Using R and Shiny to Analyze National Assessment Data

NCES Lead: Emmanuel Sikali
Team Leads: Paul Bailey & Ting Zhang
Development Team: Michael Lee, Huade Huo, Trang Nguyen, Thomas Fink, & Sun-Joo Lee
Quality Control Team: Michael Cohen, Jiayi Li, Yuqi Liao, & Qingshu Xie
September 2019
Agenda

1. Context
2. Statement of problem
3. Demo R Package - EdSurvey
4. Demo Shiny App - interactive NCES Results Display (iNRD)
Context - NCES and NAEP
The National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education in the U.S. and other nations.

The National Assessment of Educational Progress (NAEP) is the largest nationally representative and continuing assessment of what America's students know and can do in various subject areas.

Analyzing NAEP data for students achievement involves:

- Using plausible values to account for uncertainty in the estimates of student test scores; and
- Employing special methods (e.g., jackknife) that account for the NAEP complex sampling design.
Statement of Problem
Statement of Problem

1. Cost: requires specialized statistical software

2. Complexity:
   - survey data size can be unwieldy on some machines
   - custom code required to perform analyses

3. Access Limitations: impediments to downloading and reading-in data

4. Inefficiency: cognitive burden creating and reproducing analyses

5. Software Fatigue: data downloading, processing, manipulation and analysis all done in different software
Statement of Problem

Analyzing NCES data is too costly, complex, inaccessible, and burdensome for researchers.
What is EdSurvey?

An R statistical package that gives users functions to connect seamlessly with NCES data to perform analyses that take into account complex sample survey design and the use of plausible values.
Why EdSurvey?

1. **One-stop shop** for data downloading, processing, manipulation and analysis of survey data.

2. **Automated**: Weights and complex sampling design calculations are automated following standard NCES methodology.

3. **Simple**: e.g., a regression with 90 replicate weights requires only a few lines of code.

4. **Flexible**: You can use functions that rely on EdSurvey methods or get the data and use traditional R.

5. **Minimizes memory footprint** by only reading in required data.
One-stop Shop for NCES Survey Analysis

1. Understand the data

2. Clean and manipulate data
   - in EdSurvey

2. Extract and manipulate data
   - in R

3. Run analyses
One-stop Shop for NCES Survey Analysis

1. Understand the data
   - Explore: explore the codebook, see the variables with plausible values, see weights.
   - Search: search variables.
   - Expand: see variable levels, tabulate response percentages, see assessment scores by response category, summarize continuous variables.

2. Clean and manipulate data
   - in EdSurvey

2. Extract and manipulate data
   - in R

3. Run analyses
One-stop Shop for NCES Survey Analysis

1. Understand the data
   • Explore: explore the codebook, see the variables with plausible values, see weights.
   • Search: search variables.
   • Expand: see variable levels, tabulate response percentages, see assessment scores by response category, summarize continuous variables.

   in EdSurvey

2. Clean and manipulate data with built-in subset and recode features.

   2. Extract and manipulate data as a data frame (for experienced R users.)

   in R

3. Run analyses
One-stop Shop for NCES Survey Analysis

1. Understand the data
   • Explore: explore the codebook, see the variables with plausible values, see weights.
   • Search: search variables.
   • Expand: see variable levels, tabulate response percentages, see assessment scores by response category, summarize continuous variables.

   ![Diagram showing the flow from EdSurvey to R for data understanding, cleaning, and manipulation]

2. Clean and manipulate data in EdSurvey with built-in subset and recode features.

3. Extract and manipulate data as a data frame in R (for experienced R users).

2. Extract and manipulate data in R

3. Run analyses such as regression analysis, logit analysis, mixed models, show gaps, calculate achievement levels, correlate variables, calculate percentiles.
Demo - EdSurvey
Installing the EdSurvey Package

# to install the package
install.packages("EdSurvey")

# to load the package
library(EdSurvey)
Data Processing
Data Processing

- First, read in the data

```r
sdf <- readNAEP(system.file("extdata/data", "M36NT2PM.dat", package = "NAEP"), package = "NAEP")
```
Meet Your Data - print

```
print()

- Print returns detailed data file information:

```
```r
print(sdf)

# edsurvey.data.frame for 2005 NAEP (Mathematics) in USA
#
# Dimensions: 17606 rows and 302 columns.
#
# There are 1 full sample weight(s) in this edsurvey.data.frame
# 'origwt' with 62 JK replicate weights (the default).
#
# There are 6 subject scale(s) or subscale(s) in this edsurvey.data.frame
# 'num_oper' subject scale or subscale with 5 plausible values.
# 'measurement' subject scale or subscale with 5 plausible values.
# 'geometry' subject scale or subscale with 5 plausible values.
# 'data_anal_prob' subject scale or subscale with 5 plausible values.
# 'algebra' subject scale or subscale with 5 plausible values.
# 'composite' subject scale or subscale with 5 plausible values (the default).
#
# Omitted Levels: 'Multiple', 'NA', 'Omitted'
#```
Data Exploration and Cross Tabulation

- Summary table of NAEP composite mathematics performance scale scores (composite) of 8th grade students by two student factors:
  - **dsex**: gender
  - **b017451**: frequency of talk about studies at home

```r
es1 <- edsurveyTable(composite ~ dsex + b017451, data = sdf)
```

<table>
<thead>
<tr>
<th>dsex</th>
<th>b017451</th>
<th>N</th>
<th>WTD_N</th>
<th>PCT</th>
<th>SE(PCT)</th>
<th>MEAN</th>
<th>SE(MEAN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Never or hardly ever</td>
<td>2350</td>
<td>2434.844</td>
<td>29.00978</td>
<td>0.6959418</td>
<td>270.8243</td>
<td>1.057078</td>
</tr>
<tr>
<td>Male</td>
<td>Once every few weeks</td>
<td>1603</td>
<td>1638.745</td>
<td>19.52472</td>
<td>0.5020657</td>
<td>275.0807</td>
<td>1.305922</td>
</tr>
<tr>
<td>Male</td>
<td>About once a week</td>
<td>1384</td>
<td>1423.312</td>
<td>16.95795</td>
<td>0.5057265</td>
<td>281.5612</td>
<td>1.409587</td>
</tr>
<tr>
<td>Male</td>
<td>2 or 3 times a week</td>
<td>1535</td>
<td>1563.393</td>
<td>18.62694</td>
<td>0.4811497</td>
<td>284.9066</td>
<td>1.546072</td>
</tr>
<tr>
<td>Male</td>
<td>Every day</td>
<td>1291</td>
<td>1332.890</td>
<td>15.88062</td>
<td>0.5872731</td>
<td>277.2597</td>
<td>1.795784</td>
</tr>
<tr>
<td>Female</td>
<td>Never or hardly ever</td>
<td>1487</td>
<td>1517.609</td>
<td>18.20203</td>
<td>0.5078805</td>
<td>266.7897</td>
<td>1.519020</td>
</tr>
</tbody>
</table>

Linear Regression

- fits a linear model formula using sampling weights and variance estimates:
  - \( \texttt{dsex} \): gender
  - \( \texttt{b017451} \): frequency of talk about studies at home

```r
lm1 <- lm.sdf(composite ~ dsex + b017451, 
               weightVar = 'origwt', data = sdf) 
summary(lm1)
```

```
##
## Formula: composite ~ dsex + b017451
## jrrIMax: 1
## Weight variable: 'origwt'
## Variance method: jackknife
## JK replicates: 62
## full data n: 17606
## n used: 16331
##
## Coefficients:
##                               coef     se       t   dof          Pr(>|t|)
## (Intercept)                    270.41112 1.02443 263.96149 54.670 < 0.00000000000000022 ***
```

```
EdSurvey Calls Network Connection

Small Memory Footprint

Step 1
- Original data on network drive

Step 2
- Read in relevant rows and columns
- R processes the data

Step 3
- Memory is emptied after processing

UseR call:
```r
> lm(composite ~ dsex)
```
Extensive Function Library

- Correlations
- Proportion meeting expectations
- Regression
- Logit/probit
- Quantile regression
- Percentiles
- Mixed models
- Gap trend analysis
Learning EdSurvey

There are many ways to learn about how to use EdSurvey:

- Frequently offered workshops
- Reading vignettes on the EdSurvey Website (https://www.air.org/page/edsurvey-installation-and-use)
- Accessing vignettes via the package (requires internet)

```r
vignette("introduction", package="EdSurvey")
```

- R help

```r
help(package = "EdSurvey")
```
iN RD
iNRD

- A Shiny exploratory tool that visualizes results from one of the NCES data - National Assessment of Educational Progress (NAEP) - across year

- The iNRD is built using the R statistical software and it features an interactive interface with dropdowns, checkboxes and slider bars for ease of use

- Users can explore NAEP results and create customized plots including charts and maps that display NAEP national, state, and TUDA performances across multiple years, and broken down across a variety of student groups. Users can also download and save their plots to the png, pdf, or svg format.
• Some current NCES data visualization tools are cumbersome
  ○ In NAEP Data Explorer and International Data Explorer, a user has to follow certain procedure to achieve a goal. Any mistake in an intermediate step will mess up the whole process, and the user has to go back to the beginning and restart the analysis.
  ○ Does not support sophisticated data exploration
Trend in fourth-grade NAEP mathematics average scores by Gender
Virginia fourth-grade NAEP mathematics achievement levels by race (black)
Percent of Fourth-grade students at or above Proficient in Reading by IEP status (Y)