IFN Working Paper No. 712, 2007

How Should Research Performance Be Measured? A Study of Swedish Economists

Magnus Henrekson and Daniel Waldenström

Research Institute of Industrial Economics
P.O. Box 55665
SE-102 15 Stockholm, Sweden
info@ifn.se
www.ifn.se

How Should Research Performance Be Measured? A Study of Swedish Economists*

Magnus Henrekson† and Daniel Waldenström‡

December 17, 2010

Abstract: Billions are allocated annually to university research. Increased specialisation and international integration of research and researchers have sharply raised the need for comparisons of performance across fields, institutions and individual researchers. However, there is still no consensus regarding how such rankings should be conducted and what output measures to use. We rank all full professors in a particular discipline, economics, in one country using seven established, and some of them commonly used, measures of research performance. We show both that the rank order varies greatly across measures, and that depending on the measure used the distribution of total research output is valued very differently.

Keywords: Impact of research, Ranking, Research output, Research productivity, Bibliometrics, Google Scholar, *h*-index, Impact Factor, SSCI.

JEL: A11, A13, A14, B41.

* We thank Niclas Berggren, Anders Björklund, David Colander, Robin Douhan, Bruno Frey, Randall Holcombe, Henrik Jordahl, Dan Klein, Assar Lindbeck, Kevin O'Rourke, Andrew Oswald, Donald Siegel, Roy Thurik, Heinrich Ursprung, Hans-Joachim Voth, and seminar participants at Stockholm University, Örebro University and the Stockholm School of Economics for useful comments and suggestions on earlier versions of this paper. Skilful research assistance was provided by Martin Olsson and Johan Egebark. Financial support from the Jan Wallander and Tom Hedelius Re-search Foundation and the Marianne and Marcus Wallenberg Foundation is gratefully acknowledged.

† Research Institute of Industrial Economics (IFN), P.O. Box 55665, SE-102 15 Stockholm, Sweden. Ph: +8-46-6654502. E-mail: Magnus.Henrekson@ifn.se. Web: www.ifn.se/mh.

‡ Research Institute of Industrial Economics (IFN), P.O. Box 55665, SE-102 15 Stockholm, Sweden. Ph: +8-46-6654531. E-mail: Daniel.Waldenstrom@ifn.se. Web: www.ifn.se/danielw.

1 INTRODUCTION

Increased integration has made the total research market larger, but also more complex. In some disciplines, such as economics, there exists a core of general ideas and concepts that are widely acknowledged and used. At the same time, numerous global and specialized sub-disciplines have emerged and continue to emerge. Enhanced cross-border cooperation, integration and exchange of research and researchers have inspired the need for comparisons of performance across fields, institutions and individual researchers.

Considerable efforts have been made in economics to assess research output and productivity, and to rank individual researchers and institutions. However, there is still no consensus regarding how such rankings should be conducted and what output measures are appropriate to use in order for research funds to be efficiently allocated.[1]

This study aims to shed light on the extent to which the assessments depend on the bibliometric measure used. This is done by analyzing the research performances of a reasonably homogeneous population of researchers: full professors in economics in Sweden. While the scope is admittedly limited in some respects, the study's main points should apply to other countries as they face similar challenges in a globalized research environment.

Before we proceed to our analysis, a brief description of the Swedish university system is called for. The Swedish university system is part of the public sector. Professors have, by and large, been civil servants, which implies that a high degree of national uniformity has been imposed on pay schedules, rules for promotion and recruitment and other working conditions. This is still largely the case today, although it should be noted that greater flexibility in terms of pay schedules was introduced dur-

[1] These assessments have been done for economics departments in the U.S. (e.g. Conroy and Dusansky 1995; Dusansky and Vernon 1998) and their graduate programs (Grijalva and Nowell, 2008), and in recent years also for European economics departments (e.g., Kalaitzidakis et al. 1999, 2003; Combes and Linnemer 2003; Coupé 2003; Axarloglou and Theoharakis 2003; Lubrano et al. 2003; Tombazos 2005). The journal ranking of Kalaitzidakis et al. (2003) appears to have been particularly influential in Europe and especially among young economists (Oswald, 2006).

ing the 1990s.[2] In theory, rules governing promotions to full professor should be uniform across universities. But in practice, standards began to vary in line with a gradual increase of universities and university colleges and the introduction of a system of non-chaired professors (lecturers promoted to full professors) in the mid-1990s. This reform more than doubled the number of full professors in less than a decade.

In its 2008 Research Bill (2008/09:50), the Swedish government announced that effective from 2010, the allocation of public research funding across universities would be based on relative performance measured by scientific output and the ability to attract funding from non-governmental sources. In January 2009, the government commissioned the Swedish Research Council to collect the necessary data to develop and calculate a metric of scientific output and impact to govern inter-university resource allocation. The Swedish system resides thus in a formative stage, and the significance of choosing an appropriate measure(s) for resource allocation results cannot be overstated. As far as we can see, the UK Research Excellence Framework is very similar, and in an equally formative stage.[3]

Our analysis considers seven of the most established and commonly used measures of research performance. Three are based on journal publications, three draw on the number of citations to the researcher's most cited works and one counts raw, unadjusted output. Each measure has its *pros* and *cons*; in most cases, it is not obvious which is the most appropriate. We rank the Swedish professors using each of these measures and study to what extent the rankings match each other in terms of overlap. We also examine whether the individuals' performances differ across the measures and, if so, for what reasons.

In doing so, a considerable variation in the rank order across measures becomes apparent. In short, a researcher's output is valued very differently depending on the output measure used. One of them—the renowned measure of Kalaitzidakis et al. (2003)—stands out in particular, as it gives rise to a singularly skewed distribution of performances among professors: the professors at the very top are attributed a very

[2] Henrekson and Rosenberg (2001) describe the institutional setup of the Swedish university system and contrast it to the system in the United States.
[3] Higher Education Funding Council for England (2010).

large share of the total output, while the absolute contribution of the lower half of the population remains negligible.

All seven measures provide relevant information about the performance of individual researchers, although there are no doubt additional aspects that all of these measures overlook. For instance, only a small subset of economics journals are included in the weight-based measures, and most measures either ignore or give little weight to impact outside economics or on policymaking.[4] While quantitative measures are essential for assessing research, they cannot fully substitute for the careful assessment of the works of individual researchers.[5]

2 MEASURES OF RESEARCH OUTPUT

This section presents the seven measures of research output used to assess the research performance of Swedish economics professors. They comprise:

Measures based on weighted journal publications:

1. Sum of KMS weighted journal articles (Kalaitzidakis et al. 2003)
2. Sum of IF weighted journal articles (Thomson Scientific 2003)
3. Sum of KY weighted journal articles (Kodrzycki and Yu 2006)

Measures based on citations to most cited works:

4. Sum of citations of five most cited works in the SSCI
5. Sum of citations of the five most cited works in Google Scholar (GS)
6. Individual *h*-index (Hirsch 2005; Harzing 2007)

Measures based on the number of international publications:

7. Number of published works (articles, book chapters, books) in *EconLit*

The seven measures capture most of the relevant dimensions of quantifying the volume and quality of individual research performance.

[4] There are legitimate objections that can be raised against giving any weight to impact outside of academia. One could argue that impact on policy-making is a different dimension altogether; it is not usually peer-reviewed, for example, and so does not meet the first test of academic research. At the same time, there is a case to be made, especially in small countries, for the importance of local economists being engaged with policy debates and not focusing exclusively on academic research (Frey and Eichenberger 1993).

[5] Van Fleet et al. (2000) come to the same conclusions after having examined the use of journal rankings as explicit targets for researchers in management departments.

2.1 Measures Based on Weighted Journal Publications (KMS, IF, KY)

The most commonly used group of measures by far for assessing and ranking researchers, institutions and journals is average citation-weighted journal publications. While the measures do deviate from one another in a number of ways, they all emanate from the same basic set of principles. First, they only count journal articles as scientific production. Second, they define a weight, or quality, for each journal based on how other articles and journals have cited its articles these have been adjusted in various ways (i.e., correcting for self-citations, age, size, impact etc.). Third, they assume that all articles published in the same journal carry exactly the same degree of scientific merit, equal to the journal-specific citation-based weight.

When assessing individuals using these journal weights, scores are calculated by simply multiplying a person's various articles by the weight of their respective journals and then summing these figures. Institutions are assessed by summing the scores of all their affiliated researchers.

Kalaitzidakis et al. (2003) holds the honour of being one of the most widely cited rankings of economics journals during recent years. This ranking received enormous attention, not least among European economists, after being included in the 2003 special issue of the *Journal of the European Economic Association*, which described it as being "the most up-to-date set of objective journal weights available" (Neary et al. 2003, p. 1247).[6] These weights, which we call KMS, assign relative merits to each of the 159 journals in the "economics" category in the Journal Citation Reports (JCR). In constructing these weights, the authors first take the total number of citations from the journals during the ten years before 1998. Then the authors exclude within-journal citations ("self"-citations) and "older" citations, defined here as articles published before 1994 (that is to say, older than four years). In order to remove any influence of journal size on the number of citations received, Kalaitzidakis et al. compute a weight that relates each journal's annual number of pages with the average number of pages of all journals. Finally, and most importantly, citations are weighted according to their

[6] The background was an ambition to improve the knowledge regarding the status of European economics research. To this effect, the European Economic Association in 2000 invited bids for constructing journal and scholar rankings. Besides the KMS measure, there were four other measures also presented. It should be noted that Neary et al. made it clear that the EEA did not endorse any of the measures.

"impact", meaning that citations from relatively well-cited journals are given more weight than citations coming from journals that are cited less often.[7] In short, KMS is a journal weight controlling for self-citation, age, size and "impact", based on actual citations during the late 1990s. The distribution of KMS weights is skewed towards the top. For example, a single article in the *American Economic Review* is valued more highly than ten articles in the *Journal of Financial Economics*, 25 articles in the *Journal of Law and Economics*, 60 articles in the *Journal of Health Economics*, and all the 400+ articles ever published in the *Journal of Evolutionary Economics* since its first issue in 1991.

While KMS is among the most influential journal rankings in European economics, the most well-known measure across all fields of science is without doubt the *Impact Factor* (IF). Calculated by Thomson Reuters (which also runs SSCI), the IF is reported in *Journal Citation Reports*. It is defined as one year's average number of citations to a journal's articles that were published during the two preceding years.[8] In this study, we employ IF weights for journals in the JCR "economics" category, which means that the number of weighted journals is restricted to 169 (in 2003). Unlike KMS, IF weights all citations equally and does not exclude within-journal citations. Indeed, these differences lend a sizeable effect. For example, the IF score for the *Journal of Health Economics* (1.778) is only marginally lower than the IF score of the *American Economic Review* (1.938), while the *AER* outweighs the *Journal of Health Economics* in KMS by a factor of 60. Generally, even a journal ranked 150 requires fewer than 10 articles to obtain an IF score on par with one *AER* article.

Kodrzycki and Yu (2006) have constructed a third measure using weighted journal publications. The main contribution of the Kodrzycki and Yu (KY) measure is that it expands the set of journals credited for citations to include other social science disciplines as well. To mitigate this narrowness, Kodrzycki and Yu extend both the number of cit*ed* journals (to 181) and, more importantly, the number of cit*ing* journals to include all social science journals in the SSCI (currently more than 2,300). As in the

[7] This impact adjustment is made using a simple iteration algorithm that was originally suggested by Liebowitz and Palmer (1984). In the KMS case, letting C be the number of citations from journal j to journal i and Z a size adjustment factor for i, the impact at iteration round t is $I_{i,t} = [\sum_j(C_{ij})/ Z_i]\cdot I_{j,t-1}$.

[8] IF's two-year time window is arguably too short for assessing the true impact of an economics article. Still, unless this constraint has a systematically different effect across the research quality distribution, it cancels out in the comparisons across authors.

case of IF, the KY measures draw their citation information from the JCR issue of 2003. But unlike the IF's short window of the past two years' citations, KY use citations of journal articles published between 1996 and 2003. (As compared to KMS, which used citations from 1994–1998 reported in the 1998 issue of JCR.) Hence, to the extent that citation patterns change over time, which they do as persuasively shown by Kim et al. (2006), the degree of overlap between these sets of weights is reduced.

These journal-based measures have been criticized on several grounds. One is that they give the same merit to all articles in a journal regardless of their actual impact (see discussion in the next section). Moreover, they only consider journals in JCR's "economics" category (albeit KY has a somewhat extended selection), which excludes some of the journals that economists actually cite the most, e.g., the *Industrial and Labor Relations Review*, the *Journal of Finance* and the *Review of Financial Studies*. Because economists regularly interact with neighbouring disciplines such as finance, statistics, law, political science, medicine, criminology, psychology and sociology, these exclusions are quite problematic.[9]

Figure 1 displays the cumulative distribution of weights attributed to the economics journals included in the three journal weight measures. The ten most highly ranked journals in KMS comprise 50 percent of the sum of all journal weights; the corresponding figure for IF is only 25 percent. The figure also shows that the KMS and KY measures both weight the top 5 journals very highly. Since roughly 80 percent of all journals in *EconLit* are given zero weight by all three measures,[10] just one percent of the *EconLit* journals constitute 50 percent of the total KMS weight.

[Figure 1 about here]

[9] Yet another critique directed towards KMS cites its use of the number of pages per year as control for relative size of journals. Palacios-Huerta and Volij (2004) and Kodrzycki and Yu (2006) argue that the most relevant level of analysis of scholarly work is articles, and that a more appropriate size adjustment focuses on the number of articles per year.

[10] *EconLit* lists 975 journals in 2010.

2.2 Measures Based on Citations to Most Cited Works (SSCI, GS, hind)

As already mentioned, a common critique levelled against measures like KMS, IF and KY is that they assign the same value to all articles appearing in a journal based on its (adjusted) number of citations to that journal. By doing this, they disregard the fact that different articles have quite different impacts. In counting the number of citations that each of the eighteen articles in a 1981 issue of AER received over the following 25 years, Oswald (2007) finds that they ranged from 401 to 0. In other words, the publication of an article in the AER (or any other top-ranked journal) does not guarantee that the article will be widely cited.[11]

We include two measures that account for actual—rather than assumed—citations to the scholars' works. The first such measure is the sum of citations of the five most cited works of each professor, as recorded in the Social Sciences Citation Index (SSCI). We choose to sum the five most cited works in order to strike a balance between counting all citations of all works (which may give rise to spurious results at the lower end) or only of the single most cited work (which would give too little credit to scholars with a number of well-cited works). The SSCI citation database is probably the world's largest and is widely used to assess the impact of individual researchers (see, e.g., Klein and Chiang 2004). But there are some important caveats about the SSCI citations to keep in mind as well: only citations from journals in the SSCI (a minority of all existing journals) are recorded.

Our second measure of actual citations is drawn from a data source that is used less commonly: the Internet database Google Scholar (GS). GS and SSCI differ in several important respects. First, GS records citations arise from a much larger pool of publication types, including working papers, reports and books and academic journals that are available on the Internet.[12] Second, whereas SSCI only counts citations to articles appearing in SSCI journals, GS allows any of its recorded publication items to be among an author's cited works. The significance of this is illustrated by the fact that

[11] Laband and Tollison (2003) report that in their citation count over the subsequent five years of all articles published in 1996, 70 percent of the articles in 91 journals received one citation or less. Wall (2009, p. 9) shows that "large percentages of articles in the highest-ranked journals are cited less frequently than are typical articles in much-lower-ranked journals."

[12] GS also claims to exclude self-citations, although we have found examples where this is not true. For clarity reasons, we have thoroughly analysed each professor's citations in the GS database, ascertaining that we have the correct author and that there are five distinct cited works.

of the ten most cited works in GS in our sample of Swedish professors, two were textbooks, one was a monograph and one was a book chapter, which represent together 41 percent of all citations in the top 10 group.[13] Another equally interesting observation is that of the six journal articles in the top 10, only two were also among the respective author's five most cited articles in SSCI. This indicates the potential importance of citations, and thereby "outside" scientific impact, beyond the realm of the SSCI.

The *h*-index suggested by Hirsh (2005) attempts to quantify the scientific productivity and impact of a scientist based on his/her most quoted papers. A scientist with an *h*-index of *x* has published *x* papers that have at least *x* citations. The *h*-index differs from the previous two citation measures in that it emphasizes sustained productivity instead of a few successful publications.[14]

The original *h*-index does not account for the number of co-authors in a paper. We therefore implement Harzing's (2007) alternative *individual h*-index, h_{ind}, that first normalizes the number of citations for each paper by dividing the number of citations by the number of authors for that paper and then calculates the *h*-index of the normalized citation counts. The *individual h*-index is based on GS citations using the software *Publish or Perish* by Harzing (2007).[15]

2.3 Measures Based on the Number of Publications (Works)

Lastly, the seventh measure includes all kinds of internationally published output in a researcher's performance, given that the publication is listed in EconLit.[16] This is a pure quantity measure, and encompasses internationally published works such as journal articles, book chapters and monographs. Working papers and reprints of already published works are excluded. Of course, a raw output measure does not capture several of the most desired dimensions of scholarly work, including adjustment

[13] Text books are not peer-reviewed and should arguably not be included in the database for reasons of comparability. However, since we do not observe whether a work has been peer-reviewed or not (not all journal articles, even in the SSCI, have undergone strict peer review), singling out textbooks may give rise to unwarranted arbitrariness, and hence they are retained in the sample.

[14] It may do so too strongly. Two scientists may have the same *h*-index, say, $h = 25$, but one has 20 papers that have been cited more than 500 times and the other has none. Clearly, the output of the former is more valuable.

[15] The calculations were made on September 15, 2007, using *Publish or Perish*, version 2.3.

[16] No doubt, there are a number of other measures used, such as the Laband and Piette (1994) measure which is still widely used, particularly in the U.S. Coupé (2003) uses several additional measures. However, the inclusion of further measures would not strengthen the general point made in this paper.

for the quality of a publication. Nevertheless, this measure provides a useful point of reference.

3 DATA

We use a newly constructed database that contains all international research published by all active full professors in economics tenured at Swedish universities as of winter of 2007.[17] The population consists of 93 professors—87 men and 6 women—born between 1939 and 1968. Our study involves professors in Sweden and not Swedish professors; hence, foreign-born scholars appear in our sample, while Swedes tenured at foreign universities do not.

Information on individual publications was retrieved partly from *EconLit* and partly from departmental and personal websites. We have computed the researchers' scores for each of the seven measures by combining bibliographical data with the journal weights or citation counts. We adjust for co-authorship in all cases by weighting down publications with *n* co-authors by 1/*n*, which is in line with most previous work in this type of literature. Table 1 presents summary statistics for the performances in the seven cases.

[Table 1 about here]

4 EMPIRICAL ANALYSIS

4.1 To What Extent Do the Rankings Overlap?

By studying both correlations for the entire population and overlaps among the top 10 ranked professors, this section examines the degree of overlap between the professors' rankings. Tables 1a and 1b display Spearman (rank order similarities) and Pearson (absolute similarities) correlations. The results suggest a high degree of overlap between the measures. However, the overlap is substantially larger *within* the different

[17] As professors in economics we include chaired and non-chaired professors (*befordringsprofessorer*) at economics departments at Swedish universities and colleges, as well as "professors in economics" active at other university departments (e.g., industrial dynamics, economics and management). A list of all professors including affiliation, birth year, year of Ph.D. and year of promotion to full professor (for the first time at a Swedish university) is available upon request.

types of measures (i.e., between the journal weight-measures, KMS, KY and IF, on the one hand and the three citation measures, SSCI, GS and h_{ind}, on the other) than *between* the two types of measures. Still, KY and IF appear to be more correlated with SSCI, GS and h_{ind} than KMS. Moreover, the correlations of the raw, quality-unadjusted output measure, Works, are substantially smaller than in all the other measures.

The second assessment concerns the overlap of the top 10 ranked professors. Although it only involves a small group of scholars, they represent between one quarter and about one half of the aggregate performance in the analyzed measures. Figure 2 displays the ranks in all measures for the ten top-ranked professors according to KMS. There appears to be a fairly high degree of overlap between the journal weight measures (KMS, IF, KY), but considerably less so between KMS and the citation measures (SSCI, GS and h_{ind}), as well as with the unweighted output measure (Works). For example, of the top 10 scholars in the KMS ranking, only between four and five are ranked as top 20 by the citation measures. Strikingly, one of the top 10 based on KMS even ranks near the bottom of the distribution (as number 78 and 81) in terms of citations in SSCI and GS.

[Figure 2 about here]

4.2 Distribution of Performances

While analyzing the ordinal rankings of researchers across the different measures is of course essential, it is almost as important to assess the cardinal ordering across performances. Bibliometric measures that capture the absolute distances between researchers underlie hiring and funding decisions; the relative merit awarded to the top compared to the bottom can greatly affect both salaries and the size of research grants.

Figure 3 illustrates the distribution of scholar performances in the seven measures. After inspecting the cumulative relative frequencies, it appears that the GS measure gives the largest weight to the absolute top whereas the KMS measure gives the least rewards to the bottom half. At the other end, Works and h_{ind} appear to be the least "elitist" among the measures.

[Figure 3 about here]

Slightly more formal on the distributional differences, Table 2 reinforces the impression from Figure 3 that the differences across measures are sizable. In some cases the distributions are particularly skewed towards the top, while in other cases they tilt towards the middle. Most notably, professor performances according to KMS stand out to be the most skewed distribution overall, even though GS is more top heavy.

[Table 2 about here]

Specifically, KMS has the highest P90/P10 ratio of all measures by far. In contrast, GS has the highest P90/P50 ratio; GS assigns almost no value to the research output of the median professor. In terms of the share of total performances attributable to the top 10 professors, KMS had 43 percent while GS had over 50 percent. The lower half of the population based on KMS represents only 7.3 percent of total performance, whereas this share is somewhat larger for KY (8.9 percent) and about twice as large for IF, SSCI and GS, three times larger in Works (25.8 percent) and almost five times larger in h_{ind} (35.5 percent).

4.3 What Determines Success?

The above analyses found considerable variation in terms of both overlap and skewness of research performances across the bibliometric measures. It is not clear, however, whether this variation implies that there is no core set of individual characteristics that determine successful research performance, or if the fact that the measures capture different aspects of scholarly impact allows for such characteristics to play a consistent role. In order to address this issue, we regress the professors' individual performances on a set of background variables drawn from our database. The right hand side variables are the following: (i) sex (for which we have no prior regarding its effect on research performance); (ii) affiliation at an established research university (expected positive effect)[18] (iii) age when Ph.D. degree was received (no prior); (iv)

[18] We regard the universities in Gothenburg, Lund, Stockholm, Umeå and Uppsala and the Stockholm School of Economics as established research universities in economics. These universities have a long history and have had Ph.D. programs in economics for more than 40 years. Since the mid 1990s, a number of colleges have been granted university status and several of them have started Ph.D. pro-

number of years between receiving the Ph.D. and being promoted to full professor (expected negative effect) and (v) the number of years as professor up until 2007 (expected positive effect).[19] The regression equation then looks as follows:[20]

$$\ln [1 + \textit{Research performance}]_i = \alpha + \beta_1 \cdot \textit{Research university}_i + \beta_2 \cdot \textit{Age at PhD}_i + \beta_3 \cdot \textit{Years to professor}_i + \beta_4 \cdot \textit{Professor years}_i + \beta_5 \cdot \textit{Female}_i + u_i \quad (1)$$

Admittedly, this model provides a rather simple framework for analyzing the determinants of research performance; the results indicate therefore conditional correlations and not causal effects.[21]

Table 3 presents the regression results. First, receiving a Ph.D. at a relatively young age seems associated with a better performance in one's subsequent career, which might reflect that the more talented require less time to become "licensed" researchers. Second, the table shows that the more years needed to become full professor, the lower the performance. Third, there is no positive effect of having been professor a long time, suggesting that "older" professors become less productive over time. Fourth, affiliation with a research university is associated with significantly higher performance for most measures except for h_{ind} and Works, in which this seems to have no effect at all. Fifth, female professors perform worse than men. Given the small number of female Professors in our sample (six), this result should be interpreted with caution.

[Table 3 about here]

grams in recent years. Moreover, colleges that do not have the rights to grant Ph.D.s may still have full professors as a result of a mid-1990s reform. Our definition of research universities is uncontroversial.

[19] For a similar analysis of tenure success of Swedish economists, see Tasiran et al. (1997). We also tried using a squared term of this variable, capturing the potential life-cycle effect on scientific production discussed by Rauber and Ursprung (2006), but without finding any significant impact.

[20] We log one plus the research performance in order to include also those professors whose scores are zero.

[21] For example, unobservable variables are likely to be correlated with some of the included variables (e.g., being at a research university may be related to a number of personal characteristics that drive research performance) and causality could also be bi-directional in the case of affiliation at a research university.

4.4 How Should Research Performance Be Measured?

So far our analysis has documented the distributional characteristics and similarities across the seven research measures. We have yet to identify, however, which is the most encompassing and useful for universities, funding agencies and others who wish to evaluate research qualities. In order to remark on this matter—and thereby answer the question asked in the very title of our study—we propose a simple mechanism for defining the most preferable measure. In short, the most useful measure is the one that is the *most correlated with all the other measures*.[22] This approach is based on the recognition that all the measures incorporate at least some relevant dimension of research output. Therefore, the most useful measure should be the one that encompasses most of these dimensions, and hence diverges the least from all the others.

In practice, we obtain the "optimal" measure by computing the sum of correlation coefficients across all measures and then selecting the one that has the largest sum. In other words, we define an

$$\text{"Optimal" research measure } i = \max \sum_j \rho^S_{ij} , \qquad (2)$$

where ρ^S_{ij} denotes the Spearman correlation between measures i and j. Of course, the Pearson correlation coefficients ρ^P_{ij} can also be used, incorporating not only the rank order differences but also absolute distances between performances.

Table 4 presents the results of this exercise. The highest ranked measure in both correlation types is IF, and KY is second.[23] These two journal weight-based measures hence offer the most comprehensive assessment of a researcher's performance. In the case of Spearman correlation, the three citation-based measures SSCI, GS and h_{ind}, comes next. Using Pearson correlations, KMS ranks third, quite a bit behind KY but about level with SSCI. In sum, if we were to recommend one single measure of research performance, we would choose IF, since this captures more relevant dimensions of scholarly output than any of the other measures taken individually.

[22] We are grateful to one of the referees for suggesting this methodological approach.
[23] This result is highly robust. For example, removing one measure at the time from the maximization algorithm does not alter IF as the first ranked except in one single case.

[Table 4 about here]

5 CONCLUDING DISCUSSION

What should a measure of research performance capture? Citation based measures are used most frequently, and for good reasons. Still, it may not be true that the most (least) cited research is also the best (worst) research.[24] Can we assume that all important research results are published in refereed journals, or should we also include monographs, book chapters and textbooks? Is it sufficient to evaluate research based on in which journal an article is published or how many citations it gets?[25] Should we give weight to research's impact outside academia, such as influence on policymaking or the policy debate?

The importance of these questions differs depending on the issue at hand. Quantitative measures often offer guidance in hiring, tenure and promotion decisions, and in the allocation of research funds across individuals, research groups, departments, disciplines and universities, as well.

In this article we analyze seven of the most established and commonly used measures of research output by applying them to the publications of all full professors in economics in one country, Sweden. Our findings suggest large discrepancies between the measures in terms of both the rank order of professors and the absolute differences between their performances.

Relative ranking and quality-adjusted quantification of research output is no temporary fad. On the contrary, it will likely continue to gain in importance. As soon as a certain measure is widely used, researchers can be expected to adjust behaviour in order to maximize their output as defined by this measure (Holmström and Milgrom 1991; Frey and Osterloh 2006). This tendency is reinforced if universities, departments and research councils use a certain metric when making decisions about hiring,

[24] For example, famous scholars tend to receive an excess number of citations due to their past performance (Coupé 2003; Ursprung and Zimmer 2007), while articles settling academic debates or that provide important robustness checks tend to receive almost no citations at all (Mayer 2004; van Dalen and Klamer 2005).

[25] As Oswald (2007) shows, even in the top-ranked journals there are several articles that receive no or very few citations.

promotion, and the allocation of funds (Holcombe 2004; Oswald 2007; Drèze and Estevan, 2007). Therefore, the choice of measures is of great importance unless it emerges that the ranking and relative valuation of different researchers and departments is largely invariant with respect to an array of output measures. The evidence presented in this study speaks strongly against any presumption of this sort.

References

American Economic Association, "Document Types Indexed in EconLit." Online: http://www.aeaweb.org/econlit/doctypes.php (accessed October 1, 2009).

Axarloglou, Kostas, and Vasilis Theoharakis (2003), "Diversity in Economics: An Analysis of Journal Quality Perceptions." *Journal of the European Economic Association* 1(6), 1402–1423.

Combes, Pierre-Philippe, and Laurent Linnemer (2003), "Where Are the Economists Who Publish?" *Journal of the European Economic Association* 1(6), 1250–1308.

Conroy, Michael E., and Richard Dusansky (1995), "The Productivity of Economics Departments in the US: Publications in the Core Journals." *Journal of Economic Literature* 33(4), 1966–1971.

Coupé, Tom (2003), "Revealed Performances: Worldwide Rankings of Economists and Economics Departments, 1990–2000." *Journal of the European Economic Association* 1(6), 1309–1345.

Dusansky, Richard, and Clayton J. Vernon (1998), "Rankings of U.S. Economics Departments." *Journal of Economic Perspectives* 12(1), 157–170.

Drèze, Jacques H., and Fernanda Estevan (2007), "Research and Higher Education in Economics: Can We Deliver the Lisbon Objectives?", *Journal of the European Economic Association* 5(2-3), 271–304.

Frey, Bruno S., and Reiner Eichenberger (1993), "European and American Economists and Economics." *Journal of Economic Perspectives* 74), 185–193.

Frey, Bruno S., and Margit Osterloh (2006), "Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives." Working Paper, Institute for Empirical Research in Economics, University of Zürich.

Harzing, Anne-Wil (2007), "Publish or Perish." Online: http://www.harzing.com (October 4, 2007).

Henrekson, Magnus, and Nathan Rosenberg (2001), "Designing Efficient Institutions for Science-Based Entrepreneurship: Lesson from the US and Sweden". *Journal of Technology Transfer* 26(2), 207–231.

Higher Education Funding Counsil for England (2010), "Research Excellence Framework." Online: http://www.hefce.ac.uk/Research/ref/ (accessed October 3, 2009).

Hirsch, Jorge E. (2005), "An Index to Quantify an Individual's Scientific Research Output." *Proccedings of the National Academy of Sciences* 102(46), 16569–16572.

Holcombe, Randall G. (2004), "The National Research Council Ranking of Research Universities: Its Impact on Research in Economics." *Econ Journal Watch* 1(3), 498–514.

Holmström, Bengt, and Paul M. Milgrom (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7(1), 24–52.

Grijalva, Therese C., and Clifford Nowell (2008), "A Guide to Graduate Study in Economics: Ranking Economics Departments by Fields of Expertise", *Southern Economic Journal* 74(4), 971–996.

Kalaitzidakis, Pantelis, Theofanis P. Mamuneas and Thanasis Stengos (1999), "European Economics: An Analysis Based on Publications in the Core Journals." *European Economic Review* 43(4–6), 1150–1168.

Kalaitzidakis, Pantelis, Theofanis P. Mamuneas and Thanasis Stengos (2003), "Rankings of Academic Journals and Institutions in Economics." *Journal of the European Economic Association* 1(6), 1346–1366.

Klein Dan B., and Eric Chiang (2004), "Citation Counts and SSCI in Personnel Decisions: A Survey of Economics Departments." *Econ Journal Watch* 1(1) 166–174.

Kodrzycki, Yolanda K. and Pingkang David Yu (2006), "New Approaches to Ranking Economics Journals." *Contributions to Economic Analysis & Policy* 5(1), article 24.

Kim, E. Han, Adair Morse and Luigi Zingales (2006), "What Has Mattered to Economics Since 1970?" *Journal of Economic Perspectives* 20(4), 189–202.

Laband, David N., and Michael J. Piette (1994), "The Relative Impacts of Economics Journals 1970–1990." *Journal of Economic Literature* 32(2), 640–666.

Laband, David N., and Robert D. Tollison (2003), "Dry Holes in Economic Research." *Kyklos* 56(2), 161–174.

Liebowitz, Stanley J., and John P. Palmer (1984), "Assessing the Relative Impacts of Economic Journals." *Journal of Economic Literature*, 22(1), pp. 77–88.

Lubrano, Michel, Luc Baluwens, Alan Kirman and Camelia Protopopescu (2003), "Ranking Economics Departments in Europe: A Statistical Approach." *Journal of the European Economic Association* 1(6), 1367–1401.

Mayer, Tomas (2004), "Dry Holes in Economic Research: Comment." *Kyklos* 57(4), 621–626.

Neary, J. Peter, James A. Mirrlees and Jean Tirole (2003), "Evaluating Economics Research in Europe: An Introduction." *Journal of the European Economic Association* 1(6), 1239–1249.

Oswald, Andrew J. (2006), "Prestige Labels." *Royal Economic Society Newsletter*, issue No. 135, October.

Oswald, Andrew J. (2007), "An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers." *Economica* 74(293), 21–31.

Palacios-Huerta, Ignacio, and Oscar Volij (2004), "The Measurement of Intellectual Influence." *Econometrica* 72(3), 963–977.

Rauber, Michael, and Heinrich W. Ursprung (2006), "Life Cycle and Cohort Productivity in Economic Research: The Continental European Experience as Exemplified by the Case of Germany." *German Economic Review* 9(4), 431–456.

Swedish Government Bill 2008/09:50, *A Boost to Research and Innovation***.** Stockholm: Ministry of Education and Research.

Tasiran, Ali C., Ann Veiderpass and Bo Sandelin (1997), "Climbing Career Steps: Becoming a Full Professor of Economics." *Scandinavian Journal of Economics* 99(3), 471–484.

Thomson Scientific (2003), *Journal Citation Reports 2003. Social Sciences Edition,* Institute for Scientific Information, Philadelphia.

Tombazos, Christis G. (2005), "A Revisionist Perspective of European Research in Economics." *European Economic Review* 49(2), 251–277.

Ursprung, Heinrich, W., and Markus Zimmer (2007), "Who is the 'Platz-Hirsch' of the German Economics Profession." *Journal of Economics and Statistics* 227(2), 187–208.

Van Dalen, Hendrik P., and Arjo Klamer (2005), "Is Science A Case of Wasteful Competition?" *Kyklos* 58(3), 395–414.

Van Fleet, David D., Abagail McWilliams and Donald S. Siegel (2000), "A Theoretical and Empirical Analysis of Journal Rankings: The Case of Formal Lists." *Journal of Management* 26(5), 839–861.

Wall, Howard J. (2009), "Don't Get Skewed Over by Journal Rankings." *B.E. Journals of Economic Analysis & Policy* 9(1), 1–10.

Table 1: Correlations of individual performance across the seven measures

a) Spearman rank correlations

	KMS	IF	KY	SSCI	GS	h_{ind}	Works
KMS	1						
IF	0.858	1					
KY	0.955	0.922	1				
SSCI	0.543	0.680	0.627	1			
GS	0.505	0.617	0.590	0.847	1		
h_{ind}	0.490	0.633	0.561	0.674	0.788	1	
Works	0.415	0.564	0.434	0.507	0.502	0.685	1

Note: The number of observations is 93 in all cases. All coefficients are significant at the 1%-level.

b) Pearson correlations

	KMS	IF	KY	SSCI	GS	h_{ind}	Works
KMS	1						
IF	0.826	1					
KY	0.955	0.879	1				
SSCI	0.699	0.733	0.731	1			
GS	0.651	0.649	0.675	0.795	1		
h_{ind}	0.538	0.681	0.617	0.703	0.712	1	
Works	0.436	0.623	0.456	0.430	0.389	0.617	1

Note: The number of observations is 93 in all cases. All coefficients are significant at the 1%-level.

Table 2: Summary statistics and concentration estimates for performances

	Journal weight measures			Citation count measures			
Variable	KMS	IF	KY	SSCI	GS	h_{ind}	Works
Metric type	Score	score	score	cites	cites	cites	works
Mean	103.4	5.6	63.1	46.7	154.2	7.8	16.5
Median	52.0	3.9	31.1	28.0	69.7	7.0	14.6
C.V.	1.4	1.0	1.3	1.1	1.6	0.6	0.7
Min	0.0	0.0	0.0	0.0	4.5	1.0	2.0
Max	961.2	35.7	605.0	297.2	1438.0	28.0	50.7
Skewness	2.98	2.33	3.22	2.12	3.50	1.46	1.25
P90/P10	97.7	16.5	57.8	22.1	19.0	4.3	5.4
P90/P50	4.8	3.0	5.1	3.6	5.6	1.9	2.1
Share of top-10 (%)	43.2	34.8	40.9	35.6	50.6	24.1	26.9
Share of low half (%)	7.3	15.6	8.9	13.9	12.6	35.5	25.8

Note: All measures are weighted for co-authorship. C.V. stands for coefficient of variation. P90/P10 is the ratio between the 90th percentile professor (P90) and the 10th percentile professor (P10), and analogously for the P90/P50 ratio.

Table 3: Linking performance to individual background

	KMS	IF	KY	SSCI	GS	h_{ind}	Works
Age at Ph.D.	–0.08*	–0.05*	–0.09**	0.01	–0.05	–0.04**	–0.02
	(0.04)	(0.02)	(0.03)	(0.04)	(0.03)	(0.01)	(0.02)
Years to professor	–0.04	–0.04**	–0.05*	–0.07**	–0.06**	–0.04**	–0.02
	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
Professor years	0.01	0.01	0.01	0.01	0.00	–0.00	0.03**
	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
Research Univ.	1.73**	0.56**	1.35**	0.68*	0.32	–0.04	0.06
	(0.35)	(0.16)	(0.32)	(0.28)	(0.21)	(0.11)	(0.14)
Female	–1.82**	–0.42	–1.22	–0.65	–0.65	–0.52**	–0.32
	(0.67)	(0.38)	(0.67)	(0.62)	(0.46)	(0.16)	(0.17)
Constant	5.44**	3.14**	5.86**	3.39*	6.59**	3.90**	3.10**
	(1.49)	(0.72)	(1.31)	(1.34)	(1.11)	(0.50)	(0.69)
Observations	93	93	93	93	93	93	93
R–squared	0.45	0.37	0.45	0.25	0.23	0.39	0.32

Note: Robust standard errors in parentheses. For definitions of variables see the main text. * significant at 5%; ** significant at 1%.

Table 4: Selecting the "optimal" measure of research performance.

	Spearman correlation ρ^S		Pearson correlation ρ^P	
Rank	Measure	max $\sum\rho^S$	Measure	max $\sum\rho^P$
1	*IF*	5.27	*IF*	5.39
2	*KY*	5.09	*KY*	5.31
3	*SSCI*	4.88	*KMS*	5.10
4	*GS*	4.85	*SSCI*	5.09
5	h_{ind}	4.83	*GS*	4.87
6	*KMS*	4.77	h_{ind}	4.87
7	*Works*	4.11	*Works*	3.95

Figure 1: Cumulative distribution of journal weighting schemes.

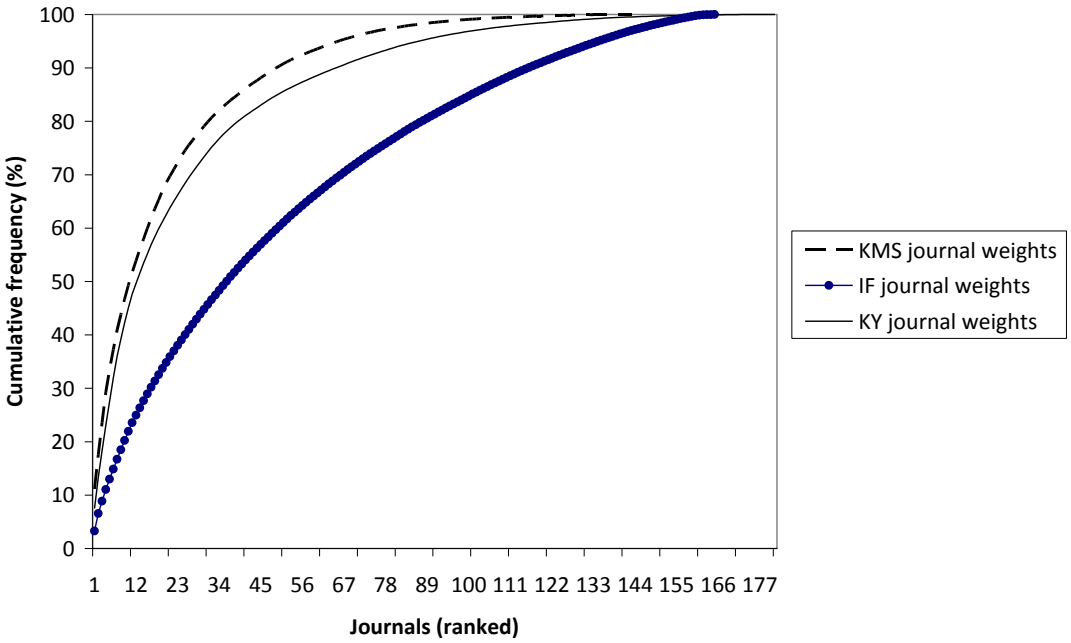

Figure 2: Rankings of the KMS top-10 professors in the other measures

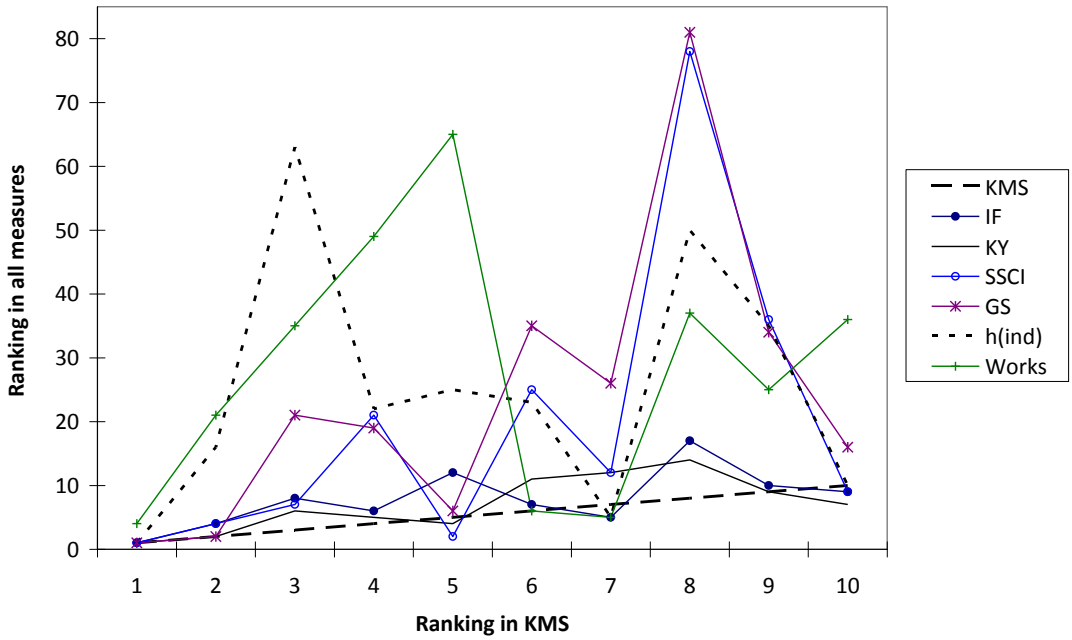

Figure 3: Distribution of professor performances in all seven measures.