NBER WORKING PAPER SERIES

THE MOBILITY OF ELITE LIFE SCIENTISTS: PROFESSIONAL AND PERSONAL DETERMINANTS

Pierre Azoulay
Ina Ganguli
Joshua S. Graff Zivin

Working Paper 21995
http://www.nber.org/papers/w21995

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2016

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. We gratefully acknowledge the financial support of the National Institutes of Health (P01-AG039347). Azoulay and Graff Zivin acknowledge the financial support of the National Science Foundation through its SciSIP Program (Award SBE-1460344). The authors also express gratitude to the Association of American Medical Colleges for providing licensed access to the AAMC Faculty Roster, and acknowledge the stewardship of Dr. Hershel Alexander (AAMC Director of Medical School and Faculty Studies). The National Institutes of Health partially supports the AAMC Faculty Roster under contract HHSN263200900009C. We thank Bruce Weinberg and participants of the NBER Innovation in an Aging Society meetings for useful discussions. All errors are our own.

The Mobility of Elite Life Scientists: Professional and Personal Determinants
Pierre Azoulay, Ina Ganguli, and Joshua S. Graff Zivin
NBER Working Paper No. 21995
February 2016
JEL No. J12,J62,O31

ABSTRACT

As scientists' careers unfold, mobility can allow researchers to find environments where they are more productive and more effectively contribute to the generation of new knowledge. In this paper, we examine the determinants of mobility of elite academics within the life sciences, including individual productivity measures and for the first time, measures of the peer environment and family factors. Using a unique data set compiled from the career histories of 10,004 elite life scientists in the U.S., we paint a nuanced picture of mobility. Prolific scientists are more likely to move, but this impulse is constrained by recent NIH funding. The quality of peer environments both near and far is an additional factor that influences mobility decisions. Interestingly, we also identify a significant role for family structure. Scientists appear to be unwilling to move when their children are between the ages of 14-17, which is when US children are typically enrolled in middle school or high school. This suggests that even elite scientists find it costly to disrupt the social networks of their children and take these costs into account when making career decisions.

Pierre Azoulay
MIT Sloan School of Management
100 Main Street, E62-487
Cambridge, MA 02142
and NBER
pazoulay@mit.edu

Joshua S. Graff Zivin
University of California, San Diego
9500 Gilman Drive, MC 0519
La Jolla, CA 92093-0519
and NBER
jgraffzivin@ucsd.edu

Ina Ganguli
Department of Economics
Thompson Hall
200 Hicks Way
University of Massachusetts
Amherst, MA 01003
iganguli@econs.umass.edu

I. Introduction

A central tenant of modern theories of labor markets is that worker mobility enhances economic productivity by allowing workers to find environments where their skills are put to greatest use. In scientific fields, where team efforts are particularly important, mobility may well increase the production of scientific knowledge (e.g. Hoisl, 2007; Agrawal, McHale, & Oettl, 2014; Ejermo and Ahlin, 2015). Yet, we know surprisingly little about what drives scientists to move in the first place.

Is mobility, in fact, driven largely by efforts to improve employer-employee match quality? What are the constraints to realizing these matches? After all, mobility can generate significant costs, even if only temporary, as a result of professional and personal dislocation. In this paper, we examine both the professional and more personal factors that influence the mobility of elite life scientists. Since many personal factors can influence productivity and vice versa, including both in the same analysis allows us to minimize concerns about statistical confounding and thus develop the most credible measures of each influence to date.

Our analysis builds upon earlier work that has shown the important role played by own-productivity in the propensity to move (e.g. Zucker, Darby & Torero, 2002; Hoisl, 2007; Crespi, Geuna & Nesta, 2007; Lenzi, 2009) to also examine the role played by the quality of the scientific environment more broadly. Science is increasingly a collaborative "team sport" (Wuchty, Jones & Uzzi, 2007), and we exploit novel measures of the quality of peers at local and distant institutions to provide the first systematic analysis of this influence on the decision to relocate.

Our analysis also extends beyond the professional to examine the role of children in shaping mobility decisions. Demographic research has shown that the presence of children in a household can limit scientific mobility (Shauman & Xie, 1996). Moreover, the social psychology literature suggests that it may be particularly costly to move children during secondary school (hereafter "high school" as it is called in the U.S.), when social bonds are strongest and thus the potential for social disruption is greatest (e.g. Fowler, Henry, & Marcal, 2014 and 2015). As such, our analyses will examine how both the number and age of children influences scientist mobility.

The mobility of elite life scientists is of interest for a number of reasons. First, these scientists are largely responsible for pushing the boundaries of the knowledge frontier in their field. Work environments that enhance the returns to their human capital and potential knowledge spillovers to their colleagues can generate sizable social returns by accelerating biomedical innovation and improving human health. Second, the conduct of research in the life sciences is a team effort that often involves expensive and highly specialized equipment, some of which is financed by external sources that are tied to institutions rather than researchers. As such, mobility may be particularly constrained in this population. Finally, the notoriety of this elite group and the public nature of their careers facilitate the collection of data on family structure that is largely unobtainable in other study populations.

We use a unique data set compiled from the career histories of over 10,000 elite life scientists to understand why and when scientists make decisions to move to new locations.[1] Our analysis confirms the importance of scientist productivity as a positive predictor of moves (Zucker, Darby and Torero, 2002; Coupé, Smeets & Warzynski, 2006; Lenzi, 2009; Ganguli, 2015). It also highlights several new professional factors that influence the propensity to move. In particular, we find that recent NIH funding serves as a deterrent to moving, likely due, in part, to the significant transaction costs associated with transferring federal research between institutions (Bernstein, 2014). We also find that the peer environment exerts a significant influence on mobility. Scientists are less likely to move when the quality of the peer environment near their home institution is high and more likely to move when the quality of the peer environment at distant institutions is high.

Turning to the non-professional side, our results reveal an important influence of family structure on mobility. We find a sizable drop in non-local mobility when scientists have children of high school age. Interestingly, scientists appear to anticipate these constraints by increasing moves just before their oldest child enters high school. Mobility accelerates once again when their youngest child is beyond high school age. Additional analyses suggest that the nature of the move – whether it generates a substantial upward or downward change in institutional rank – has little impact on the role played by these professional and personal factors in shaping scientific mobility.

The remainder of the paper is organized as follows. Section 2 describes our data and descriptive statistics. Section 3 lays out our empirical approach. Results are presented in Section 4, while Section 5 concludes.

II. Data and Descriptive Statistics

As described earlier, our analysis will focus on how both professional and personal factors impact the movement of elite life scientists across academic institutions. In this section, we begin with details on the construction of our scientist sample, including our measures of individual productivity. We then describe our measures of the productivity of a scientist's peer environment as well as the sources of data on the children of the elite scientists in our sample.

II.A. Scientist Sample

Our elite academic life scientist sample includes 12,935 individuals, which corresponds to roughly 5 percent of the entire relevant labor market. In our framework, a scientist is deemed elite if they satisfy at least one of the following criteria for cumulative scientific achievement: (1) highly funded scientists; (2) highly cited scientists; (3) top patenters; or (4) members of the National Academy of Sciences. Since these four criteria are based on extraordinary achievement over an entire scientific career, we add additional criteria to capture individuals who show great promise at the early and middle stages of their scientific careers, whether or not these episodes of

[1] We are primarily focused on understanding the determinants of the *timing* of employer changes. Because we do not observe the set of academic institutions to which a scientist could have potentially moved, our results speak to the preferences and constraints that shape the decision to leave one's current institution. We cannot say much about the choice of a specific destination.

productivity endure for long periods of time. These include: (5) NIH MERIT awardees; (6) Howard Hughes Medical Investigators; or (7) early career prize winners.[2] Additional details on this sample construction can be found in Appendix A.

For each scientist in the sample, we reconstruct their career from the time they obtained their first position as independent investigators (typically after a postdoctoral fellowship) until 2006. We do so through a combination of curriculum vitae, NIH biosketches, "*Who's Who"* profiles, accolades/obituaries in medical journals, National Academy of Sciences biographical memoirs, and Google searches. Our dataset includes employment history, degree held, date of degree, gender, and up to three departmental affiliations as well as complete list of publications, patents and NIH funding obtained in each year by each scientist. Publication counts come from the open source software PublicationHarvester. This software downloads from PubMed – an online bibliographic resource from the National Library of Medicine – the entire set of English-language articles for an elite scientist, provided they are not letters to the editor, comments, or other "atypical" articles (Azoulay, Stellman, and Graff Zivin, 2006).[3] Funding data are obtained from the Consolidated Grant/Applicant File (CGAF) from the U.S. National Institutes of Health (NIH), which records information about grants awarded to extramural researchers funded by the NIH since 1938. Patent data come from the US Patent and Trademark Office (USPTO) historical patent data files.

Our mobility data is extracted precisely from biographical records, rather than inferred from affiliation information in papers or patents (e.g. Agrawal et al. 2014) or from self-reported data (Bäker 2015). As such, we observe the exact timing of professional transitions even in the cases in which a scientist has ceased to be active in research, for example because s/he has moved into an administrative position.

Since our analytic approach requires carefully constructed measures of the quality of the research environment (as detailed below) at origin and destination institutions, we exclude all scientists who transition from academic positions to jobs in industry or to foreign academic institutions. This exclusion yields a sample of 10,004 scientists on which our study is based. Our core analysis is focused transitions that are at least 50 miles apart (based on distance between the zip codes of the institutions) to increase the likelihood that this career change leads the scientists to change their place of residence, and thus distance them from their local professional networks and disrupt the social networks of their children. We also compare our main results to the effects for scientists who are local movers (moves within the 50 mile radius).[4] The distribution of distances between institutions for those scientists that move is shown in Figure 1.

Table 1 presents summary statistics for the scientists in the sample who experience at least 1 professional transition (move) over their career and those who do not. The movers represent about 35% of the sample. Movers and stayers look similar in terms of degree type (MD or PhD). Movers are older by approximately 1 year of career age and slightly less likely to be female.

[2] We also cross-reference our list of stars with alternative measures of scientific eminence. For example, the elite subsample contains every U.S.-based Nobel Prize winner in Medicine and Physiology since 1975, and a plurality of the Nobel Prize winners in Chemistry over the same time period.

[3] More details on the assignment of publications to scientists can be found in Appendix B.

(13% vs. 16%). While we only obtain information on children for a subsample of our elite scientist population (as detailed below), the share of scientists with child information and the number of children they have is similar across the mover and stayer samples. Movers are also slightly more productive than their more static counterparts – 10 additional career publications, or roughly 7% more, on average. When accounting for the quality of publications using Journal Impact Factor (JIF)-weighted publications, movers have approximately 22 additional publications on average.

II.B. Quality of the Scientific Environment

The quality of one's peers near and far is presumably an important determinant of scientific mobility. As such, we need measures of both peers and their quality. For the latter, we follow conventions within the literature and use counts of publications and NIH funding (e.g. Azoulay, Graff Zivin, & Wang, 2010). Constructing a measure of the former is far more challenging because the set of scholars that influence the work of a scientist can take many forms. As such, we construct two distinct measures of peers, those defined by direct collaboration as coauthors and those who work in similar fields but who have not directly collaborated.

Collaborators: To identify collaborators, we use two open-source software programs: PublicationHarvester, described earlier, and the Stars/Colleagues Generator (S/CGen).[5] In the first step, the PublicationHarvester downloads from PubMED the entire set of English-language articles for an elite scientist. From this set of publications, the S/CGen strips out the list of coauthors, eliminates duplicate names, matches each coauthor with the Faculty Roster of the Association of American Medical Colleges (AAMC),[6] and stores the identifier of every coauthor for whom a match is found. The software then queries PubMed for each validated coauthor, and generates publication counts for each collaborator scientist in each year.[7]

Non-Collaborator Peers: Our measure of non-collaborating researchers that are intellectually close to the scientist of interest requires a method to delineate the boundaries of research fields. To construct such a measure, we employ a novel approach that groups scientific articles into subfields based on their intellectual content using very detailed keyword information as well as the relative frequencies of these keywords in the scientific corpus (Azoulay, Fons-Rosen, and Graff Zivin, 2015). Specifically, we use the PubMed Related Citations Algorithm (PMRA), which relies heavily on Medical Subject Headings (MeSH). The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary of 24,767 terms arranged in a hierarchical structure. The National Library of Medicine staff use it to tag all of the articles indexed by the MEDLINE

[5] PublicationHarvester and S/CGen are publicly available and can be found, along with user manuals, at http://www.stellman-greene.com/PublicationHarvester/ and http://www.stellman-greene.com/ScientificDistance/, respectively.

[6] The roster is an annual census of all U.S. medical school faculty, where each faculty is linked across yearly cross-sections by a unique identifier. We have licensed access to the AAMC data for the years 1975 through 2006.

[7] See the online appendix from Azoulay et al. (2010) for details on the matching procedure, preventing inclusion of spurious coauthors, and the approach to addressing measurement error when tallying the publication output of coauthors with common names.

database.[8] The "Related Articles" function in PubMed is use to harvest journal articles that are intellectually proximate to the elite scientists' own papers.[9] The authors of those articles are then classified as non-collaborating peers if they have never coauthored with the elite scientist of interest. More details on this approach can be found in Appendix C.

Geography: Our measure of elite scientist location is based on the detailed biographical information we have for this sample, as detailed above. Collaborating and non-collaborating peers are mapped in physical space using affiliation data from the aforementioned Faculty Roster of the Association of American Medical Colleges (AAMC). Our measure of geography distinguishes between peers that are geographically close (less than 50 miles apart) and those that are distant (more than 50 miles apart).

In Table 2, we present descriptive statistics comparing the peer environment for scientists in the sample who experience at least one professional transition (move) over their career to those who do not. The collaboration networks appear to be quite important. Movers have fewer and less accomplished local collaborators and greater and more accomplished distant collaborators than those that do not move. Interestingly, the quantity and quality of non-collaborating peers is lower for movers than stayers, regardless of whether they are local or distant.

II.C. Age of children

For each scientist in our sample, we hand-collected information on the number of children they had, each child's gender, and most importantly, each child's year of birth. To find this information, we obtained the names of the scientist's children from their "Who's Who" profile and in some cases from obituaries for the deceased. Using this data along with location information, we obtained the age of these children by cross-referencing information gathered from web-searches and online databases of public records (e.g. *People Search Now*). In the end, this yielded a subsample of slightly more than 3,000 elite scientists for whom age of children information is available.

Table 3 compares the sample of scientists for whom we have obtained information about their children to those for whom we have not. The former group is significantly older than the latter, an artifact of our reliance on public records to obtain age of children, much of which comes from databases of public records, such as state driver's license records, which necessarily oversamples scientists old enough to have kids that make them eligible to appear in these public records. While these older scientists have more publications and career NIH funding, normalizing by age suggests that they are statistically indistinguishable. The higher level of female scientists in the sample without children information may also be a reflection of this age difference across samples, as female entry into STEM fields has steadily climbed in recent decades (Ceci, Ginther, Kahn, and Williams, 2015). Interestingly, the composition of the sample without children

[8] The National Library of Medicine's explicit statement of purpose for these MeSH terms is to "…provide a reproducible partition of concepts relevant to biomedicine for the purpose of organizing knowledge and information."

[9] To facilitate the harvesting of PubMed-related records on a large scale, we have developed an open-source software tool that queries PubMed and PMRA and stores the retrieved data in a MySQL database. The software is available for download at http://www.stellman-greene.com/FindRelated/.

information is more heavily skewed toward PhDs. While all the analyses that follow will control for these demographic characteristics, caution should, nonetheless, be exercised when generalizing our findings across different populations of scientists.

III. Empirical Approach

Estimating the determinants of faculty mobility behavior requires a statistical framework that accommodates the discrete nature of the event. Since our interest lies in analyzing the dynamics associated with the timing of mobility in scientific careers, we employ discrete-time hazard rate models (Myers, Hanky and Mantel 1973, Alison 1982). The use of discrete-time models (as opposed to continuous-time models such as the Cox) is motivated by the fact that our failure time variable displays multiple events within each time period. For a researcher *i* during experience interval *t*, let the discrete time hazard rate of moving to a new academic position located at least 50 miles away be $p_{it} = \Pr[T_i=t \mid T_i \geq t, X_{it}]$, where T_i is the time at which researcher *i* experiences an event and X_{it} a vector of covariates. We use a logistic regression function to link the hazard rate with time and the explanatory covariates:

$$Ln\left[\frac{p_{it}}{1-p_{it}}\right] = \delta_t + \beta' X_{it}$$

where δ_t is a set of experience interval dummies (in actual fact, a full suite of calendar year indicator variables). In practice, we estimate a simple logit of the decision to change employer, where the observations corresponding to years subsequent to the mobility event have been dropped from the estimation sample.[10] Specifically, the vector *X* above is specified as:

$$X_{it} = \beta_0 + \beta_1 PROD_{it-1} + \beta_2 PEER_{it-1} + \beta_3 AGEKID_{it-1} + \beta_4 Z_i + f(AGE_{it})$$

Note that approximately ten percent of scientists in our sample experience multiple moves during their career. For these scientists, we analyze each job spell separately, such that the move is considered an absorbing state for a given mobility spell.[11] As such, our dependent variable is a binary variable equal to 1 if a scientist moves in a given year and 0 for all years prior to that move within the same spell. Because the overwhelming majority of mobility events take place in the summer, we adopt the following convention: a scientist is said to move from institution A to institution B in calendar year *t* whenever the actual timing of his move coincided with the summer of year *t-1*.[12]

The vector PROD includes measures of the productivity of the focal superstar scientist, including the stock and recent flow of their publications and NIH funding levels. The vector PEER

[10] Appendix D replicates this analysis using a standard ordinary least squares (OLS) framework. One advantage of the OLS framework is that we are able to include scientist-spell fixed effects, which allows us to isolate the effects of within spell changes in covariates on the likelihood of moving, while controlling for all time invariant characteristics of scientists (see discussion in Wooldridge, 2010). It is noteworthy that all of our results are very similar under this linear specification, with or without fixed effects.

[11] The most itinerant star in our sample has 4 unique job spells during our study period.

[12] This convention is adopted since most academic moves occur during the summer. Moreover, this ensures that our measure of the timing of professional transitions corresponds to the timing of the school year for children in our sample.

includes similar measures for both collaborating and non-collaborating peers as defined above. The term f(AGE) corresponds to a flexible function of the elite scientist's career age in order to capture life-cycle changes in the propensity to move that are not driven by the age of one's children. We also include demographic controls (Z) for gender and degree type (MD or PhD). All models include vintage fixed effects to account for differences across cohorts and a full set of year effects. Since we have 1,051 scientists in the sample with multiple moves, and each job spell is included in the sample separately, we cluster the standard errors at the scientist level.

Given our interest in the high social costs of moving children of high school age, we employ a variety of measures of AGEKID to probe this effect. In particular, we create variables that correspond to the case where one's oldest child is 12 or 13 years old and where one's youngest child is 18 or 19 years old. The former corresponds to the last window for a scientist to move before they have a child in high school. The latter corresponds to an 'empty nest' period in which all children should have graduated high school, regardless of whether they actually leave the nest. Alternative specifications include a simple measure of the number of children in high school and an indicator for having at least one child in high school.

Figures 2 and 3 underscore the important, and previously unmeasured, role that children play in scientist mobility. Figure 2 reveals a sizable spike in distant moves just before children in the household enter high school. Figure 3 reveals a similar spike just after all children in the household have completed high school. In both figures, the relationship between age of children and *local* moves is remarkably flat, suggesting that this is not simply a story about scientist age. The robustness of this relationship to more sophisticated statistical scrutiny that addresses potential confounders as well as the role of professional factors in shaping moves is examined in the next section.

IV. Results

Our core results on the demographic and professional factors that influence the mobility of scientists are presented in Table 4. Consistent with related work in the literature (Zucker, Darby and Torero, 2002; Coupé, Smeets and Warzynski, 2006; Lenzi, 2009; Ganguli, 2015), we find that scientists who are more productive in terms of publications are more likely to move. Interesting, we find that recent NIH funding serves as a deterrent to moving. This comports with popular accounts of the high transaction costs associated with moving federal funding across institutions (Berstein, 2014). As expected, scientists are less likely to move when the quality of the peer environment near their home institution is high and more likely to move when the quality of the peer environment at distant institutions is high. Importantly, given the focus on family factors later in the analysis, all results are nearly identical when we restrict our attention to the sample of scientists with available age of children information.

Table 5 repeats our analysis of the professional determinants of moving from Table 4, but for local moves (those within a 50 mile radius of the origin institution). The predictors of mobility are largely similar across samples, with two notable exceptions. First, female scientists are more likely to make local moves. Second, own-productivity plays a far more limited role in shaping local moves than it did in distant ones.

Our results based on an analysis of the family determinants of moving begins in Table 6. This table confirms the descriptive relationship illustrated in Figures 1 and 2. In Column 1 we see that having a child who is finishing middle school (12 or 13 years old) increases the likelihood of moving by 0.8 percentage points. In column 2, we see that when the youngest child has just completed high school, the propensity to move also increases, in this case also by 0.8 percentage points. Our results are very similar when we include both of these measures in the same regression. Using a simple indicator variable for having a child in school or the number of children in school illustrates the opposite side of this picture. Having children in high school constrains the mobility of scientists.

Table 7 presents results for the same regression as those in Table 6, but for local movers only (moves within the 50 mile radius). The picture here is very different. Having children in high school has no effect on moving and thus we see no corresponding bump in moves when the youngest child in the household is beyond high school age. While we do see a small and marginally significant effect when the oldest child in the household is of middle-school age, it is one-third the size of the effect we see for distant moves and in the opposite direction. Overall, these results are consistent with the notion that a move to an institution within 50 miles can avoid major disruptions to children's social networks and school choices.

Taken as a whole, the results in Tables 6 and 7 suggest that scientists, and in this case even elite superstar scientists, find it costly to disrupt the social networks of their children. They either postpone professional moves until after children leave high school or interestingly, anticipate these constraints by increasing moves just before their oldest child enters high school.

All of the analyses thus far have treated all moves as equivalent, but heterogeneity in the rank of institutions may mask important insights about the determinants of mobility. The drivers of moves to higher-ranked institutions may be very different than those that contribute to lateral or lesser-ranked institutions. In Tables 8 and 9 we examine how professional and family factors differentially affect mobility across different types of moves. To examine this, we categorize moves based on the differences between the quality of their origin and destination institutions, where our measure of quality is based on an institutions rank in percentiles of total NIH funding received (per grantee) in a given year. We define scientists as moving "up" if they moved to an institution that was at least 10 percentiles higher in the ranking of NIH funding than their prior institution. A move "down" is symmetrically defined as one where the new institution was at least 10 percentiles lower in the ranking.

Table 8 presents our results for the professional determinants of moves separately for those moving up (columns 1 and 3) and those moving down (columns 2 and 4). Note that in these regressions, scientists who moved to an institution that was less than 10 percentiles different (i.e. lateral movers) and scientist who moved the 'other way' are excluded from the regression such that the comparison is relative to stayers. We also include the percentile of the origin institution as a control since the direction of moves for those beginning at either end of the quality ladder will be partially constrained. The results suggest that the professional drivers of scientific mobility are insensitive to the type of move, lending support to the notion that all moves in this sample of elite scientists are 'voluntary' and driven by a desire to locate near higher quality peers, irrespective of broader institutional quality.

In Table 9, we repeat the analysis from the previous table with a focus on the family determinants of moves. Here we see that children constrain mobility for both types of moves, but the effect of having at least one child in high school is somewhat larger for downward moves. Thus, it appears that the greater reward from moving to a better institution partially offsets the social costs associated with uprooting one's family.

Appendix D repeats the analysis corresponding to Tables 4-9 using a linear probability (OLS) and fixed effects (where appropriate) framework. Reassuringly, we find that all key results are largely unchanged.

VI. Conclusion

In this paper, we examine the factors that shape the mobility of elite academics within the life sciences, including individual productivity measures and for the first time, measures of the peer environment and family determinants. As in previous literature, we find that prolific scientists are more likely to move, a likely reflection of the 'demand' for these scientists. We also provide new evidence on the 'supply' side, demonstrating that moves are driven in part by the scope for improvement in the quality of one's peers as a result of the move. We also highlight two significant constraints on moving. Elite scientists are less likely to move when they have recently received NIH funding, perhaps due to the high costs of transferring funds, equipment, and personnel to a new institution. Strikingly, we show that family structure is also important. Scientists appear to be unwilling to move when their children are in high school, suggesting that even elite scientists find it costly to disrupt the social networks of their children and take these costs into account when making career decisions.

Economists have long worried that labor market frictions can lead to job lock and foregone opportunities to increase social welfare. Set against that backdrop, the impacts of NIH funding on mobility are particularly interesting. If the purported benefit of federal research support is to generate new discoveries and push the frontiers of science, it is ironic that it also appears to limit scientists' ability to locate themselves in environments where that is most likely to happen. This is a particularly germane concern in the life sciences where public funding is the lifeblood of academic life scientists. Thus, it appears that the designation of policies and guidelines that increase the portability of research funding support can help further the mission of the NIH, and perhaps other public research funding agencies as well.

Our novel findings regarding the role of peer productivity in shaping mobility illustrate the challenges inherent in developing *causal* estimates of the impacts of mobility on knowledge production. Since mobility decisions depend upon the quality of scholars at old and new institutions, it is exceedingly difficult to infer the impacts that a scholar exerts on the productivity of his or her newly joined colleagues or those left behind. Departures may be a sign of intellectual decay even before the moving scientist left. Those s/he joined may have already been destined for greatness. In this context, our age of children results may contain a silver lining. Since scientists are reluctant to move when they have children in high school, and it is hard to imagine that they anticipate these costs more than a decade earlier when they decide to

conceive, age of children offers a potential instrumental variable for studying the causal impacts of moving. This is an area ripe for exploration in future work.

The analysis in this paper has raised as many questions as it has answered. What underlies the surprising funding results? Is mobility constrained by something inherent in the contract between the NIH and the scientist's institution or might the movement of personnel and equipment be the larger concern? On family factors, what else matters for moving? Are women, too few in our elite sample to study, more impacted by those factors than men? How should all of this change the design of institutional incentives, particularly those that explicitly grant promotion and tenure based on funding histories? Together, these questions comprise a future research agenda.

V.II. References

Agrawal, A., McHale, J., & Oettl, A. (2014). *Why stars matter*, NBER WP 20012.

Ahlin L. and Ejermo O. (2015). The patent productivity effects of mobility for a panel of Swedish inventors, DRUID, 2015.

Allison, Paul D. (1982). "Discrete-Time Methods for the Analysis of Event Histories," in Sociological Methodology. S. Leinhardt, ed. San Francisco: Jossey-Bass, pp. 61-98.

Azoulay, P. Fons-Rosen, C. and Graff Zivin, J. (2015). "Does Science Advance One Funeral at a Time?" NBER Working Paper No. 21788.

Azoulay, P., Stellman, A., and Graff Zivin, J. (2006). PublicationHarvester: An Open-Source Software Tool for Science Policy Research. *Research Policy*, *35*(7), 970-974.

Azoulay, P., Graff Zivin, J., & Wang, J. (2010). Superstar Extinction. *The Quarterly Journal of Economics*, *125*(2), 549-589.

Bäker, A. (2015). Non-tenured post-doctoral researchers' job mobility and research output: An analysis of the role of research discipline, department size, and coauthors. *Research Policy*, 44(3), 634-650.

Bernstein, R. (2014). Managing a Lab Move. *Science Careers*, September 23, 2014.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. Psychological Science in the Public Interest, 15(3), 75-141.

Coupé, T., Smeets, V., & Warzynski, F. (2006). Incentives, sorting and productivity along the career: Evidence from a sample of top economists. Journal of Law, Economics, and Organization, 22(1), 137-167.

Crespi, G. A., Geuna, A., & Nesta, L. (2007). The mobility of university inventors in Europe. The Journal of Technology Transfer, 32(3), 195-215.

Fowler, P. J., Henry, D. B., & Marcal, K. E. (2015). Family and housing instability: Longitudinal impact on adolescent emotional and behavioral well-being. *Social science research*, *53*, 364-374.

Fowler, P. J., Henry, D. B., Schoeny, M., Taylor, J., & Chavira, D. (2014). Developmental timing of housing mobility: Longitudinal effects on externalizing behaviors among at-risk youth. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(2), 199-208.

Ganguli, I. (2015). "Who Leaves and Who Stays? Evidence on Immigrant Selection from the Collapse of Soviet Science" in Aldo Geuna (ed), Global Mobility of Research Scientists: The Economics of Who Goes Where and Why, Elsevier.

Hoisl, K. (2007). Tracing mobile inventors—the causality between inventor mobility and inventor productivity. *Research Policy*, 36(5), 619-636.

Lenzi, C. (2009). Patterns and determinants of skilled workers' mobility: evidence from a survey of Italian inventors. Economics of Innovation and New Technology, 18(2), 161-179.

Myers, M.H., B.F. Hankey, and N. Mantel. 1973. "A Logistic Exponential Model for Use with Response-Time Data Involving Regressor Variables." Biometrics, 29, pp. 257-69.

Shauman, K. A., & Xie, Y. (1996). Geographic mobility of scientists: Sex differences and family constraints. *Demography*, 33(4), 455-468.

Wooldridge, J. M. (2010). Econometric Analysis Of Cross Section And Panel Data. MIT press, 2010.

Wuchty S., B. Jones, B. Uzzi. 2007. The increasing dominance of teams in production of knowledge. Science 316: 1036.

Zucker, L. G., Darby, M. R., & Torero, M. (2002). Labor Mobility from Academe to Commerce. Journal of Labor Economics, 20(3), 629-660.

Table 1. Summary Statistics of Stayers and Movers: Demographic and Individual Productivity

	Stayers		Movers	
	Mean	Std. Dev.	Mean	Std. Dev.
Female	0.156	0.363	0.132	0.338
MD	0.335	0.472	0.335	0.472
PhD	0.566	0.496	0.570	0.495
MD/PhD	0.098	0.297	0.094	0.292
Career Age	32.854	9.770	33.917	8.444
With Kids info	0.305	0.461	0.334	0.472
No. of Kids	2.484	1.120	2.512	1.108
School NIH Funding	138,992,401	115,394,909	127,110,155	107,733,856
Individual Productivity				
Career Publications	137.051	104.468	147.302	106.705
Career JIF-Weighted Pubs	591.85	537.92	614.53	544.05
Career Patents	1.982	6.688	1.771	5.052
Career NIH Amount	14,968,044	21,276,284	15,715,106	17,598,271
Observations	6,545		3,459	

Notes: For each scientist in the sample, we reconstruct their career from the time they obtained their first position as independent investigators (typically after a postdoctoral fellowship) until 2006. The numbers presented in this table are based on the final year the scientist appears in the dataset. See section IIA for more information about the sample construction and sources of data. We identify movers by extracting information on institutions from biographical records and then calculate geographic distance between the zip codes of the institutions. The stayers include scientists who do not move or only move locally (within 50 miles). We limit our empirical attention to transitions that are at least 50 miles apart to ensure that the career transition led the scientists to change their place of residence. Stars in the last column indicate the results of t-tests for the equality of means, ** $p < 0.05$, *** $p < 0.01$.

Table 2. Summary Statistics of Stayers and Movers: Peers Measures

	Stayers		Movers	
	Mean	Std. Dev.	Mean	Std. Dev.
Collaborators, Number				
Nb. of Peers, Colocated	5.374	5.858	3.037	4.081
Nb. of Peers, Close	3.748	5.608	2.032	3.565
Nb. of Peers, Distant	11.949	10.732	13.214	11.906
Non-collaborating Peers, Number				
Nb. of Peers, Colocated	3.208	4.046	2.500	3.612
Nb. of Peers, Close	7.875	10.272	5.922	8.656
Nb. of Peers, Distant	122.645	88.328	107.232	78.410
Collaborators' Productivity				
Pubs, Colocated	23.264	32.175	12.267	21.639
Pubs, Close	14.876	30.786	6.958	16.667
Pubs, Distant	57.206	70.084	58.299	69.843
Non-collaborating Peers Prod.				
Pubs, Colocated	14.888	23.895	11.126	21.263
Pubs, Close	32.775	49.253	23.642	40.791
Pubs, Distant	494.192	372.864	408.378	324.952
Observations	6,545		3,459	

Notes: The numbers presented in this table are based on the final year the scientist appears in the dataset. See section IIB for more information about the measures of the peer environment and the sources of data. We identify movers by extracting information on institutions from biographical records and then calculate geographic distance between the zip codes of the institutions. The stayers include scientists who do not move or only move locally (within 50 miles). We limit our empirical attention to transitions that are at least 50 miles apart to ensure that the career transition led the scientists to change their place of residence. Stars in the last column indicate the results of t-tests for the equality of means, *** $p < 0.01$.

Table 3. Summary Statistics: Stars with and without Kids Info

	No Kids Info		Kids Info		Difference
	Mean	Std. Dev.	Mean	Std. Dev.	
Female	0.168	0.374	0.102	0.303	0.066***
MD	0.297	0.457	0.418	0.493	-0.121***
PhD	0.607	0.489	0.483	0.500	0.124***
MD/PhD	0.096	0.294	0.098	0.298	-0.002
Career Age	31.227	9.346	37.553	7.744	-6.326***
Career Publications	133.012	98.487	157.065	117.250	-24.053***
Career NIH Amount	14,143,602	19,867,100	17,577,908	20,350,240	-3,434,305***
Career Pubs/Age	4.363	3.018	4.294	3.203	0.069
Career NIH/Age	446,756	614,629	467,250	527,525	-20,494
Observations	6,850		3,154		

Notes: The numbers presented in this table are based on the final year the scientist appears in the dataset. See section IIC for details on how the information on children was collected. Section IIA provides more information about the construction of the demographic and productivity measures presented. Stars in the last column indicate the results of t-tests for the equality of means, *** $p < 0.01$.

Table 4. Determinants of Mobility: Demographic, Productivity & Peer Measures

	Movers + Stayers				Subsample w/Kid Info
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	-0.001	-0.001	-0.000	-0.000	-0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
PhD (MD omitted)	-0.001	-0.004**	-0.007**	-0.007**	-0.009**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
MD/PhD	-0.001	-0.005**	-0.006**	-0.005**	-0.005+
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
Productivity Measures					
Ln(Pubs_t-1)		-0.001	0.001	0.000	-0.001
		(0.001)	(0.001)	(0.001)	(0.001)
Ln(Stock Pubs_t-2)		0.003**	0.005**	0.004**	0.006**
		(0.001)	(0.001)	(0.001)	(0.001)
Ln(NIH Funding_t-1)		-0.001**	-0.001**	-0.001**	-0.001**
		(0.000)	(0.000)	(0.000)	(0.000)
Ln(Stock NIH		0.001**	0.001**	0.001**	0.000+
Funding_t-2)		(0.000)	(0.000)	(0.000)	(0.000)
Collaborators					
Ln(Pubs), Colocated			-0.004**	-0.002**	-0.001*
			(0.000)	(0.000)	(0.001)
Ln(Pubs), Close			-0.002**	-0.002**	-0.002*
			(0.000)	(0.000)	(0.001)
Ln(Pubs), Distant			0.001**	0.001**	0.001+
			(0.000)	(0.000)	(0.001)
Non-collaborating Peers					
Ln(Pubs), Colocated				-0.007**	-0.006**
				(0.000)	(0.001)
Ln(Pubs), Close				-0.000	-0.000
				(0.000)	(0.000)
Ln(Pubs), Distant				0.005**	0.004**
				(0.001)	(0.001)
Nb. of Observations	190,266	174,502	174,465	174,465	61,907
Nb. of Job Transitions	10,289	10,289	10,289	10,289	3,298
Nb. of Scientists	9,389	9,389	9,389	9,389	2,960

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located at least 50 miles away. Estimation is by logit and marginal effects are reported. All specifications include full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 5. Determinants of LOCAL Mobility: Demographic, Productivity & Peer Measures

	Local Movers + Stayers				Subsample w/Kid Info
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	0.003**	0.003**	0.003**	0.002**	0.002+
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PhD (MD omitted)	-0.002**	-0.002**	-0.002**	-0.002**	-0.003**
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
MD/PhD	-0.002+	-0.002*	-0.002*	-0.002*	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Productivity Measures					
Ln(Pubs_t-1)		-0.000	-0.000	-0.000	0.001
		(0.000)	(0.000)	(0.000)	(0.001)
Ln(Stock Pubs_t-2)		0.000	0.000	0.001	0.002*
		(0.000)	(0.000)	(0.000)	(0.001)
Ln(NIH Funding_t-1)		-0.000*	-0.000*	-0.000*	-0.000**
		(0.000)	(0.000)	(0.000)	(0.000)
Ln(Stock NIH		0.000	0.000	0.000	-0.000
Funding_t-2)		(0.000)	(0.000)	(0.000)	(0.000)
Collaborators					
Ln(Pubs), Colocated			-0.001**	-0.001**	-0.001
			(0.000)	(0.000)	(0.000)
Ln(Pubs), Close			0.002**	0.000+	0.001*
			(0.000)	(0.000)	(0.000)
Ln(Pubs), Distant			-0.000+	-0.000	-0.001*
			(0.000)	(0.000)	(0.000)
Non-collaborating Peers					
Ln(Pubs), Colocated				-0.002**	-0.002**
				(0.000)	(0.000)
Ln(Pubs), Close				0.003**	0.003**
				(0.000)	(0.000)
Ln(Pubs), Distant				-0.001**	-0.002**
				(0.000)	(0.001)
Nb. of Observations	154,057	144,113	144,079	144,079	50,161
Nb. of Job Transitions	6,722	6,722	6,722	6,722	2,061
Nb. of Scientists	6,637	6,637	6,637	6,637	2,025

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Distant movers (scientists moving more than 50 miles away) are excluded from this analysis. Estimation is by logit and marginal effects are reported. All specifications include full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 6. Determinants of Mobility: Child Age

	(1)	(2)	(3)	(4)	(5)
Oldest kid 12 or 13	0.008**		0.008**		
	(0.002)		(0.002)		
Youngest kid 18 or 19		0.008**	0.009**		
		(0.002)	(0.002)		
Number of kids in high school				-0.007**	
				(0.001)	
At least one kid in high school					-0.011**
					(0.002)
Nb. of Observations	61,907	61,907	61,907	61,907	61,907
Nb. of Job Transitions	3,298	3,298	3,298	3,298	3,298
Nb. of Scientists	2,960	2,960	2,960	2,960	2,960

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located at least 50 miles away. Estimation is by logit and marginal effects are reported. All specifications include individual productivity and peer variables, as well as full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 7. Determinants of LOCAL Mobility: Child Age

	(1)	(2)	(3)	(4)	(5)
Oldest kid 11, 12, or 13	-0.003^{+}		-0.003^{+}		
	(0.001)		(0.001)		
Youngest kid 18, 19, or 20		0.001	0.000		
		(0.001)	(0.001)		
Number of kids in high school				0.001	
				(0.001)	
At least one kid in high school					0.001
					(0.001)
Nb. of Observations	50,161	50,161	50,161	50,161	50,161
Nb. of Job Transitions	2,061	2,061	2,061	2,061	2,061
Nb. of Scientists	2,025	2,025	2,025	2,025	2,025

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Distant movers (scientists moving more than 50 miles away) are excluded from this analysis. Estimation is by logit and marginal effects are reported. All specifications include individual productivity and peer variables, as well as full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 8. Moves Up vs. Down: Demographic, Professional Factors

	Moves Up (1)	Moves Down (2)
Demographics		
Female	-0.001	-0.002
	(0.002)	(0.002)
PhD	-0.004**	-0.005**
	(0.001)	(0.001)
MD/PhD	-0.004*	-0.002
	(0.002)	(0.002)
Productivity Measures		
Ln(Pubs_t-1)	0.001	0.001
	(0.001)	(0.001)
Ln(Stk Pubs_t-2)	-0.000	0.003**
	(0.001)	(0.001)
Ln(NIH Funding_t-1)	-0.000	-0.000+
	(0.000)	(0.000)
Ln(Stk NIH Funding_t-2)	0.000	0.000
	(0.000)	(0.000)
Collaborators		
Ln(Pubs), Colocated	-0.001	-0.001+
	(0.000)	(0.000)
Ln(Pubs), Close	-0.000	-0.001
	(0.000)	(0.000)
Ln(Pubs), Distant	0.001	-0.000
	(0.000)	(0.000)
Non-collaborating Peers		
Ln(Pubs), Colocated	-0.002**	-0.002**
	(0.000)	(0.000)
Ln(Pubs), Close	-0.000	-0.000
	(0.000)	(0.000)
Ln(Pubs), Distant	0.002**	0.002+
	(0.001)	(0.001)
Nb. of Observations	38,615	39,415
Nb. of Job Transitions	1,775	1,947
Nb. of Scientists	1,773	1,937

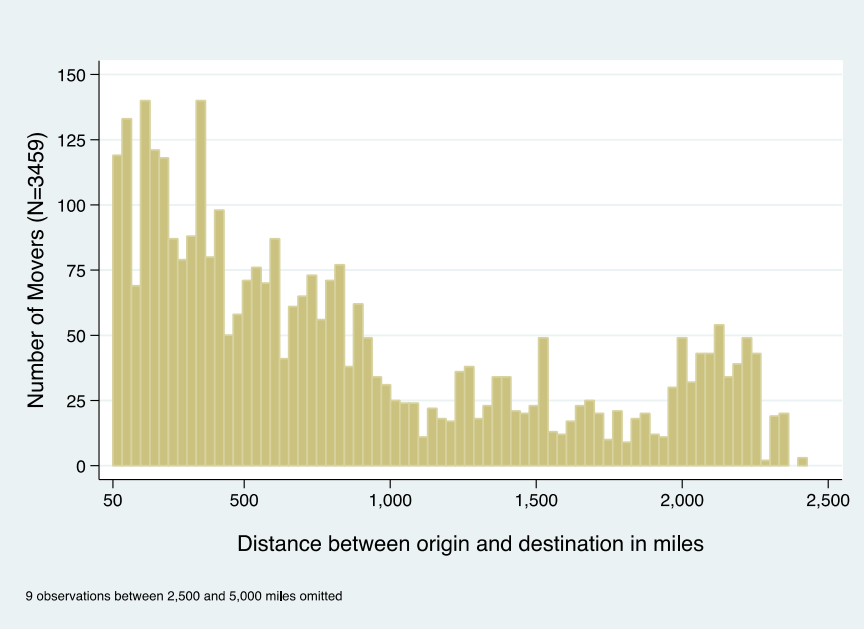
Note: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Moves up or down are based on a ranking of institutions by percentiles of total NIH funding received (per grantee). The sample includes scientists who moved to an institution that was 10 percentiles or more higher (lower) in the ranking of NIH funding and scientists who did not move. This means that scientists who moved to an institution that was less than 10 percentiles different (i.e. lateral movers) and scientists who moved down (up) are excluded. Estimation is by logit and marginal effects are reported. All regressions include full age, vintage category, and year fixed effects and controls for percentile of origin institution. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table 9. Moves Up vs. Down: Child Age

	Moves Up		Moves Down	
	(1)	(2)	(3)	(4)
Number of kids in high school	-0.0017*		-0.0020*	
	(0.0007)		(0.0010)	
At least one kid in high school		-0.0024*		-0.0040**
		(0.0010)		(0.0012)
Percentile of Origin Institution	-0.0004**	-0.0004**	0.0004**	0.0004**
	(0.0000)	(0.0000)	(0.0001)	(0.0001)
Nb. of Observations	38,615	38,615	39,415	39,415
Nb. of Job Transitions	1,775	1,775	1,947	1,947
Nb. of Scientists	1,773	1,773	1,937	1,937

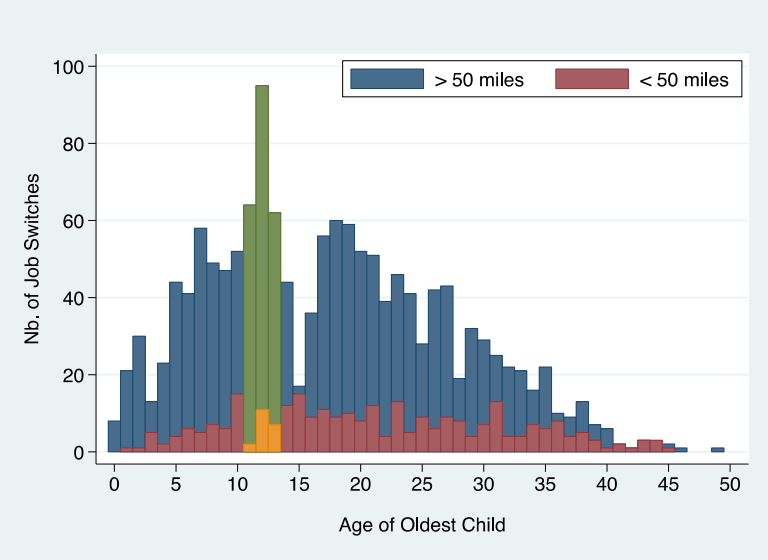
Note: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Moves up or down are based on a ranking of institutions by percentiles of total NIH funding received (per grantee). The sample includes scientists who moved to an institution that was 10 percentiles or more higher (lower) in the ranking of NIH funding and scientists who did not move. This means that scientists who moved to an institution that was less than 10 percentiles different (i.e. lateral movers) and scientists who moved down (up) are excluded. Estimation is by logit and marginal effects are reported. All regressions include full age, vintage category, and year fixed effects and controls for percentile of origin institution. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Figure 1. Distance of Moves by Elite Scientists



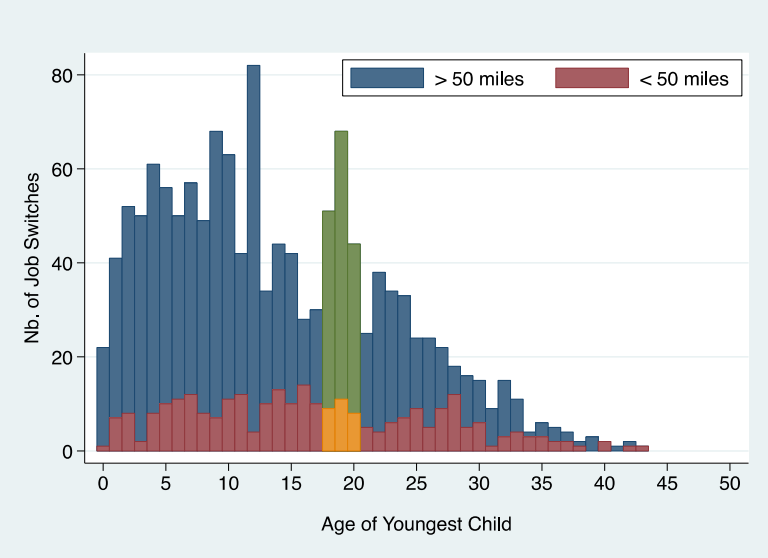
Notes: We determine moves by extracting information on institutions from biographical records and then calculate geographic distance between the zip codes of the institutions. We limit our empirical attention to transitions that are at least 50 miles apart to ensure that the career transition led the scientists to change their place of residence. Section II for a full description of how the sample and variables were constructed.

Figure 2. Age of Children and Mobility: Age of Oldest



Notes: See section IIC for details on how the information on children was collected.

Figure 3. Age of Children and Mobility: Age of Youngest



Notes: See section IIC for details on how the information on children was collected.

Appendix A: Defining Elite Life Scientists

Highly Funded Scientists: Our first data source is the Consolidated Grant/Applicant File (CGAF) from the U.S. National Institutes of Health (NIH). This dataset records information about grants awarded to extramural researchers funded by the NIH since 1938. Using the CGAF and focusing only on direct costs associated with research grants, we compute individual cumulative totals for the decades 1977-1986, 1987-1996, and 1997-2006, deflating the earlier years by the Biomedical Research Producer Price Index. We also re-compute these totals excluding large center grants that usually fund groups of investigators (M01 and P01 grants). Scientists whose totals lie above the 95th percentile of either distribution constitute our first group of elite life scientists. In this group, the least well-funded investigator garnered $10.5 million in career NIH funding and the most well-funded received $462.6 million.[13]

Highly Cited Scientists: Despite the preeminent role of the NIH in the funding of public biomedical research, the above indicator of "superstardom" biases the sample towards scientists conducting relatively expensive research. We complement this first group with a second composed of highly cited scientists identified by the Institute for Scientific Information. A Highly Cited listing means that an individual was among the 250 most cited researchers for their published articles between 1981 and 1999, within a broad scientific field.[14]

Top Patenters: We add to these groups academic life scientists who belong in the top percentile of the patent distribution among academics – those who were granted 17 patents or more between 1976 and 2004.

Members of the National Academy of Science and of the Institute of Medicine: We add to these groups academic life scientists who were elected to the National Academy of Science or the Institute of Medicine between 1970 and 2013.

MERIT Awardees of the NIH: Initiated in the mid-1980s, the MERIT Award program extends funding for up to 5 years (but typically 3 years) to a select number of NIH-funded investigators …"who have demonstrated superior competence, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas." The specific details governing selection vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application must be scored in the top percentile in a given funding cycle.

Former and Current Howard Hughes Medical Investigators (HHMIs): Every three years, the Howard Hughes Medical Institute selects a small cohort of mid-career biomedical scientists

[13] We perform a similar exercise for scientists employed by the intramural campus of the NIH. These scientists are not eligible to receive extramural funds, but the NIH keeps records of the number of "internal projects" each intramural scientist leads. We include in the elite sample the top five percentiles of intramural scientists according to this metric.

[14] The relevant scientific fields in the life sciences are microbiology, biochemistry, psychiatry/psychology, neuroscience, molecular biology & genetics, immunology, pharmacology, and clinical medicine.

with the potential to revolutionize their respective subfields. Once selected, HHMIs continue to be based at their institutions, typically leading a research group of 10 to 25 students, postdoctoral associates and technicians. Their appointment is reviewed every five years, based solely on their most important contributions during the cycle.[15]

Early Career Prize Winners: We also included winners of the Pew, Searle, Beckman, Rita Allen, and Packard scholarships for the years 1981 through 2000. Every year, these charitable foundations provide seed funding to between 20 and 40 young academic life scientists. These scholarships are the most prestigious accolades that young researchers can receive in the first two years of their careers as independent investigators.

[15] See Azoulay et al. (2011) for more details and an evaluation of this program.

Appendix B: Measuring Publication Data

The source of our publication data is PubMed, a bibliographic database maintained by the U.S. National Library of Medicine that is searchable on the web at no cost.[16] PubMed contains over 14 million citations from 4,800 journals published in the United States and more than 70 other countries from 1950 to the present. The subject scope of this database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. In order to effectively mine this publicly available data source, we designed PubHarvester, an open-source software tool that automates the process of gathering publication information for individual life scientists (see Azoulay et al. 2006 for a complete description of the software). PubHarvester is fast, simple to use, and reliable. Its output consists of a series of reports that can be easily imported by statistical software packages.

This software tool does not obviate the two challenges faced by empirical researchers when attempting to accurately link individual scientists with their published output. The first relates to what one might term "Type I Error," whereby we mistakenly attribute to a scientist a journal article actually authored by a namesake; The second relates to "Type II Error," whereby we conservatively exclude from a scientist's publication roster legitimate articles:

Namesakes and Popular Names: PubMed does not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials, and an optional suffix. This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Inconsistent Publication Names: The opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffxes (Jr., Sr., 2nd, etc.), or from multiple patronyms due to changes in spousal status.

To deal with these serious measurement problems, we opted for a labor-intensive approach: the design of individual search queries that relies on relevant scientific keywords, the names of frequent collaborators, journal names, as well as institutional affiliations. We are aided in the time-consuming process of query design by the availability of a reliable archival data source, namely, these scientists' CVs and biosketches. PubHarvester provides the option to use such custom queries in lieu of a completely generic query (e.g, "azoulay p"[au] or "graff zivin js"[au]).

As an example, one can examine the publications of Scott A. Waldman, an eminent pharmacologist located in Philadelphia, PA at Thomas Jefferson University. Waldman is a relatively frequent name in the United States (with 208 researchers with an identical patronym in the AAMC Faculty Roster); the combination "waldman s" is common to 3 researchers in the same database. A simple search query for "waldman sa"[au] OR "waldman s"[au] returns 377

[16] http://pubmed.gov/

publications at the time of this writing. However, a more refined query, based on Professor Waldman's biosketch returns only 256 publications.[17]

The above example also makes clear how we deal with the issue of inconsistent publication names. PubHarvester gives the end-user the option to choose up to four PubMed-formatted names under which publications can be found for a given researcher. For example, Louis J. Tobian, Jr. publishes under "tobian l", "tobian l jr", and "tobian lj", and all three names need to be provided as inputs to generate a complete publication listing. Furthermore, even though Tobian is a relatively rare name, the search query needs to be modified to account for these name variations, as in ("tobian l"[au] OR "tobian lj"[au]).

[17] ((((("waldman sa"[au] NOT (ether OR anesthesia)) OR ("waldman s"[au] AND (murad OR philadelphia[ad] OR west point[ad] OR wong p[au] OR lasseter kc[au] OR colorectal))) AND 1980:2013[dp])

Appendix C: Defining Peers - PubMed Related Citations Algorithm [PMRA]

Traditionally, it has been very difficult to assign to individual scientists, or articles, a fixed address in "idea space," but such a measure is critical in order to meaningfully assess the quality of peer environments at origin and potential destination institutions and thus the push and pull of match quality as a driver of moving.

This challenge is met here by the use of the PubMed Related Citations Algorithm [PMRA], a probabilistic, topic-based model for content similarity that underlies the \related articles" search feature in PubMed. This database feature is designed to help a typical user search through the literature by presenting a set of records topically related to any article returned by a PubMed search query.[18] To assess the degree of intellectual similarity between any two PubMed records, PMRA relies crucially on MeSH keywords. MeSH is the National Library of Medicine's [NLM] controlled vocabulary thesaurus. It consists of sets of terms arranged in a hierarchical structure that permit searching at various levels of specificity. There are 27,149 descriptors in the 2013 MeSH edition. Almost every publication in PubMed is tagged with a set of MeSH terms (between 1 and 103 in the current edition of PubMed, with both the mean and median approximately equal to 11). NLM's professional indexers are trained to select indexing terms from MeSH according to a specific protocol, and consider each article in the context of the entire collection (Bachrach and Charen 1978; Neveol et al. 2010). What is key for our purposes is that the subjectivity inherent in any indexing task is confined to the MeSH term assignment process and does not involve the articles' authors.[19]

Using the MeSH keywords as input, PMRA essentially defines a distance concept in idea space such that the proximity between a source article and any other PubMed-indexed publication can be assessed. The following paragraphs were extracted from a brief description of PMRA:

The neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. To carry out such a program, one must first define what a word is. For us, a word is basically an unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, new lines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next, a limited amount of stemming of words is done, but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles are added in a second time to give them a small advantage in the local weighting scheme. MeSH terms are placed in a third category, and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. If a MeSH term is starred (indicating a major concept in a document), the star is ignored. These three categories of words (or phrases in the case of

[18] Lin and Wilbur (2007) report that one fifth of "non-trivial" browser sessions in PubMed involve at least one invocation of PMRA.

[19] This is a slight exaggeration: PMRA also makes use of title and abstract words to determine the proximity of any two pairs of articles in the intellectual space. These inputs are obviously selected by authors, rather than by NLM staff. However, neither the choice of MeSH keywords nor the algorithm depends on cited references contained in publications.

MeSH) comprise the representation of a document. No other fields, such as Author or Journal, enter into the calculations.

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used, it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) the number of times the term occurs in a particular document; and 3) the number of term occurrences in the document. The first of these pieces of information is used to produce a number called the global weight of the term.

The global weight is used in weighting the term throughout the database. The second and third pieces of information pertain only to a particular document and are used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents, its weight is computed as the product of the global weight times the two local weights (one pertaining to each of the documents). The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about a document. On the other hand, a term that occurred in only 100 documents of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only 10 documents is likely to be even more informative and will receive an even higher weight.

The local weight of a term is the measure of its importance in a particular document. Generally, the more frequent a term is within a document, the more important it is in representing the content of that document. However, this relationship is saturating, i.e., as the frequency continues to go up, the importance of the word increases less rapidly and finally comes to a finite limit. In addition, we do not want a longer document to be considered more important just because it is longer; therefore, a length correction is applied.

The similarity between two documents is computed by adding up the weights of all of the terms the two documents have in common. Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are pre-computed for each document in PubMed so that when one selects Related Articles, the system has only to retrieve this list. This enables a fast response time for such queries.[20]

The algorithm uses a cut-off rule to determine the number of related citations associated with a given source article. First, the 100 most related records by similarity score are returned. Second, a reciprocity rule is applied to this list of 100 records: if Publication A is related to

[20] Available at http://ii.nlm.nih.gov/MTI/related.shtml

Publication B, Publication B must also be related to publication A. As a result, the set of related citations for a given source article may contain many more than 100 publications.[21]

Given our set of source articles, we delineate the scientific fields to which they belong by focusing on the set of articles returned by PMRA that satisfy three additional constraints: (i) they are original articles (as opposed to editorials, comments, reviews, etc.); (ii) they were published in or before 2006 (the end of our observation period); and (iii) they appear in journals indexed by the Web of Science (so that follow-on citation information can be collected).

To summarize, PMRA is a modern implementation of co-word analysis, a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., title words or phrases, or keywords) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these texts (Callon et al. 1989; He 1999). One long-standing concern among practitioners of this technique has been the "indexer effect" (Whittaker 1989). Clustering algorithms such as PMRA assume that the scientific corpus has been correctly indexed. But what if the indexers who chose the keywords brought their own "conceptual baggage" to the indexing task, so that the pictures that emerge from this process are more akin to their conceptualization than to those of the scientists whose work it was intended to study?

Indexer effects could manifest themselves in three distinct ways. First, indexers may have available a lexicon of permitted keywords which is itself out of date. Second, there is an inevitable delay between the publication of an article and the appearance of an entry in PubMed. Third, indexers, in their efforts to be helpful to users of the database, may use combinations of keywords that reflect the conventional views of the field. The first two concerns are legitimate, but probably have only a limited impact on the accuracy of the relationships between articles that PMRA deems related. This is because the NLM continually revises and updates the MeSH vocabulary, precisely in an attempt to neutralize keyword vintage effects. Moreover, the time elapsed between an article's publication and the indexing task has shrunk dramatically, though time lag issues might have been a first-order challenge when MeSH was created, back in 1963. The last concern strikes us as being potentially more serious; a few studies have asked authors to validate ex post the quality of the keywords selected by independent indexers, with generally encouraging results (Law and Whittaker 1992). Inter-indexer reliability is also very high (Wilbur 1998).

[21] The effective number of related articles returned by PMRA varies between 58 and 2,097 in the sample of 3,074 source articles published by the 452 star scientists in the five years preceding their death. The mean is 185 related articles, and the median 141.

Appendix D: OLS and Fixed Effect Results

In this Appendix, we reproduce all of our main results using a linear probability (OLS) and fixed effects (where appropriate) framework as described in Section III. They are labeled Tables A4 – A9 to mirror those labeled Tables 4 – 9 in the main body of the paper.

Table A4. Determinants of Mobility: Demographic, Productivity & Peer Measures (OLS)

	Movers + Stayers [N=9,389]				Subsample w/Kid Info [N=2,960]
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	-0.001	-0.001	-0.000	-0.000	-0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
PhD (MD omitted)	-0.001	-0.004**	-0.006**	-0.006**	-0.007**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
MD/PhD	-0.001	-0.004**	-0.005**	-0.004*	-0.004
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)
Productivity Measures					
Ln(Pubs_t-1)		-0.000	0.001	0.001	-0.000
		(0.001)	(0.001)	(0.001)	(0.001)
Ln(Stock Pubs_t-2)		0.003**	0.005**	0.004**	0.005**
		(0.001)	(0.001)	(0.001)	(0.001)
Ln(NIH Funding_t-1)		-0.001**	-0.001**	-0.001**	-0.001**
		(0.000)	(0.000)	(0.000)	(0.000)
Ln(Stock NIH		0.001**	0.001**	0.000**	0.000
Funding_t-2)		(0.000)	(0.000)	(0.000)	(0.000)
Collaborators					
Ln(Pubs), Colocated			-0.004**	-0.002**	-0.001*
			(0.000)	(0.000)	(0.001)
Ln(Pubs), Close			-0.002**	-0.002**	-0.001*
			(0.000)	(0.000)	(0.001)
Ln(Pubs), Distant			0.002**	0.001**	0.001*
			(0.000)	(0.000)	(0.001)
Non-collaborating Peers					
Ln(Pubs), Colocated				-0.007**	-0.006**
				(0.000)	(0.001)
Ln(Pubs), Close				-0.000	-0.000
				(0.000)	(0.000)
Ln(Pubs), Distant				0.005**	0.004**
				(0.001)	(0.001)
Nb. of Observations	211,161	190,583	190,541	190,541	66,462
Nb. of Job Transitions	10,289	10,289	10,289	10,289	3,298
Nb. of Scientists	9,389	9,389	9,389	9,389	2,960
R2	0.006	0.007	0.008	0.011	0.011

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located at least 50 miles away. Estimation is by Ordinary Least Squares (OLS). All specifications include full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A5. Determinants of LOCAL Mobility: Demographic, Productivity & Peer Measures (OLS)

	Local Movers + Stayers [N=6,637]				Subsample w/Kid Info [N=2,025]
	(1)	(2)	(3)	(4)	(5)
Demographics					
Female	0.003**	0.003**	0.003**	0.003**	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PhD (MD omitted)	-0.002**	-0.002**	-0.002**	-0.002**	-0.003**
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
MD/PhD	-0.002*	-0.002**	-0.002*	-0.002*	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Productivity Measures					
Ln(Pubs_t-1)		-0.000	-0.000	0.000	0.001
		(0.000)	(0.000)	(0.000)	(0.001)
Ln(Stock Pubs_t-2)		0.000	0.000	0.001	0.002*
		(0.000)	(0.000)	(0.000)	(0.001)
Ln(NIH Funding_t-1)		-0.000*	-0.000*	-0.000*	-0.000*
		(0.000)	(0.000)	(0.000)	(0.000)
Ln(Stock NIH		0.000	0.000	-0.000	-0.000
Funding_t-2)		(0.000)	(0.000)	(0.000)	(0.000)
Collaborators					
Ln(Pubs), Colocated			-0.001**	-0.001**	-0.000
			(0.000)	(0.000)	(0.000)
Ln(Pubs), Close			0.002**	0.000+	0.001*
			(0.000)	(0.000)	(0.000)
Ln(Pubs), Distant			-0.000+	-0.000	-0.001+
			(0.000)	(0.000)	(0.000)
Non-collaborating Peers					
Ln(Pubs), Colocated				-0.002**	-0.002**
				(0.000)	(0.000)
Ln(Pubs), Close				0.002**	0.002**
				(0.000)	(0.000)
Ln(Pubs), Distant				-0.000	-0.001
				(0.000)	(0.001)
Nb. of Observations	173,120	159,676	159,638	159,638	55,286
Nb. of Job Transitions	6,722	6,722	6,722	6,722	2,061
Nb. of Scientists	6,637	6,637	6,637	6,637	2,025
R2	0.002	0.002	0.003	0.006	0.007

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Distant movers (scientists moving more than 50 miles away) are excluded from this analysis. Estimation is by Ordinary Least Squares (OLS). All specifications include full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A6. Determinants of Mobility: Child Age (OLS and OLS with Fixed Effects)

	(1)	(2)	(3)	(4)	(5)
A. OLS					
Oldest kid 12 or 13	0.013**		0.013**		
	(0.003)		(0.003)		
Youngest kid 18 or 19		0.010**	0.010**		
		(0.003)	(0.003)		
Number of kids in high school				-0.007**	
				(0.001)	
At least one kid in high school					-0.011**
					(0.002)
B. OLS with Fixed Effects					
Oldest kid 12 or 13	0.017**		0.018**		
	(0.003)		(0.003)		
Youngest kid 18 or 19		0.009**	0.009**		
		(0.003)	(0.003)		
Number of kids in high school				-0.007**	
				(0.001)	
At least one kid in high school					-0.010**
					(0.002)
Nb. of Observations	66,462	66,462	66,462	66,462	66,462
Nb. of Job Transitions	3,298	3,298	3,298	3,298	3,298
Nb. of Scientists	2,960	2,960	2,960	2,960	2,960

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located at least 50 miles away. Estimation is by Ordinary Least Squares (OLS). All specifications include individual productivity and peer variables, as well as full age, vintage category, and year fixed effects. Robust standard errors are in parentheses, clustered at the individual level. Panel B regressions include scientist-spell fixed effects. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A7. Determinants of LOCAL Mobility: Child Age (OLS and OLS with Fixed Effects)

	(1)	(2)	(3)	(4)	(5)
A. OLS					
Oldest kid 11, 12, or 13	-0.002^{+}		-0.002^{+}		
	(0.001)		(0.001)		
Youngest kid 18, 19, or 20		0.001	0.000		
		(0.001)	(0.001)		
Number of kids in high school				0.001	
				(0.001)	
At least one kid in high school					0.001
					(0.001)
B. OLS with Fixed Effects					
Oldest kid 11, 12, or 13	-0.003**		-0.003**		
	(0.001)		(0.001)		
Youngest kid 18, 19, or 20		0.001	0.000		
		(0.001)	(0.001)		
Number of kids in high school				0.000	
				(0.001)	
At least one kid in high school					0.001
					(0.001)
Nb. of Observations	55,286	55,286	55,286	55,286	55,286
Nb. of Job Transitions	2,061	2,061	2,061	2,061	2,061
Nb. of Scientists	2,025	2,025	2,025	2,025	2,025

Notes: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Distant movers (scientists moving more than 50 miles away) are excluded from this analysis. Estimation is by Ordinary Least Squares (OLS). All specifications include individual productivity and peer variables, as well as full age, vintage category, and year fixed effects. Panel B regressions include scientist-spell fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A8. Moves Up vs. Down: Demographic, Professional Factors

	OLS		OLS w/FE	
	Moves Up	Moves Down	Moves Up	Moves Down
	(1)	(2)	(3)	(4)
Demographics				
Female	-0.001	-0.002		
	(0.001)	(0.001)		
PhD	-0.003**	-0.003**		
	(0.001)	(0.001)		
MD/PhD	-0.003*	-0.001		
	(0.001)	(0.002)		
Productivity Measures				
Ln(Pubs_t-1)	0.000	0.001	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Ln(Stk Pubs_t-2)	-0.001	0.003**	0.003*	0.006**
	(0.001)	(0.001)	(0.001)	(0.001)
Ln(NIH Funding_t-1)	-0.000*	-0.000+	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Ln(Stk NIH Funding_t-2)	-0.000	0.000	0.000	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
Collaborators				
Ln(Pubs), Colocated	-0.001+	-0.001+	-0.001*	-0.000
	(0.000)	(0.000)	(0.000)	(0.001)
Ln(Pubs), Close	0.000	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.001)
Ln(Pubs), Distant	0.001+	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.001)
Non-collaborating Peers				
Ln(Pubs), Colocated	-0.002**	-0.002**	-0.002**	-0.001*
	(0.000)	(0.000)	(0.000)	(0.001)
Ln(Pubs), Close	-0.000	-0.000	-0.002**	-0.002**
	(0.000)	(0.000)	(0.001)	(0.001)
Ln(Pubs), Distant	0.002**	0.001+	-0.001	-0.004**
	(0.001)	(0.001)	(0.001)	(0.001)
Ln(Pubs), Colocated	-0.002**	-0.002**	-0.002**	-0.001*
	(0.000)	(0.000)	(0.000)	(0.001)
Nb. of Observations	47,625	48,779	47,625	48,779
Nb. of Job Transitions	1,815	1,953	1,815	1,953
Nb. of Scientists	1,813	1,943	1,813	1,943
R2	0.017	0.007	0.179	0.183

Note: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Moves up or down are based on a ranking of institutions by percentiles of total NIH funding received (per grantee). The sample includes scientists who moved to an institution that was 10 percentiles or more higher (lower) in the ranking of NIH funding and scientists who did not move. This means that scientists who moved to an institution that was less than 10 percentiles different (i.e. lateral movers) and scientists who moved down (up) are excluded. Estimation is by Ordinary Least Squares (OLS). All regressions include full age, vintage category, and year fixed effects and controls for percentile of origin institution. Columns 3 and 4 include scientist-spell fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A9. Moves Up vs. Down: Child Age

	Moves Up		Moves Down	
	(1)	(2)	(3)	(4)
A. OLS				
Number of kids in high school	-0.0016*		-0.0018*	
	(0.0007)		(0.0008)	
At least one kid in high school		-0.0024*		-0.0036**
		(0.0010)		(0.0010)
Percentile of Origin Institution	-0.0006**	-0.0006**	0.0003**	0.0003**
	(0.0001)	(0.0001)	(0.0000)	(0.0000)
B. OLS with Fixed Effects				
Number of kids in high school	-0.0010		-0.0014+	
	(0.0007)		(0.0008)	
At least one kid in high school		-0.0014		-0.0029**
		(0.0010)		(0.0011)
Percentile of Origin Institution	0.0000	0.0000	0.0002**	0.0002**
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Nb. of Observations	47,625	47,625	48,779	48,779
Nb. of Job Transitions	1,815	1,815	1,953	1,953
Nb. of Scientists	1,813	1,813	1,943	1,943

Note: The dependent variable is a binary variable that takes on a value one in the year we observe the elite scientist moving to a new academic position located within 50 miles. Moves up or down are based on a ranking of institutions by percentiles of total NIH funding received (per grantee). The sample includes scientists who moved to an institution that was 10 percentiles or more higher (lower) in the ranking of NIH funding and scientists who did not move. This means that scientists who moved to an institution that was less than 10 percentiles different (i.e. lateral movers) and scientists who moved down (up) are excluded. Estimation is by Ordinary Least Squares (OLS). All regressions include full age, vintage category, and year fixed effects and controls for percentile of origin institution. Panel B regressions include scientist-spell fixed effects. Robust standard errors are in parentheses, clustered at the individual level. See Section II for a full description of how the sample and variables were constructed. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$