Is post-doc funding stimulating research performance? Evidence from Danish research policy!

Jesper W. Schneider jws@cfa.au.dk / Danish Centre for Studies in Research & Research Policy, Department of Political Science & Government, Aarhus University, Bartholins Allé 7, Aarhus, DK-8000 C (Denmark)

Thed N. van Leeuwen leeuwen@cwts.leidenuniv.nl / Centre for Science and Technology Studies (CWTS), Leiden University, Wassenaarseweg 62a, Leiden, 2333 AL (the Netherlands)

Abstract

We present main results from the bibliometric part of a recent evaluation of two different postdoctoral funding instruments used in Denmark. We scrutinize the results for robustness, sensitivity and importance, and eventually come out questioning the official conclusions inferred from these results. We specifically examine whether there is a long-term citation performance difference between groups of researchers funded by the two instruments. Through an elaborate matching process, potential differences in performance are also compared to a control group of researchers that has not received postdoctoral funding, but otherwise are comparable to the postdoc-groups. Hence, we are also able to indicate the effectiveness of being postdoc-funded, given its benefits, when it comes to citation performance. The results show that all three groups perform well above the database average impact, however, we conclude that there is no difference in citation performance between the two postdoc-groups. There is, however, a difference between the postdoc groups and the control group, but we argue that this difference is trivial and certainly not robust. Our conclusion, contradicts the official conclusion given in the evaluation, where the Research Council emphasizes the success of their funding programs and neglects to mention the rather high performance of the basically tenure-tracked control group. We speculate that reservations against tenure-track positions may lay behind this conclusion? We demonstrate and stress that indicators should come with robustness and sensitivity analyses, as well as some sort of yardstick, in order to make decisions concerning the importance of findings.

Introduction

Postdoctoral research fellowships are most often designed to facilitate the transition of a research career for young researchers. The Council for Independent Research in Denmark (DFF) has for more than a decade supported a large number of junior researchers through postdoctoral (postdoc) funding programs (DFF, 2012). Basically two instruments have been used. An instrument where individual researchers apply for and receive grants to fund their own research (for short "individual postdocs"), and an instrument where junior researchers are funded within larger

research projects, where typically more senior researchers are the principle investigators and the names of the intended postdocs can either be specified in the application, or alternatively, only the amount expected to be used to fund a postdoc position is given (for short "embedded postdocs"). In the latter case, specific junior researchers are selected following successful funding of the research project and their *curriculum vita* therefore plays no role in the funding review process.

Evidence for the effectiveness and benefits of such postdoc-programs remains vague. Attention is most often directed towards questions concerning application and selection, e.g., do the "best" young researchers apply and are the "best" selected for funding, where "best" is typically defined and measured in terms of publication performance (which consists of output and impact analyses). (e.g., Bornmann & Daniel, 2006; Bornmann, Wallon & Ledin, 2008; Böhmer, Hornsbostel & Meuser, 2008; Hornbostel et al., 2009; Bornman, Leydesdorff & van den Besselar, 2010; Neufeld & von Ins, 2011; Neufeld & Hornbostel, 2012). Discriminating between young researchers, often coming straight from a PhD-program, at the time of application based on past publication performance can be challenging due to the possible short active publishing career at that point. Further, if resources are allocated to those young researchers already on their way to a productive career, the marginal impact of the intervention may turn out to be small. In a regression-discontinuity designed case study of National Institutes of Health's (NIH) F32 individual postdoc fellowships, Jacob and Lefgren (2011) find that the F32 program does appear to increase the amount of health science research, as well as the number of individuals engaged in a biomedical research career (i.e., not leaving academia). Jacob and Lefgren (2011) argue that among benefits for those receiving funding, are a higher probability of getting funded by NIH in the future; more time to research and less teaching obligations; easier to establish networks with high performing peers; and/or a general increase of the postdoc's visibility in the domain. Again visibility is defined and measured in terms of publication performance, i.e., publication activity and citation impact.

As mentioned above, in Denmark two funding instruments have been in place for more than a decade, i.e., individual and embedded postdocs. During this period, the DFF has experienced a rising pressure on postdoc applications due to large intake of PhD-students into the Danish research system from 2004 and onwards. A general concern among the Board of Directors in DFF is whether funds are used to best effect. DFF assumes that postdoc-funding in general is beneficial to the individual researcher and his or her future career, however, at the same time DFF suspects that the funding instrument, where individuals are funded directly, is superior compared to the embedded funding instrument, and if so, money is perhaps best spent by exclusive support to the individual postdoc instrument? The rationale is that the selection process in the individual postdoc instrument, *ceteris paribus*, ensures that the "best" applications and applicants are funded, and perhaps more importantly, it is also assumed that being funded by this instrument provides the best future career opportunities, such as visibility, impact, networking and mobility, compared to the postdocs funded indirectly through larger research projects. No evidence is provided for this presumption. Hence, the DFF commissioned an evaluation of the two postdoc funding instruments (DFF, 2012).

This study presents the results of the bibliometric part of this evaluation. Besides the bibliometric analysis of publication performance, the evaluation also constitutes a survey and two register-based analyses of career paths, mobility, internationalization, subsequent funding successes and leadership opportunities.

The aim of this study is to examine and compare the longer-term productivity and impact of researchers funded by these two instruments. Additionally, in order to measure the potential effect of being funded compared to not being funded, a carefully constructed control group is also set up and their productivity and impact is likewise examined and compared to the two postdoc-funding instruments. The research design therefore enables us to answer the research question regarding long-term performance differences between the two instruments, as well as the question regarding effectiveness of being postdoc-funded at all. Notice, while the research design is considered strong due to its strict matching process, we do not claim any causal effects for funding interventions (e.g., Freedman, 2005)

The paper is organized as follows, data and methods, especially the matching process, are presented in the next section; the subsequent section presents some of the most important results; and we end with a brief discussion of the approach, the results and some consequences.

Method and data

DFF comprises five field specific research councils (FSS: health sciences, FNU: natural sciences, FTP: technical sciences, FSS: social sciences, and FKK: arts and humanities). The evaluation examines postdocs funded between 2001 and 2009; in that period 62% of the funded postdocs went directly to individual applicants and the remaining 32% went indirectly to positions embedded in larger research projects. The bibliometric evaluation comprises all research councils, but only the evaluation of FSS, FNU and FTP comprise citation analyses. In this paper we only focus on postdoc-funding from the latter three research councils comprising citation analyses.

The comparisons between the two postdoc instruments and between the instruments and the control group are carried out combined for all three research councils on the aggregated level (e.g., sub-fields aggregated to the overall council). In order to be able to compare the bibliometric performance-effect of the two postdoc instruments, in between them and between them and a comparable group of non-funded researchers, a stringent matching process is set up to deal effectively with covariates. Notice, we have access to all postdoc grants from 2001 to 2009, including grant information, names and demographic data both for individual and embedded postdocs. We also have access to a general population of researchers in Denmark for the same period.

Matching procedures

Matching methods can be considered a stronger approach than simple statistical controls when designing an observational study in the sense of selecting the most appropriate data for reliable estimation of effects when it is impossible to have full control (e.g., randomization) (Rubin, 1997; Rosenbaum, 2002). Notice, most observational studies, especially ones with strict matching procedures, cannot fulfil the required assumptions needed for frequentist inferential statistics (e.g., Gill, 2010), this is also the case in the present study. However, we do not consider this a problem at all (see Schneider, 2013).

The control group was created from the general population of researchers in an exact match (1:1) with postdocs from both groups, where exact matching criteria included research council, active subject field and year of PhD-graduation. Further matching criteria include age, where a 2-year variation in age was allowed between the matched individuals. Finally, in order to produce roughly equal-sized treatment and control groups we randomly selected eligible matched individuals. Initially around 730 researchers were included (i.e., postdocs and control group).

Bibliometric data, measures and analyses

After the matching procedure, an attempt was made to construct publication portfolios for all 730 researchers by searching the Web of Science (WoS) database on name variants, affiliations and the specified publication period. Eventually, some 60 researchers were discarded as no portfolio could be established, either due to failure of identification or dubious identification. Consequently, 670 researchers were eventually selected for the overall analysis, i.e., control group (202), embedded postdoc (228) and individual postdoc (240). The 670 researchers is the basis for the publication analyses in the overall evaluation. In this paper we only focus on the results of the citation analyses and only 632 of the 670 researchers went into the citation analyses. The reduction in the number of authors is primarily caused by the shorter publication window and for a few because they had only published proceedings papers. On the other hand, the reduction in the number of researchers seem to be just about evenly distributed among the examined groups, i.e., control group (195), embedded postdocs (207) and individual postdocs (230), the corresponding drop in publications also do not seem to alter the publication profiles of three groups.

While the number of researchers in each group turned out to be marginally different, the homogeneity between groups on matching parameters and demographic variable remained, so we considered the groups representative and suitable for comparison of publication activity and impact. Notice, the publication portfolios for some of these researchers are most likely incomplete, including false positives and missing publications. Unfortunately, we were not able to contact researchers in order to validate the publication lists. We assume these 'errors' to be equally distributed across all three groups, furthermore since we only produce results on the group level, this type of error on individual level cancels out. We should stress that what we examine is publication performance of the three groups within the constraints of Web of Science.

Bibliometric analyses require robust data. A pilot study indicated that the publication history of most postdocs were meagre around the time they received their grants, many acquiring their grants within one or two years after finishing their PhD. As a consequence, a pre-test of bibliometric performance between the funded postdocs, as well as potential applicants is not feasible in the present case; indeed we conjecture that this is the case in many circumstances. Instead, we establish more robust publication portfolios for each individual researcher and measure the longer-term impact accordingly. Notice, this implies that potential publications not related to the postdoc-funding will be included in the portfolio and as a result analyzed.

Several publication windows for the portfolios were tested, but the general homogeneity between groups, including granting years between 2001 and 2011, meant that relative differences between groups were stable regardless of the length of the publication window. We therefore report results with publication windows spanning from the granting year (estimated year after PhD graduation for the control group) to 2009, i.e., 2009 is last suitable year for citation analysis. We apply a citation window of three years, the publication year plus two consecutive years. A relative short window like this enables us to include publications from 2009. The following publication types are included in the citation analyses: Articles, articles; proceedings paper[i] , letters and reviews. We apply CWTS' version of WoS, where letters are weighted as one fourth of paper. All citation analyses are carried out with field normalizations excluding self-citations. We present productivity measures, mean (MNCS) and percentile ($PP_{top\,x\%}$) citation indicators and self-citation rates (Waltman et al., 2012). The MNCS is calculated for each researchers portfolio of publications and an overall weighted MNCS is calculated for the entire group (i.e., number of publications is used for weighting). Focus is on differences in weighted MNCSs between the three groups and results are scrutinized with bootstrapped 95% confidence intervals (Efron, 1987; Colliander & Ahlgren, 2011) and effect sizes (Cohen, 1988). In this short paper we are not able to go into details of all the analyses carried out, hence the next section will only present some of the main results directly relating to the research questions asked in this paper, i.e., whether there is a substantial difference in the long-term performance between the two postdoc-funding instruments, as well as a substantial difference in long-term performance between being postdoc-funded or not.

Results

The overall publication analyses, including 670 researchers and a publication window from the granting year (estimated year after PhD graduation for the control group) to 2011, showed that individual postdocs' average and median publication activity is 17.5 and 13; embedded postdocs' is 13.4 and 9; and finally the average and median activity for the control group is 10.2 and 8. The average publication period for all groups is 6 years and follows a Gaussian distribution. We see that the publication activity distributions for the two postdoc-groups are considerably more skewed compared to the control group. The publication activity result seems to support the presumptions of the Board of Directors at DFF, however, unfortunately publication activity often plays a too important role in peer review and policy processes at the expense of impact

analyses. A reasonable bibliometric evaluation of the two funding instruments should be based on a comparison of the impact of the groups' publications.

As stated in the methods section, the citation analysis is restricted to publications published from the year of granting (estimated year after PhD for the control group) to 2009. Table 1 below shows that the differences in publication activity between the three groups in absolute numbers are intact in the citation analysis, but more important is the marginal differences in the average and median impact (MNCS) of these publications between the groups.

Table 1. Relative indicators for the three groups.

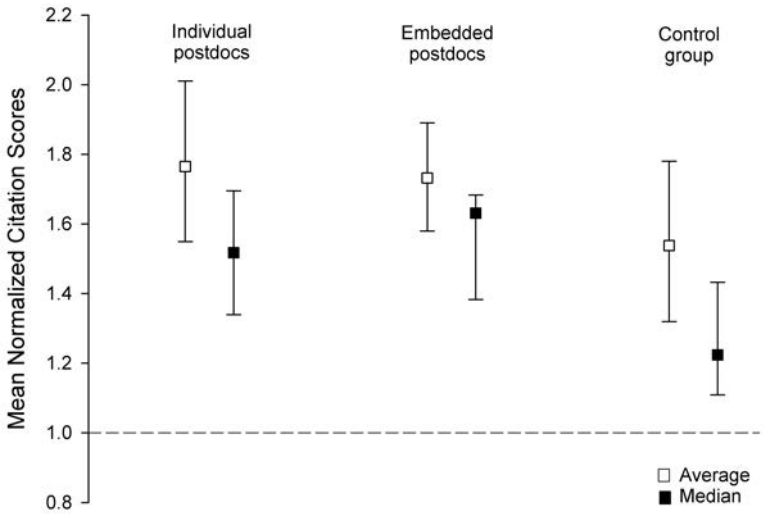
	Control group	Embedded postdocs	Individual postdocs
Total publications	1512.75	1947.75	2735.25
Self-citation rate	29%	29%	31%
Average MNCS*	1.54	1.73	1.76
Median MNCS**	1.22	1.63	1.52
$PP_{top10\%}$***	1.46	1.94	1.80
$PP_{top5\%}$***	1.61	2.13	2.19
$PP_{top1\%}$***	1.66	2.51	3.07

*Average weighted MNCS for all researchers in the group.
**Median weighted MNCS for all researchers in the group.
***Ratio of observed and expected publications among the top *x*% highly cited.

The performance of all three groups is well beyond the "database average" of one, but the publications coming from the two postdoc-instruments seem to have a larger impact on average, some 19 and 22 percentage points higher than the control group (average MNCS), whereas the performance of the two postdoc-instruments seems to be equal. In fact, if we turn to the median MNCS, we see that the embedded postdocs perform slightly better than the individual postdocs. The same pattern is visible with the percentile indicators. All three groups have more publications than expected among the most highly cited in the database, while the two postdoc-groups have larger odds-ratios of highly cited publications compared to the control group. But we also see that at the top 10% level, the embedded postdoc group have a marginally higher ratio of highly cited publications compared to the individual group; however this is reversed at the top 1% where the individual postdocs have a higher ratio. Notice the actual difference in ratios at the top 1% should not be overestimated as data is not robust at this level.

The impression we get from these indicators is that there is no difference between the two postdoc-groups, and a difference between the postdoc-groups and the control group. One should ask whether these results are robust and important. Consider Figure 1 below. Here we see a plot of the mean and median weighted MNCS indicators.

Figure 1. Differences in relative citation impact.



What is also presented in this plot is bootstrapped 95% confidence intervals (CI) surrounding the indicators. It gives no meaning to produce standard CIs as we do not have a probability sample. Instead, we can estimate the robustness of the data a hand by simulating a frequentist experimental situation, in this case with 1000 iterations (Efron, 1987). Several characteristics emerge from the CIs. All groups have rather large CIs for the weighted average MNCS and they all overlap with each other. To a certain extent this can be ascribed to the skewness of the underlying distributions and the strong dependence of highly cited papers in the calculation of averages. Nevertheless, the embedded postdoc-group seems to have the most robust weighted averages. The CIs for the median MNCS reveal asymmetric confidence limits. We notice that the embedded postdoc-group's median MNCS is slightly higher than the individual postdoc, testifying to a more stable underlying distribution in the inter quartile range. Notice again that all CIs overlap with each other. The bootstraped CIs reveal fragile underlying distributions and blurs the immediate impression that there exists a noteworthy difference between the postdoc-groups and the control group.

Further evidence for indistinctive results are provided by scrutinizing the differences by calculating standardized effect sizes for averages and medians presented in Table 2 below.

Table 2. Relative indicators for the three groups.

	Individual vs. embedded postdoc groups	Embedded postdoc group vs. control group
Hedge's g (unbiased) converted to r effect size	.01	.09
Effect size for medians	0.51	.57

We only examine the difference between individual and embedded postdocs, as well as the difference between embedded postdocs and the control group; a further comparison between individual postdocs and the control group is superfluous as the performance of the postdoc-groups is in essence the same. Effect sizes basically quantify how far we are from a null finding. We convert Hedge's *g* for differences in means to *r* for interpretive reasons, and if we compare the results of the two comparisons to Cohen's benchmarks (Cohen, 1988), we see that the differences can be characterized as "trivial". Since, the underlying distributions are skewed, we also calculate an effect size based on medians as suggest by Grissom and Kim (2012). This statistic estimates the probability that a score randomly drawn from population *a* will be greater than a score randomly drawn from population *b*. As we can see from Table 2, the probability for the two postdoc-groups is basically fifty-fifty; and with .57, the probability is only slightly in favour for embedded group compared to the control group. Consequently, the results give evidence to the claim that there is no difference in long-term citation performance between the two postdoc-funding instruments, but also inconclusive evidence for a substantial difference in long-term citation performance between being postdoc-funded or not.

Notice, it could be argued that the individual postdoc group outperforms the two other groups, and likewise that the two postdoc groups outperforms the control group, due to the considerable difference between the groups when it comes to the total number of fractionalized and field normalized citations:, 4814 (individual postdocs), 3369.6 (individual postdocs) and 2329.6 (control groups). Such an argument weights citations by publications thus emphasising the general higher average publication activity among postdocs[ii]. The argument is certainly valid, especially at the individual level when comparing relatively small *n* publication profiles, however we think the argument is more complex at the aggregate level where publication numbers are considerably larger, individual profiles much more skewed within the group, and the publication expectations between groups are unclear. In order to consider such an argument in more detail we need to do more specialized analyses that investigate differences in the distributions of author profiles between the three groups. We also need to consider that postdoc-funding is expected to yield higher publication output. We will do this in a future study.

Discussion

The bibliometric part of the overall evaluation played a dominant role in the official conclusion given by the Board of Directors at DFF. They conclude that the research councils are able to select very competent researchers for postdoc fellowships and that their performance is considerably better than the researchers in the control group, both in relation to publication activity and citation impact (DFF; 2012, p. 3). In other words, the two different postdoc-funding instruments function equally well, contrary to what was presumed, "and certainly much better that traditional tenure-track positions"! The present analysis demonstrates that this conclusion is somewhat overstated and perhaps politically motivated? As indicated above, publication activity is a poor parameter for performance evaluation in this context. Postdocs generally have more time for

research compared to tenure-track positions such as assistant professors. Consequently, it is expected that their activity is higher. When it comes to impact, the present analysis questions the robustness of the findings. First, what seems to be a "significant" difference between the performance of the postdoc-groups compared to the control group, turns out to be a "trivial" difference if we use Cohen's benchmark, or close to a fifty-fifty chance in probabilities. The general conclusion should be that all groups perform above the database average and that the postdoc-groups appear to perform slightly above the control group, but these differences are non-robust and most likely unimportant. Like Kreiman and Maunsell (2011), and others before them, we assert that bibliometric indicators, especially at micro- and meso levels should come with robustness and sensitivity analyses; but we also argue that some sort of yardstick should be provided, in order to make decisions concerning the importance of findings (e.g., Schneider, 2012). We think that this study provides some illustrations of how the robustness of indicators can be scrutinized. Finally, why should the conclusion by DFF be politically motivated? The current trend in Denmark is to terminate traditional tenure-track programs and only give support to postdoc-fellowships. No official reasons are given, but the fact that postdocs are financed by external funding is probably one of them.

References

Bornmann, L. & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review: a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68(3), 427–440.

Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European Molecular Biology Organization programs. *PLOS One*, 3(10): e3480.

Bornmann, L., Leydesdorff, L., & van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics*, 4(3), 211–220.

Böhmer, S., Hornbostel, S., & Meuser, M. (2008). Postdocs in Deutschland: Evaluation des Emmny Noether-Programs. *IFQ-Working Paper No. 3*, Bonn. http://www.forschungsinfo.de/Publikationen/Download/working_paper_3_2008.pdf [accessed 13.03–2013]

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, 5(1), 101–113.

Efron, B. (1987). Better bootstrap con?dence intervals (with Discussion). *Journal of the American Statistical Association*, 82, 171–200.

Evaluering af postdocfinansiering i Det Frie Forskningsråd [*Evaluation of postdoctoral funding within the Free Research Council*], rapport udgivet af Styrelsen for Forskning og Innovation (2012): http://fivu.dk/publikationer/2012/evaluering-af-postdocfinansiering-i-det-frie-forskningsrad.

Freedman, D.A. (2005). *Linear statistical models for causation: A critical review*. In: Everitt, B. & Howell, D., eds., *Encyclopedia of Statistcis in Behavioral Science*, New York: Wiley, 1061–1073.

Gill, J. (2010). Critical differences in Bayesian and non-Bayesian inference. In: Statistics in the social sciences : current methodological developments, Kolenikov, S. ed., Wiley: Hoboken, N.J., 135–158.

Hornbostel, S., Böhmer, S., Klingsporn, B., Neufeld, J., & von Ins, M. (2009). Funding of young scientist and scientific excellence. *Scientometrics*, 79(1), 171–190.

Jacob, B. A., & Lefgren, L. (2011). The impact of NIH postdoctoral training grants on scientific productivity. *Research Policy*, 40(6), 864–874.

Kreiman, G., & Maunsell, J.H.R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience*, 5(48): doi: 10.3389/fncom.2011.00048.

Neufeld, J. & von Ins, M. (2011). Informed peer review and uninformed bibliometrics? *Research Evaluation*, 20(1), 31–46.

Neufeld, J. & Hornbostel, S. (2012). Funding programs for young scientists – do the 'best' apply?. *Research Evaluation*, 21(4), 270–279.

Rosenbaum, P. (2002). Observational studies. 2nd edition. Springer-Verlag: New York.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.

Schneider, J.W. (2012). Testing University Rankings Statistically: Why this perhaps is not such a good idea after all. Some reflections on statistical power, effect Size, random sampling and imaginary populations. In: Archambault, É., Gingras, Y., & Larivière , V., eds., *Proceedings of 17th International Conference on Science and Technology Indicators*, 719–732.

Schneider, J.W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, 7(1), 50–62.

Waltman, L, Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., an Leeuwen, T.N., van Raan, A. F. J., Visser, M. S., Wouters, P. (2012) The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12): 2419–2432.

i Due to a change in the classification of articles in WoS, after the inclusion of the proceedings citation index, some regular articles now carry a double document type classification, i.e., articles; proceedings papers.

ii We thank one of our reviewers for bringing up this point.