When Excellence is not Excellent: The Impact of the Excellence Initiative on the Relative Productivity of German Universities

Cantner, Uwe and Grashof, Nils and Grebel, Thomas and Zhang, Xijie

Friedrich Schiller University of Jena and University of Southern Denmark, Odense, Friedrich Schiller University of Jena, Ilmenau University of Technology, Ilmenau University of Technology

29 July 2023

When Excellence is not Excellent: The Impact of the Excellence Initiative on the Relative Productivity of German Universities

Uwe Cantner[1], Nils Grashof*[2], Thomas Grebel[3], and Xijie Zhang[4]

[1]Friedrich Schiller University of Jena and University of Southern Denmark, Odense
[2]Friedrich Schiller University of Jena
[3,4]Ilmenau University of Technology

Abstract

Since the Bologna Process, universities have experienced a tremendous increase in competitive pressure. The Excellence Initiative is a path-breaking initiative to the historically-based egalitarian higher education system in Germany to boost universities' competitiveness and the competition among universities. Despite this systematic change, it remains rather unclear to what extent the Excellence Initiative influences the performance. In this paper, we investigate the effect of the Excellence Initiative on universities relative productivity in teaching and research and hence on the divergence or convergence of universities in these performance dimensions. Based on a unique dataset that combines publication and detailed university-level data, we apply a two-step approach by calculating relative productivity with a non-parametric procedure in the first step and using a difference-in-difference approach for estimating treatment effects of the Excellence Initiative in the second step. Overall, we note only a few moments when universities funded by the Excellence Initiative seem to excel and make progress compared to non-funded universities. In research, only some of the Excellence-funded universities, particularly the winners of the Graduate Schools and Clusters of Excellence funding line, manage to improve significantly, even if only for a few years. In teaching, we found no significant average treatment effect of the Excellence Initiative, but a slightly significant time-specific decline in relative teaching productivity two years after funding.

Keywords: Excellence Initiative, Universities, Relative Productivity, DEA, Difference-in-difference, Funding

JEL: C22, I20, I23, I25, I28

Acknowledgement: The authors acknowledge the received access to the publication data through the project "Kompetenzzentrum Bibliometrie" (grant number: 16WIK2101C) funded by the German Federal Ministry for Education and Research (BMBF). Furthermore, the authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG) [grantnumber: 447967785].

*Corresponding author: Carl-Zeiss-Straße 3, 07743 Jena, Tel.: +49 3641 943 203, E-mail: nils.grashof@uni-jena.de

1 Introduction

Like in many countries in Europe, German universities had traditionally been under strong influence of state authority and relied majorly on non-performance-based public funding from the state (Krücken and Meier, 2006; Frølich et al., 2010; Musselin, 2018). Little attention was paid to the performance and the competition between universities. Since the Bologna Process, starting in 1999, the autonomy of universities has increased. More and more performance-based funding programs have been launched to impose yardstick competition[1] between universities (Olivares and Wetzel, 2014; Lehmann et al., 2018; Agasisti and Johnes, 2009; Agasisti, 2009; Aghion et al., 2010), reinforced by a steady decline in basic funding budgets (Krücken, 2021; Wiener et al., 2020).[2]

One of the biggest and also most controversial competitive grant schemes in Germany is the *Excellence Initiative* (in German: *Exzellenzinitiative*). With three lines of funding, i.e. Future Concepts (*Zukunftskonzepte*), Clusters of Excellence (*Exzellenzkluster*), and Graduate Schools (*Graduiertenschule*), it aims to boost German universities to the world-class level in research (Buenstorf and Koenig, 2020; Mergele and Winkelmayer, 2021; Gawellek and Sunder, 2016). The Excellence Initiative has so far been launched in three rounds (2006, 2007 and 2012).[3] Although these programs focus on promoting research performance, universities are expected to achieve productivity gains in all of their core activities. The considerable funds allocated to the universities are intended not only as an investment in infrastructure, but first and foremost for the development of human capital, that is, excellent researchers.

The limited number of empirical studies so far investigate the effects of the German Excellence Initiative only with respect to the most prominent indicators such as publications (Menter et al., 2018; Möller et al., 2016), third-party funding (Buenstorf and Koenig, 2020; Mergele and Winkelmayer, 2021) and productivity/technical efficiency in general (Gawellek and Sunder, 2016; Wohlrabe et al., 2019). To what extent the Excellence Initiative influences all essential tasks of universities, however, remains rather unclear. To the best of our knowledge, there is no systematic analysis of all three funding rounds of the Excellence Initiative to date. This research gap we approach with a comprehensive view of two main "missions" of universities: *teaching* and *research*. We will scrutinize empirically to what extent the German Excellence Initiative has affected the relative productivity in teaching and in research of the funded universities and how this has changed the structure in the university sector.

Thus, we not only contribute to the still ongoing discussion about the role of competitive funding in the higher education system (inter alia, Auranen and Nieminen, 2010; Gawellek and Sunder, 2016; Menter et al., 2018), but also provide empirical insights for more evidence-based policies in the tertiary education and research system. Based on a unique university-level data set collected by the Federal Statistical Office of Germany (Destatis), which we combine with bibliometric information from Scopus, we use a non-parametric approach to calculate productivity scores to measure universities' relative productivity in teaching and in research, in the first step. In the second step, we estimate the causal impact of the Excellence Initiative, launched in 2006, 2007, and 2012, on universities' relative productivities, applying a difference-in-differences (DID) approach.

[1] Yardstick competition creates an "artificial" competitive environment, by rewarding actors not on their own performance, but on their relative performance in order to regulate franchised monopoly (Shleifer, 1985).

[2] At the same time, also the allocation of basic funding has transformed, now (partly) being distributed based on performance indicators (Auranen and Nieminen, 2010; Wiener et al., 2020).

[3] Since 2019, its successor scheme was called Excellence Strategy (*Exzellenzstrategie*). For the further analyses in this paper, however, we do not consider this successor initiative, as it introduces some changes and is therefore not really comparable with the previous Excellence Initiative. Moreover, we are also restricted by our data in this context.

The remainder of the article is structured as follows: In Section 2, we briefly introduce the relevant literature about competition between universities, the effects of competitive funding in academia, and the Excellence Initiative in Germany. In Section 3, we derive four main hypotheses about the impact of the Excellence Initiative. Our measurement approach of relative productivity is introduced in Section 4, while our data sources and first results on relative productivity are described in Section 5. Thereafter, we present our matching and difference-in-differences approach as well as the corresponding results about the effect of the Excellence Initiative on relative productivity in Section 6. The article ends with concluding remarks (Section 7), including a short discussion of potential policy implications, limitations as well as a research outlook.

2 A Short History of German University Funding

2.1 Competition and Funding Program in Higher Education

As in many other European countries, Germany for a long time pursued an egalitarian higher education policy in which it was primarily the federal states that financed higher education. (Buenstorf and Koenig, 2020; Mergele and Winkelmayer, 2021; Frølich et al., 2010; Musselin, 2018; Krücken and Meier, 2006). Tight budgets and the fear of losing ground in international research competition gave rise to a fundamental reform: Give more autonomy to universites and expose them to (more) competition (Aghion et al., 2010).

As before, the missions of universities have not been changed. Providing excellent education at the highest level (i.e. teaching) and perform excellent research to ensure not only international competitiveness, but also a country's innovativeness, its technical progress and economic growth in the long term (Gu, 2012; De Fraja and Iossa, 2002; Epple et al., 2006; Del Rey, 2001; Conard and Conard, 2000; Krücken et al., 2009; Lehmann et al., 2018).[4] The transformation process towards a more competitive university environment, however, comes with major challenges. The supply and demand for research and education does not follow the concept of a standard textbook competitive market. Not only is there a principle-agent dilemma on the side of government funding (Van der Meulen, 1998), as the "production process" of universities is complex; students cannot ex ante evaluate the quality of education correctly; but also there is no unique value for the quality of research output, let alone compared across disciplines (Dasgupta and David, 1994; Merton, 1973). Students instead have to rely on quality signals, such as rankings that proxy the prestige of a university (Franck and Schönfelder, 2000; Winston, 1999); researchers, for instance, have to participate in the collegiate reputation-based reward system which itself is prone to market failure (Dasgupta and David, 1994; Merton, 1973). Furthermore, for most of the goods or services a university produces, there is no market price to coordinate the players. Viewed in this light, the introduction of any competition-enhancing regulation, any incentive program, will lead to an incentive structure in a highly imperfect market whose transmission mechanism is possibly ambiguous. More money will create more output, but whether it will increase productivity is rather questionable.

Hence, competition between universities is not about maximizing profits, but rather about maximizing prestige/reputation – an intangible asset, a latent variable difficult to measure. As universities do not compete directly with each other, for deciding on which universities to fund (for excellence-enhancing), as much as it is for measuring the impacts of these excellence-promoting programs on

[4]Nevertheless, it has been emphasized that not each "mission" equally contributes to the overall reputation of a university, research appears to be the central dimension in this context instead (Krücken, 2021).

university performance, the evaluation of what is excellent has to be based on indicators associated with such a latent variable. In addition, the performance measure has to be compared to some benchmark, which can only be gauged relative to a best-practice university (or a collection of best-practice universities) in the case of a "none-price competition". Therefore, this type of competition can be called "yardstick competition" as coined by Shleifer (1985) and discussed in New Public Management (inter alia, Kettl, 2000; Pollitt, 1990), in which performance-based funding systems are essential.

Table 1
Studies of Competitive Research Funding

Studies	Grants	Countries
Abbott and Doucouliagos (2003)	Research Quantum	Australia
Butler (2003)	Research Quantum	Australia
Carayol and Matt (2006)	All public and private funding 1993-2000	France
Wiener et al. (2020)	All third-party funding	Austria
Carayol et al. (2018)	Funds from French National Agency (ANR)	France
Corsini (2022)	Initiative of Excellence (IDEX)	France
Arora et al. (2000)	Funds from National Research Council (CNR)	Italy
Cherchye and Abeele (2005)	79 research programs 1998-2000	Netherlands
Gush et al. (2018)	Marsden Fund	New Zealand
Bolli and Somogyi (2011)	Swiss National Science Foundation	Switzerland
Smith et al. (2011)	Research Excellence Framework	UK
Arora and Gambardella (2005)	National Science Foundation (NSF) grants	US
Jacob and Lefgren (2011)	National Institutes of Health (NIH) grants	US
Aghion et al. (2010)	Grants from NSF, NIH, NASA	US
Geuna and Martin (2003)	Research grants in 3 countries	UK, Netherlands, China
Bolli et al. (2016)	International public funds, private funds 1994-2003	Eight European countries

In general, competitive (research) funding has been commonly employed as instrument for yardstick competition in academia all over the world (see Table 1). The underlying idea behind all of these competitive funding schemes is that if funding is given to the best performers, better results are more likely to be achieved, and additionally, all the funding applicants will have an incentive to perform better in order to win these funds (Auranen and Nieminen, 2010; Wiener et al., 2020). Studies on this topic nonetheless find evidences for insignificant results or unintended negative consequences (Bolli and Somogyi, 2011; Bolli et al., 2016; Butler, 2003; Carayol and Matt, 2006; Strehl et al., 2007; Wiener et al., 2020), though in most studies, a positive relationship between competitive funding and productivity performance of funded individuals and institutes is detected (Butler, 2003; Aghion et al., 2010; Bolli and Somogyi, 2011; Cherchye and Abeele, 2005; Corsini, 2022; Jacob and Lefgren, 2011; Gush et al., 2018; Carayol et al., 2018; Arora and Gambardella, 2005).

2.2 The Excellence Initiative

The Excellence Initiative (ExIni) aims at making German universities into the international elite universities by promoting cutting-edge research and improving the overall performance quality (Frietsch et al., 2017; Mergele and Winkelmayer, 2021); the state secretary of the Federal Ministry of Education and Research in Germany put it: "If you want to compete in the research world, you have to have some top universities that play in the first league."(Menter et al., 2018; Morgan, 2016). To achieve this objective, the Excellence Initiative was equipped with 4.6 billion euros of additional funding, corresponding to about 4% of the total research funds in Germany (Buenstorf and Koenig, 2020; Imboden et al., 2016; Mergele and Winkelmayer, 2021). The Excellence Initiative breaks the tradition of German higher education policy in two important ways (Menter et al., 2018). Firstly, it infuses the higher education sector in Germany with competitive spirits and nurture world-class competitive universities, thereby significantly contributing to the transformation of the long-prevailing egalitarian higher education system in Germany (Buenstorf and Koenig, 2020; Mergele and Winkelmayer, 2021). Secondly, under the Excellence Initiative, the federal government has granted federal funds directly to universities that are otherwise under the control of the 16 states (Menter et al., 2018).

The analysis of the impact of Excellence Initiative contributes to the current literature of competitive research funding, because it consists of three funding rounds (2006, 2007 and 2012) and three different funding lines (see Appendix A for an overview of the winning universities of three funding lines in three rounds), which is a setting barely analyzed in the existing literature. These three funding lines encompass Graduate Schools (*Graduiertenschule*), Clusters of Excellence (*Exzellenzcluster*) and Future Concepts (*Zukunftskonzepte*). While in all three funding lines universities benefit in monetary terms, in the case of Future Concepts the winning institutions also likely experience particular reputation gains, since they are often characterized as Germany's "elite universities" (Buenstorf and Koenig, 2020; Gawellek and Sunder, 2016; Mergele and Winkelmayer, 2021). In addition, the Future Concepts provides university leaders with a great deal of flexibility in developing and implementing self-defined institutional strategies to improve their competitiveness. While in the case of Graduate Schools and Clusters of Excellence, the focus is relatively narrow and clearly defined.[5] The winners of the Future Concepts[6] in contrast have considerably more freedom in their investment decisions. Therefore, a large share of funding from the Future Concepts is distributed after internal contests within the corresponding universities (Buenstorf and Koenig, 2020; Imboden et al., 2016). Additionally, it is worth noting that only winners of the Clusters of Excellence and Graduate Schools are qualified to apply for the funding of Future Concepts (Buenstorf and Koenig, 2020).

Despite this radical and substantial (at least in monetary terms) policy intervention, the empirical evidence for the impact on the performance of universities is rather limited. Moreover, the few empirical studies that actually examine the impact of the Excellence Initiative come to rather mixed results, which can be explained with the underlying focus on important, but partial aspects. For example, Horstschräer (2012) and Fischer et al. (2017) found that the winners of the Excellence Initiative can recruit better students and improve their image of education quality among students, while in terms of additional third-party research funding, Buenstorf and Koenig (2020) as well as Mergele and Winkelmayer (2021) do not find any evidences of a driving role of the Excellence Initiative. Instead, it is shown that after winning the Future Concepts, universities gained significantly less direct funding from the federal government (Buenstorf and Koenig, 2020). Moreover, Möller et al. (2016) and Menter et al. (2018) show that Clusters of Excellence and Excellence Initiative in general only leads to a modest increase in terms of research quality. The actual productivity implications of Excellence Initiative have so far, if at all, only been investigated descriptively (e.g. Gawellek and Sunder, 2016;

[5] Graduate Schools targets to build new research supervision and funding structures on university level for scientific offsprings and Clusters of Excellence focuses on establishing new forms of cross-sector and cross-institute research collaborations (Imboden et al., 2016).

[6] Future Concepts facilitates the universities to build their own specific strategies on research and teaching (Imboden et al., 2016).

Wohlrabe et al., 2019). In this context, Gawellek and Sunder (2016) report no positive correlation between Excellence Initiative and university productivity. In sum, given the scarcity of the available literature and the mixed empirical results, the exact implications remain so far rather unclear. Hence, particularly for policy-makers it is still difficult to learn from the previous experiences with Excellence Initiative, which would be essential for more evidence-based higher education policies. Against this background, we follow the recent call by Menter et al. (2018) and examine Excellence Initiative with its three funding periods and three funding lines from a more holistic view. In this context, we also distinguish between the effects on productivity of the two main "missions" of universities, namely teaching and research (Gautier and Wauthy, 2007).

3 The Impact of Excellence Initiative

Through the Excellence Initiative, funded universities receive additional financial support to hire more staff, invest more in infrastructure, and purchase more materials and supplies. The consequences for the outcomes produced may differ between research and teaching and can be expressed in terms of their respective productivity measures. These measures relate the outcomes in an activity to the inputs used therein. We distinguish between research productivity and teaching productivity. More precisely, the productivity measure used is measured with respect to a best-practice benchmark and hence in a relative way. Therefore, the indicators used are the relative research productivity and the relative teaching productivity. These two dimensions will now be considered one after the other.

The Excellence Initiative funding is primarily meant to foster research activities. It allows funded universities to hire more (excellent) researchers, purchase further necessary research equipment and infrastructure, and initiate more research oriented activities such as collaboration with other researchers and universities. It is reasonable to assume that the quality of these additional resources is higher than that of the resources the university maintains without funding. Moreover, these additional high quality resources can lead to a productivity increase of the already invested infrastructure, such as administrative infrastructure, laboratories and equipment (Bolli and Somogyi, 2011; Robst, 2001; Bonaccorsi et al., 2006; Brinkman and Leslie, 1986; De Groot et al., 1991); they can have a boosting effect in that the existing scientific staff continues to improve. In addition, funding from the Excellence Initiative can have a sorting effect, according to which better-performing scientists become interested in the funded university and apply for open positions there; funded universities also hire academic staff according to their previous achievement (Bolli et al., 2016).

Hence, the strategy of Excellence funded universities is assumed to secure their excellence status by recruiting even more excellent researchers and more sophisticated research infrastructures on the one hand and by reallocating existing resources towards higher quality research, on the other. Being successful herein would provide for sustainable success in the yardstick competition set off by the Excellence Initiative (Agasisti, 2009; Aghion et al., 2010).

Overall, the Excellence Initiative funding should therefore give the receivers an edge in research outcomes with respect to those without such funding. An indicator of relative research productivity can be used to identify that. In order to account for differences here, an indicator of research productivity is used that expresses the distance from the best-practice research frontier. Given this setting, the supported universities compared to not supported ones should improve in research productivity, hence in their relative research productivity. This leads us to the following hypothesis:

Hypothesis 1 The Excellence Initiative improves the relative research productivity of ExIni-funded universities compared to non-funded universities.

The Excellence Initiative grants are primarily intended to boost research-oriented activities. With respect to teaching, the additional staff hired to improve research output first of all represents input without an immediate return on teaching output: neither extra teaching activities will be set up nor additional students will be expected to enroll. The latter seems plausible, because German students value factors such as offers of study subjects, cost of living, distance from home, study infrastructure, student services or student-supervisor ratio (Obermeit, 2012; Willich et al., 2011; Hachmeister et al., 2007) higher than a university's research reputation when making their decision which university to attend. Consequently, the level of teaching and the number of graduated students should remain unaffected by additional excellence funding.

Putting these arguments together, Excellence Initiative funding allows for hiring more researchers without affecting teaching output. The Excellence-funded universities will thus be at a disadvantage in teaching productivity compared to non-funded universities; the relative teaching productivity of funded universities should worsen, which leades to the following hypothesis:

Hypothesis 2 The Excellence Initiative reduces the relative teaching productivity of ExIni-funded universities compared to non-funded universities.

With respect to the impact of Excellence Initiative funding on relative research productivity, we need to distinguish between the three funding lines: Graduate Schools, Clusters of Excellence, and Future Concepts (see subsection 2.2). Graduate Schools and Clusters of Excellence are funding lines designed to stimulate research (Imboden et al., 2016). In contrast, the Future Concept funding line has a more broadly defined purpose meant to help funded universities develop new strategies, structures, and formats to promote research, develop teaching, or improve further university outcomes (e.g. transfer).

As to relative research productivity, the funds allocated to Graduate Schools or Clusters of Excellence should therefore translate directly into an improvement of funded universities. Hence, we suggest the following hypothesis:

Hypothesis 3 The funding lines *Graduate Schools* and/or *Clusters of Excellence* primarily drive the positive effect of the Excellence Initiative on funded universities' relative research productivity.

Conversely, the funding of Future Concepts should not have an immediate positive impact on the relative productivity of funded universities in any outcome dimension. (Buenstorf and Koenig, 2020; Imboden et al., 2016). Consequently, we formulate the following hypothesis:

Hypothesis 4 The funding line *Future Concepts* has no (immediate) effect on the ExIni-funded universities' relative productivity, neither on research, nor on teaching.

4 The University as Production System

4.1 The Production Process

As pointed out in the previous section, to measure university performance, we derive performance measures conceiving universities as production systems. For simplicity, we focus only on the two main missions of universities (e.g. Gautier and Wauthy, 2007), as mentioned in Section (2): "teaching" and "research". With regard to inputs and outputs of the two missions, we largely follow the literature. In a first setup, Setup 1, we focus on financial resources and human capital as our main inputs, in line with the studies by Gawellek and Sunder (2016) and Olivares and Wetzel (2014). In this setup, the inputs for both outputs,"teaching" and "research", are:

1. The number of academic staff[7] (Gralka et al., 2019; Gralka, 2018; Fandel, 2007; Başkaya and Klumpp, 2014) and

2. General expenditures (Gralka et al., 2019; Gralka, 2018; Fandel, 2007; Başkaya and Klumpp, 2014; Gawellek and Sunder, 2016; Kempkes and Pohl, 2008).

In order not to double-count the personnel input, we exclude the expenditures for personnel from the general expenditures, so that our expenditures only include capital expenditures (e.g. acquisition of land and buildings) as well as current material expenses (e.g. rents and leases for land, buildings and energy).

Since third-party funding plays a more and more important role in the financial endowment of universities (Auranen and Nieminen, 2010; Wiener et al., 2020), we consider third-party funding as a part of financial input in Setup 2,[8] additionally to human capital. Setup 2 therefore uses the following augmented measure for general expenditures:

3. General expenditures including third-party funding.

These two setups are necessary, because of data constraints that does not allow us to decompose inputs any further. Therefore, we create a corridor of results with a lower and an upper bound. Setup 1 tends to overestimate the relative productivity of universities by not including the possible received non-personnel external funding as part of the inputs. Setup 2 heals this caveat, while it tends to underestimate relative productivity by double counting the personnel expenditure, as we cannot decompose the external funding into personnel and non-personnel funding.

4.2 Measuring Relative Productivity

Having defined inputs and outputs, it remains to choose an adequate method to measure relative productivity performance. Most common in the literature is Stochastic Frontier Analysis (SFA) as a parametric approach, and Data Envelopment Analysis (DEA) as its non-parametric counterpart. In order to put the least constraints on universities production systems and to allow for structural differences, we choose DEA, which is in line with the majority of previous studies (inter alia, Liu et al., 2013; Quiroga-Martínez et al., 2018).

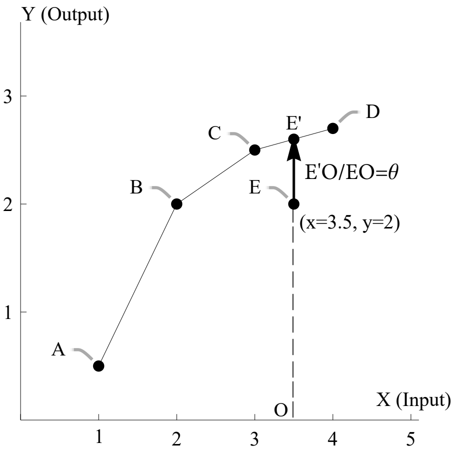
The DEA approach by Banker et al. (1984) (henceforth: BCC) assumes variable returns to scale so that only universities of "similar" size will be compared to each other. Furthermore, the output-oriented version of their model takes inputs as given (and as a measure for size) and calculates the factor by which a university's output must be multiplied in order to meet its benchmark productivity level. The BCC model represents a simple linear program (Figure 1):

The graph in the left panel of Figure 1 illustrates the basic concept of the DEA calculation. Suppose we have n universities ($i = 1, \ldots, 5$) that use $j = 1$ input ($x_{ji} = x_i$) to produce $r = 1$ output ($y_{ri} = y_i$). The size or scale of an university here is measured by its level of the input x. Within

[7]Due to data protection reasons, we could not get the separate number of professors, but the number of professors (with permanent and/or temporary positions), lecturers and assistants (with a permanent position).

[8]It is still rather controversial whether to use third-party funding as an input or output, although we are aware that in recent years it has been mainly used as an output (Barra and Zotti, 2016; Gralka et al., 2019). However, we follow the argumentation by Johnes and Johnes (1993) and argue that third-party funds are not only spent on research, but also on other facilities, which are inputs for production.

Fig. 1. Variable returns to scale model (BBC) by Banker et al. (1984)



(a) Graphical illustration

$$
\begin{aligned}
&\max \theta \\
&s.t. \sum_{i=1}^{n} x_{ji}\lambda_i \leq x_{jk} \qquad (\forall j = 1, \cdots, m) \\
&\sum_{i=1}^{n} y_{ri}\lambda_i \geq \theta y_{rk} \qquad (\forall r = 1, \cdots, s) \\
&\lambda_i \geq 0, \quad \sum_{i=1}^{n} \lambda_i = 1 \qquad \theta \geq 1
\end{aligned}
$$

(b) Linear maximization program

their respective size classes, universities A, B, C, and D achieve the highest level of productivity, given input x and output y (see Figure 1a). The productivity frontier is depicted by a convex line connecting all universities with respective highest productivity levels. University E does not belong to this group. E shows a productivity that is less than best practice in its size class (which is between university C with $x_C = 3$ and university D with $x_D = 4$). A (convex) combination of C and D, labelled E', would perform better. In order to reach its benchmark level, E with input $x = 3.5$ would need to increase its output to E'. Hence, the relative productivity of university E is measured against the convex combination E' by universities C and D.

The corresponding linear maximization problem is formulated to the right of the graph (see Figure 1b), with $i = 1, \ldots, n$ universities, $j = 1, \ldots, m$ inputs, and $r = 1, \ldots, s$ outputs. This problem is set up to determine the relative productivity θ of university k by comparing k with all universities $i = 1, \ldots, n$ (including k itself). Doing so, the algorithm maximizes k's θ given two sets of constraints.

The first set of constraints refers to inputs and forces the algorithm to compare university k only with input combinations of (at most) the same size and of being constructed by convex combinations using the inputs of all universities; this is governed by the λ-weights. Their values are non-negative and their sum is forced to be exactly 1; hence, the combinations allowed for are convex combinations. This guarantees that the best-practice point of comparison for k is of the (at most) same size as k. With k=E, in Figure 1a this is visualized by the size of E being the same as of E' which is constructed by a convex combination of the inputs of C and D.

The second set of constraints refers to the output side. They force the algorithm to compare the output of k with convex output combinations of all universities and to determine the distance via θ. For k=E, (Figure 1) this combination is again E', being constructed by the outputs of C and D. The vertical distance of E' to E, expressed as a ratio of the respective outputs ($E'O/EO$), thus determines the relative productivity performance θ of university E.

If $\theta = 1$, the university is considered best practice. If $\theta > 1$, the university has a relative productivity below best practice and needs to increase its output(s) by a factor of θ. θ is therefore a score indicating a university's relative productivity against the best practice under consideration.The excess of θ over 1 is a measure of the inefficiency in production.

5 Productivity of German Public Universities

5.1 Data Description

One source of our data stems from the Federal Statistical Office of Germany (Destatis). It provides detailed (confidential) information about all public universities in Germany. It contains non-monetary information such as the number of students, the number of graduates and the number of academic staff (including the number of doctoral students, the number of professors and other full-time researchers). Moreover, it also delivers monetary information such as general expenditure, facility and personnel expenditures, as well as third-party funding with the specification of its sources. For measuring productivity output, we enrich the database with publication data, which we retrieve from Scopus. After merging these data sets, our final database comprises a time span of 20 years from 2000 to 2019. For our later analysis, this relatively long time period will allow us to capture the effects before the first excellence initiative in 2006 and the last funding treatment in 2012. In total we identify, 80 public universities in our final database.

Table 2

Descriptive statistics – University inputs and outputs

Input/Output	Obs.	Mean	Min.	Max.	Std. Dev.	Skewness	Kurtosis
Academic staff	1,593	587.0	16.	3,316.0	461.9	1.9	9.3
General expenditure	1,594	170,089.0	1,459.1	916,315.0	189,475.0	1.5	4.6
plus Third-party funding	1,594	230,817.0	1,608.0	1,206,059	242,868.0	1.5	4.7
Graduates	1,593	2,683.4	23.	10,591.0	1,905.4	1.1	4.2
Publications	1,592	1,322.7	1.	8,431.0	1,250.0	1.6	6.5

Note: (Input) academic staff: personnel engaged in teaching and research; (input) general expenditure: basic state funding; (input) plus third-party funding: general expenditure plus third-party funding; (output) graduates: number of graduates. All input variables are from Destatis. (Output) publications: number of publications (Scopus).

As many studies in literature such as Fandel (2007); Başkaya and Klumpp (2014); Kempkes and Pohl (2008), we use the number of graduates to measure university output with respect to the mission "teaching", and research output we measure by the number of publications as in Gralka et al. (2019); Lehmann et al. (2018). Because inputs do not immediately turn into outputs, since it takes time for students to graduate and for researchers to publish their work, we forward outputs by three years.[9] Three years should be an adequate temporal adjustment, as the standard period of study for a bachelor's degree in Germany is three years (Ash, 2006). Moreover, with three years, we are also optimistic to capture the relatively large variation in the publication time across different disciplines (Björk and Solomon, 2013). Besides forwarding our outputs, we additionally use a Gaussian filter of two in order to smooth out annual jumps in the data and avoid distortions.[10]

Table 2 collects descriptive statistics about the data. The data are an unbalanced panel with few observations missing. For 79 out of 80 universities, all information is available.[11] Aside from the first

[9]As a further robustness check, we also forward our outputs by only two years. The corresponding results remain robust and can be provided upon request.

[10]As a further robustness check, we also use a Gaussian filter of three. The corresponding results remain thereby similar and can be provided upon request.

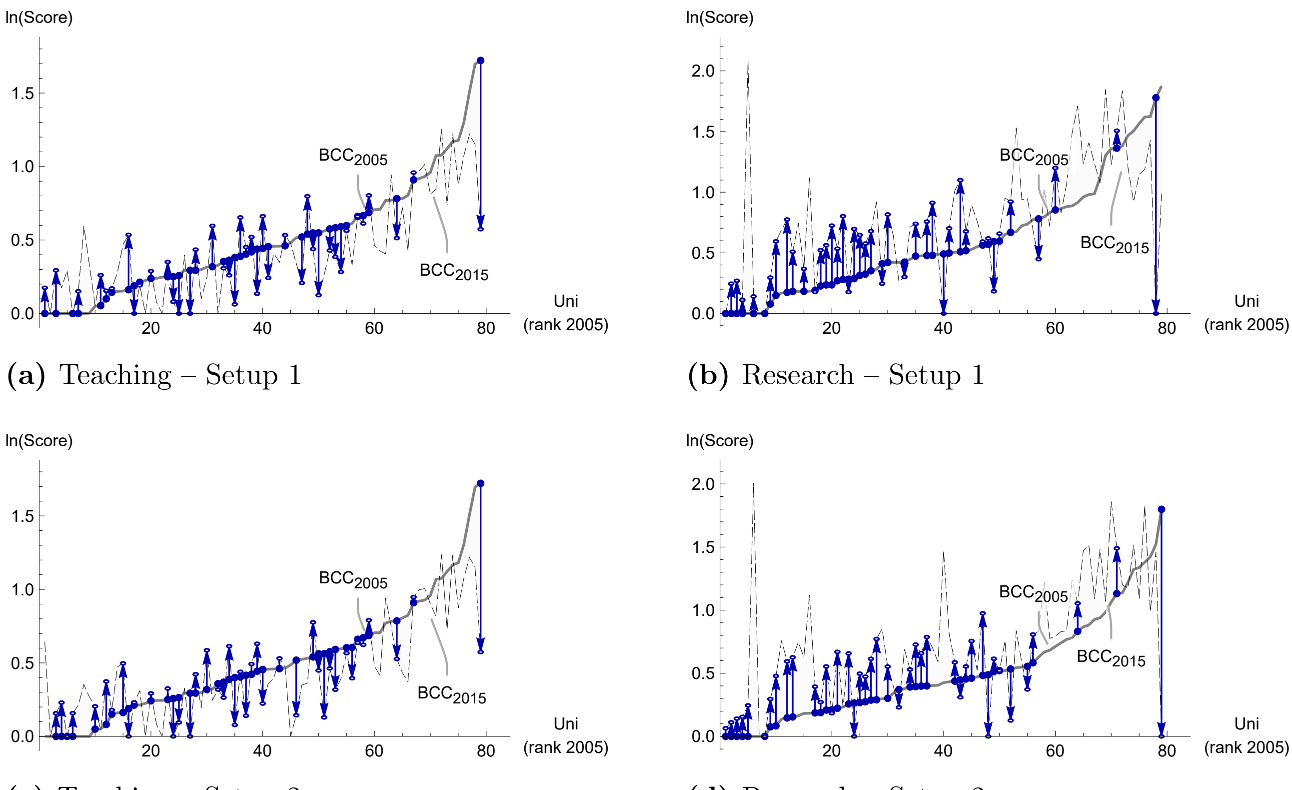
[11]From 2000 to 2007 there are 79 universities. Hence, the information for 80 universities is from 2008 to 2019.

two moments, we also include the skewness and the kurtosis of the distributions of respective variables to show that they all are right skewed and fat tailed.

5.2 Salter Curves – Changes in Relative Productivity Performance

Performing the BCC calculations, as laid out in subsection 4.2, we can now present the relative productivity scores of the 80 universities[12] and their dynamics when comparing the scores in 2005 and 2015. To recall, as we forwarded output by three years, the solid lines in the four panels in Figure 2 depict the logged relative productivity scores with inputs in 2002 and outputs in 2005; we label that line BCC_{2005}. In contrast, the dashed lines manifest the logged relative productivity scores calculated with inputs in 2012 and outputs in 2015, labelled as BCC_{2015}. The solid lines represent the university performance before the Excellence Initiative (2006, 2007, 2012) and the dashed lines after the final round of Excellence Initiative. The so-called Salter curves rank the universities in an ascending order according to the relative productivity scores of 2005. Best-practice universities with a relative productivity score $\theta = 1$ ($\ln 1 = 0$) can be found close to the origin, universities with lower relative productivity further to the right. The two left panels in the figure refer to teaching and the right for research. For both we run the two different setups, namely Setup 1 and Setup 2, for determining relative productivity.

Fig. 2. Relative productivity scores: Teaching and research



(a) Teaching – Setup 1

(b) Research – Setup 1

(c) Teaching – Setup 2

(d) Research – Setup 2

Note: Solid curves indicate relative productivity scores by universities ranked in ascending order as identified in 2005; scores are logged. BCC_{2005} labels the Salter curve calculated with inputs in 2002 and outputs in 2005, BCC_{2015} (dashed line) is calculated with inputs in 2012 and outputs in 2015, but ranked according to the order of the BCC_{2005} sequence. Thus, it is possible to see gains/losses of individual universities in their relative productivity scores. The blue points indicate universities that were awarded any of the three excellence funding programs in three rounds.

[12]From 2000 to 2007 there are 79 universities.

Sub-figure 2a shows the Salter curve constructed with relative productivity scores in teaching calculated within Setup 1. Comparing the relative productivity scores of the two years under consideration (solid and dashed lines), we observe that teaching performance has become more homogeneous from 2005 to 2015, since the dashed 2015 line mostly stays below the Salter curve in 2005. The blue points and the attached arrows mark all universities that have ever won any of the funding lines (Graduate Schools, Clusters of Excellence and Future Concepts in 2006, 2007 and 2012). As the arrows indicate, most of them either could sustain their relative productivity score or could improve it in the considered time span. When using Setup 2 instead of Setup 1, as shown in Sub-figure 2c, the general tendency of improving relative productivity scores in teaching resembles the same pattern.

The results for universities' relative research productivity, depicted in the two right hand panels with Salter curves for 2005 (solid line) and 2015 (dashed line), look different compared to the relative teaching productivity. In Sub-figure 2b, based on Setup 1, the 2015 line mostly runs above the 2005 Salter curve. In other words, the spread of the relative performance of universities has increased, i.e. the relative research performance among universities has become more heterogeneous between the two years under consideration. With regard to universities with excellence funding, most of the blue arrow point upwards, indicating a lower relative productivity; only a few of them extend downwards, i.e. they show an improvement. This pattern also shows up when Setup 2 is applied, as depicted in Sub-figure 2d.

To take a closer look at this analysis, we compare the average changes of relative productivity scores (θ_i) between ExIni-funded and not funded universities, in the following. For the full time span of the analysis 2000 to 2016,[13] three sub-periods are distinguished, 2000-05, 2006-11, and 2012-16. Whereas the first sub-period covers the time span before the launch of the Excellence Initiative, the second time span mark the beginning/ending of the first and the second round of Excellence Initiative (2006 and 2007). The third period starts with the launch of the third round of the Excellence Initiative in 2012 and ends in 2016. The calculation of the average change in relative productivity is:

$$\overline{\Delta\theta}_{J,\alpha|\omega} = \frac{1}{n_j}\sum_{i\in J}^{n_j}\left(\frac{\theta_{i,\omega}}{\theta_{i,\alpha}} - 1\right) \tag{1}$$

with $(\alpha, \omega) = \{(2000, 2005), (2006, 2011), (2012, 2016)\}$ as the start and end of the sub-period and the actual change in relative productivity $\theta_{i,\omega}/\theta_{i,\alpha}$ of university i belonging to university group J=$\{$ExIni funded $= E$, not-funded $= \bar{E}\}$ with n_j numbers of universities in group J.

Table 3 reports the corresponding average changes for teaching and research in Setup 1 and in Setup 2. For each sub-period, we compute the percentage change in relative productivity between the final and the first year of the respective sub-period. A negative change indicates an improvement of relative productivity and hence a convergence towards best-practice; contrariwise, a positive change stands for a worsening of relative productivity to be interpreted as a divergence from best-practice. The changes are computed as in Equation (1). We perform two sets of mean difference tests. Firstly, between the average changes of ExiIni-funded ($= E$) and non-funded universities ($= \bar{E}$), i.e. $H_0: \overline{\Delta\theta}_{E,\alpha|\omega} = \overline{\Delta\theta}_{\bar{E},\alpha|\omega}$ by sub-period and setup; secondly, mean difference tests between the average changes of pairwise sub-periods, such as $H_0: \overline{\Delta\theta}_{J,2000|2005} = \overline{\Delta\theta}_{J,2006|2011}$, to test whether the average changes in relative productivity differs significantly between sub-periods within a group.

Starting with teaching for the ExIni-funded universities in Setup 1 as first scenario, for period 2000-05 (before the Excellence Initiative was launched), the average change in relative productivity

[13] As we forward output by three years, we loose three years of observations, while the time span of the dataset ranges from 2000 to 2019.

Table 3

Average change in relative productivity of Excellence Initiative-funded and non-funded universities

(a) Setup 1

	Teaching				Research			
	ExIni-funded (1)		non-funded (2)		ExIni-funded (3)		non-funded (4)	
sub-period	μ	σ	μ	σ	μ	σ	μ	σ
2000-05	-0.096	0.154	-0.015	0.373	0.100**	0.242	-0.079	0.405
2006-11	°°-0.001	0.186	0.117	0.715	0.103	0.215	°°°0.185	0.359
2012-16	0.002	0.133	-0.019	0.139	°0.038**	0.129	°°°-0.034	0.143

Note: In Setup 1, the efficiency scores are calculated with "number of academic staff" and "general expenditure" as inputs. μ indicates the average change of efficiency scores at the end of the period relative to the beginning of the period, σ represents the standard deviation of these changes. A mean difference test is conducted between the average change of efficiency scores of excellence-funded universities and the ones of the non-funded universities in each of the three time periods, indicated by asterisks with significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. In addition, a mean difference test between periods was executed, indicating whether the average change in scores of a group of universities is significantly different from its average changes in the previous period, with significance levels: °°° $p < 0.01$, °° $p < 0.05$, ° $p < 0.1$.

(b) Setup 2

	Teaching				Research			
	ExIni-funded (5)		non-funded (6)		ExIni-funded (7)		non-funded (8)	
sub-period	μ	σ	μ	σ	μ	σ	μ	σ
2000-05	-0.111	0.154	-0.030	0.347	0.082	0.22	-0.023	0.335
2006-11	°°°-0.004	0.203	0.098	0.540	0.078	0.194	°°0.145	0.301
2012-16	-0.007	0.125	-0.010	0.148	0.044***	0.111	°°°-0.029	0.125

Note: In Setup 2, the efficiency scores are calculated with "number of academic staff", "general expenditure" and "third-party funding" as inputs. μ indicates the average change of efficiency scores at the end of the period relative to the beginning of the period, σ represents the standard deviation of these changes. A mean difference test is conducted between the average change of efficiency scores of excellence-funded universities and the ones of the non-funded universities in each of the three time periods, indicated by asterisks with significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. In addition, a mean difference test between periods was executed, indicating whether the average change in scores of a group of universities is significantly different from its average changes in the previous period, with significance levels: °°° $p < 0.01$, °° $p < 0.05$, ° $p < 0.1$.

between 2000 and 2005 is $\overline{\Delta\theta}_{E,2000|2005} = -9.6\%$ indicating an improvement of average relative teaching productivity of ExIni-funded universities and hence convergence. In the following period 2006-11, the average change drops to $\overline{\Delta\theta}_{E,2006|2011} = -0.1\%$, very close to no change on the average at all. For 2012-16 it changes to $\overline{\Delta\theta}_{E,2012|2016} = 0.2\%$, again a value close to no change on the average at all. Comparing these average relative productivity changes across the three sub-periods shows that only the values of 2000-05 and of 2006-11 are significantly different. (i.e. $\overline{\Delta\theta}_{E,2000|2005} \neq \overline{\Delta\theta}_{E,2006|2011}$) This is indicated by the two circles (5% significance level) before -0.001 in the row labeled 2006-2011.

Looking at the development in relative teaching productivity of the non-funded universities, a process of convergence in sub-period 2000-05 is followed by an insignificant change towards divergence during 2006-11. For 2012-16 a re-switch to convergence is identified. When the average changes of the three sub-periods are compared pairwise, none of them is significantly different from the others within the group of non-funded universities.

The comparison of the average changes of relative teaching productivity of the ExIni-funded and the non-funded universities period by period shows these changes to be not significantly different.

Interpreting all sub-tables as a whole, the following three results can be obtained: A first result addresses the direction of change in relative productivity. With respect to teaching these changes are negative, with few exceptions. Conversely, for research average changes are primarily positive for ExIni-funded universities, whereas for non-funded universities, there appears to be a converging tendency before and after the two initial rounds of the Excellence Initiative. Hence, overall, within the considered time span from 2000 to 2016 we find an average convergence in productivity for teaching, but a divergence in research performance. These results match well with the analysis of the Salter curve for 2005 and 2015 in Figure 2, and show that the pattern holds for the entire period of analysis.

A second result refers to the comparison between ExIni-funded and non-funded universities. Irrespective of the setup, the differences in the average changes of relative productivity between ExIni-funded and non-funded universities are not significant with respect to teaching. Concerning research in Setup 1 as well as Setup 2, out of the six comparisons three deliver significant differences. In those cases it is evident that the group of ExIni-funded universities diverge from best-practice, whereas non-funded universities show convergence. Comparing the standard deviations between the two groups of universities in all eight scenarios and each sub-period reveals that the spread of the changes of non-funded universities is notably higher than for supported ones. This result already indicates that in order to compare ExIni-funded and non-funded universities, an approach with proper comparison groups needs to be applied (see Section 6).

The comparison of average changes in relative productivity over time yields a third result. In teaching, the trend toward convergence is significantly weakening from 2000-05 to 2006-11. This mainly holds true for ExIni-funded universities. For research, the divergence for funded universities is weakly significant in Setup 1 when comparing changes in relative productivity in 2006-11 with 2012-16; hence, the divergence continues from 2006-11 to 2012-16. For non-funded universities, we detect a significant divergence in 2006-2011, and a significant convergence in 2012-16 – in both setups.

Although the different stages of development suggest that the Excellence Initiative plays a role here, these descriptive results do not yet provide a convincing picture of the expected impact of the Excellence Initiative in terms of hypotheses 1 and 2. To deliver more reliable insights (on the basis of e.g. proper comparison groups), we will take a closer look at the impact of the Excellence Initiative, in the following. A difference-in-difference approach will help identify robust results.

6 The Effect of the Excellence Initiative

In the second step of our analysis, we execute a difference-in-difference (DID) analysis to scrutinize the effectiveness of excellence funding on universities' productivity performance. We now use the efficiency scores which we calculated for both missions, teaching and research (see section 5), as a dependent variable to compute the contribution of excellence funding to universities' relative productivity performance.

6.1 Matching and difference-in-differences approach

In this section, we draw on the Difference-in-differences (DID) approach by Callaway and Sant'Anna (2021).[14] The Difference-in-differences (DID) analysis has become a pervasive tool, particularly for evaluating policy interventions (Callaway and Sant'Anna, 2021). Despite its relatively long tradition, it has been advanced in many ways to cope with several caveats in its original form (e.g. Goodman-Bacon, 2021). One of the main challenges in DID, as is the case in our study, is to find an appropriate control group in order to establish the counterfactual (Callaway, 2022; Scandura, 2016). In well-designed experiments, treatment and control group are randomized to avoid a selection bias. In the case of observational data, however, randomized groups are difficult to establish, as an ex post randomization is infeasible (Scandura, 2016; Stuart, 2010). Regarding our research context, the universities that won the Excellence Initiative, i.e. the treatment group, were selected by the corresponding funding agency (the German Research Foundation, called DFG). Hence, they were not randomly selected into treatment. Instead, the selection committee had to select universities for funding based on particular characteristics, such as previous funding success or publication records. Consequently, there is a selection bias. To circumvent such a selection bias, we follow Almus and Czarnitzki (2003) and Stuart (2010) by identifying statistical twins, i.e. universities that are not treated but are in a statistical sense very similar to the treated universities. The downside of the procedure is that we have to exclude universities that do not qualify as a counterfactual twin (Almus and Czarnitzki, 2003; Grashof and Kopka, 2022; Leusin, 2022), implying a loss of information. We conduct a propensity score matching (PSM) based on the nearest neighbour algorithm as, for instance, in Abadie and Imbens (2016) or

[14] In the general format of DID, there are two time periods and two groups. In the first period, both groups receive no treatment. In the second period, some units are treated (the treated group) while others remain untreated (the control group). One crucial assumption of the DID approach is the parallel trend assumption, meaning that in the absence of the treatment, the average outcomes of the actually treated group and the control group would follow parallel paths over time (i.e. a common trend). If the parallel trend assumption holds true, the average treatment effect for the treated group (ATT) can be estimated by comparing the average change in outcomes in the treated group with the average change in outcomes in the control group (Callaway and Sant'Anna, 2021; Callaway, 2022).

Rosenbaum and Rubin (1983).[15] The propensity score matching has been frequently used to estimate causal effects by accounting for observable covariates that predict the receipt of the corresponding treatment (Caliendo and Kopeinig, 2008; Scandura, 2016). The challenge is to detect a sufficient set of covariates to find a match between a treated and an untreated university as control.[16] The resulting propensity score can then be interpreted as the probability to receive funding from the Excellence Initiative (our treatment). Due to space limitations, the detailed propensity score matching process and the corresponding results are presented in Appendix B.

After the matching, we now turn back to the DID identification strategy. In the canonical version of DID, the most common approach to estimate the treatment effect is to use a two-way fixed effects (TWFE) panel regression with unit and time fixed effects (Callaway, 2022). However, recent studies noted substantial drawbacks in using TWFE, especially when the treatment occurs at multiple time periods, which potentially leads to heterogeneous treatment effects (inter alia, Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). With a staggered treatment, more heterogeneous groups can be detected than in the canonical version, such as the never treated, the not yet treated, and the already treated group. In this case, a simple TWFE DID design would lead to a bias as soon as the treated observations are compared to the already treated observations (Ardiyono, 2022; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021). This issue is also relevant for the Excellence Initiative, as it was executed in several rounds. In a TWFE DID setting, some of the funded universities in 2007, for instance, would be compared to universities that had already been treated in 2006 (first round). To cope with this issue, we follow Callaway and Sant'Anna (2021), who suggest an extended DID approach that allows for multiple treatments at different time periods.[17] It extends the classical DID approach by two additional assumptions: (1) The Staggered Treatment Adoption Assumption: once an entity becomes treated, it remains the status of a treated entity in future time periods (Callaway and Sant'Anna, 2021; Callaway, 2022). (2) It allows for assuming covariate-specific trends in outcomes across groups, time, and/or length of exposure to the treatment. For instance, one may need to assume parallel trends with respect to the never-treated group and/or the not-yet-treated group. The latter is especially useful when the never-treated control group is too small, as it allows using more groups for comparison (Callaway and Sant'Anna, 2021). Nevertheless, since we have a sizeable group of control units in our sample, it suffices to apply the

[15] As indicated by Caliendo and Kopeinig (2008), there is not one best solution in terms of choosing the corresponding matching algorithm, but a trade-off between bias and efficiency, depending on the data at hand. Particularly, for small samples the choice of the matching algorithm can be crucial (Heckman, 1997). Since in our concrete research setting, the propensity score distribution (before the actual matching) is different between the treated and the potential control group (see Figure 7), we used the nearest neighbour algorithm with replacement. By allowing an untreated university to be matched multiple times, we reduce the number of mismatches (i.e the case when high score treated universities are matched with low score untreated universities) and therefore minimize potential biases at the price of higher variance (Caliendo and Kopeinig, 2008). In line with Guerzoni and Raiteri (2015), as a further robustness check, we have also implemented the nearest neighbour algorithm with a "caliper" threshold of 0.09, corresponding approximately to 0.25 times the standard deviation of the propensity scores recovered with the probit regression (Rosenbaum and Rubin, 1985).

[16] With their introduction in 1983, (Rosenbaum and Rubin, 1983), the propensity scores facilitated the construction of matched samples by not requiring close or exact matches on all covariates, thereby overcoming the so-called *curse of dimensionality*, i.e. the higher the number of covariates and the more values these can take, the lower is the probability of finding good matches with similar characteristics in finite samples. Instead of directly considering all covariates, propensity scores summarize all of the covariates into one scalar, namely the probability of being treated, i.e. winning the Excellence Initiative (Huber, 2021; Stuart, 2010; Zhao, 2008).

[17] For the actual implementation we make use of the Stata module "csdid" (Rios-Avila et al., 2022).

never-treated option, similar to previous studies (e.g. Leusin, 2022). Hence, for the estimation of the group-time average treatment effect $ATT(g,t)$ on the group that is treated at time g, we obtain at time t:

$$ATT(g,t) = E(Y_t - Y_{g-1}|G = g) - E(Y_t - Y_{g-1}|C = 1) \tag{2}$$

where $E(Y_t - Y_{g-1}|G = g)$ is the average potential outcome for treatment group $G = g$ at time t, with G denoting the time when a unit is first treated. It indicates which treatment group a unit belongs to, and $t > g$. $E(Y_t - Y_{g-1}|C = 1)$ is the average outcome for control group C that remains never-treated (Callaway and Sant'Anna, 2021; Callaway, 2022; Di Matteo, 2022). Since all group-time average treatment effects in Equation (2) have to be estimated by group and time, the resulting coefficients need to be aggregated in order to be interpretable. Callaway and Sant'Anna (2021) propose different aggregation schemes, of which we use the aggregation schemes "heterogenous group treatment effects", "dynamic treatment effects", and "simple weighted average". The ATT for all post-treatment periods varying across groups is estimated in the following way:

$$\theta_{sel}(\tilde{g}) = \frac{1}{\tau - \tilde{g} + 1} \sum_{t=\tilde{g}}^{\tau} ATT(\tilde{g}, t) \tag{3}$$

with $\theta_{sel}(\tilde{g})$ as average treatment effect for units in group $\tilde{g}$. The aggregate thereby includes all post-treatment periods until the last period τ. From Equation (3), we can now derive θ^o_{sel} as the ATT across all groups over all post-treatment periods:

$$\theta^o_{sel} = \sum_{g \in G} \theta_{sel}(g) P(G = g|G \leq t) \tag{4}$$

where $P(G = g|G \leq t)$ denotes the probability of an observation ending up in group g. θ^o_{sel} is therefore the average treatment effect experienced by all units across all the treated groups in all the post-treatment time periods. Its interpretation is most similar to the ATT on the treated in the traditional DID setup (Callaway and Sant'Anna, 2021). To shed light on the dynamics of treatment effects varying across treatment exposure time or event-time, we need to compute:

$$\theta_{es}(e) = \sum_{g \in G} 1\left\{g + e \leq \tau\right\} P(G = g|G \leq \tau) ATT(g, g + e) \tag{5}$$

where e denotes the event-time, with $e = t - g$ indicating the elapsed time since the start of the treatment. Hence, $\theta_{es}(e)$ indicates the average treatment effect of all groups having participated e periods in the treatment. The ATT for each period is thereby measured relative to the first period of treatment across all cohorts (Rios-Avila et al., 2022). For the estimation of the above equations (3, 4, and 5), we apply the doubly robust DID estimators based on stabilized inverse probability weighting and ordinary least squares (Sant'Anna and Zhao, 2020; Callaway and Sant'Anna, 2021).

6.2 Average treatment Effects of the Excellence Initiative

Since different treatment groups and control groups are required for testing different hypotheses, we create an overview of the hypotheses and the corresponding test groups in Table 4. For hypotheses 1 and 2, we examine the influence of the Excellence Initiative on the relative productivity of teaching and of research. The universities that won any of the three funding lines (Graduate Schools, Clusters of Excellence and Future Concepts) in three periods (2006, 2007 and 2012) constitute the treatment group. The control group is derived from PSM among the universities that did not win any of the funding lines. To investigate hypotheses 3 and 4, we need to test the effects of the Future Concepts and the Graduate Schools and/or Clusters of Excellence respectively, before we can compare both effects with the effects of the Excellence Initiative in general. To test the impact of the Future Concepts, the treatment group consists of universities that won the Future Concepts funding line. The corresponding control group is constructed by PSM among all the universities that did not win the Future Concepts, which may also include the universities that won Graduate Schools and/or Clusters of Excellence. Finally, in order to examine the impact of Graduate Schools and/or Clusters of Excellence, the universities that won Graduate Schools and/or Clusters of Excellence but not the Future Concepts form the treatment group. Accordingly, the universities that did not win any of these funding lines become the control group (similar to the control group for hypotheses 1 and 2). It is worth noting that for all the hypotheses 1 to 4, we conduct the analyses for the two main missions of universities (teaching and research) and the two setups described in Section 4.1.

Table 4

Overview of hypotheses and the corresponding test groups

Hypotheses	Treatment groups	Control groups
Hypotheses 1 & 2	Excellence Initiative Uni.	non-ExIn. Uni. PSM
Hypothesis 3	Graduate Schools/Clusters of Excellence Uni.	
Hypothesis 4	Future Concepts Uni.	non-FuCon. Uni. PSM

Notes: "non-ExIn. Uni. PSM": control group derived from PSM (Propensity Score Matching) among the universities that did not win any of the three funding lines of Excellence Initiative. "non-FuCon. Uni. PSM": control group derived from PSM among the universities that did not win Future Concepts, including the universities that won Graduate Schools and/or Clusters of Excellence and the universities that did not win any of the three funding lines.

6.2.1 Average treatment effects of the Excellence Initiative in general

In a first step, we test whether an overall effect of the Excellence Initiatives (i.e. all winners of all three funding lines) can be detected (see hypotheses 1 and 2). For doing so, we follow the DID approach from Callaway and Sant'Anna (2021) specified in subsection 6.1 (see equation three). After running the regressions for each of the two missions and the two setups, with the universities that won the Excellent Initiative as the treatment group and the rest of the universities (matched by PSM) as the control group, we aggregate the partial treatment effects to generate the results in Table 5.

As shown in Table 5, the coefficients for teaching are positive, which indicates a deterioration of relative productivity for universities with excellence support, but the results are not significant.[18] Contrary to hypothesis 2, across all groups and all time periods the Excellence Initiative does therefore not lead to a statistically significant lower relative teaching productivity of the treated universities.

With respect to research, we can only detect a significantly negative effect of the Excellence Initiative in setup 1. This supports hypothesis 1, since the significant coefficient alludes to an average increase of funded universities' research efficiency by 21.3% which is also significantly different to the non-funded universities. As emphasized in subsection 4.1, this coefficient includes a positive

[18] We run several robustness, such as an alternative PSM with average share of third-party funding from DFG, average number of students and dummy of technical university. The robustness checks shown in Appendix C and Appendix D do not deliver significant results in this respect, either.

Table 5
Overall ATT of the Excellence Initiative (across all groups and all time periods)

	Teaching Productivity		Research Productivity	
	Setup 1	Setup 2	Setup 1	Setup 2
ATT of Excellence Initiative	0.057	0.033	-0.213*	-0.129
	(0.072)	(0.071)	(0.113)	(0.094)
Number of observations	867	867	867	867

Clustered standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes: The treatment group consists of all the universities that ever won any of the funding lines of Excellence Initiative in the three rounds (2006, 2007, 2012). Relative productivity scores in teaching and research are the dependent variables respectively. In setup 1, they are calculated with "number of academic staff" and "general expenditure" (2000-2016); in setup 2 with "number of academic staff", "general expenditure" and "third-party funding" (2000-2016) as inputs. Both setups employ "number of graduates" (2003- 2019) as output in teaching and "number of publications" (2003- 2019) as output in research.

bias by not counting the non-personnel third-party funding as a part of the input. Setup 2, which double counts the personnel expenditure by including the third-party funding without excluding the embedded personnel funding, should have a negative bias. The ATT in setup 2, has a negative sign, but is insignificant (p = 0.168). Evidence for an significant influence on the research efficiency can therefore only be found in the basic setup. When considering additional financial resources coming from third-party funding, this influence becomes statistically insignificant. These two findings are also confirmed when we use alternative propensity score matching approaches (see Table 15 in Appendix C)[19] as well as specifically account for third-party funding coming from the DFG (see Table 17 in Appendix D).

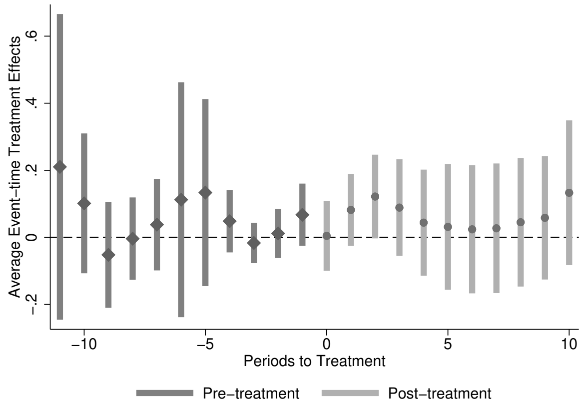
Hence, while the empirical results in the case of teaching efficiency are quite clear, they are so far rather ambiguous with regard to research efficiency. Besides the rather "technical" reason (i.e. potential over- and underestimation), the different results from setup 1 and 2 might also indicate that the efficiency gains from the Excellence Initiative, found in our basic setup, are partially offset by third-party funding coming from other sources. This would also go in line with recent findings by Buenstorf and Koenig (2020), claiming that winning universities in the Future Concepts funding line received significantly less direct funding from the federal government.

To understand the influence of the Excellence Initiative even more precisely, we aggregate individual treatment effects, as suggested by Callaway and Sant'Anna (2021) introduced in subsection 6.1 equation 5, to event-time ATTs (see Figure 3). The darker bars show the distribution of relative productivity scores before treatment; the lighter bars depict the change in relative productivity after treatment. If a lighter bar does not touch the dashed horizontal line, the treatment effect is considered statistically significant. In setup 1 for teaching (see Sub-figure 3a), the relative productivity scores of funded universities are only weakly significant two years after treatment. Since the effect on the scores is positive, the relative productivity in teaching of funded universities has been relatively lower against the non-funded universities after treatment. Similar results are found, in setup 2 (Sub-figure 3b), when additional third-party funding is taken into account. In both setups, we find that two years after having received funding from the Excellence Initiative, the teaching relative productivity is

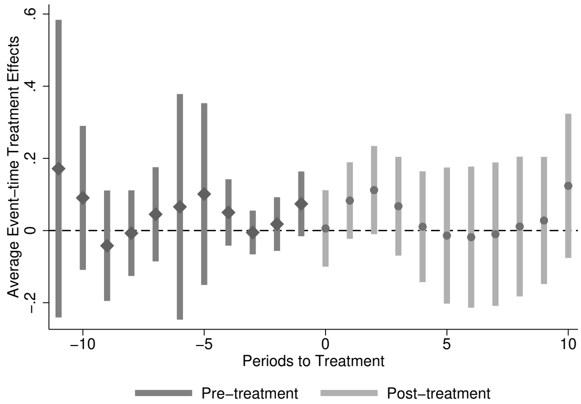
[19]The results for the second setup in the robustness check with alternative matching variables are the only exception here. The corresponding coefficient is slightly below the p-value of 0.1 and is therefore significant at conventional levels.

significantly decreased (indicated by a positive ATT on the teaching efficiency scores). In other words, there is slight empirical evidence that the Excellence Initiative can also imply negative effects on the relative teaching productivity of universities, although on average this effect turns to be insignificant. Hypothesis 2 can therefore not fully be rejected.

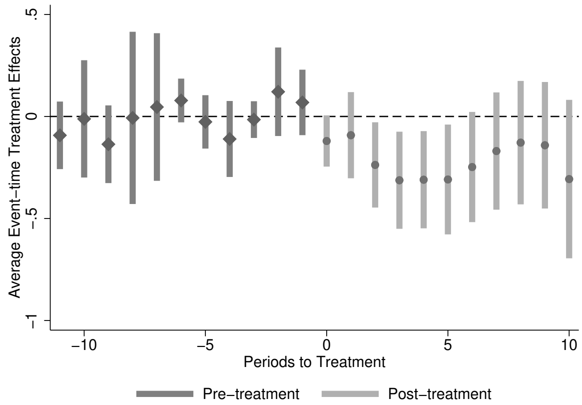
Fig. 3. Dynamic event-time treatment effect of the Excellence Initiative



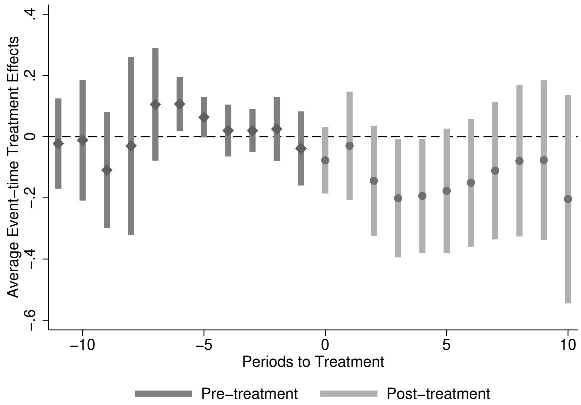
(a) Teaching – Setup 1



(b) Teaching – Setup 2



(c) Research – Setup 1



(d) Research – Setup 2

Note: The treatment group consists of all the universities which ever won any of the funding lines of Excellence Initiative in three rounds (2006, 2007, 2012). Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with "number of academic staff" and "general expenditure" (2000-2016) in setup 1; with "number of academic staff", "general expenditure" and "third-party funding" (2000-2016) in setup 2 as inputs. Both setups employ "number of graduates" (2003-2019) as output in teaching and "number of publications" (2003-2019) as output in research. The bars mark the 95% confidence interval.

The event study analysis also offers interesting insights regarding the relative research productivity. Sub-figure 3c shows the dynamic treatment effect of the Excellence Initiative on the relative research productivity of the treated universities (in our basic setup). Two years after the treatment we do observe an significant negative impact, which lasts until five years after the treatment. Hence, without considering third-party funding as an additional input (setup 1), the Excellence Initiative increases the relative research productivity of the treated universities in the short to medium term (compared to the never-treated control group). A somewhat similar pattern can even be found in the case of setup 2 (Sub-figure 3d), where the overall ATT across all groups and all time periods (see Table 5) is actually insignificant. Although the effect of the Excellence Initiative loses significance (compared to our first setup), a significant influence of the Excellence Initiative can still be detected for three to five years after the treatment. Consequently, we find evidence that the Excellence Initiative has increased the relative research productivity of the treated universities in research in the short to medium run, with or without taking the third-party funding into account as a part of the input.

6.2.2 Average treatment effects of Future Concepts – elite university status

Nevertheless, these results hold only true for the impact of the Excellence Initiative in general (including all three funding lines). However, as indicated in section 3, there are formal differences between the three funding lines, which should also be investigated in order to further disentangle the impact of the Excellence Initiative.

In this section, we investigate the ATT of the Future Concepts funding line (see Hypothesis 4). The universities that won the Future Concepts make up our treatment group. In order to construct our control group, we follow the same matching approach as described in section 6.1 for universities that did not win Future Concepts (including universities that won the Graduate Schools or Clusters of Excellence, and universities that did not win any of the funding lines)[20]. Table 6 shows the overall ATTs (across all groups and all time periods) on teaching and research relative productivity for both setups. The corresponding results indicate that the gained flexibility and reputation for winners of Future Concepts do not translate into relative productivity gains. In both setups, we do not find evidences indicating a significant impact of Future Concepts on neither teaching nor research productivity of treated universities. Hence, winning the Future Concepts and thereby receiving flexibility gains in investments decisions and the title of "elite university" do not result in an productivity premium, which is in line with Hypothesis 4.

Table 6

Average treatment effects of Future Concepts (across all groups and all time periods)

	Teaching Productivity		Research Productivity	
	Setup 1	Setup 2	Setup 1	Setup 2
ATT of Excellence Initiative	0.059	0.056	-0.278	-0.264
	(0.081)	(0.082)	(0.238)	(0.245)
Number of observations	425	425	425	425

Clustered standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes: The treatment group consists of the universities which won Future Concepts in three rounds (2006, 2007, 2012). The control group is derived from PSM within the universities that won the Graduate Schools or Clusters of Excellence, and universities that did not win any of the funding lines. Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with "number of academic staff" and "general expenditure" (2000-2016) in setup 1; with "number of academic staff", "general expenditure" and "third-party funding" (2000-2016) in setup 2 as inputs. Both setups employ"number of graduates" (2003-2019) as output in teaching and "number of publications" (2003-2019) as output in research.

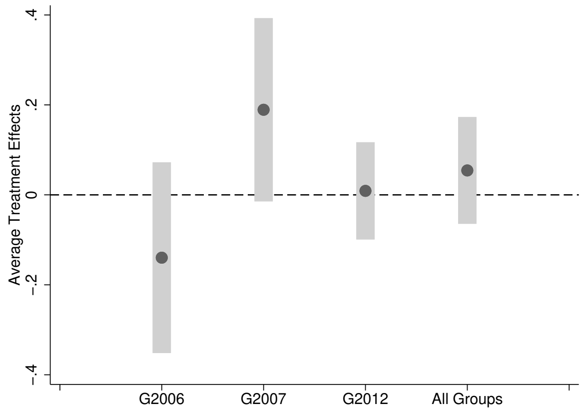
Instead, by estimating the ATT for each group, averaged across all post-treatment periods (as described in subsection 6.1 equation 3), we can even find a significant negative influence on the relative

[20]The corresponding results of the matching approach are presented in Appendix E. As a further robustness check, we follow previous studies (e.g. Buenstorf and Koenig, 2020) and also use universities that did not win the Future Concepts, but the Graduate Schools and/or Clusters of Excellence as a "natural" control group. The corresponding results remain robust to those shown in this section and are presented in Appendix F as well as Appendix G.

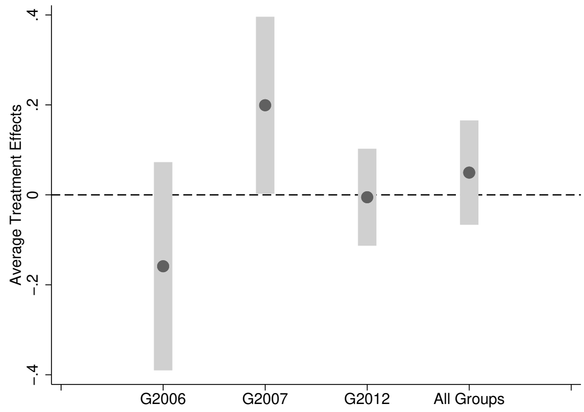
teaching productivity of the second-round (in 2007) Future Concepts' winners. As can be seen in subfigures 5a and 5b, the ATT for the second group (group 2007) on relative teaching productivity is significant positive. This means that winning the Future Concepts significantly decreases the relative teaching productivity of the second group.

At the same time, however, we find no significant impact of Future Concepts on relative research productivity for any group in both setups. The only exception in this context is the first group (in our first setup), where we find a slightly significant negative effect (p-value = 0.092).[21] Thus, the relative research productivity gained through the Excellence Initiative in general (shown in section 6.2.1) is not driven by the funding from Future Concepts or in other words the reputation of "elite universities". Instead, it appears that the relative productivity is particularly driven by universities that have not won the Future Concepts.

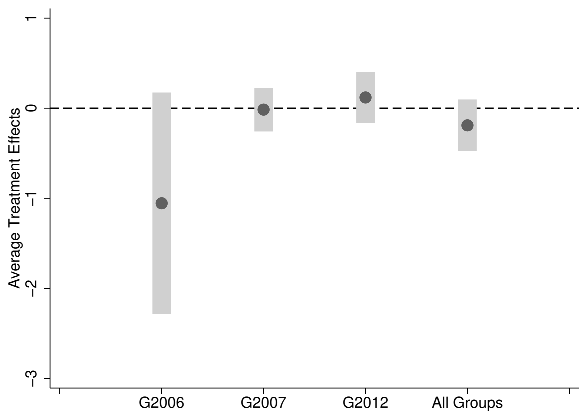
Fig. 5. Average treatment effects across groups of Future Concepts



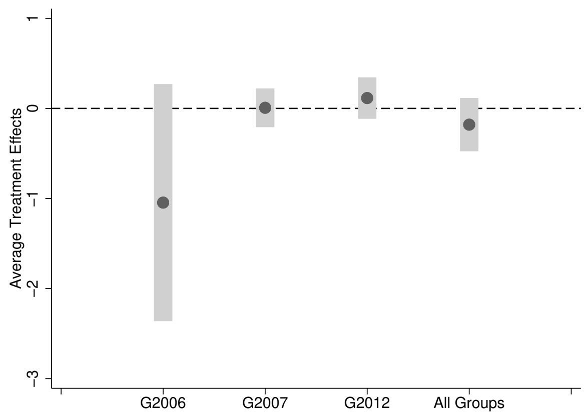
(a) Teaching – Setup 1



(b) Teaching – Setup 2



(c) Research – Setup 1



(d) Research – Setup 2

Note: The treatment group consists of the universities which won Future Concepts in three rounds (2006, 2007, 2012). The control group is derived from PSM within the universities that won the Graduate Schools or Clusters of Excellence, and universities that did not win any of the funding lines. Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with "number of academic staff" and "general expenditure" (2000-2016) in setup 1; with "number of academic staff", "general expenditure" and "third-party funding" (2000-2016) in setup 2 as inputs. Both setups employ "number of graduates" (2003-2019) as output in teaching and "number of publications" (2003-2019) as output in research. The bars mark the 95% confidence interval.

[21] However, also in the further robustness checks (e.g. with different matching variables) we do not find such a significant influence for any group. The corresponding results can be provided upon request.

6.2.3 Average treatment effects of Graduate Schools and Clusters of Excellence

To further empirically support this claim, we now focus on universities that either won the Graduate Schools and/or Clusters of Excellence funding line, but that have not won the Future Concept funding line (see Hypothesis 3 in Table 4). The control group is derived from PSM within universities that did not win any of the three funding lines. To construct our control group, we had to adapt our matching variables, as the quality of the previous matching approach (explained in section 6.1) did not meet the criteria suggested by Rubin (2001). We therefore used the following matching variables: Average share of third-party funding from the DFG in total third-party funding, a dummy variable capturing whether a university is a technical university or not, and the average number of students from 2000 to 2005.[22] Based on the control group derived from PSM we estimate the overall ATT across all groups (see equation 4) on teaching and research relative productivity.[23]

As Table 7 shows, the results are quite similar to those for the impact of the overall Excellence Initiative (see Table 5). In line with the impact of the overall Excellence Initiative, Graduate Schools and Clusters of Excellence turn out to have a significant negative effect on the relative research productivity in setup 1, where third-party funding is not included in the inputs. This means that the Graduate Schools and/or Clusters of Excellence funding lines increase relative research productivity compared to untreated universities. Moreover, the impact on the relative teaching productivity is again positive and insignificant.

Table 7

ATT of Graduate Schools/Clusters of Excellence (across all groups and all time periods)

	Teaching Productivity		Research Productivity	
	Setup 1	Setup 2	Setup 1	Setup 2
ATT of Graduate School/Clusters of Excellence	0.032	0.023	-0.145*	-0.083
	(0.097)	(0.094)	(0.088)	(0.069)
Number of observations	629	629	629	629

Clustered standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes: The treatment group consists of the universities which won Graduate Schools and/or Clusters of Excellence in three rounds (2006, 2007, 2012), but did not win any Future Concepts. The control group is derived from PSM within universities that did not win any of the funding lines. Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with "number of academic staff" and "general expenditure" (2000-2016) in setup 1; with "number of academic staff", "general expenditure" and "third-party funding" (2000-2016) in setup 2 as inputs. Both setups employ "number of graduates" (2003-2019) as output in teaching and "number of publications" (2003-2019) as output in research.

Hence, the overall results of the Excellence Initiative are rather driven by the average treatment effects of Graduate Schools and Clusters of Excellence, thereby supporting Hypothesis 3. This goes in line with the rather descriptive findings of Menter et al. (2018). By differentiating between first-tier universities ("elite" universities that won the Future Concepts), second-tier (non-"elite" excellence universities that won the Graduate Schools and/or Clusters of Excellence only) and third-tier universities (universities without any funding from the Excellence Initiative), they show that even though the number of research fellows increase more in the first-tier than in the second-tier universities (after the first funding round), the relatively strong increase in the number of scientists at first-tier universities

[22]Due to space limitations, the results of the corresponding matching approach are presented in Appendix H.

[23]The results for the dynamic event-time ATTs and the across-group ATTs are presented and discussed in Appendix I.

is not reflected in a higher number of publications. In contrast, it is the second tier universities that are able to drastically increase their publication output per scientist (Menter et al., 2018).

The rather high degree of freedom and flexibility in the case of Future Concepts therefore seem to have negative implications for research productivity. Since there are often internal discussions and contests within the Future Concepts funded universities (Buenstorf and Koenig, 2020; Imboden et al., 2016), it is not clear from the beginning for what specific projects the funds shall be invested. It is therefore reasonable that the increase of scientists, which is probably spurred by the mere title of an "elite" university, is not directly associated with specific research projects. As a result, it is not directly translated into a higher publication output. On the contrary, in the case of Graduate Schools and Clusters of Excellence, the funded universities have a rather fixed and concrete focus on specific projects (given by the corresponding project proposals submitted during the application process). As a consequence, the newly hired scientists can immediately start in a specific project, which most likely leads faster to first research outputs and thereby to a higher relative research productivity.

7 Conclusion

Since the Bologna Process, the autonomy of universities has increased and along with it the competition for and relevance of third-party funding (Krücken, 2021; Lehmann et al., 2018; Wiener et al., 2020). This paper therefore aimed to contribute to the ongoing discussion about the role of competitive funding in the higher education system (e.g. Gawellek and Sunder, 2016; Menter et al., 2018) by investigating the impact of the Excellence Initiative, being one of the biggest and most controversial grant schemes in Germany, on the relative teaching and research productivity of the funded public universities. We thereby follow the recent call by Menter et al. (2018) and systematically examine the Excellence Initiative with its three funding periods and three funding lines, allowing us to take a more holistic view than previous literature. Moreover, we additionally enrich previous literature about the Excellence Initiative by differentiating between the, so far ignored, effects on relative teaching and research productivity, being the two main "missions" of universities (Gautier and Wauthy, 2007). Based on a unique university-level database encompassing detailed information from the Federal Statistical Office of Germany and Scopus, we firstly conducted a relative productivity analysis (using non-parametric Data Envelopment Analysis) and secondly we used the DID approach suggested by Callaway and Sant'Anna (2021) to estimate the average treatment effect of the Excellence Initiative.

All in all, the Excellence Initiative has had only a slight structuring effect on German universities, and it does not seem to be long-lasting - it evaporates and dissipates over time. With respect to research, our results indicate that the Excellence Initiative significantly and positively affected the relative research productivity of the ExIni-funded universities compared to non-funded ones, particularly in a short to medium term. However, when considering financial resources from third-party funding as an additional input to research, the overall average treatment effect becomes insignificant. This implies that there were potential compensation effects from other funding sources which benefited universities that were not successful in the Excellence Initiative (e.g. Buenstorf and Koenig, 2020). Hence, there is no persistent tendency for ExIni-funded and non-funded universities to diverge. If we further differentiate the effect of the funding lines, we also find that it is particularly those universities that did not receive the Future Concept funding line but did receive the Graduate Schools and/or Clusters of Excellence funding line that experienced an increase in relative research productivity. In other words, the given advantages of winning the Future Concept funding line (i.e. reputation and flexibility gains) do not (yet) result in a higher relative research productivity compared to non-excellence universities.

As to teaching performance, we do not find a significant influence of the Excellence Initiative on the relative teaching productivity across all groups and all time periods. Hence, on average the Excellence Initiative (and also its specific funding lines) appears not to reduce the relative teaching productivity of the corresponding universities. However, we can detect time-specific and group-specific negative influences in this context. Across all groups, we find slight evidence that two years after the funding, the Excellence Initiative significantly decreases the relative teaching productivity. In the case of "Excellence" universities, i.e. winners of the Future Concepts funding line, our results show that the second round of winners experienced a significant decline in relative teaching productivity across all time periods.

Nevertheless, when considering our results, some limitations must be discussed that may provide the starting point for future research. Due to data limitations,[24] the focus of our empirical analysis lies on the university-level. Thus, we might miss some discipline-related effects. Although a lower disaggregation level might cause multicollinearity problems (Olivares and Wetzel, 2014), future research should deal with this issue and further disentangle the effect of the Excellence Initiative across different disciplines. Furthermore, our two output variables for teaching and research are only quantitative in nature (i.e. number of graduates and number of publications). Therefore, it might be promising to consider more qualitatively oriented output variables in future research. Particularly in the case of teaching, it would be highly relevant to find and use information about the actual jobs of graduates in order to estimate the quality of teaching at the corresponding university. Despite its advantages, DEA has also some drawbacks, such as its sensitivity to potential outliers. Future research could therefore apply, for example, partial frontier analysis (e.g. Gnewuch and Wohlrabe, 2018) or the Malmquist Index (e.g. Cantner et al., 2007; Caves et al., 1982) to shed light on further aspects of the competitive dynamics. Moreover, instead of restricting our attention to the Excellence Initiative, as one of the biggest and most controversial competitive grant schemes in Germany, a comparison of the Excellence Initiative with similar competitive grant schemes from other countries (e.g. the Initiative D'Excellence in France) could be an instructive future research avenue.

Overall, our results suggest that the Excellence Initiative has (i) mixed effects on relative research and teaching productivity, (ii) funding round and time specific effects, as well as (iii) funding line specific effects. Beyond the specific empirical setting, consideration of these three points can also help better understand the role and performance implications of other competitive funding programs in the higher education system. Moreover, they are also highly relevant for evidence-based higher education policies. For example, the rather mixed effects of the Excellence Initiative, depending on the specific outcome variable of interest, emphasize that policy makers should take a rather holistic view in order to avoid (unintentional) negative effects of the corresponding funding program. Moreover, the positive influence of the Excellence Initiative on relative research productivity is more likely to play out in the short to medium term. Hence, it seems reasonable that future policy interventions in the university sector follow a rather long-term perspective in order to achieve efficiency gains. Lastly, the Future Concepts funding line needs to be put into perspective concerning its impact on relative productivity. While the basic idea of promoting an institutional strategy seems promising, especially in terms of a broad impact on the university as a whole. The current implementation tends not to promote relative research productivity – at least not in the short or medium run. Although long-run relative productivity advantages of the Future Concepts funding cannot be ruled out, due to current data limitations, the results for the other two funding lines give cause for consideration as to whether the

[24]Due to data privacy reasons, the data from Destatis is not fully available on a lower disaggregation level, such as faculties.

flexibility of the Future Concepts funding line should be further restricted, for example by requiring more concrete project plans.

Overall, our results indicate that the gains in relative research productivity induced by the Excellence Initiative are not so much driven by the "excellent" universities, but rather by second-tier universities that use their acquired resources from Graduate Schools and Clusters of Excellence funding line in a more targeted and efficient way. In other words: Excellence(-funded) universities are not necessarily excellent.

Appendices

A Universities Funded by Excellence Initiative

	Excellence Initiative in 2006	
	Future Concepts	
Ludwig-Maximilians-Universität München	Technische Universität München	Universität Karlsruhe
	Graduate Schools	
Albert-Ludwigs-Universität Freiburg	Freie Universität Berlin	Friedrich-Alexander-Universität Erlangen-Nürnberg
Humboldt-Universität zu Berlin	Julius-Maximilians-Universität Würzburg	Justus-Liebig-Universität Gießen
Ludwig-Maximilians-Universität München	Rheinische Friedrich-Wilhelms-Universität Bonn	Ruhr-Universität Bochum
Ruprecht-Karls-Universität Heidelberg	RWTH Aachen	Technische Universität Berlin
Technische Universität Dresden	Technische Universität München	Universität Bremen
Universität Karlsruhe	Universität Mannheim	
	Excellence Cluster	
Christian-Albrechts-Universität zu Kiel	Georg-August-Universität Göttingen	Johann Wolfgang Goethe-Universität Frankfurt am Main
Justus-Liebig-Universität Gießen	Ludwig-Maximilians-Universität München	Rheinische Friedrich-Wilhelms-Universität Bonn
Ruprecht-Karls-Universität Heidelberg	RWTH Aachen	Technische Universität Dresden
Technische Universität München	Universität Karlsruhe	Universität Konstanz
	Excellence Initiative in 2007	
	Future Concepts	
Albert-Ludwigs-Universität Freiburg	Freie Universität Berlin	RWTH Aachen
Universität Göttingen	Universität Heidelberg	Universität Konstanz
	Graduate Schools	
Christian-Albrechts-Universität zu Kiel	Freie Universität Berlin	Friedrich-Schiller-Universität Jena
Humboldt-Universität zu Berlin	Johannes Gutenberg-Universität Mainz	Technische Universität Darmstadt
Universität Bayreuth	Universität Bielefeld	Universität Bonn
Universität Bremen	Universität des Saarlandes	Universität Göttingen
Universität Heidelberg	Universität Konstanz	Universität Leipzig
Universität Stuttgart	Universität Ulm	Universität zu Lübeck
	Excellence Cluster	
Christian-Albrechts-Universität zu Kiel	Eberhard Karls Universität Tübingen	Freie Universität Berlin
Friedrich-Alexander-Universität Erlangen-Nürnberg	Humboldt-Universität zu Berlin	Johann Wolfgang Goethe-Universität Frankfurt am Main

Leibniz Universität Hannover	Ruprecht-Karls-Universität Heidelberg	RWTH Aachen
Technische Universität Berlin	Technische Universität Darmstadt	Universität Bielefeld
Universität Bremen	Universität des Saarlandes	Universität Freiburg
Universität Hamburg	Universität Stuttgart	Universität zu Köln
Westfälische Wilhelms-Universität Münster		

Excellence Initiative in 2012

Future Concepts

Eberhard Karls Universität Tübingen	Freie Universität Berlin	Humboldt-Universität zu Berlin
Ludwig-Maximilians-Universität München	Ruprecht-Karls-Universität Heidelberg	RWTH Aachen
Technische Universität Dresden	Technische Universität München	Universität Bremen
Universität Konstanz	Universität zu Köln	

Graduate Schools

Albert-Ludwigs-Universität Freiburg	Christian-Albrechts-Universität zu Kiel	Eberhard Karls Universität Tübingen
Freie Universität Berlin	Friedrich-Alexander-Universität Erlangen-Nürnberg	Friedrich-Schiller-Universität Jena
Georg-August-Universität Göttingen	Humboldt-Universität zu Berlin	Johannes Gutenberg-Universität Mainz
Julius-Maximilians-Universität Würzburg	Justus-Liebig-Universität Gießen	Karlsruher Institut für Technologie
Ludwig-Maximilians-Universität München	Otto-Friedrich-Universität Bamberg	Rheinische Friedrich-Wilhelms-Universität Bonn
Ruhr-Universität Bochum	Ruprecht-Karls-Universität Heidelberg	RWTH Aachen
Technische Universität Berlin	Technische Universität Darmstadt	Technische Universität Dresden
Technische Universität München	Universität Bayreuth	Universität Bielefeld
Universität Bremen	Universität des Saarlandes	Universität Konstanz
Universität Mannheim	Universität Regensburg	Universität Stuttgart
Universität Ulm	Universität zu Köln	

Excellence Cluster

Albert-Ludwigs-Universität Freiburg	Carl von Ossietzky Universität Oldenburg	Christian-Albrechts-Universität zu Kiel
Eberhard Karls Universität Tübingen	Freie Universität Berlin	Friedrich-Alexander-Universität Erlangen-Nürnberg
Georg-August-Universität Göttingen	Gottfried Wilhelm Leibniz Universität Hannover	Heinrich-Heine-Universität Düsseldorf
Humboldt-Universität zu Berlin	Johann Wolfgang Goethe-Universität Frankfurt am Main	Johannes Gutenberg-Universität Mainz
Justus-Liebig-Universität Gießen	Ludwig-Maximilians-Universität München	Rheinische Friedrich-Wilhelms-Universität Bonn
Ruhr-Universität Bochum	Ruprecht-Karls-Universität Heidelberg	RWTH Aachen
Technische Universität Berlin	Technische Universität Dresden	Technische Universität München
Universität Bielefeld	Universität Bremen	Universität des Saarlandes
Universität Hamburg	Universität Konstanz	Universität Stuttgart
Universität zu Köln	Universität zu Lübeck	Westfälische Wilhelms-Universität Münster
Wilhelm Leibniz Universität Hannover		

B Propensity score matching

The propensity score for university i can be defined as the probability of receiving funding from the Excellence Initiative (our treatment) given a set of observed pre-treatment covariates X: $p_i(X_i) = Pr(D = 1|X_i)$. In order to perform the propensity score matching, two important assumptions must be met. First, the Conditional Independence Assumption (CIA), or unconfoundedness, meaning that the potential outcomes of the treatment are independent of the treatment conditional on a set of observable pre-treatment covariates. In other words, the variables that determine the matching between treated and untreated university should predict the same probability of receiving funding before the treatment phase (Autio and Rannikko, 2016; Huber, 2021; Scandura, 2016). This condition can be relaxed, if we instead look at the average treatment in the case of the ATT, i.e. the average treatment effect of the treated subpopulation, this condition can be relaxed so that mean independence is sufficient (Autio and Rannikko, 2016; Heckman et al., 1997). Second, the Common Support Condition, which means that for each value conditioning on all the covariates there is a positive probability of being treated or untreated. It ensures that the vector of relevant covariates does not perfectly predict whether a university receives or does not receive the treatment (Huber, 2021; Scandura, 2016). Using the built-in function of the 'psmatch2' Stata module, we impose common support as treated observations whose propensity score is higher than the maximum or less than the minimum propensity score of the control observations are dropped (Leuven and Sianesi, 2003).

Based on the Stata module psmatch2 (Leuven and Sianesi, 2003) and in line with previous empirical studies (e.g. Almus and Czarnitzki, 2003; Scandura, 2016), we then estimate the propensity score (i.e. probability that a university receives funding from the Excellence Initiative conditional on a set of relevant characteristics), *p(X)*, through probit regression. We follow the practical guidance of Caliendo and Kopeinig (2008) for the selection of corresponding matching variables. In order to avoid a violation of the CIA, we match universities according to their observable characteristics prior to the first funding round of the Excellence Initiative (Caliendo and Kopeinig, 2008; Kaiser and Kuhn, 2012). Moreover, since our sample is rather small, we cannot include a very large set of independent variables, a procedure previous studies investigating the causal effect of the Excellence Initiative also adhere to (e.g. Buenstorf and Koenig, 2020). This holds particularly true, since our outcome variables, i.e. relative productivity scores in teaching and research, have already been calculated with different input and output variables.

Therefore, we focus on two main variables that are likely to influence both treatment assignment and outcome: the average share of third-party funding from the DFG in total third-party funding and the average number of academic staff from 2000 to 2005.[25] The share in DFG funding captures the previous experience in raising research-oriented funds, which, in turn, should increase the likelihood of receiving funding from the Excellence Initiative. In addition, it provides information about a university's strategic orientation towards research. The higher the ratio of DFG third-party funding, the more likely the university is to engage in basic research (Laudel, 2006). The second variable, in line with previous studies (inter alia, Buenstorf and Koenig, 2020; Gralka et al., 2019; Fandel, 2007), takes the human capital of universities into account. In addition to financial resources, human capital is also an essential driver of teaching as well as research success. For this reason, we use the average number of academic staff from 2000 to 2005 as a further matching variable.

[25] As a further robustness check, we also included more and alternative matching variables, such as a dummy variable capturing whether a university is a technical university or not, and the average number of students from 2000 to 2005. The corresponding results, also with respect to the DID approach (despite one case which is in section 6.2 discussed), remain robust and are shown in Appendix C.

Table 9 shows the corresponding results of the probit regression. Both matching variables influence the probability of receiving excellence funding from either of the three funding periods significantly. The pseudo R^2 underlines that the utilized variables have a relatively high explanatory power.

Table 9
Probit regression results

	Excellence Funding dummy
Average share of third-party funding from DFG	8.026***
	(2.465)
Average number of academic staff	0.002***
	(0.001)
Constant	-3.662***
	(0.871)
Observations	79
Pseudo R^2	0.458
LR	50.09***

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Unfortunately, we cannot formally test whether the CIA holds true,[26] yet we test for the balancing property, i.e. whether treatment and control group no longer differ significantly in terms of their observed characteristics after the matching (Kaiser and Kuhn, 2012; Vanino et al., 2019). Table 10 reports the corresponding results. While prior to the matching, the means between treated and control group differ significantly, after the matching, we cannot reject the Null of equal means. In general, the bias after matching for all considered covariates is reduced below the critical threshold of 25% (Vanino et al., 2019). Consequently, we argue that our matching procedure satisfies the balancing property in general and is therefore of good quality.[27]

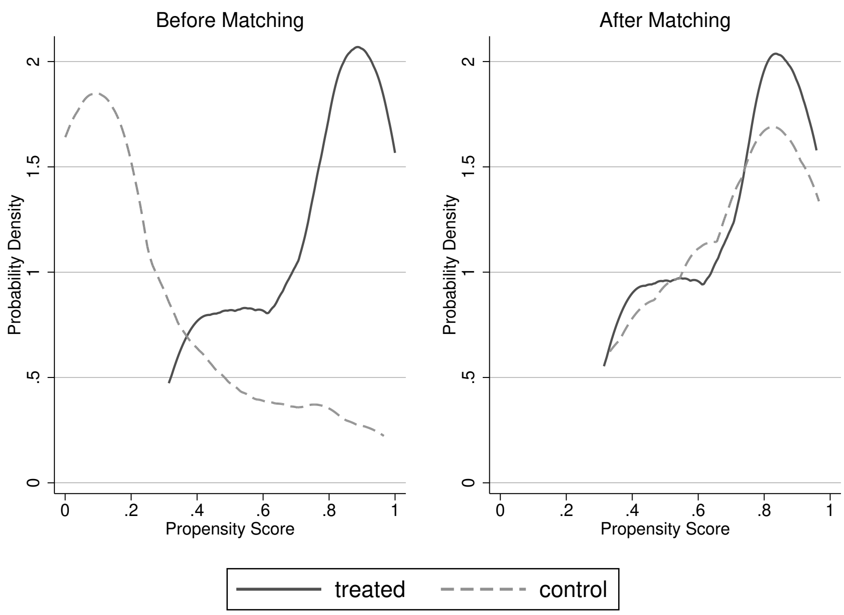
Table 10
Matching balancing test

		Mean		Bias	t-test	
Variable	Sample	Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	unmatched	0.359	0.224	140.9	6.32	0.000
	matched	0.356	0.364	-8.2	-0.52	0.603
Average number of academic staff	unmatched	850.83	329.65	121.4	5.32	0.000
	matched	694.9	673.07	5.1	0.33	0.745

[26] Nevertheless, the high explanatory power of our two matching variables as well as the overall robustness when using additional and/or alternative matching variables (see Appendix C) makes us confident to assume that the CIA is indeed satisfied.

[27] Moreover, due to the common support requirements we only lose six observations, representing less than 15% of the size of the treated group.

Fig. 7. Comparison of propensity score density with Excellence Initiative universities as treated



C Alternative Propensity Score Matching Approaches and Diff-in-Diff results (ATT of Excellence Initiative)

To further check the robustness of our results, we conduct alternative approaches for propensity score matching. First, we used the following alternative matching variables: (i.) the average share of third-party funding from the DFG in total third-party funding, (ii.) a dummy variable capturing whether a university is a technical university or not, and (iii.) the average number of students from 2000 to 2005. As described in section 6.1, we implemented in this context the nearest neighbour algorithm with replacement. Second, we used the same matching variables as indicated in section 6.1, but implemented the nearest neighbour algorithm with a "caliper" threshold of 0.09, which corresponds approximately to 0.25 times the standard deviation of the propensity scores recovered with the probit regression (Rosenbaum and Rubin, 1985). The corresponding results of these alternative matching processes are illustrated in Table 11 - 14.

As it can be seen in Table 11, the average share of third-party funding from the DFG as well as the size of the university, measured by the average number of students, both significantly influence the probability of receiving Excellence Funding from either of the three funding periods. However, whether a university is a technical university or not does not seem to significantly influence the probability of receiving the treatment. Moreover, with respect to the balancing property, in both cases, i.e. the matching with alternative variables (see Table 12) and the matching with a caliper threshold of 0.09 (see table 14), the differences between the treated and control group existing prior to the matching could be significantly reduced after the matching process.

Table 11
Probit regression results with alternative matching variables

	Excellence Funding dummy
Average share of third-party funding from DFG	8.069***
	(2.378)
Average number of students	0.0001***
	(0.00001)
Technical University dummy	0.089
	(0.467)
Constant	-3.339***
	(0.831)
Observations	79
Pseudo R^2	0.398
LR	43.49***

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 12
Matching balancing test with alternative matching variables

Variable	Sample	Mean		Bias	t-test	
		Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	Unmatched	0.359	0.224	141.2	6.33	0.000
	Matched	0.344	0.362	-18.4	-1.36	0.177
Average number of students	Unmatched	23119	10336	125.3	5.54	0.000
	Matched	19689	17099	25.4	1.16	0.250
Technical University dummy	Unmatched	0.122	0.158	-10.2	-0.46	0.650
	Matched	0.156	0.125	8.9	0.35	0.724

Table 13
Probit regression results with a caliper threshold of 0.09

	Excellence Funding dummy
Average share of third-party funding from DFG	8.026***
	(2.465)
Average number of academic staff	0.002***
	(0.001)
Constant	-3.661***
	(0.872)
Observations	79
Pseudo R^2	0.458
LR	50.09***

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 14
Matching balancing test (caliper threshold of 0.09)

		Mean		Bias	t-test	
Variable	Sample	Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	unmatched	0.359	0.224	141.2	6.33	0.000
	matched	0.356	0.365	-9.3	-0.60	0.553
Average number of academic staff	unmatched	850.83	330.31	121.3	5.32	0.000
	matched	694.9	670.9	5.6	0.35	0.724

Table 15
Robustness check: ATT of Excellence Initiative on research productivity (across all groups and all time periods)

	Dependent variable: Research Productivity					
	Original PSM		Alternative PSM (Alt. matching variables)		Alternative PSM (Caliper threshold of 0.09)	
	Setup 1	Setup 2	Setup 1	Setup 2	Setup 1	Setup 2
ATT	-0.213*	-0.129	-0.363**	-0.263*	-0.219**	-0.142
	(0.113)	(0.094)	(0.167)	(0.139)	(0.105)	(0.090)
Observations	867	867	884	884	901	901

Clustered standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes: 1. Setup (Inputs: academic staff, general expenditures); 2. Setup (Inputs: academic staff, general expenditures incl. third-party funding).

Table 16
Robustness check: ATT of Excellence Initiative on teaching productivity (across all groups and all time periods)

	Dependent variable: Teaching Productivity					
	Original PSM		Alternative PSM (Alt. matching variables)		Alternative PSM (Caliper threshold of 0.09)	
	Setup 1	Setup 2	Setup 1	Setup 2	Setup 1	Setup 2
ATT	0.057	0.033	0.138	0.126	0.070	0.052
	(0.072)	(0.071)	(0.099)	(0.101)	(0.065)	(0.065)
Observations	867	867	884	884	901	901

Clustered standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Notes: 1. Setup (Inputs: academic staff, general expenditures); 2. Setup (Inputs: academic staff, general expenditures incl. third-party funding).

D Alternative inputs for efficiency analysis and Diff-in-Diff results (ATT of Excellence Initiative)

As a further robustness check, we broke down the third-party funding (used in the 2. Setup) into more research- and teaching-specific funding. For the research oriented funding we used third-party funding coming from the DFG, while funding from the Quality Pact for Teaching (”Qualitätspakt Lehre”) from the Federal Ministery of Education and Research is used as a proxy for teaching oriented funding. Consequently, the inputs for the relative productivity analysis slightly differ between teaching and research. In the former case, we use the number of academic staff and the general expenditures including third-party funding coming from the Quality Pact for Teaching. While we also use the number of academic staff in the case of research, we additionally consider the general expenditures including third-party funding coming from the DFG. The corresponding results are presented in Table 17.

Table 17
Robustness Check: ATT of Excellence Initiative in general on teaching and research relative productivity (across all groups and all time periods) with alternative efficiency scores

	Dependent variable:					
	Teaching	Research	Teaching	Research	Teaching	Research
	Original PSM		Alternative PSM (Alternative matching variables)		Alternative PSM (Caliper threshold of 0.09)	
	3. Setup	3. Setup	3. Setup	3. Setup	3. Setup	3. Setup
ATT of Excellence Initiative	0.060 (0.086)	-0.141 (0.107)	0.127 (0.101)	-0.287* (0.157)	0.070 (0.076)	-0.153 (0.101)
Observations	867	867	884	884	901	901
Note:	3. Setup (Inputs for teaching: Acad. staff, General Expenditures incl. third-party funding from QPL Inputs for research: Acad. staff, General Expenditures incl. third-party funding from DFG) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01					

Moreover, we also forward our two main outputs by only two years, instead of three years (see Table 18). The corresponding results remain robust, despite slightly changes in the first setup of the research relative productivity, where the ATT of the Excellence Initiative becomes insignificant (p-value = 0.131).

Table 18
Robustness Check: ATT of Excellence Initiative in general on teaching and research relative productivity (across all groups and all time periods) with efficiency scores (where outputs are forwarded by two years)

	Dependent variable:			
	Teaching	Teaching	Research	Research
	Original PSM		Original PSM	
	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Excellence Initiative	0.140 (0.093)	0.125 (0.090)	-0.215 (0.143)	-0.135 (0.126)
Observations	918	918	918	918
Note:	1. Setup (Inputs: Acad. staff, General Expenditures) 2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01			

Furthermore, besides changing the timing of our outputs, we use a Gaussian filter of three as an additional robustness check (see Table 19). The corresponding results remain robust, despite the second setup of research efficiency, where we now find a significant influence.

Table 19
Robustness Check: ATT of Excellence Initiative in general on teaching and research efficiency (across all groups and all time periods) with Gaussian filter of three

	Dependent variable:			
	Teaching Efficiency	Teaching Efficiency	Research Efficiency	Research Efficiency
	Original PSM		Original PSM	
	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Excellence Initiative	0.031	0.010	-0.234**	-0.167**
	(0.073)	(0.071)	(0.107)	(0.081)
Observations	867	867	867	867
Note:	1. Setup (Inputs: Acad. staff, General Expenditures)			
	2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding)			
	Clustered standard erros			
	*p<0.1; **p<0.05; ***p<0.01			

E Propensity Score Matching (Treatment group: Future Concepts winner)

For the propensity score matching, we followed the same approach as described in Appendix B. The corresponding results of the probit regression are shown in Table 20. Similar to the results in Table 9, the average share of third-party funding from the DFG in total third-party funding and the average number of academic staff from 2000 to 2005 both significantly increase the probability of receiving funding from the Future Concepts funding line across all three funding periods. However, the pseudo R^2 is lower than in the case of the overall Excellence Initiative (i.e. all three funding lines), even though it is still reasonable high. The results for the balancing property are reported in Table 21. As can been seen, prior to the matching, the means between treated and control group differ significantly, whereas after the matching we cannot reject the Null of equal means. In addition, overall the bias after matching for all considered covariates is reduced below the critical threshold of 25% (Vanino et al., 2019), namely 15.1%. Hence, we are again confident that our matching approach satisfies the balancing property in general and is therefore of good quality.

Table 20

Probit regression results (Future Concepts funding)

	Future Concepts Funding dummy
Average share of third-party funding from DFG	4.368^{*}
	(2.305)
Average number of academic staff	0.001^{***}
	(0.0004)
Constant	-3.086^{***}
	(0.863)
Observations	79
Pseudo R^2	0.219
LR	16.17^{***}

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 21

Matching balancing test (Future Concepts funding)

		Mean		Bias	t-test	
Variable	Sample	Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	unmatched	0.364	0.279	80.8	2.58	0.012
	matched	0.364	0.352	10.9	0.40	0.695
Average number of academic staff	unmatched	1043	505.13	102.9	3.94	0.000
	matched	1043	1071.7	-5.5	-0.12	0.908

F Alternative Propensity Score Matching Approaches and Diff-in-Diff results (ATT of Future Concepts)

To further check the robustness of our results, we conducted alternative propensity score matching approaches. First, similar to the previous approach described in Appendix C, we used alternative matching variables, namely the average share of third-party funding from the DFG in total third-party funding, a dummy variable capturing whether a university is a technical university or not, and the average number of students from 2000 to 2005. The corresponding results of this matching approach are presented in Table 22 and Table 23.

Table 22
Probit regression results with alternative matching variables (Treatment group: Future Concept winners)

	Future Concepts Funding dummy
Average share of third-party funding from DFG	4.620*
	(2.653)
Average number of students	0.0001***
	(0.00001)
Technical University dummy	0.547
	(0.560)
Constant	-3.557***
	(1.100)
Observations	79
Pseudo R^2	0.229
LR	16.93***

Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 23
Matching balancing test with alternative matching variables (Treatment group: Future Concept winners)

Variable	Sample	Mean		Bias	t-test	
		Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	Unmatched	0.364	0.279	80.8	2.58	0.012
	Matched	0.355	0.345	9.9	0.34	0.737
Average number of students	Unmatched	27451	14712	109.5	3.90	0.000
	Matched	24076	23113	8.3	0.26	0.799
Technical University dummy	Unmatched	0.143	0.123	5.7	0.20	0.842
	Matched	0.167	0.333	-48.0	-0.92	0.368

Moreover, in line with previous studies (Buenstorf and Koenig, 2020), we also used universities that did not win the Future Concept, but the Graduate Schools and/or Clusters of Excellence funding line as a "natural" control group. The corresponding DID results are presented in Table 24 (ATT of Future Concepts on research efficiency) and in Table 25 (ATT of Future Concepts on teaching efficiency).

Table 24
Robustness Check: ATT of Future Concepts on research relative productivity (across all groups and all time periods)

	Dependent variable:					
	Research	Research	Research	Research	Research	Research
	Original PSM		Robustness Check with alternative PSM (Alternative matching variables)		Robustness Check with "natural" control group (Graduate school and Cluster of Excellence winners)	
	1. Setup	2. Setup	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Future Concepts	-0.278 (0.238)	-0.264 (0.244)	-0.153 (0.224)	-0.181 (0.230)	-0.293 (0.224)	-0.270 (0.233)
Observations	425	425	374	374	697	697
Note:	1. Setup (Inputs: Acad. staff, General Expenditures); 2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01					

Table 25
Robustness Check: ATT of Future Concepts on teaching relative productivity (across all groups and all time periods)

	Dependent variable:					
	Teaching	Teaching	Teaching	Teaching	Teaching	Teaching
	Original PSM		Robustness Check with alternative PSM (Alternative matching variables)		Robustness Check with "natural" control group (Graduate school and Cluster of Excellence winners)	
	1. Setup	2. Setup	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Future Concepts	0.059 (0.081)	0.056 (0.082)	0.109 (0.077)	0.089 (0.078)	0.075 (0.079)	0.069 (0.081)
Observations	425	425	374	374	697	697
Note:	1. Setup (Inputs: Acad. staff, General Expenditures); 2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01					

G Alternative inputs for relative productivity analysis and Diff-in-Diff results (ATT of Future Concepts)

As described in Appendix D, we also further differentiate the third-party funding (used in the 2. Setup) into more research- and teaching-specific funding. The corresponding results for the ATT of Future Concepts are presented in Table 17.

Table 26

Robustness Check: ATT of Future Concepts on teaching and research relative productivity (across all groups and all time periods) with alternative efficiency scores

	Dependent variable:					
	Teaching	Research	Teaching	Research	Teaching	Research
	Original PSM		Robustness Check with alternative PSM (Alternative matching variables)		Robustness Check with "natural" control group (Graduate school and Cluster of Excellence winners)	
	3. Setup	3. Setup	3. Setup	3. Setup	3. Setup	3. Setup
ATT of Future Concepts	0.024 (0.081)	-0.235 (0.243)	0.059 (0.467)	-0.130 (0.571)	0.047 (0.077)	-0.268 (0.250)
Observations	425	425	374	374	697	697
Note:	3. Setup (Inputs for teaching: Acad. staff, General Expenditures incl. third-party funding from QPL Inputs for research: Acad. staff, General Expenditures incl. third-party funding from DFG) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01					

Moreover, we also forward our two main outputs by only two years, instead of three years (see Table 27). The corresponding results remain robust, despite slightly changes in the first setup of the research relative productivity, where the ATT of the Excellence Initiative becomes insignificant (p-value = 0.131).

Table 27

Robustness Check: ATT of Future Concepts on teaching and research relative productivity (across all groups and all time periods) with efficiency scores (where outputs are forwarded by two years)

	Dependent variable:			
	Teaching	Teaching	Research	Research
	Original PSM		Original PSM	
	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Future Concepts	0.066 (0.077)	0.065 (0.075)	-0.380 (0.329)	-0.364 (0.339)
Observations	450	450	450	450
Note:	1. Setup (Inputs: Acad. staff, General Expenditures) 2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding) Clustered standard erros *p<0.1; **p<0.05; ***p<0.01			

Furthermore, besides changing the timing of our outputs, we use a Gaussian filter of three as an additional robustness check (see Table 28). The corresponding results remain robust, despite the second setup of research relative productivity, where we now find a significant influence.

Table 28

Robustness Check: ATT of Future Concepts on teaching and research relative productivity (across all groups and all time periods) with Gaussian filter of three

	Dependent variable:			
	Teaching	Teaching	Research	Research
	Original PSM		Original PSM	
	1. Setup	2. Setup	1. Setup	2. Setup
ATT of Future Concepts	0.052	0.052	-0.246	-0.226
	(0.074)	(0.075)	(0.182)	(0.189)
Observations	425	425	425	425
Note:	1. Setup (Inputs: Acad. staff, General Expenditures)			
	2. Setup (Inputs: Acad. staff, General Expenditures incl. third-party funding)			
	Clustered standard erros			
	*p<0.1; **p<0.05; ***p<0.01			

H Propensity Score Matching (Treatment group: Graduate Schools and Clusters of Excellence winner)

For the propensity score matching, we followed the same approach as described in Appendix C, because in the original matching approach (with two variables) the bias after matching is not reduced below the critical threshold of 25% (Vanino et al., 2019). Hence, our matching variables are: the average share of third-party funding from the DFG in total third-party funding, a dummy variable capturing whether a university is a technical university or not, and the average number of students from 2000 to 2005. The corresponding results of the probit regression are shown in Table 29. Similar to the results in Table 11, the average share of third-party funding from the DFG in total third-party funding as well as the average number of students both increase the probability of receiving funding from the Graduate Schools and/or Clusters of Excellence funding line. In general, the pseudo R^2 is also relatively high, indicating that the utilized variables have a relatively high explanatory power.

The results for the balancing property are reported in Table 30. As can been seen, prior to the matching, the means between treated and control group differ significantly, whereas after the matching we cannot reject the Null of equal means. In addition, overall the bias after matching for all considered covariates is reduced below the critical threshold of 25% (Vanino et al., 2019), namely 14.5%. Hence, we are again confident that our (adapted) matching approach satisfies the balancing property in general and is therefore of good quality.

Table 29

Probit regression results (Graduate school and Clusters of Excellence funding)

	Graduate Schools/Cluster Funding dummy
Average share of third-party funding from DFG	8.963***
	(2.791)
Average number of students	0.00004*
	(0.00002)
Technical University dummy	-0.046
	(0.551)
Constant	-3.544***
	(0.940)
Observations	65
Pseudo R^2	0.359
LR	31.65***

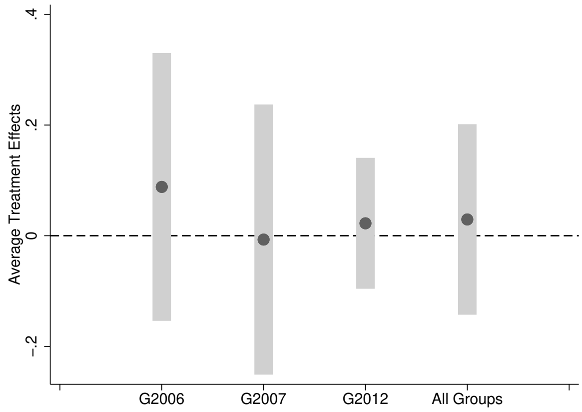
Notes: Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 30
Matching balancing test with alternative matching variables (Treatment group: Graduate school/Cluster of Excellence winners)

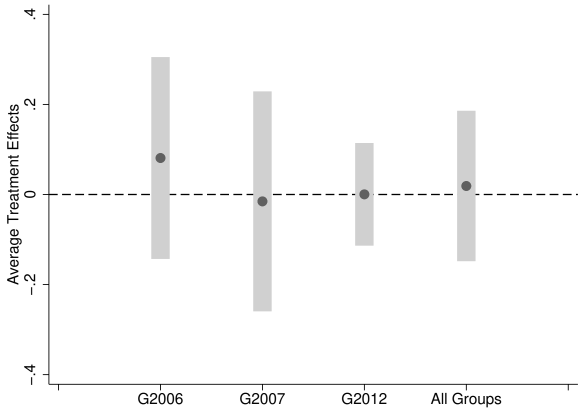
Variable	Sample	Mean		Bias	t-test	
		Treated	Control	Perc.	t-value	p-value
Average share of third-party funding from DFG	Unmatched	0.356	0.224	146.8	5.51	0.000
	Matched	0.344	0.352	-8.3	-0.54	0.595
Average number of students	Unmatched	20872	10336	109.0	4.41	0.000
	Matched	18839	17370	15.2	0.51	0.612
Technical University dummy	Unmatched	0.111	0.132	-6.2	-0.24	0.808
	Matched	0.136	0.182	-13.7	-0.40	0.689

I Dynamic event-time ATTs and the across-group ATTs (Treatment group: Graduate Schools and Clusters of Excellence winner)

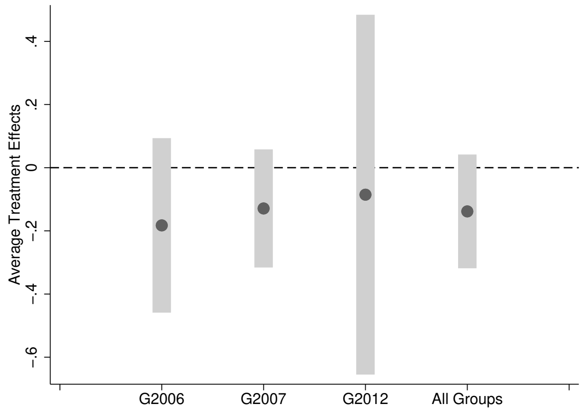
Fig. 8. Across-group ATTs of Graduate Schools and Clusters of Excellence



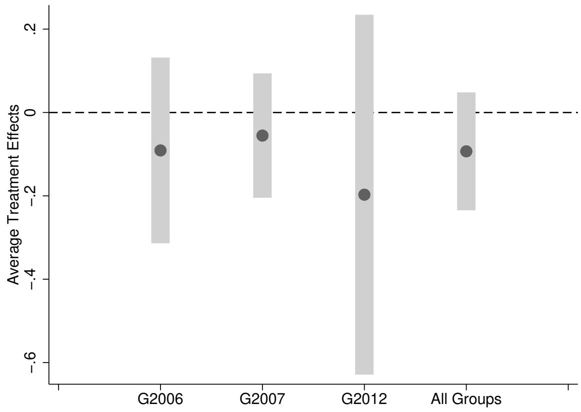
(a) Teaching – Setup 1



(b) Teaching – Setup 2



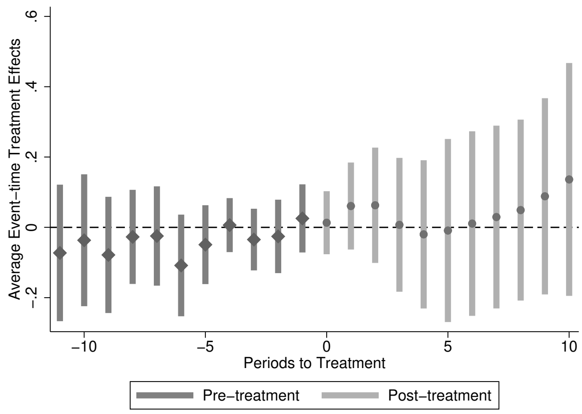
(c) Research – Setup 1



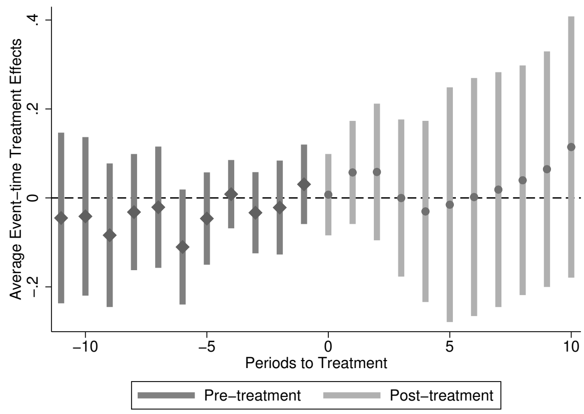
(d) Research – Setup 2

Note: The treatment group consists of the universities which won Graduate Schools and/or Clusters of Excellence in three rounds (2006, 2007, 2012), but did not win any Future Concepts. The control group is derived from PSM within the universities that did not win win any of the funding lines. Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with “number of academic staff” and “general expenditure” (2000-2016) in setup 1; with “number of academic staff”, “general expenditure” and “third-party funding” (2000-2016) in setup 2 as inputs. Both setups employ“number of graduates” (2003-2019) as output in teaching and “number of publications” (2003-2019) as output in research. The bars mark the 95% confidence interval.

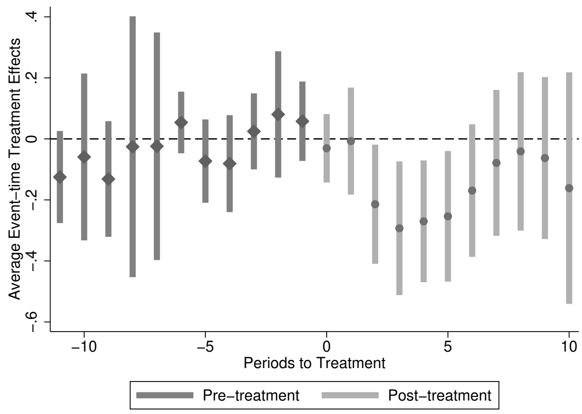
Fig. 10. Event-time ATTs of the Graduate Schools and Clusters of Excellence



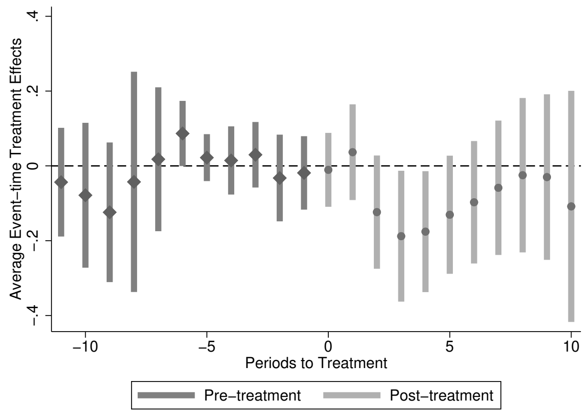
(a) Teaching – Setup 1



(b) Teaching – Setup 2



(c) Research – Setup 1



(d) Research – Setup 2

Note: The treatment group consists of the universities which won Graduate Schools and/or Clusters of Excellence in three rounds (2006, 2007, 2012), but did not win any Future Concepts. The control group is derived from PSM within the universities that did not win win any of the funding lines. Relative productivity scores in teaching and research are the dependent variables respectively. They are calculated with “number of academic staff” and “general expenditure” (2000-2016) in setup 1; with “number of academic staff”, “general expenditure” and “third-party funding” (2000-2016) in setup 2 as inputs. Both setups employ“number of graduates” (2003-2019) as output in teaching and “number of publications” (2003-2019) as output in research. The bars mark the 95% confidence interval.

References

Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

Abbott, M. and Doucouliagos, C. (2003). The efficiency of australian universities: a data envelopment analysis. *Economics of Education review*, 22(1):89–97.

Agasisti, T. (2009). Market forces and competition in university systems: theoretical reflections and empirical evidence from italy. *International Review of applied economics*, 23(4):463–483.

Agasisti, T. and Johnes, G. (2009). Beyond frontiers: comparing the efficiency of higher education decision-making units across more than one country. *Education economics*, 17(1):59–79.

Aghion, P., Dewatripont, M., Hoxby, C., Mas-Colell, A., and Sapir, A. (2010). The governance and performance of universities: evidence from europe and the us. *Economic policy*, 25(61):7–59.

Almus, M. and Czarnitzki, D. (2003). The effects of public r&d subsidies on firms' innovation activities: the case of eastern germany. *Journal of Business & Economic Statistics*, 21(2):226–236.

Ardiyono, S. K. (2022). Covid-19 pandemic, firms' responses, and unemployment in the asean-5. *Economic Analysis and Policy*, 76:337–372.

Arora, A., David, P. A., and Gambardella, A. (2000). Reputation and competence in publicly funded science: estimating the effects on research group productivity. In *The Economics and Econometrics of Innovation*, pages 141–176. Springer.

Arora, A. and Gambardella, A. (2005). The impact of nsf support for basic research in economics. *Annales d'Economie et de Statistique*, pages 91–117.

Ash, M. G. (2006). Bachelor of what, master of whom? the humboldt myth and historical transformations of higher education in german-speaking europe and the us 1. *European Journal of Education*, 41(2):245–267.

Auranen, O. and Nieminen, M. (2010). University research funding and publication performance—an international comparison. *Research policy*, 39(6):822–834.

Autio, E. and Rannikko, H. (2016). Retaining winners: Can policy boost high-growth entrepreneurship? *Research policy*, 45(1):42–55.

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9):1078–1092.

Barra, C. and Zotti, R. (2016). Measuring efficiency in higher education: an empirical study using a bootstrapped data envelopment analysis. *International Advances in Economic Research*, 22(1):11–33.

Başkaya, S. and Klumpp, M. (2014). International data envelopment analysis in higher education: how do institutional factors influence university efficiency. *Journal of Business and Economics*, 5(11):2085–2090.

Björk, B.-C. and Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of informetrics*, 7(4):914–923.

Bolli, T., Olivares, M., Bonaccorsi, A., Daraio, C., Aracil, A. G., and Lepori, B. (2016). The differential effects of competitive funding on the production frontier and the efficiency of universities. *Economics of Education Review*, 52:91–104.

Bolli, T. and Somogyi, F. (2011). Do competitively acquired funds induce universities to increase productivity? *Research Policy*, 40(1):136–147.

Bonaccorsi, A., Simar, D. C., et al. (2006). Advanced indicators of productivity of universities. an application of robust nonparametric methods to italian data, scientometrics, vol. 66, no. 2, 389-410.

Brinkman, P. T. and Leslie, L. L. (1986). Economies of scale in higher education: Sixty years of research. *The review of higher education*, 10(1):1–28.

Buenstorf, G. and Koenig, J. (2020). Interrelated funding streams in a multi-funder university system: Evidence from the german exzellenzinitiative. *Research Policy*, 49(3):103924.

Butler, L. (2003). Explaining australia's increased share of isi publications—the effects of a funding formula based on publication counts. *Research policy*, 32(1):143–155.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72.

Callaway, B. (2022). Difference-in-differences for policy evaluation. *arXiv preprint arXiv:2203.15646*.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Cantner, U., Krüger, J., and Hanusch, H. (2007). *Produktivitäts-und Effizienzanalyse: der nicht-parametrische Ansatz*. Springer-Verlag.

Carayol, N., Lanoë, M., et al. (2018). The impact and design of project-based funding in science: Lessons from the anr experience. Technical report.

Carayol, N. and Matt, M. (2006). Individual and collective determinants of academic scientists' productivity. *Information Economics and policy*, 18(1):55–72.

Caves, D. W., Christensen, L. R., and Diewert, W. E. (1982). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica: Journal of the Econometric Society*, pages 1393–1414.

Cherchye, L. and Abeele, P. V. (2005). On research efficiency: A micro-analysis of dutch university research in economics and business management. *Research policy*, 34(4):495–516.

Conard, M. J. and Conard, M. A. (2000). An analysis of academic reputation as perceived by consumers of higher education. *Journal of Marketing for Higher Education*, 9(4):69–80.

Corsini, A. (2022). Scientific networks and idex funding. *Doctoral dissertation*.

Dasgupta, P. and David, P. A. (1994). The new economics of science. *Research Policy*, 23(5):487–521.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

De Fraja, G. and Iossa, E. (2002). Competition among universities and the emergence of the elite institution. *Bulletin of economic Research*, 54(3):275–293.

De Groot, H., McMahon, W. W., and Volkwein, J. F. (1991). The cost structure of american research universities. *The Review of economics and statistics*, pages 424–431.

Del Rey, E. (2001). Teaching versus research: a model of state university competition. *Journal of Urban Economics*, 49(2):356–373.

Di Matteo, D. (2022). Does high-speed rail matter for tourism? evidence from italy. *Research in Transportation Business & Management*, page 100881.

Epple, D., Romano, R., and Sieg, H. (2006). Admission, tuition, and financial aid policies in the market for higher education. *Econometrica*, 74(4):885–928.

Fandel, G. (2007). On the performance of universities in north rhine-westphalia, germany: Government's redistribution of funds judged using dea efficiency measures. *European Journal of Operational Research*, 176(1):521–533.

Fischer, M., Kampkötter, P., et al. (2017). Effects of german universities' excellence initiative on ability sorting of students and perceptions of educational quality. *Journal of Institutional and Theoretical Economics*, 173(4):662.

Franck, E. P. and Schönfelder, B. (2000). On the role of competition in higher education—uses and abuses of the economic metaphor. *Schmalenbach Business Review*, 52(3):214–237.

Frietsch, R., Schubert, T., and Rothengatter, O. (2017). An analysis of the excellence initiative and its effects on the funded universities. Technical report, Studien zum deutschen Innovationssystem.

Frølich, N., Coate, K., Mignot-Gerard, S., and Knill, C. (2010). Einheit von forschung und lehre: implications for state funding of universities. *Higher Education Policy*, 23(2):195–211.

Gautier, A. and Wauthy, X. (2007). Teaching versus research: A multi-tasking approach to multi-department universities. *European Economic Review*, 51(2):273–295.

Gawellek, B. and Sunder, M. (2016). The german excellence initiative and efficiency change among universities, 2001-2011. Technical report, Working Paper.

Geuna, A. and Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva*, 41(4):277–304.

Gnewuch, M. and Wohlrabe, K. (2018). Super-efficiency of education institutions: an application to economics departments. *Education Economics*, 26(6):610–623.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Gralka, S. (2018). Persistent inefficiency in the higher education sector: evidence from germany. *Education Economics*, 26(4):373–392.

Gralka, S., Wohlrabe, K., and Bornmann, L. (2019). How to measure research efficiency in higher education? research grants vs. publication output. *Journal of Higher Education Policy and Management*, 41(3):322–341.

Grashof, N. and Kopka, A. (2022). Artificial intelligence and radical innovation: an opportunity for all companies? *Small Business Economics*, pages 1–27.

Gu, J. (2012). Spatial recruiting competition in chinese higher education system. *Higher Education*, 63(2):165–185.

Guerzoni, M. and Raiteri, E. (2015). Demand-side vs. supply-side technology policies: Hidden treatment and new empirical evidence on the policy mix. *Research Policy*, 44(3):726–747.

Gush, J., Jaffe, A., Larsen, V., and Laws, A. (2018). The effect of public funding on research output: the new zealand marsden fund. *New Zealand Economic Papers*, 52(2):227–248.

Hachmeister, C.-D., Harde, M. E., and Langer, M. F. (2007). Einflussfaktoren der studienentscheidung. *Eine empirische Studie von CHE und EINSTIEG. Arbeitspapier*, 95.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, pages 441–462.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Horstschräer, J. (2012). University rankings in action? the importance of rankings and an excellence competition for university choice of high-ability students. *Economics of Education Review*, 31(6):1162–1176.

Huber, M. (2021). Causal analysis: Impact evaluation and causal machine learning with applications in r.

Imboden, D., Lütjen-Drecoll, E., Bargmann, S., Casper, G., Gächter, S., Kratky, C., Nahrstedt, K., Bech Nosch, M., Pauss, F., and Scheidegger, D. (2016). Internationale expertenkommission zur evaluation der exzellenzinitiative–endbericht. *Internationale Expertenkommission Exzellenzinitiative.*

Jacob, B. A. and Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of public economics*, 95(9-10):1168–1177.

Johnes, G. and Johnes, J. (1993). Measuring the research performance of uk economics departments: an application of data envelopment analysis. *Oxford economic papers*, pages 332–347.

Kaiser, U. and Kuhn, J. M. (2012). Long-run effects of public–private research joint ventures: The case of the danish innovation consortia support scheme. *Research Policy*, 41(5):913–927.

Kempkes, G. and Pohl, C. (2008). Do institutions matter for university cost efficiency? evidence from germany. *CESifo economic studies*, 54(2):177–203.

Kettl, D. F. (2000). *The global public management revolution: A report on the transformation of governance*. Brookings Institution Press.

Krücken, G. (2021). Multiple competitions in higher education: a conceptual approach. *Innovation*, 23(2):163–181.

Krücken, G., Blümel, A., and Kloke, K. (2009). Towards organizational actorhood of universities: Occupational and organizational change within german university administrations.

Krücken, G. and Meier, F. (2006). Turning the university into an organizational actor. *Globalization and organization: World society and organizational change*, pages 241–257.

Laudel, G. (2006). The art of getting funded: how scientists adapt to their funding conditions. *Science and Public Policy*, 33(7):489–504.

Lehmann, E. E., Meoli, M., Paleari, S., and Stockinger, S. A. (2018). Approaching effects of the economic crisis on university efficiency: A comparative study of germany and italy. *Eurasian Business Review*, 8(1):37–54.

Leusin, M. E. (2022). The development of al in multinational enterprises-effects upon technological trajectories and innovation performance.

Leuven, E. and Sianesi, B. (2003). Psmatch2: Stata module to perform full mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.

Liu, J. S., Lu, L. Y., Lu, W.-M., and Lin, B. J. (2013). A survey of dea applications. *Omega*, 41(5):893–902.

Menter, M., Lehmann, E. E., and Klarl, T. (2018). In search of excellence: A case study of the first excellence initiative of germany. *Journal of Business Economics*, 88(9):1105–1132.

Mergele, L. and Winkelmayer, F. (2021). The relevance of the german excellence initiative for inequality in university funding. *Higher Education Policy*, pages 1–19.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Möller, T., Schmidt, M., and Hornbostel, S. (2016). Assessing the effects of the german excellence initiative with bibliometric methods. *Scientometrics*, 109(3):2217–2239.

Morgan, J. (2016). Germany: an alternative route to excellence. *Times Higher Education. https://www. timeshighereducation. com/features/germany-an-alternative-route-to-excellence.*

Musselin, C. (2018). New forms of competition in higher education. *Socio-Economic Review*, 16(3):657–683.

Obermeit, K. (2012). Students' choice of universities in germany: structure, factors and information sources used. *Journal of Marketing for higher education*, 22(2):206–230.

Olivares, M. and Wetzel, H. (2014). Competing in the higher education market: empirical evidence for economies of scale and scope in german higher education institutions. *CESifo Economic Studies*, 60(4):653–680.

Pollitt, C. (1990). *Managerialism and the public services: The Anglo-American experience*. Blackwell.

Quiroga-Martínez, F., Fernández-Vázquez, E., and Alberto, C. L. (2018). Efficiency in public higher education on argentina 2004–2013: Institutional decisions and university-specific effects. *Latin American Economic Review*, 27(1):1–18.

Rios-Avila, F., Sant'Anna, P., and Callaway, B. (2022). Csdid: Stata module for the estimation of difference-in-difference models with multiple time periods.

Robst, J. (2001). Cost efficiency in public higher education institutions. *The Journal of higher education*, 72(6):730–750.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Discussion of "on state education statistics": A difficulty with regression analyses of regional test score averages. *Journal of Educational Statistics*, 10(4):326–333.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Scandura, A. (2016). University–industry collaboration and firms' r&d effort. *Research Policy*, 45(9):1907–1922.

Shleifer, A. (1985). A theory of yardstick competition. *The RAND journal of Economics*, pages 319–327.

Smith, S., Ward, V., and House, A. (2011). 'impact'in the proposals for the uk's research excellence framework: Shifting the boundaries of academic autonomy. *Research Policy*, 40(10):1369–1379.

Strehl, F., Reisinger, S., and Kalatschan, M. (2007). Funding systems and their effects on higher education systems.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Van der Meulen, B. (1998). Science policies as principal–agent games: Institutionalization and path dependency in the relation between government and science. *Research policy*, 27(4):397–414.

Vanino, E., Roper, S., and Becker, B. (2019). Knowledge to money: Assessing the business performance effects of publicly-funded r&d grants. *Research policy*, 48(7):1714–1737.

Wiener, M., Maresch, D., and Breitenecker, R. J. (2020). The shift towards entrepreneurial universities and the relevance of third-party funding of business and economics units in austria: A research note. *Review of Managerial Science*, 14(2):345–363.

Willich, J., Buck, D., Heine, C., and Sommer, D. (2011). Studienanfänger im wintersemester 2009/10: Wege zum studium, studien-und hochschulwahl, situation bei studienbeginn [first-year students in the winter semester 2009/10: Avenues to study, choice of field of study and university, situation when starting to study]. *Hannover: HIS Hochschul-Informations-System GmbH.*

Winston, G. C. (1999). Subsidies, hierarchy and peers: The awkward economics of higher education. *Journal of economic perspectives*, 13(1):13–36.

Wohlrabe, K., Bornmann, L., Gralka, S., and de Moya Anegon, F. (2019). Wie effizient forschen universitäten in deutschland, deren zukunftskonzepte im rahmen der exzellenzinitiative ausgezeichnet wurden? ein empirischer vergleich von input-und output-daten. *Zeitschrift für Evaluation*, 18(1):9–27.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, 98(3):309–319.