ISSN 1755-5361

University of Essex

Department of Economics

Discussion Paper Series

No. 757 December 2014

A Tale of Two Metrics: Research Assessment vs Recognised Excellence

Pierre Regibeau and Katharine E Rockett

A Tale of Two Metrics: Research Assessment vs Recognised Excellence

Pierre Régibeau[1] and Katharine E. Rockett[2]

December 2014

Abstract

We build an economics department entirely composed of Nobel Prize winners and evaluate it using standard research assessment metrics. Performing the same evaluation on existing departments, we find that the rating of our Nobel Prize department does not stand out from other good departments. Compared to recent research evaluations, our Nobel Prize department's ranking is less stable. This suggests a significant effect of score "targeting" induced by the rankings exercise. We find some evidence that modifying the assessment criteria to increase the total number of publications considered can help distinguish the top.

JEL Codes: H4, I23, L51, O38

Keywords: Research Evaluation, Research Excellence Framework, Nobel Prize

[1] Charles River Associates and Imperial College, London

[2] University of Essex and CEPR

We would like to thank Liutauras Petrucionis for research assistance. All remaining errors are our own.

1. INTRODUCTION

Over the last twenty years, many countries have set up systems to evaluate the performance of their institutions of higher learning as an incentive scheme for improving the quality of provision and as a tool for allocating government funds. The quality of the research performed at these institutions has been a particular point of focus. The manner in which academic research is assessed varies both across countries and over time. In the UK, an initially quite discrete and non-discriminating system was gradually replaced by finer divisions, evolving into continuous rankings of departments in the 2008 research assessment exercise. Primarily journal-based rankings of research outputs have been complemented by other measures, such as impact and the research environment in more recent periods. The system continues to evolve, with discussions of citations measures to be used in the next assessment in 2020. Systems in other countries have been reviewed in EC (2008). These systems vary widely in how and whether they evaluate quality and how many research outputs they evaluate. In some systems, fewer than one publication per full time equivalent member of staff is submitted for review, while in others all publications are counted. In some assessments, a publication's quality surmounts the threshold if it appears in a JCR journal whereas others have a fine ranking of output quality or rely on peer review panels to evaluate contribution.

The current diversity of approaches reflects the different purposes to which research evaluations are put and the different contexts in which research operates in different countries. At the same time, one would wish that any methodology could, as a minimum, be validated against some widely accepted evaluation criteria. In many cases, no such external and widely accepted criteria exist. Our basic assumption here, however, is that a department composed solely of Nobel Prize winners could be considered as an absolute standard of excellence in research, irrespective of which precise criteria are proposed. In other words, we struggle to imagine an institution which, if populated solely by those destined to win Nobel prizes, should be judged to be anything other than top quality in research. Indeed, one would think that the research conducted by Nobel laureates over their career would not only be judged to be of high academic merit but that it would also tend to have quite significant "impact" in the sense defined in the last UK research excellence framework (REF).

In view of this, we consider that a fictitious department assembled from future Nobel-winners would be a good measure of the "ideal" pursued by academic researchers so that it can be used as benchmark to gauge the reliability of the measures that policy makers apply to actual departments. We use this insight to conduct a brief investigation of research assessment exercise-style evaluation of research outputs, including not only the "top 4" system employed in the UK, but also some variants to reflect the differences among countries in the weighting applied to number and quality of outputs. While we concentrate on Economics, our methodology is quite general and could be applied to other disciplines.

This note attempts to address three questions. Firstly, how well would an "ideal" Nobel department fare under current measures and how sensitive would its performance be to the simple randomness of publication, the size of the department, and the age profile of that department? Secondly, how does our Nobel department compare to some leading UK departments when some standard criteria used for research evaluation are applied? Does this help us shed light on the nature of the – not always observable – "adjustments" that actual department use in order to perform well? Finally, are there simple

changes in the criteria applied that would help our Nobel department stand out compared to other high performing but non-Nobel populations?

In addressing these questions, we restrict ourselves to a rather mechanical application of research assessment. This is for both practical and conceptual reasons. If research assessment exercises tend to involve not only area specialists but also the screening and organising activities of deans of research and other university offices interested in maximising departmental scores, then some of the softer aspects of the research evaluation exercise can be lost in the face of the more mechanical aspects of ranking research outputs. Similarly, we ignore the environment and impact aspects of the exercise, as it has been applied in the UK. Indeed, we have no objective way of evaluating what the environment in a department purely populated with Nobel Prize laureates would be: this depends on interpersonal skills as well as pure research quality and so we are unable to judge this portion of the exercise. As an aside, environment is weighted far less than research outputs in the current version of the research evaluation framework in the UK (and in other countries), so we concentrate on the main element of the evaluation – individual research outputs -- only.

To preview our results, in answer to our first question we find in our simulations that the same underlying Nobel population generates a wide range of "ratings" for a department of Nobel quality individuals. Indeed, for a small department, the same population can generate a research assessment score of as low as 1.5 and as high as 4. Simulations of larger departments generate less variance, clearly, but still one would expect the ranking of the same population to vary significantly across research assessment cycles. Further, in order to raise the rating of our simulated department to various cut-off assessment scores, we find that for high scores (3.5 and 3.8) a very large proportion of faculty members would need to be dropped from the exercise in order to hit the target score. This large variance in final score and potentially large number of faculty left un-submitted simply due to random variation with no change in underlying departmental quality has implications for how one should "action" individual assessment results. We do not pursue this implication, but we leave it as a point to consider.

In response to the second question, we find that the current REF-like metric fails to distinguish between our "Nobel" departments and one of the highest ranked UK departments, suggesting a lack of discriminating power at the top. We also observe much less variance in real research assessments than in our simulations, which suggests a variety of effects that the rankings themselves may be having on targeting behaviour of both hiring institutions and those they hire. When we take our Nobel population, assuming that all our individuals are of equal "Nobel" quality, to be stable across our simulations we observe a large change in ranking of simulated departments when our Nobels are aged one period (and replaced with a random selection of juniors from the same sample and so of the same average expected quality). This change in ranking is simply due to random variations in the publications cycle. The inertia that we observe in the actual rankings would need to overcome this by targeting individuals with particular publications outcomes and by inducing targeted movements of employees in response to these outcomes. This certainly confirms casual observation.

In response to the third question, we attempt other criteria for research assessment such as including more, or all, (quality weighted) publications in the research assessment period or doing a simple count of outputs. We find that we need a rather drastic increase in the number of individual publications considered in order to obtain better separation of the Nobel Prize group from highly ranked UK departments without Nobel Prizes. In particular, moving to all quality weighted publications or including a weight on total number of publications seems to improve the relative performance of the Nobel group.

Our paper fits into a general stream of literature on research rankings. Recently, this journal carried a special issue on journal rankings and the Research Excellence Framework (REF). These papers review and produce a meta-ranking of journals that takes into account the audience for which the ranking is targeted (Hudson, 2013), evaluate the role of citations in conducting quality evaluations (Laband, 2013), and discuss how peer review panels should combine output counts with citations to generate an overall view of quality (Sgroi and Oswald, 2013). Other literature has looked at research assessment more generally, investigating accusations of bias (Clerides et al, 2011), and a lack of stability in rankings when the balance of quality/quantity weightings and citations systems change (Frey and Rost, 2010)[3]. We use the Hudson meta-ranking in our analysis, but we do not address citations for reasons we outline below. Our concern with lack of stability in rankings is quite different from Frey and Rost's, as we address stability over time rather than in response to a change in ranking methodology. Indeed, our rankings are relatively robust to the variants we attempt.

In contrast to these papers, our main focus is on the top quality of research output and how this is identified in rankings. Top quality output has been considered specifically in a few papers. Gans and Shepherd (1994), focussing on Nobel and Clark prize winners, comment anecdotally that even discipline-based reviewers may not easily distinguish high quality work, a point echoed by Starbuck (2005, 2006). Abramo et al (2009) add to this the concern that selection procedures by universities of which outputs to submit to the research assessment, which may or may not be conducted primarily by experts in the field depending on who and how this selection is conducted, is the weakest stage of the evaluation process.

Our work could be thought of as making a similar point to Gans and Shepherd but where we show that a standard research assessment metric – the quality of four top papers in an assessment period -- may not easily distinguish high quality work from very good work as opposed to peer review, which is their focus. We can add a few points to this, however. Specifically, we can comment on the stability of rankings predicted by our underlying population and compare these to the stability we observe in actual rankings in order to make an observation about the effects of rankings on "targeting" behaviour of universities. Even if a university's underlying quality remains the same, our simulations using our Nobel data suggest that rankings should shift a lot from one assessment period to the next due to random variation in the publishing cycle for individuals. Hence, targeting is having more of an effect than keeping the underlying quality of the population the same at an institution: it is maintaining the same quality rating within the period. In other words, an individual with a good publication "run" can move to a department that targets the quality level of this "run". This is not quite the same as hiring an individual of the same underlying quality regardless of the most recent publication "run" and may generate the greater stability that we observe. Our contribution here, other than to raise the point, is to put numbers on this effect.

Given that the previous literature indicates that high quality output is poorly detected by peer review and internal selection procedures, we restrict ourselves to a mechanistic view of research evaluation. In other words, our view is that the "softer" element of such exercises as the REF, mainly peer review through reading of the outputs, will not systematically improve the accuracy of the mechanistic process that we specify -- whereby outputs are evaluated purely based on journal rankings. We also have a more practical reason to limit ourselves to evaluating quality using journal rankings only: we will be basing our analysis on a long history of Nobel Prize winner output, which spans a period during which citations

[3] Frey and Osterloh (2011) provide an extensive review and evaluation of the literature on ranking systems for academic work.

patterns and peer views changed markedly. This makes it difficult to perform either a qualitative or a citations analysis of their output. We discuss modifications to our system in part 6 of the paper, below.

The rest of his note is organised as follows. Section 2 presents our methodology. Section 3, then, examines how a "Nobel" department would rate under the type of assessment used in the UK of four top papers. The comparison between this ideal department and some of the leading UK department is performed in section 4. Section 5 discusses various robustness issues and suggests some directions for further work. Section 6 concludes.

2. METHODOLOGY

To create our "Nobel" departments, we start with all recipients of a Nobel Prize in Economics since its inception in 1969 to 2013. From this set, we exclude a few outliers whose performance is known to have been affected by illness or exceptional events. This leaves us 68 individuals for whom we collect lifetime publications, date of birth (and death when relevant) as well as the date at which they entered academic work. We can then partition the career of these laureates into intervals of six years each and allocate their publications to each of these periods. Maybe because of the fairly advanced age of Nobel laureates so far, each individual in our set was professionally active for at least six periods of six years. However, most of our results will be based only on 4 "middle" periods for each individual, roughly corresponding to ages from the early 30s to about 60. Ignoring the first period makes sense since the REF has special provisions for faculty members who have graduated recently. We also ignore the last period as some individuals effectively retire at some point during this period, while others go on publishing even in retirement. In fact, this restriction has the added advantage that it limits us to periods before the individuals received their Nobel Prize. We might expect that their publications changed markedly after this event. Our interest is not in the effect of the prize on publications but rather the output of those who merit the prize. Hence, our focus on productive periods before winning the prize is appropriate for our purposes.

We used standard indexing sources (such as JSTOR and RePec) to generate a list of economics publications for each individual. We rated each publication according to the meta-ranking of Hudson (2013) in this journal. Where this ranking was not available, we used the Keele rankings table[4]. These are on a four point scale, with 4 the maximum quality rating. We then aggregated these into an average score for each individual in each period, using the top four publications, the top five, the top eight, and also the total score over all publications, the total unweighted count of publications and the average score over all publications.

We did not earmark publications by subfields. In this respect, we believe that relying on a population of Nobel laureates is again useful as, presumably, the distribution of individuals – or their publications – across fields is likely to reflect the profession's opinion of "what matters" at the time. This does not, of course, mean that field does not account for some variations in output across Nobel laureates but these variations should be seen as "incidental", i.e. as not reflecting any difference in the underlying quality of the research profiles.

[4] We include books, book chapters, and reports as output, but as a practical matter university press books were the main non-journal outputs that received weight. These received a rating of 3. Other details of our implementation of the rankings are available from the authors.

Given this database, we can generate "Nobel" departments of any size. We start with the number *N* of individuals in the department and randomly draw N cells out of the set of 4x68 "six years publication periods" that we have. It is therefore entirely possible that we will draw the same individual twice. We can then compute the average REF rating of the N individual department. We repeat this process for 1000 repetitions to obtain a distribution of scores for department size N. We can, of course, constrain our choices to be for particular age distributions across the department as well as for different department sizes. Unless otherwise mentioned, we constrain the age distribution to be uniform over the four publication periods for all departments and generally use a department size of 48[5].

3. HOW DOES RECOGNISED EXCELLENCE FARE?

3.1. ASSESSED QUALITY AND VARIATION

Although a thousand repetitions may seem many, there are still some differences between the distributions obtained from two independent 1000 repetitions. These differences are, however, small so that a "typical" department distribution can be seen from one such exercise. Figure 1 presents such a "typical" probability density function of departmental REF scores when we create departments with 48 members from our Nobel population and an even distribution of ages across our "middle" four employment periods.

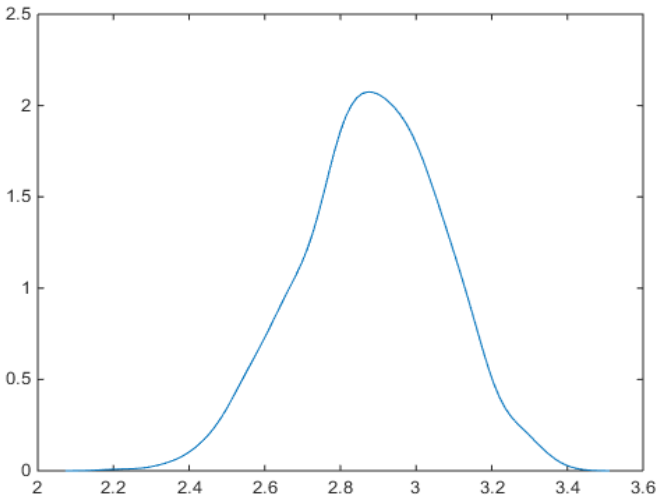

Figure 1: Probability density of departmental scores for Nobel Prize simulations

[5] All simulations are conducted in Matlab. The programming and data sets are available upon request.

Two features are striking. Firstly, the distribution is centred around a rather low average of 2.88. This means that, in the absence of the various adjustments that the actual REF allows for special circumstances, the type of metric used leaves plenty of "room at the top". Secondly, the variance is substantial[6] so that we should expect significant variations of assessed performance across departments even if those are all drawn from a population where every member attains the highest standard of economic research excellence. This raises questions about the suitability of REF-type measures to actually *rank* departments based on numerical outcome and allocate funds based on such a ranking since departments which attain the highest possible standard of research excellence can still differ widely due to different publication patterns across both fields and individuals within these fields.[7]

This variance also suggests that the relative ranking as well as the absolute rating of Nobel departments should not be very stable across successive assessments despite the fact that if we take all Nobel prize winners to be equally "good", our population's underlying quality has not changed and, given the way we run the simulations, neither has its overall age distribution or size.[8] We investigate this point and its implications in section 4, below.

3.2. ADJUSTING THE NUMBER OF MEMBERS SUBMITTED

Of course we know that actual departments do not submit all of their faculty members to the REF. We therefore ask how many members our Nobel departments would have to drop in order to reach some specific target score. Figures 2.a. and 2.b. show the distribution of the number of faculty members that need to be dropped to achieve a target of 3.5 and 3.8 respectively. To reach the lower of these two scores, our department must shed between 1 and 23 members, i.e. between 2% and 48% with an average of about 23%. In order to secure the higher score, departments have to exclude between 21% and 71% of their faculty, with an average of about 42%. We can therefore conclude that a characteristic of large Nobel departments is that they can raise their REF scores to levels beyond 3.5 only with quite significant cuts in the personnel submitted. This is important only to the extent that a university "actions" non-submission to a research assessment exercise. Given that our entire population is individuals who eventually won Nobel prizes for the research they conducted during the periods we include in our sample, moving *any* of these individuals away from a research contract would have been an "incorrect" decision in the sense of reducing the probability that Nobel-quality research would have been conducted. Our point here is only potentially cautionary, of course, since most faculty are not in this high end of the distribution of qualities.

[6] The variance increases, clearly, as we restrict department sizes to be smaller. For example, the typical range of values for a department size of 12 is from 1.5 to 4.

[7] To repeat, our underlying assumption is that there are no "good" or "bad" Nobel Prize winners: each laureate has reached a level beyond which differences of measurement should have little policy relevance.

[8] The issue of the stability of relative rankings over time is addressed more precisely in the next section.

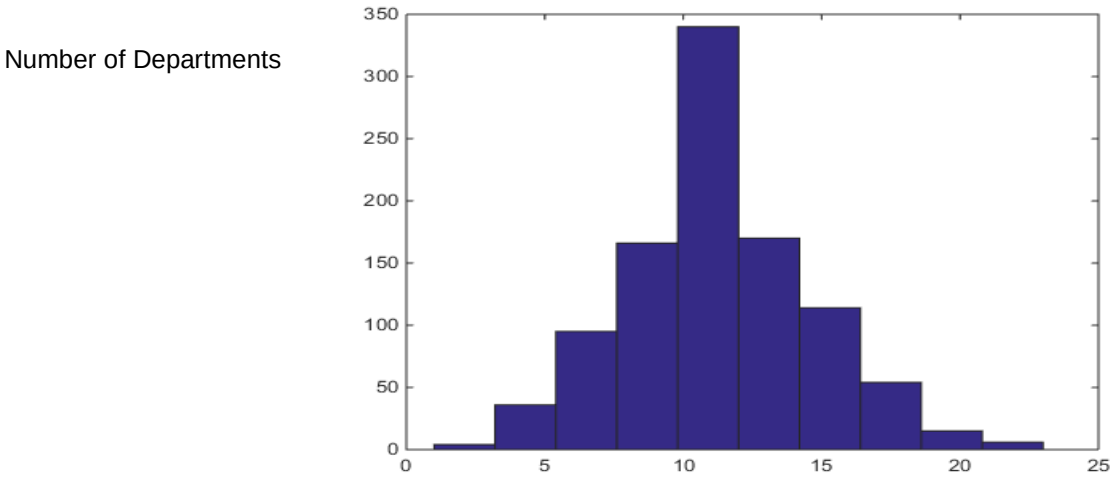

Figure 2.a.: Distribution of omitted Nobel faculty members with 3.5 target assessment level

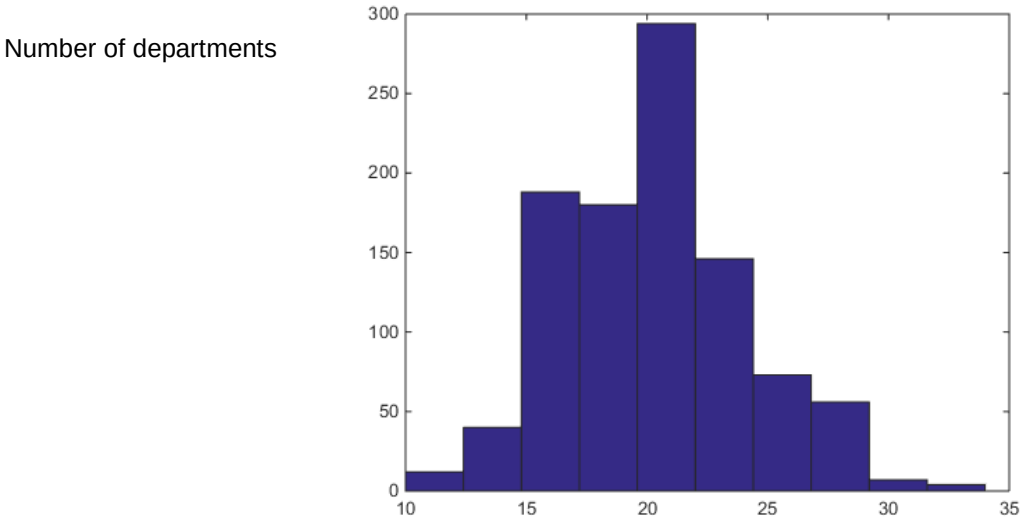

Figure 2.b: Distribution of omitted Nobel faculty members with 3.8 target assessment level

3.3. DEPARTMENTAL SIZE AND AGE PROFILE

We repeat the department generation experiment for different sizes of the department, i.e. N = 36, 24 or 12. There is of course no reason to expect the size of the department to affect the average performance of the department but it should affect the variance of the distribution so that one might expect the performance of smaller departments to be less stable across successive exercises. The question, then, is how much of an increase in the variance of the distribution do we see? The distributions of departmental averages for sizes of 12 and 48 are shown in Figure 3, with the middle department sizes falling predictably between these two extremes.

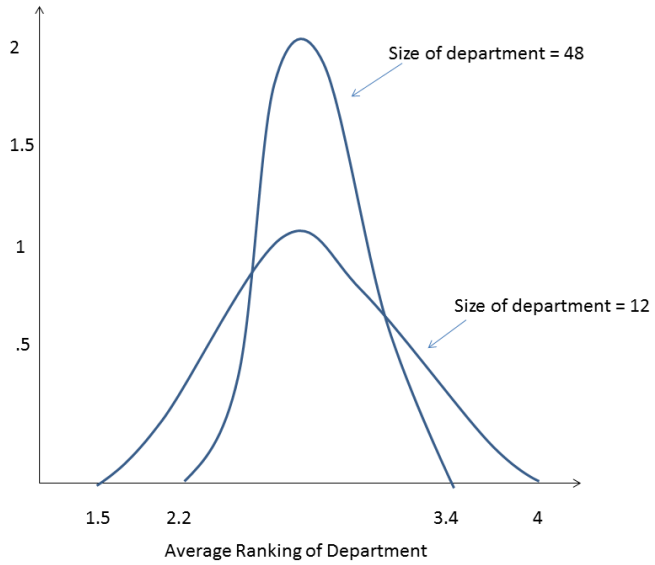

Figure 3: Probability density of departmental scores in Nobel simulations for sizes 12 and 48

The shape of the density function changes quite markedly as we reduce the size of the department so that, for the same underlying quality, we should expect the ranking of small departments to fluctuate substantially more between evaluations. This might have prescriptive content for deans and others, of course, but again if we take our underlying population as uniformly of the highest quality, it does suggest that the reaction to a low rating needs to be measured: our quality does not change across cycles here but the rating can change a lot with each draw, especially for smaller departments. Hence, our simulations based on these random draws from the population suggest that one would be concerned about drastic funding cuts or other strong reactions to even relatively large swings in rating. We will have more to say about the stability of rankings in section 4, where we will look at the change in the performance of *a given* Nobel department which is resubmitted six years later *with the same individual members.* This is, of course, a tighter test for stability in the ranking of a given set of individuals as opposed to draws from a given population.

We also consider the possible impact of the age profile of the department. As explained above, we do not consider the first assessment for which a young faculty member is eligible. We can however control the

distribution of faculty members over the four central six-year periods on which we rely. Table 2 shows the average rating for Nobel laureates only in their first, second, third or fourth "central" periods respectively, with standard deviations in parentheses.

Career Stage	first	second	third	fourth
Average score[9]	2.96 (1.4)	3.06 (1.12)	2.8 (1.29)	2.71 (1.43)

Table 1: Average research evaluation scores for Nobel department in single age range

These averages clearly differ. In particular, the average is higher in the earlier periods, particularly the second (corresponding roughly to the early to mid-40s) and declines slightly over the career.

We do not consider the sensitivity of the department's performance to variations in the fields of specialisation as there are simply not enough Nobel laureates to give us a critical mass in each relevant field. On the other hand, to make sure that our "Nobel" departments' performance is not unduly affected by variations in the professional propensity to publish from the beginning of the earliest Nobel "career" period to the present, we also computed the average from all of the "individual/career stage" observations over two distinct periods: post 1960 and pre-1960. While the first of these stages has a lower average, it also is composed of relatively few observations compared to the total data set (22% of the total).

Periods	Pre-1960	Post 1960	Whole Population
Averages	2.59 (1.36)	2.96 (1.27)	2.88 (1.30)

Table 2: Difference in average research assessment in two eras for Nobel population

A 48 individual department composed only from the post-1960 observations looks similar to the full data set, but shifted slightly to the right[10]. The observations made in this section would still hold qualitatively for such a restricted population.

[9] Standard errors are indicated in parentheses in this table and in other subsequent tables.

[10] If we restrict the data further to more recent periods, the average continues to rise but our sample gets very restricted. We chose a compromise of 1960 partly based on our own reading of the style of publishing, moving away from debate in article-response-comment format and discussions and toward full length articles of a very modern "look".

4. REAL-WORLD BENCHMARK AND CHANGES IN THE EVALUATION PROCESS

There are several reasons why we cannot compare the results that we get for our Nobel departments to the ratings obtained by UK departments in the actual assessment processes. Firstly, as explained above, we do not consider publications at either the very beginning or at the very end of a scholar's career. Secondly, while the successive assessments in the UK allow for "extenuating" circumstances that make it possible to reduce the number of papers to be submitted for a given faculty member, we do not have the required information to make similar adjustments. Thirdly, we assess the four papers per scholar based on Hudson's meta-ranking, while the UK review makes at least some attempt to introduce some independent evaluation of the papers' merit. Finally, each successive generation of UK assessment has also introduced other criteria besides the quality of publications, including "research environment" and "impact".

What we are interested in is both more limited and more precise: how would actual departments compare to our Nobel departments in a system that would strictly be based on an assessment of publication quality that is well approximated by the available journal rankings? To answer this question, we collected publication data for two leading UK departments over the six-year period from 2006 to 2012. The fact that these years do not correspond to an actual assessment period is intentional: we use these departments simply as comparators and do not try to understand how the scores that they received for an actual assessment exercise were generated. The two institutions are UCL and the University of Warwick, mainly for convenience: it was somewhat easier to identify individuals in these universities who were associated with the economics programme than at some other institutions with more diffuse organisations. We collected data only for individuals who were listed as associated with the economics department on the department's website.at these institutions[11].

The logic of this section is that, without questioning the high quality of the two chosen departments, it seems reasonable to assume that a correct assessment of research quality should put them some distance behind our Nobel departments. Whether we find such a relative ranking as well as the magnitude of the distance between Nobel departments and our two leading UK departments can therefore give us some indication of the validity of an assessment based on the quality of four publications per submitted member.

The following gives information about the number of faculty members considered in each of the two departments, and the age structure of this group within the "middle" years. Similar information is provided for our set of Nobel Prize winners in row three, with a draw size that is comparable to the department sizes for the two comparison departments. The fourth row gives the results when we restrict attention to entries for the Nobel group post-1960.

[11] Clearly we have no contract information, so we count all individuals as "fully" within the department. This over-counts any individual on a partial contract.

Department	Number of individuals in middle periods	Composition of individuals in four "middle" year periods	Average score (standard deviation)
Warwick	29	1 8 2 6 3 7 4 8	2.9 (.87)
UCL	36	1 14 2 8 3 10 4 4	3.27(.97)
Nobel Data, uniform age distribution, 36 member department	36	1 9 2 9 3 9 4 9	2.88 (.2)
Post 1960 Nobel Data, uniform age distribution, 36 member department	36	1 9 2 9 3 9 4 9	2.99 (.21)

Table 3: Average performance for middle years at actual departments and Nobel department

While the ratings we obtain for the two comparison departments clearly differ from their actual REF ratings, they have the advantage of having been constructed in the same way as our Nobel sample's REF rating. We also note that our comparison departments have relatively similar age distributions to our uniform distribution in our Nobel department. We note that, without any adjustment, the Nobel group does not fare particularly well compared to the two departments, even if we restrict attention to the post-1960 observations. Furthermore, they do not obtain a rating near the top of the REF scale.

The next table gives us the proportion of faculty members that each of the two actual departments – and our simulated Nobel department -- would have to sacrifice to reach scores of 3.5 and 3.8 respectively. We can then see if allowing departments to "drop" members but making the percentage dropped public might

give us a different relative ranking of underlying research quality. While this measure fares somewhat better at improving the Nobel department's relative ranking, it still ranks far from the top.

Department	Number of individuals	Percentage (actual or average) dropped to reach 3.5 average	Percentage (actual or average) dropped to reach 3.8 average
Warwick	29	41	66
UCL	36	11	28
Nobel Data, uniform age distribution, 36 member department	36	24	43
Post 1960 Nobel Data, uniform age distribution, 36 member department	36	20	39

Table 4: percentage dropped to attain target scores in actual and Nobel departments

In a more prescriptive perspective, we can ask whether the relative ranking of Nobel and actual departments would get closer to what we expect if the precise measure used to assess research performance were modified. In this spirit, the following table shows the average rating of UCL, Warwick and Nobel laureates if we increase the number of publication submitted from 4 to 5 and then to 8. We also list the results if we report the total score for all articles receiving a positive rating during the review period, if we count the number of articles receiving a positive rating during the review period, or if we simply average the rating over all publications.

Department	Number of individuals	Average score for 5 top publications	Average score for 8 top publications	Average score for Total of all publications	Average score for Number of publications	Average score over all publications
Warwick	29	2.66 (.99)	2.11 (1.15)	21.45 (17.37)	7.1 (5.7)	3.04 (.52)
UCL	36	3.12 (1.07)	2.62 (1.19)	28.35 (18.92)	9.61 (6.51)	3.01 (.67)
Nobel Data, uniform age distribution, 36 member department	36	2.71 (.22)	2.29 (.22)	27 (4.17)	9.25 (1.45)	2.74 (.15)
Post 1960 Nobel Data, uniform age distribution, 36 member department	36	2.82(.21)	2.4 (.22)	28.4 (4.3)	9.7 (1.5)	2.77(.15)

Table 5: Scores for alternative metrics in actual and Nobel departments

We find that only a sharp increase in the number of publications counted and/or some explicit weight put on the total count of qualifying publications would bring the relative ranking of the higher ranked of our two UK departments and our post 1960 Nobel departments closer to what might be appropriate based on our assumptions about underlying research quality.

Finally, we turn to the issue of the persistence of rankings. As we noted above, for the same population of Nobel individuals, and when pegging the same age distribution and department size, we obtain a wide variance in realised score over our 1000 simulation repetitions. While the underlying population quality remains the same, the identities of the individuals in the 1000 departments differ. We now perform a more targeted exercise to evaluate how the ranking of the same set of individuals might compare across periods.

We shall compare the rankings of our Nobel departments to the *actual* rankings obtained by the UK assessment exercises. The table below lists the top 20 ranked departments in each of the assessment exercises conducted in the UK in 2001 and 2008, based on the Times Higher Education table of aggregate marks and listed in the order of their 2008 ranking:

Department	2008 research assessment exercise score	2001 research assessment exercise score
LSE	3.55	5*
UCL	3.5	5*
Essex	3.35	5*
Oxford	3.35	5
Warwick	3.35	5*
Bristol	3.15	4
Nottingham	3.15	5
Queen Mary	3.15	5
Cambridge	3.05	5
Manchester	3.05	n/a
Glasgow	3.00	4
Royal Holloway	3.00	4
Southampton	3.00	5
Edinburgh	2.95	4
Exeter	2.95	5
Leicester	2.90	5
Kent	2.90	4
Birkbeck	2.85	5
Aberdeen	2.85	3a
Sheffield	2.80	3a

Table 6: Rankings of actual UK economics departments in 2001 and 2008 research assessments

While the systems of marking departments changed from a rather discrete system to a more continuous system so that the rankings are not completely comparable, the ranking still looks relatively stable[12]. To get a benchmark to which this degree of stability can be compared, we also generated twenty Nobel departments, determined their ratings – and hence their ranking -- and then performed the same calculations on the same twenty departments aged one period.

Clearly, some members "retire" between assessment exercises if we do this, so that we must replace the retired members with new members if we wish to maintain the same department size. If we replace the retired members with randomly selected juniors, so that the age distribution also remains the same across the original and the "aged" department, we can illustrate the change in ranking from one period to the next for our population for 20 simulated departments. This is listed in the table, below. The rank correlation between the two departments' average ranking is approximately .28, a rather modest score[13] and considerably lower than the correlations among actual departments across two periods.

Original population, 48 member department, score and ranking		Aged population, 48 member department, score and ranking	
3.21	1	3.13	3
3.05	4	3.09	4
2.52	18	3.3	1
2.72	15	2.93	12
2.48	20	2.84	14
2.98	7	3.25	2
2.91	11	3.06	6
2.86	12	3.07	5
2.97	8	2.98	10
2.91	10	3.06	7
2.93	9	2.94	11

[12] While a Spearman rank correlation can be computed for these two periods, the discrete nature of the 2001 exercise compared to the 2008 exercise reduced our confidence in how we should rank the 2001 departments: in other words, within the 5* group, it was unclear to us which should be listed first and which fourth. As a result, we computed the correlations in the most favourable way (with rankings within the 2001 categories remaining stable) and the least favourable way (with those 2001 rankings reversed within categories). The range of values was a low of .49 and a high of .86.

[13] A t-test on this score suggests that it is different from zero at 10% significance level.

2.56	17	2.58	19
2.72	16	2.71	18
3.04	6	3.02	8
2.84	13	2.87	13
3.05	5	2.53	20
3.07	3	3.01	9
2.5	19	2.81	16
2.82	14	2.82	15
3.13	2	2.78	17

Table 7: Variation in same Nobel department's score and ranking over 20 repetitions[14]

The fact that actual departmental rankings remain much more stable over time than the ranking of our Nobel departments is quite striking. This can be explained by a number of factors. Firstly, it might be that – contrary to our Nobel departments which are drawn from the same population – the underlying differences in quality between UK departments in the top 20 is quite significant. While this is certainly a factor, the closeness of the actual scores obtained suggests that this cannot be the full answer. Secondly, it is also possible that the other factors used in actual assessments, e.g. impact, are much more stable over time than publication quality so that actual rankings would show more inertia than rankings based only on the quality of publication. Still, these other factors have not received heavy weight in past exercises, so this might not be very likely. Thirdly, previous rankings themselves might have some form of *causal* effects on future rankings. At least three such "endogenous" mechanisms come to mind: the additional resources coming with a higher ranking have a self-reinforcing effects, it is easier to attract high quality faculty members to a department which already enjoys a high ranking, and departments (or deans) might be mostly concerned about "not losing their spot".

The degree of instability observed in the ranking of our Nobel departments should in principle foster a more relaxed attitude with respect to actual rankings since they indicate that a department can easily lose a few places, even in a world where, by definition, the department cannot possibly have done anything "wrong" between successive assessments. The degree of stability in the ranking of actual departments might generate queries about whether they generate considerable inertia.

[14] We obtain the same qualitative result if we target a score and rank departments by the percentage of staff who must be dropped to obtain this target.

5. DISCUSSION

The main limitation of this note is that we have only considered assessments based on the number and quality of publications, ignoring other metrics – such as impact, grants and citation measures – which are also used (often together) in a number of OECD countries. There are two main reasons for this choice. Firstly, so far at least, the type of measure that we consider has been the main element of the assessment processes conducted in the UK and many other countries. The second reason is that applying our "Nobel" methodology to other metrics raises specific difficulties. The obvious problem with impact is not only that it is rather ill-defined but that we cannot possibly go back and meaningfully assess the impact of publications for our whole set of Nobel laureates: comparisons would be very difficult across time. Using grants would raise similar problems as information dating back many years is unlikely to be forthcoming. Moreover, the mode of financing of research has changed appreciably over time and differs significantly across the countries where Nobel laureates have worked. Using Nobel laureates as a benchmark for some form of citation metric is more promising. It would, however, require some adjustment for the increase in "citations opportunities" as the size of the profession and publication outlets has increased over time.

Other limitations concern the results that we have chosen to present in the note. Although these results might sometimes seem rather specific, the reader should keep in mind that we have conducted a large number of robustness exercises. For example, while the note focusses mainly on departments drawn from the middle four periods of a scholar's life, we have run extensive simulations that also include the first period, the last one or both. We also have conducted simulations for different "epochs", many different age profiles, and many different department sizes. Our data is of course available to any reader who would want to run any other type of simulation.

6. CONCLUSION

We have used simulations based on a dataset composed purely of Nobel Prize winners to make a series of points about research assessment, using a mechanistic approach of evaluating research output using journal rankings. As a comparison, we subject some actual departments to the same evaluation procedure we use for the Nobel population. Based on our results, we can make several observations.

First, taking the Nobel population as the desiderata of high quality, we note that the Nobel population does not stand out compared to other top departments if we use the average ranking of the top four publications in each of our 6 year evaluation periods, suggesting that the mechanistic view of the review process that we adopt does not distinguish the top of the quality distribution terribly well. If other papers had indicated greater confidence in qualitative evaluation by peer panels, we would not be concerned with our finding. Unfortunately, we do not have good evidence that peer review panels tend to improve the accuracy of the mechanistic portion of research evaluation exercises.

Second, we note that the same underlying population generates a wide set of rankings in our simulations due to variance across the careers of the individuals in our data set. This variance in average ranking increases as the size of the department falls, with a support ranging from a low of about 1.5 to 4 for a 12 member department to 2.2 to 3.4 for a larger department. When we conduct a more specific experiment of "aging" a department of constant size, we find that the rankings of those aged departments changes

and changes more than the rankings we have observed in the last few actual REF exercises in the UK, suggesting a significant impact of "targeting" an REF ranking by actual departments. Indeed, this is perhaps the finding that we find most intriguing: that the rankings generate inertia in themselves as universities target specific positions in the ranking and attract staff with particular qualities of publication "runs", and who also target rankings of their employers. This generates a pattern of hiring behaviour in the pre-assessment period that we casually observe.

Third, we note that other measures of output quality and quantity, such as including more publications or taking the total assessment of all publications in the period only seem to improve on taking the top four publications if the scope of the metric is extended drastically. Of course, as significantly increasing the number of publications considered might also remove a number of taxing administrative measures involved with selecting publications, such an expansion of the metric might be worth considering. We do not investigate citations, due to the difficulty in comparing citations propensity across time for our Nobel group. This dimension has been treated elsewhere, however, as we mentioned in the introduction.

In sum, the "top four" metric of research assessment does not seem to distinguish the top of the profession very well, using the Nobel Prize as an accurate measure of "the top". An increase in the number of publications considered per individual could improve matters. We also show that even departments composed of only the best possible researchers would show significant "natural" variation overtime, suggesting that targeting behaviour is affecting the rankings significantly.

7. REFERENCES

Abramo, G, D'Angelo, C., and Caprasecca, A. (2009) "Allocative efficiency in public research funding: Can bibliometrics help?" *Research Policy*, vol 38, 206-215.

Clerides, S., Pashardes, P., and Polycarpou, A. (2011) "Peer review vs. metric-based assessment: testing for bias in the RAE ratings of UK Economics Departments", *Economica,* vol 78, pp. 565-583.

European Commission Expert Group on Assessment of University-Based Research (2008), *Assessing Europe's University-Based Research.* Eur 24187 EN.

Frey, B., and Osterloh, M. (2011) "Ranking Games*", University of Zurich Department of Economics Working Paper 39.*

Frey and Rost (2010) "Do Rankings Reflect Research Quality?" *Journal of Applied Economics,* XIII(1), 1-38.

Gans, J. and Shepherd, G. (1994) "How are the mighty fallen: Rejected classic articles by leading economists" *Journal of Economic Perspectives,* vol 8, pp. 165-179.

Hudson, J. (2013) "Ranking journals", *The Economic Journal,* vol. 123, pp. F202-F222.

Laband, D. (2013) "On the use and abuse of economics journal rankings", *The Economic Journal*, vol 123, pp. F223-F254.

Sgroi, D. and Oswald, A. (2013) "How should peer-review panels behave?" *The Economic Journal*, vol. 123, pp. F255-F278.

Starbuck, W. (2005) "How much better are the most prestigious journals? The statistics of academic publication", *Organization Science*, vo. 16, pp. 180-200.

Starbuck, W. (2006) *The production of knowledge: The challenge of social science research*. Oxford: Oxford University Press.

Times Higher Education Table of Excellence in RAE 2008: The Results. Available at: *http://www.timeshighereducation.co.uk/404786.article.*