CYFS nebraska center for research children youth families & schools

research methodology series

An Introduction to Secondary Data Analysis

Natalie Koziol, MA

CYFS Statistics and Measurement Consultant

Ann Arthur, MS

CYFS Statistics and Measurement Consultant

Outline

- Overview of Secondary Data Analysis
- Understanding & Preparing Secondary Data
- Brief Overview of Sampling Design
- Analyzing Secondary Data
- Illustration of Secondary Data Analysis
- Other Logistical Considerations

Overview of Secondary Data Analysis

What is Secondary Data Analysis?

- "In the broadest sense, analysis of data collected by someone else" (p. ix; Boslaugh, 2007)
- Analysis of secondary data, where "secondary data can include any data that are examined to answer a research question other than the question(s) for which the data were initially collected" (p. 3; Vartanian, 2010)
- In contrast to primary data analysis in which the same individual/team of researchers designs, collects, and analyzes the data

Local Examples of Research Involving Secondary Data Analysis

- Starting Off Right: Effects of Rurality on Parent's Involvement in Children's Early Learning (Sue Sheridan, PPO)
  - Data from the Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) were used to examine the influence of setting on parental involvement in preschool and the effects of involvement on Kindergarten school readiness.
- Testing Thresholds of Quality Care on Child Outcomes Globally & in Subgroups: Secondary Analysis of QUINCE and Early Head Start Data (Helen Raikes, PPO)
  - Data from two secondary datasets were used to examine the potentially non-linear relationship between quality of child care and children's development

What are Secondary Data?

- Come from many sources
  - Large government-funded datasets (the focus of this presentation)
  - University/college records
  - Statewide or district-level K-12 school records
  - Journal supplements
  - Authors’ websites
  - Etc.!
- Available for a seemingly unlimited number of subject areas
- Quantitative (the focus of this presentation) and qualitative
- Restricted and public-use
- Direct (e.g., biomarker data) and indirect observation (e.g., self-report)

Where Can I Find Secondary Data?

Searching for secondary datasets:

- Inter-University Consortium for Political and Social Research
  - http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/index.jsp
- Data.gov
  - http://www.data.gov
- National Center for Education Statistics
  - http://nces.ed.gov
- U.S. Census Bureau
  - http://www.census.gov
- Simple Online Data Archive for Population Studies (SodaPop)
  - http://sodapop.pop.psu.edu/data-collections

Examples of Large Secondary Datasets for Education & Social Sciences Research

- Common Core of Data (CCD)
- Current Population Survey (CPS)
- Early Childhood Longitudinal Study (ECLS): Birth (ECLS-B) and Kindergarten (ECLS-K) Cohort
- General Social Survey (GSS)
- Head Start Family and Child Experiences Survey (FACES)
- Monitoring the Future (MTF)
- National Assessment of Educational Progress (NAEP)
- National Education Longitudinal Study (NELS)
- National Household Education Surveys (NHES)
- National Longitudinal Study of Adolescent Health (Add Health)
- National Longitudinal Survey of Youth (NLSY)
- National Survey of American Families (NSAF)
- National Survey of Child and Adolescent Well-Being (NSCAW)
- National Survey of Families and Households (NSFH)
- NICHD Study of Early Child Care and Youth Development (SECCYD)
- Programme for International Student Assessment (PISA)
- Progress in International Reading Literacy Study (PIRLS)
- Trends in International Mathematics and Science Study (TIMSS)
- U.S. Panel Study of Income Dynamics (PSID): Child Development Supplement (CDS)

Advantages of Secondary Data Analysis

- Study design and data collection already completed
  - Saves time and money
    - Access to international and cross-historical data that would otherwise take several years and millions of dollars to collect
    - Ideal for use in classroom examples, semester projects, masters theses, dissertations, supplemental studies
- Data may be of higher quality
  - Studies funded by the government generally involve larger samples that are more representative of the target population (greater external validity!)
  - Oversampling of low prevalence groups/behaviors allows for increased statistical precision
- Datasets often contain considerable breadth (thousands of variables)

Disadvantages of Secondary Data Analysis

- Study design and data collection already completed
  - Data may not facilitate particular research question
  - Information regarding study design and data collection procedures may be scarce
- Data may *potentially* lack depth (the greater the breadth the harder it is to measure any one construct in depth)
  - Constructs may be operationally defined by a single survey item or a subset of test items which can lead to reliability and validity concerns
  - ‘Post hoc’ attempts to construct measurement models may be unsuccessful (survey items may not hang together)
- Certain fields or departments (e.g., experimental programs) may place less value on secondary data analysis
- May require knowledge of survey statistics/methods which is not generally provided by basic graduate statistics courses

Understanding & Preparing Secondary Data

Understanding Secondary Data

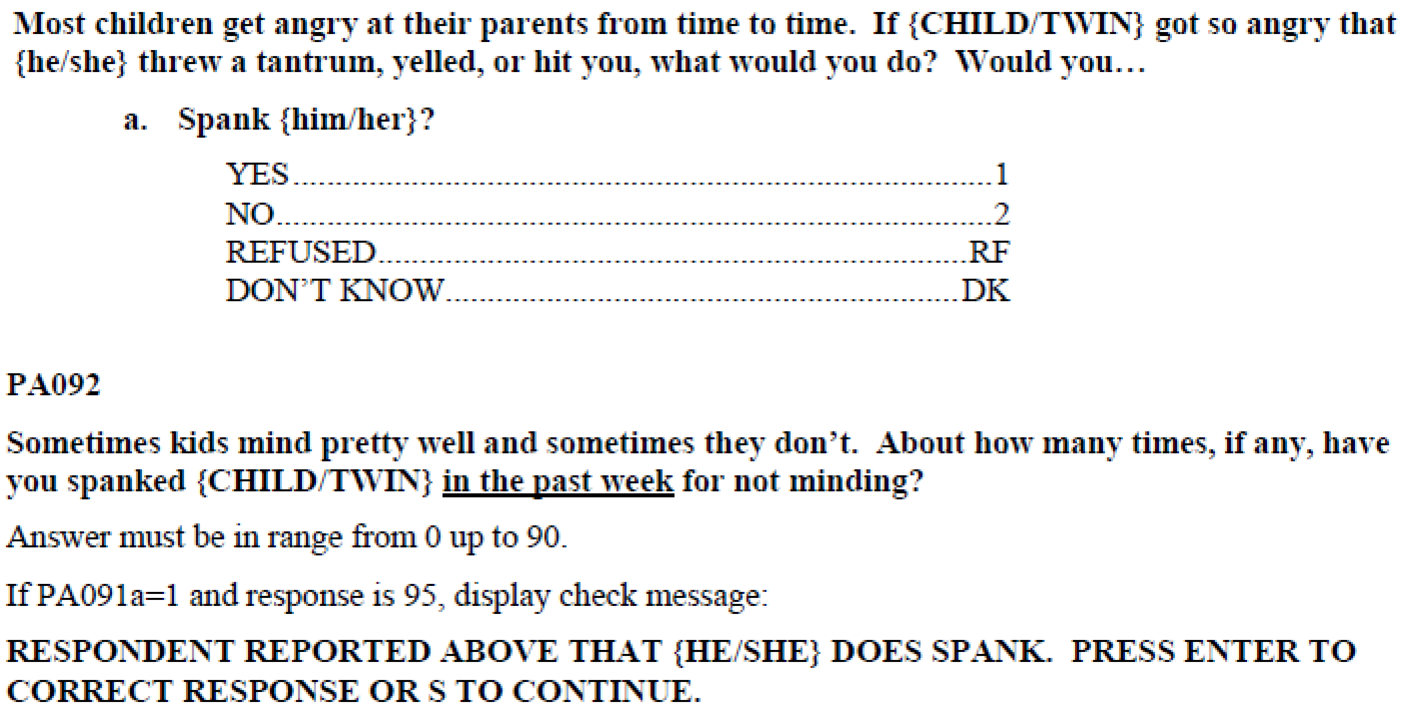
Familiarize yourself with the original study and data!

- Read all User’s/Technical manuals
  - To whom are the results generalizable?
    - E.g., ECLS-B analyses involving data from kindergarten wave can be used to make inferences about children born in the U.S. in 2001 as they enter kindergarten (not to make inferences about U.S. kindergarteners)
  - How are missing data handled?
  - What are the appropriate analysis weights?
  - What is the appropriate method (and what variables are necessary) for computing adjusted standard errors?
  - What composite variables are available and how are they constructed?

Understanding Secondary Data

Familiarize yourself with the original study and data!

- Examine questionnaires and interview protocols when available
  - Identify skip patterns to determine coding of missing data; example from the ECLS-B preschool parent interview:



Most children get angry at their parents from time to time. If {CHILD/TWIN} got so angry that {he/she} threw a tantrum, yelled, or hit you, what would you do? Would you…

a. Spank {him/her}?

YES	1
NO	2
REFUSED	RF
DON'T KNOW	DK

PA092

Sometimes kids mind pretty well and sometimes they don't. About how many times, if any, have you spanked {CHILD/TWIN} in the past week for not minding?

Answer must be in range from 0 up to 90.

If PA091a=1 and response is 95, display check message:

RESPONDENT REPORTED ABOVE THAT {HE/SHE} DOES SPANK. PRESS ENTER TO CORRECT RESPONSE OR S TO CONTINUE.

Understanding Secondary Data

Familiarize yourself with the original study and data!

- Examine questionnaires and interview protocols when available
  - For examining trends or growth, determine whether the same construct is being measured across time
    - Interview questions may be modified across time
      - Example from an Opinion Research Business (ORB) survey on conflict deaths in Iraq (Spagat & Dougherty, 2010):

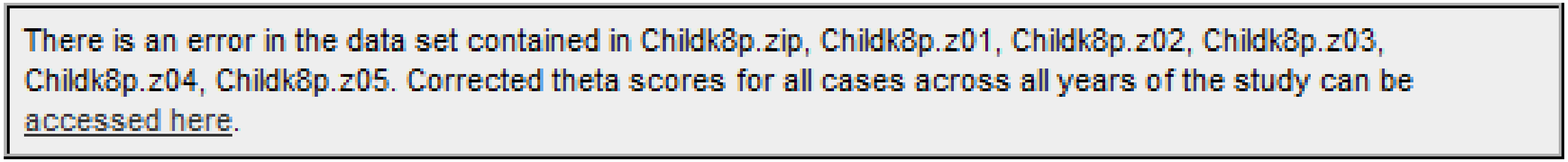
        Yes/No: There has been a "murder of a member of my family/relative" (February 2007)

        Yes/No: There has been a death "as a result of conflict/violence of a household member" (August 2007)
    - Respondents (e.g., parent/guardian) may change over time
    - Different scales may be used across time (e.g., different cognitive measures are used for infants and kindergarteners)

Understanding Secondary Data

Familiarize yourself with the original study and data!

- Check study website frequently for errors and/or updates
  - Example from http://nces.ed.gov/ecls/dataproducts.asp:



There is an error in the data set contained in Childk8p.zip, Childk8p.z01, Childk8p.z02, Childk8p.z03, Childk8p.z04, Childk8p.z05. Corrected theta scores for all cases across all years of the study can be accessed here.

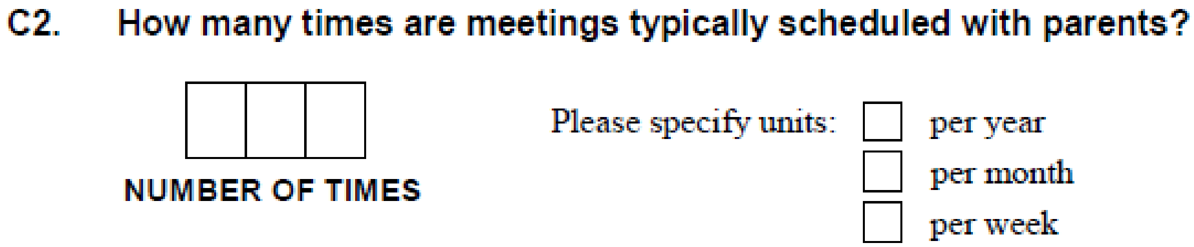
  - Ongoing panel (i.e. longitudinal) studies generally provide new datasets after each wave of data collection
    - Always use the most up-to-date file! Scores developed using item response theory may be recalibrated at each wave to permit investigation of growth

Preparing Secondary Data

- Document everything!
  - Save all syntax
  - Create an abridged codebook describing the original and recoded variables of interest
- Step 1: Transfer all potential data of interest to a new file in preferred base program
  - Electronic codebooks (ECBs) greatly facilitate this process
  - Never alter the original datafile!
- Step 2: Address missing data
  - Identify/label missing values in software program
  - When possible, use knowledge of skip patterns to recode missing data as meaningful values
  - Select method for handling missing data (e.g., multiple imputation, full-information maximum likelihood [FIML])

Preparing Secondary Data

- Step 3: Recode variables
  - Reverse code negatively worded items if creating scale scores
  - Dummy code dichotomous variables into values of 0, 1 (original dataset may use values of 1, 2)
  - Recode other categorical variables (e.g., dummy or effect coding)
  - Combine separate but like variables
    - E.g., ECLS-B contained 2 kindergarten waves (only 75% of children were in kindergarten in 2006); to analyze kindergarteners, need to combine variables from waves 4 and 5 using “if-else” commands
  - Recode variables so that all responses are based on the same units
    - Example from ECLS-B Preschool Center Director Questionnaire:



C2. How many times are meetings typically scheduled with parents?

□□□

NUMBER OF TIMES

Please specify units: □ per year
□ per month
□ per week

Preparing Secondary Data

- Step 4: Create new variables
  - May need to recreate composite variables if disagree with original conceptualization
    - E.g., An "SES" variable in the original datafile may be constructed from "income" and "parent education" variables; secondary researcher may want to construct new SES variable
  - Psychometric work
    - Create scores from individual items using factor analysis or item response theory
      - Unfortunately, individual survey items do not always hang together
      - To avoid potentially biased variance estimates,

        (a) incorporate measurement models directly into analysis, or
        (b) output "plausible values" (e.g., Mislevy et al., 1992)

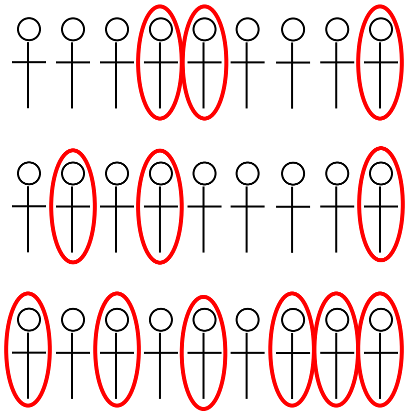
A *Brief* Overview of Sampling Design

Sampling Design

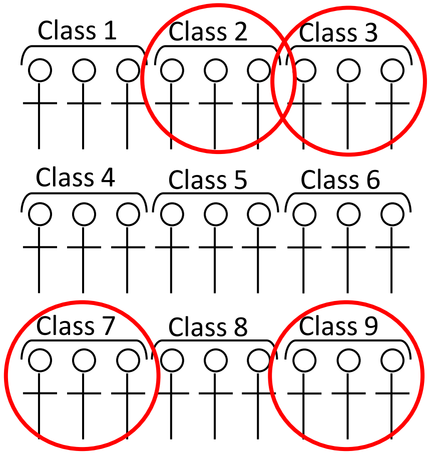
- Ideally, we want a sample that is perfectly representative of our target population (we want to use sample results to make inferences, or generalizations, about a larger population)
- Types of probability sampling
  - Simple random sampling
    - Randomly sample individuals
  - Stratified sampling
    - Divide population into strata (groups); within *each* stratum, randomly sample individuals
  - Cluster sampling
    - Population contains naturally occurring groups (e.g., classrooms); randomly sample groups

Sampling Design

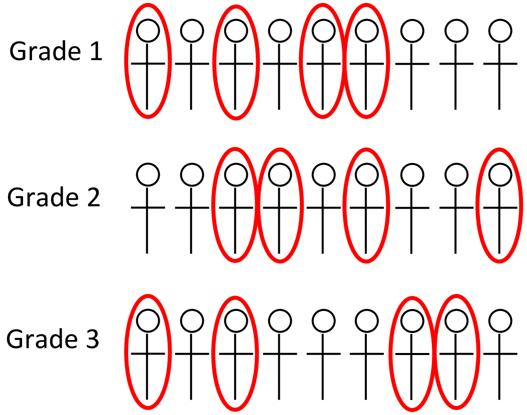
Simple Random Sampling



Cluster Sampling



Stratified Sampling



Sampling Design

- Simple random sampling
  - Assumed when performing conventional statistical analyses
  - No guarantee of a representative sample
  - May not be feasible (e.g., costly, impractical)
- Stratified sampling
  - More control over representativeness
  - Allows for intentional oversampling which permits greater statistical precision (i.e., decreases standard errors)
- Cluster sampling
  - May be necessary (e.g., educational interventions may only be possible at the classroom level)
  - Decreases statistical precision (individuals within groups tend to be more similar so we have less unique information)

Sampling Design

- Statistical analyses should reflect sampling design
  - Point estimates (e.g., means) should be adjusted to take into account unequal sampling probabilities
  - Standard errors should be adjusted to ensure correct level of confidence in point estimates
- Different statistical approaches exist for handling complex sampling designs
  - Multilevel modeling
  - Application of weights and alternative methods of variance estimation
    - Common approach when analyzing large secondary datasets due to complexity of sampling design
  - Combination of approaches

Sampling Design

- Sampling weights
  - “The reciprocal of the inclusion probability…the number of population units represented by unit *i*” (p. 39; Lohr, 2010)
    - $w_i = \frac{1}{\pi_i}$ where π_i is the probability that unit *i* is in the sample
  - Necessary for obtaining accurate/generalizable point estimates
  - Construction of sampling weights is complex (based on multiple stages of sampling, non-response, post-stratification, etc.)
    - Thankfully, large secondary datasets generally have pre-constructed weights
    - However, multiple weights may exist for any one dataset
      - Appropriate selection and application of weights is the responsibility of the secondary data analyst!

Sampling Design

- Variance estimation
  - Alternative estimation necessary for computing correct standard errors which influence tests of statistical significance
  - Does not influence point estimates
  - Multiple approaches
    - Taylor series linearization method
      - Involves specifying cluster and stratum variables
    - Replication methods
      - Balanced repeated replication (BRR)
      - Jackknife replication (JK1, JK2, JKn)
      - Choice of method depends on sampling design
      - Involves specifying series of replicate weights
    - Other methods (e.g., use of generalized variance functions)
    - Use the approach recommended in the User’s manual

Analyzing Secondary Data

Analyzing Secondary Data

- Based on research question, identify appropriate statistical analysis
- Select software package that will implement analysis and account for complex sampling
- Examine unweighted descriptive statistics to identify coding errors and determine adequacy of sample size
- Identify weights
  - Make sure missing weights are set to 0
- Identify variance estimation method (and corresponding variables)
- Conduct diagnostic analyses (identify outliers, non-normality, etc.)
- Conduct primary analysis and interpret results!

Analyzing Secondary Data

- Other considerations
  - Inclusion of covariates
    - E.g., Age at time of child assessment (not always possible to collect data at target age)
  - Analysis of subpopulations (see Lohr, 2010 for more information)
    - Additional specifications necessary (e.g., “domain”, “subpop”)
    - Don’t delete cases
  - Protecting confidentiality
    - Some restricted-use datasets require unweighted sample sizes to be rounded and/or estimates based on small sample sizes to be suppressed
  - Analysis involving multiple imputation or plausible values (Asparouhov & Muthén, 2010; Enders, 2010)

Analyzing Secondary Data

Software

- Software specifically developed for analyzing complex survey data
  - Generally free
  - Generally user-friendly but may lack flexibility (limited to certain datasets, limited statistical analyses)
  - Useful for initial data exploration (particularly restricted data)
  - Examples
    - NCES tools for computing descriptive statistics, regressions
      - PowerStats: http://nces.ed.gov/datalab/
      - Data Analysis System (DAS): http://nces.ed.gov/das/
    - AM Statistical Software
      - http://am.air.org/
      - Descriptives, regression, some latent variable estimation
      - Relatively easy to incorporate plausible values

Analyzing Secondary Data

Software

- General-purpose software that can account for complex sampling
  - Can be expensive (R is free)
  - Generally syntax-based rather than drop-down menu
  - More flexible
  - Examples:
    - SAS (certain analyses require SUDAAN add-on)
      - http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_introsamp_sect001.htm
      - http://www.rti.org/sudaan/onlinehelp/sudaan10/default.htm
    - Stata
      - http://www.stata.com/features/survey-data/

Analyzing Secondary Data

Software

- General-purpose software that can account for complex sampling
  - Examples:
    - SPSS (requires Complex Samples add-on)
      - http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fsyn_csglm_.htm
    - R
      - http://cran.fhcrc.org/web/packages/survey/index.html
    - M*plus*
      - http://www.statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v6.pdf (pp. 499-505, 521)
    - Other software options
      - http://www.hcp.med.harvard.edu/statistics/survey-soft/

Example SAS Syntax

```
*Taylor series linearization method;
PROC SURVEYMEANS data=yourdata varmethod=taylor;
strata stratavar;
cluster clustervar;
var varofinterest;
weight wtvar;
run;

PROC SURVEYREG data=yourdata varmethod=taylor;
strata stratavar;
cluster clustervar;
model outcomevar=predictorvar;
weight wtvar;
run;

*Jackknife method;
PROC SURVEYMEANS data=yourdata varmethod=jk;
repweights repwt1-repwtn;
var varofinterest;
weight wtvar;
run;

PROC SURVEYREG data=yourdata varmethod=jk;
model outcomevar=predictorvar;
repweights repwt1-repwtn;
weight wtvar;
run;
```

*Jackknife syntax varies by type (JK 1, 2, or n)

Example Stata Syntax

```
/*Taylor series linearization method*/
svyset [pweight = wtvar], psu(clustervar) strata(stratavar) vce(linearized)
svy: mean varofinterest
svy: regress outcomevar predictorvar

/*Jackknife method*/
svyset [pweight = wtvar], jkrw(repwt1 - repwtn) vce(jack) mse
svy: mean varofinterest
svy: regress outcomevar predictorvar
```

*Jackknife syntax varies by type (JK 1, 2, or n)

Example R Syntax

```
#Taylor series linearization method
library(survey)
design1 <- svydesign(id=~clustervar, strata=~stratavar, weights=~wtvar,
          data=yourdatafile)
svymean(~varofinterest, design.1)
regmodel <- svyglm(outcome~predictor, design=design.1)

#Jackknife method
library(survey)
design.2 <- svrepdesign(repweights=yourdatafile[,repwt1:repwtn], type="JK1",
          weights=yourdatafile$wtvar)
svymean(~varofinterest, design.2)
regmodel <- svyglm(outcome~predictor, design=design.2)
```

*Jackknife syntax varies by type (JK 1, 2, or n)

Example M*plus* Syntax

```
!Taylor series linearization method;
TITLE: Example complex sample syntax;
DATA: FILE=yourfile;
VARIABLE:
NAMES=outcomevar predictorvar stratavar clustervar wtvar;
USEVARIABLES=outcomevar predictorvar stratavar clustervar wtvar;
STRATIFICATION=stratavar;
CLUSTER=clustervar;
WEIGHT=wtvar;
ANALYSIS:
TYPE=COMPLEX;
MODEL:
outcomevar ON predictorvar;
OUTPUT: SAMPSTAT;

!Jackknife method;
TITLE: Example complex sample syntax;
DATA: FILE=yourfile;
VARIABLE:
NAMES=outcomevar predictorvar wtvar repwt1-repwtn;
USEVARIABLES=outcomevar predictorvar wtvar repwt1-repwtn;
WEIGHT=wtvar;
REPWEIGHTS=repwt1-repwtn;
ANALYSIS:
TYPE=COMPLEX;
REPSE=JACKKNIFE1;
MODEL:
outcomevar ON predictorvar;
OUTPUT: SAMPSTAT;
```

*Jackknife syntax varies by type (JK 1, 2, or n)

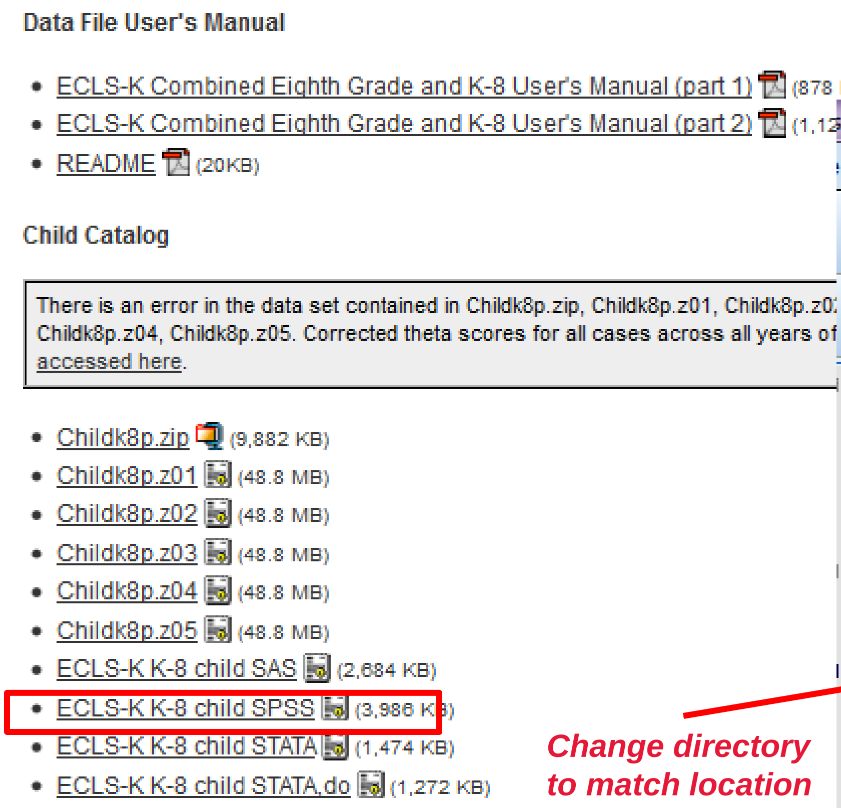
An Illustration of Secondary Data Analysis

Illustration

- Early Childhood Longitudinal Study – Kindergarten Class of 1998-99 (ECLS-K)
- Two research questions:
  - Does the number of children’s books in the home predict a child’s “Tell Stories” score as measured in the fall of kindergarten?
  - What is the average trajectory of math achievement as measured in kindergarten through 8th grade?

Illustration: Download Data & Import into Base Program

*Download data from http://nces.ed.gov/ecls/dataproducts.asp



Data File User's Manual

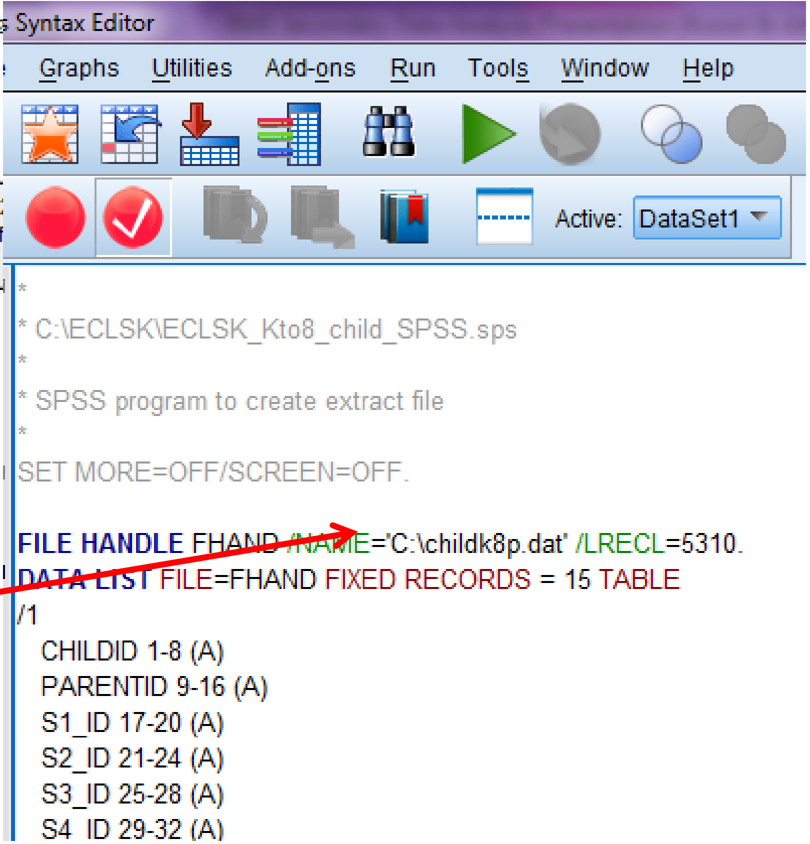
- ECLS-K Combined Eighth Grade and K-8 User's Manual (part 1) (878 KB)
- ECLS-K Combined Eighth Grade and K-8 User's Manual (part 2) (1,12
- README (20KB)

Child Catalog

There is an error in the data set contained in Childk8p.zip, Childk8p.z01, Childk8p.z0
Childk8p.z04, Childk8p.z05. Corrected theta scores for all cases across all years of
accessed here.

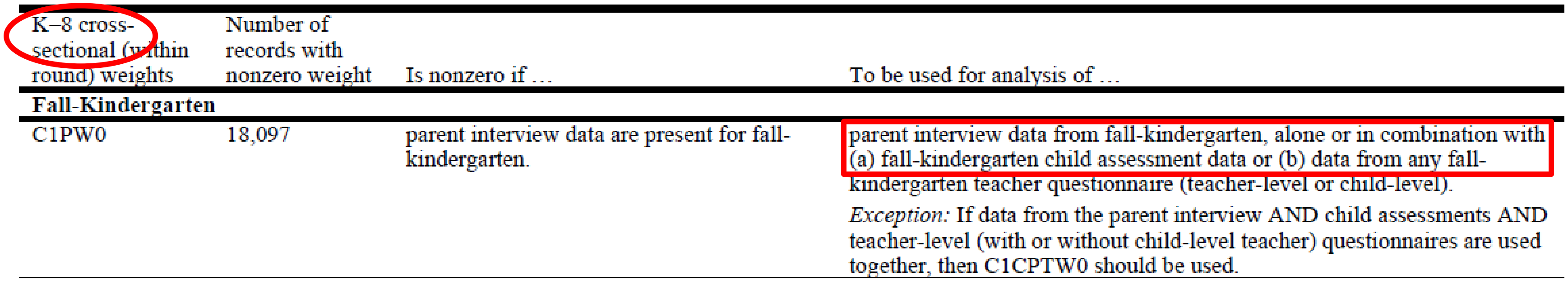
- Childk8p.zip (9,882 KB)
- Childk8p.z01 (48.8 MB)
- Childk8p.z02 (48.8 MB)
- Childk8p.z03 (48.8 MB)
- Childk8p.z04 (48.8 MB)
- Childk8p.z05 (48.8 MB)
- ECLS-K K-8 child SAS (2,684 KB)
- ECLS-K K-8 child SPSS (3,986 KB)
- ECLS-K K-8 child STATA (1,474 KB)
- ECLS-K K-8 child STATA.do (1,272 KB)

Change directory to match location of datafile

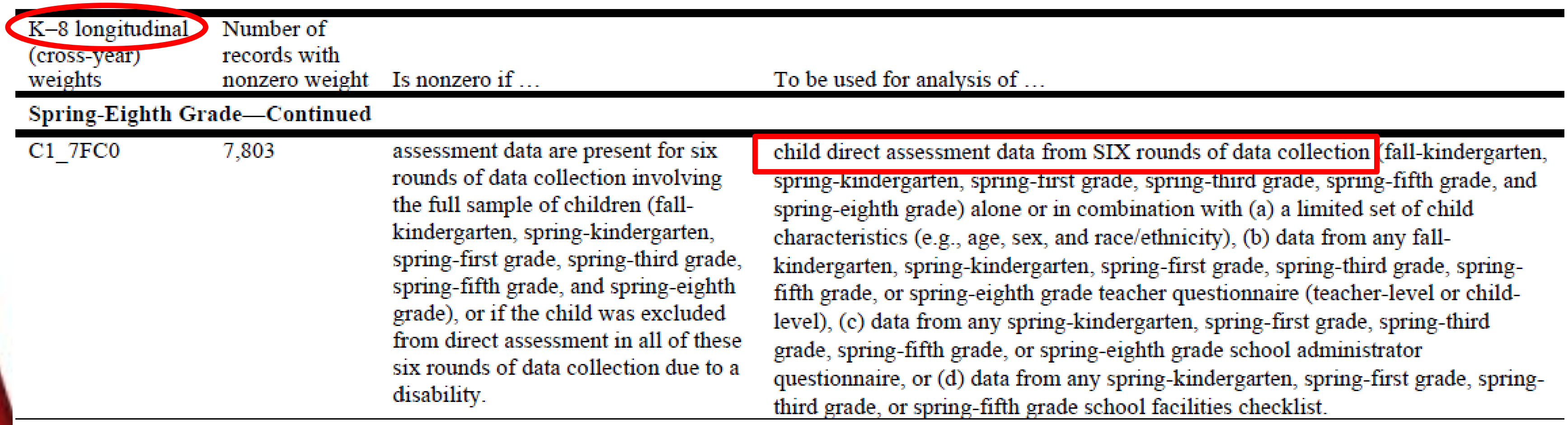


```
*
* C:\ECLSK\ECLSK_Kto8_child_SPSS.sps
*
* SPSS program to create extract file
*
SET MORE=OFF/SCREEN=OFF.

FILE HANDLE FHAND /NAME='C:\childk8p.dat' /LRECL=5310.
DATA LIST FILE=FHAND FIXED RECORDS = 15 TABLE
/1
  CHILDID 1-8 (A)
  PARENTID 9-16 (A)
  S1_ID 17-20 (A)
  S2_ID 21-24 (A)
  S3_ID 25-28 (A)
  S4_ID 29-32 (A)
```

Illustration: Identify Weights



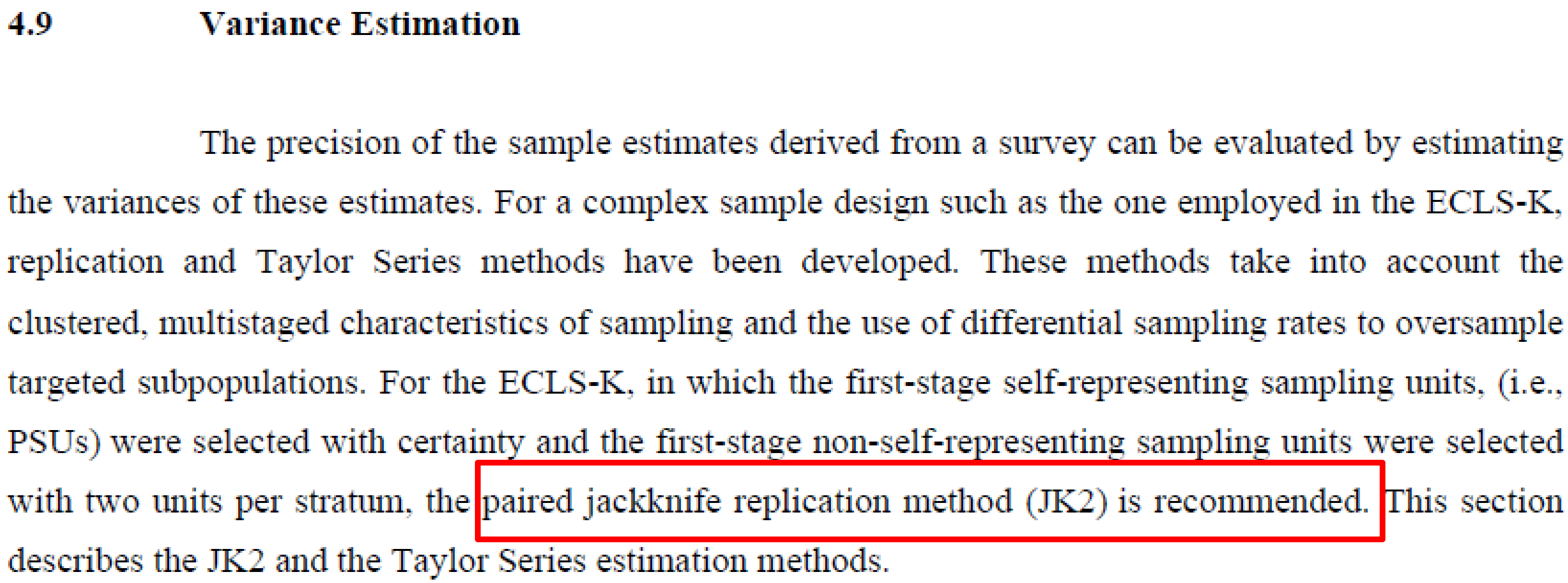
K–8 cross-sectional (within round) weights	Number of records with nonzero weight	Is nonzero if …	To be used for analysis of …
Fall-Kindergarten			
C1PW0	18,097	parent interview data are present for fall-kindergarten.	parent interview data from fall-kindergarten, alone or in combination with (a) fall-kindergarten child assessment data or (b) data from any fall-kindergarten teacher questionnaire (teacher-level or child-level). *Exception:* If data from the parent interview AND child assessments AND teacher-level (with or without child-level teacher) questionnaires are used together, then C1CPTW0 should be used.



K–8 longitudinal (cross-year) weights	Number of records with nonzero weight	Is nonzero if …	To be used for analysis of …
Spring-Eighth Grade—Continued			
C1_7FC0	7,803	assessment data are present for six rounds of data collection involving the full sample of children (fall-kindergarten, spring-kindergarten, spring-first grade, spring-third grade, spring-fifth grade, and spring-eighth grade), or if the child was excluded from direct assessment in all of these six rounds of data collection due to a disability.	child direct assessment data from SIX rounds of data collection (fall-kindergarten, spring-kindergarten, spring-first grade, spring-third grade, spring-fifth grade, and spring-eighth grade) alone or in combination with (a) a limited set of child characteristics (e.g., age, sex, and race/ethnicity), (b) data from any fall-kindergarten, spring-kindergarten, spring-first grade, spring-third grade, spring-fifth grade, or spring-eighth grade teacher questionnaire (teacher-level or child-level), (c) data from any spring-kindergarten, spring-first grade, spring-third grade, spring-fifth grade, or spring-eighth grade school administrator questionnaire, or (d) data from any spring-kindergarten, spring-first grade, spring-third grade, or spring-fifth grade school facilities checklist.

*Descriptions from ECLS-K User’s Manual

Illustration: Determine Variance Estimation Method



4.9 Variance Estimation

The precision of the sample estimates derived from a survey can be evaluated by estimating the variances of these estimates. For a complex sample design such as the one employed in the ECLS-K, replication and Taylor Series methods have been developed. These methods take into account the clustered, multistaged characteristics of sampling and the use of differential sampling rates to oversample targeted subpopulations. For the ECLS-K, in which the first-stage self-representing sampling units, (i.e., PSUs) were selected with certainty and the first-stage non-self-representing sampling units were selected with two units per stratum, the paired jackknife replication method (JK2) is recommended. This section describes the JK2 and the Taylor Series estimation methods.

*Description from ECLS-K User's Manual

Illustration: Research Question 1

- Conduct simple linear regression in SAS

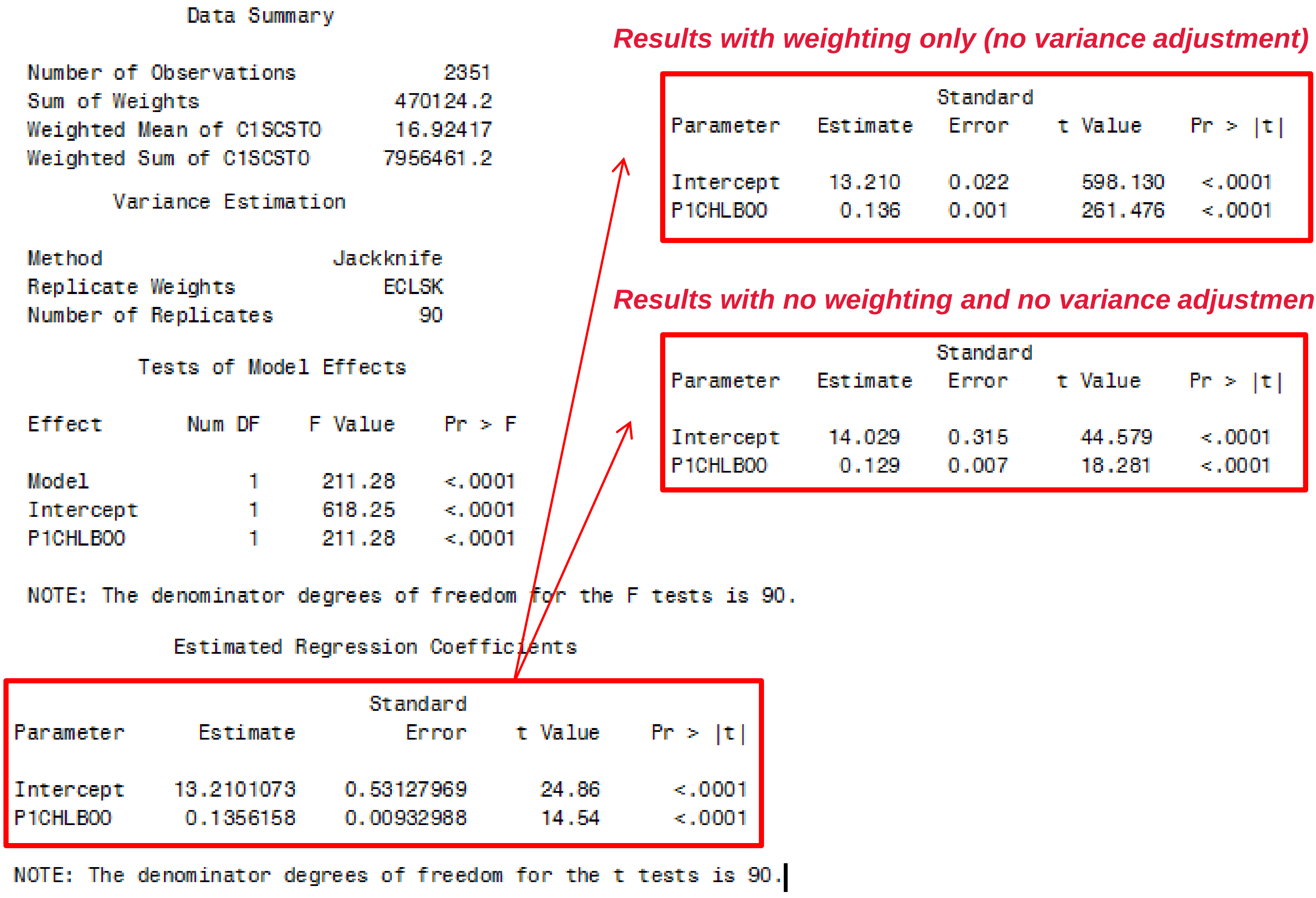
```
LIBNAME spsslib SPSS "C:\Users\nkoziol\Documents\CYFS\ECLSK_SAS_File.por";
DATA work.eclsk;
      SET spsslib.spssfile;
RUN;

PROC SURVEYREG data=eclsk varmethod=jk ;
      model c1scsto=p1chlboo;
      repweights c1pw1-c1pw90 / jkcoefs = 0.999999999;
      weight c1pw0;
run;
```

Necessary for JK2 method

Child 'Tell Stories' variable

Number of children's books

Illustration: Research Question 1



```
                    Data Summary

Number of Observations               2351
Sum of Weights                   470124.2
Weighted Mean of C1SCSTO         16.92417
Weighted Sum of C1SCSTO         7956461.2

             Variance Estimation

Method                         Jackknife
Replicate Weights                  ECLSK
Number of Replicates                  90

               Tests of Model Effects

Effect          Num DF     F Value     Pr > F

Model                1      211.28     <.0001
Intercept            1      618.25     <.0001
P1CHLBOO             1      211.28     <.0001

NOTE: The denominator degrees of freedom for the F tests is 90.

                  Estimated Regression Coefficients

                                   Standard
Parameter       Estimate              Error     t Value     Pr > |t|

Intercept     13.2101073         0.53127969       24.86       <.0001
P1CHLBOO       0.1356158         0.00932988       14.54       <.0001

NOTE: The denominator degrees of freedom for the t tests is 90.
```

Results with weighting only (no variance adjustment)

```
                           Standard
Parameter     Estimate       Error      t Value       Pr > |t|

Intercept      13.210       0.022       598.130       <.0001
P1CHLBOO        0.136       0.001       261.476       <.0001
```

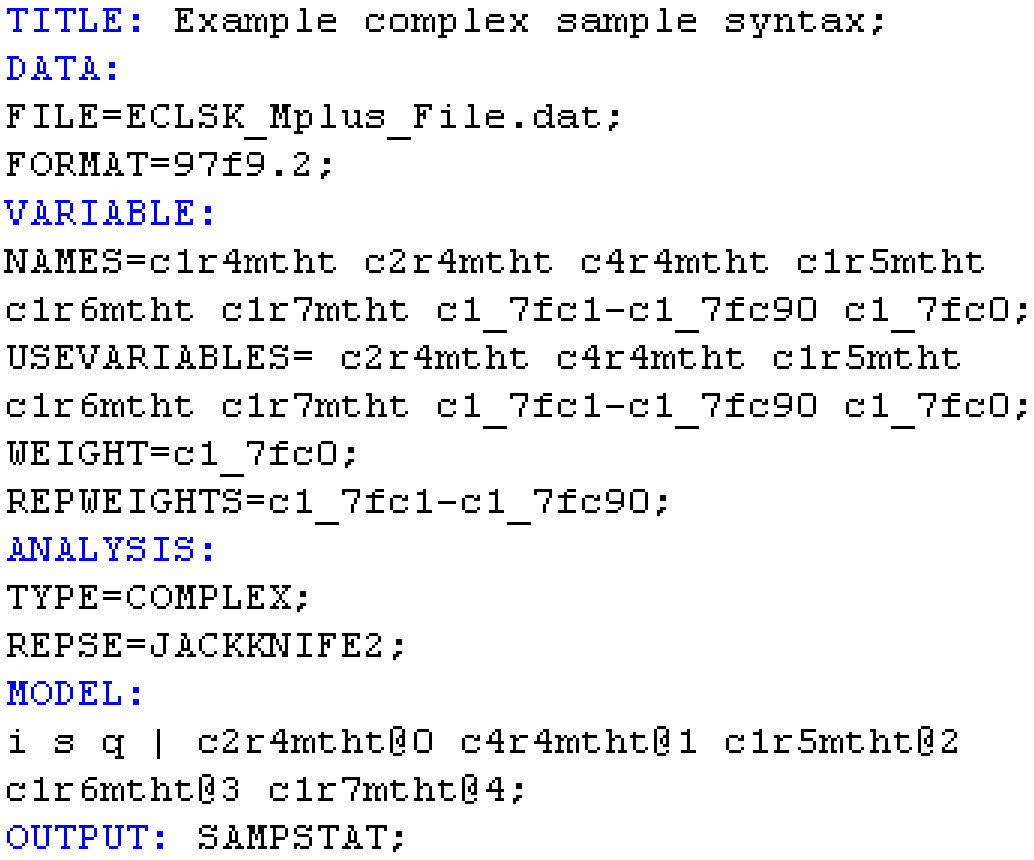
Results with no weighting and no variance adjustment

```
                           Standard
Parameter     Estimate       Error      t Value       Pr > |t|

Intercept      14.029       0.315        44.579       <.0001
P1CHLBOO        0.129       0.007        18.281       <.0001
```

*Results are for illustration purposes only; please do not cite or distribute.

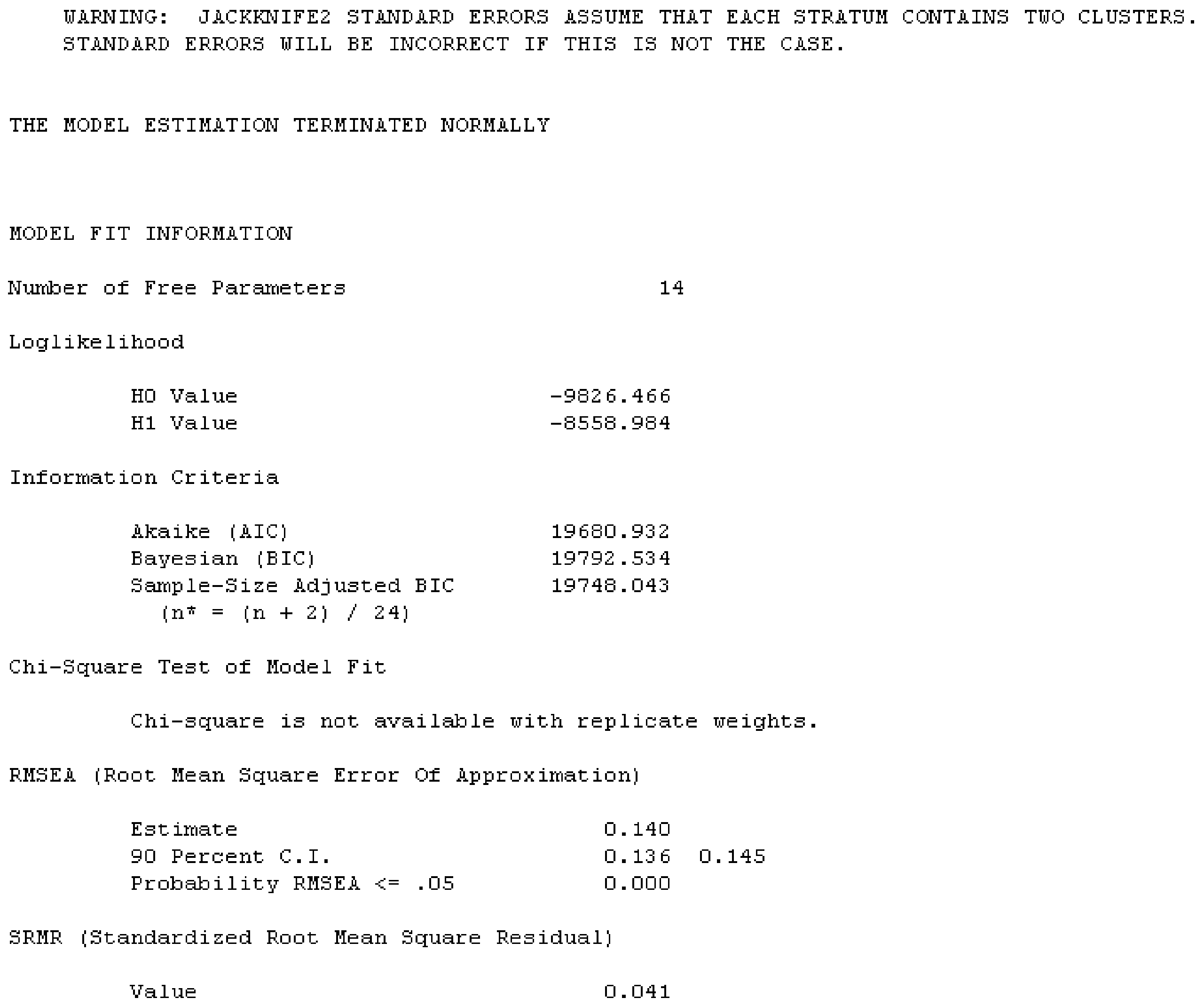
Illustration: Research Question 2

- Conduct 2nd order (quadratic) latent growth model in M*plus*



```
TITLE: Example complex sample syntax;
DATA:
FILE=ECLSK_Mplus_File.dat;
FORMAT=97f9.2;
VARIABLE:
NAMES=c1r4mtht c2r4mtht c4r4mtht c1r5mtht
c1r6mtht c1r7mtht c1_7fc1-c1_7fc90 c1_7fc0;
USEVARIABLES= c2r4mtht c4r4mtht c1r5mtht
c1r6mtht c1r7mtht c1_7fc1-c1_7fc90 c1_7fc0;
WEIGHT=c1_7fc0;
REPWEIGHTS=c1_7fc1-c1_7fc90;
ANALYSIS:
TYPE=COMPLEX;
REPSE=JACKKNIFE2;
MODEL:
i s q | c2r4mtht@0 c4r4mtht@1 c1r5mtht@2
c1r6mtht@3 c1r7mtht@4;
OUTPUT: SAMPSTAT;
```

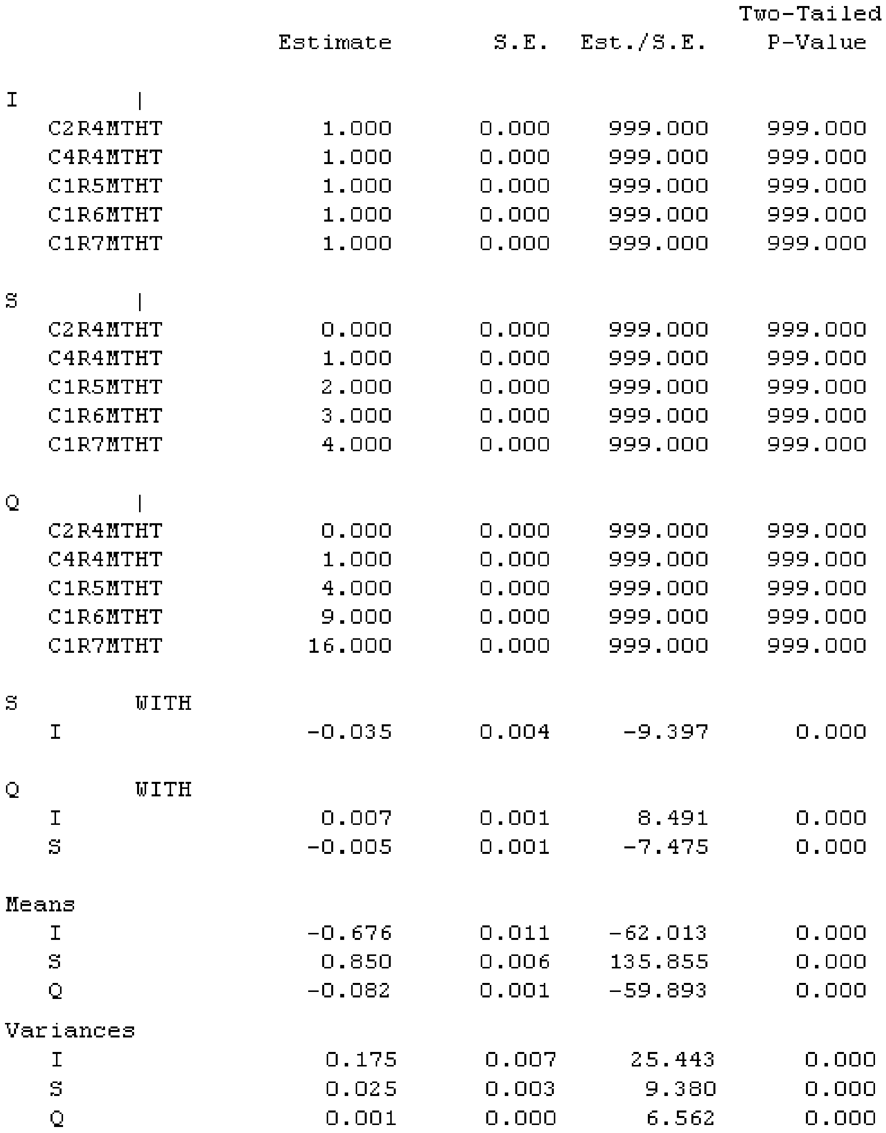
Illustration: Research Question 2



```
     WARNING:  JACKKNIFE2 STANDARD ERRORS ASSUME THAT EACH STRATUM CONTAINS TWO CLUSTERS.
     STANDARD ERRORS WILL BE INCORRECT IF THIS IS NOT THE CASE.


THE MODEL ESTIMATION TERMINATED NORMALLY



MODEL FIT INFORMATION

Number of Free Parameters                         14

Loglikelihood

         H0 Value                        -9826.466
         H1 Value                        -8558.984

Information Criteria

         Akaike (AIC)                    19680.932
         Bayesian (BIC)                  19792.534
         Sample-Size Adjusted BIC        19748.043
           (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

         Chi-square is not available with replicate weights.

RMSEA (Root Mean Square Error Of Approximation)

         Estimate                            0.140
         90 Percent C.I.                     0.136  0.145
         Probability RMSEA <= .05            0.000

SRMR (Standardized Root Mean Square Residual)

         Value                               0.041
```

*Results are for illustration purposes only; please do not cite or distribute.

Illustration: Research Question 2



```
                                                      Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 I        |
    C2R4MTHT           1.000      0.000    999.000    999.000
    C4R4MTHT           1.000      0.000    999.000    999.000
    C1R5MTHT           1.000      0.000    999.000    999.000
    C1R6MTHT           1.000      0.000    999.000    999.000
    C1R7MTHT           1.000      0.000    999.000    999.000

 S        |
    C2R4MTHT           0.000      0.000    999.000    999.000
    C4R4MTHT           1.000      0.000    999.000    999.000
    C1R5MTHT           2.000      0.000    999.000    999.000
    C1R6MTHT           3.000      0.000    999.000    999.000
    C1R7MTHT           4.000      0.000    999.000    999.000

 Q        |
    C2R4MTHT           0.000      0.000    999.000    999.000
    C4R4MTHT           1.000      0.000    999.000    999.000
    C1R5MTHT           4.000      0.000    999.000    999.000
    C1R6MTHT           9.000      0.000    999.000    999.000
    C1R7MTHT          16.000      0.000    999.000    999.000

 S        WITH
    I                 -0.035      0.004     -9.397      0.000

 Q        WITH
    I                  0.007      0.001      8.491      0.000
    S                 -0.005      0.001     -7.475      0.000

 Means
    I                 -0.676      0.011    -62.013      0.000
    S                  0.850      0.006    135.855      0.000
    Q                 -0.082      0.001    -59.893      0.000

 Variances
    I                  0.175      0.007     25.443      0.000
    S                  0.025      0.003      9.380      0.000
    Q                  0.001      0.000      6.562      0.000
```

Other Considerations

Training Opportunities

- 2-5 day government- or other institution-sponsored workshops
  - AERA Institute on Statistical Analysis & AERA Faculty Institute
    - http://www.aera.net/grantsprogram/res_training/stat_institute/SIFly.html
    - http://www.aera.net/grantsprogram/res_training/stat_institute/SIFacFly.html
  - IES sponsored workshops
    - http://ies.ed.gov/whatsnew/conferences/?cid=2
  - AIR
    - http://airweb.org/?page=1803
- ICPSR 1 week offerings
  - http://www.icpsr.umich.edu/icpsrweb/sumprog/
- 1 day pre-annual meeting workshops
  - E.g., “Early Childhood Surveys at NCES: The ECLS and NHES Data Users Workshop” (2011 SRCD biennial meeting)

Funding Opportunities

- AERA Dissertation and Research Grants
  - http://www.aera.net/grantsprogram/
- NIH Grants
  - http://grants.nih.gov/grants/guide/
  - E.g., “R40 Maternal & Child Health Research Secondary Data Analysis Studies Grants”
  - R21 grant mechanism (exploration)
- IES Grants (exploration; Goal 1)
  - http://ies.ed.gov/funding/
  - NAEP Secondary Analysis Grants
  - RFP’s that encourage use of secondary data, for example, the “Social and Behavioral Context for Academic Learning” RFP
- AIR Grants
  - http://www.airweb.org/?page=1622

References

AIR, NCES, & NSF (2011). National Summer Data Policy Institute training materials. Washington, D.C.

Asparouhov, T., & Muthén, B. (2010). Plausible values for latent variables using M*plus*. Technical Report. www.statmodel.com

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide*. New York, NY: Cambridge.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd Ed.). Boston, MA: Brooks/Cole.

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data*. Newbury Park, CA: SAGE.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

McCall, R. B., & Appelbaum, M. I. (1991). Some issues of conducting secondary analyses. *Developmental Psychology*, *27*, 911-917.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

NCES (2011). ECLS-B database training seminar materials. Washington, D.C.

References

Spagat, M., & Dougherty, J. (2010). Conflict deaths in Iraq: A methodological critique of the ORB survey estimate. *Survey Research Methods, 4*, 3-15.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, *42*, 517-540.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks* (NCES 2009-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Trzesniewski, K. H., Donnellan, M. B., & Lucas, R. E. (Eds) (2011). *Secondary data analysis: An introduction for psychologists*. Washington, D.C.: APA.

Vartanian, T. P. (2011). *Secondary data analysis*. New York, NY: Oxford.

Thanks!

For more information, please contact:
Natalie Koziol, nak371@neb.rr.com