About the Data Files

I have two purposes in posting these data files. The first is to allow interested readers to test whether the conclusions presented in *Facing Reality* are robust when analyzed in alternative ways. The second is to share databases that were laborious to assemble and have far more potential than I could make use of in *Facing Reality.*

The problem is that I have not yet done any extensive analyses of the database of violent crime by zip code or the combined longitudinal studies. In my experience, cleaning large databases is an iterative process. One may try hard to check all possible inconsistencies and coding errors, but some errors remain hidden until the database has been wrung out by being subjected to many analyses. I have to assume this is true of these datasets as well. I request users who discover anomalies to let me know so that I can either fix the problem and post updated files, or in some cases can explain the apparent anomaly. I can be reached at cmurray@aei.org.

About Downloading the Data Files

One of the data files, *Nationally Representative Studies.xlsx*, consists of small datasets that some users will want to analyze using Excel. The sheets include information on sources and variables. If you want to export them to a statistical package, you should strip each sheet of its descriptive header rows and source information. The easiest method is to copy a block of data, paste it directly into your statistical package, and assign your own variable names.

The other files are suitable for analysis using a statistical package. I have saved them as Excel files with variable names in the first line and no other explanatory information. I recommend that you use Excel’s option to save them as .csv files. You can then easily import them into the major statistical packages. I don’t know how SPSS or SAS does it, but in Stata you should use the File/Import/Text data (delimited, *csv, …) option.

Nationally Representative Studies.xlsx

This file has three sheets with self-explanatory variables with one exception.

Inventory of g*-loaded Studies* shows the means, standard deviations, and sample sizes for all of the IQ standardizations and for the federally-sponsored longitudinal surveys, plus the calculations of the Black-White, Latino-White, and Asian-White differences in test scores. The title of the sheet is slightly misleading because two of the federally-sponsored surveys, NELS-88 and ELS-02 administered only math and reading tests, as indicated in the variable **test**.

The only variable that requires some explanation is **periodcohort**. A coding of "C" means that the data in that line should be used only in cohort analyses, a coding of "P" means that the data should be used only in period analyses, and a coding of "PC" means that the data may be used for both types of analyses.

Inventory of M&R Studies includes all of the NAEP, PISA, and PIAAC results. For the method of calculating race differences in the NAEP data, see *Facing Reality,* pp. 29–30 and note 9, pp. 133–34.

Detailed NAEP Data brings together on one sheet all the means and standard deviations for all the administrations of the NAEP, both LTT and standard, for all three age/grade groups, and all race breakdowns.

Combined Longitudinal Studies.xlsx

This file consists of 20,203 observations. Variables are defined as follows.

study (string)

NLS-72 National Longitudinal Study-1972. The original data are downloadable at icpsr.umich.web/pages.

NLSY-79 National Longitudinal Survey of Youth-1979.

NLSY-97 National Longitudinal Survey of Youth-1997. The original data for both NLSY-79 and NLSY-97 are downloadable at nlsinfo.org/investigator/pages/login.

studyid (numeric)

Each study assigned a unique ID number to each participant. **studyid** can be used to merge additional variables from any of the three downloadable databases.

idcombined (numeric)

These unique ID numbers were assigned as the line number when the three surveys were sorted by **study** and **studyid**.

dob (numeric)

Date of birth expressed as the last two digits of the year plus a fraction of that year.

sex (string)

Female

Male

birthyr (numeric)

Birth year.

testyr (numeric)

Year when the cognitive test was administered. When administration stretched over more than one calendar year, the year when most of the tests were administered.

1972	NLS-72
1980	NLSY-79
1997	NLSY-97

followupyr (numeric)

The year of the followup survey.

1986	NLS-72
2004	NLSY-79
2017	NLSY-97

testwgt (numeric)

Sample weights for **testyr.** The variable names are TBWT for NLS-72, R0614700 for NLSY-79, and R1236101 for NLSY-97

fuwgt (numeric)

Sample weights for **followupyr**. The variable names are FU5WT for NLS-72, R8495700 for NLSY-79, and U1855400 for NLSY-97

religion (string)

Religion in which the subject was raised (answers were given in the first survey, when subjects were in their teens or early 20s).

Protestant
Roman Catholic
Other Christian
Judaism
Other
None

momed (numeric)

Years of mother's completed education.

poped (numeric)

Years of father's completed education

parincile (numeric)

The parents' percentile in the national income distribution of family income in the year of the first survey. The national distribution was determined through the Current Population Survey (CPS) variables for family income and household weights. The percentile represents the bottom of the range, so that the 10th percentile includes every one from 10.0 until but not including 11.0. The NLS-72 reported family income in 10 categories instead of number of dollars, so there are only 10 values of **parincile** and they are far more imprecise measures than the ones for NLSY cohorts.

marstatus (string)

Marital status of the subject as of **followupyr**.

Never married
Married
Sep/Div [Separated or Divorced]
Widowed

degree (string)

The subject’s terminal educational degree as of **followupyr**.

0 LTHS	Less than high school
1 HS	High school diploma or GED
2 AA	Completion of an Associate’s degree or a two-year vocational course.
3 BA	Bachelor’s degree (BS as well as BA)
4 MA	Master’s degree
5 Prof	PhD, MD, JD, DDS, DVM

iq (numeric)

The subject’s score on the cognitive test battery administered for each study expressed in the IQ metric (mean=100, SD=15). Very low and very high scores were assigned a floor and ceiling of three standard deviations from the mean (55 and 145 respectively). See the online *Notes to the Text of Facing Reality*, pp. 31–32, for the method used for the NLSY cohorts. For NLS-72, I took the mean of the six subtests in the cognitive battery, all of which were scaled to a mean=50 and SD=10, calculated standardized scores, and converted them to the IQ metric. The resulting distribution was slightly right-skewed but so close to normal that I did not judge a correction for skew to be appropriate.

latinobin (numeric)

Coded 1 for all who self-identified as Latino of whatever race for NLSY-79 and NLSY-97. NLS-72 did not identify Latinos independently of race.

basicrace (string)

Racial self-identification, ignoring Latino self-identification (a person of any race can self-identify as a Latino).

Amerind	American Indian, Native American, Alaskan Native
Arab	Arab
Asian	Asians combined
E Asian	East Asian (1997 cohort only)
S Asian	South Asian (1997 cohort only)
Black	Black, African American
Filipino	Filipino
HiPac	Hawaiian or other Pacific Islands
White	Whites
Other	Some other single race, mixed race, or otherwise not classified

racefr (string)

Racial identification using the definitions in *Facing Reality.* See the online *Notes to the Text of Facing Reality*, p. 6.

1 Eur	European
2 Afr	African
3 Lat	Latin
4 Asi	Asian (includes Hawaiian and Pacific islanders for NLS-72)
5 Oth	Other

sector (string)

19 occupational categories

job (string)

130 occupational labels. NLS-72 used the 1970 version of the Census Occupation Code list in the 1986 followup while NLSY-79 and NLSY-97 used the 2002 version. When the 1970 version did not have a code that coincided with the 2002 version, NLS-72 subjects were coded as missing.

jobcode (numerical)

Numerical codes for the 130 occupational labels assigned by me. They do not correspond to any of the Census Occupation Code versions.

jobn (numerical)

Number of subjects in a given occupation.

jobiqmean (numerical)

Unweighted mean IQ of subjects holding a given job.

jobsd (numerical)

Standard deviation of IQ scores for a given job.

wagepctile (numerical)

The subject's percentile in the national income distribution of wage and salary income in the year prior to **followupyr**. (wage and salary data in are reported for the preceding calendar year in all three studies). The percentile represents the bottom of the range, so that the 10th percentile includes every one from 10.0 until but not including 11.0.

anomalies (numerical)

Coded 1 if the reported occupation is radically inconsistent with **degree** or **iq**. See the online *Notes to the Text of Facing Reality,* pp. 58–59.

Violent Crime by Zip Code.xlsx

This file consists of data on 506,238 arrests in New York City NY, Los Angeles CA, Washington DC, Chandler AZ, Charleston SC, Fayetteville NC, Fort Lauderdale FL, and Tucson AZ.

All of the variables listed under *Arrest Data* below were downloaded from public files, as described in the online *Notes to the Text of Facing Reality,* p. 49. Note that the datasets with zip codes are a subset of the total arrest data used to create the tables in Chapter 4. They will not produce exactly the same arrest ratios.

Arrest Demographics consists of demographic variables for the 367 zip codes represented in Arrest Data. The data are based on the five-year compilation of the American Community Survey (ACS) from 2014–2018, downloaded from socialexplorer.com.

Arrest Data

---

zip (numeric)

Zip codes for Chandler, Charleston, Fayetteville, Los Angeles, New York City, and Washington were reverse-geocoded from **latitude** and **longitude** by TAMU GeoServices at Texas A&M (geoservices@tamu.edu). The Fort Lauderdale and Tucson databases supplied **zip** but not **latitude** and **longitude.**

Zip codes can refer to a specific geographic area or to a post office. Sometimes arrests are coded for a post office zip. In those cases, I converted the zip to the geographic zip code surrounding the post office.

Police departments occasionally make arrests that are geographically coded in zip codes outside their jurisdiction. I deleted these from the database. You may assume that a zip code with a low number of arrests was within the jurisdiction of the police department of the cities included in the database. In the case of Los Angeles, Beverly Hills and Santa Monica each has its own police department despite being surrounded by Los Angeles zip codes. LAPD arrests in those jurisdictions have been deleted as not indicative of the total level of arrests in those zip codes. The Charleston Police Department has jurisdiction over both Charleston and North Charleston.

---

city (string)

Each city was defined in terms of the area that falls within the boundaries of that city as a legal entity.

---

age (numeric)

For all the cities but New York and Fayetteville, **age** is age of the arrestee in years. Fayetteville did not report the age of the arrestee. Note that Washington reported only arrests involving suspects age 18 and older.

agecat (string)

New York City reported only age categories, coded as follows:

1	<18
2	18-24
3	25-44
4	45+

sex (string)

Female

Male

Uncoded

race (string)

The cities used a variety of codes for race and ethnicity.

Chandler: Codes for race were Amerind, Asian, Black, HI/Pacific, White, Unknown. An ethnicity variable had 20 codes, or ethnicity were detailed, with a single category labeled Hispanic/Latino/Mexican.

Charleston: A single race variable was coded Asian, Amerind, Black, White, Other, and Unknown.

Fayetteville: A single race variable was coded Asian, Amerindian, Black, Latino, White, Other, and Unknown.

Fort Lauderdale: A race variable was coded Asian, Amerind, Black, White, Other, and Unknown. An ethnicity variable was coded Latino, Not Latino, and Unknown.

Los Angeles*:* A single "descent" variable had 19 categories, including a single code for "Hispanic/Latino/Mexican."

New York: A single race variable was coded Asian/Pacific Islander, Amerind, Black, Black Hispanic, White, White Hispanic, Other, and Unknown.

Tucson: The race variable was coded Amerind, Asian, Black, Latino, HI/Pacific, Mixed, White, Other, and Unknown. An ethnicity variable was coded Latino, Not Latino, and Unknown.

Washington: A race variable was coded Asian, Black, White, Unknown. An ethnicity variable was coded Latino, Not Latino, Unknown (with 30% of cases coded as Unknown, which means that the results involving Latinos should be treated with reservations).

As noted in the text of *Facing Reality* (p. 139, note 4), Fayetteville included the names of the arrestees, which revealed that the count of Latinos among the arrests for violent crime (just 28) missed many people with Latin names. I did not report results Latinos in Fayetteville in the text. In the datafile, **race** for Fayetteville includes codes for Latino if the Fayetteville PD coded the arrestee as Latino or if the arrestee was coded as White and the name of the arrestee was highly likely to be Latino. Last names that qualified were Acosta, Alvardado, Aragon, Arand, Ayerbe, Barraza, Barroso, Bellis, Binetez, Bermudez, Burney, Candia, Carrillo, Castillo, Castro, Chavere, Colon, Corretjer, Cortes, Cruz, Cubacus, Deltoro, Durazo, Echeverria, Fayson, Ferrell, Figueroa, Flores, Fuentes, Garcia, Geddie, Gomez, Goncalves, Gonzalez, Guadalupe, Guzman, Hernandez, Ilarraza, Leon, Lopez, Loano, Marcellus, Martinez, Medina, Molina, Montero, Morales, Murillo, Navarro, Nieve, Olivares, Ortega, Ortiz, Perez, Portugues, Ramierez, Ramos, Recinos, Ribeiro, Ricardo, Rios, Rivas, Rivera, Roca, Rodriguez, Rosser, Ruiz, Santiago, Segarra, Serrano, Sierra, Silva, Sisneres, Socorro, Tirado, Toro, Trescatro, Valdez, Vallejo, Varela, Vega, Velez, Whitted, Zapata. First names that qualified were Angel, Carlos, Celestino, Cristobal, Eduardo, Enrique, Fernando, Francisco, Guillermo, Hector, Jaime, Jesus, Jorge, Jose, Juan, Juanita, Loyola, Luis, Mario, Miguel, Pedro, and Rodrigo.

I used the available information to code for four groups that had large enough sample sizes to analyze separately; Asian, Black, Latino, and White, with everyone else grouped under “Other.” When possible, I followed the coding rules for Latinos in *Facing Reality*: Those who self-identify Blacks or Asians but also self-identify as Latino are classified as Black and Asian respectively; Amerinds, Whites, those of an “other” single race, or mixed race who self-identify as Latinos are classified as Latinos. In all the cities, it was possible to discriminate between White Latinos (classified as Latino) and Non-Latino Whites (classified as Whites).

year (numeric)

The year during which the arrest took place.

crime (string)

1 Murder

2 Rape

3 Robbery

4 Agg Ass [Aggravated Assault]

latitude & longitude (numeric)

Note that the despite the potential geographic precision of the coordinates, police routinely entered the coordinates of an intersection near which the arrest occurred or some other generic code (e.g., the entrance to large park or public housing project).

location (string)

police — Denotes the coordinates of a police station or substation.

jail — Denotes the coordinates of a jail or other detention facility.

courthouse — Denotes the coordinates of a facility with courtrooms or, in the case of Fayetteville, a probation office.

hospital — Denotes the coordinates of a medical facility.

Examination of the distribution of **latitude** and **longitude** combinations revealed that the location of many arrests was a police station, correctional facility, courthouse, or medical facility. This poses a problem when using the zip code as the unit of analysis. It is likely that a large majority of the offenses that resulted in an arrest at a hospital, jail, or courthouse did not occur within the zip code where the hospital, jail, or courthouse is located. For arrests that are located in a police station, many probably involve offenses that occurred in the same zip code as the police station, but usually not all of them.

A variety of approaches might be devised for dealing with this problem. My own inclination is to exclude arrests at hospitals, jails, and courthouses from analyses using the zip code as the unit of analysis and try to devise plausible methods of estimating the proportions of arrests at police stations that should be assigned to the zip code. In making those choices, much depends on the specific city, the size the local zip codes, and the number of police substations within the jurisdiction. I also plan to make use of information in **addresspd** (see below) that can be used to identify arrests with an address that is clearly not that of the law enforcement or medical facility where the arrest was recorded.

New York City poses the biggest analytic challenges. The other cities made much less use of police stations as the location of arrests than New York.

location could not be coded for Fort Lauderdale or Tucson because the arrest databases did not include the latitude and longitude coordinates. Tucson did supply GIS coordinates for some arrests that could presumably be converted to latitude and longitude, but with so many missing values that I did not try to do so.

---

datepd (string)

The date of the arrest supplied by police database. Different cities use different formats.

---

chargepd (string)

The description of the offense recorded by the police department and used to classify offenses as falling into one of the **crime** categories. Note that the classification for some cities employed information from additional variables in the police database (e.g., classification of the offense as a felony) or the legal definition of 1st, 2nd, and 3rd degree offenses in a given state.

---

addresspd (string)

Some cities (not New York, unfortunately) included street addresses in their database. Used in conjunction with **latitude** and **longitude**, **addresspd** can be used to identify arrests coded in **location** for "police/jail/court" or "hospital" that probably were offenses that occurred in the zip code in question, not at the facility.

---

arrests (numeric)

The total number of arrests for violent offenses in the zip code during the years covered by the database.

Arrest Demographics

---

area (numeric)

Geographic area of **zip** expressed in square miles.

---

density (numeric)

Number of people per square mile, calculated by the Census Bureau.

---

hh (numeric)

Number of households in **zip**. A household can contain more than one family.

sescentile (numeric)

The percentile of zip on an index of socioeconomic status (SES) with a potential range from 0.001 to 99.999. In this dataset, the range goes from 0.054 to 99.999. The index combines standardized score for median family income and the percentage of persons who have at least a bachelor's degree expressed as the percentage persons ages 25 and older. Calculation of the SES index is described in Charles Murray, *Coming Apart: The State of White America 1960–2010*, pp. 315–317. The data for **sescentile** in this datafile come from the ACS combined datafile for 2014–2018.

When interpreting results, note that **sescentile** represents where individuals within a zip code fit within the national population of individuals, not where the zip code as a whole fits within the national set of zip codes.

mfi (numeric)

Median family income as reported in the 2014–2018 combined ACS, expressed in 2018 dollars.

mhhi (numeric)

Median household income as reported in the 2014–2018 combined ACS, expressed in 2018 dollars.

Educational Attainment Variables (numeric)

Each of six variables expresses the number of persons in the zip code whose highest degree falls in that category. The categories are:

hsorless. No more than a high school diploma or GED. Includes persons who did not attain either.

somecollege. Includes persons who attained an AA or some other intermediate educational credential or who attended but did not graduate from a four-year college or university.

ba. Persons with a bachelor's degree

ma. Persons with a master's degree.

phd. Persons with a PhD.

profdeg. Persons with an MD, JD or equivalent, DDS, or DVM.

Population Variables (numeric)

pop. Total population of the zip code.

pop25. Population ages 25 or older. The percentage of persons with a BA that is part of **sescentile** uses this variable as the denominator for the sum of **ba**, **ma**, **phd**, and **profdeg**.

Race of Persons Who Do Not Identify as Latinos (number of persons in the zip code)

amerind. Amerindian or Alaskan native.

asian. Asian

black. Black

hipac. Native Hawaiian or other Pacific Islander

white. White

other. Some other race alone.

mixed. Two or more races.

Race of Persons Who Identify as Latinos (number of persons in the zip code)

lamerind. Amerindian or Alaskan native.

lasian. Asian

lblack. Black

lhipac. Native Hawaiian or other Pacific Islander

lwhite. White

lother. Some other race alone.

lmixed. Two or more races.