, which could be explained by grade inflation or learning-effects. However, as mentioned before, all coefficient estimates of the cross-sectional comparison, the before-after comparison, and the DiD approach are suspected of being biased due to unobserved heterogeneity.
In contrast to the first three estimators, the Fixed-Effects DiD estimator does not pool the observations but compares the performance of a student over time. It is remarkable that unlike the other estimators, the FEDiD approach shows a negative, though small and insignificant, treatment effect. Although it cannot be concluded that students benefit from WrC visitation, this confutes the result of the PDiD approach, indicating a negative effect of WrC participation on students' grades. Nevertheless, the finding that WrC participants did not benefit from their visits and do not differ significantly from non-participants with respect to their grades is noteworthy.

We suspect the difference between the FEDiD estimator and the other estimators to be caused by unobserved differences between WrC visitors and non-visitors. Assume that the individuals differ in regard to their unobserved writing talent (WrT). This personal trait might feature a special case of "normal" cognitive talent and describes the individual's ability of self-organization and explaining complex and self-contained trains of thoughts.[6] If a student's writing talent is correlated with both his/her grades and his/her decision to participate in the WrC, the estimated coefficients in the cross-

[6]In contrast to cognitive talent and motivation, which might be proxied with *AvGrade*, we assume the writing talent to be mostly uncorrelated with *AvGrade*, since the exams composing the average grade are multiple-choice exams or feature (very) short written answers.

Table 2: Estimation Results – Effect on Written Assignments

	CSC Coeff/StdE	BAC Coeff/StdE	PDiD Coeff/StdE	FEDiD Coeff/StdE
Female	0.087	0.298**	0.107**	–
	(0.07)	(0.13)	(0.05)	
German	−0.232*	−0.250	−0.433***	–
	(0.13)	(0.34)	(0.10)	
Age	0.043**	0.022	0.031**	−0.065
	(0.02)	(0.04)	(0.01)	(0.07)
Bachelor	0.309***	0.049	0.304***	–
	(0.08)	(0.17)	(0.06)	
Long-term student	−0.163*	−0.138	−0.189**	−0.044
	(0.09)	(0.16)	(0.07)	(0.10)
Average Grade	0.700***	0.644***	0.723***	–
	(0.05)	(0.12)	(0.04)	
Thesis	−0.549***	−0.809***	−0.506***	−0.301***
	(0.06)	(0.15)	(0.06)	(0.06)
Writing Center Participant (WrCP)	0.100	–	−0.116	–
	(0.10)		(0.08)	
Introduction Writing Center (IntWrC)	–	0.068	−0.242***	−0.254***
		(0.13)	(0.05)	(0.07)
WrCP*IntWrC	–	–	0.230**	−0.076
			(0.11)	(0.12)
Constant	−0.328	0.436	0.316	–
	(0.46)	(1.06)	(0.37)	
Adjusted R^2	0.221	0.200	0.234	0.063
F Statistic	49.79***	12.13***	72.09***	34.49***
Observations	1274	229	2414	2414

*Notes: – Significant at: *** 1% level; ** 5% level; * 10% level. – Cluster-robust (CSC, BAC, and PDID) and robust (FEDiD) standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1.0 to 5.0 such that lower values indicate a higher score in the written assignment.*

sectional comparison, the before-after comparison and the pooled DiD approach are biased. This bias can be exemplified by the treatment effect of the cross-sectional comparison:

$$E(\widetilde{\beta}_{WrCP}) = \underset{(-)}{\beta_{WrCP}} + \underset{(-)}{\beta_{WrT}} * \underset{(-)}{\rho} \, . \tag{5}$$

The treatment effect $E(\widetilde{\beta}_{WrCP})$ consists of the true parameter of the treatment effect in the absence of unobserved heterogeneity, β_{WrCP}, and the product $\beta_{WrT} * \rho$, where β_{WrT} displays the effect of the unobserved writing talent on the student's grade, while $\rho = Corr(WrCP, WrT)$. As equation 5 shows, the treatment effect $E(\widetilde{\beta}_{WrCP})$ will be biased in the case of $(\beta_{WrT} * \rho) \neq 0$.

If we consider a high writing talent to improve grades, the coefficient β_{WrT} will be negative. In addition, we expect a negative correlation ρ between

$WrCP$ and WrT due to self-selection of students – the more students feel they lack the ability to write, the higher is their chance to visit the Writing Center. Therefore, although the real treatment effect β_{WrCP} is expected to be negative, the omitted variable WrT might lead to turning it positive. This may explain the unexpected results of all estimators not eliminating unobserved fixed effects. However, note that the FEDiD estimator can only eliminate the bias due to WrT if it is constant over time, i.e. if it equally affects the written examination grades in all semesters. For this reason, we have to consider WrT to be a cognitive talent that cannot be improved by training.

5. Conclusion

In the past, the vast majority of evaluation efforts of educational action relied on qualitative methods. While allowing persons concerned to express individual, content-related feedback, these evaluation strategies can be highly susceptible to imperceptible and unrelated factors possibly causing wrong self-assessment and self-perception. Thus, in many cases, they might be inappropriate for institutional assessment of real causal effects or cost-benefit-relationships.

For this reason, we apply a quantitative approach when evaluating the effectiveness of the introduction of a Writing Center at a large German university. Our aim is to isolate the real causal effect of WrC participation on students' academic success, which might be independent from the students' self-assessment. The Writing Center was introduced to improve the ability of students to write scientific works such as dissertations and theses and can be attended on a voluntary basis by students prior to or during the writing process. Based on administrative and survey data, we were able to construct extensive data allowing us to employ different quantitative methods for evaluation, each with its own advantages and disadvantages based on the circumstances and the availability of limited observational data.

In contrast to qualitative findings of WrC participation, which were based on student's course evaluations attesting it to be very effective, our quantitative analyses could not find a significant effect of WrC visitation on students' writing ability, as measured by their written examination grades. We attribute this difference to a wrong self-perception of students hindering them to give valid feedback for evaluating the effectiveness of educational programs. This finding underlines the need of educational institutions to also

rely on quantitative methods for evaluation purposes, as qualitative methods may be biased by unobserved and unconsidered factors.

There are several possible reasons for the suspected ineffectiveness of the facility. Students facing severe language- or self-organizational problems may not benefit significantly when attending Writing Center consultancy only once or twice. Furthermore, the observed self-selection phenomenon alludes to an unsuitable participation group, namely students who would benefit the most might simply not participate in the program because of lack of motivation. Moreover, the small sample size makes it impossible to evaluate the two treatments (i.e. workshop attendance and face-to-face support) separately or to evaluate the effect of WrC participation for different subgroups (such as males and females) although different treatment effects may exist. However, another reason could be the limited number of observations of participants generating imprecise estimations.

6. Lessons learned

Evaluation design for educational programs should rely both on quantitative methods, which allow the measurement of effectiveness, and qualitative feedback providing valuable insights into student attitudes and explanations for possible ineffectiveness. Ideally, in advance to the introduction of new services the requirements of the randomized controlles trial should be taken into account and the program participation should be made random in order to allow the comparison of mean values of participants and non-participants as a feasible method of evaluation. In case this is not possible, the implementation of the evaluation strategies used in this article has to be considered. Their valid utilization requires a careful research design in advance to the study as special attention has to be paid to the selection of an appropriate evaluation method by taking all specific characteristics of a situation into consideration. Researchers have to think *a priori* about which information is needed in order to carry out adequate estimations later on. In order to get more precise results, the conceptual design of the evaluation program should allow researchers to collect data for longer periods. However, this creates additional costs and can extend the period of time to collect data raising the necessity to balance program complexity and (financial) constraints.

In our evaluation, e.g., the lack of information concerning the number of visits of WrC participants prevents further analysis of time-related differences of program participation. Due to the short observation period covering the

existence of the WrC, it is further not possible to control for the fact that treatment quality might rise over time. Assuming that WrC employees need to familiarize with their tasks, we might pick up a weak treatment effect in the first semesters.

References

Bauer, T. K., Fertig, M., Schmidt, C. M., 2009. Empirische Wirtschaftsforschung: Eine Einführung. Springer, Berlin.

Cameron, A. C., Trivedi, P. K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge.

Cashin, W. E., 1995. Student ratings of teaching: The research revisited (IDEA Paper No. 32). Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University.

Centra, J. A., 2003. Will teachers receive higher student evaluations by giving higher grades and less course work? Research in Higher Education 44 (5), 495–518.

Davies, M., Hirschberg, J., Lye, J., Johnston, C., McDonald, I., 2007. Systematic influences on teaching evaluations: The case for caution. Australian Economic Papers 46 (1), 18–38.

Dynarski, S. M., 2003. Does aid matter? Measuring the effect of student aid on college attendance and completion. The American Economic Review 93 (1), 279–288.

Francis, C. A., 2011. Student course evaluations: Association with pre-course attitudes and comparison of business courses in social science and quantitative topics. North American Journal of Psychology 13 (1), 141–154.

Hanushek, E. A., Wößmann, L., 2006. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. The Economic Journal 116 (510), C63–C76.

Jaarsma, A. D. C., de Grave, W. S., Muijtjens, A. M. M., van Beukelen, P., 2008. Perceptions of learning as a function of seminar group factors. Medical Education 42 (12), 1178–1184.

Kluve, J., Card, D., Fertig, M., Góra, M., Jacobi, L., Jensen, P. L. R., Nima, L., Patacchini, E., Schaffner, S., Schmidt, C. M., van der Klaauw, B., Weber, A., 2007. Active labor market policies in Europe: Performance and perspectives. Springer, Berlin.

Lo, C. C., 2010. How student satisfaction affect perceived learning. Journal of the Scholarship of Teaching and Learning 10 (1), 47–54.

Marsh, H. W., 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research 11 (3), 253–388.

Marsh, H. W., Roche, L. A., 2000. Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias validity, or innocent bystanders? Journal of Educational Psychology 92 (1), 202–228.

McNatt, D. B., 2010. Negative reputation and biased student evaluations of teaching: Longitudial results from a naturally ocurring experiment. Academy of Management Learning & Education 9 (2), 224–242.

McPherson, M. A., Jewell R. Todd, Kim, M., 2009. What determines student evaluation scores? A random effects analysis of undergraduate economics classes. Eastern Economic Journal 35 (1), 37–51.

Meghir, C., Palme, M., 2005. Educational reform, ability, and family background. The American Economic Review 95 (1), 414–424.

Remedios, R., Lieberman, D. A., 2008. I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching. British Educational Research Journal 34 (1), 91–115.

Schlotter, M., Schwerdt, G., Wößmann, L., 2011. Econometric methods for causal evaluation of education policies and practices: A non-technical guide. Education Economics 19 (2), 109–137.

Walker, D. J., Palmer, E., 2009. The relationship between student understanding, satisfaction and performance in an australian engineering programme. Assessment & Evaluation in Higher Education 36 (2), 157–170.

Wooldridge, J. M., 2010. Econometric Analysis of Cross Section and Panel Data, 2nd Edition. MIT Press, Cambridge, Massachusetts.