# Newly developed datasets

## National school counselor survey (nscs.csv and nscs\_missing.csv)

To gain a broad picture of the counselor experience during the early months of COVID-19 school closures, [Savitz-Romer and Rowan-Kenyon (2021)](https://doi.org/10.3886/E145681V1-98740) created an 80-question online survey to collect data from 1,060 school counselors and educators in adjacent roles (e.g., college counselors, adjustment counselors, counseling directors) in 48 states and Puerto Rico in the US. For simplicity reasons, we only include in our dataset the key variables as following (dropping incomplete observations resulted in N = 787):

* id, counselor identification number
* grades, school level, 0 = elementary, 1 = middle school, 2 = high school, 3 = mixed
* schooltype, school type, 1 = public, 2 = charter, 3 = private
* urbanicity, school urbanicity level, 1 = urban, 2 = suburban, 3 = rural
* female, counselor gender coded one for female and zero for others
* exp, counselor experience measured by years
* ooo, dichotomous variable coded one if the counselor had less time for one-on-one counseling due to COVID
* group, dichotomous variable coded one if the counselor had less time for group counseling due to COVID
* sel, dichotomous variable coded one if the counselor had less time for supporting student social emotional needs due to COVID

## School principal data (principal.csv)

This dataset was drawn from Meta Krüger’s (1994) study investigating gender differences in school leadership in Netherlands. The study implemented a matching procedure to generate pairs of comparable schools (N=98) with the only difference being that one school principal was female and the other male, then surveyed the school principals, teachers, and students. Due to deletion of incomplete observations, our dataset is a bit smaller than the one used in the article, it contains 94 school principals and 800 students.

The dataset contains nine variables, detailed below.

* schid, school identification number
* stuid, student identification number within each school
* female, coded one for female principals and zero for male principals
* stufemale, coded one for female students and zero for male students
* match, coded one if student and principal are the same gender
* q1, students’ rating on a survey item, “Sometimes the principal talks to me”, on a 4-point likert scale, 1=low, 4=high
* q2, students’ rating on a survey item, “I think the principal knows who I am”, on a 4-point likert scale, 1=low, 4=high
* q3, students’ rating on a survey item, “School principal knows how well I am doing”, on a 4-point likert scale, 1=low, 4=high
* q4, students’ rating on a survey item, “The school principal is nice”, on a 4-point likert scale, 1=low, 4=high

## College GPA data (gpa.csv and gpa\_above.csv)

The GPA data is a longitudinal dataset, where 200 college students were followed for six consecutive semesters. The data was simulated. Note that this dataset contains incomplete observations.

For gpa.csv, key variables are:

* stuid, student identification number
* time, a factor variable indexing the six consecutive semesters
* gpa, student’s GPA at the end of semester
* female, coded one for female students and zero for male
* hsgpa, high school GPA

For gpa\_above.csv, key variables are:

* stuid, student identification number
* gpa\_above, a binary variable coded one for students whose overall college GPA was above 3.0 and zero otherwise
* female, coded one for female students and zero for male
* hsgpa, high school GPA

## GALO data (particularly for Unit 5 dealing with missing data) (galo.csv)

The GALO data is originally described by Peschar (1975) and analyzed by Dronkers and Schijf (1994) among other studies. The 1959 cohort consists of 1270 school children in the sixth grade of 37 elementary schools in the city of Groningen (Netherlands).

Variables are detailed below.

* schid, school identification number
* female, coded one for female students and zero for male
* galo, student’s achievement score on GALO test
* advice, teacher’s advice about secondary education; 0=“no subsequent school”, 1=“lowest”,…, 6=“highest”, 999=“missing”
* medu, mother’s highest education; 1=“lowest”,…, 9=“highest”, 999=“missing”
* fedu, father’s highest education; 1=“lowest”,…, 9=“highest”, 999=“missing”
* focc, father’s occupational status; 1=“lowest”,…, 6=“highest”, 9=“missing”

## Binge Eating in Men

This dataset is the original dataset used in a study conducted by Dr. Nichole Kelly and her colleague ([Kelly, Cotter, & Guidinger, 2018](https://pubmed.ncbi.nlm.nih.gov/29990652/)). They examined whether perceived overeating patterns were uniquely associated with eating- and weight-related comorbidities and found that young men who engage in both subjective (SBEs) and objective binge eating episodes (OBEs) may be at the highest risk for chronic disease and psychological concerns.

Analytic Sample. This dataset is individual-level data for 1114 young men (age 18-30), who were recruited from across US and finished the survey through a link sent via email. The data contains 343 incomplete observations.

The data set contains 13 variables, detailed below.

* *none*, dummy variable coded 1 for individuals without disordered eating
* *oe*, dummy variable coded 1 for individuals with OE(s) only; note: OE means overeating without presence of loss of control
* *sbe*, dummy variable coded 1 for individuals with SBE(s) only; note: SBE means subjective binge eating
* *obe*, dummy variable coded 1 for individuals with OBE(s) only; note: OBE means objective binge eating
* *sobe*, dummy variable coded 1 for individuals with both OBE(s) and SBE (s)
* *dietary*, mean dietary restraint score (higher scores = higher restraint)
* *emotion*, measure of emotion dysregulation problems (higher scores = more problems)
* *image*, body image concern measure, higher = greater internalization of thin and muscular ideal
* *race*, Racial/ethnic identity (four categories)
* *bmi, Body mass index (in kg/m-squared)*
* *age*
* *edu, education group (nine categories)*
* *employ*, employment status (four categories)

## CEPS

I have access to and knowledge about [China Education Panel Survey (CEPS)](http://ceps.ruc.edu.cn/English/Overview/Overview.htm), China’s nationally representative, longitudinal survey of middle school students. Two-year data on 9,449 students from the initial 7th grade cohort is publicly available. CEPS data contains 400+ school-level, 200+ teacher-level, 300+ parent-level, and 400+ student-level variables. We could talk more about what we need for this class, including lectures and assignments, and I can work on the datasets accordingly.

# Datasets used across EDUC 641 & 643

## 1. Teacher Professional Development Study (PDS)

This dataset is drawn from the NCRECE [Teacher Professional Development Study (PDS)](https://www.icpsr.umich.edu/web/ICPSR/studies/34848) database. This study was a randomized controlled evaluation of two forms of professional development (PD) - coursework (phase 1) and consultancy (phase 2) - delivered to about 490 early childhood education teachers across the nation. These PD supports aimed to improve teachers’ implementation of language/literacy activities and interactions with children, as well as promote gains in children’s social and academic development.

Our dataset is student-level data for 440 preschool students. Since PDS student outcome data has not been collected during phase 1, we focus on phase 2 intervention. Specifically, the sample students were in phase 1 treatment group and in either the treatment or control group during phase 2. Observations with missing values on any of the key variables were deleted for simplification reasons.

The data set contains 11 variables, detailed below.

* *childid*, unique identification number for the student
* *tchid*, unique identification number for the student’s teacher
* *schid*, unique identification number for the student’s school
* *cohort*, coded 1 for cohort 1 students (who started the experiment in spring 08) and 2 for cohort 2 students (who started the experiment in spring 09)
* *treat*, coded 1 for students in phase 2 treatment group and 0 for those in phase 2 control group
* *vocabulary*, the student’s post phase 2 score on a receptive vocabulary test (Peabody Picture Vocabulary Test-3rd edition)
* *female*, coded 1 for female students and 0 for male students
* *language\_eng*, coded 1 for students whose primary language is English and 0 otherwise
* *disability*, coded 1 for students with a disability and 0 otherwise
* *ethnicity*, eight-category variable documenting the student’s ethnicity
* *mother\_edu*, the student’s mother’s total years of education

## 2. Add Heath

This dataset is drawn from the third and fourth waves of National Longitudinal Study of Adolescent to Adult Health ([Add Health](https://addhealth.cpc.unc.edu/data/#public-use)) public-use data. Add Health is a longitudinal study of a nationally representative sample of over 20,000 adolescents (the public-use data sample is much smaller in size) who were in grades 7-12 during the 1994-95 school year, containing rich demographic, social, familial, socioeconomic, behavioral, psychosocial, cognitive, and health information. When the third (2001) and fourth (2008) waves took place, the participants were aged from 18-26 and from 24-32.

[Kraft et al (2021)](https://www.edworkingpapers.com/ai21-441) took advantage of this data and examined an understudied area - the pathway through which mentoring relationship between students and school personnel influences human capital development. Particularly, they found that having a school-based mentor is associated with an increase of 0.24 GPA points at high-school, a full year of additional education, and an increase of $60,600-$92,400 in lifetime earnings.

The dataset you’ll be using is individual-level data for Add Health participants who have no missing data on our variables of interested in both waves 3 and 4. Note that for simplifying reasons, we use smaller datasets than Kraft et al does and the measures of predictor and outcome variables are slightly different. Also we do not ask you to fit the fixed-effects models in Kraft et al, therefore all the results you get from this exercise cannot be interpreted as causal effects.

**Key variables:**

* id*, the individual’s unique identification number.*
* mentor*, binary variable, coded 1 for individuals who reported to have a school-based mentor (teachers/guidance counselors and coaches/athletic directors). This variable exists only in ah01.csv dataset.*
* mentee\_age*, the individual’s age when the mentor started to have an impact in their life. This variable exists only in ah02.csv dataset.*
* education*, the individual’s total years of education (recoded from Add Health data using the same approach in Kraft et al p.14).*
* income\_log*, the individual’s log-transformed annual earnings (including salary and all types of other incomes).*
* gpa*, the individual’s high school final GPA points.*
* gpa\_3*, binary variable coded 1 for individuals whose high school final GPA was above 3.0*

## 3. Stanford Education Data Archive (SEDA)

The data set we’ll be using in our final project was drawn from the Stanford Education Data Archive ([SEDA](https://edopportunity.org)) version 4.1. SEDA was launched in 2016 to provide nationally comparable, publicly available test score data for U.S. public school districts, allowing scientific inquiries on the relationships between educational conditions, contexts, and outcomes (especially student math/ELA achievements) at the district-level across the nation. It contains rich variables including measures of academic achievement and achievement gaps for school districts and counties, as well as district-level measures of racial and socioeconomic composition, racial and socioeconomic segregation patterns, and other features of the schooling system. Some descriptive findings can be found [here](https://edopportunity.org/discoveries/).

Due to the large size of SEDA full data set, we focus on the school year 2017-18 data for the state of Oregon. Specifically, our data set is district-level data for 103 Oregon school districts. Observations with missing values on any of the key variables were deleted for simplification reasons.

The data set contains 13 variables, detailed below.

* *district*, name of the district
* *subject*, coded “mth” fo Mathematics and “rla” for English/Language Arts
* *grade*, coded 3, 4, 5, 6 for grades 3-6
* *achievement*, grade-level average achievement test score
* *gap\_gender*, grade-level male-female gender gap on achievement test
* *percent\_ell*, district-level percentage of ELL students
* *percent\_sped*, district-level percentage of students in special education program
* *percent\_frl*, district-level percentage of students eligible for free or reduced school lunch
* *percent\_native*, district-level percentage of Native American students
* *percent\_asian*, district-level percentage of Asian students
* *percent\_hispanic*, district-level percentage of Hispanic students
* *percent\_black*, district-level percentage of African American students
* *percent\_white*, district-level percentage of white students