Scientometrics (2016) 108:83–111
DOI 10.1007/s11192-016-1947-9

Optimal research team composition: data envelopment analysis of Fermilab experiments

Slobodan Perović[1] · Sandro Radovanović[2] · Vlasta Sikimić[3] · Andrea Berber[1]

Received: 5 October 2015 / Published online: 27 April 2016

Abstract We employ data envelopment analysis on a series of experiments performed in Fermilab, one of the major high-energy physics laboratories in the world, in order to test their efficiency (as measured by publication and citation rates) in terms of variations of team size, number of teams per experiment, and completion time. We present the results and analyze them, focusing in particular on inherent connections between quantitative team composition and diversity, and discuss them in relation to other factors contributing to scientific production in a wider sense. Our results concur with the results of other studies across the sciences showing that smaller research teams are more productive, and with the conjecture on curvilinear dependence of team size and efficiency.

Keywords Social epistemology of science · Team size · Team diversity · Data envelopment analysis · High energy physics · Fermilab

Introduction

When it comes to determining the optimal distribution of human resources of a research group with respect to its performance, a number of diverse investigations have been conducted in various fields, e.g. psychology (Jackson 1996; Kozlowski and Bell 2003), economics (Page 2011), and philosophy (Kitcher 1990; Zollman 2007). Various tools of scientometric research have also been applied in the study of the structure of scientific research teams and their performance across the sciences (Milojević 2014; Abbasi et al.

✉ Slobodan Perović
perovicslobodan@gmail.com

[1] Department of Philosophy, University of Belgrade, Belgrade, Serbia

[2] Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia

[3] Department of Formal Languages, Technische Universität Wien, Vienna, Austria

2011; Bonaccorsi and Daraio 2005; Horta and Lacy 2011; van der Wal et al. 2009; Von Tunzelmann et al. 2003; Qurashi 1991; Nieva et al. 1985). We investigate the question of optimal research quantitative group composition (the number of researchers and the number of teams) in Fermilab, a major high-energy physics laboratory. We apply an input-oriented variable returns to scale (VRS) Data Envelopment analysis (DEA) model based on the data for three inputs (team size, number of teams, length of the experiment performance) and six outputs (based on the papers valued by citations, divided into six categories) for twenty-seven experiments over the course of a decade and a half. The primary goal of our data-driven analysis is to find out whether there is an optimal quantitative team composition in the context of modern particle physics laboratories, and the reasons underlying it. We analyze the relationship between the quantitative team composition and the importance of diversity of teams and projects (experiments), the seniority and institutional affiliations of team members, and the nature of relevant collaborations for the efficiency of scientific communities working in the given context.

The organization of the paper is as follows. In the following section we provide the motivation for our research by presenting the debate on optimal quantitative team composition, pointing out the ways in which it is interrelated with team diversity, and stating its limitations. We also explain the focus of our analysis on the organizational structure of high-energy physics laboratories and Fermilab in particular, and explain the main characteristics of the teams and the projects in this context. In the third section we explain the methodology we use, provide a brief introduction to DEA, and explain its advantages and limitations. In the fourth section we explain our choices of data and the type of analysis, and present the results. The fifth section analyzes the results and relates them to other relevant studies, while the concluding the last section suggests possible policy implications and points to new directions for future research.

Motivation

The issue: an optimal research team size, an optimal number of teams, and team diversity

The research teams and their quantitative composition

Studying the size of the teams in order to understand and test team performance was initiated in and is still a focus of industrial economics (Brinkman and Leslie 1986). Yet one focus of recent research has been the possibility of an optimal research team size in biology laboratories (Cook et al. 2015).[1] Also, there is still a long-standing controversy over whether there are good reasons, conceptual or empirical, for the claim that any given research project can be best performed by a particular number of teams of a particular size. The team size has been identified (Kozlowski and Bell 2003) as one of several dimensions of team composition, along with the demographic characteristics of the team members, their skills and personalities. It has been also characterized as one of three structural factors of team composition, along with resources and time spent on research (West and Anderson 1996).

Very different and often opposed conclusions about the optimal team size have been drawn. Based on empirical studies, some researchers suggest that seven members is the optimal size of

[1] Results of this study show that biology labs should ideally have between ten and fifteen members.

a team (Scharf 1989; West and Anderson 1996). Regarding research groups in academia, Qurashi (1991) suggests that the optimal size of a scientific research team is between five and nine. Some have suggested a curvilinear relationship between team size and performance (Nieva et al. 1985) including research in academia (Andrews 1979); the curvilinear relationship means that performance increases with an increase in team size up to a certain fixed point. Some claim that there is a limit, where increasing the size of a team negatively affects research (Hackman and Vidmar 1970; Martz et al. 1992) which also applies to the research in academic settings (Bonaccorsi and Daraio 2005), while others (Campion et al. 1993) claim the opposite. We will discuss these and other relevant findings and various aspects of team composition when we present and analyze the results of our study.

It should be noted that one question that divides authors in the debate is the extent to which the context of research determines the answer to these questions, and to what extent one can give reasonable recommendations given the size of a project (including overall number of researchers) irrespective of its content. Regardless of the resolution of this dilemma, it is generally beneficial to properly account for the specific properties of any projects analyzed. This is why data-dependent analysis is an invaluable tool. Thus, satisfactory simulations and quantitative methods should ideally analyze the actual data.[2]

Diversity of research teams

The issue of team size is inherently related to another issue in studying performance of research teams, in which interest has recently expanded—namely the diversity of research groups. We will explain the nature of this relationship in the next section, but for now let us briefly review the motivation for examining the diversity aspect of research teams.

After a shift in social and applied epistemology of science from single-agent knowledge to the examination of group beliefs and knowledge acquisition, interdisciplinary interest in the research of multi-agent dynamics arose (Valentin et al. 2016; Wang et al. 2015). In general, various empirical studies have shown the importance of diversity for innovation,[3] which is especially important in scientific research. Different studies in economics and management concerning the efficient division of labor have shown that cognitive diversity reflects positively on a team's efficiency.[4] Page (2007) argues that there is a link between cognitive differences among team members and better collective outcomes at specific tasks, namely those involving problem solving and prediction. Similarly, Kitcher (1990, 1993) and Strevens (2003) have argued that it is useful for the scientific community to have a significant number of scientists working on research programs that are considered less promising.[5]

Various methods, mostly standard sociological techniques, have been used to test different hypothesis concerning diversity in scientific communities (Olson et al. 2007). Some authors used computer simulations instead to examine the issue. Weisberg and Muldoon

[2] Moreover, from a data-driven analysis both general and particular conclusions can be extracted, while, as we will see in more detail shortly, this is not the case with data-independent, i.e. hypotheses-driven analyses.

[3] For example, see Kimberly (1981) and Agrell and Gustafson (1996).

[4] For example, see Bantel and Jackson (1989) and Olson et al. (2007).

[5] Kitcher (1990) writes: "I claim, simply, that we sometimes want to maintain cognitive diversity even in instances where it would be reasonable for all to agree that one of two theories was inferior to its rival, and we may be grateful to the stubborn minority who continue to advocate problematic ideas." Both Kitcher and Strevens use examples from history of science to underline the positive epistemic effect of diversity in methods used for tackling scientific problems. They have investigated the best way of assigning credit within the scientific community in order to achieve an optimal division of labor.

(2009) have used a computer simulation to investigate one type of cognitive diversity in science, namely variation in patterns and strategies of intellectual engagement with respect to the other research teams. They considered three different research strategies: the *follower* strategy—a strategy of being biased towards already explored and known results; the strategy of working independently; and the *maverick* strategy—deliberately avoiding explored areas of research. The results show that the population composed purely of scientists using the follower strategy does less well than the population of scientists working independently. However, the population of scientists using the maverick strategy vastly outperform populations that use the other two strategies. Finally, the most interesting finding is that populations composed of a few mavericks and many followers are the most successful and most frequently discover significant scientific results. Alexander et al. (2015) argue against these conclusions.[6]

The work of Zollman (2007, 2010) displays similar aims and uses similar techniques, but it is perhaps the most elaborate use of computer simulations in social epistemology of science. Zollman uses simulations to study a model representing learning situations in scientific communities and examines the probability of successful learning in three types of social networks, which are usually represented using graphs known as cycle, wheel, and the complete graph.[7] The cycle represents a social network in which the agents are arranged in a circle and only communicate with those agents who are on either side of them. This is supposed to represent loosely related research groups, each of which uses a different method (i.e. different assessment of the neighbours' results). The wheel is a cycle in which only one of the agents is connected to everyone else. It resembles a centralized and hierarchical setup of laboratories. Finally, the complete graph is social network where everyone communicates with everyone else. The computer simulation has shown that a cycle is more reliable than the wheel, and that both of these do better than the complete graph, in which each agent is directly informed of everyone else's results. Zollman's result might *prima facie* seem unexpected, since his analysis shows that the community of scientists is more reliable when its members are less frequently and often indirectly made aware of their colleagues' experimental results. His results show that there is an optimal amount of intercommunication that maximizes the reliability of the results. Yet there is a trade-off between the speed of reaching consensus in a scientific community—the speed being increased with communication, on the one hand, and the reliability of the results on the other.

[6] They point to shortcomings in Weisberg and Muldoon's approach in order to undermine their attempts to provide epistemic reasons for a division of cognitive labor. Using an epistemic landscape model they show that in some cases homogeneous populations may be more successful than heterogeneous ones. Furthermore, they offer a general complaint against Weisberg and Muldoon's particular model: the necessity of basing simulations on assumptions concerning the specific nature of the epistemic landscape is problematic because specific features of the epistemic landscape may be beyond our knowledge. They did not argue against any epistemic reason for cognitive diversity, but instead claimed that Weisberg and Muldoon did not succeed in showing it.

[7] He investigates the following issue: imagine that there are two competing methods, M1 and M2, for tackling a scientific problem. It may happen that initial experiments favor M1 even though M2 is the correct method. In this situation a consensus can be reached too fast; scientists who believe that M2 is the correct method may cancel further research once they become aware of their colleagues' experimental results and thus the wrong methodology may become consensual. Zollman uses computer simulation to explore whether there is a correlation between the structure of a communication network and convergence towards the right hypothesis. The structure of a communication network represents connections between agents, or more precisely the connections between them and those with whom they share information.

Generally speaking, there are at least two levels of diversification of research teams. At the level of the team composition individual team members can contribute different education, training, and experience, which can result in diversity of the team's approaches to the task at hand. But, the point most important to our study, diversity is also brought about with a looser connection between the team members, as less connectedness means more autonomy in deciding and developing one's ideas and approaches to the task at hand, which ultimately results in a diversity of approaches to the task. Now, another point important for our study is that these two kinds of diversity apply to teams working on a project and to different laboratories pursuing a common goal as much as they apply to individual team members.[8]

The general questions

Our study is motivated by the following general questions that arise with respect to the quantitative composition and performance of research teams across sciences, which we will ask with respect to the research in High Energy Physics laboratories: First, we are interested in how variations in the number of researchers affect the efficiency of research outcomes. In other words: *What is the optimal number of researchers in a given project?* Second, how many teams should constitute a project? Should one basically stick to one research master-team, or should one break the master-team into smaller groups? How many should there be? It is clear that one needs to break researchers into teams in more complex projects, but is there an optimal number of teams for a given project? Thus, *how do variations in the number of teams involved in a project affect the outcomes?* Third, *does the time it takes to complete a project of a certain composition affect the outcomes?*

Also, the *size factor of team composition*, as we have pointed out, is *intrinsic to structural diversity*: too small a team apparently leads to a lack of diversity of viewpoints (Jackson 1996). Too large teams become cumbersome and prevent effective communication, thus stifling the benefits of diversity (Poulton 1995). Large teams face motivational issues at some point, and the more the groups hang together the more they lose the ability to offer internal and external critical feedback, again stifling diversity (Katz 1982; Kozlowski and Hulls 1986; Bantel and Jackson 1989; Jackson 1996). Now, the variations in the number of researchers (i.e. size of the master-team) and in the number of teams in a project will nudge teams to either hierarchical or less centralized structures as larger teams tend to be more hierarchical (Heinze et al. 2009) and thus determine the extent of diversification through the autonomy and looseness of connections between the team members and between the teams. Thus, getting right the number of researchers and the number of teams is coextensive with achieving sufficient and optimal diversity in a project, and also for properly balancing it with a hierarchical organization. Our guiding idea, then, is that the quantitative analysis of team composition (*the first goal of our study*) is a very helpful, and perhaps even necessary first step in the study of diversity in a scientific community (*our second goal*) both of which we will study in the context of modern particle physics laboratories. In the analysis of our results we will tackle this coextensive relationship between the quantitative structural factors of team composition (number of researchers and teams) and diversity.

[8] In other words, the nodes of a graph representing a scientific network can represent individuals, teams, or laboratories.

The case of Fermilab: teams and experiments in high-energy physics laboratories

Fermilab

This study is part of a larger quantitative and qualitative assessment of the organizational structure and performance of high-energy physics (HEP) laboratories in particle physics. Particle physics laboratories have been transformed to an unprecedented extent in the last six or seven decades. They started as laboratories with a few scientists involved in individual experiments at the end of the 19th century, and steadily grew into laboratories that employ hundreds of permanent staff and thousands of physicists working on each experiment. The number of physicists grew by about a hundred times after WWII, while the resources available for particle physics grew even more (Kragh 2002, Ch. 2). The experiments themselves are currently performed on equipment that sometimes spans tens of kilometers in length. The discovery papers are coauthored by thousands of physicists. Given their cost, the importance that the results they produce have for the development of fundamental physics and physics in general, and technological innovations as their byproduct, it is very important to assess the organizational structure and efficiency of these laboratories (Martin and Irvine 1984).

Fermilab has been one of the most successful laboratories in the history of particle physics (Hoddeson 1997; Hoddeson et al. 2008). A number of breakthrough discoveries such as the discoveries of the bottom and top quarks, the key elements of the Standard Model of particle physics, have been some of its major achievements. It is also the first physics mega-laboratory in the US built for non-military purposes.

The development of Fermilab took place in three distinct phases (Hoddeson et al. 2008). During the first phase (late 1960s to late 1970s) it functioned as a site for experimental teams to assemble, within an allocated time period, the apparatus at the on-sight facilities, often with the help of equipment borrowed from other laboratories, and run the experiment. It was not centralized like, e.g., CERN (Herman et al. 1987) or Lawrence Berkley Lab (Hoddeson et al. 2008, p. 40; Galison and Hevly 1992; Heilbron and Seidel 1989) were from their inception.[9] It housed multiple teams that performed hundreds of experiments. In the second phase (late 1970s to mid-1990s), the number of experiments was substantially reduced while their running time was substantially extended (spanning from a few years to half a decade). In the third phase, with the emergence of the colliding of beams of particles technique the number of the approved experiments was substantially reduced again, and took a much longer time (up to a decade) to design, commission, and perform an experiment, also requiring much larger teams. During this phase the experiments performed in the laboratory started to resemble the current experiments at the Large Hadron Collider at CERN that take a decade to realize and employ thousands of physicists. The contrast between the organizational structure of the third phase on the one hand, and the other two phases is much more substantial than between the first two phases (Boisot 2011).

The three phases represent the dominant ways of organizing the laboratory, but the transitions between them were gradual. Now, the experiments we analyzed were performed during the period that spans from the very beginning of the second phase to the point when it started merging with the third phase. Thus our analysis captures a fairly homogenous

[9] See also Westfall (1997) and Greenberg (1999) for an analysis of the policies that underlined the organization of big physics laboratories.

organizational structure when compared to the other two phases, especially the third. Also, focusing on a period when experiments were still of rather moderate size seems to be a natural first step in the sort of analysis of HEP laboratories we are undertaking. The analysis of the third phase would provide more recent data, but the changes in the composition of teams and overall experiments that emerged in the third phase are very substantial in terms of their size and structure. This would require a rethinking of our methodology, possibly including more suitable quantitative methods. Yet we offer ideas regarding the analysis of data about how the results may be relevant to the third, collider phase of the experimental work in Fermilab and HEP laboratories in general, which requires analysis of its own. Moreover, a vast majority of experiments in particle physics laboratories are still of a moderate size compared to the mega-experiments performed at CERN, which makes the policy implications of the analysis relevant to them, too.

Projects

There are a number of organizational specificities that have to be taken into account when studying team composition, organization, and performance in HEP laboratories.

Many scientists directly involved in experiments at HEP laboratories are affiliated with external institutions, mostly universities and university-affiliated institutes. This has been true of Fermilab from its beginnings (Hoddeson et al. 2008). Thus the number of those employed at an HEP laboratory may not be very informative if it is compared to more traditional laboratories. This is why it is more beneficial to analyze *individual experiments* around which researchers are organized rather than the laboratory as such, especially if one focuses on the performance in a narrow sense (i.e. based on publications and citations)—as our quantitative analysis does. Thus, although currently Fermilab employs around 1750 staff,[10] we focused our analysis on the organization of individual experiments, since external researchers often form part of the teams and often can be the only primary scientists in some experiments (See Table 5). Thus, for instance, a coordinator of statistical analysis in a major experiment may spend little or none of her time in the laboratory. More generally speaking, given the results of a study by Bonaccorsi and Daraio (2005) that collected data across the sciences, in the analysis of productivity of research institutes "the level at which increasing returns apply is not the institute, but the research team" (p. 107). In our case, we decided to focus on the level of individual experiments performed in the laboratory by external and internal members of the experimental team.

In Fermilab, researchers are essentially organized around individual experiments. The research group gathers and proposes an experiment, which is then either approved or rejected by the laboratory management, which is in turn overseen by the Board of directors (Ibid.). Yet the process of establishing the project has varied throughout the three phases of the development of the laboratory. It also differs considerably from the same process in other more centralized laboratories such as CERN (Kragh 2002; Krige 1993; Herman et al. 1987; Irvine and Martin 1985). Fermilab was envisioned as a user-empowered laboratory, where individual master-teams would have full control over the choice, design, and performance of experiments. This was in tune with the general spirit of the laboratory, where the management insisted on "maintaining the focus of the laboratory squarely on outside users having numerous quickly done experiments rather than concentrating manpower, money, and beam time on a few larger experiments" (Hoddeson et al. 2008, p. 182).

[10] http://www.fnal.gov/pub/about/.

Teams

The physicists, engineers, and technical staff who gather to design and perform an experiment will share a common goal for the experiment, e.g. to measure a certain physical value precisely, search for a particle, or test a new experimental technique. Yet typically, the physicists' mission is much more closely tied to the fulfillment of this main goal, whereas the technical staff and engineers' main mission is rather to make sure the assembled equipment works properly. Thus the careers and performance of the latter will not be directly tied to the success of the experiment. (The experiment can fail for a number of reasons, even if the equipment works impeccably.) The technical staff are essential to the performance too, but in quite a different way (see "The data, the choices, and the limitations" section). Moreover, unlike in some other laboratories in particle physics, in Fermilab physicists rather than staff and engineers are in charge of almost all the tasks—or at least they were during the first two phases—including administration and maintenance of the laboratory equipment (Hoddeson et al. 2008). This was actually a deliberate policy that aimed at preventing the emergence of a "technocratic" management of the laboratory that was believed to negatively affect performance, and indeed did affect it that way in other laboratories (Ibid., Part 2). This is why we focus on the performance of physicists directly involved in the design, commissioning, and performance of the experiments.

The physicists working on an experiment will divide the master-team into smaller teams, each in charge of various aspects of design of the equipment, calibration, or the detection process. Thus, the composition and the expertise of each team will basically match one of the subsystems of the experimental equipment (Ibid.). It is very important to realize, however, that although specialized, the teams will not only coordinate their tasks but constantly share their expertise while coordinating. (We point out these and other details of the structure of the teams, and the limits it imposes on our study in "The data, the choices, and the limitations" section.) Thus, physicists work on their main area of expertise, but this is never exclusive; they will constantly share ideas within other collaborators' domains of expertise. This is particularly important during the design of the experiment and when it comes to choices concerning the exact hypothesis (or hypotheses) that is going to be tested. This sort of on-going collaboration and cross-pollination of knowledge improves the overall performance and has been an essential part of the culture of team-work in physics laboratories for some time (Heinze et al. 2009, p. 617). This cross-fertilization of ideas is encouraged so that the autonomy of the approach is preserved. Such networks seem similar to the kind of loose connectedness that Zollman's (2007, 2010) simulations suggest are optimal, rather than to the wheel structure or the maximally connected networks. But, as essential as such cross-expertise may be, it may become increasingly hard to maintain with the increase of the number of teams within a research project, and with the increase in the number of researchers. The main question we will address is whether and to what extent an increase in the size and the number of teams affects performance in a major HEP laboratory, as it turns out to do, as we will see, in many other cases across the sciences.

The methods

Our general methodology

The main idea of our approach is to establish what made certain experiments in Fermilab more efficient than others; more specifically, whether changing the team composition along two structural quantitative dimensions (team size and the number of teams) and completion time may have beneficial effects on a project's efficiency expressed as the number of papers associated with each experiment, where each paper is weighted based on citations. Ultimately, getting the team composition right over time enables appropriate diversification at the level of teams, and also along various other dimensions of team composition, such as seniority, affiliation, skills, personalities, and demographics.

In terms of the choice of method of quantitative analysis, although e.g. Zollman's technique and his results provide a good general argument and motivation for further research, the application of the results to concrete cases in science is inevitably too tenuous. While such hypotheses-driven simulations may provide general arguments in favor of a particular hypothesis, e.g. diversification of scientific networks, our analysis is data-driven. It seeks to determine an optimal number of teams and researchers in an experiment of a particular kind based on actual data. Thus using data envelopment analysis and distance-based analysis we evaluate the efficiency of each research project (experiment) and subsequently, in the analysis, identify the projects from which valid conclusions about diversity can be extracted. The most relevant advantage of the application of DEA is that the determinants of the inefficiency of each individual unit tested can be evaluated (He et al. 2015). The method has proven to reveal parameters relevant to the efficiency of the units. We argue that DEA is beneficial when studying the structure of research teams in physics mega-laboratories. Agasisti and Johnes (2015) for similar reasons chose DEA as most suitable for investigating higher education in US. Ideally, such a method could provide predictions for optimizing the resources of a scientific laboratory by testing various team compositions.[11]

Finally, our approach suggests that after a number of different analyses of this type one could be in a position to draw normative conclusions that would be applicable to social epistemology studies of physics laboratories similar to the one we examine.

Data envelopment analysis (DEA)

DEA is a mathematical programming technique developed by Charnes et al. (1978). It has become an increasingly popular tool in operational research. It has been applied in many different fields such as information and communication technologies, management evaluation, education (schools, universities), the banking industry (banks, branches), health care (hospitals, doctors), courts, manufacturing, and regional economics.[12] DEA assesses the relative efficiency of complex entities that convert multiple inputs into multiple outputs. The entities that are assessed, in our case experiments conducted in Fermilab, are usually called Decision Making Units (DMUs). Assessed entities are called "decision making", because the process of converting resources into outcomes is deliberate and directed and can be revised (Cvetkoska 2011). The creators of DEA defined the efficiency

[11] Furthermore, one can use this information as a predictor about the project, with the help of machine-learning algorithms (neural networks, support vector machines, and logistic regression).

[12] For some examples of the application of DEA see Dokas et al. (2014) and Emrouznejad et al. (2014).

of a DMU as the ratio of the sum of its weighted outputs and the sum of its weighted inputs (Charnes et al. 1978).

Data Envelopment Analysis measures comparative efficiency, where efficiency is defined as the ratio of the weighted sum of decision-making units (DMU) output and the weighted sum of its inputs. This means that DEA measures efficiency with reference to the set of units that are compared with each other rather than with reference to some external measure. Commonly, the efficiency score is expressed as either a number between 0 and 100 % (in case of an input-oriented model) or from 100 % and beyond (in case of an output-oriented model). In our research we used the input-oriented model. The reasons for this choice are explained in the next section. If DMU in an input-oriented model has a score below 100 % this means that it is inefficient with respect to other units (Cvetkoska 2011). Unlike a typical statistical approach, which evaluates each producer relative to an average producer, DEA is an extreme point method and compares each DMU with only the "best" DMU.

Now we will explain the basic idea behind this extreme point method. If a given producer, A, is capable of producing Y(A) units of output with X(A) inputs, and if producer B is capable of producing Y(B) units of output with X(B) inputs, then other producers should also be able to do the same if they operate efficiently. Producers A, B, and others can then be combined to form a composite producer with composite inputs and composite outputs. A composite producer is usually called a virtual producer because it does not have to exist. DEA finds the "best" virtual producer for each real producer. The original producer (producer whose efficiency is being estimated) is deemed inefficient in case that virtual producer gives more output with the same input (or gives the same output with less input) for input-oriented and output-oriented models respectively, than the original producer (http://www.emp.pdx.edu/dea/homedea.html). The procedure of finding the best virtual producer can be formulated as the following linear programming task:

$$\begin{gathered}\max \mu^T Y_k + u_* \\ \mu, v \\ s.t. \\ v^T X_k = 1 \\ u_* e^T + \mu^T Y - v^T X \leq 0 \\ \mu^T \geq \epsilon, v^T \geq \epsilon\end{gathered}$$

Here μ (vector of output weights) and v (vector of input weights) are only variables in a model. X and Y are matrices of input and output, respectively. From this model, called the weighted model, we can derive a data envelopment model such as the one presented below. A data envelopment model is a dual problem[13] of the above-mentioned model and it is preferred when the number of DMUs is greater than the number of inputs and the number of outputs. Thus we get the following:

$$\begin{gathered}\max Z - \epsilon(e^T s^+ + e^T s^-) \\ \theta, \lambda \\ s.t. \\ Y\lambda - s^+ = Y_k \\ ZX_k - X\lambda - s^- = 0 \\ \lambda, s^+, s^-, \epsilon \geq 0\end{gathered}$$

[13] Each problem in linear programming, a so-called primal problem, can be converted into a dual problem that provides an upper limit for an optimal value of the primal problem.

where Z is vector of factors' intensities, which explains the possibility of a DMU's reducing its inputs. Dual variables $s+$ and $s-$ show the possibility of a DMU's increasing its outputs and reducing its inputs, respectively. Dual variable λ presents the weight or importance of a DMU. In our application, ϵ is set at 0.0001.

Applying DEA to Fermilab data

The data, the choices, and the limitations

The choice of DEA model and data

When using the DEA technique it is important to make several choices (Koetter and Meesters 2013). We had to decide whether an input- or an output-oriented model was more appropriate for our research. An input-oriented model is more appropriate when one wants to establish a way of maximizing the results (outcomes) using existing resources (inputs), while an output-oriented model is suitable when one wants to establish how to achieve equally good results with fewer resources (Koetter and Meesters 2013). The policy-minded motive of our research is to determine whether it is possible to improve the results within physics laboratories with existing resources. In such a case, we are dealing with input-oriented models in which inputs are fixed. We used the variable returns to scale model because the output does not linearly increase with the input. For instance, if a 2-year project is extended for 6 months, this does not imply that the number of weighted publications will increase by 25 %.

When it comes to the choice of a particular set of inputs and a particular set of outputs, certain limitations were imposed based on the availability of the statistical data.[14] We used the following as inputs: (1) the number of researchers within a project (experiment), (2) the number of teams within a project, and (3) the duration of the project expressed in hours.[15]

Although the size of a research unit is ideal in many respects for the analysis of the performance, the contribution of labour of various elements of a research unit such as senior staff, students, technical and other support staff is never equal, and various elements contribute to various aspects of the team performance (Horta and Lacy 2011; Von Tunzelmann et al. 2003). Thus, in our study we focus on the performance of primary researchers, while the technicians and support staff are not included in the analysis. These are the researchers listed on the proposals of the experiments by Fermilab: the number of researchers provided by Fermilab refers to the designers of the experiment who also oversee its performance, are actively involved in the analysis of results, data corrections, and in at least the key elements of calibrations of the equipment.[16] What really matters in our analysis are data on the actual differences in team sizes across experiments with respect

[14] Our choice of data was based on historical records, the HEP database we discuss in the following subsection, and interviews with the proposers of experiments in Fermilab, a former head of its Research and Development Department (during the period when some of the experiments we analyzed were performed), as well as several physicists currently and formerly involved in various experiments performed at Fermilab and CERN.

[15] It would be interesting to include the costs as well as more fine-grained distribution of labor within a project, such as the number of researchers within each team. However, these data were not publicly available.

[16] See Hoddeson et al. (2008, esp. section 7) on the constitution and labour distribution among the teams in Fermilab.

to primary researchers, rather than the actual number of all the people involved: we are interested here in the operation and efficiency of the primary research team, so adding support staff and technicians to each experiment would not change the analysis since they are typically not involved, or are much less significantly involved, in key decisions concerning the design, commissioning, and performance phases of the experiments. (As mentioned above, much of the technical and even administrative work at Fermilab was done by physicists themselves, so the technical staff—in the sense of independently hired engineers who take care of the equipment—are often quite minimal.) This is why some teams may seem surprisingly small given the complexity of high-energy physics experiments, in some cases comprising only two researchers. Also, in some cases the initial researchers or an entire sub-team will be replaced, but this does not significantly affect the overall size of the teams.[17]

The number of teams is a limited input compared to the number of researchers. Researchers involved in a project function as a master-team but they are also broken down into smaller groups based on various criteria. Some groups focus on a particular subject within the master-project but the members of such groups also collaborate substantially across the groups at different stages of the experiment, as we have explained. This is why we opted for academic affiliations in order to delineate teams. The researchers working on an experiment gather into small teams at their home institution and as such join the experiment. They also typically collaborate as a team in many other projects, being involved together in planning and designing each joint collaboration, and they share their expertise at each stage of the experiments (if for no other reason, simply due to the proximity to each other at the home institution). In experimental and applied physics, they often pioneer a particular approach or an instrument and offer it to larger projects. Potentially, other fine-grained team connections could be tested as well, but we opted for a fairly obvious delineation given the nature of the study.

When it comes to the output, we based it on the number of papers associated with each experiment, where each paper is weighted based on citations divided into six categories.[18] The number (and the list) of the associated papers, as well as this citation-based categorization of them had been provided by the Fermilab archives. The categories are the following: renowned papers (500 + citations), famous papers (250–499 citations), very well-known papers (100–249 citations), well-known papers (50–99 citations), known papers (10–49 citations), less known papers (1–9 citations), and unknown papers (0 citations). Now, for each experiment every citation-based category is used as a separate output variable, where the value of a variable presents the number of papers in that citation-based category (associated to the experiment). It is also important to notice that citations present a global number of citations, excluding self-citations. We excluded self-citations since they affect the objectivity of the output (MacRoberts and MacRoberts 1989; Fowler and Aksnes 2007).

[17] This is why the occasional replacement of researchers and graduate students, or those added to the project at later stages, are indicated on a secondary list of collaborators in Fermilab's archives, not in the proposal section of the archive. This secondary list also contains the names of researchers who worked and published based on the results of the performance after the experiment was finalized, but who were not necessarily directly involved in it. Collaborations of this sort are undertaken for the purpose of producing papers based on the results of the experiment and do not necessarily overlap with the collaboration that designs and performs the experiment, which interests us here. We discuss the relationship between the primary collaborations and these wider collaborations in "Analysis" section.

[18] The categories are a property of the database we used. Please see next section for the explanation concerning the use of this particular scaling.

Now, such an output is a reasonably appropriate way to assess the efficiency of projects in physics. The reason for this is that consensus in the community of physicists is reached relatively quickly and it remains stable (Kragh 2002; Franklin 1990).[19] We decided to use citation analysis as a measure of project efficiency because citation is a very good indicator of the impact and perceived significance of a project within the scientific community (Garfield et al. 1964). It is important to note that citation is not necessarily a measure of the inherent value or quality of a paper. On the one hand, some high-quality work can go unnoticed and uncited for various reasons, perhaps because it is ahead of its time or because it has been published in a less widely read language journal. Also, low-quality papers can have high citation rates because they are mistaken or controversial (Martin and Irvine 1984). Thus, keeping in mind the fact that it is not possible to obtain an absolute measurement of the quality of scientific research, we decided to use citation analysis as the most straightforward available evidence of papers' impact on the scientific community.

As far as the kind of the experiments we analyzed goes, we excluded the calibrations of instruments, precision measurements, and the so-called strings of experiments conducted in Fermilab during the designated time period. Calibrations of instruments only serve the purpose of testing and adjusting new instruments; we therefore do not consider them physics experiments in a narrow sense. Precision measurements are used for establishing the exact properties of a known phenomenon and we therefore did not consider them discoveries in the same sense as other experiments. Finally, the treatment of strings of experiments would be problematic for practical reasons, because one experiment, its design and performance, is built on previous ones. It is not therefore clear how we could evaluate their results, i.e. jointly or separately.[20] Since the considered experiments are not linked into strings, we essentially only compare experiments that start from scratch. Moreover, after excluding calibrations of instruments, we are left with scientific experiments, whose success rate and outcomes were unknown when they were proposed. Finally, all projects were performed in the same laboratory during a fixed time-period. For all these reasons, the listed experiments can be safely compared. Thus, since all studied research projects are hosted by the same institution, extend over the same time-period (and belong to the second phase of the development of the laboratory or its edges), and since projects were filtered to measure similar effects, we believe that the projects (DMUs) are homogeneous.

Finally, physics is particularly suitable for data analysis of the sort we performed given that the convergence of results in physics, including particle physics, is generally relatively fast and reliable (Kragh 2002; Franklin 1990), which guarantees that the outputs of the research process, i.e. the results, can be compared in a rather non-controversial way. In addition, developments in modern high-energy physics, characterized by vast investments and mega-experiments with highly complex organizations are almost a natural target of studies like ours.[21] We will return to this issue in the analysis of the results.

[19] The experiment associated with a number of prominent papers is not necessarily an indication that it was the breakthrough event. It could be a part of a larger research trend instead. But given the citation pattern it was certainly successful in addressing the physical phenomenon within the prominent trend, and which continued to be prominent after the experiment was performed. In other words, the high citation record indicates minimally that the goal of the experiment and the performance has successfully become part of the trend.

[20] Some experiments are related closely in terms of their content, but they are not strings of experiments where essentially the same team applies for the next phase of the same experiment.

[21] The value of the data from Fermilab is undeniable, since it is one of the biggest and most successful physics laboratories in the world as we have pointed out. In more practical terms, since the number of individual experiments and laboratories in high-energy physics is very small relative to, e.g., biology, a

The sources

Our main source of data was INSPIRE, the high-energy physics information system. Physicists working in particle physics have used this database for decades as their own measure of productivity in terms of publications and citations. Thus, basing our analysis on this particular database means that the results should be relevant to productivity as it is determined by the studied research community, although, as we will see, this sort of criteria of productivity has its substantial limitations. A different methodology for gathering relevant data can certainly be used to assess productivity in this narrow sense of paper production and citations (e.g. the Google scholar search engine or the Thomson Reuters database), but the INSPIRE database is custom-made and developed for the specific purpose of assessing publications in HEP. It continuously tracks the relevant papers coming out of experiments, as well as it collects citations in a very comprehensive manner. The other databases are, comparatively speaking, quite generic so it is not clear that it would be beneficial to use these other databases even as supplements.

The database contains records of all Fermilab experiments. These experiment records include information such as the name and number of the experiment, when it was proposed, when it was approved, when it started, when it was completed, and the name of the spokesperson. These records also include links to all papers related to each experiment, including the original proposal for the experiment.[22] Each experiment proposal contains the following data, which we used as inputs: the number of the researchers and the institutions with which they were associated, and the duration of the experiment, expressed in hours. It also contains an overview of the goals and methods of the experiment, which we used in "Results" section to analyze the results of the DEA simulation.

When it came to the output data, i.e. the number of papers and their citation analysis, INSPIRE is a thorough database as well. It uses Invenio[23]—an integrated digital library system developed at CERN and with other content sources: The Astrophysics Data System, CERN, Oxford University Press and INSPIRE users.[24] For every experiment on the INSIRE website there is a citation summary[25] that contains data we used as outputs: total number of papers divided into six categories based on the citations.[26] This particular scaling of citations is a property of the Inspire database—and thus an integral part of the criterion of productivity the HEP community has used for decades. We thus assimilated this property of the database in our DEA; DEA tuned to this particular scaling, rather than a

Footnote 21 continued
laboratory such as Fermilab was one of the natural options, both because of the number of experiments performed within it over the years, and because of the a high quality of record of experiments in its electronic archives.

[22] Here is an example record for a Fermilab experiment (E-104): http://inspirehep.net/record/1110215. If you click on the "HEP articles associated with FNAL-E-0104" link, it will take you to the list of all the papers associated with the experiment, which should give you an overview of the results of the experiment. There are usually links to the full text of the papers, and the earliest item in that list is usually the experiment proposal.

[23] Invenio is an open source software package which provides framework and tools for building an autonomous digital library server. For more details see: http://invenio-software.org/.

[24] For the full list of INSPIRE content sources see: https://inspirehep.net/info/general/content-sources.

[25] For the description of citation metrics available in Invenio see: http://inspirehep.net/help/citation-metrics#citesummary.

[26] As an example of the citation summary of an experiment see: http://inspirehep.net/search?ln=en&ln=en&p=693__e%3AGNO&of=hcs.

different one, makes the results useful to the HEP community on its own terms of productivity (as does the use of INSPIRE database in general). This is why fixing a scale at a *somewhat different* count of citations would not be beneficial, while it would not substantially change the outcomes of our (or a similar sort of) analysis either. Also, when we consider the range of the numbers of citations (between zero and several hundred at most) six different scales applied to citation rates for each paper ensure that transitions from one scale to another are not too large, thus eliminating the extent of arbitrariness that would characterize a scale of only two or perhaps three categories.

The general limitations

In general, despite their efficiency and clarity, quantitative methods of analysis that focus on team composition and citation records have their limitations. The productivity of researchers and laboratories is multidimensional and cannot be reduced to publications. It also includes educational practice and external relationships, as well as wider benefits of the researchers' work (Torrisi 2014). In recent years research projects have also encouraged the collaboration of research teams, organization of conferences, open source libraries for others, or for profit, as extra outcomes. The factors that influence productivity are wide-ranging as well. Besides the education and experience of team members, they also include family, the institution's location, social context, women's scientific productivity, and perceived satisfaction and organisational atmosphere (ibid. 757). Although our DEA does not focus on these factors, we take relevant available data concerning some of them (seniority of team members and collaboration patterns) into account and discuss them in the analysis.[27]

It should be mentioned that one of the concerns in DEA models, especially for variable return to scale models (Cooper et al. 2011) is the problem of overfitting. Our research does satisfy rule-of-thumb constraints for the number of DMUs, which is $n \geq \max\{ms, 3(m+s)\}$, where n is the number of DMUs, m is the number of inputs, and s is the number of outputs (in our case $n = 27$, $m = 3$, $s = 6$). Also, we set ϵ to be greater or equal to 0.0001. Additionally, we created more constraints for the DEA model in order to ensure our efficiency score did not overfit. Those constraints are:

$$\mu_k - \mu_{k+1} \geq 0, \quad \forall k,\ k = 1, \ldots, s-1$$

This set of constraints ensures that more important papers have higher weight in the DEA model, while preventing projects with a higher number of lower quality papers from emerging as efficient projects.

One major drawback of the DEA is that it is an extremal method in which all extreme points are characterized as efficient. This means that outliers could distort efficiency scores. In order to prevent this, a sensitivity analysis is conducted. We performed univariate and multivariate outlier detection. The univariate outlier test was performed using the standard Z score. Several samples had high scores (between 2 and 2.5). However, none of the samples had unusually high score (Z score $\geq |2.5|$). For multivariate outlier detection we performed the Mahalanobis D^2 score, which is multidimensional version of a Z score. It measures the distance of a case from the multidimensional mean of the distribution, given

[27] Some of the available data concerning individual team members are either not complete or are not suitable to be used in DEA, the way that we utilized the data on the factors we tested. Also, the data on the financial resources of the projects could be beneficial as well but they are not available for individual experiments we included in our analysis.

the covariance of the distribution (Ben-Gal 2005). None of the probability scores were ≤ 0.001, meaning that none of the samples were outliers.

Additionally, we provided results from a Jackknife re-sampling DEA, which estimates variance and the bias of our dataset. It works as follows: one DMU is dropped at a time and the remaining DMUs are used to compute DEA scores. This procedure is repeated $M - 1$ time (each DMU is dropped exactly one time) (Wu 1986). When the procedure is finished the average DEA scores and standard deviation are calculated and analyzed. This procedure provides an indication about the presence of outliers, where dropping a DMU may change the scores significantly.

Results

The results of the DEA are based on data from twenty-seven experiments conducted in Fermilab whose requests were submitted in the period from September 1981 to August 1995, as presented in Table 1. The experiments are indexed in the first column. The second column represents decision-making units, i.e. Fermilab projects.[28] In the third column the efficiency of each project is presented. Finally, from the fourth column we can read off the benchmarks, i.e. how much each project should be improved with respect to the other indexed projects.

Based on the results of the quantitative analysis, six experiments turned out to be efficient, while the remaining twenty-one are inefficient to various extents relative to those six. Results from the Jackknife DEA are presented in Fig. 1. On the X axis we present the mean DEA efficiency, while on the Y axis we present DMU. The mean efficiency score is presented in black lines, while the 95 % confidence interval is presented in red lines. As can be observed, standard deviation is small, leading to the conclusion that there are no outliers (as stated in the previous section) in the dataset. The highest standard deviation is seen in project 0743, where the standard deviation is 0.069.

The following six experiments stand out as efficient: Experiment 0882 addressed the long-standing fundamental issue of the possible existence of monopol particles. It aimed at discovering the energy range within which such particles could exist, so the results, in the case of a successful run, were potentially ground-breaking for any future experiment addressing the issue. Experiment 0854 successfully introduced a new experimental technique that was widely applicable across particle searches. 0792 tested the microphysical properties by bombarding molecules of gold, which potentially has wide applicability across the physical sciences. 0774 was a discovery experiment that used the Fermilab's edge over other laboratories in having the highest energy of electron beams to test the possibility of the existence of particles that were postulated based on an anomaly previously detected in electron/positrons collisions. The experiment definitely eliminated this possibility. 0770 was more of a re-run, rather than a continuation of 0744 (i.e. a string), so we included it in the study. It looked at di-muon decays, which are crucial in indirect neutrino detection (neutrinos are not charged so cannot be captured by an electromagnetic field). It was one of the initial experiments that brought about a renaissance of the so-called fixed target technique (as opposed to colliding beams) at the achievable energy edge of the time (around 1TeV). 0713 was a search that definitely eliminated the possibility of highly ionizing exotic particles (other than monopoles).

As far as the least efficient experiments go, number 0745 yielded no considerable results given the considerable resources it used over 5 years. Number 0711 provided "thin"

[28] The full list of considered Fermilab projects with their details can be found at the following address: http://ccd.fnal.gov/techpubs/fermilab-reports-proposal.html.

Table 1 DEA efficiency scores and benchmarks

	DMU	Efficiency	Benchmarks			
1	FERMILAB-PROPOSAL-0882	1.0000				5
2	FERMILAB-PROPOSAL-0871	0.4188			7 (0.4038)	12 (0.5962)
3	FERMILAB-PROPOSAL-0868	0.3276		5 (0.6202)	7 (0.3103)	12 (0.0694)
4	FERMILAB-PROPOSAL-0866	0.4019				12 (1.0000)
5	FERMILAB-PROPOSAL-0854	1.0000				16
6	FERMILAB-PROPOSAL-0802	0.5022		5 (0.0045)	7 (0.0045)	12 (0.9910)
7	FERMILAB-PROPOSAL-0792	1.0000				14
8	FERMILAB-PROPOSAL-0789	0.3002		5 (0.5935)	7 (0.2010)	12 (0.2055)
9	FERMILAB-PROPOSAL-0774	1.0000				5
10	FERMILAB-PROPOSAL-0773	0.2814		5 (0.8072)	7 (0.1255)	12 (0.0672)
11	FERMILAB-PROPOSAL-0772	0.9432				12 (1.0000)
12	FERMILAB-PROPOSAL-0770	1.0000				24
13	FERMILAB-PROPOSAL-0769	0.3066		5 (0.1227)	7 (0.5331)	12 (0.3441)
14	FERMILAB-PROPOSAL-0761	0.2758		5 (0.7059)	7 (0.1032)	12 (0.1908)
15	FERMILAB-PROPOSAL-0760	0.5000	1 (0.0031)	5 (0.0179)	9 (0.0170)	12 (0.9620)
16	FERMILAB-PROPOSAL-0756	0.4261			7 (0.1303)	12 (0.8697)
17	FERMILAB-PROPOSAL-0747	0.2102			7 (0.9421)	12 (0.0579)
18	FERMILAB-PROPOSAL-0745	0.1447		5 (0.3952)	7 (0.5920)	12 (0.0128)
19	FERMILAB-PROPOSAL-0744	0.3641		5 (0.4387)	7 (0.4566)	12 (0.1047)
20	FERMILAB-PROPOSAL-0743	0.1649		5 (0.0817)	7 (0.8140)	12 (0.1043)
21	FERMILAB-PROPOSAL-0735	0.5000	1 (0.0828)	5 (0.1049)	9 (0.0090)	12 (0.8033)
22	FERMILAB-PROPOSAL-0733	0.2843		5 (0.7699)	7 (0.1372)	12 (0.0929)
23	FERMILAB-PROPOSAL-0713	1.0000				12 (1.0000)
24	FERMILAB-PROPOSAL-0711	0.2009			7 (0.4108)	12 (0.5892)
25	FERMILAB-PROPOSAL-0706	0.1667	1 (0.0001)	5 (0.2005)	9 (0.0029)	12 (0.7966)
26	FERMILAB-PROPOSAL-0705	0.2500	1 (0.0007)	5 (0.5857)	9 (0.0024)	12 (0.4112)
27	FERMILAB-PROPOSAL-0704	0.2087				12 (1.0000)

results as well, and its hypothesis seems to have turned out problematic. The results of 0706 have still not been fully provided given the excessive amount of data it generated likely due to problems with data analysis. 0704 introduced the novel technique of beam tagging for measuring gluon contribution to the spin of the proton—a technique potentially widely applicable across high-energy experiments—but its results have still not been fully disclosed, again perhaps due to data analysis problems or problems with the equipment. 0743 was a precision-measurement experiment of sorts, but also transferred new experimental technique from another experiment, and did so comparatively unsuccessfully (compared to e.g. 0854).

Now, one way to measure the influence of research project inputs on the outputs is by inspecting correlations. This is presented in Table 2. The values in Table 2 have been calculated using the Pearson correlation coefficient and they are presented in bold if the value is statistically significant, meaning that an increase or decrease in one variable significantly relates to increases or decreases in the second variable, at 0.05 level. Similarly, if the value is presented in bold and italic then the correlation coefficient is

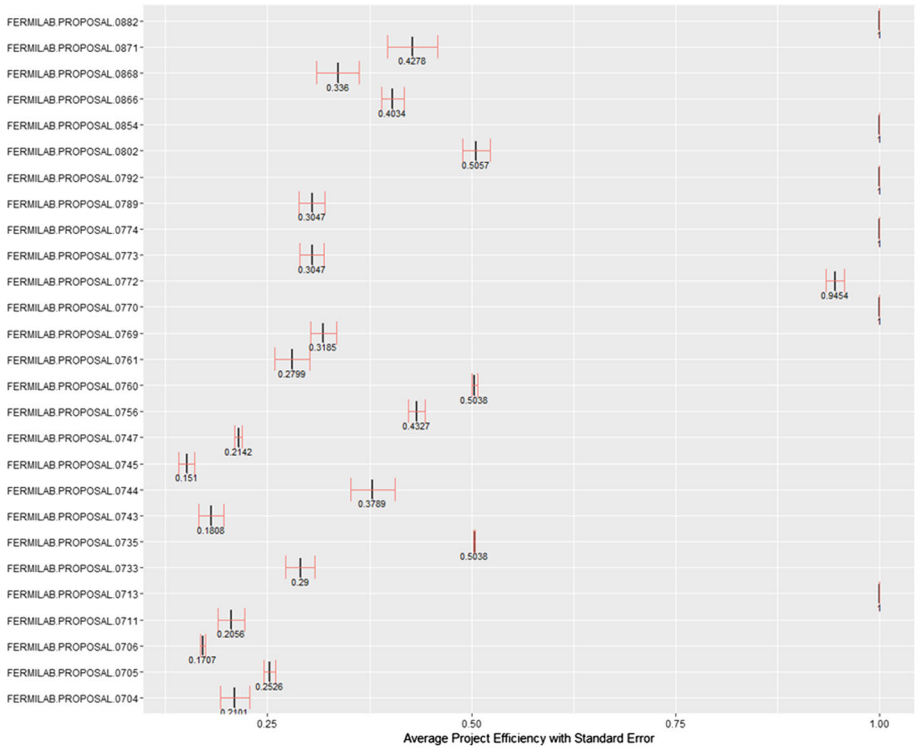

Fig. 1 Jackknife DEA results

statistically significant, at 0.01 level. If the correlation coefficient is close to 1, the box is green, and if the correlation coefficient is close to −1, the box is red.

Also, as expected, the length of a project displays a positive correlation with the number of papers, more specifically in the category of *well-known papers* (WK papers), *known papers* (K papers), and *unknown papers* (U papers). However, the length of the project does not correlate with the number of *famous papers* (F papers) and *very well-known papers* (VWK papers).

After the efficiency calculation we calculated the correlation coefficient of each input and output variable to efficiency (Table 3). The coefficient was expected to show a negative correlation to inputs and positive correlation to outputs. In our experiment we obtained only two statistically significant correlations. More specifically, efficiency is negatively correlated with the number of teams and the number of researchers in a project, meaning that higher the efficiency the lower the number of teams and researchers.

Analysis

The smaller teams are more efficient

It is striking that *all six efficient experiments were very small in terms of the number of researchers*, despite the fact that their goals—i.e. new experimental technique, application,

Table 2 Correlation between variables in DEA model

	Time {I}	Teams {I}	Res.{I}	F Papers {O}	VWK Papers {O}	WK Papers {O}	K Papers {O}	LK Papers {O}	U Papers {O}
Time {I}	1.0000	−0.0066	0.3777	−0.0117	−0.1094	**0.4310**	***0.5880***	0.1295	**0.3990**
Teams {I}		1.0000	***0.7270***	0.0613	−0.1214	0.0215	−0.0370	0.0775	0.2257
Researchers {I}			1.0000	0.0730	−0.0341	0.3480	0.3179	0.2668	***0.5080***
F Papers {O}				1.0000	***0.5700***	**0.4540**	0.1487	0.2750	**0.3830**
VWK Papers {O}					1.0000	0.3670	**0.4340**	**0.4670**	0.3001
WK Papers {O}						1.0000	***0.6470***	**0.4240**	***0.8060***
K Papers {O}							1.0000	***0.6080***	***0.6180***
LK Papers {O}								1.0000	***0.6290***
U Papers {O}									1.0000

Bold - Correlation is significant at the 0.05 level (2-tailed).

Bold Italic - Correlation is significant at the 0.01 level (2-tailed).

Bold—correlation is significant at the 0.05 level (2-tailed)

Bold Italic—correlation is significant at the 0.01 level (2-tailed)

Table 3 Correlation of variables with efficiency

Variable	Corr.
Time {I}	−0.271
Teams {I}	***−0.699***
Researchers {I}	***−0.682***
Famous Papers {O}	0.199
Very Well Known Papers {O}	0.0106
Well Known Papers {O}	−0.058
Known Papers {O}	−0.225
Less Known Papers {O}	−0.091
Unknown Papers {O}	−0.165

Bold Italic—correlation is significant at the 0.01 level (2-tailed)

discovery—were very different.[29] In half of these cases a small number of researchers was

[29] Sometimes an experiment falls in a grey area in terms of its goals, combining e.g. the introduction of new techniques and their wide application across physical sciences. This is typically not the case. Also, sometimes an experiment can aim at a particular goal, e.g. the testing of a new technique, but result in the major discovery of a new particle (e.g. discovery of J/PSI particle). This happens very rarely, however, as most experiments give results within their set goals.

divided into two teams.[30] In contrast, *the inefficient experiments are all very large (sometimes several dozen primary scientists) and divided into several teams* (between six and eleven). Although this effect is certainly not necessary, perhaps one would expect the small experiments to be somewhat more efficient. But the extent of this is surprising. The small ones very simply proportionally widely outperform the bigger ones in terms of citations and publications.[31]

What is also striking is that *the length of both the efficient and the least efficient experiments varies greatly*. The efficient ones took between 1 and 7 years, while the inefficient ones took between 1 and 9 years. It seems then that whether one ends an experiment fast or extends it considerably does not make much of a difference in terms of the outcome. [This result might be explained in part by an unsurprising tendency of the researchers to save the unfruitful projects by asking for extensions instead of quitting (Hoddeson et al. 2008, p. 280).] Overall then, in seeking the optimal composition of research teams, if teams are sufficiently small and there are many of them one can let them run their course for a reasonably long period of time.

The results concur with the results of a wide study across the sciences, the conclusion of which is that "contrary to common wisdom, in almost all areas of the most productive institutes are not found in largest size classes, but in the small one. In agriculture, environment, chemistry and physics, the most productive institutes have 5–6 researchers" (Bonaccorsi and Daraio 2005, pp. 103–104). A study of a large-scale biological institute (Carayol and Matt 2004) revealed similar results. This is also true at the level of production in a broader sense, including the teaching time and other contributions of an individual researcher—those in smaller labs will be more productive (Carayol and Matt 2006).

Our results also agree with the conjecture of a curvilinear relationship between team size and efficiency (Nieva et al. 1985). While a project run by a single researcher is not feasible in high-energy physics, an overly increased team size reflects negatively on efficiency. This negative effect may not be too surprising, since one would expect the coordination of a large group to be significantly more complex and less efficient than the coordination of a smaller group for the reasons we have specified earlier. Yet an important question is whether there is a threshold beyond which inefficiency jumps significantly. Finally, a study by Horta and Lacy (2011) shows that when a wider performance beyond publication and citations, and within a wider organizational structure are taken into account, the scientific profile of individual academics will be affected in a curvilinear manner.

It is expected that factors other than team composition contribute to particular groups of experiments' emerging as efficient. Thus, experiments that might have wide application across physics sub-fields are to be preferred. Yet not all experiments within this or any other group are equally efficient; we are interested in why some experiments within a given group turned out to be efficient. Nor are all or most of one group more efficient than any in another group. Moreover, none of the experiments from these groups could have been necessarily efficient, nor turned out to be a template of efficiency. Thus, it is quite

[30] As we pointed out, the indicated team members really worked with the support staff and technicians so the actual team size was larger. But again, this is not important since we only need data on the differences between the primary members of teams.

[31] Changing the relative weight of the publications (i.e. the weight of a large number of less significant publications compared to the weight of a smaller number of more significant publications) would not change this significantly.

reasonable to assume, based on the results, that tested factors are a major contributing factor to the efficiency of the experiments, or set a limit on the efficiency.

One could also group the experiments differently based on the three major groups, namely those with wide applicability, those that introduce new experimental techniques, and those that provide new insights into particle properties—and thus test efficiency within each group. However, this is a further development of the project warranted by the plausibility of our current results. Also, the goal of such analysis would be somewhat different, since the overall efficiency of experiments performed in the same laboratory would not be tested in the same way as we do here.[32]

The size and diversity of teams

When explaining how exactly the team size brings about a difference in its performance, one of the aspects we ought to look at based on the previous studies we have outlined is how team size facilitates diversity. The results seem to show how intricate may be the connection between team composition and the generally beneficial requirement for diversification. It may be that in the circumstances of a high-energy laboratory too small a team is not an obstacle to efficient performance, and that on the whole thus it does not prevent diversification. But given that all the teams draw on the same pool of researchers, namely international physicists, are not larger teams more diverse simply because they can house more members? Smaller teams—just in virtue of the number of researchers—cannot harbor the same extent of diversity as larger teams. Also given the expected contribution of diversity to the performance, should not we expect larger teams to perform better, not worse?

The smaller teams may not match diversity at the level of individual members to that of larger teams but neither do they suffer from the difficulties in communication that the increased size brings about. Also, the benefit of very efficient cross-expertise, a staple of physics experimentation, in smaller teams may outweigh the individual diversity of bigger teams: although larger master-teams working on individual experiments have the potential for fertile cross-expertise, the extent to which they can realize it diminishes substantially—probably in a curvilinear manner with the increase in size.

Also, at the level of overall exploration in a particle physics laboratory the strategy of letting a pool of smaller teams make decisions and perform experiments may be beneficial. The second kind of diversification we discussed earlier, namely diversification through looser connectedness between research teams, allows for autonomy in approaches to the physical phenomena of interest. It is likely that with a multitude of small experiments diversification happens at a higher level of the community of physicists when project choices are made. Thus, smaller teams enable more researchers to pick hypotheses they deem worthy of experimental testing than would be the case were they all gathered into bigger research teams that could test only a few hypotheses. It is thus possible that small teams are better at picking their research goals and hypotheses (i.e. those hypotheses that are more likely to be successfully tested) then big hierarchical teams. This may not be surprising give the findings by Heinze et al. (2009). They found the correlation between small-sized teams and creativity (as the qualitative trait of performance) in teams working on nanotechnologies. Moreover, although the results for the quantity of the performance are not conclusive, the quality of performance of research groups seems to be affected in a

[32] A future study could also look at the size of each sub-team of the master-team in the experiment, along with the overall number of researchers, and thus determine optimal size of each team within the experiment.

curvilinear manner across the board (ibid., 612; Andrews 1979). What is lost in the possibility of diversification in large teams is more than made up by the functional relationships in smaller teams at the level of overall exploration: "Small groups typically show a lack of hierarchical decision-making in relation to research activities. The flat structure of communication, with no difference in communication between formal hierarchical levels, fuelled the dynamics regarding creative research accomplishments. Furthermore, small-group size fosters productive mentor-students relationships that larger groups have difficulty to establish and maintain" (Heinze et al. 2009, p. 617). Similarly, smaller teams may not suffer from the "atmospheric consequences" (lower commitment) and communication distortions that typically plague large scientific research teams (van der Wal et al. 2009, p. 319).

Other contributing factors to team and project performance

Even though our results indicate that the size and number of teams itself substantially affects performance, there may be other factors contributing to the inefficiency of low-performing experiments. In terms of the relevant properties of individual researchers, for instance, there is an important concern outside the scope of DEA application that is worth exploring further: The trade-off between research and teaching may significantly affect the performance of academic researchers in their laboratory work (Horta and Lacy 2011; Olsen and Simmons 1996).[33] In addition, sometimes larger teams have more researchers without PhDs, which affects the performance negatively because the team members have less experience, or have extensive doctoral dissertation obligations that are external to the project.[34] We compared the inefficient projects to efficient ones based on the percentage of senior[35] and junior researchers[36] (Table 4). The difference in seniority levels between efficient and inefficient experiments is of a much lower order than the difference in terms of the team size and number of teams, but it cannot be excluded as a secondary contributing factor. The data are indicative of the possibility that somewhat lower seniority may have contributed to the inefficiency of the experiments.

The specific nature of collaborations within and in-between teams is often an important determinant of team performance as well (Milojević 2014). Tightly-knit in-house teams are often better performing, and their performance may depend on their previous experience of working together. In our case, across the most successful and the most unsuccessful experiments some master-teams were composed exclusively of members external to the

[33] One could, generally speaking, focus on the performance of individual academics instead of teams (see e.g. Horta and Lacy 2011).

[34] This is in fact the main reason why Fermilab management, as well as management in other HEP laboratories, insist that very specialized work done by graduate students in the lab can be turned into doctoral dissertations at their home universities (Hoddeson et al. 2008).

[35] We considered as senior researchers scientists that had obtained their PhD degrees by the time the project was proposed, as well as researchers who had been academically active for at least 12 years before the project was proposed and had at least 24 publications in the high-energy physics database <http://inspirehep.net/?ln=en> before the project in question was proposed and for whom we could not find the year they received PhD degree. Publications in the High-Energy Physics database include articles, books, conference proceedings, and project proposals. Specifically, for 75 % of senior researches we established that they had a PhD when the project was proposed, while the seniority of the remaining 25 % was determined indirectly. This choice was made due to data limitations and the fact that at the time experiments were conducted, in some countries, the degrees other than PhD or related research statuses were awarded instead.

[36] As junior researchers we characterised those for whom we conclusively established that they did not have a PhD degree at the time project was proposed, nor a publication record prior to the experiment.

Table 4 Types of researchers (in %) per experiment

	Proposal	Senior researcher (%)	Junior researcher (%)	Unknown (%)
Inefficient	745	61	9	30
Inefficient	711	67	25	8
Inefficient	706	66	31	3
Inefficient	704	69	2	29
Inefficient	743	74	5	21
Efficient	882	100	0	0
Efficient	854	100	0	0
Efficient	792	33	33	33
Efficient	774	100	0	0
Efficient	770	100	0	0
Efficient	744	61	26	13
Efficient	713	100	0	0
Inefficient		**67**	**14**	**18**
Efficient		**85**	**8**	**7**

laboratory (i.e. externally affiliated with Fermilab only during the project), some exclusively of internal (in-house) members, and some mixed (Table 5). Also, both the external and internal basic team units within individual experiments typically had a history of previous collaborations and co-authorship both in efficient and inefficient experiments. Mutual co-authorship is a good indication of prior collaboration, in our case and in general (See Abbasi et al. 2011). Yet the larger experiments gathered multiple basic teams that typically had not collaborated previously. Thus, the mutual adjustment between these teams took some time and possibly created various negative atmospheric effects over time, and as a result could have substantially affected the performance.

Table 5 Percentage of Fermilab employees in the experiments

	Proposal	Fermilab researchers (%)
Inefficient	745	6
Inefficient	743	0
Inefficient	711	33
Inefficient	706	21
Inefficient	704	0
Efficient	882	5
Efficient	854	100
Efficient	792	0
Efficient	774	100
Efficient	770	100
Efficient	744	26
Efficient	713	0
Inefficient		**11**
Efficient		**47**

It is also possible that the teams in the most efficient experiments substantially benefited from collaborations external to the experiments during the period of collaboration in their Fermilab project or even over a longer period of time. This external exchange can be a very important factor for the performance of scientific teams (Carillo et al. 2013). This concern may have to do with the issue of agglomeration and embedding of research teams into larger economic and institutional contexts we will discuss shortly.

The papers that came out of the experiments we looked at were written by wider collaboration groups. Thus, the composition of these wider collaborations—which were gathered over the years after the experiment was finalized—that analyzed the results, drew theoretical conclusions, and published them may have substantially affected the citation patterns. These collaborations typically only somewhat overlapped with the teams that designed and performed the experiment, which we focused on here. But their efficiency may have contributed to the overall efficiency of the primary experimental team we looked at, although their performance inevitably dovetails with the efficiency of this primary experimentalist team, i.e. those physicists who performed the experiment well. Thus, no extent of ingenuity on the part of the wider collaborations keen to churn out papers could substantially raise the efficiency of the primary experimentalist team if the experiment itself was bad—there would simply be not much to publish on, let alone cite. But an exceptionally good collaboration working on a paper based on results coming out of a good experiment 10 years after its completion could somewhat raise the number of citations, i.e. the factor measuring the efficiency of the primary team.

Possible policy implications

Our results show that groups of moderate size outperform larger ones in a major particle physics laboratory. Moreover, they support the hypothesis that the relationship between team size and efficiency is curvilinear. This is why it is valuable to compare relatively homogenous group of experiments and determine which ones are more efficient in terms of publications, irrespective of whether changing their structure is possible immediately. Thus, a possible way of making a particle physics laboratory more efficient, if efficiency is understood in the terms we set up in the model, is to establish it as a turn-around site for many smaller experiments rather than bet on a few large long-standing teams performing only a few experiments. In fact, as we have mentioned, this is exactly how Fermilab was deliberately organized during the tenure of R.R. Wilson (Hoddeson et al. 2008, Part 2), its first director, in the first half of the 1970s. Later on, it became a site for a comparatively small number of long-running strings of experiments.

A drop in efficiency with enlargement is informative and has policy implications, but the policy-change payoff may be expected only in the longer run. We are fully aware that there may be physical and technological limitations that may force experimenters to gather in large and thus much less efficient groups. It is often argued (e.g. Weinberg 2012) that certain particle physics experiments, especially those expected to break through can only be conducted if large numbers of researchers are involved due to physical and technological constraints. It would therefore be interesting to compare the efficiencies of experiments with the number of researchers higher than a specific fixed point. Yet the extent of these limits and the necessities that lead to decisions to go down the road of ever bigger experiments has to be carefully examined, both empirically and qualitatively. Alternative existing and potential experimental techniques and phenomena addressing

fundamental physical questions have to be entertained to the fullest extent in order to maximize the efficiency of research.

Now, even if the experiments tackling the same phenomena had to be as big as they are, it does not necessarily follow that they ought to be less productive. Yet the bigger teams are apparently less efficient—they churn out fewer and less cited publications, to put it simply, not only per individual researcher but overall. This is not what one would expect or desire given that they employ ten or more times more people and resources. It could have turned out that they were of the same efficiency, at least. Thus, the results are informative in any case with respect to the relation between size and performance: size is not expected to correlate with inefficiency, at least not to such an extent, but it does. This is the result that is informative to the overall setup of the exploration and it indicates that one should go smaller whenever one can in a particle physics laboratory.

It is usually possible to substantially reorganize research in physics through innovation, but this is often painstaking or impossible once the laboratory is already built. It is a matter of early choices when planning exploration and when setting up a big lab (Perovic 2011). In general, the same experimental goals can be pursued by various means and in various ways. One particle can be typically discovered by different techniques that require very different equipment and size of the experiment, and accordingly different team sizes and numbers of researchers. For instance, the first director of Fermilab, Wilson, a champion of small-science particle physics, "argued for Spartan guidelines as well in the design of experimental facilities and research equipment … because offering expensive facilities may tend to paralyze better developments later on" (Hoddeson et al. 2008, p. 66). This policy turned out to be beneficial on numerous occasions. For instance, "Wilson's alternate 200 GeV design, including research equipment and shielding, totaled less than $100 million, and the accelerator could, he estimated be brought to completion in 3 years. The contrast to Berkley's design, estimated at $348 million with completion over 7 years, was sharp" (Ibid.). Similarly, a British physicist proposed a design for another key piece of the accelerator that would cost three times less. In both cases the structure and the size of the teams that would perform experiments on the equipment would be substantially different, and based on our results, more efficient, than those on the original Berkley proposals (ibid., 67).

It is also plausible to assume, however, that even the existing inefficient experiments we analyzed could have been restructured so that they favored smaller teams in the same lab. For instance, it is quite plausible that #0704 and #0706 would not have got stuck at the level of data analysis had been teams restructured in a timely manner, guided by the benchmarks, or if the tasks were divided differently to enable operation with less personal and smaller teams. There is a number of ways in which each of the listed experiments could have been different, including restructuring of the teams, dividing them into smaller teams, and sometimes into a few individual experiments. Our analysis indicates the points at which such restructuring would have been beneficial.

In general, Fermilab, as we have mentioned, was restructured towards bigger and longer lasting experiments in the late 1970s, but it is possible that the same general goals of exploration could have been effectively pursued within the existing structure. In fact, the outgoing Director insisted on this strategy as he thought it was both physically plausible and effective (Hoddeson et al. 2008, p. 216). A fairly straightforward division into smaller independent experimental tasks is often possible, without waiting for revolutionary innovative experimental techniques. But we should bear in mind that the funding pressures and the pressure for fast, if selective, results is important when contemplating restructuring. The funding agencies had to be presented with firm results, even if that meant narrowing

the field of phenomena that were explored and negatively affecting research in the long run. Restructuring certainly involves both long-term and short-term tradeoffs between the desired goals and results.

Thus, first, our results may be relevant to numerous physics laboratories organized in a similar fashion. In fact, this is the majority of laboratories in physics and even in particle physics. The current organization of experimentation at CERN (Boisot 2011)—as much as it may play a central role in the discipline is an exception in terms of organization (highly centralized, with a few long-lasting experiments exploring a fairly small number of phenomena, where multiple teams perform specialized interrelated tasks pursuing a common set of research goals) and funding.

Second, the extent to which the size of the experiments mattered in the second, transitional period when the experimental teams were relatively small compared to the third, collider phase can provide guidelines for the assessment of efficiency across HEP laboratories as they are currently structured. Thus, if smaller teams are more efficient during the second, less centralized phase, this may indicate that efficiency in the third more centralized collider phase involving even bigger teams has been further reduced. Moreover, if this is the case in a fairly decentralized laboratory such as Fermilab, it may be even more likely in other laboratories, including the current leading laboratory CERN, which has been centralized and organized top-down from the very beginning (Krige 1993, Herman et al. 1987).

In fact, the centralization of scientific research across the sciences suffers from limitations brought about by various social factors that lead to a decrease in productivity. The performance of individual researchers, their publication rate and quality in large research sites that aggregate scientists at a particular location, are negatively affected (van der Wal et al. 2009). Instead of centralization providing a fall in the cost per research unit while increasing productivity, productivity in fact decreases, and the researchers from small sites were most likely to publish breakthrough papers in a few top science journals (Nature and Science) (ibid., 317). The aggregation may have the same effect across the sciences, including the particle physics mega-labs. Nor does the agglomeration of research in a highly developed and urban setting aiming to benefit from a wider economy necessarily have positive effects on productivity. In fact, in biomedical research agglomeration it does not have a positive effect on productivity (Bonaccorsi and Daraio 2005). Actually, the benefits may come from doing things the other way around: the scientists themselves build an environment from scratch, which will eventually start benefiting from agglomeration (Ibid., 97). Fermilab may be a good example of this, as it was built in a remote location while gathering external groups that created a high-quality research-conducive environment (Hoddeson et al. 2008). These are the reasons why our future research will compare the performance of a highly centralized HEP laboratory composed of specialized teams with the performance of a few independent laboratories exploring the same phenomenon.

Finally, when it comes to possible benefits with respect to the methodology we used, data-driven quantitative analysis and simulations are more advantageous than hypotheses-driven ones when testing the epistemic efficiency of concrete real-science situations, since their predicted results regarding the optimal distributions and structure of labor can be unambiguously applied. We used data envelopment analysis (DEA) in part as an illustration of the power of data-driven approaches. More specifically, we presented a novel application of DEA in analyzing the efficiency of a series of experiments in a HEP laboratory, but this approach could be applied widely. Predictive analysis is another methodological tool the potential of which needs to be explored with respect to the cases such as the one we explored. Note that a predictive data-driven analysis is only possible

after establishing project efficiency, since such an analysis is built upon efficiency results. Two obvious candidates for an automated predictive analysis are decision trees[37] and logistic regression[38] models. In the next stage of the research more elaborate models could be employed as well.

Acknowledgments This work was supported in part by the Project "Dynamic Systems in Nature and Society: philosophical and empirical aspects" (#179041) financed by the Ministry of Education, Science, and Technological Development of Serbia. The work of the third author was supported by the FWF project W1255-N23. We would like to thank the Fermilab History & Archives Project, Fermilab's lnformation Resources Department and the Fermilab Program Planning Office for providing us the necessary data and explanations about the INSPRE-HEP website. In particular we would like to thank Heath O'Connell, Adrienne W. Kolb, Valerie Higgins and Roy Rubinstein for their assistance. We would also like to thank Lilian Hoddeson for putting us in contact with Fermilab stuff, and Milan Ćirković for his important initial suggestions. Finally, we thank the two anonymous reviewers for their outstanding effort.

References

Abbasi, A., Hossain, L., Uddin, S., & Rasmussen, K. J. (2011). Evolutionary dynamics of scientific collaboration networks: Multi-levels and cross-time analysis. *Scientometrics, 89*(2), 687–710.

Agasisti, T., & Johnes, G. (2015). Efficiency, costs, rankings and heterogeneity: The case of US higher education. *Studies in Higher Education, 40*(1), 60–82.

Agrell, A., & Gustafson, R. (1996). Innovation and creativity in work groups. In M. A. West (Ed.), *Handbook of work group psychology* (pp. 317–344). Chichester: Wiley.

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science, 82*(3), 424–453.

Andrews, F. M. (Ed.). (1979). *Scientific productivity: The effectiveness of research groups in six countries*. Cambridge: Cambridge University Press.

Bantel, K. A., & Jackson, S. E. (1989). Top management and innovations in banking: Does the demography of the top team make a difference? *Strategic Management Journal, 10*, 107–124.

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rockach (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers* (pp. 131–146). Kluwer Academic Publishers/Springer.

Boisot, M. (2011). *Collisions and collaboration: The organization of learning in the ATLAS experiment at the LHC*. Oxford: Oxford University Press.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics, 63*(1), 87–120.

Brinkman, P. T., & Leslie, L. L. (1986). Economies of scale in higher education: Sixty years of research. *Review of Higher Education, 10*(1), 1–28.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel psychology, 46*(4), 823–847.

Carayol, N., & Matt, M. (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy, 33*(8), 1081–1102.

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

Carillo, M. R., Papagni, E., & Sapio, A. (2013). Do collaborations enhance the high-quality output of scientific institutions? Evidence from the Italian Research Assessment Exercise. *The Journal of Socio-Economics, 47*, 25–36.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European Journal of Operational Research, 2*(6), 429–444.

Cook, I., Grange, S., & Eyre-Walker, A. (2015). Research groups: How big should they be? *PeerJ, 3*, e989. doi:10.7717/peerj.989.

[37] Decision trees are graph structures used for identifying best strategies for a fixed goal.

[38] Logistic regression is a statistical model used for predicting the optimal division of resources.

Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). *Handbook on data envelopment analysis* (Vol. 164). New York: Springer.
Cvetkoska, V. (2011). Data envelopment analysis approach and its application in information and communication technologies. In M. Salampasis & A. Matopoulos (Eds.), *Proceedings of the international conference on information and communication technologies for sustainable agri-production and environment (HAICTA 2011)*, Skiathos, pp. 421–430.
Dokas, I., Giokas, D., & Tsamis, A. (2014). Liquidity efficiency in the Greek listed firms: A financial ratio based on data envelopment analysis. DEA window analysis approach for measuring the efficiency of Serbian Banks based on panel data. *Management* (1820–0222), (65).
Emrouznejad, A., Banker, R., Lopes, A. L. M., & de Almeida, M. R. (2014). Data envelopment analysis in the public sector. *Socio-Economic Planning Sciences, 48*(1), 2–3.
Fowler, J. H., & Aksnes, D. W. (2007). Does self-citation pay? *Scientometrics, 72*(3), 427–437.
Franklin, A. (1990). *Experiment, right or wrong*. Cambridge: Cambridge University Press.
Galison, P., & Hevly, B. W. (1992). *Big science: The growth of large-scale research*. Stanford: Stanford University Press.
Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: The Institute for Scientific Information.
Greenberg, D. S. (1999). *The politics of pure science*. Chicago: University of Chicago Press.
Hackman, J. R., & Vidmar, N. (1970). Effects of size and task type on group performance and member reactions. *Sociometry, 33*, 37–54.
He, F., Xu, X., Chen, R., & Zhang, N. (2015). Sensitivity and stability analysis in DEA with bounded uncertainty. *Optimization Letters, 10*(4), 1–16.
Heilbron, J. L., & Seidel, R. W. (1989). *Lawrence and his laboratory: A history of the Lawrence Berkeley laboratory* (Vol. 1). Berkeley: University of California Press.
Heinze, T., Shapira, P., Rogers, J. D., & Senker, J. M. (2009). Organizational and institutional influences on creativity in scientific research. *Research Policy, 38*(4), 610–623.
Herman, A., Krige, J., Mersits, U., & Pestre, D. (1987). History of CERN, vol. 1. *Launching the European Organization for Nuclear Research*. Amsterdam/New York: North-Holland Physics Pub.
Hoddeson, L. (1997). *The rise of the standard model: A history of particle physics from 1964 to 1979*. Cambridge: Cambridge University Press.
Hoddeson, L., Kolb, A. W., & Westfall, C. (2008). *Fermilab: Physics, the frontier, and megascience*. Chicago: University of Chicago Press.
Horta, H., & Lacy, T. A. (2011). How does size matter for science? Exploring the effects of research unit size on academics' scientific productivity and information exchange behaviors. *Science and Public Policy, 38*(6), 449–460.
Jackson, S. E. (1996). The consequences of diversity in multidisciplinary work teams. In M. A. West (Ed.), *Handbook of work group psychology* (pp. 53–76). Chichester: Wiley.
Katz, R. (1982). The effects of group longevity or project communication and performance. *Administrative Science Quarterly, 27*, 81–104.
Kimberly, J. R. (1981). Managerial innovation. In P. C. Nystrom & W. H. Starbuck (Eds.), *Handbook of organizational design: Adapting organizations to their environments* (pp. 4–104). Oxford: Oxford University Press.
Kitcher, P. (1990). The division of cognitive labor. *Journal of Philosophy, 87*(1), 5–22.
Kitcher, P. (1993). *The advancement of science*. New York: Oxford University Press.
Koetter, M., & Meesters, A. (2013). Effects of specification choices on efficiency in DEA and SFA. In *Efficiency and productivity growth: Modelling in the financial services industry*, pp. 215–236.
Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology* (Vol. 12, pp. 333–375). New York: Industrial and Organizational Psychology.
Kozlowski, S. W. J., & Hulls, B. M. (1986). Joint moderation of the relation between task complexity and job performance for engineers. *Journal of Applied Psychology, 71*, 196–202.
Kragh, H. (2002). *Quantum generations: A history of physics in the twentieth century*. Princeton: Princeton University Press.
Krige, J. (1993). Some socio-historical aspects of multinational collaborations in high-energy physics at CERN between 1975 and 1985. In *Denationalizing science* (pp. 233–262). Springer, Netherlands.
MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*, 342–349.
Martin, B. R., & Irvine, J. (1984). CERN: Past performance and future prospects: I. CERN's position in world high-energy physics. *Research Policy, 13*(4), 183–210.

Martin, B. R., & Irvine, J. (1985). Basic research in the East and West: A comparison of the scientific performance of high-energy physics accelerators. *Social Studies of Science, 15*(2), 293–341.
Martz, W. B., Vogel, D. R., & Nunamaker, J. F. (1992). Electronic meeting systems: Results from the field. *Decision Support Systems, 8*(2), 141–158.
Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences, 111*(11), 3984–3989.
Nieva, V. F., Fleishman, E. A., & Reick, A. (1985). *Team dimensions: Their identity, their measurement, and their relationships (Research Note 85–12)*. Washington, DC: U. S. Army, Research Institute for the Behavioral and Social Sciences.
Olsen, D., & Simmons, A. (1996). The research versus teaching debate: Untangling the relationships. *New Directions for Institutional Research, 1996*(90), 31–39.
Olson, B. J., Parayitam, S., & Bao, Y. (2007). Strategic decision making: The effects of cognitive diversity, conflict, and trust on decision outcomes. *Journal of Management, 33*(2), 196–222.
Page, S. E. (2007). Making the difference: Applying a logic of diversity. *The Academy of Management Perspectives, 21*(4), 6–20.
Page, S. E. (2011). *Diversity and Complexity*. Princeton: Princeton University Press.
Perovic, S. (2011). Missing experimental challenges to the Standard Model of particle physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics, 42*(1), 32–42.
Poulton, B. C. (1995). *Effective multidisciplinary teamwork in primary health care*. Unpublished doctoral thesis, Institute of Work Psychology, University of Sheffield, Sheffield, England.
Qurashi, M. (1991). Publication-rate and size of two prolific research groups in departments of inorganic chemistry at Dacca University (1944–1965) and Zoology at Karachi University (1966–84). *Scientometrics, 20*(1), 79–92.
Scharf, A. (1989). How to change seven rowdy people. *Industrial Management, 31*, 20–22.
Strevens, M. (2003). The role of the priority rule in science. *Journal of Philosophy, 100*(2), 55–79.
Torrisi, B. (2014). A multidimensional approach to academic productivity. *Scientometrics, 99*(3), 755–783.
Valentin, F., Norm, M. T., & Alkaersig, L. (2016). Orientations and outcome of interdisciplinary research: The case of research behavior in translational medical science. *Scientometrics, 106*(1), 67–90.
van der Wal, R., Fischer, A., Marquiss, M., Redpath, S., & Wanless, S. (2009). Is bigger necessarily better for environmental research? *Scientometrics, 78*(2), 317–322.
Von Tunzelmann, N., Ranga, M., Martin, B., & Geuna, A. (2003). *The effects of size on research performance: A SPRU review*. Brighton: SPRU.
Wang, J., Thijs, B., & Glanzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS One, 10*(5), e0127298.
Weinberg, S. (2012). The crisis of big science. *The New York Review of Books*, *59*(8). www.nybooks.com/articles/2012/05/10/crisis-big-science/. Accessed 10 May.
Weisberg, M., & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of Science, 76*(2), 225–252.
West, M., & Anderson, N. (1996). Innovation in top management teams. *Journal of Applied Psychology, 81*(6), 680–693.
Westfall, C. (1997). Science policy and the social structure of big laboratories, 1964–1979. In Hoddeson 1997.
Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*, 1261–1295.
Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 574–587.
Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis, 72*(1), 17–35.