3-2014

Assessing the Effect of High Performance Computing Capabilities on Academic Research Output

Amy Apon
Clemson University, aapon@clemson.edu

Linh B. Ngo
Clemson University, lngo@clemson.edu

Michael E. Payne
Clemson University

Paul W. Wilson
Clemson University

Recommended Citation

Assessing the Effect of High Performance Computing Capabilities on Academic Research Output

Amy W. Apon Linh B. Ngo Michael E. Payne
Paul W. Wilson[*]

March 2014

Abstract

This paper uses nonparametric methods and some new results on hypothesis testing with nonparametric efficiency estimators and applies these to analyze the effect of locally-available high performance computing (HPC) resources on universities' efficiency in producing research and other outputs. We find that locally-available HPC resources enhance the technical efficiency of research output in Chemistry, Civil Engineering, Physics, and History, but not in Computer Science, Economics, nor English; we find mixed results for Biology. Our research results provide a critical first step in a quantitative economic model for investments in HPC.

Keywords: Efficiency, frontier, nonparametric, inference
JEL Classification: C12, C14, C44, H52

Forthcoming, *Empirical Economics*

[*] Apon: School of Computing, Clemson University, Clemson, SC 29634; email: aapon@clemson.edu.

Ngo: School of Computing, Clemson University, Clemson, SC 29634; email: lngo@clemson.edu.

Payne: School of Computing, Clemson University, Clemson, SC 29634; email: mpayne3@clemson.edu.

Wilson: Department of Economics and School of Computing, Clemson University, Clemson, SC 29634; email: pww@clemson.edu (corresponding author).

Research support from U.S. National Science Foundation grant no. SMA-1243436 is gratefully acknowledged. Any remaining errors are solely our responsibility.

1 Introduction

High performance computing (HPC) instrumentation has become increasingly important in scientific research, and its use in research has expanded across diverse disciplines in recent years.[1] Numerous reports have emphasized the importance of investment in HPC instrumentation and its impact to science. For example, the President's Information and Technology Advisory Committee articulates the need for boosting federal funding levels for fundamental research in networking and information technology (PITAC, 1999). According to the National Academy of Sciences (NAS, 2008), increased investments in HPC are central to the nation's safety (i.e., national security) and competitiveness. The President's Council of Advisors on Science and Technology (PCAST, 2010) reports that, "By enabling transformative advances in science and technology, HPC helps maintain our historical leadership for future generations of Americans."

Brooks (1986) and Leyden and Link (1992), among others, have cited federal support for research as the foundation of innovation in the United States. Locally-available HPC instrumentation for academic research is funded by a number of federal agencies, including the U.S. National Science Foundation, the U.S. Department of Energy, the National Institutes of Health, etc. Universities may also fund HPC through endowment income, donations, etc. However, as a result of recent economic events, federal funds for basic and applied research are increasingly targeted for cuts, while at the same time universities' budgets face similar pressure. Difficult decisions must be made regarding the allocation of scarce funds—both federal and non-federal—to areas of basic research. Furman et al. (2002) note that at the federal level, decisions to support funding of HPC instrumentation that supports a variety of research—versus funding of narrowly-focused research activities—is a critical decision with

[1] HPC instrumentation is commonly referred to as "supercomputers," but is heterogeneous in nature. By "HPC instrumentation," or simply "HPC," we mean specialized computer systems designed to solve challenging computational problems that cannot be solved using desktop or similar, commonly-available machines. As such, what constitutes HPC instrumentation necessarily evolves over time; what was regarded as HPC in 1980, 1990, or even 2000 would be unremarkable today aside from historical interest. Today, HPC instruments typically consist of large clusters, where hundreds or thousands of central processing units (CPUs) in close proximity to each other are linked by a fast, tightly integrated network, but may also include specialized machines capable of producing extraordinary numbers of computations per second. Our use of "HPC" does not include grid computing systems such as Condor systems, which may offer high-throughput computing for certain types of problems, but which are not well-suited for applications that require a large amount of communication between CPUs.

far-reaching impacts, not only to the institutions receiving funds, but also, potentially, to U.S. national competitiveness. Even at the campus level, institutional decision makers must make funding decisions for information technology infrastructure and instrumentation in the face of competing priorities and scarce resources. The cost of a reasonably well-equipped campus HPC instrument (e.g., a cluster providing a few hundred trillion floating-point operations per second) is roughly $1 million in 2014. While the purchase of an HPC instrument is often treated as a capital expense, the relatively short lifetime (i.e., high rate of depreciation) of HPC instrumentation means that maintaining an HPC instrument that can support modern research is a recurring cost that repeats every three to five years. When additional costs for staffing, power, and building infrastructure are considered, the costs of operating a campus HPC instrument can be equivalent to the cost of operating a small or moderate-size academic department (or, depending on the discipline, perhaps even a large-size academic department).

Although the costs of investment in HPC instrumentation are quantifiable (some costs are hard to quantify, e.g., staffing that is shared across different administrative units), the benefits are understood less precisely. While many would agree that HPC is important for research, to date there are few quantitative measures of the effect of HPC and other information technology infrastructure on research output. Consequently, decision-makers at federal agencies that allocate funding, as well as research university administrators, are impeded from making cost-benefit comparisons when allocating scarce funds for HPC. Moreover, it is difficult to know whether past allocation decisions have been efficient in an economic sense.

This paper analyzes department-level data from research universities in order to understand and quantify the effect of locally-available HPC on research output. While many, in addition to the reports cited above, would agree that HPC has a positive effect on research output, until now the level of understanding about the relation between HPC and research output has been *assumed*, and *qualitative.* Our analysis provides statistical evidence of the (positive) direction of the relationship, as well as some quantification. This is a first step toward a cost-benefit analysis of HPC funding; here, we quantify the effect of HPC on research output, while leaving for subsequent studies the problem of placing values on research output. Nonetheless, having an idea of the quantitative relationship between HPC and research output provides an important component of information that decision-makers have lacked until now.

Our analysis is at the level of university departments in a given discipline. We examine eight disciplines: Chemistry, Civil and Environmental Engineering, Computer Science, Ecology and Evolutionary Biology, Economics, English, History, and Physics. We estimate, within each discipline, departments' technical efficiency in producing research output, and whether and how this efficiency is affected by locally-available HPC instrumentation. Of course, researchers can obtain access to distant HPC resources, either through collaboration with researchers at another institution, or by accessing nationally funded HPC resources such as those managed through the U.S. Department of Energy's INCITE program or from various instrumentation programs funded by the U.S. National Science Foundation (NSF).[2] However, almost anyone who uses HPC resources would likely agree that it is typically easier to use local resources than distant ones, perhaps in part due to the fact that support personnel for local resources are closer and have greater availability to local campus users. We do not expect that the presence of on-campus HPC instruments would affect all disciplines equally, or even that it would have an effect on all disciplines. Although our mix of disparate disciplines results in part from data availability as discussed below in Section 4, our mix provides controls for judging whether our results might be believable.

Measurement of the research performance of universities is not new, but as noted by Toutkoushian and Webber (2011, p. 140), "the state of the art in measuring institutional research productivity... is still fairly primitive." Many studies have ignored the scale of operation, and have focused on total research produced as opposed to efficiency or productivity in producing research output. Only a few papers have examined the relationship between HPC and research output. Among those that have done so, Kepner (2004), Sterling (2004), and Tichenor and Reuther (2006) examine accounting ratios, but do not explicitly model a production process. Apon et al. (2010) employ two-stage least squares, but this study examines research output at the institutional level, thereby aggregating across various diverse disciplines. However, different academic disciplines have different criteria for the production and evaluation of research products. Broadly-applied sets of institution-wide measurements to evaluate productivity potentially mask the effects of specific types of investments in indi-

[2] Examples of such programs funded by NSF are the Blue Waters project and the Extreme Science and Engineering Discovery Environment (XSEDE). One can also purchase time on HPC systems from Amazon and other providers.

vidual academic departments.

Parametric approaches to the measurement of research productivity are problematic due to simultaneity issues, as discussed by Bonaccorsi and Daraio (2003). In addition, multiple outputs and the lack of corresponding observable, economically meaningful prices (which prevent us from estimating cost functions) limit the usefulness of parametric approaches for our purposes. We avoid these problems by modeling explicitly the production set faced by departments within a given discipline, and employ non-parametric distance-function estimators to examine departments' technical efficiency in producing research output. Our method easily allows consideration of both multiple inputs and multiple outputs, and avoids endogeneity issues arising from simultaneity or reverse-causality by focusing on estimates of distance to boundaries of production sets, as seen below in Section 2. Using recent results obtained by Kneip et al. (2014, 2013), we are able to test whether mean efficiency in a given discipline among departments with local access to HPC is greater than mean efficiency among those without local access.[3]

The remainder of the paper unfolds along the following lines: Section 2 describes our statistical model and the corresponding efficiency estimators while briefly summarizing their relevant properties. Section 3 discusses the various hypotheses we wish to test, and the statistical tests that are employed to do so. Data used for estimation and inference are discussed in Section 4, and our empirical results are presented in Section 5. Conclusions and directions for future work are discussed in Section 6.

2 Statistical Model and Estimators

In order to distill the salient features of universities' activities, and in order to address the questions of interest, let $\boldsymbol{x} \in \mathbb{R}_+^p$ denote a vector of p input quantities, and let $\boldsymbol{y} \in \mathbb{R}_+^q$ denote a vector of q output quantities. The production set

$$\mathcal{P} = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^{p+q} \mid \boldsymbol{x} \text{ can produce } \boldsymbol{y}\} \tag{2.1}$$

[3] As is well known, the nonparametric efficiency estimators described below in Section 2 do not permit measurement error. Then again, the normally-distributed error term typically included on the right-hand side of fully parametric models allows only for measurement error in the left-hand side variable; any measurement error in the right-hand side variables that is correlated with the error terms leads to endogeneity problems. We do not claim that our choice of estimators is perfect; however, our choice addresses two obvious problems. See the recent survey by Simar and Wilson (2013) for a discussion of other estimation methods that might be used and the various tradeoffs that are involved.

is the set of feasible combinations of inputs and outputs. The technology, or efficient frontier of $\mathcal{P}$, is defined by

$$\mathcal{P}^{\partial} = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P} \mid \left(\gamma^{-1}\boldsymbol{x}, \gamma\boldsymbol{y} \right) \notin \mathcal{P} \text{ for all } \gamma > 1 \right\}. \tag{2.2}$$

The Shephard (1970) output-oriented measure of technical efficiency is given by

$$\lambda(\boldsymbol{x}, \boldsymbol{y}) = \inf\{\lambda > 0 \mid (\boldsymbol{x}, \lambda^{-1}\boldsymbol{y}) \in \mathcal{P}\}. \tag{2.3}$$

By construction, $\lambda(\boldsymbol{x}, \boldsymbol{y}) \in (0, 1]$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$. This measure gives the feasible, proportionate increase in output levels, holding input levels constant, for a decision-making unit (DMU) operating at $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$. If $\lambda(\boldsymbol{x}, \boldsymbol{y}) = 1$, the DMU is said to be technically efficient in the output direction, while if $\lambda(\boldsymbol{x}, \boldsymbol{y}) < 1$ the firm is said to be technically *in*efficient. Similar measures have been defined to measure technical efficiency in the input direction, in a hyperbolic direction, or in an arbitrary, linear direction toward the frontier; see Simar and Wilson (2013) for a recent survey and discussion.

The first three assumptions regarding $\mathcal{P}$ are standard in microeconomic theory of the firm; see, for example, Shephard (1970) and Färe (1988).

Assumption 2.1. *$\mathcal{P}$ is closed.*

Assumption 2.2. *All production requires use of some inputs: $(\boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{P}$ if $\boldsymbol{x} = 0$ and $\boldsymbol{y} \geq \boldsymbol{0}$, $\boldsymbol{y} \neq \boldsymbol{0}$.*[4]

Assumption 2.3. *Both inputs and outputs are strongly disposable; i.e., for $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$, $0 \leq \widetilde{\boldsymbol{y}} \leq \boldsymbol{y}$, if $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ then $(\widetilde{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{P}$ and $(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) \in \mathcal{P}$.*

Assumption 2.1 ensures that the frontier $\mathcal{P}^{\partial}$ exists and is contained in $\mathcal{P}$; i.e., that it is feasible for a DMU to produce on the frontier. Assumption 2.2 means that there can be no "free lunches." Assumption 2.3 amounts to an assumption of weak monotonicity for the frontier, and is standard in micro-economic theory of the firm; e.g., see Afriat (1972). This property also admits the feasibility or possibility of wasting resources (i.e., the possibility of producing less with more resources); of course, this is undesirable, but the assumption allows for the possibility, due perhaps to incompetence, bureaucratic waste, or other factors.

[4] Throughout, inequalities involving vectors are assumed to hold element by element; e.g., $\boldsymbol{a} \leq \boldsymbol{b}$ denotes $a_j \leq b_j$ for each $j = 1, \ldots, k$, where k is the length of $\boldsymbol{a}$ and $\boldsymbol{b}$.

The production set $\mathcal{P}$, the frontier $\mathcal{P}^{\partial}$, and the efficiency measure $\lambda(\boldsymbol{x}, \boldsymbol{y})$ are unknown, and hence must be estimated from a random sample $\mathcal{S}_n = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^{n}$ of n input-output pairs. The next assumption describes how such a sample is generated.

Assumption 2.4. *(i) The sample observations $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ in $\mathcal{S}_n$ are realizations of identically, independently distributed (iid) random variables $(\boldsymbol{X}, \boldsymbol{Y})$ with probability density function f, which has support over $\mathcal{D} \subset \mathcal{P} \subset \mathbb{R}_{+}^{p+q}$; and (ii) f is continuously differentiable on $\mathcal{D}$.*

Given a sample $\mathcal{S}_n$ of observations on input-output pairs, several nonparametric estimators of the production set $\mathcal{P}$ are available. Deprins et al. (1984) proposed estimating $\mathcal{P}$ by the free disposal hull (FDH) of the sample observation in $\mathcal{S}_n$, i.e., by

$$\widehat{\mathcal{P}}_{\text{FDH}}(\mathcal{S}_n) = \bigcup_{(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in \mathcal{S}_n} \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_{+}^{p+q} \mid \boldsymbol{x} \geq \boldsymbol{X}_i,\ \boldsymbol{y} \leq \boldsymbol{Y}_i\}. \tag{2.4}$$

Alternatively, the production set can be estimated by the convex hull of $\widehat{\mathcal{P}}_{\text{FDH},n}$, i.e., by

$$\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p+q} \mid \boldsymbol{y} \leq \underline{\boldsymbol{Y}}\boldsymbol{q},\ \boldsymbol{x} \geq \underline{\boldsymbol{X}}\boldsymbol{q},\ \boldsymbol{i}_n'\boldsymbol{q} = 1,\ \boldsymbol{q} \in \mathbb{R}_{+}^{n}\} \tag{2.5}$$

where $\underline{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 & \ldots, \boldsymbol{X}_n \end{bmatrix}$ and $\underline{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Y}_1 & \ldots, \boldsymbol{Y}_n \end{bmatrix}$ are $(p \times n)$ and $(q \times n)$ matrices (respectively) whose columns are the input-output combinations in $\mathcal{S}_n$, $\boldsymbol{q}$ is an $(n \times 1)$ vector of weights, and $\boldsymbol{i}_n$ is an $(n \times 1)$ vector of ones. This estimator was proposed by Banker et al. (1984). Yet another estimator of $\mathcal{P}$ is the conical hull $\widehat{\mathcal{P}}_{\text{CRS}}(\mathcal{S}_n)$ of $\widehat{\mathcal{P}}_{\text{FDH}}(\mathcal{S}_n)$ obtained by dropping the constraint $\boldsymbol{i}_n'\boldsymbol{q} = 1$ in (2.5); this estimator was used by Farrell (1957) and Charnes et al. (1978).

Nonparametric estimators of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ are obtained by replacing $\mathcal{P}$ in (2.3) with either $\widehat{\mathcal{P}}_{\text{FDH}}(\mathcal{S}_n)$, $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n)$, or $\widehat{\mathcal{P}}_{\text{CRS}}(\mathcal{S}_n)$ to obtain estimators $\widehat{\lambda}_{\text{FDH}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$, $\widehat{\lambda}_{\text{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$, $\widehat{\lambda}_{\text{CRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ (respectively) of $\lambda(\boldsymbol{x}, \boldsymbol{y})$. The estimators $\widehat{\lambda}_{\text{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ and $\widehat{\lambda}_{\text{CRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ are known in the literature as data envelopment analysis (DEA) estimators, and can be computed using familiar programming methods. Simple numerical methods can be used to compute $\widehat{\lambda}_{\text{FDH}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$. See Simar and Wilson (2013) for examples and discussion.

Some additional technical assumptions are needed to establish statistical consistency and limiting distributions of the nonparametric efficiency estimators introduced above. Together with Assumptions 2.1–2.4, the next assumptions define a statistical model in the context of which technical efficiency can be estimated consistently from data.

Assumption 2.5. *(i)* $\mathcal{D}^* := \{(\boldsymbol{x}, \lambda(\boldsymbol{x},\boldsymbol{y})^{-1}\boldsymbol{y}) \mid (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}\} \subset \mathcal{D}$*; (ii)* $\mathcal{D}^*$ *is compact; and (iii)* $f(\boldsymbol{x},\boldsymbol{y}) > 0$ *for all* $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}^*$.

Assumption 2.6. $\lambda(\boldsymbol{x},\boldsymbol{y})$ *is three times continuously differentiable on* $\mathcal{D}$.

Assumption 2.7. $\mathcal{D}$ *is* almost strictly convex*; i.e., for any* $(\boldsymbol{x},\boldsymbol{y})$, $(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) \in \mathcal{D}$ *with* $\left(\boldsymbol{x}, \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\right) \neq \left(\boldsymbol{x}, \frac{\widetilde{\boldsymbol{y}}}{\|\widetilde{\boldsymbol{y}}\|}\right)$*, the set* $\{(\boldsymbol{x}^*,\boldsymbol{y}^*) \mid (\boldsymbol{x}^*,\boldsymbol{y}^*) = (\boldsymbol{x},\boldsymbol{y}) + \alpha((\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}}) - (\boldsymbol{x},\boldsymbol{y}))$ *for some* $0 < \alpha < 1\}$ *is a subset of the* interior *of* $\mathcal{D}$.

Assumptions 2.4–2.6 are similar to assumptions needed by Kneip et al. (2008) to establish the limiting distribution of the VRS efficiency estimator in the input direction, except that there, the efficiency measure was only required to be twice continuously differentiable. The compact set $\mathcal{D}$ is introduced in Assumption 2.4 is needed in Kneip et al. (1998) to establish statistical consistency of DEA estimators; the assumption is mild and rules out use of infinite quantities of one or more inputs. Assumption 2.5 is a regularity condition sufficient for proving consistency of the VRS estimator as in Kneip et al. (1998); it says among other things that the probability of observing DMUs in any open neighborhood of the frontier is strictly positive—quite a reasonable property since microeconomic theory indicates that with competitive input and output markets, DMUs which are inefficient will, in the long run, be driven from the market. The addition of Assumption 2.7 and the additional smoothness of $\lambda(x,y)$ in Assumption 2.6 are analogous to assumptions needed by Kneip et al. (2014) to establish results on the bias, variance, and covariances of VRS estimators and results on the limiting distribution of sample means of nonparametric efficiency estimators.[5]

Under Assumptions 2.1–2.7, $\widehat{\lambda}_{\mathrm{FDH}}(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{S}_n)$ is a consistent estimator of $\lambda(\boldsymbol{x},\boldsymbol{y})$ with

$$\widehat{\lambda}_{\mathrm{FDH}}(\boldsymbol{x},\boldsymbol{y} \mid \mathcal{S}_n) = \lambda(\boldsymbol{x},\boldsymbol{y}) + O_p\left(n^{-1/(p+q)}\right) \tag{2.6}$$

with a limiting Weibull distribution as proved by Park et al. (2000).[6] As proved by Kneip

[5] Kneip et al. (1998) and Kneip et al. (2008, 2014) work in the input direction, but the results of each extend trivially to the output direction after appropriate changes in notation. The characterization of smoothness in Assumption 2.6 is stronger than required for the consistency of the nonparametric estimators. Kneip et al. (1998) require only Lipschitz continuity of the efficiency scores, which is implied by the simpler, but stronger requirement presented here. However, derivation of limiting distributions of the nonparametric estimators has been obtained by Kneip et al. (2008) only with the stronger assumption made here.

[6] The proof given by Park et al. (2000) is for the input-oriented estimator, but the result extends almost trivially to the output orientation. See Daouia et al. (2013) for a much simpler proof relying on results from extreme value theory.

et al. (1998), under the additional assumption of convexity of $\mathcal{P}$,

$$\widehat{\lambda}_{\text{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n) = \lambda(\boldsymbol{x}, \boldsymbol{y}) + O_p\left(n^{-2/(p+q+1)}\right). \tag{2.7}$$

Under the additional assumptions of convexity of $\mathcal{P}$ and globally constant returns to scale for $\mathcal{P}^{\partial}$,

$$\widehat{\lambda}_{\text{CRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n) = \lambda(\boldsymbol{x}, \boldsymbol{y}) + O_p\left(n^{-2/(p+q)}\right) \tag{2.8}$$

as proved by Park et al. (2010).[7]

It is important to note that the statistical model defined by the assumptions listed above does not include a regression-equation structure. Moreover, consistency of the FDH and DEA estimators does not rely on an assumption of an error term, conditional on some regressors, having zero expectation, nor are any assumptions about causality (beyond the existence of the production set $\mathcal{P}$) required. The assumptions other than Assumption 2.4, which defines the sampling model, impose restrictions on the joint density of inputs and outputs, and the support of this density, but do not impose conditions on causality, confirming the claim in Section 1 that our estimation approach avoids simultaneity issues. On the other hand, the lack of a regression-type structure prevents us from investigating causal relationships, such as whether local HPC infrastructure is endogenous with respect to research performance.

FDH and DEA estimators do not require functional form assumptions, contrary to parametric estimators, and hence avoid possible specification errors. However, this advantage also comes at a cost: the results in (2.6), (2.7), and (2.8) demonstrate that the three estimators of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ suffer from the well-known curse of dimensionality that affects most nonparametric estimators; i.e., the convergence rate of each estimator becomes slower as the number of dimensions $(p + q)$ increases. On the other hand, while parametric estimators typically converge at rate $n^{1/2}$, this typically holds only if the parametric model that is estimated is correctly specified. Robinson (1988, p. 933) refers to parametric estimators in misspecified models as "$n^{1/2}$-inconsistent"; i.e., a root-n consistent estimator in a misspecified model will converge quickly to something that is perhaps meaningless. There are always tradeoffs.

The limiting distributions of $\widehat{\lambda}_{\text{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ and $\widehat{\lambda}_{\text{CRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ are given by Kneip et al. (2008) and Jeong et al. (2010). Unfortunately, however, the limiting distributions of $\widehat{\lambda}_{\text{VRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$, $\widehat{\lambda}_{\text{CRS}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$, as well as that of $\widehat{\lambda}_{\text{FDH}}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$, involve unknown

[7] The frontier $\mathcal{P}^{\partial}$ exhibits globally CRS if and only if for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$, $(t\boldsymbol{x}, t\boldsymbol{y}) \in \mathcal{P}\ \forall\ t \in [0, \infty)$.

constants that are difficult to estimate. The only practical way to make inference about $\lambda(\boldsymbol{x}, \boldsymbol{y})$ is to use bootstrap methods. However, the usual, naive bootstrap based on resampling from the empirical distribution of the observations in $\mathcal{S}_n$ does not provide consistent inference as discussed by Simar and Wilson (1998, 1999a, 1999b, 2000, and 2008). Kneip et al. (2008) prove consistency of two bootstrap methods for making inference about $\lambda(\boldsymbol{x}, \boldsymbol{y})$ using the VRS estimator. One method involves smoothing both the empirical distribution of the data as well as the initial frontier estimate, but this presents formidable computational difficulties. Kneip et al. (2011) developed a computationally efficient method that avoids the need for smoothing the empirical distribution by resampling from a uniform distribution in a neighborhood of a smoothed version of the initial frontier estimate, and drawing from the empirical distribution of the data outside this neighborhood. Kneip et al. (2008) also proved consistency of a subsampling bootstrap for making inference about $\lambda(\boldsymbol{x}, \boldsymbol{y})$, but did not offer a method for choosing the size of the subsamples. Monte Carlo experiments reported by Kneip et al. (2008) indicate that the results are very sensitive to the choice of subsample size. This problem is overcome by methods proposed by Simar and Wilson (2011a); the existence of limiting distributions for each of the three estimators permits the subsampling method to be used in each case.

3 Testing Hypotheses

As discussed previously in Section 1, one of the primary goals of this paper is to examine whether locally-available HPC infrastructure enhances academic departments' production of research output. The discussion that follows is based on Kneip et al. (2013), which can be consulted for additional details.

We face groups labeled G_1 and G_2 of departments in a given discipline; departments in G_1 have local access to HPC ("haves"), while departments in G_2 do not ("have-nots"). Let $\mu_\lambda^j = E(\lambda(\boldsymbol{X}, \boldsymbol{Y}) \mid (\boldsymbol{X}, \boldsymbol{Y}) \in G_j)$ for $j \in \{1,\ 2\}$; then we wish to test the null hypothesis $\mu_\lambda^1 = \mu_\lambda^2$ versus the alternative hypothesis $\mu_\lambda^1 > \mu_\lambda^2$.

Of course, implementation of the test requires consistent estimation of the efficiencies $\lambda(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ for departments in both groups 1 and 2. Three different efficiency estimators are introduced above in Section 2, requiring varying levels of restrictions depending on which estimator is used. The FDH estimator is the least restrictive estimator, since it can accommodate situations regardless of whether the production set $\mathcal{P}$ is convex, but among the three estimators, the FDH estimator has the slowest convergence rate. If $\mathcal{P}$ is strictly convex, then the VRS estimator can be used with its faster convergence rate. This estimator remains consistent under CRS, but the CRS estimator is also valid in this case, and may have smaller variance and less bias than the VRS estimator. In order to choose an appropriate estimator, one might first test whether $\mathcal{P}$ is convex; if convexity is rejected, the FDH estimator should be used. If convexity is not rejected, one might subsequently test the null hypothesis of CRS against the alternative of non-CRS; if CRS is rejected, the VRS estimator should be used, but if CRS is not rejected, one might use the CRS estimator.[8]

Figure 1 shows the various tests to be implemented and the various decisions to be made after each test. The test of equivalent means can only be implemented after choosing the appropriate estimator, which requires first testing convexity of $\mathcal{P}$, and then testing CRS versus VRS if convexity is not rejected. It is well-known that sequential testing complicates inference. Here, the test of CRS versus VRS is performed only if the first test fails to reject convexity, so the test is conditional on the outcome of the first test. The effect of this conditioning distorts the size of the test of returns to scale, causing the true size to differ from the a-priori chosen nominal size of the test. If the tests were independent, it would be possible to work out the true size of the returns-to-scale test subsequent to the convexity test, and the true size would be smaller than the nominal size. However, the tests are not independent since they are conducted using the same data. Unfortunately, this is a common problem in statistical testing, and there seems to be no useful solution other than to caution the reader and to make appropriate caveats when discussing results, which we do when discussing our empirical results later. As will be seen below in Section 5, in five of the six cases where we test CRS versus VRS, the resulting p-values are quite large and we do not

[8] Alternatively, one might want to test CRS versus non-increasing returns to scale (NIRS). The DEA, NIRS estimator $\widehat{\mathcal{P}}_{\text{NIRS}}(\mathcal{S}_n)$ of $\mathcal{P}$ is obtained by changing the constraint $\boldsymbol{i}_n'\boldsymbol{q} = 1$ in (2.5) to $\boldsymbol{i}_n'\boldsymbol{q} \leq 1$. By construction, $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n) \subseteq \widehat{\mathcal{P}}_{\text{NIRS}}(\mathcal{S}_n) \subseteq \widehat{\mathcal{P}}_{\text{CRS}}(\mathcal{S}_n)$. Kneip et al. (2013) establish that the VRS estimator of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ remains consistent under CRS (which is no surprise), but with the faster CRS convergence rate $n^{2/(p+q)}$ (which is surprising). Properties of the NIRS estimator of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ remain unknown, and are likely complicated in view of the result of Kneip et al. (2013) for the VRS estimator under CRS. Since $\widehat{\mathcal{P}}_{\text{VRS}}(\mathcal{S}_n) = \widehat{\mathcal{P}}_{\text{NIRS}}(\mathcal{S}_n)$ in the limit if $\mathcal{P}^{\partial}$ is NIRS, it is easy to imagine that the VRS estimator of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ remains consistent if $\mathcal{P}^{\partial}$ is NIRS, though its convergence rate might depend on where the point $(\boldsymbol{x}, \boldsymbol{y})$ lies in $\mathcal{P}$. Since the properties of the NIRS estimator have not been worked out, we test CRS versus VRS instead of testing CRS versus NIRS.

come close to rejecting CRS in favor of VRS. In the one case where this does not happen, we take a conservative approach and use the VRS estimator to test for differences in mean efficiency between the haves and the have-nots.

3.1 Testing convexity

We first describe a test of $H_0\colon \mathcal{P}$ is convex, versus $H_1\colon \mathcal{P}$ is not convex. We assume that departments in both groups 1 and 2 face the same frontier $\mathcal{P}^{\partial}$, and that to the extent that locally-available HPC might affect production, it does so only through the distribution of efficiency. This is the "separability" assumption discussed by Simar and Wilson (2007, 2011b), who noted that it constitutes a restriction that should be tested with data. To date, however, no such test exists, although work on this difficult problem is progressing.

Recall from the discussion in Section 1 that researchers at institutions without on-campus HPC infrastructure can access HPC environments through the U.S. Department of Energy INCITE program or the NSF XSEDE program, by purchasing time on HPC systems from commercial providers, or perhaps other avenues. In principle, any U.S. academic researcher who wants to can work on HPC systems, but more effort may be required if the HPC systems are not local (this has been the authors' experiences). The increased effort required when working on a remote system may result from lack of personal relationships with support staff, more complicated procedures for obtaining allotments of processing time, greater difficulty in modifying parameters for job queues, job scheduling, etc., as well as other factors. We view all of this as adversely affecting the ability of researchers at have-not institutions to work near a common frontier $\mathcal{P}^{\partial}$, i.e., as affecting technical efficiency. Alternatively, if one believes there are two different frontiers for haves and have-nots, our assumption of a common frontier might be viewed as a meta-frontier along the lines described by O'Donnell et al. (2008). In this case, it is easy to imagine that one would be likely to reject convexity of the meta-frontier.

Let $\mathcal{S}_n$ be a sample of input-output pairs of n departments (both groups G_1 and G_2) in a given discipline. In addition, let $\widehat{\lambda}_{\bullet}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n)$ represent either the FDH, VRS, or CRS estimator of $\lambda(\boldsymbol{x}, \boldsymbol{y})$ using the sample $\mathcal{S}_n$, and let $\kappa_{\bullet} = 1/(p+q)$, $2/(p+q+1)$, or $2/(p+q)$, respectively.[9] Then the results in (2.6)–(2.8) can be summarized by writing

[9] Throughout, the subscript "$\bullet$" will be used when either "FDH," "VRS," or "CRS" is applicable.

$\widehat{\lambda}_{\bullet}(\boldsymbol{x}, \boldsymbol{y} \mid \mathcal{S}_n) = \lambda(\boldsymbol{x}, \boldsymbol{y}) + O_p\left(n^{-\kappa_{\bullet}}\right)$ for the FDH, VRS, or CRS cases, respectively.

Now consider the sample mean

$$\widehat{\mu}^{\text{full}}_{\bullet,n} = n^{-1} \sum_{i=1}^{n} \widehat{\lambda}_{\bullet}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_n). \tag{3.9}$$

In order to test convexity of $\mathcal{P}$, it would be tempting to use the statistic $\left(\widehat{\mu}^{\text{full}}_{\text{FDH},n} - \widehat{\mu}^{\text{full}}_{\text{VRS},n}\right)$, and to reject H_0 if this is suitably large. However, Kneip et al. (2014) show that while $\widehat{\mu}^{\text{full}}_{FDH,n}$ $\widehat{\mu}^{\text{full}}_{VRS,n}$ are consistent estimators of $\mu_{\lambda} = E\left(\lambda(\boldsymbol{X}, \boldsymbol{Y})\right)$ under the null hypothesis of convexity, conventional central limit theorem results cannot be used to make inference about μ_{λ} unless $\kappa_{\bullet} > 1/2$; moreover, for the application in this study, $p = q = 2$, and hence $\kappa_{\bullet} = 1/4$ and $2/5$ for the FDH and VRS cases, respectively.

As seen from Kneip et al. (2014, Theorem 4.1), the problem arises from the fact that the nonparametric estimators of $\lambda(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ are biased; in particular, Theorem 4.1 of Kneip et al. (2014) establishes that

$$\sqrt{n}\left(\widehat{\mu}^{\text{full}}_{\bullet,n} - \mu_{\lambda} - C_{\bullet} n^{-\kappa_{\bullet}} - o(n^{-\kappa_{\bullet}})\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{\lambda}) \tag{3.10}$$

where $\sigma^2_{\lambda} = \text{VAR}(\lambda(\boldsymbol{X}, \boldsymbol{Y}))$. If $\kappa_{\bullet} > 1/2$, the bias term $C_{\bullet}$ vanishes asymptotically, but if $\kappa_{\bullet} = 1/2$, the bias remains constant, and if $\kappa_{\bullet} < 1/2$, the bias explodes as $n \to \infty$. Kneip et al. (2014) note that replacing the scale factor $\sqrt{n}$ in (3.10) with another power of n, e.g., n^{γ} with $\gamma \in (0, \kappa_{\bullet})$, would cause the bias to vanish as $n \to \infty$, but this would also cause the variance to converge to zero whenever $\kappa_{\bullet} \leq 1/2$, making inference impossible.

To overcome this problem, Kneip et al. (2014) provide new central limit theorem results for means of the nonparametric efficiency estimators introduced above, and Kneip et al. (2013) use these results to develop tests of convexity versus non-convexity, CRS versus VRS, and differences in means. To implement the test of convexity, the difference in convergence rates of the FDH and VRS estimators can be exploited by setting $n_1^{2/(p+q+1)} = n_2^{1/(p+q)}$ and $n_1 + n_2 = n$ for a given sample size n, and then solving for n_1 and n_2 by writing $n - n_1 - n_1^{2(p+q)/(p+q+1)} = 0$ and finding the root n_1 using the bisection method. Let n_1 be the integer part of this solution, and set $n_2 = n - n_1$ so that $n_1 < n_2$.

Next, let the sample observations $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ in $\mathcal{S}_n$ be randomly ordered; for purposes of testing convexity of $\mathcal{P}$, we do not consider the presence or absence of local HPC infrastructure. Split the sample into two independent subsamples $\mathcal{S}_{1,n_1}$, $\mathcal{S}_{2,n_2}$ containing the first n_1

and the last n_2 observations so that $\mathcal{S}_{1,n_1} \cup \mathcal{S}_{2,n_2} = \mathcal{S}_n$ and $\mathcal{S}_{1,n_1} \cap \mathcal{S}_{2,n_2} = \emptyset$. Using the first subsample, compute

$$\widehat{\mu}_{\text{VRS},1,n_1} = n_1^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{1,n_1}} \widehat{\lambda}_{\text{VRS}}\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{1,n_1}\right) \tag{3.11}$$

and

$$\widehat{\sigma}^2_{\text{VRS},1,n_1} = n_1^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{1,n_1}} \left[\widehat{\lambda}_{\text{VRS}}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{1,n_1}) - \widehat{\mu}_{\text{VRS},1,n_1}\right]^2, \tag{3.12}$$

as well as

$$\widehat{\mu}_{\text{FDH},2,n_2} = n_2^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{2,n_2}} \widehat{\lambda}_{\text{FDH}}\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}\right). \tag{3.13}$$

In addition, let

$$\widehat{\sigma}^2_{\text{FDH},2,n_2} = n_2^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{2,n_2}} \left[\widehat{\lambda}_{\text{FDH}}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}) - \widehat{\mu}_{\text{FDH},2,n_2}\right]^2. \tag{3.14}$$

By Kneip et al. (2014, Theorem 4.1), both $\widehat{\mu}_{\text{VRS},1,n_1}$ and $\widehat{\mu}_{\text{FDH},2,n_2}$ are consistent estimators of μ_λ under the null hypothesis of convexity, and both $\widehat{\sigma}^2_{\text{VRS},1,n_1}$ and $\widehat{\sigma}^2_{\text{FDH},2,n_2}$ consistently estimate the variances of the VRS and FDH efficiency estimators.

In order to estimate the biases of $\widehat{\mu}_{\text{VRS},1,n_1}$ and $\widehat{\mu}_{\text{FDH},2,n_2}$, let $\kappa_1 = 2/(p+q+1)$ and $\kappa_2 = 1/(p+q)$. Then, following Kneip et al. (2014), estimate the biases by splitting each of the two subsamples $\mathcal{S}_{\ell,n_\ell}$, $\ell \in \{1,\ 2\}$, into sub-subsamples $\mathcal{S}^{(1)}_{\ell,m_{\ell,1}}$ and $\mathcal{S}^{(2)}_{\ell,m_{\ell,2}}$ of sizes $m_{\ell,1} = [n_\ell/2]$ and $m_{\ell,2} = n_\ell - m_{\ell,1}$, where $[a]$ denotes the integer part of a, so that $\mathcal{S}^{(1)}_{\ell,m_{\ell,1}} \cup \mathcal{S}^{(2)}_{\ell,m_{\ell,2}} = \mathcal{S}_{\ell,n_\ell}$ and $\mathcal{S}^{(1)}_{\ell,m_{\ell,1}} \cap \mathcal{S}^{(2)}_{\ell,m_{\ell,2}} = \emptyset$.

For sub-subsample $j \in \{1,\ 2\}$ of subsample 1, let

$$\widehat{\mu}^{(j)}_{\text{VRS},1,m_{1,j}} = (m_{1,j})^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}^{(j)}_{1,m_{1,j}}} \widehat{\lambda}_{\text{VRS}}\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}^{(j)}_{1,m_{1,j}}\right). \tag{3.15}$$

For sub-subsample $j \in \{1,\ 2\}$ of subsample 2, let

$$\widehat{\mu}^{(j)}_{\text{FDH},2,m_{2,j}} = (m_{2,j})^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}^{(j)}_{2,m_{2,j}}} \widehat{\lambda}_{\text{FDH}}\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}^{(j)}_{2,m_{2,j}}\right). \tag{3.16}$$

Note that the efficiency estimates under the summations in (3.15)–(3.16) are computed using only the observations in a given sub-subsample; similarly, the summations are over these same observations. Then the bias correction for the first subsample, where VRS estimators are used, is given by

$$\widetilde{B}_{\text{VRS},1,\kappa_1,n_1} = (2^{\kappa_1} - 1)^{-1} \left[0.5 \left(\widehat{\mu}^{(1)}_{\text{VRS},1,m_{1,1}} + \widehat{\mu}^{(2)}_{\text{VRS},1,m_{1,2}} \right) - \widehat{\mu}_{\text{VRS},1,n_1} \right]. \tag{3.17}$$

Similarly, the bias correction for the second subsample, where FDH estimators are used, is

$$\widetilde{B}_{\text{FDH},2,\kappa_2,n_2} = (2^{\kappa_2} - 1)^{-1} \left[0.5 \left(\widehat{\mu}^{(1)}_{\text{FDH},2,m_{2,1}} + \widehat{\mu}^{(2)}_{\text{FDH},2,m_{2,2}} \right) - \widehat{\mu}_{\text{FDH},2,n_2} \right]. \tag{3.18}$$

Continuing to follow Kneip et al. (2014), note that each of the two subsamples $\mathcal{S}_{\ell,n_\ell}$, $\ell \in \{1,\ 2\}$, can be split a large number of ways; we split each randomly 100 times and then average the two sets of 100 bias estimates obtained from (3.17) and (3.18) to obtain bias corrections $\widehat{B}_{\text{VRS},1,\kappa_1,n_1}$ and $\widehat{B}_{\text{FDH},2,\kappa_2,n_2}$.

Finally, let $\kappa = \kappa_2 = 1/(p+q)$. Let $n_{\ell,\kappa} = n_\ell^{2\kappa} = n_\ell^{1/2}$ for $(p+q) = 4$. For $\ell \in \{1,2\}$, let $\mathcal{S}^*_{\ell,n_{\ell,\kappa}}$ be a random subset of size $n_{\ell,\kappa}$ from $\mathcal{S}_{\ell,n_\ell}$. Using subsample 1, let

$$\widehat{\mu}_{\text{VRS},1,n_{1,\kappa}} = n_{1,\kappa}^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}^*_{1,n_{1,\kappa}}} \widehat{\lambda}_{\text{VRS}}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{1,n_1}). \tag{3.19}$$

Using subsample 2, let

$$\widehat{\mu}_{\text{FDH},2,n_{2,\kappa}} = n_{2,\kappa}^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}^*_{2,n_{2,\kappa}}} \widehat{\lambda}_{\text{FDH}}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}). \tag{3.20}$$

Then for $(p+q) \geq 3$, under the null hypothesis of convexity of $\mathcal{P}$,

$$\widehat{\tau}^*_n = \frac{\left(\widehat{\mu}_{\text{FDH},2,n_{2,\kappa}} - \widehat{\mu}_{\text{VRS},1,n_{1,\kappa}}\right) - \left(\widehat{B}_{\text{FDH},2,\kappa_2,n_2} - \widehat{B}_{\text{VRS},1,\kappa_1,n_1}\right)}{\sqrt{\frac{\widehat{\sigma}^2_{\text{FDH},2,n_2}}{n_{2,\kappa}} + \frac{\widehat{\sigma}^2_{\text{VRS},1,n_1}}{n_{1,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.21}$$

by Kneip et al. (2014, Theorem 4.4). The null hypothesis of convexity of $\mathcal{P}$ is rejected if the p-value $\widehat{p} = 1 - \Phi(\widehat{\tau}^*_n)$, where $\Phi()$ denotes the standard normal distribution function, is less than an appropriately small value, e.g., .1, .05, or .01.

3.2 Testing constant versus variable returns to scale

If the null hypothesis of convexity of $\mathcal{P}$ is not rejected, it is natural to test $H_0\colon \mathcal{P}^\partial$ is CRS versus $H_1\colon \mathcal{P}^\partial$ is not CRS while maintaining the assumption that $\mathcal{P}$ is convex under the alternative. To implement a test, one might consider comparing a sample mean of CRS efficiency estimates based on (2.8) with a sample mean of VRS efficiency estimates based on (2.7). However, the problem is similar to that of testing convexity of $\mathcal{P}$; as shown by Kneip et al. (2014), standard central limit theorems cannot be used due to the bias of the efficiency estimators and their corresponding sample means. On the other hand, as shown by Kneip et al. (2013), the problem of testing CRS versus VRS is simpler than that of testing convexity of $\mathcal{P}$ due to (i) the fact that the convergence rates of the CRS and VRS estimators are faster than the convergence rate of the FDH estimator, and (ii) the fact that the VRS efficiency estimator in (2.7) achieves the same convergence rate as the CRS estimator in (2.8) *under the null hypothesis of CRS,* as proved by Kneip et al. (2013, Theorem 3.1).

Here again, we consider a sample $\mathcal{S}_n$ of n randomly-ordered observed input-output pairs, ignoring whether a particular department is a "have" or "have-not." Since the VRS and CRS estimators converge at the same rate under the null, the initial split of $\mathcal{S}_n$ into two parts $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ should be done while setting $n_1 = [n/2]$ and $n_2 = n - n_1$. To implement the test of returns to scale described by Kneip et al. (2013), define

$$\widehat{\mu}_{\text{CRS},2,n_2} = n_2^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{2,n_2}} \widehat{\lambda}_{\text{CRS}} \left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}\right) \tag{3.22}$$

and

$$\widehat{\sigma}^2_{\text{CRS},2,n_2} = n_2^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{2,n_2}} \left[\widehat{\lambda}_{\text{CRS}}(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}) - \widehat{\mu}_{\text{CRS},n_2}\right]^2, \tag{3.23}$$

analogous to (3.11)–(3.12). Next, repeatedly (100 times) split each subsample into sub-subsamples as described above in Section 3.1 and compute the bias correction $\widehat{B}_{\text{VRS},\kappa,n_1}$ as before, but while setting $\kappa = \kappa_1 = 2/(p+q)$. For each of the 100 splits, also compute

$$\widetilde{B}_{\text{CRS},2,\kappa_2,n_2} = (2^{\kappa_2} - 1)^{-1} \left[0.5 \left(\widehat{\mu}^{(1)}_{\text{CRS},2,m_{2,1}} + \widehat{\mu}^{(2)}_{\text{CRS},2,m_{2,2}}\right) - \widehat{\mu}_{\text{CRS},2,n_2}\right] \tag{3.24}$$

where

$$\widehat{\mu}^{(j)}_{\text{CRS},2,m_{2,j}} = (m_{2,j})^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}^{(j)}_{2,m_{2,j}}} \widehat{\lambda}_{\text{CRS}} \left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}^{(j)}_{2,m_{2,j}}\right), \tag{3.25}$$

analogous to (3.15) and (3.16). Compute $\widehat{\mu}_{\text{VRS},1,n_1,\kappa}$ using subsample 1 and (3.11) with $\kappa = 2/(p+q)$, and compute

$$\widehat{\mu}_{\text{CRS},2,n_2} = n_2^{-1} \sum_{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\in\mathcal{S}_{2,n_2}} \widehat{\lambda}_{\text{CRS}} \left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid \mathcal{S}_{2,n_2}\right). \tag{3.26}$$

using subsample 2, again with $\kappa = 2/(p+q)$. Then average the values $\widetilde{B}_{\mathrm{CRS},2,\kappa_2,n_2}$ to obtain the CRS bias correction $\widehat{B}_{\mathrm{CRS},2,\kappa_2,n_2}$.

Under the null hypothesis of CRS, for $(p+q) \leq 5$,

$$\widehat{\tau}_n^{**} = \frac{(\widehat{\mu}_{\mathrm{CRS},2,n_2} - \widehat{\mu}_{\mathrm{VRS},1,n_1}) - \left(\widehat{B}_{\mathrm{CRS},2,\kappa,n_2} - \widehat{B}_{\mathrm{VRS},1,\kappa,n_1}\right)}{\sqrt{\frac{\widehat{\sigma}^2_{\mathrm{CRS},2,n_2}}{n_2} + \frac{\widehat{\sigma}^2_{\mathrm{VRS},1,n_1}}{n_1}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.27}$$

as demonstrated by Kneip et al. (2013) using Theorem 4.2 of Kneip et al. (2014). [10] The null is rejected in favor of VRS if the p-value $\widehat{p} = 1 - \Phi(\widehat{\tau}_n^{**})$ is less than, e.g., .1, .05, or .01.

3.3 Testing for differences in means

As noted above and as illustrated in Figure 1, the outcome of the convexity test, and subsequently the outcome of the returns to scale test if convexity is not rejected, determine which among the FDH, VRS, and CRS efficiency estimators should be used to test $H_0\colon \mu_\lambda^1 = \mu_\lambda^2$ versus $H_1\colon \mu_\lambda^1 > \mu_\lambda^2$. In any of these cases, consider iid samples $\mathcal{S}_{1,n_1}$ and $\mathcal{S}_{2,n_2}$ of n_1 and n_2 input-output pairs from G_1 (the "haves") and G_2 (the "have-nots"), respectively. In the language of Sections 3.1 and 3.2, one might think of these as subsamples from a combined sample $\mathcal{S}_n$ of size $n = n_1 + n_2$.

Let $\widehat{B}_{\bullet,j,\kappa,n_j}$ for group G_j, $j \in \{1,\ 2\}$, denote the bias expressions given previously, with $\bullet$ denoting either VRS, CRS, or FDH with corresponding values of $\kappa = 2/(p+q+1)$, $2/(p+q)$, or $1/(p+q)$, respectively. Using the result in Kneip et al. (2014, Theorem 4.3), Kneip et al. (2013) show that for cases where either VRS or CRS estimators are used, with $(p+1) \leq 4$ or 5 (respectively),

$$\widehat{\delta}^*_{\bullet,n_1,n_2} = \frac{(\widehat{\mu}_{\bullet,1,n_1} - \widehat{\mu}_{\bullet,2,n_2}) - \left(\widehat{B}_{\bullet,1,\kappa,n_1} - \widehat{B}_{\bullet,2,\kappa,n_2}\right) - (\mu_\lambda^1 - \mu_\lambda^2)}{\sqrt{\frac{\widehat{\sigma}_{\bullet,1,n_1}}{n_1} + \frac{\widehat{\sigma}_{\bullet,2,n_2}}{n_2}}} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.28}$$

with $\kappa = 2/(p+q+1)$ or $2/(p+q)$ as $\bullet$ signifies either VRS or CRS (respectively), provided $n_1/n_2 \to c > 0$ as $n_1,\ n_2 \to \infty$ for some $c \in \mathbb{R}^1_+$. Then the null hypothesis of equal means is rejected whenever $\widehat{\delta}^*_{n_1,n_2}$ exceeds an appropriate critical value from the standard normal distribution.

[10] Note that the result in (3.27) is valid only for $(p+q) \leq 5$. See Kneip et al. (2013) for an alternative test statistic for cases where $(p+q) > 5$.

In cases where convexity of $\mathcal{P}$ is rejected by the initial test, the FDH estimator of efficiency must be used, as both the VRS and CRS estimator are inconsistent when $\mathcal{P}$ is non-convex. In the FDH case, Theorem 4.3 of Kneip et al. (2014) holds only for $(p+q) \leq 2$. However, Kneip et al. (2013) use results from Kneip et al. (2014, Theorem 4.4) to establish that

$$\widehat{\delta}^{**}_{\bullet,n_{1,\kappa},n_{2,\kappa}} = \frac{\left(\widehat{\mu}_{\bullet,1,n_{1,\kappa}} - \widehat{\mu}_{\bullet,2,n_{2,\kappa}}\right) - \left(\widehat{B}_{\bullet,1,\kappa,n_1} - \widehat{B}_{\bullet,2,\kappa,n_2}\right) - \left(\mu^1_\lambda - \mu^2_\lambda\right)}{\sqrt{\frac{\widehat{\sigma}_{\bullet,1,n_{1,\kappa}}}{n_{1,\kappa}} + \frac{\widehat{\sigma}_{\bullet,2,n_{2,\kappa}}}{n_{2,\kappa}}}} \xrightarrow{\mathcal{L}} N(0,1) \qquad (3.29)$$

for all $(p+q)$ and $\kappa = 1/(p+q)$, again provided $n_1/n_2 \to c > 0$ as $n_1,\ n_2 \to \infty$ for some $c \in \mathbb{R}^1_+$. The null hypothesis of equal means is rejected whenever $\widehat{\delta}^{**}_{n_1,n_2}$ exceeds an appropriate critical value from the standard normal distribution.

4 Data Description

A primary goal of this paper is to examine whether locally-available HPC infrastructure enhances academic departments' production of research output. Our empirical results are obtained by applying the theoretical results discussed in the previous section using carefully validated data. Numerous sources of data are available for higher education institutional research, including the Integrated Postsecondary Education Data System from the National Center for Educational Statistics, UCLA's Higher Education Research Institute, Integrated Science and Engineering Resources Data System from the U.S. National Science Foundation (NSF), the Institute for Scientific Information's Web of Science, US News and World Report, and the National Research Council (NRC). Each of these sources of data has strengths and weaknesses. Given that we wish to examine the effect of HPC on research productivity at the level of individual departments within various disciplines, across a range of U.S. academic institutions, we are led to use survey data from the 2005-2006 academic year collected by the NRC; among the various sources that are available, only the NRC data contain the level of granularity needed for our study. Wherever possible, we compared the NRC data with data from other sources in order to verify its consistency and veracity.

The NRC survey covers 212 institutions, including 177 universities with "high" or "very-high" research levels, in addition to 12 medical schools and a small number of other degree-granting institutions. (see Ostriker et al., 2011 for additional details on the survey and sampling frame). The data contain detailed information on U.S. university departments and cover six broad areas: agricultural sciences, biological and health sciences, engineering, physical and mathematical sciences, social and behavioral sciences, and humanities. These broad areas are broken into 62 specific fields. All together, the data contain information on more than 5,000 academic programs at 212 higher education institutions in the U.S.

Of the 62 academic disciplines represented in the data, we examine the eight disciplines that provide the largest number of observations (our unit of observation is a university-department). Other disciplines provide too few observations to likely make analysis meaningful from a statistical viewpoint, given the slow convergence rates of our estimators as discussed above in Section 2. As noted in Section 1, the eight disciplines we examine include Chemistry, Civil and Environmental Engineering, Computer Science, Ecology and Evolutionary Biology, English, History, Economics, and Physics.

The data contain a variety of variables describing attributes of university departments. Broadly speaking, universities combine faculty effort, physical plant, and other inputs to produce outputs such as research and education. Of course, universities are complicated organizations; unlike profit-maximizing firms, their objectives may be varied and are perhaps sometimes unclear. At the same time, any useful model must necessarily distill a complicated reality into something tractable. Balancing these considerations against the curse of dimensionality of our nonparametric estimators described earlier in Section 2, we consider $p = 2$ inputs: total number of faculty (measured in full-time equivalents), and the average graduate record examination (GRE) scores of the department's incoming graduate students. Average GRE scores give a measure of the aptitude, and therefore the capability, of incoming graduate students. We expect that GRE scores from one year to the next are highly correlated, and hence this should give a measure of the capability of students in a department's graduate program. We also consider $q = 2$ outputs: total publications for the academic year, and the number of Ph.D. degrees awarded. Of course, a department's output is a dynamic process, and publications and degrees granted in a given year usually reflect work over not just the current year, but also prior years. However, both of these measures are likely correlated through time.[11] Summary statistics for our inputs and outputs are given,

[11] The NRC data count publications by departments according to authors' affiliations, with no adjustments for coauthor relationships. Hence a publication coauthored by two authors in the same department is counted only once, while a publication coauthored by two authors in different departments is counted twice, once for each department. While this is not perfect, it is far from clear what would be an ideal weighting scheme. Deans, department chairs, and authors of articles on department rankings have used a variety of

by discipline, in Table 1.

As noted above, our specification of inputs and outputs attempts to find some balance between comprehensiveness and the curse of dimensionality. Our input-output specification is not dissimilar to specifications others have used. In choosing the variables in our specification, we have opted in favor of those that are more easily quantifiable, and therefore (hopefully) less contaminated by measurement error, as opposed to other factors that are more difficult to measure precisely. An example of the latter, that we have not included, is teaching load. Faculty required to teach introductory courses incur an opportunity cost in terms of lost research time. However, for some, teaching 50–60 students in a class is little different from teaching 300; moreover, not all faculty in a given department teach such courses. In addition, in some departments, introductory courses are often taught by graduate students. Dividing number of undergraduates, or number of undergraduate credit hours, by the number of faculty in a given department would likely result in a very "noisy" measure. Similar concerns surround department budgets, which are often not allocated equally among a department's faculty members.

To examine the effect of HPC on research productivity, we require additional data regarding HPC capability at each institution. To date, only two comprehensive sources of information on U.S. universities' HPC capability exist. The Science and Engineering Research Facilities Survey conducted biannually by the NSF's National Center for Science and Engineering Statistics collects detailed information about the overall technical configurations of cyberinfrastructure at individual institutions in the United States. Data from these surveys are available for 2005, 2007, 2009, and 2011. The 2005 survey, which most closely matches the time frame of our input-output data, contains responses to 15 questions. Of these 15 questions, 10 focus on different aspects of the institution's network bandwidth; only one question relates to the institution's computational capability, and this question asks for the number of HPC systems that were physically located at the institution during 2005. It

schemes, and there seems to be little agreement on what might be the "best" weighting scheme. Given that we examine a set of eight disparate disciplines, we have little choice but to use the data that are available; the alternative would be to search through all publications in our eight disciplines appearing in our time frame, identify the affiliation of each publication's authors, and then apply what would still be an arbitrary weighting scheme. This would present a formidable computational burden, perhaps requiring machine learning and other methods. In addition, one would likely want to use different weighting schemes across different disciplines; e.g., in Economics, it is somewhat uncommon for graduate assistants to be listed as coauthors, whereas this is more common in Computer Science and perhaps other disciplines.

is not clear how one would answer this question, and the survey leaves for the respondent to define what constitutes an HPC system. Examination of these data reveals a number of inconsistencies, and we do not use these data for our study.

The second source of information on institutions' HPC capabilities is the "Top 500" list of supercomputer sites maintained by Top500 (2013). The Top 500 list has ranked the fastest 500 HPC systems in the world twice each year since 1993 in June and November. The list includes HPC systems at both academic and non-academic institutions. The rankings are determined by the time required for systems to successfully complete a computationally intensive benchmarking program (i.e., the LINPACK benchmark, which requires solution of a dense system of linear equations) developed by Dongarra et al. (1976). The benchmark permits code optimization suitable for a given machine, and measures how quickly a machine can compute a particular, given problem as opposed to the machine's theoretical performance, which might not be achievable in practice for a variety of technical reasons. The list is based on voluntary submissions by system owners who choose to run the benchmark program; hence only institutions that choose to participate are ranked.[12] In addition, the benchmarking program does not support systems that utilize ad-hoc or opportunistic work management schemes.

Systems appearing in the Top 500 lists are affiliated with an HPC center, which may be located at an academic institution. For our purposes, we match HPC centers with universities (e.g., the Texas Advanced Computing Center, which serves the University of Texas at Austin).[13] We use data from the Top 500 lists for 2000-2006 to group departments within each discipline into haves (with locally available HPC) and have-nots (without locally available HPC). In particular, a department is included with the haves if the department's university appears on any of the Top 500 lists during 2000–2006; otherwise, it is included with the have-nots. We use this strategy because the Top 500 list evolves over time; what

[12] We suspect that in most cases, universities have substantial incentives to submit benchmark times. Appearing on the Top 500 list is a marketing tool for faculty and graduate student recruitment, and may provide prestige for administrators, while the opportunity cost is a few days of down-time for a system, and perhaps a few disgruntled faculty users. On the other hand, benchmark times for systems owned by secretive government agencies or private companies may be less likely to be submitted; for private companies, the system down-time required to run the benchmark translates into foregone profits. So, while the Top 500 list may be unreliable for gauging HPC capacity of governments and private companies, it likely gives a much better idea of HPC capacity for universities.

[13] These data are available upon request.

would be the 500th fastest system in one year is likely to be the $(500 + j)$-th fastest system the following year, where j is a positive integer, unless the system is upgraded substantially. However, systems that fall off the Top 500 list are typically not shut down, at least not right away.

Using appearances on the Top 500 list as our measure of HPC capability has the advantages of being measurable and has a clear meaning. We recognize that there are an unknown number of HPC systems at universities that would be ranked, for example, 501 to 1000 if there were a "Top 1000" list available. If a university has the 501st HPC system, our strategy causes us to treat its departments as have-nots instead of haves. To the extent that some of our have-not departments have local access to HPC resources not on the Top 500 list, our results will be biased in favor of finding no difference between haves and have-nots, or finding smaller differences than might otherwise be the case. In this sense, our tests for differences in mean efficiencies and for stochastic dominance can be regarded as conservative tests. Table 2 gives the number of observations for each discipline represented in our sample, and breaks these numbers into haves and have-nots for each discipline.

We examine the effect of a researcher having local, campus access to a Top 500 system. We do not use system rankings on the Top 500 list. In many cases, universities that have higher ranked systems share these resources externally with researchers from outside the university. In addition, departments within a university compete for access to HPC resources, and we do not have measures of the relative share of the system that is allocated to a department.[14] Most computational problems do not require an entire system of the size of those on the Top 500 list, although there are some such problems. To the extent that most systems on the Top 500 list during 2000–2006 and beyond are massively parallel systems, faster system may be able to process more jobs that use, say, 2^{10} or 2^{11} threads in a given window of time than a slower system, but there are also likely more users competing for time on the faster system than on the slower system.

[14] Nor do we have information on queuing times for jobs that run on Top 500 systems. In principle, we could look at the number of processors, nodes, or theoretical performance (in terms of floating point operations per second) per faculty member, but this would be misleading for several reasons. For example, on some systems (e.g., the Palmetto cluster at Clemson University) faculty who contribute funds are given ownership rights, and have higher priority than other uses. In addition, some systems contribute resources to the XSEDE program or sell time to commercial users. Many systems exhibit peak-load characteristics; e.g., usage may be higher near the ends of academic semesters and in the weeks before NSF submission deadlines than at other times.

As noted earlier in footnote 3 the efficiency estimators used in this paper do not permit measurement error; hence it is reasonable to consider the quality of the data used in this analysis and the extent to which it might be subject to measurement error. Ostriker et al. (2011, pp. 45–47) discuss at length how the NRC data used to define inputs and outputs were validated and cleaned; after identifying outliers, corresponding institutions were contacted to clarify and cross-check suspicious values. Total number of faculty is a rather objective measure and is a budget item for universities; although there might be some remaining errors of measurement, given the data collection described by Ostriker et al. (2011), it is doubtful that any remaining errors are large. Average GRE scores are equally an objective measure, and moreover, are an *average*; since there is no reason to believe any measurement error in individual scores would be systematic, averaging likely tends to cancel out such errors. Ph.D. degrees awarded and number of publications are equally objective measures. Of the two, publications are more likely to be mis-counted; to the extent that publications are mis-counted, the counts are likely to be too low in view of the discussion by Ostriker et al. (2011, pp. 45–47). This would tend to lead to over-estimates of efficiency, but there is no reason to believe that the mis-counts would not be uniformly distributed between haves and have-nots, and thus should not overly impact our test statistics which involve comparisons of means of efficiency estimates across groups.

5 Empirical results

Figure 1 gives an overview of our testing strategy. In this section, we discuss the test results and their implications.

We first apply the test of convexity versus non-convexity of the production set described in Section 3 to the data for each discipline. As discussed earlier, for this purpose, we ignore the distinction between haves and have-nots. The resulting p-values are given in the second column of Table 3. We clearly fail to reject the null hypothesis of convexity for Chemistry, Civil and Environmental Engineering, Computer Science, Ecology and Evolutionary Biology, and Economics, with p-values greater than 0.72 in each case. For Physics, the p-value is 0.0463, and so convexity is just rejected at the 5-percent level, but not at the 1-percent level. By contrast, for English and History, convexity is soundly and clearly rejected, with p-values well below 0.01.

Our data do not allow us to investigate why we find non-convexity for English and History, but it is easy to speculate why this result might arise. Non-convexity of the production sets for English and History departments might result if, for example, output rises slowly as faculty are added to very small departments, and adding still more faculty to become a medium-size department results in formation of factions that bicker with each other, tending to reduce productivity. As departments move from medium- to large-size, factions may become larger along with department budgets; if this reduces bickering between factions, then one might find non-convexity of the production set. The summary statistics in Table 1 indicate that the ranges of department sizes in our sample, in terms of number of faculty, are similar across the eight disciplines in our sample. However, we doubt anyone who has worked in academia for more than a few years would argue that cultures in English and History departments resemble those in the sciences. Non-convexity in the technology for Physics departments might result, at least in part, from other reasons. For example, if laboratory space and equipment can be shared by several researchers, this might lead to non-convexities. Faculty could be added until facilities become fully used, then more facilities may be needed before additional faculty can be added and made productive.

We next turn to the tests of constant versus non-constant returns to scale described in Section 3 for the five disciplines for which the first test did not reject convexity. Here again we ignore the distinction between the haves and the have-nots. The resulting p-values for these tests are given in the third column of Table 3, and these results indicate that the null hypothesis of CRS should not be rejected for any of the five disciplines where we test CRS. The p-values are quite large—0.4022 for Chemistry, and greater than 0.92 for the other four disciplines. As noted earlier, this test is conditional on the outcome of the first test, and this distorts the true size of the test from its nominal value, causing the true size to be smaller than the nominal size. One would worry if the test of CRS resulted in a p value of 0.045, for example, suggesting mild evidence against the null. Here, however, all of our p values are far above any level where one might consider rejecting the null hypothesis of CRS, and any pre-test bias that might exist does not seem to affect our conclusions.

The results of the first two tests determine how the null hypothesis of equivalent mean efficiency among haves and have-nots in each discipline should be tested, i.e., which estimator should be used to implement the test of equivalent mean efficiency. For English and History, we use the FDH estimator of efficiency since convexity of the production set $\mathcal{P}$ was clearly rejected. For Physics, we also use the FDH estimator of efficiency. Although convexity was just rejected for Physics at the 5-percent level, and not at the 1-percent level, the FDH estimator remains consistent regardless of whether $\mathcal{P}$ is convex or non-convex. For the remaining disciplines, where our tests did not come close to rejecting convexity nor CRS, we use the CRS estimator of efficiency, exploiting its faster rate of convergence (as opposed to the rates of the FDH and VRS estimators).

The fourth column of Table 3 gives p-values for our tests of equivalent mean efficiencies across the haves and the have-nots in each discipline. For the first five disciplines, where convexity of the production set could not be rejected, we tested equivalent mean efficiencies using both the CRS and VRS estimators (recall that the VRS estimator remains consistent and achieves the CRS rate under constant returns); for these departments, the first p-value shown corresponds to the test using the CRS estimator, while the second corresponds to the test using the VRS estimator. The p-values are close to 1 for Computer Science and Economics; for these disciplines, the results are clearly in favor of not rejecting the null of equivalent means in favor of higher mean efficiency for the haves. For Ecology and Evolutionary Biology, the p value is almost 0.64 when the CRS estimator is used, but is less than 0.01 when the VRS estimator is used, and so the results are in conflict with each other. Consequently, some caution is warranted for readers who may be inclined to reject the null in this case. For English, the p value is 0.1412, and so here we find no evidence that the null should be rejected. On the other hand, the resulting p-values for Chemistry, Civil and Environmental Engineering, and Physics are many orders of magnitude smaller than 0.01, and so the null is clearly rejected by our tests in favor of higher mean efficiency among haves than among have-nots for these disciplines. The null hypothesis is also rejected for History, but less dramatically than for the others.

Rejection of equivalent means in favor of greater mean efficiency among haves than have-nots is not surprising for Chemistry and Physics, where much of modern research relies on HPC. It is also perhaps not surprising that we reject the null for Civil and Environmental Engineering. Among computer scientists, biologists, and economists, some use HPC in their research, but many do not. Consequently, failure to reject equivalence between haves and have-nots for these disciplines is also not surprising.

Rejection of equivalence for History is surprising at first glance. However, researchers in History have begun to use HPC; HPC is used to analyze digitized texts, and some historians use HPC to analyze global positioning system data. On the other hand, some careful consideration suggests that the results of the means tests for this discipline could be spurious. Recall that convexity of the production set was rejected for History, and hence our means test for History is based on the FDH efficiency estimators. Consequently, the results from Kneip et al. (2014, 2013) which lead to the test statistic given in (3.29) that we use to test mean differences for these two disciplines require that we employ means of sub-samples of the haves and have-nots. Careful consideration reveals that with $(p+q)=4$, the sample means in the first term in parentheses in the denominator of (3.29) are computed using $n_j, \kappa = \sqrt{n_j}$ observations for $j \in \{1,\ 2\}$, where n_1 is the number of haves and n_2 is the number of have-nots. Table 2 indicates that for History, $\sqrt{n_j}$ is 7 and 9 for haves and have-nots, respectively. In the end, it is conceivable that too-few observations are available to give meaningful results for History, and that the tests yield spurious results for History because the FDH estimator must be used. On the other hand similar observations can be made for English and Physics, where convexity was also rejected. In the case of Physics, however, HPC is widely used, and we reject the null hypothesis of equivalent mean efficiency across haves and have-nots within Physics. We fail to reject the null in the case of English, but this is perhaps expected and hence not surprising.

Further insight is obtained by turning to Table 4, where we report sample means of estimated efficiencies among haves and have-nots in each of the eight disciplines we examine.[15] As in Table 3, we report two sets of the results—based on the CRS and VRS estimators, respectively—for the first five departments in Table 4. Table 4 also shows estimated 95-percent confidence intervals for the expected value of efficiency in case. Confidence intervals for the first five disciplines— Chemistry, Civil and Environmental Engineering, Computer Sciences, Ecology and Evolutionary Biology, and Economics were estimated using the central limit theorem result in Theorem 4.3 of Kneip et al. (2014). Since the FDH estimator is used to estimate efficiency for English, History, and Physics, the central limit theorem result in Theorem 4.4 of Kneip et al. (2014) was used to estimate confidence intervals for the last

[15] In each case, the sample means are computed using all observations in a given group. As made clear by Kneip et al. (2014), the "full" sample mean remains a consistent estimator, although it is not useful for making inference if $(p+q)$ is too large.

three disciplines in Table 4. However, in each case, the sample means of estimated efficiencies reported in Table 4 use all available efficiency estimates within a particular group.[16]

As discussed above and in Kneip et al. (2014, 2013), the CRS, VRS, and FDH efficiency estimators are biased, as are sample means of estimates obtained with these estimators. The estimated confidence intervals in Table 4, however, explicitly incorporate corrections for these biases. Consequently, in each case represented in Table 4, the point estimates lie outside of and to the right of the corresponding estimated confidence intervals.

Recall from Table 3 that the null hypothesis of equivalent mean efficiency between haves and have-nots was strongly rejected for Chemistry, Civil and Environmental Engineering, and Physics. In Table 4, estimated mean efficiencies among the haves are much larger than among the have-nots for Chemistry and Civil and Environmental Engineering; the same is true for Physics, though the difference in the two means is less than for the other two disciplines. There is also a large gap between upper bound of the estimated 95-percent confidence interval for the have-nots and the lower bound of the interval for the haves in each of these three disciplines. This is in line with the tiny p values reported for these disciplines in Table 4.

On the other hand, the two sample means for History are very close, but the two estimated confidence intervals do not overlap. This is also consistent with rejection of the null of equivalent means for History in Table 4, and is consistent with the p-value that is less than 0.01 but that is also far larger than the p-values obtained for Chemistry, Civil and Environmental Engineering, and Physics. For Ecology and Evolutionary Biology and for English, the means in Table 4 are also very close; the mean for the haves is only slightly larger than the mean for the have-nots, and the confidence intervals overlap, again consistent with the results of the means test in Table 3.

The results for Computer Science and Economics shown in Table 4 are rather different from the results for the other disciplines. The means test failed to reject equivalent means in favor of higher mean efficiency for haves than for have-nots for both of these departments. Table 4 shows that for Computer Science, the point estimate for the haves is slightly larger

[16] For English, History, and Physics, Theorem 4.4 of Kneip et al. (2014) as well as the test of equivalent mean efficiency across haves and have-nots require a sample mean of a subset of estimated efficiencies within each group since the FDH estimator is used, and its convergence rate of $1/(p+q) = 1/4$ is less than $1/3$. The sample mean based on all of the available efficiency estimates remains, however, a consistent estimator of the corresponding population mean. See Kneip et al. (2014) for additional discussion.

than for the have-nots, but the reverse is true for Economics. Moreover, in both cases, the estimated confidence intervals for the haves lies entirely to the left of the estimated confidence intervals for the have-nots in these two disciplines. For Computer Science and Economics, the results suggest that HPC is associated with *lower* mean efficiency in these disciplines. For Computer Science, one might suppose that local access to HPC leads to and enables more experimental systems research, or perhaps interdisciplinary research with other computational science disciplines by faculty at those institutions as compared to research that is less systems-oriented or is more theoretically-based. More research is needed to know if this is the case, and if so, whehter such research is less efficient in terms of scientific advancement.

As a check on our results, we tested for differences in mean efficiency between haves and have-nots at the institutional level, as opposed to the departmental level; we do not include tables to save space, but the specific results are available on request. Aggregating the NRC that we have used so far across departments is problematic due to missing observations at the department level. Hence, for the aggregate analysis, we used different data obtained from the National Science Foundation. In these data, we do not observe GRE scores at the institutional level; we used instead NSF's Federal Expenditures to replace the GRE input that we use in the paper. For 2003–2006 (roughly the same time period covered by our NRC department-level data), we find that institutions with HPC are more efficient than institutions those without HPC. Specifically, using the institutional level data, we failed to reject convexity as well as constant returns to scale. Using the CRS efficiency estimator, we reject equivalence of mean efficiency across haves and have-nots in favor of greater mean efficiency among the haves. So at both the departmental level as well as at the institutional level, we find evidence that local HPC tends to enhance research efficiency on average, though not in all disciplines.

Figure 2 shows, for each discipline in our study, kernel estimates of the densities of estimated efficiency for haves and have-nots where efficiency has been estimated separately for the two groups in each discipline. The solid curves give density estimates for the haves, while the dotted curves give density estimates for the have-nots.[17]

[17] Recall from the discussion in Section 2 that by construction, the efficiency estimates are bounded below at 0 and above at 1. Ordinary kernel density estimators are biased and inconsistent near support boundaries. This problem is addressed by using the reflection method described by Silverman (1986) and

The density estimates shown in Figure 2 are in accordance with the sample means shown in Table 4. For example, among Chemistry departments in our sample, mean efficiency is greater for haves than for have-nots (significantly so, given the earlier results in Table 3). Similar observations hold for Civil and Environmental Engineering and for Physics. Computer Science and Ecology and Evolutionary Biology show similar patterns in Figure 2, although results of the means tests in Table 3 fail to reject the null of equivalent means for these two disciplines. Economics exhibits the reverse pattern, again consistent with the results in Table 4.

The plots for Chemistry, Civil and Environmental Engineering, and Physics shown in Figure 2 suggest that the efficiency distributions of the haves might stochastically dominate the efficiency distributions of the have-nots. Recall that for two random variables A and B, first-order stochastic dominance of B by A occurs if and only if $\Pr(A \geq t) \geq \Pr(B \geq t) \; \forall \; t$ and $\Pr(A \geq t) > \Pr(B \geq t)$ for some t. A necessary, but not sufficient, condition for this is $E(A) > E(B)$. Unfortunately, however, standard tests for stochastic dominance are not easily adapted to the present situation due to the fact that only estimates of efficiency, as opposed to the true efficiencies, are observed. The same issues described by Kneip et al. (2014, 2013) that complicate inference about mean efficiency and testing for differences in mean efficiency also affect testing for stochastic dominance, where the problem is more difficult. More research is needed on this.

6 Conclusions and Directions for Future Work

Large sums of money are allocated each year by government agencies and by academic research institutions to encourage basic research, including funds that are directed at providing HPC resources. Yet these allocation decisions are made outside competitive markets, where prices would provide information about marginal costs and marginal benefits that might lead to *efficient* allocations. We have found clear evidence that research departments in Chemistry, Civil and Environmental Engineering, and Physics (and to a lesser extent, History) are more efficient in universities where HPC is locally available. This is likely true for other disciplines as well, but for which we lack sufficient data. However, it remains to

Simar and Wilson (1998) to produce the density estimates shown in Figure 2.

attach an economic value to this greater efficiency, and this is difficult since research output is not traded in markets, at least not directly. One could perhaps shed some light on this by looking at data on salaries of faculty researchers, but no easily-available data exist to do this. Moreover, academic job markets are arguably not perfectly competitive.

While we have found that the presence of HPC resources seems to increase research output, we have not yet considered how it might change the nature of the research that is done. Obviously, problems that require HPC resources cannot be solved without HPC. Availability of local HPC resources may cause some researchers to work on *different* problems than they would otherwise, perhaps even to think of problems that might not occur to them if HPC were not available. Our ongoing research is using methods from machine-learning to categorize research in Chemistry into identifiable types, with the hope that this can be used to examine how the increasing availability as well as capability of HPC over time may have influenced the mix of types of research being produced. This work, however, is at present in its early stages.

References

Afriat, S.: 1972, Efficiency estimation of production functions, *International Economic Review* **13**, 568–598.

Apon, A., Ahalt, S., Dantuluri, V., Gurdgiev, C., Limayem, M., Ngo, L. and Stealey, M.: 2010, High performance computing instrumentation and research productivity in U.S. universities, *Journal of Information Technology Impact* **10**, 87–98.

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

Bonaccorsi, A. and Daraio, C.: 2003, A robust nonparametric approach to the analysis of scientific productivity, *Research Evaluation* **12**, 47–69.

Brooks, H.: 1986, National science policy and technological innovation, *in* R. Landau and N. Rosenberg (eds), *The Positive Sum Strategy*, Vol. 1, National Academy Press, Washington, DC, pp. 119–165.

Charnes, A., Cooper, W. W. and Rhodes, E.: 1978, Measuring the efficiency of decision making units, *European Journal of Operational Research* **2**, 429–444.

Daouia, A., Simar, L. and Wilson, P. W.: 2013, Measuring firm performance using nonparametric quantile-type distances. Discussion paper, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Deprins, D., Simar, L. and Tulkens, H.: 1984, Measuring labor inefficiency in post offices, *in* M. M. P. Pestieau and H. Tulkens (eds), *The Performance of Public Enterprises: Concepts and Measurements*, North-Holland, Amsterdam, pp. 243–267.

Dongarra, J. J., Bunch, J., Moler, C. and Stewart, G.: 1976, *LINPACK Users' Guide*, SIAM Publications, Philadelphia, PA.

Färe, R.: 1988, *Fundamentals of Production Theory*, Springer-Verlag, Berlin.

Farrell, M. J.: 1957, The measurement of productive efficiency, *Journal of the Royal Statistical Society A* **120**, 253–281.

Furman, J. L., Porter, M. E. and Stern, S.: 2002, The determinants of national innovative capacity, *Research Policy* **31**, 899–933.

Jeong, S. O., Park, B. U. and Simar, L.: 2010, Nonparametric conditional efficiency measures: asymptotic properties, *Annals of Operations Research* **173**, 105–122.

Kepner, J.: 2004, High performance computing productivity model synthesis, *The International Journal of High Performace Computing Applications* **18**, 505–516.

Kneip, A., Park, B. and Simar, L.: 1998, A note on the convergence of nonparametric DEA efficiency measures, *Econometric Theory* **14**, 783–793.

Kneip, A., Simar, L. and Wilson, P. W.: 2008, Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models, *Econometric Theory* **24**, 1663–1697.

Kneip, A., Simar, L. and Wilson, P. W.: 2011, A computationally efficient, consistent bootstrap for inference with non-parametric DEA estimators, *Computational Economics* **38**, 483–515.

Kneip, A., Simar, L. and Wilson, P. W.: 2013, Testing hypotheses in nonparametric models of production. Discussion paper #2013/48, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Kneip, A., Simar, L. and Wilson, P. W.: 2014, When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores, *Econometric Theory* . Forthcoming.

Leyden, D. P. and Link, A. N.: 1992, *Government's Role in Innovation*, Kluwer Academic Publishers, Boston, MA.

NAS: 2008, *The Potential Impact of High-End Capability Computing on Four Illustrative Fields of Science and Engineering*, The National Academies Press, Washington, DC. National Research Council of the National Academies.

O'Donnell, C. J., Rao, D. S. P. and Battese, G. E.: 2008, Metafrontier frameworks for the study of firm-level efficiencies and technology ratios, *Empirical Economics* **34**, 231–255.

Ostriker, J. P., Holland, P. W., Kuh, C. V. and Voytuk, J. A.: 2011, A data-based assessment of research-doctorate programs in the united states (2011). Available online at http://www.nap.edu/rdp/.

Park, B. U., Jeong, S.-O. and Simar, L.: 2010, Asymptotic distribution of conical-hull estimators of directional edges, *Annals of Statistics* **38**, 1320–1340.

Park, B. U., Simar, L. and Weiner, C.: 2000, FDH efficiency scores from a stochastic point of view, *Econometric Theory* **16**, 855–877.

PCAST: 2010, Report to the president and congress—designing a digital future: Federally funded research and development in networking and information technology. Executive Office of the President, President's Council of Advisors on Science and Technology.

PITAC: 1999, Information technology research: Investing in our future. President's Information Technology Advisory Committee Report to the President.

Robinson, P. M.: 1988, Root-n-consistent semiparametric regression, *Econometrica* **56**, 931–954.

Shephard, R. W.: 1970, *Theory of Cost and Production Functions*, Princeton University Press, Princeton.

Silverman, B. W.: 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Simar, L. and Wilson, P. W.: 1998, Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* **44**, 49–61.

Simar, L. and Wilson, P. W.: 1999a, Some problems with the Ferrier/Hirschberg bootstrap idea, *Journal of Productivity Analysis* **11**, 67–80.

Simar, L. and Wilson, P. W.: 1999b, Of course we can bootstrap DEA scores! But does it mean anything? Logic trumps wishful thinking, *Journal of Productivity Analysis* **11**, 93–97.

Simar, L. and Wilson, P. W.: 2000, Statistical inference in nonparametric frontier models: The state of the art, *Journal of Productivity Analysis* **13**, 49–78.

Simar, L. and Wilson, P. W.: 2007, Estimation and inference in two-stage, semi-parametric models of productive efficiency, *Journal of Econometrics* **136**, 31–64.

Simar, L. and Wilson, P. W.: 2008, Statistical inference in nonparametric frontier models: Recent developments and perspectives, *in* H. Fried, C. A. K. Lovell and S. Schmidt (eds), *The Measurement of Productive Efficiency*, 2nd edn, Oxford University Press, Oxford, chapter 4, pp. 421–521.

Simar, L. and Wilson, P. W.: 2011a, Inference by the m out of n bootstrap in nonparametric frontier models, *Journal of Productivity Analysis* **36**, 33–53.

Simar, L. and Wilson, P. W.: 2011b, Two-Stage DEA: Caveat emptor, *Journal of Productivity Analysis* **36**, 205–218.

Simar, L. and Wilson, P. W.: 2013, Estimation and inference in nonparametric frontier models: Recent developments and perspectives, *Foundations and Trends in Econometrics* **5**, 183–337.

Sterling, T.: 2004, Productivity metrics and models for high performance computing, *The International Journal of High Performace Computing Applications* **18**, 433–440.

Tichenor, S. and Reuther, A.: 2006, Making the business case for high performance computing: A benefit-cost analysis methodology, *CTWatch Quarterly* **2**, 2–8.

Top500: 2013, Top 500 Supercomputer Sites. Available online at http://www.top500.org/.

Toutkoushian, R. K. and Webber, K.: 2011, Measuring the research performance of postsecondary institutions, *in* J. C. Shin, R. K. Toutkoushian and U. Teichler (eds), *University Rankings*, Springer, pp. 123–144.

Table 1: Summary Statistics for Inputs and Outputs by Discipline

	Mean	Std. Dev.	Minimum	Maximum
Chemistry				
Faculty	23.65	11.63	1.81	76.76
Average GRE Score	717.18	33.05	612.50	797.14
Publications	61.97	52.59	3.71	305.05
PhDs granted	11.54	10.04	1.00	60.40
Civil and Environmental Engineering				
Faculty	19.45	13.62	2.46	73.49
Average GRE Score	748.74	36.53	626.67	800.00
Publications	22.48	19.11	1.77	101.33
PhDs granted	5.68	5.76	1.00	31.20
Computer Sciences				
Faculty	26.92	17.41	5.05	143.38
Average GRE Score	771.89	27.84	593.00	800.00
Publications	50.20	47.35	4.60	332.51
PhDs granted	7.22	6.79	1.00	37.20
Ecology and Evolutionary Biology				
Faculty	23.23	12.26	1.59	65.08
Average GRE Score	676.99	43.07	540.00	767.67
Publications	33.73	23.44	2.14	136.27
PhDs granted	4.98	4.29	1.00	26.40
Economics				
Faculty	25.41	12.20	6.54	59.36
Average GRE Score	757.72	42.14	579.17	800.00
Publications	14.27	10.85	0.87	56.00
PhDs granted	7.44	5.73	1.00	26.40
English Language and Literature				
Faculty	35.40	13.45	5.12	69.85
Average GRE Score	635.92	51.76	387.00	748.00
Publications	369.38	206.91	45.47	1083.55
PhDs granted	7.91	4.64	1.00	24.20
History				
Faculty	29.41	15.12	1.18	74.36
Average GRE Score	607.30	50.21	482.86	740.00
Publications	338.28	250.37	0.00	1260.65
PhDs granted	6.06	5.30	1.00	29.20
Physics				
Faculty	29.62	16.86	3.00	88.19
Average GRE Score	760.00	29.06	658.00	800.00
Publications	122.94	106.90	4.16	539.79
PhDs granted	7.09	6.08	1.00	29.60

Table 2: Number of Observations for each Discipline

		Have HPC?	
	n	Y	N
Chemistry	180	61	119
Civil and Env. Engineering	130	53	77
Computer Science	127	52	75
Ecology and Evol. Biology	94	39	55
Economics	117	53	64
English	119	44	75
History	137	56	81
Physics	161	64	97

Table 3: Hypothesis Test Results

	p-values		
Department	Convexity	RTS	Means
Chemistry	0.8367	0.4022	8.20×10^{-172} 6.11×10^{-101}
Civil and Env. Engineering	0.9144	0.9257	2.83×10^{-80} 2.92×10^{-32}
Computer Science	0.8134	0.9999	0.9999 0.9999
Ecology and Evol. Biology	0.9811	0.9999	0.6391 0.0001
Economics	0.7259	0.9961	0.9999 0.9999
English	2.46×10^{-6}	—	0.1412
History	4.33×10^{-7}	—	0.0009
Physics	0.0463	—	8.48×10^{-40}

NOTE: Two *p*-values for the test of equivalent mean efficiency between haves and have-nots are given for the first five disciplines where neither convexity nor CRS could be rejected. In each case, the first *p*-value was obtained using the CRS estimator, while the second *p*-value was obtained using the VRS estimator.

Table 4: Mean Efficiencies for Haves and Have-Nots

	Mean Efficiency	
	Haves	Have-Nots
Chemistry	0.6252 (0.5566, 0.6091)	0.3485 (0.1902, 0.2036)
	0.6714 (0.4567, 0.5054)	0.4623 (0.1861, 0.2057)
Civil and Env. Engineering	0.6307 (0.5059, 0.5499)	0.4674 (0.2776, 0.3002)
	0.7010 (0.4507, 0.4953)	0.5538 (0.3001, 0.3285)
Computer Sciences	0.5271 (0.2466, 0.2753)	0.4976 (0.3483, 0.3721)
	0.6445 (0.2221, 0.2513)	0.5814 (0.2648, 0.2879)
Ecology and Evol. Biology	0.6949 (0.5310, 0.5578)	0.6239 (0.5342, 0.5617)
	0.7727 (0.5695, 0.6030)	0.6967 (0.5314, 0.5695)
Economics	0.5605 (0.4425, 0.4704)	0.6443 (0.5496, 0.5785)
	0.6402 (0.4189, 0.4485)	0.7007 (0.5495, 0.5808)
English	0.9359 (0.7729, 0.7918)	0.9238 (0.7561, 0.7843)
History	0.9429 (0.7965, 0.8169)	0.9372 (0.7699, 0.7882)
Physics	0.9453 (0.8111, 0.8428)	0.8639 (0.5454, 0.5962)

NOTE: Estimated 95-percent confidence intervals are given in parentheses. For the first five disciplines, two sets of results are given; in each case, the first set were obtained with the CRS estimator, while the second set was obtained with the VRS estimator.

Figure 1: Hypotheses to be Tested

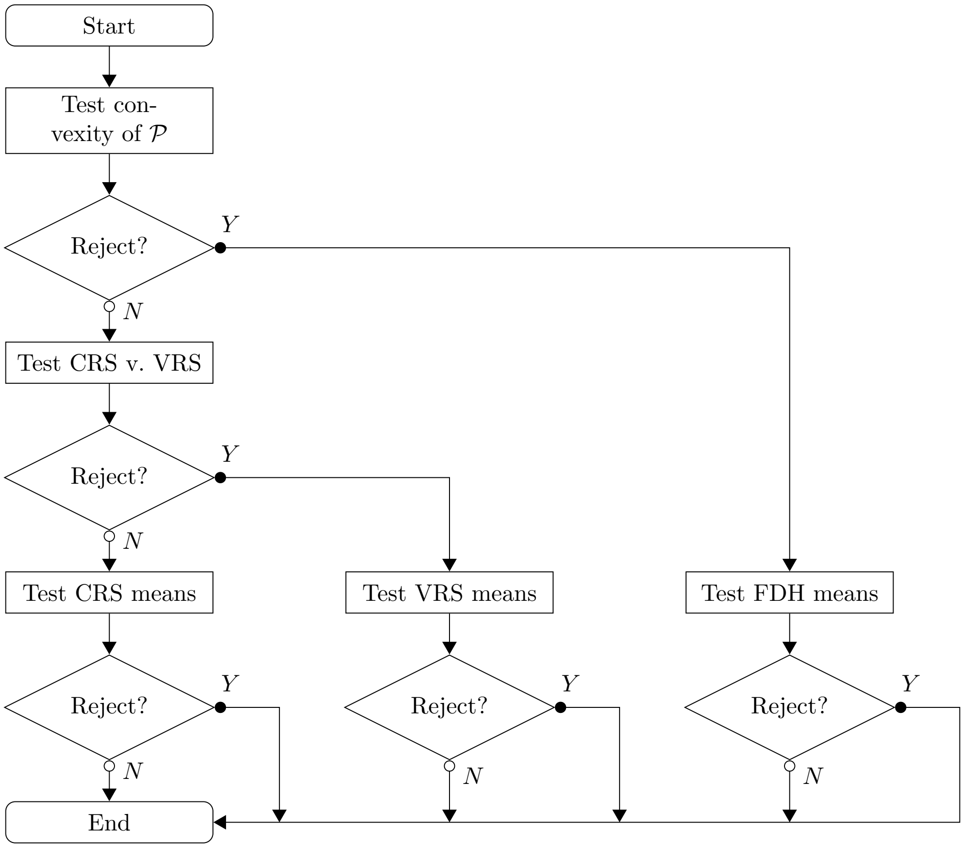
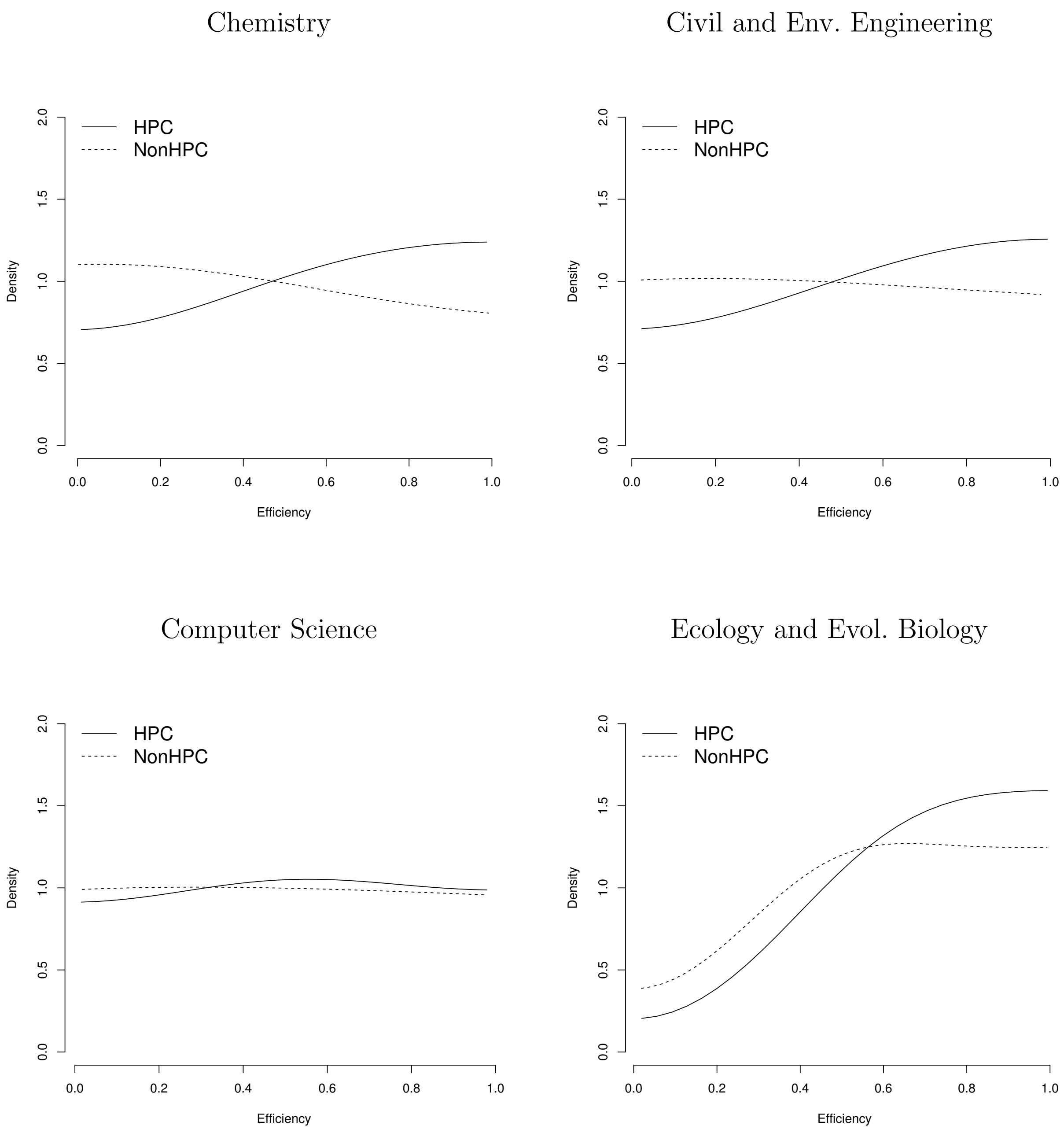


Figure 2: Estimated Densities of Efficiencies by Department and Group



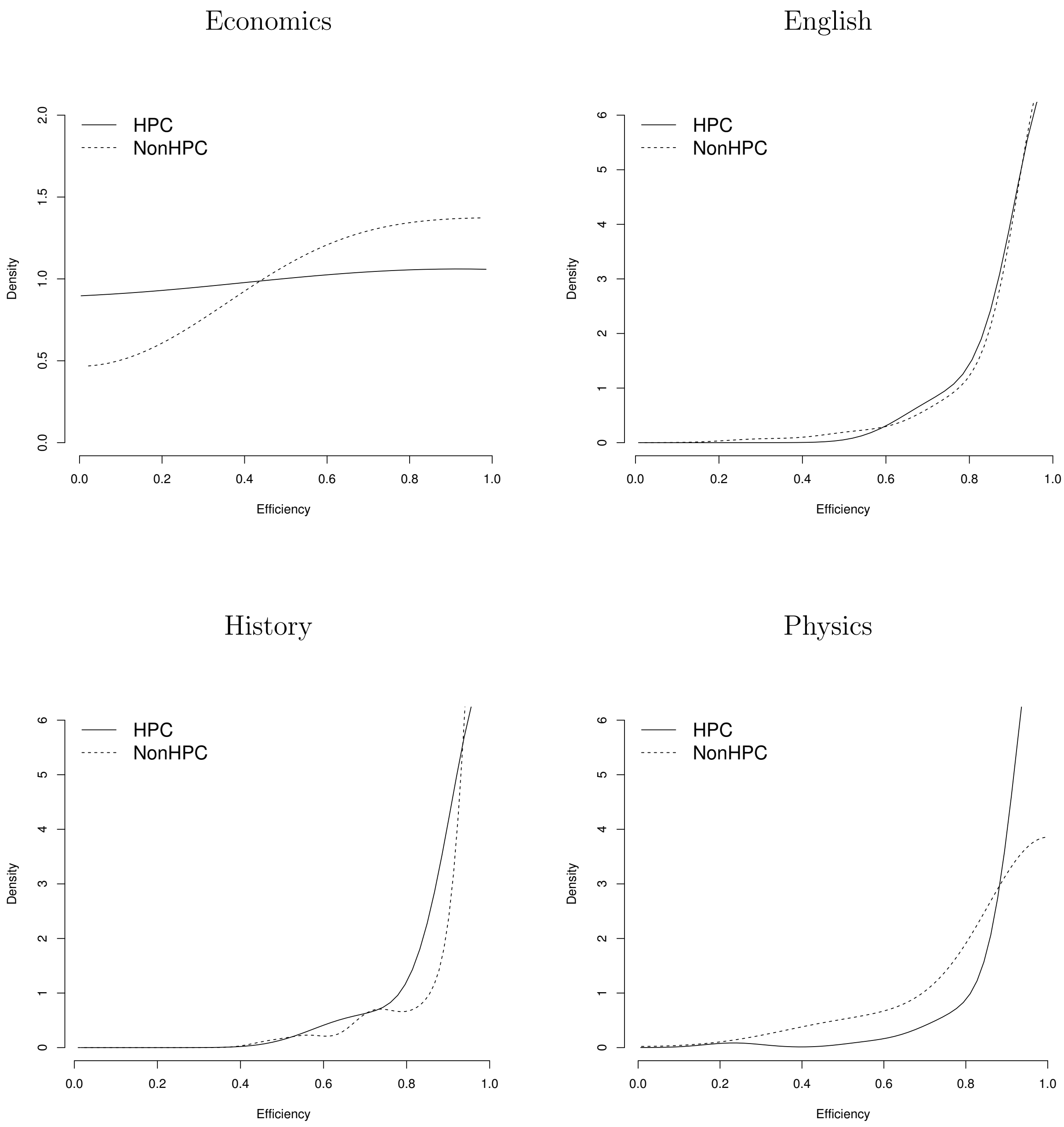
Chemistry

Civil and Env. Engineering

Computer Science

Ecology and Evol. Biology

Figure 2: Estimated Densities of Efficiencies by Department and Group (continued)



Economics

English

History

Physics

Note: Vertical axes in density plots for English, History, and Physics (where the FDH estimator was used) range from 0 to 6; vertical axes in the remaining plots (where the CRS estimator was used) range from 0 to 2.