Codebook to Accompany Stata Dataset: Science_NIH_deidentified.dta

This codebook lists the variables, definitions, and tabulations of the de-identified data that can be used to replicate key findings in "Race, Ethnicity, and NIH Awards." Information about the construction of the data used in the paper can be found in the supplemental online material that accompanies the paper. The analytical files used in the studies described in this manuscript contain personal information from individuals who have submitted applications and in some cases have received awards from the NIH. Many of these application records have been matched to records included in the Survey of Earned Doctorates as maintained by the National Science Foundation and to records included in the Faculty Roster maintained by the Association of American Medical Colleges. The information is therefore protected by the Privacy Act of 1974 as amended (5 U.S.C. 552a) and the National Science Foundation Act of 1950 as amended (42 U.S.C. 1873(i). In order to facilitate replication of the analysis, we created a de-identified version of the data that preserved key variables used in the research including race, type of grant, employer characteristics, previous NIH grants, publications and citations.

Researchers interested in access to the confidential analytical files should review the NSF/SRS Restricted-Use Data Procedures Guide available at www.nsf.gov/statistics/license/forms/pdf/srs_license_guide_august_2008.pdf and the NSF Data and Tools website at http://www.nsf.gov/statistics/database.cfm. Upon completing this review, they should call W. T. Schaffer at 301-402-2725 to discuss the security clearance requirements necessary for using the data.

Replication:
Researchers using the de-identified data for replication purposes should be aware that estimates from the de-identified data will not be fully comparable to estimates reported in "Race, Ethnicity, and NIH Awards." This is the result of several variables having been omitted and many variables having been recoded to preserve the privacy of NIH applicants.

We have provided a STATA 11 data set Science_NIH_deidentified.dta. We have also provided a delimited ASCII dataset Science_NIH_deidentified.csv. Information on variable definitions and tabulations of variables appear below.

We have provided STATA code to accompany the data.

- Science_descriptive_deident1.do: This generates the estimates for versions of Figure 1 and Tables S2, S11, S12, and S13 using the de-identified data.
- Science_probit_deidentified1.do: This estimates the probit models for versions of Table S5, Figure 3, Table S14, and Table S15 using the de-identified data.

Please note when replicating the tables, that the top panel of Table S2 uses all of the data, the middle panel of Table S2 limits the data to all R01 Type 1 awards (r01_type1==1), and the bottom panel of S2 and the remaining Tables in the paper limits the data to the PhD analysis sample (phdsamp==1).

We have included a set of tables (Science_deidentified_tables.xls) that use the de-identified data to estimate tables similar to those reported in the supplemental online material using Science_descriptive_deident1.do and Science_probit_deidentified1.do.

```
               Dataset:  Science_NIH_deidentified.dta
             Last saved:  16 Jun 2011 23:40
                 Author:  Donna K. Ginther, Laurel Haak, Joshua Schnell

                  Label:  De-identified dataset to accompany “Race, Ethncity, and
                          NIH Awards”
   Number of variables:  32
Number of observations:  205,106
                   Size:  22,151,448 bytes ignoring labels, etc.
```

VARIABLE **DEFINITION**

---

id_num — Sample Grant ID Number

```
                  type:  numeric (long)

                 range:  [1,205106]                  units:  1
         unique values:  205106                   missing .:  0/205106
```

Note: Generated for De-identified dataset. Grant is the unit of analysis.

---

r01 — R01 grant application

```
                  type:  numeric (byte)
                 label:  dummy

                 range:  [0,1]                       units:  1
         unique values:  2                        missing .:  0/205106

            tabulation:  Freq.   Numeric  Label
                          74147         0  No
                         130959         1  Yes
```

---

rpgawd — Research Grant Awarded

```
                  type:  numeric (byte)
                 label:  dummy

                 range:  [0,1]                       units:  1
         unique values:  2                        missing .:  0/205106

            tabulation:  Freq.   Numeric  Label
                         136435         0  No
                          68671         1  Yes
```

VARIABLE **DEFINTION**

---

r01awd R01 Grant Awarded

```
                 type:  numeric (byte)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                    missing .:  74147/205106

            tabulation:  Freq.   Numeric  Label
                         86397         0  No
                         44562         1  Yes
                         74147         .  Missing
```

Note: Defined only for R01 grants

---

org_high Employer Higher Education Institution

```
                 type:  numeric (byte)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                    missing .:  0/205106

            tabulation:  Freq.   Numeric  Label
                         40588         0  No
                        164518         1  Yes
```

---

fund_rank1 Employer Ranked 1-30 NIH Funding

```
                 type:  numeric (byte)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                    missing .:  0/205106

            tabulation:  Freq.   Numeric  Label
                        136923         0  No
                         68183         1  Yes
```

VARIABLE **DEFINTION**

---

fund_rank2 Employer Ranked 31-100 NIH Funding

---

type: numeric (byte)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 0/205106

tabulation: Freq. Numeric Label
135,257 0 No
69849 1 Yes

---

fund_rank3 Employer Ranked 101-200 NIH Funding

---

type: numeric (byte)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 0/205106

tabulation: Freq. Numeric Label
175,090 0 No
30016 1 Yes

---

fund_rank4 Employer Ranked >200 NIH Funding

---

type: numeric (byte)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 0/205106

tabulation: Freq. Numeric Label
168,048 0 No
37058 1 Yes

VARIABLE **DEFINTION**

---

hs_y Application includes Human subjects

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         60211         0  No
                         37969         1  Yes
                        106926         .  Missing
```

Note: Defined only for PhD Analysis Sample

---

priorgrant Prior NIH Grants

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         23871         0  No
                         74309         1  Yes
                        106926         .  Missing
```

Note: Defined only for PhD Analysis Sample

---

cmte_c Served on NIH Review Committee

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         49856         0  No
                         48324         1  Yes
                        106926         .  Missing
```

Note: Defined only for PhD Analysis Sample

VARIABLE **DEFINTION**

---

pub_badmatch — Publication Data not Matched

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         83267        0  No
                         14913        1  Yes
                        106926        .  Missing
```

Note: Defined only for PhD Analysis Sample

---

phdsamp — PhD Analysis Sample

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         14992        0  No
                         83188        1  Yes
                        106926        .  Missing
```

Note: Defined only for PhD Analysis Sample

---

scored — Grant has priority score

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                         82541        0  No
                        122,565       1  Yes
```

VARIABLE **DEFINTION**

---

roleftk F, T, or K Recipient

---

type: numeric (float)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 0/205106

tabulation:

Freq.	Numeric	Label
107,729	0	No
97377	1	Yes

---

pubq1 Publications 1st Quartile (<=3)

---

type: numeric (float)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 106926/205106

tabulation:

Freq.	Numeric	Label
57876	0	No
40304	1	Yes
106926	.	Missing

Note: Defined only for PhD Analysis Sample

---

pubq2 Publications 2nd Quartile (4-7)

---

type: numeric (float)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 106926/205106

tabulation:

Freq.	Numeric	Label
80069	0	No
18111	1	Yes
106926	.	Missing

Note: Defined only for PhD Analysis Sample

VARIABLE **DEFINTION**

---

pubq3 Publications 3rd Quartile (8-18)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                        missing .:  106926/205106

            tabulation:  Freq.   Numeric  Label
                         78583         0  No
                         19597         1  Yes
                        106926         .  Missing
```

Note: Defined only for PhD Analysis Sample

---

pubq4 Publications 4th Quartile (>18)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                        missing .:  106926/205106

            tabulation:  Freq.   Numeric  Label
                         78012         0  No
                         20168         1  Yes
                        106926         .  Missing
```

Note: Defined only for PhD Analysis Sample

---

citq1 Citations 1st Quartile (<=5)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
         unique values:  2                        missing .:  106926/205106

            tabulation:  Freq.   Numeric  Label
                         61560         0  No
                         36620         1  Yes
                        106926         .  Missing
```

Notes: Excludes Self-Citations
Defined Only for PhD Analysis Sample

VARIABLE **DEFINTION**

---

citq2 Citations 2nd Quartile (6-24)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         78732        0  No
                         19448        1  Yes
                        106926        .  Missing
```

Notes: Excludes Self-Citations
Defined Only for PhD Analysis Sample

---

citq3 Citations 3rd Quartile (25-84)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         77076        0  No
                         21104        1  Yes
                        106926        .  Missing
```

Notes: Excludes Self-Citations
Defined Only for PhD Analysis Sample

---

citq4 Citations 4th Quartile (>84)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         77172        0  No
                         21008        1  Yes
                        106926        .  Missing
```

Notes: Excludes Self-Citations
Defined Only for PhD Analysis Sample

VARIABLE **DEFINTION**

---

citq5 Citations above Median (>24)

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  106926/205106

           tabulation:  Freq.   Numeric  Label
                         56068         0  No
                         42112         1  Yes
                        106926         .  Missing
```

Notes: Excludes Self-Citations
Defined Only for PhD Analysis Sample

---

r01_type1 R01 Type 1 Application

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                    missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                         98738         0  No
                        106,368        1  Yes
```

---

race Race / Ethnicity Category

```
                 type:  numeric (float)
                label:  racecat

                range:  [1,5]                        units:  1
        unique values:  5                    missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                         28274         1  Asian
                          2942         2  Black
                          6954         3  Hispanic
                        135594         4  White
                         31342         5  Other
```

Notes: Other includes Native Americans, Multiple Race, Race Unknown

VARIABLE **DEFINTION**

---

asian Asian

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                       176,832        0  No
                        28274         1  Yes
```

---

white White

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                        69512         0  No
                       135,594        1  Yes
```

---

black Black

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                       202,164        0  No
                         2942         1  Yes
```

---

hispanic Hispanic

```
                 type:  numeric (float)
                label:  dummy

                range:  [0,1]                        units:  1
        unique values:  2                          missing .:  0/205106

           tabulation:  Freq.   Numeric  Label
                       198,152        0  No
                         6954         1  Yes
```

VARIABLE **DEFINTION**

---

other Other Race

---

type: numeric (float)
label: dummy

range: [0,1] units: 1
unique values: 2 missing .: 0/205106

tabulation: Freq. Numeric Label
173,764 0 No
31342 1 Yes

Notes: Includes Native Americans, Multiple Race, and Unknown Race/Ethnicity

Tabulation of Variables

```
. tab r01

  R01 grant |
application |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     74,147       36.15       36.15
        Yes |    130,959       63.85      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab rpgawd

   Research |
      Grant |
    Awarded |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    136,435       66.52       66.52
        Yes |     68,671       33.48      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab r01awd

  R01 Grant |
    Awarded |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     86,397       65.97       65.97
        Yes |     44,562       34.03      100.00
------------+-----------------------------------
      Total |    130,959      100.00

. tab org_high

   Employer |
     Higher |
  Education |
Institution |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     40,588       19.79       19.79
        Yes |    164,518       80.21      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab fund_rank1

   Employer |
Ranked 1-30 |
NIH Funding |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    136,923       66.76       66.76
        Yes |     68,183       33.24      100.00
------------+-----------------------------------
      Total |    205,106      100.00
```

```
. tab fund_rank2

   Employer |
     Ranked |
 31-100 NIH |
    Funding |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    135,257       65.94       65.94
        Yes |     69,849       34.06      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab fund_rank3

   Employer |
     Ranked |
101-200 NIH |
    Funding |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    175,090       85.37       85.37
        Yes |     30,016       14.63      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab fund_rank4

   Employer |
Ranked >200 |
NIH Funding |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    168,048       81.93       81.93
        Yes |     37,058       18.07      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab hs_y

Application |
   includes |
      Human |
   subjects |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     60,211       61.33       61.33
        Yes |     37,969       38.67      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab priorgrant

  Prior NIH |
     Grants |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     23,871       24.31       24.31
        Yes |     74,309       75.69      100.00
------------+-----------------------------------
      Total |     98,180      100.00
```

```
. tab cmte_c

  Served on |
 NIH Review |
  Committee |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     49,856       50.78       50.78
        Yes |     48,324       49.22      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab pub_badmatch

Publication |
   Data not |
    Matched |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     83,267       84.81       84.81
        Yes |     14,913       15.19      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab phdsamp

        PhD |
   Analysis |
     Sample |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     14,992       15.27       15.27
        Yes |     83,188       84.73      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab scored

  Grant has |
   priority |
      score |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     82,541       40.24       40.24
        Yes |    122,565       59.76      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab roleftk

 F, T, or K |
  Recipient |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    107,729       52.52       52.52
        Yes |     97,377       47.48      100.00
------------+-----------------------------------
      Total |    205,106      100.00
```

```
. tab pubq1

Publication |
      s 1st |
   Quartile |
      (<=3) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     57,876       58.95       58.95
        Yes |     40,304       41.05      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab pubq2

Publication |
      s 2nd |
   Quartile |
      (4-7) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     80,069       81.55       81.55
        Yes |     18,111       18.45      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab pubq3

Publication |
      s 3rd |
   Quartile |
     (8-18) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     78,583       80.04       80.04
        Yes |     19,597       19.96      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab pubq4

Publication |
      s 4th |
   Quartile |
      (>18) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     78,012       79.46       79.46
        Yes |     20,168       20.54      100.00
------------+-----------------------------------
      Total |     98,180      100.00
```

```
. tab citq1

  Citations |
        1st |
   Quartile |
      (<=5) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     61,560       62.70       62.70
        Yes |     36,620       37.30      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab citq2

  Citations |
        2nd |
   Quartile |
     (6-24) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     78,732       80.19       80.19
        Yes |     19,448       19.81      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab citq3

  Citations |
        3rd |
   Quartile |
    (25-84) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     77,076       78.50       78.50
        Yes |     21,104       21.50      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab citq4

  Citations |
        4th |
   Quartile |
      (>84) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     77,172       78.60       78.60
        Yes |     21,008       21.40      100.00
------------+-----------------------------------
      Total |     98,180      100.00
```

```
. tab citq5

  Citations |
      above |
     Median |
      (>24) |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     56,068       57.11       57.11
        Yes |     42,112       42.89      100.00
------------+-----------------------------------
      Total |     98,180      100.00

. tab r01_type1

 R01 Type 1 |
Application |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     98,738       48.14       48.14
        Yes |    106,368       51.86      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab race

     Race / |
  Ethnicity |
   Category |      Freq.     Percent        Cum.
------------+-----------------------------------
      Asian |     28,274       13.79       13.79
      Black |      2,942        1.43       15.22
   Hispanic |      6,954        3.39       18.61
      White |    135,594       66.11       84.72
      Other |     31,342       15.28      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab asian

      Asian |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    176,832       86.21       86.21
        Yes |     28,274       13.79      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab white

      White |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |     69,512       33.89       33.89
        Yes |    135,594       66.11      100.00
------------+-----------------------------------
      Total |    205,106      100.00
```

```
. tab black

      Black |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    202,164       98.57       98.57
        Yes |      2,942        1.43      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab hispanic

   Hispanic |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    198,152       96.61       96.61
        Yes |      6,954        3.39      100.00
------------+-----------------------------------
      Total |    205,106      100.00

. tab other

 Other Race |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |    173,764       84.72       84.72
        Yes |     31,342       15.28      100.00
------------+-----------------------------------
      Total |    205,106      100.00
```