

# *Matching the SDR to Publications in the 1980s: Experiences and Outcomes*

Paula Stephan

Georgia State University and NBER

[pstephan@gsu.edu](mailto:pstephan@gsu.edu)

NSF

February 2008

# [ A short history ]

- 1980 Alan Fechter, NSF program officer, made observation that U.S. scientific workforce was aging as a result of slowdowns in hiring.
- Between 1973 and 1987 median age of PhD scientists rose from 41 to over 44.
- If “science is a young person’s game,” this could have serious productivity effects
- Suggested need for an econometric study examining relationship between age and productivity

# [ P.A.M. Diract ]

Age is, of course, a fever chill  
That every physicist must fear.  
He's better dead than living still  
When once he's past his thirtieth year.



# [ SDR ]

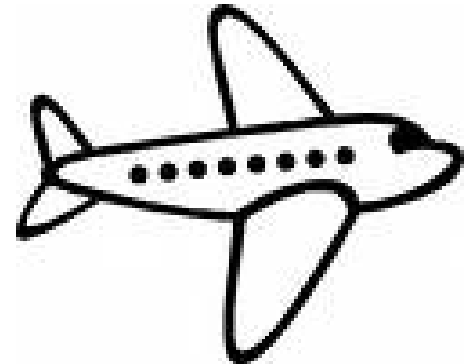
- SDR was most appropriate data to use
- But did not have publication data and no one had matched it to publication data
- Had numerous meetings with NSF staffers
- Wrote grant proposal(s)
- Finally funded in 1983 by NSF, Sloan and Exxon; encouraged our meeting Harriet Zuckerman and Coles
- Sharon Levin co-PI and co-author

# Publication Data

- Publication data for individuals in four fields of science came from ISI
  - Physics
  - Earth Sciences
  - Physiology
  - Biochemistry
- Match occurred at NRC, SDR contractor at time
  - 15,238 records in four fields during 1973-79
  - 9.6 entries in ISI Source Index
  - Matches made for two-year period; lag of one year; 1973 records linked with 1974 & 1975 ISI data, etc.

# [ Use of data ]

- Data analysis occurred at the Green Building (NRC location)
- Or, on rare occasions, tapes were flown to St. Louis by an NRC employee and Sharon worked over a weekend.



# [ Match Criteria ]

- Matching procedure was a variant of what was developed for Jones et al NRC report in 1983
- Used last name, first and middle initial (when present) as well as address and information on coauthors.
- After match was completed, random samples were drawn and then actual publication was found to assess if match were correct.

# [ What verification showed ]

- If first and middle initial match and zip matches reliability was extremely high
- Reliability began to decline if zip code was only at 3-digit level match
- Match was significantly more reliable if there was a middle initial than no middle initial
- Reliability could be increased in two latter cases by “cleaning” data, eliminating matches in which the publication appeared to be outside the journal set for a fine field. **Field of scientist taken from SDR.**

# Reliability of match for “prolific” authors

- An unanticipated finding during verification was that match reliability appeared related to number of articles attributed to a scientist
- Also realized that many prolific authors had a common last name (something that is also abundantly clear if one studies ISI’s list of most cited individuals)
- Checked reliability for “common” prolific authors

# What constitutes a common name?

- Used DRF; cumulative name frequency was generated from 1920-1985 DRF file.
- 770,624 individuals
- 38.32% of population have a last name and first initial that is unique
- Slightly more than 70% share last name and first initial with at most 10 other persons.
- 10% share a name with 50 or more
- 20% share a name with 20 or more
- 644 scientists share common last name, first initial.

# Findings

Distribution location	N of Matches	N Correct	% Correct
Top 10 pubs; top 10 names	539	69	12.8
Top 10 pubs; top 20 names	314	99	31.6
Top 20 pubs; top 10 names	414	233	56.3
Top 20 pubs; top 30 names	130	93	71.5
Top 10 pubs; top 30 names	317	244	76.9
Top 20 pubs Top 20 names	200	168	84.0
Top 10 pubs; Top 40 names	105	105	100.0

[ But . . . ]

- Eliminating matches for which there was zip code disagreement would have increased accuracy of prolific authors with common names to 75%. (first three rows).
- Good news because a real challenge for the current match is tremendous increase in common names because of ethnic distribution of U.S. PhDs awarded; need to include in a study rather than “throw away”

# [ Empirical Findings ]

- Controlled for age, vintage and time effects
- Vintage effects created by surveys and interviews
- Also controlled for age, time effects and fixed effects.
- Found that with exception of particle physicists employed in PhD-granting departments, true aging effects are present in a fully specified model.
- Vintage matters, but not in the sense that the “latest educated are best educated.”

# [ Conclusions ]

---

1. For average PhD scientists age matters but it does not matter a great deal
2. Stronger relationship between age and ability to do path-breaking work.

# [ Findings ]

- Published in *American Economic Review*, *Social Studies of Science*, and others
- Book: *Striking the Mother Lode in Science*
- Data was basis for another study that focused on gender funded by NSF.
- Study was easier to get funded but harder to get access to the data. Took approximately a year to work out arrangement to use the data matched for earlier study.

# [ Comments/Questions ]

---

- [pstephan@gsu.edu](mailto:pstephan@gsu.edu)