University of Zurich^UZH

Department of Business Administration

UZH Business Working Paper Series

Working Paper No. 367

Do research training groups operate at optimal size?

Kerstin Pull, Birgit Pferdmenges and Uschi Backes-Gellner

February 2017

University of Zurich, Plattenstrasse 14, CH-8053 Zurich,
http://www.business.uzh.ch/forschung/wps.html

Published in *Schmalenbach Business Review* (2017). The final publication is available at Springer via http://dx.doi.org/10.1007/s41464-017-0029-8.

UZH Business Working Paper Series

Contact Details

Prof. Dr. Kerstin Pull (corresponding author)

University of Tuebingen

School of Business and Economics

P_O@uni-tuebingen.de

Dr. Birgit Pferdmenges

University of Tuebingen

School of Business and Economics

Prof. Dr. Uschi Backes-Gellner

University of Zurich

Department of Business Adminstration

Do research training groups operate at optimal size?

Kerstin Pull[1]; Birgit Pferdmenges[2]; Uschi Backes-Gellner[3]

JEL Codes: I23, I21, I28

Abstract: In this paper, we analyze whether structured PhD programs operate at optimal size and whether there are differences between different disciplinary fields. Theoretically, we postulate that the relation between the size of a PhD program and program performance is hump shaped. For our empirical analysis, we use hand-collected data on 86 Research Training Groups (RTGs) funded by the German Research Foundation (*DFG*). As performance indicators, we use (a) the number of completed PhDs and (b) the number of publications by RTG students (PhD students and postdoctoral researchers). Applying DEA with constant and variable returns to scale, we find that the optimal team size varies between 10 and 16 RTG students in the humanities and social sciences. In contrast, our empirical analysis does not uncover a systematic relation between size and performance for RTGs in the natural and life sciences.

Forthcoming in *Schmalenbach Business Review* (2017). The final publication is available at Springer via http://dx.doi.org/10.1007/s41464-017-0029-8.

[1] School of Business and Economics, University of Tuebingen, Email: P_O@uni-tuebingen.de, Tel.: +49 (0)7071 29 78186, Fax : +49 (0)7071 29 5077.
[2] School of Business and Economics, University of Tuebingen.
[3] Department of Business Administration, University of Zurich.

1. Introduction

PhD education is being reformed in many countries (see Powell and Green 2007; Sadlak 2004), increasingly turning away from an individual student-supervisor relationship and toward structured programs. Until now, only very few studies have analyzed the determinants of PhD program performance. Regarding program performance in terms of PhD completion, Pull, Pferdmenges and Backes-Gellner (2016) study the effect of group composition with respect to the diversity of the fields of study and cultural backgrounds of PhD students. Regarding program performance in terms of publication output, Bedeian et al. (2010) analyze the effect of a program's prestige on researchers' later publication success, and Breuninger, Pull and Pferdmenges (2012) study the link between PhD student publication output and the publication output of their supervisors. In our paper, we measure program performance in terms of publication output *and* PhD completion rates and we concentrate on the *size* of PhD groups as a potential determinant of program performance. Specifically, we ask if PhD programs operate at an optimal size and whether there are differences between different disciplinary fields.

Whereas size effects in academia have been studied repeatedly at the department, school and university levels (for recent overviews, see Albers 2015 and Clermont, Dirksen, and Dyckhoff 2015), there is hardly any literature on potential size effects at the PhD program level. We are aware of only one study, the study by Bowen and Rudenstine (1992) who, however, only calculate and descriptively compare the doctoral completion rates of comparatively larger and comparatively smaller PhD programs without further analyzing the differences. Also, to the best of our knowledge, it has not yet been studied whether size effects in PhD education differ between different disciplinary fields, despite the fact that the literature largely agrees on the importance of acknowledging the distinct collaborative practices in different disciplinary fields. In our paper, we distinguish between the natural and life sciences on the one hand and the humanities and social sciences on the other. It goes without saying that these two disciplinary areas cannot capture the complexity of disciplinary fields in academia and that the sub-disciplines within the areas remain rather diverse. However, as, e.g., Pull, Pferdmenges and Backes-Gellner (2016) argue, the two broad disciplinary areas may well serve as proxies for two profoundly different processes of knowledge production — characterized by, among others, differing degrees of codification and specialization and by varying degrees to which interdependent tasks and group work are the norm rather the exception.

From a theoretical perspective, we argue that there is good reason to believe that the relation between the size of a PhD program and its performance is hump shaped. That is, we argue that program performance first increases with size until an internal optimum has been reached and then decreases with size. In light of the many differences between the natural and life sciences on the one hand and the humanities and social sciences on the other, we further explore whether there are differences in the optimal program sizes of the two disciplinary areas.

For our empirical analysis, we hand-collected data from the reports of 86 Research Training Groups (RTGs) funded by the German Research Foundation (*DFG*). We use the number of completed PhDs and the number of RTG students' publications as performance indicators for the RTGs. Whereas completed PhDs represent an obvious measure of RTG performance, we also regard the number of publications to account for the fact that RTGs also comprise postdoctoral researchers and RTGs were explicitly established to train the "next generation of researchers" requiring a certain scientific visibility. Taking more than one performance dimension into account, we use data envelopment analysis (DEA) to comparatively assess the relative performance of RTGs within the two disciplinary fields.

Regarding completed PhDs and publication outputs as output dimensions and using DEA with constant and variable returns to scale, our empirical results suggest that the optimal team size varies between approximately 10 and 16 students in the humanities and social sciences. In contrast, our empirical analysis does not uncover a systematic relation between size and performance for RTGs in the natural and life sciences – one potential explanation being that the long-standing tradition of working in teams in the natural and life sciences has resulted in more informed applications and funding decisions, with essentially only those RTGs that operate at (almost) optimal size being funded.

The remainder of the paper is organized as follows: Section 2 reviews the relevant literature on (a) the link between size and performance in academia and (b) the differences between disciplinary areas. Section 3 introduces the data and method, section 4 describes the results, section 5 displays a few robustness checks and section 6 provides a discussion of our results. Section 7 discusses the implications and limitations of our research.

2. Literature

2.1 Size and performance

From a *theoretical* perspective, the link between the size of a research entity and its performance is not clear. On the positive side, larger groups might profit from synergies or joint resource utilization. On the negative side, larger groups might suffer from reduced flexibility resulting from a larger degree of bureaucratization (see Brown 1996; Kyvik 1995; Laband and Tollison 2000). Because there is good reason to believe that the additional benefits of an increasing group size (in terms of synergies and joint resource utilization) become smaller and smaller the more group size is increased (i.e., we assume decreasing marginal benefits of an increasing group size) and that the additional costs of an increasing group size (in terms of reduced flexibility and bureaucratization) become larger and larger the more group size is increased (i.e., we assume increasing marginal costs of an increasing group size), we argue that there is an interior optimum for the size of a research group and that the relation between research group size and performance is hump shaped.

However, the *empirical* evidence on the link between size and performance in academia is mixed (for recent overviews, see Albers 2015 and Clermont, Dirksen, and Dyckhoff 2015). Concerning publication performance, e.g., Carayol and Matt (2004), Verbree et al. (2015), van der Wal et al. (2009) and Brandt and Schubert (2013) find a *negative relation* between group size and publication performance. In contrast, Kyvik (1995), Cohen (1981), Johnes and Johnes (1995), Bonaccorsi and Daraio (2005), Ahn, Dyckhoff and Gilles (2007) and Dyckhoff, Rassenhövel and Sandfort (2009) find *no significant relation* between size and publication performance.[4] Lastly, Albers (2015) finds a positive relation between size and publication performance, and Cohn, Rhine and Santos (1989), De Groot, McMahon and Volkwein (1991), Laband and Lentz (2003), as well as Lloyd, Morgan and Williams (1993) who also include teaching outcomes, do likewise find a positive link between size and performance.

For PhD programs, there is hardly any literature on potential size effects. Bowen and Rudenstine (1992) simply calculate and compare the doctoral completion rates of comparatively larger (Berkeley, Chicago, Columbia) and comparatively smaller PhD programs (Cornell, Harvard, Princeton, Stanford)

[4] When including further performance dimensions, the latter, however, do find a significant link between size and performance.

across different disciplines (English, History, Political Science, Economics, Mathematics, Physics) and find smaller programs to have higher rates than larger programs. It is unclear, however, whether the measured differences are statistically significant, whether there are disciplinary differences and whether their results can be generalized beyond the highly selective programs studied.

2.2 Disciplinary Differences

As Pull, Pferdmenges and Backes-Gellner (2016) argue in detail, different disciplinary areas are characterized by distinct collaborative practices. In our paper, we follow Snow (1964) and his famous work on "the two cultures," and distinguish the natural and life sciences from the humanities and social sciences.

One first important difference between the two disciplinary areas under consideration concerns the fact that the humanities and social sciences are less paradigmatic (see Biglan 1973). Whereas in the humanities and social sciences, there is a plurality of theoretical and methodical approaches (see Wanner, Lewis and Gregorio 1981: 249), the natural sciences in particular are often dominated by one central research paradigm and hence less open to different methodologies and competing theoretical explanations (see Nuijten 2011; Biglan 1973). Further, knowledge in the humanities and social sciences is codified to a lower degree than knowledge in the Natural Sciences (see Audretsch, Lehmann and Warning 2004), and as a result, implicit and tacit knowledge is more important. In addition, graduate education in the humanities and social sciences is broader and less specialized (see Audretsch, Lehmann and Warning 2004; Hagstrom 1964), and research projects are less narrowly defined (see Hagstrom 1964). Moreover, research projects in the humanities and social sciences are often culture specific and follow a more “interpretative approach” to research (see Stanford University 2014). Further, in the humanities and social sciences, PhD students cannot rely on a quasi-universal language (such as “mathematics”). Lastly and most importantly for our research question, task interdependence in the two disciplinary areas differs distinctively. Unlike PhD students in the humanities and social sciences, PhD students in the natural and life sciences often rely on the cooperativeness of others in their research (see Warning, 2004; Knorr-Cetina, 1992), rendering cooperation not a choice but a must (see Breneman, 1976; Stephan, 1996; Wanner, Lewis and Gregorio, 1981).

The above described dissimilarities lead to differences in the scientific production technology, which is mirrored in the qualification process of young researchers. In the humanities and social sciences, research is more often conducted “in isolation” (see Black and Stephan 2010; Gellert 1993), and resource requirements are lower (see Stephan 1996; Wanner, Lewis and Gregorio 1981). In contrast, research projects in the natural and life sciences often require high computational capacity to undertake extensive simulations and complex laboratory experiments. As a result, scientists in the natural and life sciences have a long tradition of working together in teams and larger research units (see Black and Stephan 2010; Warning 2004; Knorr-Cetina 1992).

Empirically, only little is known about the potentially differential success of PhD programs in different disciplinary fields. Descriptively, Bowen and Rudenstine (1992) find doctoral completion rates in the natural sciences to be considerably higher than those in the humanities and in the social sciences. In contrast, Unger, Pull and Backes-Gellner (2010) find that the average doctoral completion rate in PhD programs belonging to the humanities and social sciences is almost as high as that in the natural and life sciences. Counting all types of publication outputs (monographs, editorships, journal articles, chapters in edited books, conference proceedings, discussion papers, published abstracts, and reviews) and adjusting for the number of co-authors, Unger, Pull and Backes-Gellner (2010) find that the publication output in PhD programs from the humanities and social sciences is considerably larger than that in programs from the natural and life sciences. Referring to the descriptive results from Bowen and Rudenstine (1992), Main (2014) analyzes the potential determinants of the comparatively lower doctoral completion rates and comparatively longer time to degree completion in the humanities and in the humanistic social sciences but does not account for program size.

Concerning potentially different size effects, Bowen and Rudenstine (1992) find the size of PhD programs to be negatively related to performance – irrespective of the disciplinary field that was analyzed (natural sciences, humanities, social sciences). Cherchye, Vanden Abeele (2005) study size effects at a sub-disciplinary level (i.e., at the level of "specialization areas") and find size effects to differ between specialization areas. However, the authors analyze their data at the university level and not at the level of PhD programs.

In conclusion, there is not much knowledge on the optimal size of PhD programs in general and whether the optimal size differs between different disciplinary areas.

3. Data, method and variables

3.1 Data

Our empirical analysis is based on a data set of 86 RTGs funded by the German Research Foundation. The German Research Foundation established RTGs as a new form of governance for PhD education in Germany in the early 1990s. RTGs are run by a group of cooperating researchers and include a study program for the PhD students, who are supervised by a team of senior researchers. The study program is compulsory for RTG students and is held to provide them with methodological skills and specialized knowledge in a particular field of research. The German Research Foundation grants fellowships to RTG students as well as funds for travel expenses and equipment. Until March 2003, a grant consisted of an initial funding for a period of three years that could be renewed twice. Since April 2003, a grant consists of a funding for 4.5 years that can be renewed only once.

Our data were collected manually from the RTG reports within a multi-year project in which the German Research Foundation temporally granted us access to their files. Our hand-collected data set comprises information on 86 RTGs[5] – 28 from the humanities and social sciences and 58 from the natural and life sciences – with 2,086 PhD students and postdoctoral researchers in total, representing a full sample of all RTGs from the humanities and social sciences and from the natural and life sciences funded by the German Research Foundation that were in their second funding period and had submitted an application for a third funding period between October 2004 and October 2006. The sub-disciplines in our data set are quite diverse, comprising crystallography as well as ethnology and paleontology as well as immunology.

When collecting the data, we chose not to sample RTGs that were in their first funding period and that had submitted an application for a second funding period because applications for the second funding period are submitted at a point in time when there is little performance to be reported (i.e., few doctoral theses completed and few publications). In contrast, when applying for a third funding period, RTGs that had already been funded for one full first funding period and that were in the mid-

[5] All of the RTGs are based at German universities. Four of them are located at more than one German university ("*Gemeinschaftskollegs*"), 20 are part of an international cooperation network ("*internationale Kollegs*").

dle of the second funding period (a) had performance data on which to report and (b) had an incentive to fully report their performance data because this would clearly increase their chances to be funded in the third funding period. Further, we did not include RTGs that reported on their third funding period because these final reports apparently often contain incomplete information on RTG performance. Likewise, we did not include RTGs that completed only one or two funding periods and did not apply for a second or third one because the respective RTGs – again – would not have an incentive to fully report their performance.

Since we have a *full* sample of all applications from RTGs that were in their second funding period, it also includes the unsuccessful applications. A total of 17 applications in our sample were unsuccessful, including five out of 28 applications from the humanities and social sciences and twelve out of 58 applications from the natural and life sciences.

3.2 Method

To analyze size effects at the PhD program level, we use data envelopment analysis (DEA), as developed by Charnes, Cooper and Rhodes (1978). DEA is a nonparametric method for estimating frontier production functions with multiple inputs and outputs. It is used to measure the relative (rather than absolute) efficiency of decision-making units (or DMUs), in our case, PhD programs. It does so by comparing their inputs and outputs without imposing any prices or weights. DMUs for which none of the outputs can be further increased (or, likewise, none of the inputs can be further decreased) without decreasing one of the other outputs or increasing one of the remaining inputs are defined to be 100% efficient in relation to the other DMUs in the data (see, e.g., Dyckhoff et al. 2013).

With the help of DEA, we can simultaneously account for different (and differently sized) input and output dimensions without any a priori imposed weighting factors. DEA measures relative efficiency by assuming research group-specific weighting factors in the most favorable way for each research group. This advantage of DEA opened up a wide field of applications in higher education (see Warning 2007: 175ff. for an overview). However, most of the analyses undertaken so far are at the university level (see e.g., Abbott and Doucouliagos 2003, Athanassopoulus and Shale 1997, Fandel 2007, McMillan and Datta 1998, Nazarko and Saparauskas 2014, Ng and Li 2000 and Warning 2004). Some are at the level of departments or smaller research groups (Groot and Garcìa-Valderrama 2006, Korhonen, Tainio and Wallenius 2001). Clermont, Dirksen and Dyckhoff (2015), Dyckhoff, Rassenhövel and Sandfort (2009) and Dyckhoff et al. (2013) use DEA to assess the relative efficiency of Business Schools. Concerning PhD education, we are aware of only one study using DEA, the study by Unger, Pull and Backes-Gellner (2010). Unger, Pull and Backes-Gellner (2010) distinguish between different disciplinary fields, but they do not analyze potential size effects. Dyckhoff, Rassenhövel and Sandfort (2009) and Clermont, Dirksen and Dyckhoff (2015), in contrast, do analyze size effects by applying DEA models with constant and variable returns to scale and by calculating scale efficiencies; however, they do so not with respect to PhD programs but with respect to Business Schools.

For the RTGs under consideration, we use the numbers of PhD and postdoctoral positions as inputs and the number of completed PhDs and the number of publications as outputs (see Unger, Pull, and Backes-Gellner 2010 for an analogous procedure). The number of completed PhDs is an obvious measure of RTG output, and the number of RTG students publications is added to account for RTGs having been established to train the “next generation of researchers” requiring a certain scientific visibility. Acknowledging differing modes of publication between the disciplinary fields (see Dundar and Lewis 1998; Stephan 1996; Unger, Pull and Backes-Gellner 2010), we counted all types of publications: monographs, editorships, journal articles, book sections in edited books, conference proceedings, discussion papers, published abstracts, and reviews. In light of the fact that publication patterns also vary among sub-disciplines, we refrained from imposing any quality weighting of the different publication outputs because in some sub-disciplines conference, proceedings might be regarded as a very important research output whereas in others, they might be regarded only as an intermediate output. Likewise, in some sub-disciplines, monographs and book sections might be important, whereas in others, only journal publications or even only publications in certain journals might be regarded as countable outputs. In addition, we refrained from quality adjusting journal article publications because we are not aware of any established and comprehensive ranking of journals across (sub-)disciplines. Given varying citation patterns, impact factor-based weightings might also be considered problematic. What we did is adjust publication outputs according to the number of co-authors and allocate a fraction of 1/n to each author (see Egghe, Rousseau and Hooydonk 2000).

To determine an efficiency measure, DEA classifies RTGs with a comparatively favorable input-output ratio to be efficient and calculates the level of relative inefficiency for the remaining RTGs. Thus, DEA identifies potential for improvement for the relatively inefficient RTGs resulting from the comparison with the input-output structures of the efficient reference units. We calculate an output-oriented DEA model, because the size of an RTG is fixed in the short term, whereas the outputs can be influenced. That is, we view the RTGs as aiming to maximize their outputs with given resources.

In search of scale effects, we employ a model with constant returns to scale (CCR) and one with variable returns to scale (BCC). The CCR model assumes that (a) publication activities and the completion of doctoral degrees are proportional to the number of young scientists and (b) average costs are independent of the output produced so there is no optimal size for RTGs (Charnes, Cooper and Rhodes, 1978: 437). In contrast, the BCC model captures possible scale effects (Banker, Charnes and Cooper, 1984: 1086). If an RTG is efficient in both models, CCR and BCC, it works at optimal size. An RTG that is efficient in the BCC model but not in the CCR model does not produce at optimal size – it is locally technically efficient but not globally technically efficient (see Cooper, Seiford and Tone 2006: 140). Whether an RTG operates at optimal size or can increase its efficiency by increasing or decreasing its size can be determined by calculating its scale efficiency. Scale efficiency is defined as the ratio of the CCR efficiency to the BCC efficiency and has an optimum of 1 (see Cooper, Seiford and Tone 2006: 140f.). An RTG is scale efficient if modifications in size lead to lower efficiency grades.

3.3 Descriptives

Table 1 displays the descriptives of the input and output variables used for the DEA – separately for the humanities and social sciences (Panel a) and for the natural and life sciences.

Table 1: Input and output data for the DEA

Panel (a): RTGs in the humanities and social sciences

	Min	Max	Mean	Std. dev.
No. of PhD positions	6.17	23.93	13.68	4.21
No. of postdoctoral positions	0	2.75	0.88	0.78
No. of publications per year	2.18	38.12	14.36	8.13
No. of completed PhDs per year	0	5.48	1.84	1.41

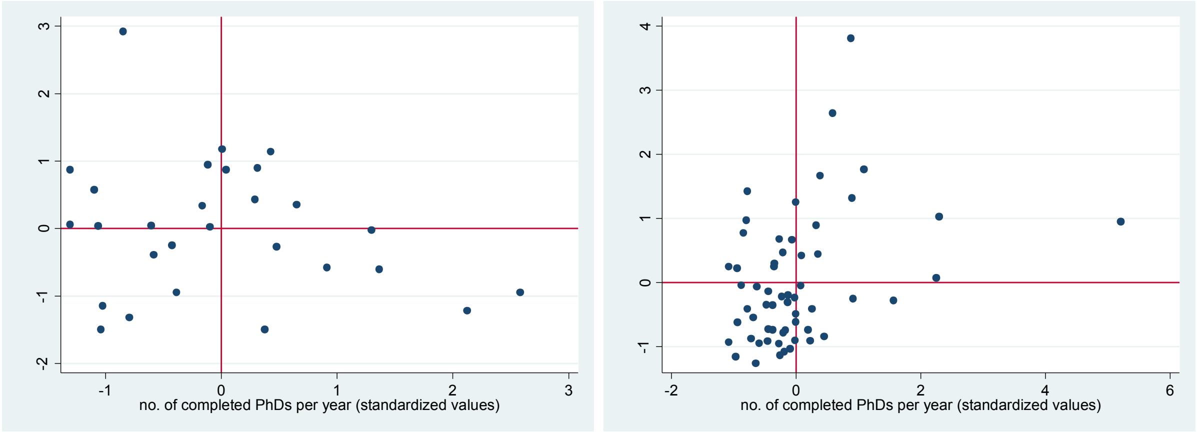
Panel (b): RTGs in the natural and life sciences

	Min	Max	Mean	Std. dev.
No. of PhD positions	4.86	42.62	16.01	7.63
No. of postdoctoral positions	0	6.73	1.20	1.33
No. of publications per year	0.19	28.17	7.15	5.52
No. of completed PhDs per year	0	15.75	2.70	2.50

Source: Own data.

Whereas for the humanities and social sciences, the two output dimensions (number of publications per year and number of completed PhDs per year) are not related, there is a positive (r=0.347) and statistically highly significant ($p < 0.01$) relation between the two dimensions in the natural and life sciences – potentially reflecting the latter's stronger tradition of publication-based PhDs. Figure 1 plots the realized outputs at RTG level separately for the humanities and social sciences (Panel a) and for the natural and life sciences (Panel b). We z-standardized the values of the two output dimensions by setting the means equal to zero and standard deviations equal to one (see Hamilton 2006: 331). On the horizontal axis, we plot the (standardized) number of completed PhDs per year, and on the vertical axis, we plot the (standardized) number of publications per year. It can be seen that in both disciplinary areas, some RTGs appear to be rather concentrated on the number of publications, whereas others appear to be rather concentrated on PhD completion rates – making data envelopment analysis the method of choice for assessing their comparative performance.

Figure 1: Plotting RTGs with their (standardized) number of completed PhDs per year (x-axis) and their (standardized) number of publications per year (y-axis)



Panel (a): Humanities and social sciences

Panel (b): Natural and life sciences

Source: Own data.

4. Results

4.1 Humanities and social sciences

In the CCR model, four out of 28 RTGs reach an efficiency value of 100%. Whereas two of these four RTGs focus on publications in particular (at varying numbers of completed PhDs), one of the RTGs concentrates on PhD completion while being rather weak in publications. Only the fourth RTG is successful in both performance dimensions. Looking at the remaining 24 RTGs, there appears to be considerable scope for efficiency improvement because they reach an average efficiency of only 53.26%. Because the weights are determined endogenously in the DEA and because each of the RTGs is assigned an individually optimal weighting vector, a poor rating for an RTG is not due to an unfavorable determination of the weights.

When calculating the BCC model, that is, when allowing for variable returns to scale, we find six efficient RTGs (100% efficiency level) and an average efficiency of 75.37% (CCR: 59.93%). The least efficient RTG raises its efficiency from 19.77% in the CCR model to 35.61% in the BCC model. Table 2 displays the descriptive statistics for the efficiency values. It is plausible that the efficiency values and the number of efficient units increase from the CCR model to the BCC model. Whereas inefficiencies in the CCR model can be caused by inefficient resource utilization in the production process as well as by a non-optimal size of an RTG, inefficiencies in the BCC model are caused only by inefficient resource utilization. Thus, the efficiency value of an RTG in the BCC model cannot be lower than the efficiency value in the CCR model.

Table 2: Efficiency values in the Humanities and Social Sciences

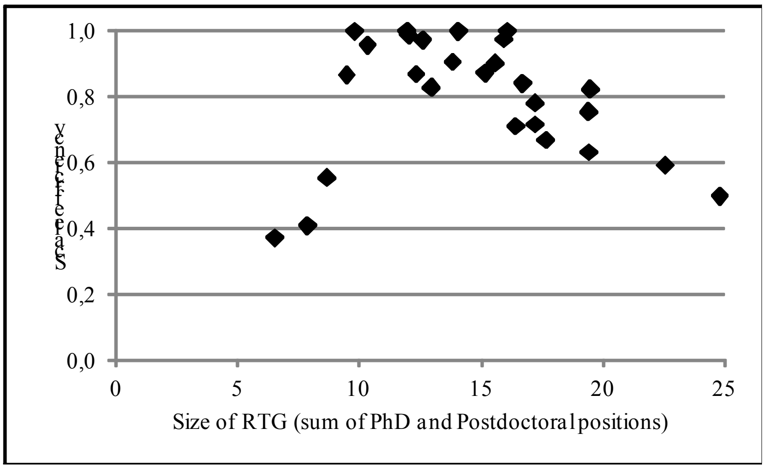
	CCR model	*BCC model*
Number of efficient RTGs	4	6
Mean	0.5993	0.7537
Standard deviation	0.2115	0.2024
Minimum	0.1977	0.3561
Maximum	1	1
Mean scale efficiency	0.8037	

Source: Own data.

In a next step, we calculate the scale efficiency as the ratio of the CCR efficiency value to the BCC efficiency value (see Cooper, Seiford and Tone 2006: 140). The four RTGs with efficiency values of 1 in both models operate at optimal size (i.e., they operate at global technical efficiency), whereas the two RTGs that are efficient under the BCC model but not the CCR model do not operate at optimal size. Likewise, no RTG with higher efficiency values in the BCC model than in the CCR model operates at optimal size. Only those RTGs that have the same efficiency value in the CCR and BCC models are scale efficient, i.e., they work at optimal size. The average scale efficiency in the humanities and social sciences is 80.37%. A correlation analysis between the CCR and BBC efficiency values yields r=0.65 as the correlation coefficient and 0.01 as the level of significance. Still, a large proportion of inefficiencies is not attributable to size effects but hints at other inefficiencies in the production process.

Figure 2 shows the relation between scale efficiency and RTG size, visualized as the *sum* of PhD and postdoctoral positions. The figure displays a hump-shaped relation between RTG size and scale efficiency. RTGs in the area of increasing scale efficiency – ranging up to approximately 10 person-years – reach an average scale efficiency of 55.20%. That is, they could optimize by further increasing their size. The optimum for the RTGs in the humanities and social sciences appears to lie in the range of 10 to 16 PhD and postdoctoral positions. In this range, we find (nearly) constant returns to scale. RTGs of that size are (nearly) scale efficient (average scale efficiency: 94.78%), as confirmed by a correlation coefficient of r=0.98 between the CCR and BCC efficiency values (level of significance: 0.01). In contrast, RTGs with more than 16 PhD and postdoctoral positions have decreasing scale efficiency values (average scale efficiency: 70.27%). These RTGs could reduce their inefficiency by downsizing. When comparing the average scale efficiencies of RTGs with 10 to 16 PhD and postdoctoral positions with the average scale efficiencies of smaller and larger RTGs with the help of a t-test, we found the average scale efficiency of the RTGs with 10 to 16 PhD and postdoctoral positions to be significantly higher than those of the smaller and the larger groups with our results being robust to slightly shifting the thresholds from the left or the right. None of the five RTGs from the humanities and social sciences that were unsuccessful in their application for a third funding period operated at optimal size in the second funding period.

Figure 2: Size effects in the Humanities and Social Sciences



Source: Own data.

4.2 Natural and life sciences

In the CCR model, four out of 58 RTGs in the natural and life sciences are efficient (Table 3)[6]: One of them draws its efficiency particularly from its publication activities, a second is the frontrunner in PhD completion, and two RTGs are successful in both output dimensions. The number of efficient units and the low average efficiency of all RTGs in the natural and life sciences (61.67%) indicate a considerable potential for efficiency increases. Compared to the CCR model, the calculation of the BCC model increases the number of efficient RTGs from four to nine, and the average efficiency value increases from 61.67% to 65.24%. It is worth noting that – in contrast to the humanities and social sciences – most of the RTGs in the natural and life sciences hardly raise their efficiency values from the CCR to the BCC model. That is, RTGs in the natural and life sciences are (almost all) scale efficient, which means they work at (nearly) optimal size. Correspondingly, the average scale efficiency is very high, with a value of 0.95. A correlation analysis of the CCR and BCC efficiency values shows a correlation coefficient of r=0.97 (level of significance: 0.01).

[6] It is important to note that we cannot compare the efficiency values of RTGs in the humanities and social sciences with those of RTGs in the natural and life sciences since the number of units included in the analysis influences the efficiency values and since there are more observations from the natural and life sciences in our data set (58 as opposed to 28 from the humanities and social sciences).

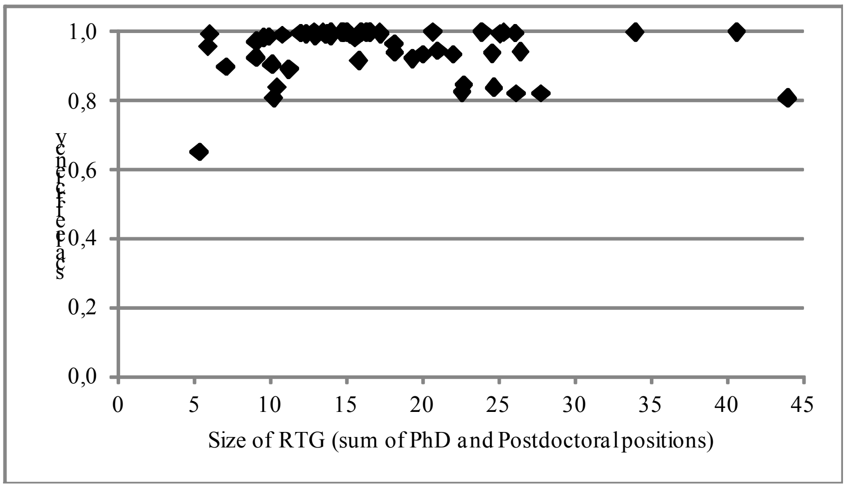
Table 3: Efficiency values in the Natural and Life Sciences

	CCR model	*BCC model*
Number of efficient RTGs	4	9
Mean	0.6167	0.6524
Standard deviation	0.2348	0.2489
Minimum	0.0966	0.0994
Maximum	1	1
Mean scale efficiency	0.9496	

Source: Own data.

Figure 3 shows the relation between scale efficiency and RTG size, again visualized as the sum of PhD and postdoctoral positions. Apparently, RTGs of almost *any* team size achieve high values of scale efficiency. None of the twelve RTGs from the natural and life sciences that were unsuccessful in their application for a third funding period operated at optimal size in the second funding period.

Figure 3: Size effects in Natural and Life Sciences



Source: Own data.

5. Robustness

We undertook a series of robustness checks to validate our results.

(a) We varied the inputs of the DEA and ran two alternative analyses. First, we ran a DEA including the *sum* of PhD and Postdoc positions as the one and only input variable (instead of PhD and post-doctoral positions as two separate inputs). The correlation coefficients between the scale efficiencies of our main model and the scale efficiencies of the model of this first robustness check are around r=.85*** for both, the natural and life sciences and the humanities and social sciences. As a second robustness check, we ran a DEA including the sum of PhD and postdoctoral positions as one input *and* the number of principal investigators of the RTG as a second input. The correlation coefficients between the scale efficiencies of our main model and the scale efficiencies of the model in this second robustness check are r=.78*** for the humanities and social sciences and .79***, for the natural and life sciences, respectively. The results of both of these robustness checks are robust in the sense that in the humanities and social sciences only medium-sized RTGs are scale efficient, while in the natural and life sciences RTGs of almost *any* team size achieve high values of scale efficiency.

(b) We varied the outputs of the DEA. Following the suggestion of one of the anonymous reviewers we ran two alternative analyses where in a first one we only included monographs and journal articles as publication outputs and in a second one we included monographs, book sections and journal articles. Again, in both additional analyses we find that in the humanities and social sciences only medium-sized RTGs are scale efficient while we find scale efficient RTGs in the natural and life sciences over the whole size distribution.

(c) Our last robustness check concerns the sensitivity of our results with respect to outliers. In light of the comparatively low average efficiency of the CCR model for the humanities and social sciences we followed the suggestion of an anonymous reviewer and calculated super efficiencies to detect outliers. In fact, we found one RTG to vastly outperform all the others (super efficiency of 384%). When we exclude this RTG from the efficient frontier, our results again remain robust to this alteration: there still seems to be a hump-shaped relation between RTG size and scale efficiencies.

6. Discussion and Interpretations

There are several competing explanations for our results. First, it is of course conceivable that there is no hump-shaped relation between size and performance for RTGs in the natural and life sciences, whereas there is such a relation in the humanities and social sciences. However, such an interpretation might be regarded as naïve or premature since the data we are using clearly suffer from an endogeneity problem, as we only observe data for RTGs that have been funded by the German Research Foundation. Assuming informed funding decisions, only those RTGs that operate at (nearly) optimal size will ultimately be funded and hence included in our sample. Hence, the fact that we observe hardly any RTGs in the natural and life sciences that operate at a sub- or above-optimal size might result from the German Research Foundation simply not being likely to fund RTGs with non-optimal size in the natural and life sciences.

However, why, then, should the German Research Foundation fund RTGs in the humanities and social sciences that operate at a sub- or above-optimal size when it does not in the natural and life sciences? The argument could run as follows: In the natural and life sciences, with its long-standing tradition of working in teams, the information base on the optimal group size, was – at the time our data come from – far better than in the humanities and social sciences for both applicants and reviewers. Hence, in the natural and life sciences, we might not observe RTGs that survived the review process and that operated at a sub-optimal or above-optimal size because researchers in the natural and life sciences, with its long tradition of research groups, have sufficient information on the optimal size of a research group and how it varies under specific conditions. In contrast, in the humanities and social sciences, where working in groups of jointly supervised PhD students and postdoctoral researchers was not widespread at the time that our data come from, there was very little common knowledge on the optimal group size, let alone how it may vary according to specific conditions. Hence, the humanities and social sciences were in a "trial and error" phase; therefore, we can expect to observe RTGs that – even though they survived the review process – operated at a sub-optimal or above-optimal group size, because apparently neither the applicants nor the reviewers were aware of the fact that the scale efficiency of the RTG in question could have been increased by scaling up or sizing down, respectively.

Another reason we find few RTGs in the natural and life sciences that operate at a sub- or above-optimal size might be the fact that our sample is rather small (however, the sample for the natural and life sciences is larger than that for the humanities and social sciences). Potentially, a larger sample might have rendered more data points where RTGs in the natural and life sciences operate at a sub- or above-optimal size. With respect to the humanities and social sciences, the fact that here we observe more non-optimal program sizes might be explained by the humanities and social sciences being characterized by a larger degree of heterogeneity with reference to the sub-disciplines contained in that disciplinary area. As one of our anonymous reviewers noted, relative efficiency might then simply be a measure of difference. In contrast, programs in the natural and life sciences might – on average – be more homogenous. Again, more data points would potentially help in assessing these effects.

7. Limitations and Implications

Since we can observe only *funded* RTGs, that is, RTGs that were positively reviewed, our data might suffer from a selection and self-selection bias. This bias will arguably be stronger when reviewers possess better information, leading to only very strong applications with a good understanding of the optimal group size being positively evaluated and ultimately funded. In addition, this will be anticipated by potential applicants, leading to only those applications that have a good understanding of the optimal group size being sent out in the first place. In light of our results, we argue that in the natural and life sciences, with its long-standing tradition of working in teams, the information base on the optimal group size is presumably much better than that in the humanities and social sciences for both applicants and reviewers. Hence, in the natural and life sciences, we might not observe RTGs that survived the review process *and* operated at a sub-optimal or above-optimal size since researchers in the field have information on the optimal size of a research group and how it varies under specific conditions. In contrast, in the humanities and social sciences, where working in groups of jointly supervised PhD students and postdoctoral researchers is not that widespread, we observe RTGs that – even though they survived the review process – might still operate at sub-optimal and above-optimal group sizes.

As to the implications for research policy, we conclude that there is evidence for an optimal size of RTGs – which will, however, vary with respect to the specific contexts in which the RTG is set up. In a situation in which applicants and reviewers know the optimal group size and how it varies with the specific context, we will observe only RTGs that operate at optimal size. Whereas the natural and life sciences are exemplary of such a situation, the humanities and social sciences are not – or at least were not in the time from which our data come. It would be interesting to see whether today we would still observe RTGs in the humanities and social sciences that do not operate at optimal size given that today the informational basis on optimal group size – even in a disciplinary area that is as heterogeneous as the humanities and social sciences – will arguably be much better than at the beginning of the century.

In the late nineties and the beginning of the new millennium, the humanities and social sciences only just started to organize their PhD education in groups, and even today, the individual student-advisor relationship in PhD education is widespread. Being able to observe non-optimal sizes of PhD programs in the humanities and social sciences hints at a “trial and error process” being underway. This trial and error process, however, might be very productive, because the process itself generates the missing information on the optimal group size in contexts that do not have a long-standing tradition in working in groups. That is, from the perspective of research policy, it is, first, important to leave room for experiments, e.g., to be willing to establish the first RTG in a field that is still dominated by individual student-advisor relationships in PhD education without knowing what its optimal size may be and to experiment with different group sizes in succeeding funding decisions. Second, the reviewer process should be constructed as a learning system where information on the relative performance of differently sized RTGs is reflected and then fed back to the scientific community.

Our study is not without limitations. The first limitation is, of course, the comparatively small data base and the fact that it renders only a snapshot at a certain point in time. Hence, we can only speculate on whether and how things might have changed until the time of our data collection. In addition, our level of aggregation is quite high in that we can distinguish between the natural and life sciences on the one hand and the humanities and social sciences on the other but cannot – as a result of the small data base – disaggregate our data to the level of sub-disciplines. In light of the fact that the broad disciplinary areas are still rather diverse, a more disaggregate look at the data may be needed. Further, the method we applied (DEA) is, not undisputed. Although DEA has certain distinct advantages that explain its widespread use, including in analyses of higher education, it also has its disadvantage, for instance, as discussed by Albers (2015). Further, we regard RTGs as research entities not knowing whether there are potentially alternatively funded researchers that are also part of the research group and thus effectively increase the group size. Unfortunately, there are no such data, and we hence cannot account for this in our analysis. Last but not least, it should be mentioned that the outputs we regard (PhD completion rates and publications) are not quality adjusted; that is, we count the number of completed PhDs without assessing the quality of the work that has been done, and we count the number of publications without a quality assessment. Although it might be possible to quality-adjust the two measures for a specific sub-discipline and to then comparatively assess the relative efficiency of the RTGs within that same sub-discipline, a quality adjustment of RTG outputs for the whole set of sub-disciplines covered in our analysis would seem beyond reach.

Literature

Abbott, Malcolm, Chris Doucouliagos. 2003. The efficiency of Australian universities: A data envelopment analysis. *Economics of Education Review* 22: 89-97.

Ahn, Heinz, Harald Dyckhoff, Roland Gilles. 2007. Datenaggregation zur Leistungsbeurteilung durch Ranking: Vergleich der CHE- und DEA-Methodik sowie Ableitung eines Kompromissansatzes. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 77: 615–643.

Albers, Sönke. 2015. What drives publication productivity in German business faculties? *Schmalenbach Business Review* 67: 6-33.

Athanassopoulos, Antreas D., Estelle Shale. 1997. Assessing the comparative efficiency of higher education institutions in the UK by means of Data Envelopment Analysis. *Education Economics* 5: 117-134.

Audretsch, David B., Erik E. Lehmann, Susanne Warning. 2004. University spillovers: Does the Kind of science matter? *Industry and Innovation* 11: 193-205.

Banker, Rajiv D., Abraham Charnes, William W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30: 1078-1092.

Bedeian Arthur G. et al. 2010. Doctoral degree prestige and the academic marketplace: A study of career mobility within the management discipline. *Academy of Management Learning and Education* 9: 11-25.

Biglan, Anthony. 1973. Relationships between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology* 57: 204-213.

Black, Grant C., Paula E. Stephan. 2010. *The economics of university science and the role of foreign graduate students and postdoctoral scholars.* American universities in a global market. University of Chicago Pres. 129-161.

Bonaccorsi, Andrea, Cinzia Daraio. 2005. Exploring Size and Agglomeration Effects on Public Research Productivity. *Scientometrics* 63: 87–120.

Bowen, William G., Neil L. Rudenstine. 1992. *In pursuit of the Ph.D.* Princeton: Princeton University Press.

Brandt, Tasso, Torben Schubert. 2013. Is the university model an organizational necessity? Scale and agglomeration effects in science. *Scientometrics* 94: 541–565.

Breneman, David W. 1976. The Ph.D. production process. *In Education as an Industry*, eds. Froomkin, Joseph T., Dean T. Jaminson, Roy Radner, 3-52. Cambridge, Mass.: Ballinger.

Breuninger, Susanne, Kerstin Pull, Birgit Pferdmenges 2012. Like father(s), like son(s) – Does the relation between advisor and student productivity persist on group level? *German Journal of Research in Human Resource Management – Zeitschrift für Personalforschung* 26: 331-345.

Brown, Kathryn S. 1996. The key to Academic bliss can be found in large or small departments. *The Scientist* 10(1996)21: 15-16.

Carayol, Nicolas, Mireille Matt. 2004. Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy* 33: 1081-1102.

Charnes, Abraham, William W. Cooper, Edward Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429-444.

Cherchye, Laurens, Piet Vanden Abeele. 2005. On research efficiency. A micro-analysis of Dutch university research in Economics and Business Management. *Research Policy* 34: 495-516.

Clermont, Marcel, Alexander Dirksen, Harald Dyckhoff 2015: Returns to scale of business administration research performance in Germany. *Scientometrics* 103: 583-614.

Cohen, Joel E. 1981. Publication rate as a function of laboratory size in a biomedical research institution. *Scientometrics* 3

Cohn, Elchanan, Sherry L.W. Rhine, Maria C. Santos. 1989. Institutions of Higher Education as Multi-Product Forms: Economies of Scale and Scope. *Review of Economics and Statistics* 71: 284–290.

Cooper, William W., Lawrence M. Seiford, Kaoru Tone. 2006. *Introduction to data envelopment analysis and its uses*. New York: Springer.

Dundar, Halil, Darrell R. Lewis. 1998. Determinants of research productivity in higher education. *Research in Higher Education* 39: 607-631.

Dyckhoff, Harald et al. 2013. Measuring balanced effectiveness and efficiency of German business schools' research performance. *Zeitschrift für Betriebswirtschaft* Special Issue 3/2013: 39-60.

Dyckhoff, Harald, Sylvia Rassenhövel, Kirsten Sandfort. 2009. Empirische Produktionsfunktion betriebswirtschaftlicher Forschung: Eine Analyse der Daten des Centrums für Hochschulentwicklung, *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 61: 22–56.

Egghe, Leo, Ronald Rousseau, Guido van Hooydonk. 2000. Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science* 51: 145–157.

Fandel, Günter. 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: 521–533.

Gellert, Claudius. 1993. The conditions of research and training in contemporary German universities. In *The research foundations of graduate education*, ed. Burton R.Clark, 45-66. Berkeley: University of California Press.

Groot, Tom, Teresa García-Valderrama. 2006. Research quality and efficiency. An analysis of assessments and management issues in Dutch economics and business research programs. *Research Policy* 35: 1362-1376.

De Groot, Hans, Walter W. McMahon, J. Fredericks Volkwein. 1991. The Cost Structure of American Research Universities. *Review of Economics and Statistics* 73: 424–451.

Hagstrom, W.O. 1964. Anomy in Scientific Communities. *Social Problems* 12: 186-195.

Hamilton, Lawrence C. (2006): Statistics with STATA. Belmont, Calif. [u.a.]: Thomson, Brooks/Cole.

Johnes, Jill, Geraint Johnes. 1995. Research funding and performance in U.K. university departments of Economics: A frontier analysis. *Economics of Education Review* 14: 301-314.

Knorr-Cetina, Karin. 1992. The couch, the cathedral, and the laboratory: On the relationship between experiment and laboratory science. In *Science as practice and culture*, ed. Andrew Pickering, 113-138. Chicago, London: University of Chicago Press.

Korhonen, Pekka, Risto Tainio, Jyrki Wallenius. 2001. Value efficiency analysis of academic research. *European Journal of Operational Research* 130: 121-132.

Kyvik, Svein. 1995. Are big university departments better than small ones? *Higher Education* 30: 295-304.

Laband, David N., Bernard F. Lentz. 2003. New Estimates of Economies of Scale and Scope in Higher Education. *Southern Economic Journal* 70: 172–183.

Laband, David N., Robert D. Tollison. 2000. Intellectual collaboration. *Journal of Political Ecomnomy* 108: 632-662.

Lloyd, Peter J., Margaret H. Morgan, Ross A. Williams. 1993. Amalgamations of Universities: Are There Economies of Size and Scope? *Applied Economics* 25: 1081–1092.

Main, Joyce B. 2014. Gender homophily, Ph.D. completion, and time to degree in the Humanities and Humanistic Social Sciences. *The Review of Higher Education* 37: 349-375.

McMillan, Melville L., Debasish Datta. 1998. The relative efficiencies of Canadian universities: A DEA Perspective. *Canadian Public Policy - Analyse de Politiques* 24: 485-511.

Nazarko, Joaniscjusz, Jonas Saparauskas. 2014. Application of DEA method in efficiency evaluation of public higher education institutions. *Technological and Economic Development of Economy* 20: 25-44.

Ng, Ying Chu, Sung Ko Li. 2000. Measuring the research performance of Chinese higher education institutions: an application of Data Envelopment Analysis. *Education Economics* 8: 139-156.

Nuijten, E. 2011. Combining Research Styles of the Natural and Social Sciences in Agricultural Research. *NJAS-Wageningen Journal of Life Sciences* 57: 197–205.

Powell, Stuart; Howard Green (eds.) 2007. *The doctorate worldwide*. Maidenhead: Open University Press.

Pull, Kerstin, Birgit Pferdmenges, Uschi Backes-Gellner. 2016. Composition of junior research groups and PhD completion rate: disciplinary differences and policy implications. *Studies in Higher Education* (in print).

Sadlak, Jan. 2004. *Doctoral studies and qualifications in Europe and the United States: Status and prospects.* Bukarest: UNESCO-CEPES.

Snow, Charles P. 1964. *The two cultures: and a second look. An expanded version of the two cultures and the scientific revolution*. Cambridge: University Press.

Stephan, Paula E. 1996. The Economics of Science. *Journal of Economic Literature* 34: 1199-1235.

Unger, Birgit, Kerstin Pull, Uschi Backes-Gellner. 2010. The performance of German research training groups in different disciplines: An empirical analysis. In *Governance and performance in the German public research sector: disciplinary differences*, ed. Dorothea Dordrecht Jansen, 93-106. Springer.

Verbree, Maaike et al. 2015. Organizational factors influencing scholarly performance: a multivariate study of biomedical research groups. *Scientometrics* 102: 25-49.

Van der Wal, René et al. 2009. Is bigger necessarily better for environmental research? *Scientometrics* 78: 317–322.

Wanner, Richard A., Lionel S. Lewis, David I. Gregorio. 1981. Research productivity in academia: a comparative study of the Sciences, Social Sciences and Humanities. *Sociology of Education* 54: 238-253.

Warning, Susanne. 2004. Performance differences in German higher education: empirical analysis of strategic groups. *Review of Industrial Organization* 24: 393-408.

Warning, Susanne. 2007. *The Economic Analysis of Universities. Strategic Groups and Positioning*. Cheltenham, Northhampton: Elgar.