Age Dynamics in Scientific Creativity: Supporting Information

Benjamin F. Jones[1,3], Bruce A. Weinberg[2,3]

1. Kellogg School of Management, Northwestern University, Evanston IL 50208 USA
2. Department of Economics, Ohio State University, Columbus OH 43210 USA
3. National Bureau of Economic Research, Cambridge, MA 02138 USA

October 2011

Nobel Prize Data

One advantage of studying Nobel Prize winners is the wealth of information available. The Nobel Foundation's website, nobelprize.org, is a particularly rich source of data. We collected data on dates of birth, the highest earned degree, the year or range of years in which each laureate's prize-winning work was performed, and whether the work contained an important theoretical component. We were able to obtain dates of birth for 526 of the 528 Nobel Prize winners (99.4%), and the period of key research for all but 1. People who received more than one prize were included for their first prize. In cases where the Nobel Foundation's web-site did not accurately identify the year or period of key research, other sources were consulted, including: (1) Schlessinger, B. and Schlessinger, J. *The Who's Who of Nobel Prize Winners*, 1901-1995. Oryx Press, Phoenix AZ 1996; (2) Daintith, J. and Gjertsen, D. *The Grolier Library of Science Biographies*. Vols. 1-10. Grolier Educational, Danbury CT 1996; (3) Debus, A.G. ed. *World Who's Who in Science: A Biographical Dictionary of Notable Scientists from Antiquity to the Present*. Marquis Who's Who Inc., Chicago 1968; (4) Kragh, H. *Quantum Generations: a History of Physics in the Twentieth Century*. Princeton: Princeton UP, 1999. (5) McMurray, E.J., Kosek, J.K., and Valade, R.M. *Notable Twentieth-Century Scientists*. Vols. 1-4. Gale Research, Detroit 1995; (6) Williams, T.I. ed. *Biographical Dictionary of Scientists*. John Wiley and Sons, New York 1974.

If, after analyzing these sources, additional information was required, individual biographies were consulted. When a range of years was identified as being the most important period, we consulted the Science Citation Index to identify the year in which the single most important contribution was made. Where a single year could not be identified, the estimates use the middle year of the research period to define the age at great achievement. The three measures are closely related, with the correlation between our middle years and early years being .998 and the correlation between our middle years and late years being .997.

Kragh [1999] also identifies the years (or range of years) in which physicists do their prize-winning work. The correlation between our work year and year he identifies (or the midpoint if he specifies a range) is .995. Stephan and Levin (1993) have collected data on the year in which Nobel laureates in all three fields began and stopped working on the broad research agenda for which they received the Nobel Prize. The correlation between our work years on the one hand and their beginning and ending years on the other, are 0.969 and 0.974 despite the difference in constructs (we focus on when people did the specific work for which they received the Nobel Prize, whereas Stephan and Levin focus on when the broad research agenda begins and ends).

To assess the extent to which each laureate's prize-winning work was deductive versus inductive, we determine whether the work had an important theoretical component. This classification was done using the biographical sources (discussed above). Kraugh [1999] also classifies the physics laureates, and we reconciled individual cases against his classification. In classifying research, we identified whether their primary contribution was empirical, theoretical, or both empirical and theoretical. Works were classified as having an important theoretic component if their primary contribution was theoretical or if it combined theoretical and empirical work, (only 21 of the 525 laureates in our sample are classified as having received the prize for a combination of theoretical and empirical work).

Century of Science and Web of Science Data

We use Thomson Reuters' Institute for Scientific Information (ISI) Web of Science and Century of Science databases providing coverage from 1900 to the present. The Web of Science database, which we use from 1955 to the present, indexes 20 million articles. The Century of Science database indexes a smaller sample of articles, indexing the journals from the early 20th century that contained preeminent scientific contributions.

To analyze citation age, we consider the top 100 papers in each year over the 20th century in each of the three Nobel fields and in an "other" category comprising all other fields of science and engineering. For each paper, we calculate the mean duration between the paper's publication year and the publication years of all the papers the given paper cites. In analyzing the dynamics, we calculate these citation ages for four fields:

(i) Physics, defined as the those papers which ISI assigns to field categories "physics, applied", "physics, condensed matter", and "physics, multidisciplinary"

(ii) Chemistry, defined as ISI field categories "chemistry, analytical", "chemistry, applied", "chemistry, inorganic", "chemistry, medicinal", "chemistry, multidisciplinary", "chemistry, organic", "chemistry, physical"

(iii) Medicine, defined as ISI field categories "anatomy and morphology", "biochemistry and molecular biology", "cardiac and cardiovascular system", "cell biology", "clinical neurology", "dermatology", "endocrinology", "genetics", "immunology", "medical laboratory technology", "medicine general and internal", "medicine research", "neurosciences", "nutrition", "obstetrics and gynecology", "ophthalmology", "orthopedics", "otorhinolaryngology", "pathology", "pediatrics", "pharmacology", "psychology", "radiology", "surgery", "urology", "psychiatry", "psychology, experimental", "psychology, multidisciplinary" (Note that the medicine category, like the Nobel Prize in that discipline, encompasses a wide variety of areas.)

(iv) Other, which is the other 133 ISI field categories within Science and Engineering.

The analysis considers the deviation between a paper's mean citation age and the mean citation age for the "Other" category in that publication year, divided by the standard deviation of the mean citation age for the "Other" category in the publication year. This method purges the citation age dynamics from the background trends in citations over the 20th century and puts the deviations on a common scale. Formally, our measure for the age of citations in field *f* at time *t* is

$$CiteAge_{ft} = \frac{\frac{1}{N_{ft}}\sum_i CiteAge_{ift} - \mu_{Ot}^{CiteAge}}{\sigma_{Ot}^{CiteAge}},$$

where $CiteAge_{ift}$ gives the mean age of the citations in paper i in field f at time t; N_{ft} denotes the number of top papers (100 in our analysis) from field f at time t; and $\mu_{Ot}^{CiteAge}$ and $\sigma_{Ot}^{CiteAge}$ give the mean and standard deviation of citation ages of the papers in the other field category in year t.

Although related, our measure differs from a citation half-life insofar as half-lives measure durability using forward citations, whereas our measure captures reliance on previous work using backward citations. Our measure is also distinct from conventional citation metrics for research performance in that it measures the amount of foundational knowledge in a field at a point in time as opposed to identifying important papers or researchers (e.g., the H-index).

The regressions in Supporting Table 4 show the dynamics in citation age with and without author fixed effects in the regression. To construct author identifiers, we employ the author name information employed in the Century of Science and Web of Science databases. We create individual author identifiers as a unique name (last name and first initial) in the given field (physics, chemistry, medicine, and other) for the top papers in each field and year. The regressions in Supporting Table 4 include only those authors that appear at least twice in the sample – i.e. produce at least two of the mostly highly cited papers. Inclusion of name fixed effects eliminates systematic differences between individuals, to focus the citation dynamics within scientists' careers. Thus, these estimates identify whether individuals themselves are shifting their behavior (i.e. the field is changing) as opposed to citation dynamics driven by a shifting set of individuals in the field.

Population Data

We estimate the age distribution for subsets of the US population, using data from the Census IPUMS (Steven Ruggles and Matthew Sobek et. al. *Integrated Public Use Microdata Series: Version 2.0* Minneapolis: Historical Census Projects, University of Minnesota, 1997). We use the 1% samples for 1870, 1880, and 1900-2000 (no samples are available for 1890; for 1970, we use the Form 1 State Sample). Person weights are used with the 1940 and 1950 samples, which are weighted samples (the samples for the other years are unweighted / flat samples). We interpolate population shares linearly between the census years. (For year t, between census years t_0 and t_1, we estimate $\hat{\rho}(Age|t) = \frac{t_1 - t}{t_1 - t_0}\rho(Age|t_0) + \frac{t - t_0}{t_1 - t_0}\rho(Age|t_1)$, where $\rho(Age|t)$ denotes the share of the population at time t that is Age years old.) We have one observation in 2001 and linearly extrapolate using data for 1990 and 2000 according to $\hat{\rho}(Age|2001) = 1.1\rho(Age|2000) - .1\rho(Age|1990)$.

Our population subsets are: (1) the entire population; (2) the employed population (labforce=2); (3) people employed in professional and technical occupations (labforce=2 and occ1950 between 0 and 100); (4) people employed as natural scientists, engineers, or physicians (labforce=2 and occ1950 equal to 007, 012-026, 401, 49, 61-69, or 75); and (5) people employed as natural scientists or engineers (labforce=2 and occ1950 equal to 007, 012-026, 401, 49, or 61-69).

Supporting Table 1: Summary Statistics for the Nobel Laureates

This table presents summary statistics for the Nobel laureates. Standard deviations are given in parentheses.

	All		Chemistry		Medicine		Physics	
Mean Age of Prize-Winning Research	39.0	(8.54)	40.2	(8.24)	39.9	(7.86)	37.2	(9.20)
Mean Age of Highest Degree	26.1	(3.42)	25.5	(3.22)	26.5	(3.56)	26.2	(3.37)
Frequency of Prize-Winning Work with Important Theoretical Component	.185	(.388)	.190	(.393)	.074	(.262)	.297	(.458)
Frequency of Prize-Winning Research by Age 30	.124	(.330)	.092	(.289)	.079	(.270)	.178	(.399)
Frequency of Prize-Winning Research by Age 40	.564	(.496)	.490	(.502)	.537	(.500)	.654	(.477)
Frequency of Highest Degree by Age 25	.350	(.478)	.399	(.491)	.305	(.462)	.357	(.480)
Mean Year of Prize-Winning Work	1947	(28.2)	1948	(29.2)	1947	(27.3)	1947	(28.5)
Observations	525		153		190		182	

Supporting Figure 1: Underlying Data and Additional Estimates of Dynamics

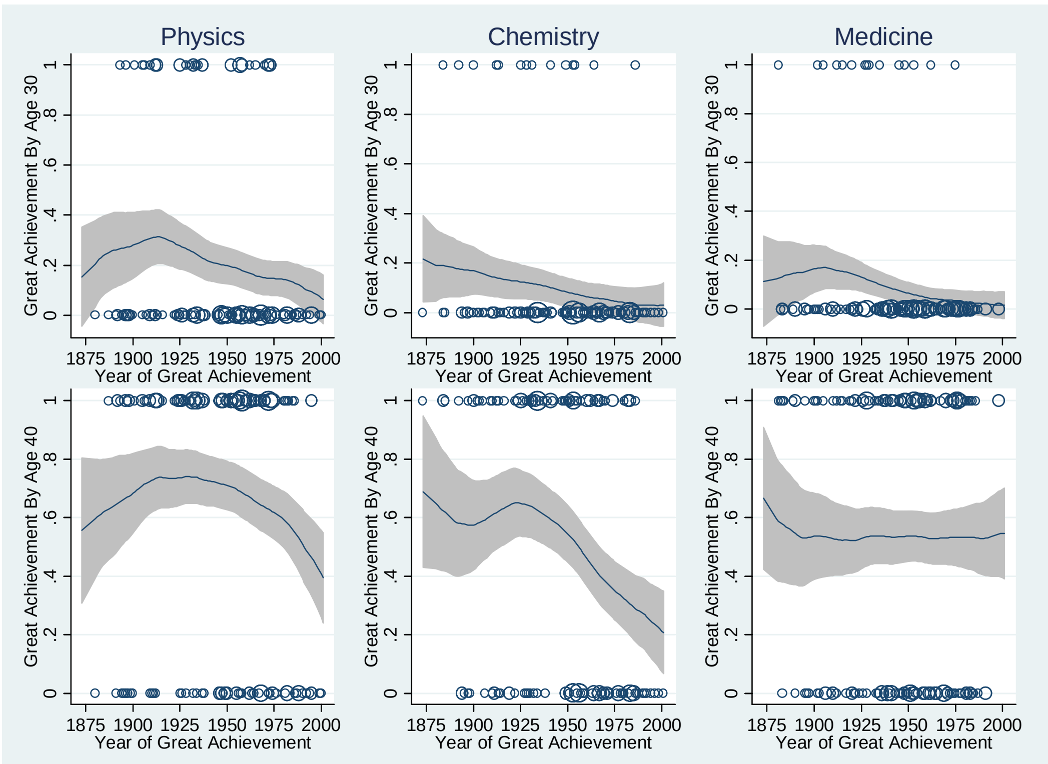
This section presents our underlying data and further examines dynamics in the age at great achievement, in theoretical work, and in foundational knowledge. We also reproduce our estimates using kernel regressions as a further robustness check. The fractional polynomial regressions used in the text are a global estimator where the functional form that is chosen to match the data is determined by all observations. Kernel regressions are a local estimator, providing estimates at a given point in time based only on the data in a neighborhood around that time. In the case of the age of great achievement, for a time *t* and a bandwidth *h*, the predicted age at *t* is a weighted average of the ages within a radius *h* of *t*. Formally, $\hat{Age}_h(t) = \frac{\sum_{i=1}^{N} K_h(t - t_i) Age_i}{\sum_{i=1}^{N} K_h(t - t_i)}$, where t_i denotes the time at which laureate *i* made his or her prize winning contribution; Age_i is a measure of laureate *i*'s age (e.g. below 30 or 40 or age measured continuously) at the time of his or her prize-winning contribution; and $K_h(t - t_i)$ denotes the weight applied to observations that are a distance $t - t_i$ from *t*. The numerator gives a weighted sum of observations and the denominator gives the sum of the weights, so the estimator is a weighted average, with the weights declining from the point in question according to the kernel. We use the standard Epanechnikov kernel, defined as

$$K_h(\delta) = \frac{3}{4h}\left(1 - \left(\frac{\delta}{h}\right)^2\right) I\left(\left|\frac{\delta}{h}\right| \leq 1\right)$$

and a bandwidth of 15 years. Analogous procedures are used for the other variables.

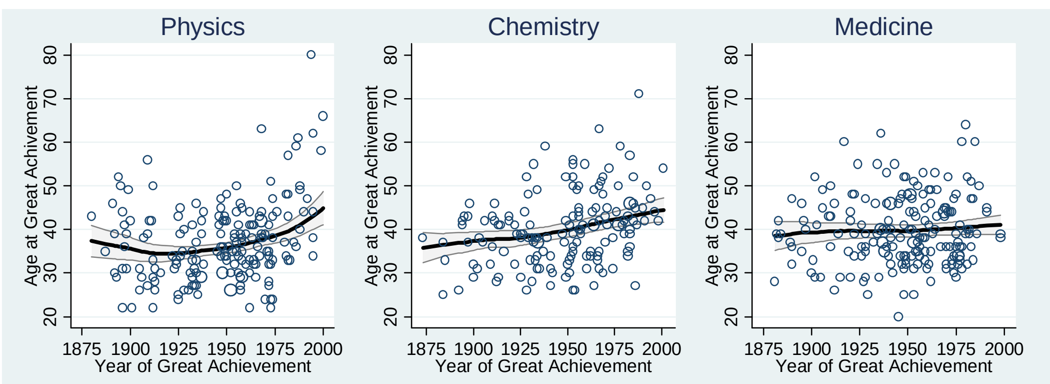
Estimates of the probability that work is done before ages 30 and 40 (and 95% confidence intervals) are shown in Supporting Figure 1A. The figure also shows the underlying binary indicator for whether each laureate was at least age 30 or 40 at the time of his or her prize-winning contribution (1 = above the age threshold). In the case that multiple laureates do prize-winning work above or below the age threshold in the same year, the circles are scaled in proportion to the number of people they represent. The dynamics using this non-parametric approach show the same core features as the fractional polynomial method. Physics shows hump-shaped patterns similar to the fractional polynomial estimates. For chemistry, the under 30 propensity declines steadily to zero, while the under 40 propensity fluctuates before declining for most of the period. For medicine the under 30 pattern has a small initial increase and then declines to zero, while the under 40 pattern is flatter, showing some convexity.

Supporting Figure 1A. Kernel Estimates of Trends in Age at Great Achievement by Ages 30 and 40.



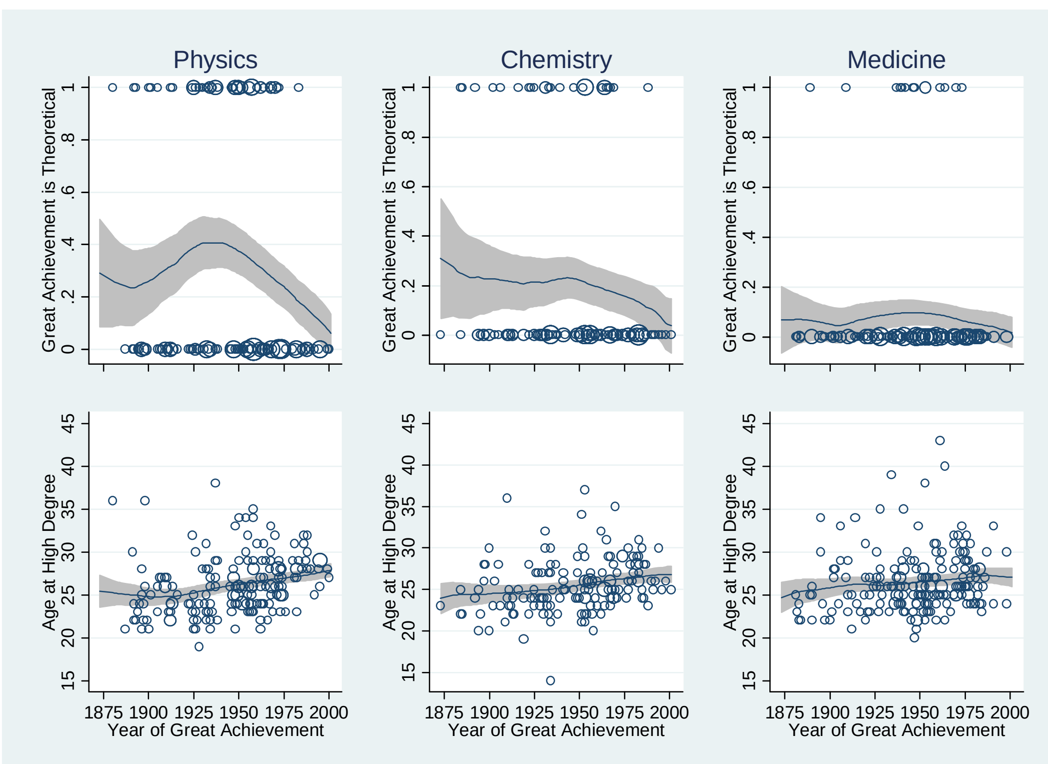
To further examine the age dynamics and enable the reader to see our underlying data, Supporting Figure 1B reports kernel estimates (and 95% confidence intervals) treating age as a continuous variable. The underlying data is also shown (with circles scaled in proportion to the number of observations they represent). As discussed in the text, most of the variation in the age at which people do their Prize-winning work is idiosyncratic at the level of the individual (i.e. within a field at a given point in time), but there are strong trends in ages within each field and these are quite consistent with the other estimation approaches. Ages are hump shaped in Physics, with a global minimum in the 1920s. Chemistry shows a steady increase in ages, while medicine is quite flat. See also Jones (2010) for non-parametric mean age analysis.

Supporting Figure 1B. Kernel Estimates of Trends in Mean Age at Great Achievement.



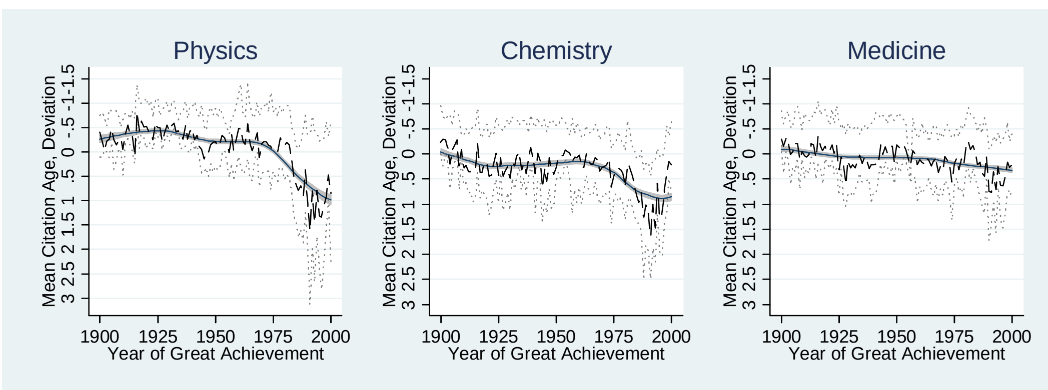
Supporting Figure 1C reports kernel estimates of the frequency of theoretical work (top panel) and the age at high degree (bottom panel) and 95% confidence intervals. The procedures follow those described above. The estimates are again quite similar to those reported in the text. For the frequency of theoretical work, physics shows a hump shape; chemistry is flat initially and then declines; and medicine is quite low, with a slight hump. For the age at high degree, following the analysis in Jones (2010), physics shows a U-shape; and both chemistry and medicine decline. The underlying data clearly show the reduction in high degrees before age 25 by the end of the period, especially in physics and chemistry.

Supporting Figure 1C. Kernel Estimates of Frequency of Theoretical Work and Age at High Degree.



Supporting Figure 1D reports kernel estimates of backward citation ages and 95% confidence intervals. The procedures follow those described above, but there are 100 observations per year in each field. The volume of data increases the precision of the estimates. To summarize the data, the figure plots the mean for each year (dashed line) and the 25th and 75th percentiles of the backward citation ages in each year (dotted lines), which give a sense of the dispersion in the data. Here too, the estimates are similar to those reported in the text (and, as in the text, we have inverted the axis.) Backward citation ages decrease in physics and then increase. Both chemistry and medicine show smaller increases in backward citation ages that are consistent with those reported in the text.

Supporting Figure 1D. Kernel Estimates of Trends in Backward Citation Ages.



Supporting Analysis: Controlling for the Age Distribution of the Population

Our main results examine the probability that prize-winning work is done by people beneath ages 30 and 40. In general, shifts in the age at great achievement can be due to productivity shifts across the life-cycle and/or demographic shifts in the underlying age distribution (10). This section shows that demographic shifts are too small to explain the dynamics in the share of young scientists doing Nobel Prize winning work.

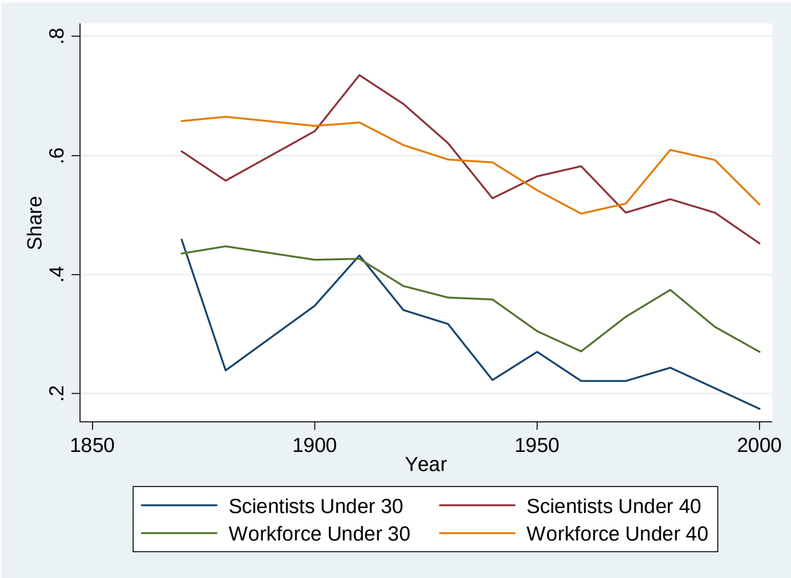
We outline our framework for the age 30 threshold (the age 40 case is directly analogous), building from (10). The probability that a prize winning contribution is made by someone under 30 is

$$\Pr\left[Age < 30 \middle| Contribution, t\right] = \frac{\int_0^{30} \Pr\left[Contribution \middle| Age, t\right] \Pr\left[Age \middle| t\right] dAge}{\int_0^{\infty} \Pr\left[Contribution \middle| Age, t\right] \Pr\left[Age \middle| t\right] dAge}.$$

Changes in the share of prize-winning contributions made by people under 30 may be due to changes in the probability that contributions are made by people of different ages, which we refer to as changes in the age-productivity relationship. These changes are represented by the function $\Pr\left[Contribution \middle| Age, t\right]$ shifting over time. Alternatively, shifts in shares of prize-winning contributions done before age 30 may be due to changes in the age distribution of the population, which we refer to as changes in the age distribution. These shifts are represented by the function $\Pr\left[Age \middle| t\right]$.

Supporting Figure 2 presents the share of scientists and engineers and the workforce under ages 30 and 40 from 1870 to 2000 in the United States. The data show a general decline in the share of scientists (and all workers) under 30 and 40. Notably the share of young scientists and engineers rises between 1880 and 1910 as US universities expand. The share of young workers also increases as the baby boom enters the labor market during the 1970s.

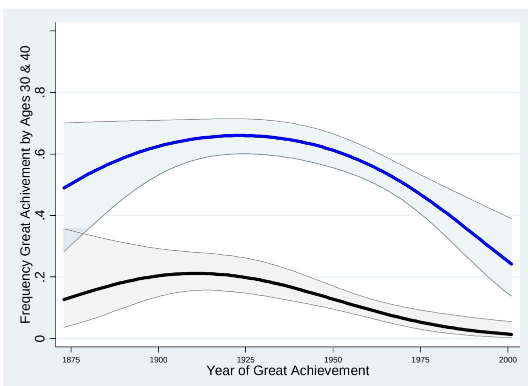
Supporting Figure 2: The Age Distribution of Scientists and Engineers and the Workforce in the United States



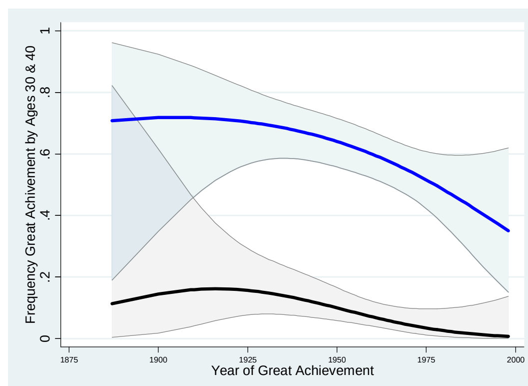
Supporting Figure 3 shows the share of Nobel laureates doing their prize winning work beneath ages 30 and 40 across all fields. Separate estimates are shown for people doing their prize winning work in the United States. The dynamics are quite similar, although only 4 of the 71 contributions made in or before 1910 were made in the United States, limiting the precision of the early age trends in the United States.

Supporting Figure 3: Age Dynamics for All Fields, World and USA.

World



USA



Examining these figure together, we see that the share of scientists and engineers under age 30 falls to 17.4% (Supporting Figure 2) whereas the share of people doing Nobel Prize winning work by age 30 falls to nearly zero across the three fields (Supporting Figure 3). Similarly, the share of the scientists and engineers under age 40 remains at 45.2% in 2000, also above the declining share of Nobel Prize winning achievements in that age range. The share of young scientists and engineers also increases in the 1970s following the post-war baby boom, yet great achievements by younger scholars become increasingly rare during this period, further suggesting that the aging phenomenon is not driven by such demographic shifts.

To formally estimate the extent to which trends in the age of great achievement are due to changes in the age-productivity relationship as opposed to changes in the age distribution of scientists, we parametrize the probability that contributions are made by people of different ages, $\Pr\left[Contribution \middle| Age, t\right]$ flexibly. We assume that,

$$\Pr\left[Contribution \middle| Age, t\right] \equiv \exp\left\{\alpha_0 Age \cdot \left(Year - \overline{Year}\right) + \alpha_1 Age + \alpha_2 Age^2\right\}.$$

Here α_1 and α_2 govern the shape of the age-productivity curve in the mean year of great achievement ($\overline{Year}$), which is 1957. The parameter α_2 is expected to be negative so that the age-productivity profile peaks at $Age = -\frac{\alpha_1 + \alpha_0\left(Year - \overline{Year}\right)}{2\alpha_2}$. The parameter α_0 governs shifts in the peak of the age-productivity curve over time, where the peak increases by $-\frac{\alpha_0}{2\alpha_2}$ per year. This simple formulation was chosen to minimize the number of parameters while still allowing for hump-shaped age-producitivy profiles and can be viewed as a simple approximation to an arbitrary function.

We estimate this model using maximum likelihood, searching over values of α_0, α_1, and α_2. The likelihood for observation *i* is

$$L^i = \frac{\exp\left\{\alpha_0 Age_i \cdot \left(Year_i - \overline{Year}\right) + \alpha_1 Age_i + \alpha_2 Age_i^{\,2}\right\}\rho\left(Age_i \middle| Year_i\right)}{\sum_{Age=0}^{90} \exp\left\{\alpha_0 Age \cdot \left(Year_i - \overline{Year}\right) + \alpha_1 Age + \alpha_2 Age^2\right\}\rho\left(Age \middle| Year_i\right)},$$

where Age_i denotes the age at which laureate *i* did his or her prize winning work; $Year_i$ denotes the year of the laureate's prize winning work; and $\rho\left(Age \middle| Year_i\right)$ gives the observed share of the population that is *Age* years old in $Year_i$. The log likelihood function is

$$\sum_{i=1}^{I} \ln\left(\frac{\exp\left\{\alpha_0 Age_i \cdot \left(Year_i - \overline{Year}\right) + \alpha_1 Age_i + \alpha_2 Age_i^{\,2}\right\}\rho\left(Age_i \middle| Year_i\right)}{\sum_{Age=0}^{90} \exp\left\{\alpha_0 Age \cdot \left(Year_i - \overline{Year}\right) + \alpha_1 Age + \alpha_2 Age^2\right\}\rho\left(Age \middle| Year_i\right)}\right)$$

where *I* gives the number of observations.

To implement this framework, we use population data for the United States from the Census IPUMS (described above) and data on people who did their Prize-winning work in the United States. We present 5 sets of estimates measuring the population in different ways. The population measures are (1) the entire population; (2) the employed population; (3) people employed in professional and technical occupations; (4) people employed as natural scientists, engineers, or physicians; and (5) people employed as natural scientists or engineers. Supporting Table 2 reports the results. The first column reports the implied annual change in the age at which the age-productivity profiles peak

$-\frac{\alpha_0}{2\alpha_2}$ (with the standard error of these estimates constructed using the delta method). The estimates indicate that the peak of the age-productivity profile increases by roughly 1 year per decade (.0971-.1362 years of age per calendar year). These estimates are quite precise and robust to the population measure. Thus, there is clear evidence that the probability that any given young person will do Nobel Prize winning work has declined over time and that the trends shown in the text are not due to changes in the age distribution of the population.

The previous estimates minimize the number of parameters that need to be estimated, but impose symmetry on the age-productivity profiles. To allow for an asymetric age-productivity profile, we have estimated models including a cubic in age,

$$\Pr\left[Contribution \middle| Age, t\right] \equiv \exp\left\{\alpha_0 Age \cdot \left(Year - \overline{Year}\right) + \alpha_1 Age + \alpha_2 Age^2 + \alpha_3 Age^3\right\}.$$

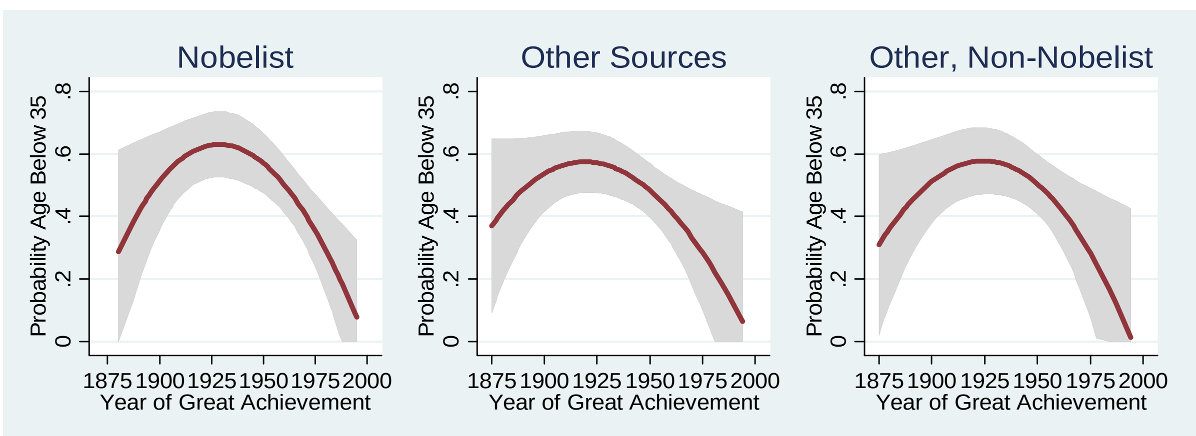
In addition to adding another parameter, including a cubic term implies that the rate of change in the peak of the age-productivity profiles changes over time, but the imputed trend is similar to those from the quadratic specification. When the science and engineering workforce is used as the population measure the peak of the age-productivity profile is imputed to increase, for example, by .0707 years per year that passes in 1957 (by .0660 years per year in 1937 and by .0767 years per year in 1977) compared to .1088 (S.E.=.0413) for the comparable quadratic specification.

Supporting Table 2. Maximum Likelihood Career Productivity Patterns.

	Implied Trend	α_0	α_1	α_2
		Population		
Estimate	.0971	.0015	.610	-.0077
Std. Err.	(.0311)	(.0006)	(.059)	(.0007)
		Employed		
Estimate	.1015	.0014	.566	-.0071
Std. Err.	(.0332)	(.0006)	(.0593)	(.0007)
		Professional Technical Occupations		
Estimate	.1087	.0015	.547	-.0067
Std. Err.	(.0350)	(.0006)	(.059)	(.0007)
		Natural Scientists, Engineers, Physicians		
Estimate	.1362	.0015	.446	-.0055
Std. Err.	(.0411)	(.0006)	(.059)	(.0007)
		Natural Scientists, Engineers		
Estimate	.1088	.0012	.462	-.0057
Std. Err.	(.0413)	(.0006)	(.060)	(.0007)

Supporting Figure 4: Age Dynamics for Physicists, Using Alternative Sources

Given the remarkable and unusual age dynamics in physics, we further explored the age pattern using alternative data sources to the Nobel Prize. To gather an alternative dataset, we considered numerous sources, described below, which collectively produced 160 famous physicists who did not win the Nobel Prize. Each graph below presents the evolution of the probability of great achievement by age 35. The leftmost graph uses the Nobelist data, as in the text. The middle graph uses the achievements defined by the alternative data sources, which include non-Nobelists and Nobelists. The rightmost graph uses only the 160 physicists who did not win the Nobel. We see that the dynamics are robust across data sources.



<u>Alternative Data Sources for Physicists</u>

1. Reinhardt, Joachim. *AIP Center for History of Physics*. 19 June 2007 <http://th.physik.uni-frankfurt.de/~jr/portraits.html>.

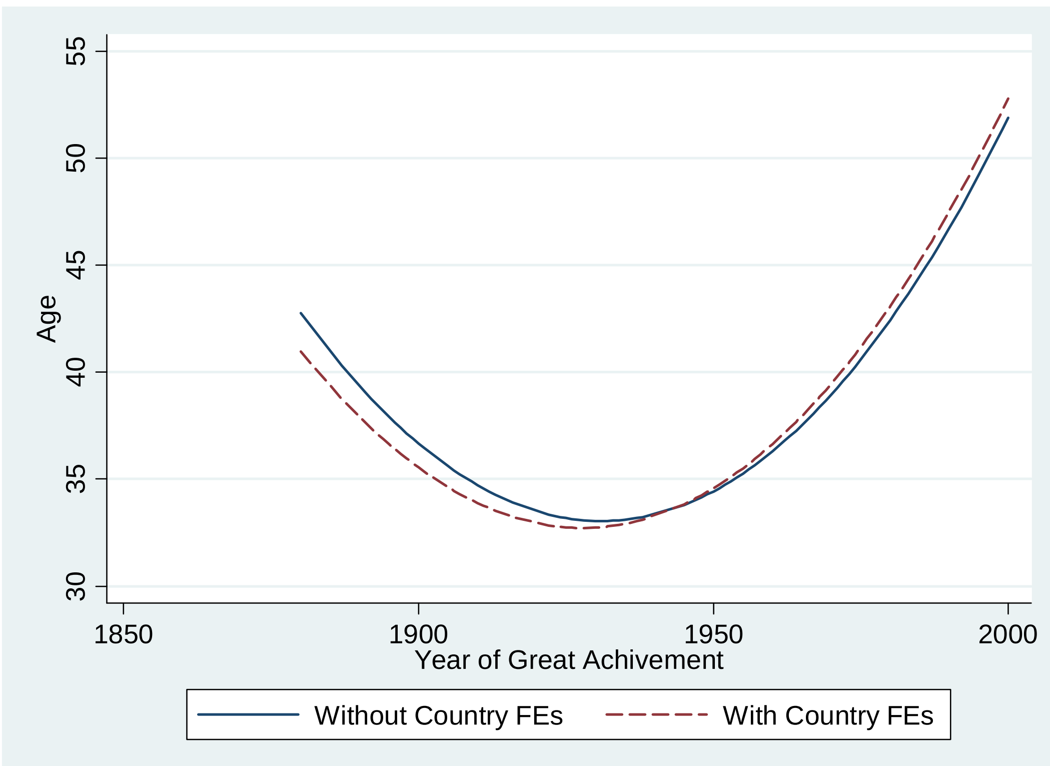
2. Abbott, David. *Physicists*. New York: Bedrick Books, 1984.

3. Bernal, J. D., and Andrew Brown. *The Sage of Science*. Vol. XIV. Oxford: Oxford UP, 2005. 1-562.
4. Brennan, Richard P. *Heisenberg Probably Slept Here: the Lives, Times, and Ideas of the Great Physicists of the 20th Cenutry*. New York: John Wiley & Sons, Inc., 1997.
5. Bromley, Allan. *A Century of Physics*. New Haven: Springer, 2002.
6. Gonzalo, Julio A., and Carmen A. Lopez. *Great Solid State Physicists of the 20th century*. Toh Tuck Link: World Scientific Co., 2003.
7. Hargittai, Magdolna, and Istvan Hargittai. *Candid Science IV: Conversations with Famous Physicists*. London: Imperial College P, 2004. 3-695.
8. Kragh, Helge. *Quantum Generations: a History of Physics in the Twentieth Century*. Princeton: Princeton UP, 1999.
9. Nye, Mary J. *Physics, War, and Politics in the Twentieth Century*. Cambridge: Harvard UP, 2004. 1-255.
10. Österman, Jonny, and Carl Nordling. "Famous Physicists in Appendix D of Physics Handbook." *Physics Handbook*. 1999. 05 July 2007 <http://hem1.passagen.se/asystem/PH/FamPhys.htm>.
11. Pelletier, Paul A. *Prominent Scientists: an Index to Collective Biographies*. New York: Neal-Schuman, 1980.
12. Reinhardt, Joachim. "Pioneers of Quantum Theory." *AIP Center for History of Physics*. 09 Sept. 1999. 23 June 2007 <http://th.physik.uni-frankfurt.de/~jr/physpicquant.html>.

13. Renn, Jürgen, and Kostas Gavroglu. *Positioning the History of Science*. Vol. VII. Dordrecht : Springer, 2007. 1-188.
14. Weisstein, Eric. "Eric Weisstein's World of Biography / Physicists" *Wolfram Research*. 08 July 2007 <http://scienceworld.wolfram.com/biography/topics/Physicists.html>.
15. "Selected Papers of Great American Physicists." *American Institute of Physics*. 2007. 06 July 2007 <http://www.aip.org/history/gap/>.
16. "Biographical Memoirs." *National Academy of Sciences*. 06 July 2007 <http://www.nasonline.org/site/PageServer?pagename=MEMOIRS_A>.
17. "List of Physicists." *Wikipedia*. 19 June 2007 <http://en.wikipedia.org/wiki/Physicists>. Physicists exclusive of Nobel Prize winners.

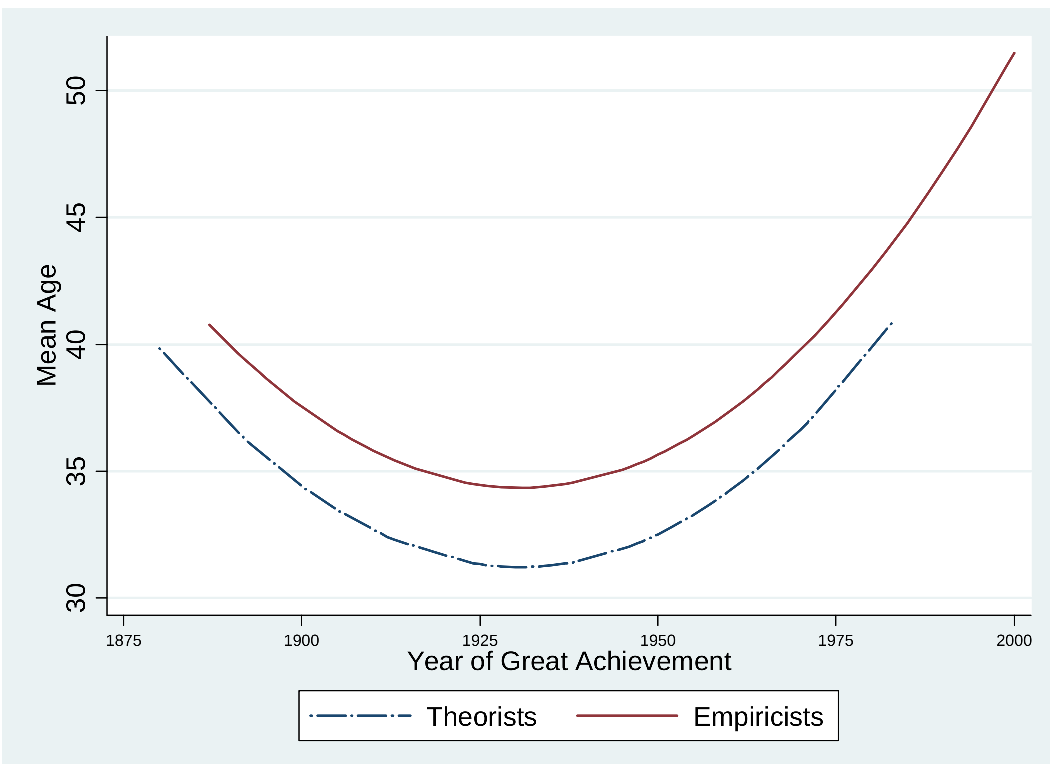
Supporting Figure 5: Age of Achievement over Time Controlling for Region of Birth

This figure shows how the mean age of great achievement in physics varies over time controlling for 8 regions of birth (the United Kingdom, Germany, Russia, other Eastern Europe, the rest of Europe, the United States, European offshoots, Japan, and the rest of the world) using country / region fixed effects (FEs). Time is captured using a fractional polynomial regression. The figure plots the implied curves with and without dummy variables for region of birth, showing that the dynamics are similar.



Supporting Figure 6: Age of Achievement in Physics for Theorists vs Empiricists over Time

This figure shows how the mean age of great achievement varies with whether a physics laureate's great achievement had an important theoretical component, while controlling flexibly for time using a fractional polynomial regression. Nobel laureates who received the prize for works with an important theoretical component did their work 3.13 years (standard error +/-1.37 years) younger than Nobel laureates who received the prize for empirical work. The figure plots the implied curves for theorists and empiricists. The regression predictions show that the age gap between theorists and empiricists is large, but that a sizeable U-shape in time remains.



Supporting Table 3: Predictors of Age of Great Achievement

The following panels report regressions predicting the age of great achievement based on (i) the theoretical nature of the work and (ii) the age at Ph.D. Panel A uses probit models to predict great achievement by age 35. Here we use $Theoretical_i$, a binary indicator equal to 1 if laureate *i*'s contribution had an important theoretical component; and $PhD\ by\ Age\ 25_i$, a binary indicator equal to 1 if laureate *i*'s Ph.D. was received before age 25 to predict the probability that laureate *i*'s great achievement wage made by age 35, where Age_i denotes the age at which laureate *i* made his or her contribution. We also flexibly control for the field and time when achievement *i* was made with dummy variables for the field, quadratics in time, and interactions between the two (captured by $\mathbf{FT}_i$). Formally, we use the Probit model,

$$\Pr[Age_i < 35] = \Phi(\beta_0 + Theoretical_i \beta_1 + PhD\ by\ Age\ 25_i \beta_2 + \boldsymbol{\theta}\mathbf{FT}_i),$$

where Φ denotes the cumulative density function of a normal distribution.
The table reports the marginal effects of a discrete change in $Theoretical_i$ and $PhD\ by\ Age\ 25_i$ from 0 to 1 on the mean probability that the laureates' great achievements will be made by age 35. In the case of $Theoretical_i$ ($PhD\ by\ Age\ 25_i$ is directly analogous) the reported estimate is,

$$\begin{aligned}&\Pr[Age_i < 35] \\ &= \frac{1}{I}\sum_i \Phi(\beta_0 + 1\beta_1 + PhD\ by\ 25_i \beta_2 + \boldsymbol{\theta}\mathbf{FT}_i) - \Phi(\beta_0 + 0\beta_1 + PhD\ by\ 25_i \beta_2 + \boldsymbol{\theta}\mathbf{FT}_i)\end{aligned}$$

where *I* denotes the number of laureates in the data.

Panel B use ordinary least squares regressions to predict the mean age at great achievement. Here the model is

$$Age_i = \gamma_0 + \gamma_1 Theoretical_i + \gamma_2 PhD\ Age_i + \mathbf{\theta FT}_i + \varepsilon_i$$

where *PhD* Age_i is the age at which laureate *i* received his or her Ph.D. In panel B, the coefficients give the relationship between each variable and the mean age of great achievement.

In both panels, column (1) considers theory alone, column (2) considers training alone, and column (3) considers both together. Column (4) further includes field fixed effects for each of physics, chemistry, and medicine. Column (5) further includes time controls, which are field-specific quadratics in the calendar year of the achievement.

Depending on the specification, the probit models for the probability of great achievement by 35 reported in Panel A, show that receiving a Ph.D. by age 25 is associated with a 13-15 percentage point increase in great achievement by age 35 (a 38-45 percent increase in the baseline rate). Independently, a theoretical contribution is associated with a 17-24 percentage point increase in great achievement by age 35 (a 51-73 percent increase in the baseline rate). The linear models reported in Panel B show that both theoretical research and Ph.D. age have substantial explanatory power for the achievement age. People whose contributions were theoretical were 2.930 to 4.546 years younger at the time of their great achievement and the age of great achievement increases by .223 to .326 years with every year of age at Ph.D. Robust standard errors are given in parentheses. ** indicates significance at 5%; *** indicates significance at 1%.

Panel A: Models to Predict Probability of Great Achievement by Age 35

	(1)	(2)	(3)	(4)	(5)
Theoretical	0.245*** (0.055)		0.231*** (0.055)	0.201*** (0.057)	0.174*** (0.058)
PhD by Age 25		0.151*** (0.044)	0.135*** (0.044)	0.141*** (0.044)	0.125*** (0.046)
Field Fixed Effects	No	No	No	Yes	Yes
Time Controls	No	No	No	No	Yes
No. of Observations	525	525	525	525	525
Mean of Dependent Variable	0.33	0.33	0.33	0.33	0.33
Regression Chi2	20.47	12.21	29.72	40.32	47.98

Panel B: Models to Predict Mean Age of Great Achievement

	(1)	(2)	(3)	(4)	(5)
Theoretical	-4.546*** (0.925)		-4.434*** (0.907)	-3.999*** (0.932)	-2.930*** (0.921)
Age at PhD		0.325*** (0.094)	0.304*** (0.094)	0.326*** (0.094)	0.223** (0.094)
Field Fixed Effects	No	No	No	Yes	Yes
Time Controls	No	No	No	No	Yes
No. of Observations	525	525	525	525	525
Mean of Dependent Variable	39.04	39.04	39.04	39.04	39.04
Regression F-statistic	24.15	12.08	16.96	12.62	11.20

Supporting Table 4: Citation Age Dynamics in Physics

This table reports regressions that estimate the citation age dynamics in physics. Observations are at the paper level (see discussion of ISI data above for details). Citation age is the mean duration between the paper's publication year and the publication years of the papers it cites. The dependent variable in the regression is the normalized citation age for a given paper, defined formally above and calculated as the deviation from the mean citation age of all other papers published that year and divided by the standard deviation in citation age among other papers in that year. Other papers are defined as the 100 most cited articles annually in the Century of Science and Web of Science databases outside the fields of physics, chemistry, and medicine. The first column considers the citation age dynamics for all individuals who write at least 2 papers in physics. To assess the extent to which our estimates indicate general changes in the knowledge space itself, not simply changes in which physicists were active, the second column repeats this regression but includes researcher fixed effects, thus netting out any fixed individual tendency to cite old or new work. In order to implement the fixed effect model, we employ quadratic polynomials as time controls. The estimates imply that the tendency to cite recent papers peaks in 1920 in physics. The citation data cover the period 1900-2000. Robust standard errors are in parentheses, clustered by researcher name. *** indicates significant at 1%

	(1)	(2)
Year	-1.56***	-0.74
	(0.19)	(0.54)
Year ^ 2	0.0197***	0.0184***
	(0.0017)	(0.0044)
Individual Fixed Effects	No	Yes
Observations	17440	17440
R-squared	0.04	0.54
Year of Minimum	1939.71	1920.15
	(1.81)	(10.32)