INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

When the Brightest are not the Best

Marco Valente °

° University of L'Aquila, Italy, and Institue of Economics, Scuola Superiore Sant'Anna, Pisa, Italy

2015/13 **April 2015**

ISSN (online) 2284-0400

When the Brightest are not the Best

Marco Valente
Università dell'Aquila and LEM, Pisa
marco.valente@univaq.it

Abstract

Selection procedures for new recruits in research organizations, supposedly aiming at identifying the candidates with the highest potential, relies necessarily on indirect information concerning the quality of a researcher. It is safe to assume that this information is correlated to, but not coinciding with, the un-observable future contributions of the candidates to a position. We show that using exceedingly selective criteria operating on observable proxy indicators of research quality may hinder the overall goal to ensure the highest expected research quality in the hiring organization. The paper presents a simple abstract model showing how pursuing the absolute best is a strategy very likely to produce results worse than alternative approaches, humbly aiming at identifying the good.

Keywords: Simulation models, Research assessment, Management of academic institutions.

JEL-classification: A14, H10, C63

1 Introduction

When the demand for resources exceeds the available supply it is necessary to ration some of the potential receivers. Among research institutions asking for funds the selection process is likely to include, at least to some degree, the assessment of the quality of past research as a proxy for the quality promised by alternative uses of the scarce resources. For this reason, along the increasing diffusion of assessment-based systems to distribute research funds there is a growing attention on the different methods to assess the quality of research and on their overall effects on the institutions involved (Geuna and Martin, 2003; Hicks, 2012).

In this literature, however, there is the tendency to focus mostly on the theoretical and empirical properties of indicators of the quality of research, giving far less attention to the uses of these indicators, that is, to how exactly the decisions on how the funds should be spent. That is, it seems that the main preoccupation of the literature has been to find a reliable way to identify, in the context of evaluating people, the *brightest* researchers. This work, on the contrary, will focus exclusively on the procedural issue using those indicators for the typical purpose of hiring new recruits to become members of a research group, such a university department. Designing a selection procedure to hire new researchers the goal is to identify the *best* candidate, defined as those most likely to produce high-level research during their tenure. As we will show in an idealized experiment, there are highly likely conditions under which seeking to identify the brightest (i.e. candidates with the best credentials) runs in opposite direction to find the best recruits, those most likely to provide high quality research. Before moving to clarify under which conditions this apparently contradictory result may emerge, and how we support this claim, we discuss briefly the central assumption of our contribution: the indicators available to measure the quality of a scientific contribution are necessarily a noisy approximation of their actual "true" quality, which is an un-observable quality.

The topic of assessing the quality of research is becoming a hotly debated issue, and here we limit to only sketchily summarize the most prominent positions. To assess the quality of research one has to rely on published contributions (journal articles, books, conference proceedings, etc.). There are essentially two broad classes of assessment methods: bibliometric indicators or peer review evaluations. The latter approach has the advantage of exploiting experts' judgment, but is subject to the accusation of discrimination and its costs limit the applicability. The former can exploit the increasingly diffused ICT platforms to perform ever more sophisticated analyses (Glänzel, 2010) on essentially unlimited number of cases, but it is also heavily criticised. Automatic indicators needs to rely on questionable assumptions, such that a high number of citations means higher quality papers, and subjective, hence arbitrary, choices such as journals rankings, undermining the credibility of their results.

A recent contribution performed an interesting comparison between the two methods (Bertocchi *et al.*, 2015) using the unique opportunity offered by a dataset containing data on the same subjects produced with both methods. The authors conclude that the two approaches provide essentially compatible results, implicitly supporting the use of bibliometric indicators, though some observers (without access to the raw data), contested heavily the result (Baccini, 2014).

Independently from ones' opinion concerning the relative advantages and weaknesses of the indicator used, the debate reflects a common understanding that systematic assessment strongly affects the type of activities carried on within research institutions, and not necessarily for the better (Martin, 2011). The reason is that any indicator risk to bias

the results favoring, for example, the quantity of research output rather than the quality of their content, or pushing researchers to choose less interesting areas only because they are more likely to be appreciated by reviewers. In short, any assessment of a piece of research should be considered as a noisy approximation of the true quality of the product.

These considerations produces consequences not only within the debate concerning the design of quality indicators for research output or people, but also concerning the design of the procedures for allocating resources based on those indicators. Let's consider, for example, the hypothetical case of a research institution, such as a university department, considering the hiring of a new member[1].

One problem with using the assessment as guidance to hire new staff is that the very assessment method risks to bias the recruiting towards the research preferences of the evaluators, hence the recommendation to simply ignore the evaluations at greater risk of personal judgment (Gillies, 2014). To be noted that even adopting "preference neutral" bibliometric indicators it is possible that highly regarded journals accepts more easily papers concerning specific research areas and/or methodological approaches, so that we find the problem of assessment biases even avoiding peer reviews. However, this is not the only problem, as we will show in the remaining of the paper.

Let's consider an hypothetical (and, in some respect, ideal) world in which the quality of research is the only variable relevant for the production of the members of a department, and that there is no questioning about the interpretation of quality, other than the exact value of the quality of each candidate for recruitment in the department cannot be known with certainty. On its stead, recruiters are provided with an estimation of the un-observable quality, an indicator correlated to, but not coinciding with, the true quality, so that the distance between the available indicator and the hidden quality is a random error known to be symmetrically distributed around zero.

The question we discuss is whether, in these ideal conditions, it is comparatively better to adopt selection criteria highly competitive or, conversely, it is more advantageous to adopt looser hiring protocols. In the first case, it is more likely that the department hires candidates with the highest scores, as measured by the indicators proxying the true quality. In the latter even candidate with apparent worse indicators have non-negligible chances of being picked. If the indicators reflected perfectly the true quality of candidates than the highly competitive mechanism would, on average, perform clearly better than more tolerant approaches. When the error is, instead, strictly positive, is it still the case that highly competitive selection mechanisms are worth the effort?

The question we pose is relevant because running a highly competitive hiring scheme is costly, requiring candidates to produce a lot of evidence and the selection commissioners to carefully weight all the evidence. Moreover, knowing that the probability of being hired strictly depends on your score constitutes an objective incentive to make any effort to improve the appearance of high research quality, independently from the substance. If the advantage provided by high competition levels is small, or even non-existent (as we will prove below) even in our idealistic virtual world, then we have a clear case suggesting to explore alternative hiring procedures not relying on a simple indicator, however reliable it appear to be.

The next section describes informally an abstract model build on purpose to answer the question we posed above. We then provide a formal implementation of the proposed model, implemented in terms of a simple agent-based model. Section 4 presents and

[1] Evaluating an individual researcher is only one possible application of the assessment of research quality. Besides the importance of these events, it seems also a fundamental building block for other assessment goals, such as funding distribution across institutions or individuals (Ioannidis, 2011).

discusses the results provided by simulating the model, showing the conditions required for the high and low competitive hiring system to perform better. Before concluding, we sketch briefly an alternative hiring mechanism not relying on a specific estimation of a candidate's quality.

2 Informal model description

The model, described formally in the next section, is a stylized representation of a generic organization, such as a department of a research institution, regularly requiring new recruits to replace retiring members. The model is designed to test the outcome produced by different selection procedures under different external conditions, represented by the nature of the candidates for the positions. The model is meant to highlight a generic properties of a hiring method, and therefore ignores as many details as possible in order to make evident the relevant consequences for a hypothetical decision maker.

We assume that there exist a correct measure of quality of a researcher[2] that, however, is not directly observable by the modeled decision makers. It is instead possible to collect data (such as, e.g., the publications' record, education, references, etc.) collectively providing an indicator supposedly approximating the (hidden) quality. One of the control parameters of the model is the gap, assumed stochastic, between the observable indicator available to the simulated agents and the un-observable true quality, that we control as modelers.

The model represents an ideal "department" that needs to hire new staff in order to replace its retiring members. The new recruit can be chosen only on the base of the available indicators, and the goal is to maximise the average quality of the department. For simplicity, to highlight the impact of the hiring practices on the quality of an organization, we assume the quality of a hired researcher to remain constant throughout the tenure of selected researchers within the organization, whose length is assumed for simplicity constant and identical for all members. Allowing for the endogenous dynamics of the skills of researchers would complicate the interpretation of the results and therefore, at this stage, we prefer to keep this option as a possible future extension.

The model assumes that the hiring method consists in choosing randomly a single new recruit from a set of candidates with probabilities proportional to their observable indicators, so that a "better" candidate is always more likely of being hired than a competitor showing a poorer indicator. The use of a stochastic choice, rather than a deterministic one, reflects the fact that small differences in the indicators' values may have little relevance, similarly to what have been shown to the distributional properties of the Hirsch's *h-index* (Baccini *et al.*, 2012). The same assumption may be supported that, in reality, other considerations enter in the selection procedures besides the research quality indicator. For example, the selecting committee may additionally consider the specific area of specialization of candidates, logistic considerations, reliability, teaching qualities, etc., so that when the indicator of research quality is very similar among two short-listed quality the probability of being selected is roughly similar, independently from small differences. For modeling purposes we therefore assume that good credentials on research quality provide an advantage for being hired, but that the actual choice is probabilistic. We study a range of different practices differing by the degree of *selectivity* on (observable) research

[2]We personally do not believe that it is possible to represent all the necessarily multidimensional and non-measurable aspects of a researcher by a single value, or even a whole set of values. This assumption is adopted on a *a fortiori* basis, and the results presented would be even more robust admitting the non-measurable nature of research quality.

skills, i.e. the relative concentration of the probability distribution assigned to the candidates. A more selective practice assigns far higher probabilities to the (apparently) best candidates, while a more tolerant practice is represented by smaller probability differences, giving relatively more chances to less qualified candidates.

As last element of the model we implement a sort of "personal orientation" of researchers that determine whether they have a tendency towards improving the true quality of their research or towards improving their indicator, where the two orientations are mutually exclusive: pursuing one necessarily deteriorates the other. The implementation of this characteristic is designed so as to verify its relevance to the eventual results.

We allow researchers to "invest" a given stock of available effort pursuing the two alternative goals. On the one hand they may aim at improving their true research quality, for example pursuing ambitious and risky research projects, or daring to explore radically novel areas of research. In these cases the quality of researchers improves by a random amount in respect of the "original" endowment of quality, but their proxy indicators are reduced by the same amount, under the assumption that pursuing risky projects or breaking new grounds have, at least initially, a negative impact on the observable indicators, tuned to capture established research careers. Alternatively, researchers may chase improvements in their indicators, such as submitting papers to prestigious journals with minor revisions of known results. In these cases, the efforts aimed at improving the public measures of research quality will somewhat damage the actual value of quality. This may be due to two, independent, reasons. Firstly, we may imagine that there exist a sort of diminishing returns on research within a given subject using the same approach. Secondly, researchers tailoring their research to fulfill requirements diverse from pure scientific curiosity may be expected to comparatively reduce their research skills, specializing in reinforcing past results but hardly likely to break new grounds.

The model is implemented as a simulation agent-based model using extensively independent random values, so that the average values collected over many time steps ensures a reliable appreciation of the expected results, and of their underlining motivations, presented in the section on results.

3 Formal model description

The model contains a group of $N = 100$ "researchers" composing a department, whose generic member is represented by four values:

- $q_i \in [0, 1]$: research quality, unobservable by the agents within the model and whose average over all the members of the department is the main result, measuring the department true quality; the goal of the selection mechanism is ensure the highest value of these variables for all the members of the department.
- $i_i \in [0, 1]$: proxy indicator of research quality, observable by the model agents, correlated to the quality q_i as described below.
- $o_i \in [-1, 1]$: preferential orientation ranging from -1 (maximal effort to maximise the public indicator) to 1 (maximal effort to maximise the true research quality).
- $Age_{i,t} \in \{0, 1, 2, ..., T\}$: age of the researcher, starting from 0 for newly hired members and reaching T, when the researcher retires triggering the department to open a new position.

Apart the age, these values are fixed when the researcher is hired and are not modified during their life in the department, assumed to last 100 time steps. At each time step the simulation replaces the retiring members (those reaching $Age_{i,t} = 100$) launching as many independent "calls", each producing eventually one new recruit starting its membership with $Age_{i,t} = 0$[3]. Each call is answered by exactly 100 candidates whose values are determined as follows, indicating with the symbol $*$ the variables for candidates.

The first variable determined for a candidate is the true quality of research, q_i^*, drawn from a random value distributed according to a power law. This distribution reflects the evidence that research skills are distributed in an asymmetric way, with few excellent candidate increasing proportions of candidates for decreasing levels of quality. The function adopted produces values for quality levels q distributed according to the distribution $Prob(q = x) = e^{-\alpha \times x}$, with $x \in [0, 1]$. The higher α the more concentrated the distribution, i.e. smaller the share of top quality values, producing a highly skewed distribution. Conversely lower values for α produce a more even distribution. A value $\alpha = 0$, at the extreme, would produce a uniformly distributed random distribution spanning evenly over the range$[0 : 1]$.

The raw proxy indicator i_i^* for the candidates is derived from their true quality with a stochastic choice determined by parameter δ. The procedure draws a random value with uniform distribution in the range around the true quality: $i_i^* = U(q_i^* - \delta/2, q_i^* + \delta/2)$. In case the extreme of the range extends beyond the permitted range [0,1] this is shifted to ensure that the resulting value is always within unitary interval. That is, if $q_i^* < \delta/2$, then $i_i^* = U(0, \delta/2)$. Symmetrically, for $q_i^* > 1 - \delta/2$ we use $i_i^* = U(1 - \delta/2, 1)$.

The orientation of the candidate is obtained using a uniformly distributed random draw in the range indicated, that is $o_i^* = U(-1, 1)$. After the orientation is determined, the values previously drawn for the quality of research and the proxy indicator are modified as follows: $q_i^* = q_i^* + o_i^* \times \gamma$ and $i_i^* = i_i^* - o_i^* \times \gamma$. Both q_i^* and i_i^* are replaced with the closest boundary if they exceed the range [0,1].

A time step in the simulation run consists in creating $N = 100$ candidates according to the rules described above. Then the simulation counts the number of current members of the department reaching retirement age, and replace them with the same number of candidates chosen according to the probability $p_i = \frac{i_i^{*\sigma}}{\sum_{j=1}^{N} i_j^{*\sigma}}$. The parameter σ represents the intensity of the selection process. Higher values of σ represent higher differences in probability, while lower levels indicate less marked differences. In case at the same time step more than one member needs to be replaced, the system repeats the a fresh draw as many times as necessary, each time using the same probability distribution.

We run a simulation exercise for 10,000 time steps, and collect the average values of the relevant variables across all these steps. The large number of time steps ensures that the stochastic volatility introduced by the large number of random event is absorbed, eventually providing stable results.

In summary, the model is controlled by the following parameters, whose indicated values will be used for the simulation results presented in the next section:

- $N = 100$, number of, both, members of the department and potential candidates for hiring at any given time step.
- $T = 100$, retirement age for members of the department, after which the researcher is replaced by a new one chosen as indicated among the candidates. Initial age values

[3]Note that each open position draws a fresh call, so that the distributional properties affecting each new recruit are independent from the number of calls at each period.

are chosen randomly, uniformly distributed between 0 and T, to ensure a roughly regular departmental turn over rate.

- $\alpha = 20$, the degree of skewness of the distribution of true quality.
- δ, range of error for the indicators i_i around the true quality, before adjusting for the effects of the orientation of candidates. We explore the case for a maximum error of 10% of the possible range of values between true quality and proxy indicator, setting $\delta = 20\%$.
- γ, maximum change induced by the orientation of the researcher in respect of research quality and proxy indicator. The simulation runs are repeated for 8 different values of this parameter, ranging from 0 to 0.7. As mentioned, this maximum range is obtained when the orientation takes the extreme values (1 or -1).
- σ, intensity of selectivity. Higher values represent a higher concentration of probability in favor of the candidates with the higher i_i, while lower values represent selection mechanisms still proportional to the indicators, but more tolerant for less than stellar values. We consider 10 values from 1 to 10.

4 Simulation results

The simulation described in this section consists in 10,000 time steps during which retiring members of the department are replaced with candidates chosen in the set of candidates. To minimize the possibility of distortions due to rare random choice of values, the whole set of candidates is fully redrawn at each time step, therefore smoothing away any volatility due to possible extreme values.

The results presented consist in three statistics computed as averages over the researchers hired in the simulated department that, in turn, are again averaged over all the 10,000 time steps. The values collected are: average true quality Q; average proxy indicators I; average orientation value O. Notice that we have imposed the random distributions of these variables only among candidates. The observed values are instead the averages of the *selected* candidates, and therefore the distortions between the mean values from the original distributions and the averages computed in the simulations depend solely on the selection procedure. To understand how relevant is the selectivity under different conditions concerning the opportunity of candidates to orient their research attitude, we replicate a simulation run for each combination of the values for the parameters σ (10 values expressing different levels selectivity from 1 to 10), and γ (8 values from 0.0 to 0.7).

We present the results under the setting $\delta = 0.2$, meaning that the maximal difference the proxy indicator and the true quality of a candidate is 10%, before distortions due to the effects on quality and indicator due to personal orientation. Figure 1 reports the average indicator values across all time steps of the researchers hired in the department. The figure clearly shows that increasing the selectivity, giving higher probability to researchers with higher proxy indicators, does indeed increase the average level of this indicator, reflecting the fact that the selectivity works as expected increasing the average value of the variable considered more relevant in the selection of candidates.

The series marked with the value 0, referring the cases in which $\gamma = 0$, can be considered as a sort of benchmark, since the orientation chosen by researchers, in this case, has no effect. All other cases are ordered along increasing values of γ, indicating that the stronger the effect of orientation the higher is the average indicator value. This is obvious since

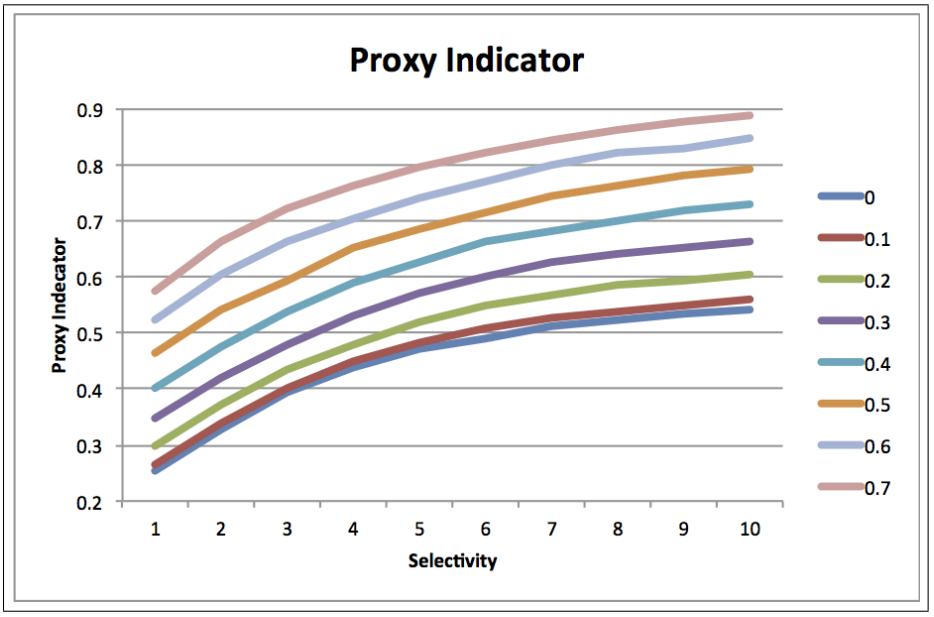

Figure 1: Average proxy indicator I for the quality of researchers for different levels of selectivity levels σ. The different series refer to different values of parameter γ indicating the maximum effect of orientation.

when the effect of orientation is stronger (higher γ's) the higher will be the values i_i^* of the indicators for the candidates, and therefore the selection procedure will have larger pool of (apparently) high quality candidates to choose from.

Figure 2 shows that the apparent quality of the department as reported by the proxy indicators paints a far rosier picture than the actual (un-observable) quality of the department as reported by the true quality. This is shown by observing that the quality levels for all cases are sensibly lower than those reported by the indicators. If the scaling was the only effect, than it still would not matter in terms of the choice of selectivity level (and of a possible incentive policy aiming at influencing personal orientation). But this is not the case.

The ordering of the series for the true qualities is the reverse of the one computed over the indicator. The benchmark case (no effect of orientation) leads the group, while the results produced with the stronger effect of orientation is, by far, the worst. This result is easily explained by the fact that promoting researchers with the best (public) score favors those pushing harder to improve their visible standing, even at the cost of damaging their actual research capacity. Remember that the model is built to study the selection process, not behavioural ones. It means that all cases in the same series (same maximum effect of orientation) you have the same distribution of candidates, and therefore the differences depend only on the rigidity of the selection.

Judging from the true quality, the importance of the selection pressure is actually much reduced, increasing in general quality less than the increment in proxy indicators it produces. Actually, in several cases the average quality provided by even high levels of selectivity remains *below* the expected values from the power law distribution of qualities (about 0.142). This means that a selection committee would do better by picking candidates randomly without any criterion at all, rather than looking at the proxy indicator.

Possibly worse of all from the perspective of a designer of hiring procedures, the in-

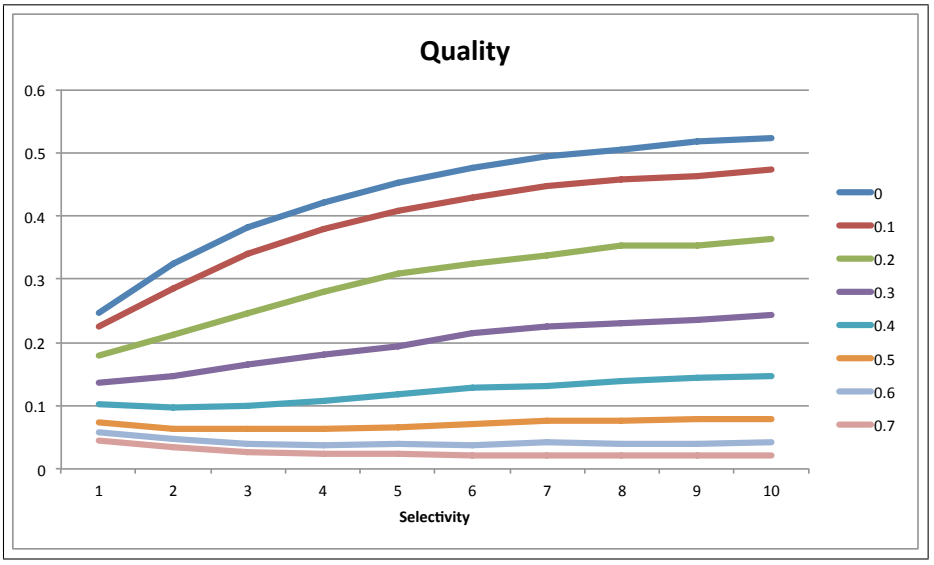

Figure 2: Average (true) quality of researchers for different levels of selectivity levels σ. The series refer to different values of parameter γ indicating the maximum effect of orientation.

crement of selectivity is, in some cases, *negatively* related to the quality. Meaning that adopting more stringent selective procedures provides worse results than using a softer intensity of selectivity.

Figure 3 concludes the results showing the average "orientation" of the researchers in the department. We see that the benchmark case shows a null average orientation at any level of selectivity, as can be expected since, when orientation has no effect, we cannot but obtain the expected value of a uniformly distributed random variable in the [-1,1] range, that is 0. For all other cases the results consistently show a strong average negative orientation, meaning that, on average, researchers hired by the department are biased towards improving their public indicators with the consequence of worsening their true research quality. This negative result is accentuated by increasing selectivity pressure, as shown by negative slope of the series. This result means that selectivity, even when increasing moderately the average quality of hired researchers, does so at the cost of selecting those with the stronger orientation towards focusing on the appearance, rather than substance, of their research.

We can conclude that a recruiting procedure based on selection operated on an indicator of quality systematically overestimates performance expected from candidates recruited, and frequently fails even to exceed the average performance provided by random choice of candidates. The severity of the selection, represented by the differential in hiring probabilities for candidates with different indicators, is shown to be either poorly and even negatively correlated with the overall average quality for the department, suggesting that hiring criteria should be designed with care to avoid wasting resources (obtaining and elaborating information is costly) to obtain counter-productive results.

The next section discusses the features of possible hiring procedures producing higher expected recruits than those based on indicators only.

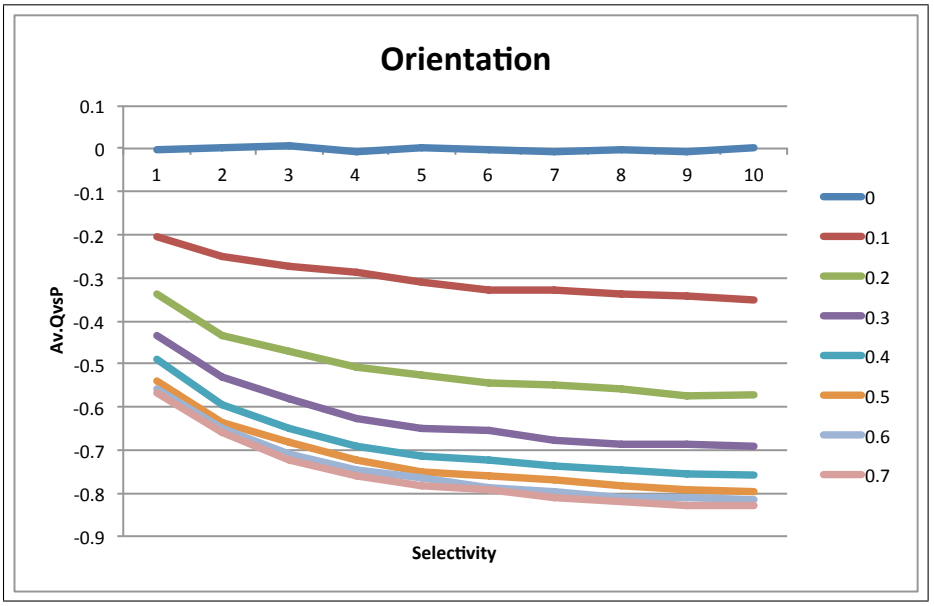

Figure 3: Average orientation between increasing efforts on research (values close to 1) or on improving the proxy indicators (values close to -1). Values produced for different levels of selectivity levels σ. The series refer to different values of parameter γ indicating the maximum effect of orientation.

5 The best are the best

The benchmark result of a hiring procedure is the quality provided by picking randomly any candidate irrespective of any information, since any procedure providing worse results may be fruitfully replaced by the mere blind choice of candidates from an urn. The previous section shows a negative result, that is, that relying on a proxy indicator, even when correlated with the true quality, is frequently no guarantee of success in beating the benchmark, leading to worse-than-random overall performance. In this section we sketch a positive proposal, suggesting how an alternative hiring procedure may ensure that the average quality of recruits is consistently better than purely random choices.

Proxy indicators of quality, such as CV's, lists of publications, and the like, are data immediately accessible, but imprecise and severely subject to distortions. On the contrary *personal* direct information is something requiring an investment in time to accumulate, requiring the evaluator to have an in-depth familiarity with the evaluated, and provides generally far more precise assessment of the true quality. Assuming that candidates have had previously worked with some of the current member of the department, it is possible to devise a hiring procedure relying on this knowledge.

A simple mechanism of co-optation based on personal judgment, though potentially able to exploit the best information possible, such as the direct knowledge of the quality of a researcher, suffers from two problems. Firstly, outsiders lacking internal sponsors are obviously cut-off from the access to a club admitting new members by personal connections only. Secondly, current members may have incentives to recommend candidates for reasons different from sincere admiration of their objective qualities as researchers. The first problem may be solved, at least partially, by trial periods and by requesting recommendation from sources external to the department, such as, for example, asking for recommendation letters. The second problem is more subtle, generating a *tragedy of the*

commons case, where each member would rely on others to bring in the best researchers while personally recommending people who may be privately rewarding to hire (relatives, friends, etc.) irrespective of their quality. The sketch of a solution for this second problem is discussed below.

We start from the assumption that recommending the recruitment of a known poor candidate provides a positive private outcome, rewarding private objective at the cost of diminishing the overall quality of the department. To discourage such behavior it is necessary to introduce a private cost, assuming that the collective one, due to the weakening of the department caused by adding a poor element, is generally not sufficient. This cost may be introduced in many ways using the concept of *responsibility*, that is, that the recommending person is considered somewhat responsible of his or her past recommendation when decisions on new recruits has to be taken. That is, recommending high quality researchers (as revealed with time) increases the chances of being more influential in future hiring decisions, while, on the opposite, "cheating" by making public statements in contrast with private information reduces the influence of one's opinion on future decisions.

There are many possible implementation of such a principle, all of them requiring essentially the publicity of one position, such as a formal statement by members of the department concerning new recruits. Notice that such a principle deals identically for both good-faith errors, as those performed by people sincerely unable to assess the quality of potential candidates, and the bad-faith sort, made by people willfully lying about their private knowledge of candidates. Establishing a link between the influence of the department in choosing new recruits and their past performance as recommending advisors would automatically exploit the private and (time) costly information about candidates removing the risks of poor judgment caused by either private interest or pure inability.

6 Conclusions

Most of the literature on the assessment of research quality focuses on the problems arising from attaching an estimate to either a researcher or on his/her output. However, the inner mechanisms of the actual procedures using those estimations may be as much, if not more, relevant. This paper has discussed how the selection of new staff may be heavily affected by the selectivity intensity adopted in the hiring procedure, providing counter-intuitive results by means of an agent-based model.

This paper discusses the abstract problem of the procedure to adopt by an hypothetical department in order to hire new members. The problem arises because the interest of the department as a whole is to maximize the quality of the its researchers, but this variable is not directly observable.

A common practice is to adopt visible indicators supposedly proxying the quality of a researchers, such as the number of papers written, the journals on which a candidates published, memberships to editorial boards, honors and awards, etc. However, any measure reducing the output of scientific research to one or a few indicators can be expected to approximate the true quality of a researchers with some slack, in that candidate researchers have indicators likely to either over- or under-estimate their actual research capacity.

This paper explores the effects of hiring procedures implemented as competitive selection based on the indicators, measuring their performance in terms of the average true quality of the resulting department. The stronger result we discuss is that, under rather general assumptions, the quality provided by such selection may be pretty poor, even poorer than mere random choice. Moreover, increasing the selection pressure (giving high importance to small differences in indicators) may even lead to *worsen* the performance

in terms of average quality of research.

These negative results show that a theoretically perfect system may produce results opposite to the expectations when introducing an apparently minor distortion, in our case that public information is strongly correlated, but not identical, to the true values. In short, a department is better off by not choosing necessarily the (apparently) brightest people it may find, but needs searching more sophisticated hiring systems, provingly robust against the biases induced by poor information.

The paper suggests that there may be ways to exploit the personal information on candidates by current members, gained with experience and collaboration. To discouraged and neutralized "errors" in recommendations (either in good or bad faith), a hiring system would requires publicity of statements and responsibility for past actions.

This work may be extended in both empirical direction and theoretical one. All the values and distributions adopted in the paper reflect plausible assumptions, but hard data should not be difficult to find. Hence, it may be possible to calibrate a version of the model around data from a specific reality and test the effectiveness of different hiring policies under different assumptions for un-observed data (e.g. distribution of research qualities). On the theoretical approach, it could be possible to explore a more detailed version of the model including more sophisticated behaviors by both commissioners sitting sitting on selection committees.

References

Baccini, A. (2014), "Lo strano caso delle concordanze della VQR", http://www.roars.it/online/lo-strano-caso-delle-concordanze-della-vqr/.
URL: *http://www.roars.it/online/lo-strano-caso-delle-concordanze-della-vqr/*

Baccini, A., Barabesi, L., Marchesi, M. and Pratelli, L. (2012), "Statistical inference on the h-index with an application to top-scientist performance", *Journal of Informetrics*, **6**(4), pp. 721–728.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. and Peracchi, F. (2015), "Bibliometric evaluation vs. informed peer review: Evidence from Italy", *Research Policy*, **44**, pp. 451–466.

Geuna, A. and Martin, B. R. (2003), "University Research Evaluation and Funding: An International Comparison", *Minerva*, **41**(4), pp. 277–304.

Gillies, D. (2014), "Selecting applications for funding: why random choice is better than peer review", *RT. A Journal on Research Policy and Evaluation*, **2**(1).

Glänzel, W. (2010), "On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper", *Journal of Informetrics*, **4**(3), pp. 313–319.

Hicks, D. (2012), "Performance-based university research funding systems", *Research Policy*, **41**(2), pp. 251–262.

Ioannidis, J. (2011), "More time for research: Fund people not projects", *Nature*, **477**, pp. 529–531.

Martin, B. R. (2011), "The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?", *Research Evaluation*, **20**(3), pp. 247–254.