# Unit 11: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):
***READING***, a continuous variable, test score, mean = 47 and standard deviation = 9
Predictor Variables (aka Independent Variables):
Question Predictor-
***RACE***, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White
Control Predictors-
***HOMEWORK***, hours per week, a continuous variable, mean = 6.0 and standard deviation = 4.7
***FREELUNCH***, a proxy for SES, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
***ESL***, English as a second language, a dichotomous variable, 1 = ESL, 0 = native speaker of English

- Unit 11: What is measurement error, and how does it affect our analyses?
- Unit 12: What tools can we use to detect assumption violations (e.g., ***outliers***)?
- Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?
- Unit 14: How do we deal with violations of the ***homoskedasticity*** assumption?
- Unit 15: What are the correlations among reading, race, ESL, and homework, controlling for SES?
- Unit 16: Is there a relationship between reading and race, controlling for SES, ESL and homework?
- Unit 17: Does the relationship between reading and race vary by levels of SES, ESL or homework?
- Unit 18: What are sensible strategies for building complex statistical models from scratch?
- Unit 19: How do we deal with violations of the ***independence*** assumption (using ANOVA)?

# Introductory Data Analysis: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):

*READING*, a continuous variable, test score, mean = 47 and standard deviation = 9

Predictor Variables (aka Independent Variables):

*FREELUNCH*, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
*RACE*, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White

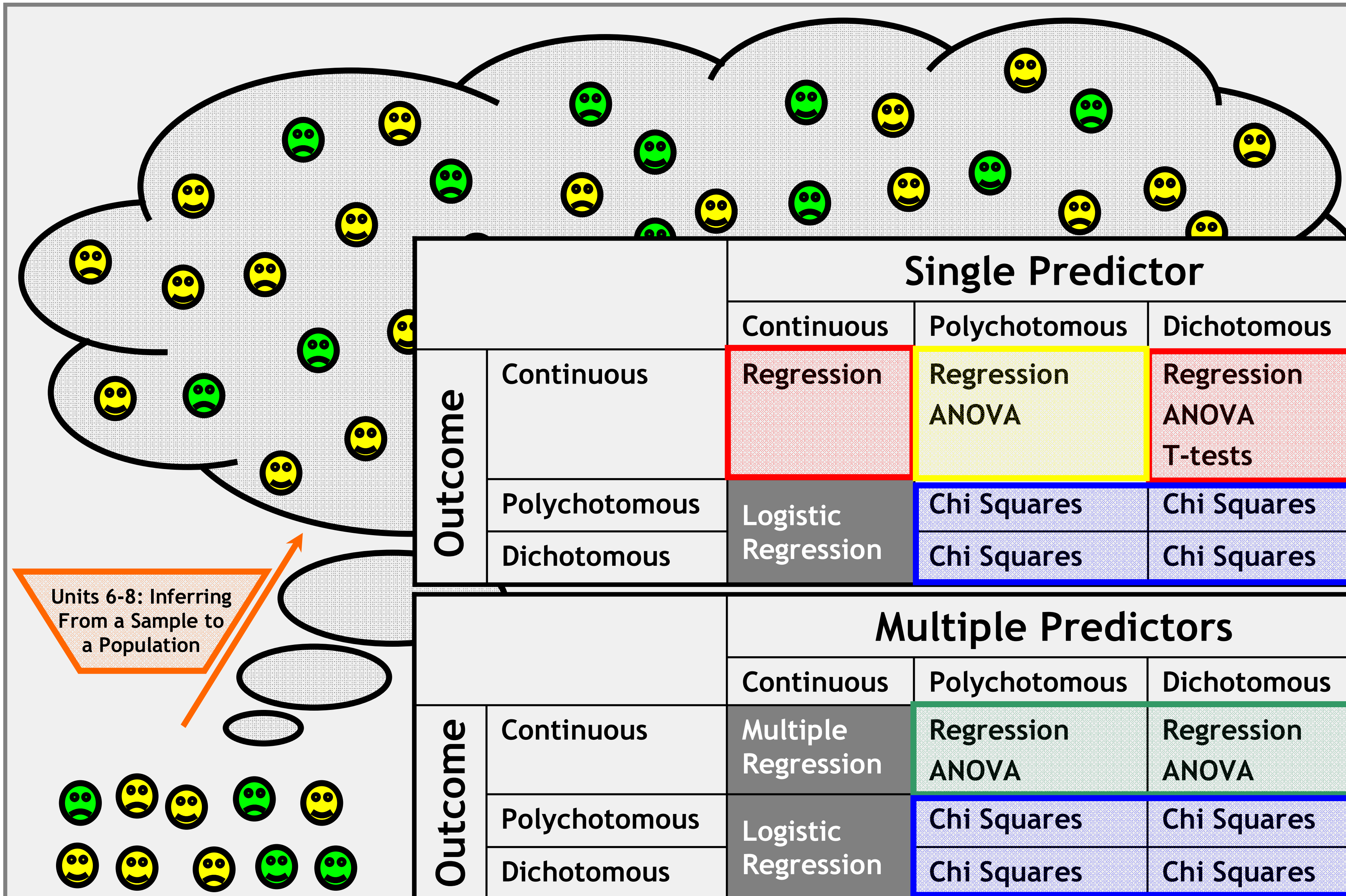
- Unit 1: In our sample, is there a relationship between reading achievement and free lunch?
- Unit 2: In our sample, what does reading achievement look like (from an outlier resistant perspective)?
- Unit 3: In our sample, what does reading achievement look like (from an outlier sensitive perspective)?
- Unit 4: In our sample, how strong is the relationship between reading achievement and free lunch?
- Unit 5: In our sample, free lunch predicts what proportion of variation in reading achievement?
- Unit 6: In the population, is there a relationship between reading achievement and free lunch?
- Unit 7: In the population, what is the magnitude of the relationship between reading and free lunch?
- Unit 8: What assumptions underlie our inference from the sample to the population?
- Unit 9: In the population, is there a relationship between reading and race?
- Unit 10: In the population, is there a relationship between reading and race controlling for free lunch?
- Appendix A: In the population, is there a relationship between race and free lunch?

# Intermediate Data Analysis: Road Map (VERBAL)

Nationally Representative Sample of 7,800 8th Graders Surveyed in 1988 (NELS 88).
Outcome Variable (aka Dependent Variable):
***READING***, a continuous variable, test score, mean = 47 and standard deviation = 9
Predictor Variables (aka Independent Variables):
Question Predictor-
***RACE***, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White
Control Predictors-
***HOMEWORK***, hours per week, a continuous variable, mean = 6.0 and standard deviation = 4.7
***FREELUNCH***, a proxy for SES, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
***ESL***, English as a second language, a dichotomous variable, 1 = ESL, 0 = native speaker of English

- Unit 11: What is measurement error, and how does it affect our analyses?
- Unit 12: What tools can we use to detect assumption violations (e.g., ***outliers***)?
- Unit 13: How do we deal with violations of the ***linearity*** and ***normality*** assumptions?
- Unit 14: How do we deal with violations of the ***homoskedasticity*** assumption?
- Unit 15: What are the correlations among reading, race, ESL, and homework, controlling for SES?
- Unit 16: Is there a relationship between reading and race, controlling for SES, ESL and homework?
- Unit 17: Does the relationship between reading and race vary by levels of SES, ESL or homework?
- Unit 18: What are sensible strategies for building complex statistical models from scratch?
- Unit 19: How do we deal with violations of the ***independence*** assumption (using ANOVA)?

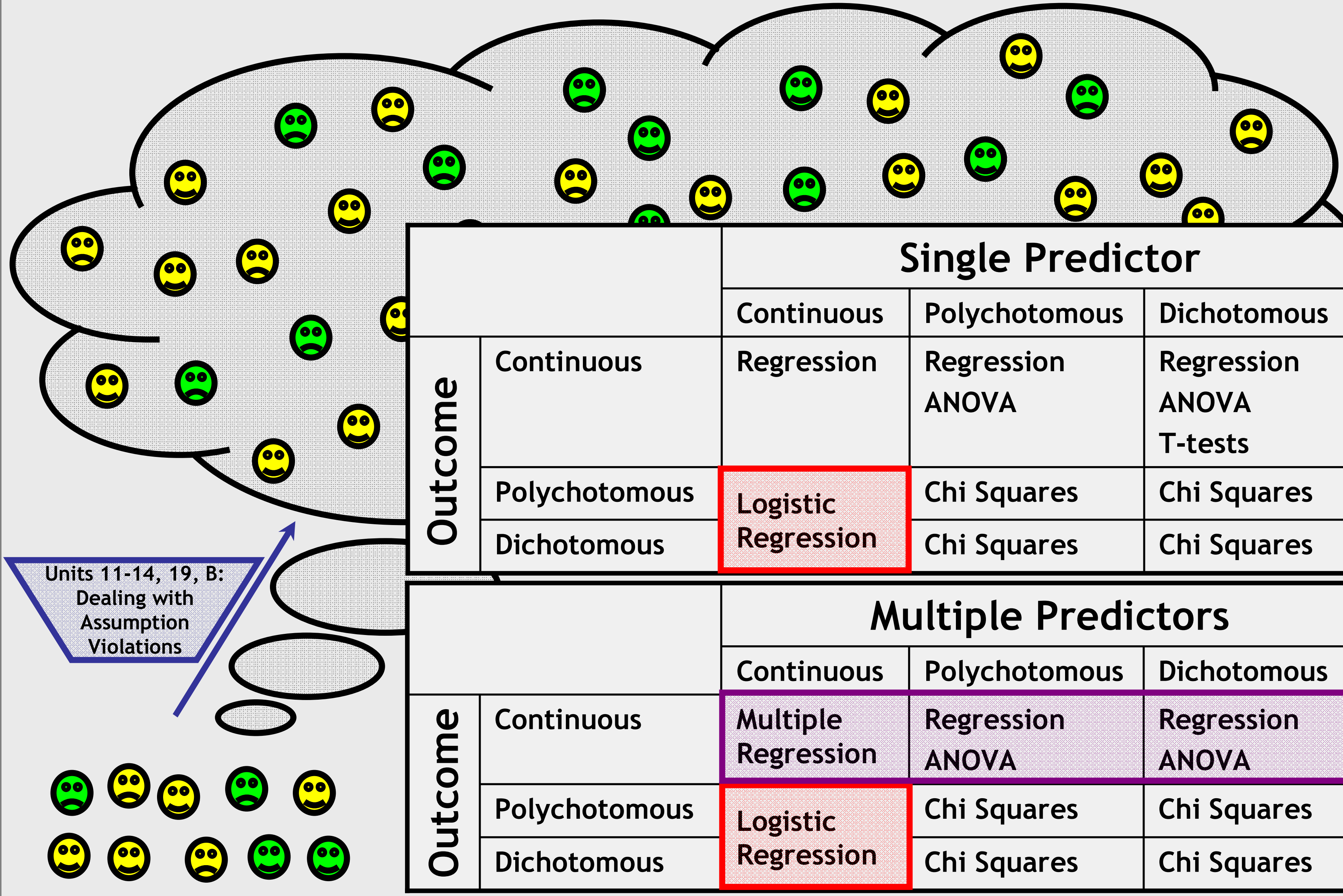
# Introductory Data Analysis: Road Map (Schematic)




| | | Single Predictor | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Regression | Regression<br>ANOVA | Regression<br>ANOVA<br>T-tests |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

| | | Multiple Predictors | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Multiple Regression | Regression<br>ANOVA | Regression<br>ANOVA |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

# Intermediate Data Analysis: Road Map (Schematic)




| | | Single Predictor | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Regression | Regression<br>ANOVA | Regression<br>ANOVA<br>T-tests |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

| | | Multiple Predictors | | |
|---|---|---|---|---|
| | | Continuous | Polychotomous | Dichotomous |
| Outcome | Continuous | Multiple Regression | Regression<br>ANOVA | Regression<br>ANOVA |
| | Polychotomous | Logistic Regression | Chi Squares | Chi Squares |
| | Dichotomous | | Chi Squares | Chi Squares |

# Introductory Data Analysis: Roadmap (SPSS Output)

**Statistics**

| | | READING | FREELUNCH |
|---|---|---|---|
| N | Valid | 7800 | 7800 |
| | Missing | 0 | 0 |
| Mean | | 47.4940 | .3354 |
| Std. Deviation | | 8.56944 | .47216 |
| Minimum | | 23.96 | .00 |
| Maximum | | 63.49 | 1.00 |
| Percentiles | 25 | 41.2400 | .0000 |
| | 50 | 47.4300 | .0000 |
| | 75 | 53.9300 | 1.0000 |

Unit 3

Unit 2

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .267[a] | .071 | .071 | 8.25952 |

a. Predictors: (Constant), FREELUNCH

Unit 5

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 40744.322 | 1 | 40744.322 | 597.251 | .000[a] |
| | Residual | 531977.541 | 7798 | 68.220 | | |
| | Total | 572721.864 | 7799 | | | |

a. Predictors: (Constant), FREELUNCH

b. Dependent Variable: READING

Unit 9

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | 49.118 | .115 | | 428.169 | .000 | 48.893 | 49.342 |
| | FREELUNCH | -4.841 | .198 | -.267 | -24.439 | .000 | -5.229 | -4.453 |

a. Dependent Variable: READING

Unit 1 Unit 8 Unit 4 Unit 6 Unit 7

# Introductory Data Analysis: Roadmap (SPSS Output)

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .221[a] | .049 | .049 | 8.35882 |

a. Predictors: (Constant), BLACK, ASIAN, LATINO

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 28016.721 | 3 | 9338.907 | 133.662 | .000[a] |
| | Residual | 544705.143 | 7796 | 69.870 | | |
| | Total | 572721.864 | 7799 | | | |

a. Predictors: (Constant), BLACK, ASIAN, LATINO

b. Dependent Variable: READING

**Unit 9**

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 48.338 | .110 | | 438.242 | .000 | 48.122 | 48.554 |
| | ASIAN | 1.034 | .383 | .030 | 2.697 | .007 | .283 | 1.786 |
| | LATINO | -4.418 | .306 | -.161 | -14.447 | .000 | -5.017 | -3.818 |
| | BLACK | -4.889 | .339 | -.161 | -14.423 | .000 | -5.554 | -4.225 |

a. Dependent Variable: READING

# Introductory Data Analysis: Roadmap (SPSS Output)

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .314[a] | .099 | .098 | 8.13954 |

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 56485.879 | 7 | 8069.411 | 121.799 | .000[a] |
| | Residual | 516235.985 | 7792 | 66.252 | | |
| | Total | 572721.864 | 7799 | | | |

a. Predictors: (Constant), FREELUNCHxBLACK, FREELUNCHxASIAN, FREELUNCHxLATINO, ASIAN, FREELUNCH, LATINO, BLACK
b. Dependent Variable: READING

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound |
| 1 | (Constant) | 49.439 | .127 | | 389.587 | .000 | 49.191 | 49.688 |
| | FREELUNCH | -3.882 | .238 | -.214 | -16.293 | .000 | -4.349 | -3.415 |
| | ASIAN | 1.491 | .431 | .043 | 3.461 | .001 | .646 | 2.335 |
| | LATINO | -3.250 | .426 | -.119 | -7.624 | .000 | -4.086 | -2.415 |
| | BLACK | -3.406 | .504 | -.112 | -6.764 | .000 | -4.393 | -2.419 |
| | FREELUNCHxASIAN | -2.472 | .865 | -.037 | -2.858 | .004 | -4.167 | -.776 |
| | FREELUNCHxLATINO | -.363 | .606 | -.010 | -.600 | .549 | -1.550 | .824 |
| | FREELUNCHxBLACK | -.501 | .678 | -.013 | -.739 | .460 | -1.829 | .828 |

a. Dependent Variable: READING

**Unit 10**
**Multiple Regression**

# Introductory Data Analysis: Roadmap (SPSS Output)

**Tests of Between-Subjects Effects**

Dependent Variable:READING

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 56485.879[a] | 7 | 8069.411 | 121.799 | .000 |
| Intercept | 6087049.770 | 1 | 6087049.770 | 91877.151 | .000 |
| RACE | 14705.701 | 3 | 4901.900 | 73.989 | .000 |
| FREELUNCH | 16141.887 | 1 | 16141.887 | 243.644 | .000 |
| RACE * FREELUNCH | 559.039 | 3 | 186.346 | 2.813 | .038 |
| Error | 516235.985 | 7792 | 66.252 | | |
| Total | 1.817E7 | 7800 | | | |
| Corrected Total | 572721.864 | 7799 | | | |

a. R Squared = .099 (Adjusted R Squared = .098)

**Unit 10**
**ANOVA**

**FREELUNCH * R^S RACE/ETHNIC BACKGROUND Crosstabulation**

**Appendix A**

| | | | R^S RACE/ETHNIC BACKGROUND | | | | |
|---|---|---|---|---|---|---|---|
| | | | Asian | Latino | Black | White | Total |
| FREELUNCH | 0 | Count | 391 | 400 | 279 | 4114 | 5184 |
| | | Expected Count | 344.3 | 570.9 | 451.9 | 3816.9 | 5184.0 |
| | | Std. Residual | 2.5 | -7.2 | -8.1 | 4.8 | |
| | 1 | Count | 127 | 459 | 401 | 1629 | 2616 |
| | | Expected Count | 173.7 | 288.1 | 228.1 | 1926.1 | 2616.0 |
| | | Std. Residual | -3.5 | 10.1 | 11.5 | -6.8 | |
| | Total | Count | 518 | 859 | 680 | 5743 | 7800 |
| | | Expected Count | 518.0 | 859.0 | 680.0 | 5743.0 | 7800.0 |

# Unit 11: Roadmap (SPSS Output)

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. | 95% Confidence Interval for B Lower Bound | 95% Confidence Interval for B Upper Bound |
|---|---|---|---|---|---|---|---|---|
| 1 | (Constant) | 48.338 | .110 | | 438.242 | .000 | 48.122 | 48.554 |
| | ASIAN | 1.034 | .383 | .030 | 2.697 | .007 | .283 | 1.786 |
| | BLACK | -4.889 | .339 | -.161 | -14.423 | .000 | -5.554 | -4.225 |
| | LATINO | -4.418 | .306 | -.161 | -14.447 | .000 | -5.017 | -3.818 |
| 2 | (Constant) | 43.878 | .280 | | 156.558 | .000 | 43.328 | 44.427 |
| | ASIAN | .727 | .377 | .021 | 1.929 | .054 | -.012 | 1.465 |
| | BLACK | -4.796 | .333 | -.158 | -14.412 | .000 | -5.448 | -4.144 |
| | LATINO | -4.123 | .301 | -.151 | -13.715 | .000 | -4.712 | -3.534 |
| | L2HOMEWORKP1 | 1.766 | .102 | .188 | 17.254 | .000 | 1.565 | 1.967 |
| 3 | (Constant) | 45.381 | .284 | | 159.528 | .000 | 44.823 | 45.938 |
| | ASIAN | .461 | .441 | .013 | 1.045 | .296 | -.404 | 1.325 |
| | BLACK | -3.622 | .331 | -.119 | -10.956 | .000 | -4.270 | -2.974 |
| | LATINO | -3.311 | .366 | -.121 | -9.035 | .000 | -4.029 | -2.592 |
| | L2HOMEWORKP1 | 1.603 | .100 | .170 | 15.974 | .000 | 1.406 | 1.799 |
| | ESL | .218 | .363 | .009 | .600 | .548 | -.494 | .930 |
| | FREELUNCH | -3.867 | .199 | -.213 | -19.452 | .000 | -4.256 | -3.477 |
| 4 | (Constant) | 45.358 | .288 | | 157.560 | .000 | 44.794 | 45.923 |
| | ASIAN | -.377 | .668 | -.011 | -.564 | .573 | -1.687 | .933 |
| | BLACK | -3.447 | .498 | -.113 | -6.922 | .000 | -4.423 | -2.471 |
| | LATINO | -2.779 | .517 | -.102 | -5.371 | .000 | -3.793 | -1.765 |
| | L2HOMEWORKP1 | 1.591 | .100 | .169 | 15.866 | .000 | 1.394 | 1.788 |
| | ESL | -.876 | .638 | -.035 | -1.373 | .170 | -2.126 | .374 |
| | FREELUNCH | -3.574 | .235 | -.197 | -15.208 | .000 | -4.035 | -3.113 |
| | ESLxASIAN | 3.245 | .999 | .080 | 3.249 | .001 | 1.287 | 5.202 |
| | ESLxBLACK | 5.872 | 1.885 | .036 | 3.115 | .002 | 2.177 | 9.568 |
| | ESLxLATINO | .446 | .858 | .013 | .520 | .603 | -1.235 | 2.127 |
| | FREELUNCHxASIAN | -2.769 | .853 | -.041 | -3.245 | .001 | -4.442 | -1.096 |
| | FREELUNCHxBLACK | -.751 | .666 | -.019 | -1.127 | .260 | -2.058 | .555 |
| | FREELUNCHxLATINO | -.437 | .604 | -.012 | -.724 | .469 | -1.622 | .747 |

a. Dependent Variable: READING

Unit 18

Unit 9

Unit 16

Unit 17

Unit 8

Unit 11

Unit 12

Unit 13

Unit 14

Unit 19

# Unit 11: Roadmap (SPSS Output)

**Correlations**

| | | READING | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | ESL | FREELUNCH |
|---|---|---|---|---|---|
| READING | Pearson Correlation | 1.000 | .183** | -.053** | -.267** |
| | Sig. (2-tailed) | | .000 | .000 | .000 |
| | N | 7800.000 | 7800 | 7800 | 7800 |
| NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | Pearson Correlation | .183** | 1.000 | .005 | -.092** |
| | Sig. (2-tailed) | .000 | | .648 | .000 |
| | N | 7800 | 7800.000 | 7800 | 7800 |
| ESL | Pearson Correlation | -.053** | .005 | 1.000 | .093** |
| | Sig. (2-tailed) | .000 | .648 | | .000 |
| | N | 7800 | 7800 | 7800.000 | 7800 |
| FREELUNCH | Pearson Correlation | -.267** | -.092** | .093** | 1.000 |
| | Sig. (2-tailed) | .000 | .000 | .000 | |
| | N | 7800 | 7800 | 7800 | 7800.000 |

**. Correlation is significant at the 0.01 level (2-tailed).

Unit 4

**Correlations**

| Control Variables | | | READING | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | ESL |
|---|---|---|---|---|---|
| FREELUNCH | READING | Correlation | 1.000 | .165 | -.029 |
| | | Significance (2-tailed) | . | .000 | .009 |
| | | df | 0 | 7797 | 7797 |
| | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | Correlation | .165 | 1.000 | .014 |
| | | Significance (2-tailed) | .000 | . | .222 |
| | | df | 7797 | 0 | 7797 |
| | ESL | Correlation | -.029 | .014 | 1.000 |
| | | Significance (2-tailed) | .009 | .222 | . |
| | | df | 7797 | 7797 | 0 |

Unit 15

# Unit 11: GLM and Assumptions About Measurement Error

**Unit 11 Post Hole:**

Note the threats to validity posed by measurement error in the outcome and predictor(s) of a model.

**Unit 11 Technical Memo and School Board Memo:**

Fit a simple linear regression model (with a continuous outcome and a predictor of your choice), interpret your results, and discuss the threats to validity posed by measurement error in the outcome and predictor.

**Unit 11 Review:**

Review Units 1 and 2.

**Unit 11 Reading:**

Meyers et al., Chapter 2.

# Unit 11: Technical Memo and School Board Memo

Work Products (Part I of I):

I. Technical Memo

A. Introduction

i. State a theory (or perhaps hunch) for the relationship—think causally,  be creative. (1 Sentence)

ii. State a research question for the theory (or hunch)—think correlationally, be formal. Now that you know the statistical machinery that justifies an inference from a sample to a population, begin each research question, “In the population,…” (1 Sentence)

iii. List the two variables, and label them “outcome” and “predictor,” respectively.

iv. Describe your outcome, noting validity threats due to measurement error. (1 Paragraph)

v. Describe your predictor, noting validity threats due to measurement error. (1 Paragraph)

vi. Include your theoretical model.

X. Exploratory Data Analysis. Explore your data using outlier resistant statistics.

i. For  each variable, use a coherent narrative to  convey the results of your exploratory univariate analysis of the data. Don’t lose sight of the substantive meaning of the numbers. (1 Paragraph Each As Per Unit 1)

ii. For each relationship between your outcome and predictor, use a coherent narrative to convey the results of your exploratory bivariate analysis of the data. (1 Paragraph Each As Per Unit 2)

II. School Board Memo: Concisely, precisely and plainly convey your key findings to a lay audience. Note that, whereas you are building on the technical memo for most of the semester, your school board memo is fresh each week. (Max 200 Words)

III. Memo Metacognitive

# Unit 11: Research Question I

Theory: Reading achievement requires many complex skills acquired over a lengthy time period. Therefore, commitment to school work, particularly homework, is essential to reading development.

Research Question: Are students' reading scores positively correlated with the number of hours per week spent on homework?

Data Set: National Education Longitudinal Study (NELS88.sav)

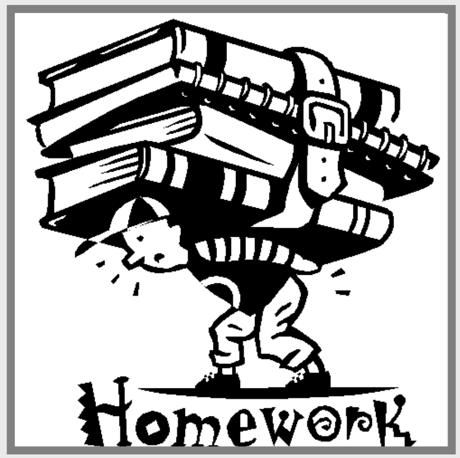
Variables:

Outcome—8th Grade Reading Achievement Score (*READING*)

Predictor—Self-reported Hours Per Week Spent on HW (*HOMEWORK*)

Model:

$$READING = \beta_0 + \beta_1 HOMEWORK + \varepsilon$$



# Unit 11: Research Question II

Theory: Reading achievement requires many complex skills acquired over a lengthy time period. Therefore, persistent educational risk factors, such as low socioeconomic status (SES), will tend to signal low reading achievement.

Research Question: Are students' reading scores negatively correlated with eligibility for free/reduced lunch (a proxy for SES)?

Data Set: National Education Longitudinal Study (NELS88.sav)

Variables:
- Outcome—8$^{th}$ Grade Reading Achievement Score (*READING*)
- Predictor—Eligibility for Free/Reduced Lunch as Determined by an Annual Family Income Cut-Off of $25,000 (*FREELUNCH*)

Model:

$$READING = \beta_0 + \beta_1 FREELUNCH + \varepsilon$$

# NELS88.sav Codebook

**National Education Longitudinal Study**

**Source: U.S. Department of Education**

**Summary: Here are select variables from the NELS88 data set.**

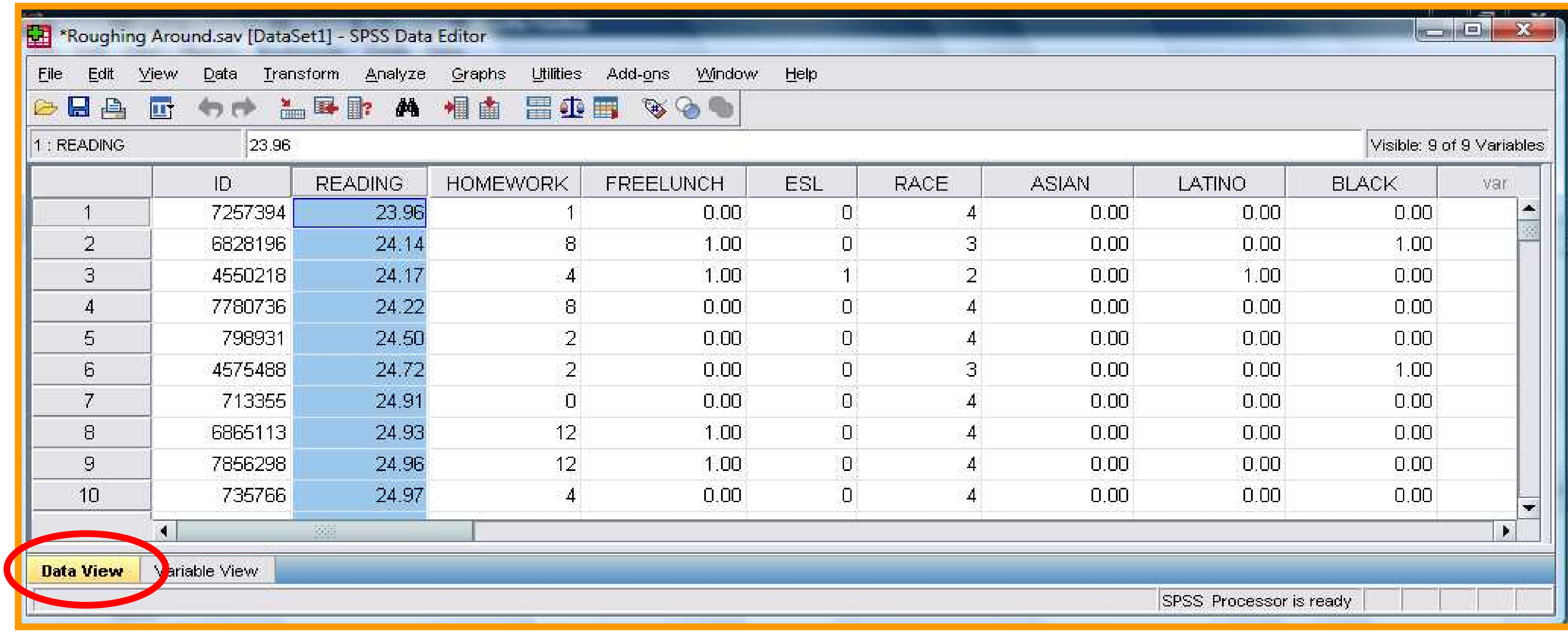
**Notes: I created the FREELUNCH variable based on annual family income of less than $25,000. converted the HOMEWORK variable from an ordinal/categorical variable to a continuous variable, which is why it is so "binny." I removed from the data set students who self-identified as other than Asian, Black, Latino, or White. I then created a set of indicator variables from RACE: ASIAN, BLACK AND LATION with WHITE as an (implicit) reference category.**

**Sample: A nationally representative sample of 7,800 8th graders.**

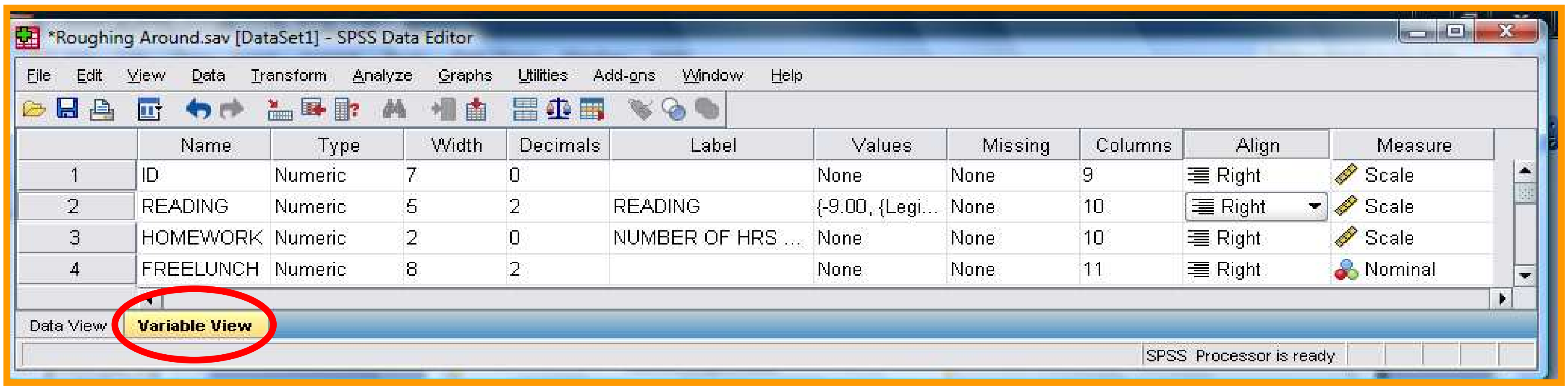
**Variables:**

***READING***, a continuous variable, test score, mean = 47 and standard deviation = 9
***FREELUNCH***, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
***HOMEWORK***, hours per week, a continuous variable, mean = 6.0 and standard deviation = 4.7
***FREELUNCH***, a proxy for SES, a dichotomous variable, 1 = Eligible for Free/Reduced Lunch and 0 = Not
***ESL***, English as a second language, a dichotomous variable, 1 = ESL, 0 = native speaker of English
***RACE***, a polychotomous variable, 1 = Asian, 2 = Latino, 3 = Black and 4 = White
***ASIAN*,** a dichotomous variable, 1 = Self-Identifies as Asian and 0 = Not
***LATINO*,** a dichotomous variable, 1 = Self-Identifies as Latino and 0 = Not
***BLACK*,** a dichotomous variable, 1 = Self-Identifies as Black and 0 = Not

# Select Variables from the NELS Data Set




| | ID | READING | HOMEWORK | FREELUNCH | ESL | RACE | ASIAN | LATINO | BLACK | var |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7257394 | 23.96 | 1 | 0.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 2 | 6828196 | 24.14 | 8 | 1.00 | 0 | 3 | 0.00 | 0.00 | 1.00 | |
| 3 | 4550218 | 24.17 | 4 | 1.00 | 1 | 2 | 0.00 | 1.00 | 0.00 | |
| 4 | 7780736 | 24.22 | 8 | 0.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 5 | 798931 | 24.50 | 2 | 0.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 6 | 4575488 | 24.72 | 2 | 0.00 | 0 | 3 | 0.00 | 0.00 | 1.00 | |
| 7 | 713355 | 24.91 | 0 | 0.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 8 | 6865113 | 24.93 | 12 | 1.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 9 | 7856298 | 24.96 | 12 | 1.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |
| 10 | 735766 | 24.97 | 4 | 0.00 | 0 | 4 | 0.00 | 0.00 | 0.00 | |




| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align | Measure |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Numeric | 7 | 0 | | None | None | 9 | Right | Scale |
| 2 | READING | Numeric | 5 | 2 | READING | {-9.00, {Legi... | None | 10 | Right | Scale |
| 3 | HOMEWORK | Numeric | 2 | 0 | NUMBER OF HRS ... | None | None | 10 | Right | Scale |
| 4 | FREELUNCH | Numeric | 8 | 2 | | None | None | 11 | Right | Nominal |

# Univariate Exploratory Data Analyses

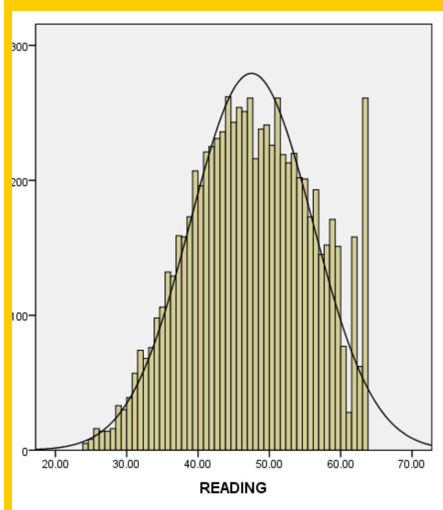
**Statistics**

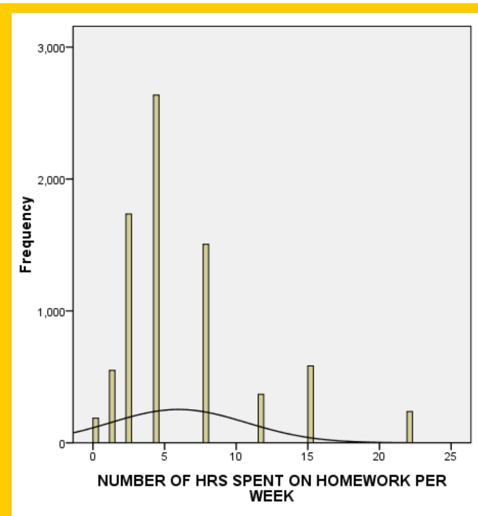
| | | READING | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | FREELUNCH |
|---|---|---|---|---|
| N | Valid | 7800 | 7800 | 7800 |
| | Missing | 0 | 0 | 0 |
| Mean | | 47.4940 | 5.97 | .3354 |
| Std. Deviation | | 8.56944 | 4.730 | .47216 |
| Minimum | | 23.96 | 0 | .00 |
| Maximum | | 63.49 | 22 | 1.00 |
| Percentiles | 25 | 41.2400 | 2.50 | .0000 |
| | 50 | 47.4300 | 4.25 | .0000 |
| | 75 | 53.9300 | 8.00 | 1.0000 |

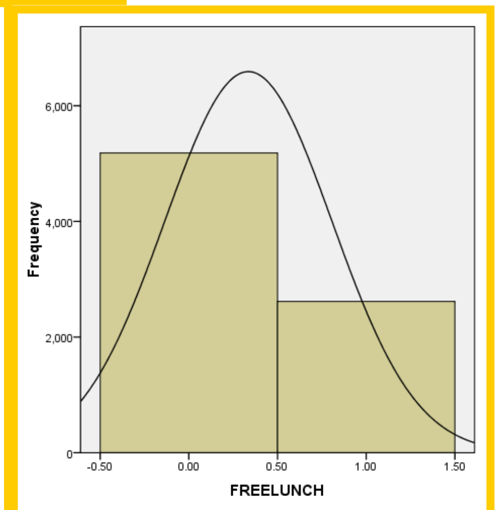
**The distribution of reading scores is approximately normal, except for an obvious ceiling effect at the top score of 63. In our sample, the reading scores range from 24 points to 63 points, thus a range of 39 points, with a midspread of 13 points and a median of 47 points. There does not appear to be any outliers.**

**The distribution of hours spent on homework is positively skewed. We should note that respondents had only eight choices: about 0, 1.25, 2.5, 4.25, 8, 11.75, 15, or 22 hours. The median and mode is 4.25. Half the students report studying between 2.5 and 8 hours per week.**

**In our sample of 7,800 8th graders, 2616 (33.5%) were eligible for free/reduced lunch.**

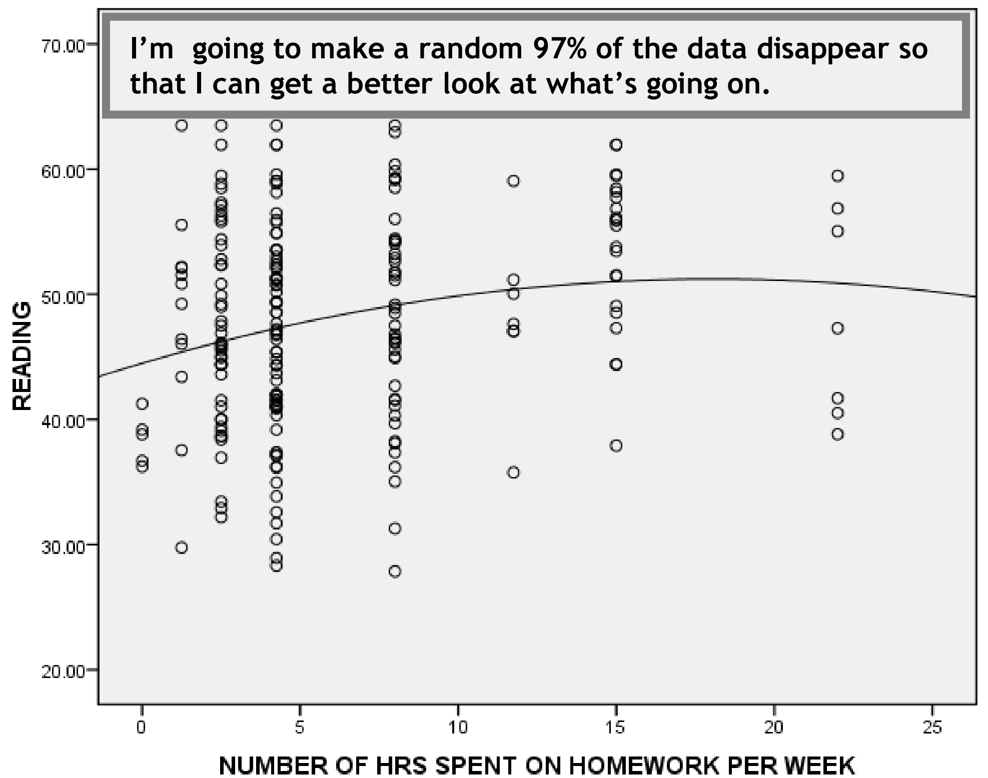



**SPLASH: SPread, Location, SHape**

# Bivariate Exploratory Data Analysis: *READING* vs. *HOMEWORK*




The relationship between reading achievement and homework hours is non-linear such that the relationship is positive from 0 hours to about 12 hours, but beyond 12 hours there appears to be no relationship or perhaps a slight negative relationship. This pattern may signal a law of diminishing returns for homework. The relationship is weak, with no apparent outliers. If we were to fit a linear model, we would conclude that a one-hour difference per week in homework preparation is associated with 0.33 points on the reading achievement test; students who study more tend to score higher.

Recall how we fit a straight line to data using OLS regression. http://www.dynamicgeometry.com/JavaSketchpad/Gallery/Other_Explorations_and_Amusements/Least_Squares.html

**Coefficients**[a]

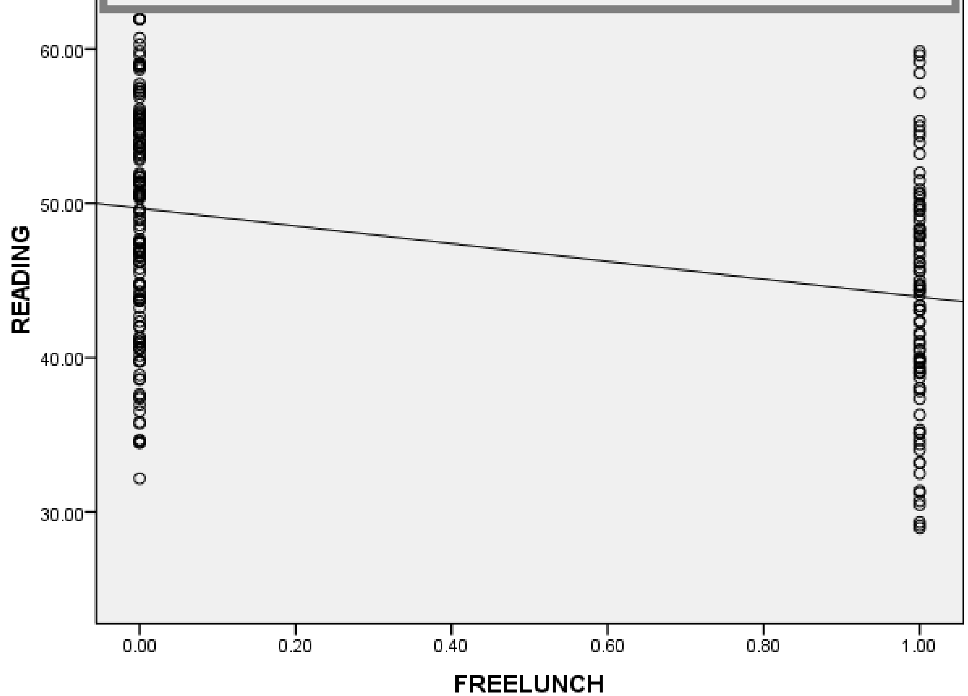
| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 45.514 | .154 | | 296.359 | .000 |
| | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | .332 | .020 | .183 | 16.451 | .000 |

a. Dependent Variable: READING

**DOLMAS:**
Direction,
Outliers,
Linearity,
Magnitude
And
Strength.

# Bivariate Exploratory Data Analysis: *READING* vs. *FREELUNCH*

Again, we reduce the data by 97% for EDA purposes.




The relationship between reading achievement and eligibility for free/reduced lunch is negative in our sample. On average, students who are eligible for free/reduced lunch score 5 points less than their ineligible counterparts. The average eligible student attains a reading score of 44, and the average eligible student attains a reading score of 49. However, due to broad variation among individuals, the relationship is weak, so these group averages should not be the basis for judgments about individuals. There are no apparent outliers.

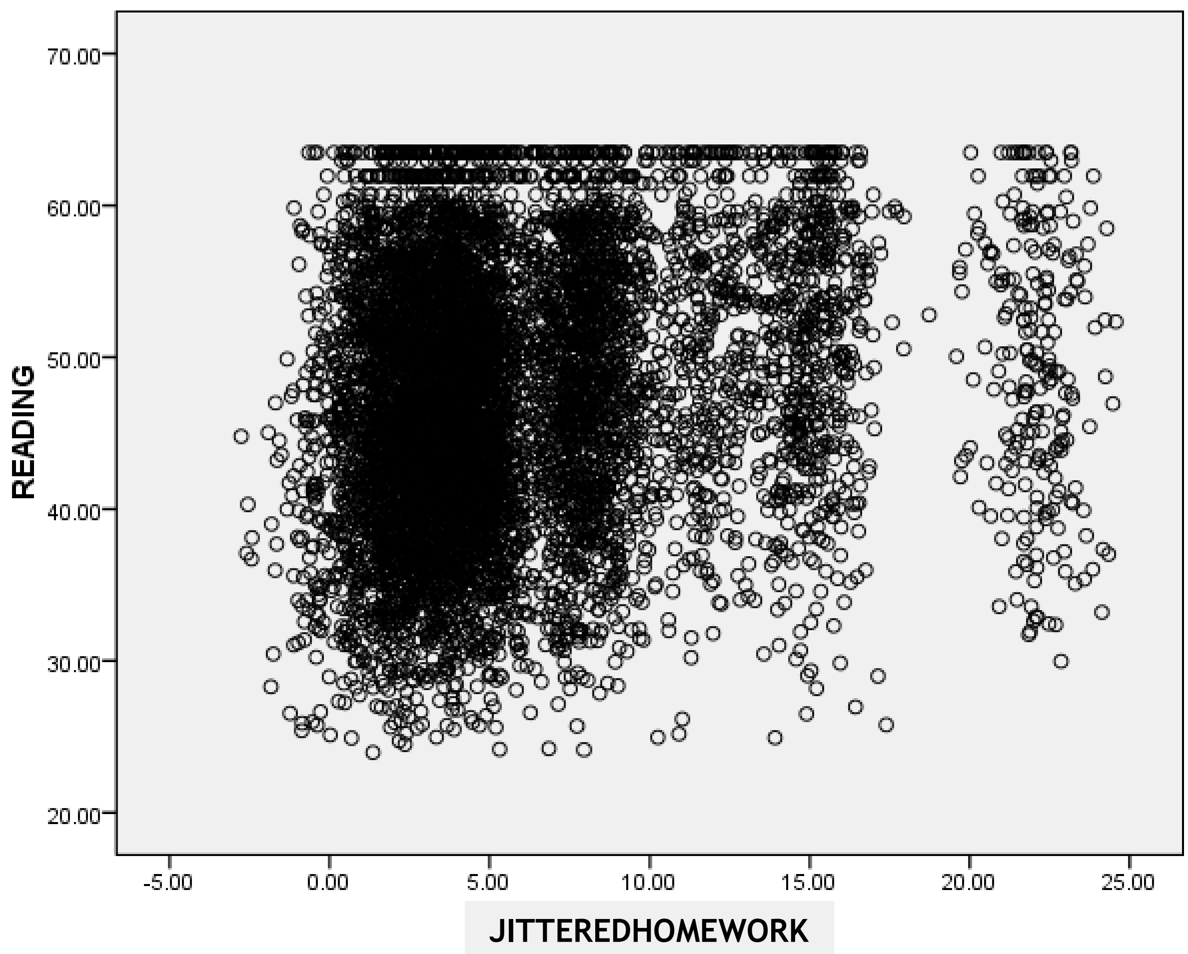
Notice that linearity is not worth mentioning. There are only two conditional averages, so a straight line will always fit best.

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 49.118 | .115 | | 428.169 | .000 | 48.893 | 49.342 |
| | FREELUNCH | -4.841 | .198 | -.267 | -24.439 | .000 | -5.229 | -4.453 |

a. Dependent Variable: READING

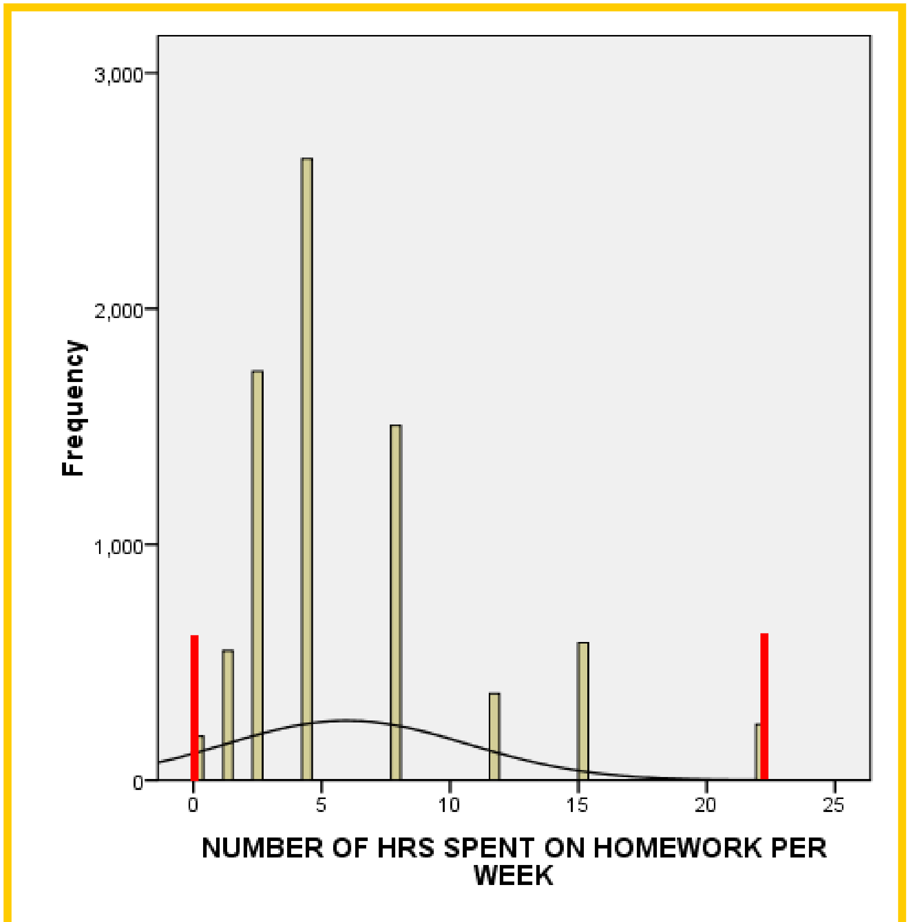
# Adding Measurement Error for the Sake of EDA (Jittering)

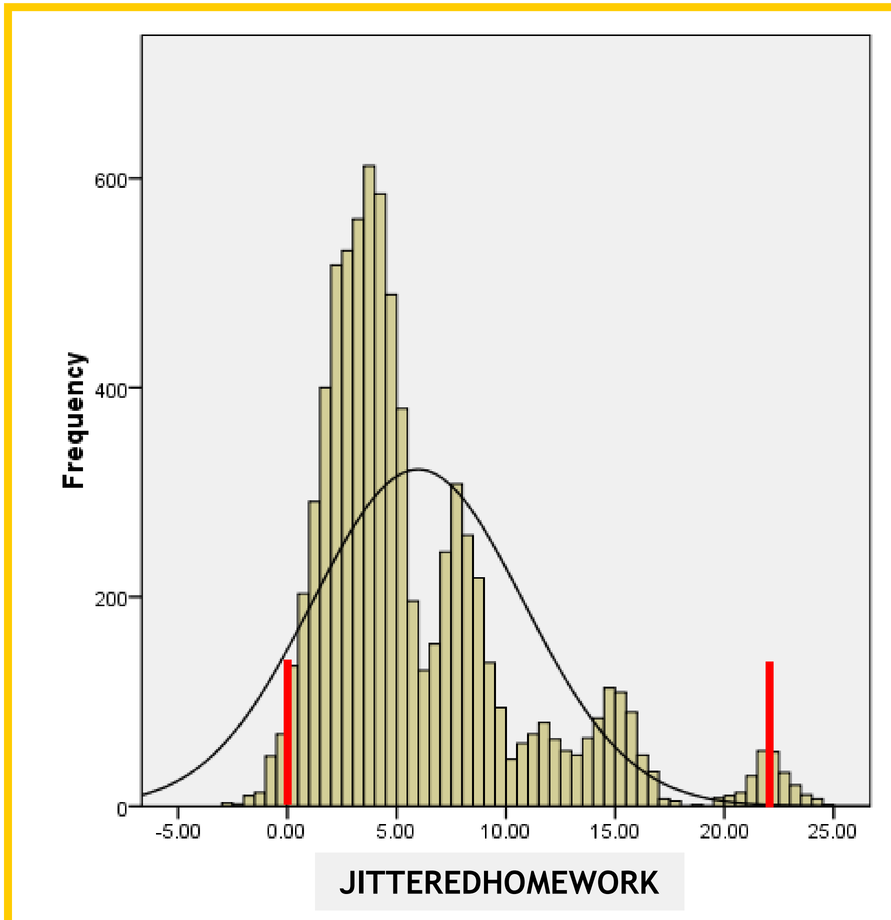



Riddle me this: What would happen if we jittered the *READING* data as well?
Riddle me that: What would happen if we jittered *JITTEREDHOMEWORK*?

Think of a random number from a distribution with a mean of 0 and a standard deviation of 1. Here are a few: 0.3, -1.2, -0.8 and 3.1. We could come up with a list of 7,800 such random numbers. When we jitter data, we add a little randomness to each datum. Our *HOMEWORK* data consist of 7,800 observations. Here are a few: 8, 2.5, 0, and 8. We can add a random number to each observation: 8.3, 1.3, -0.8 and 11.1. Thus, we make a new variable that we can call, "*JITTEREDHOMEWORK*." This adding of random smidges is a funky transformation called "jittering." Because we jittered our X variable, our scatterplot spreads out horizontally. Generally, we do not want to add noise to our data, but sometimes it helps in EDA because, by giving the scatterplot a good shake, it topples the stacks of points, making the individual points more visible to the eye. Do you see?

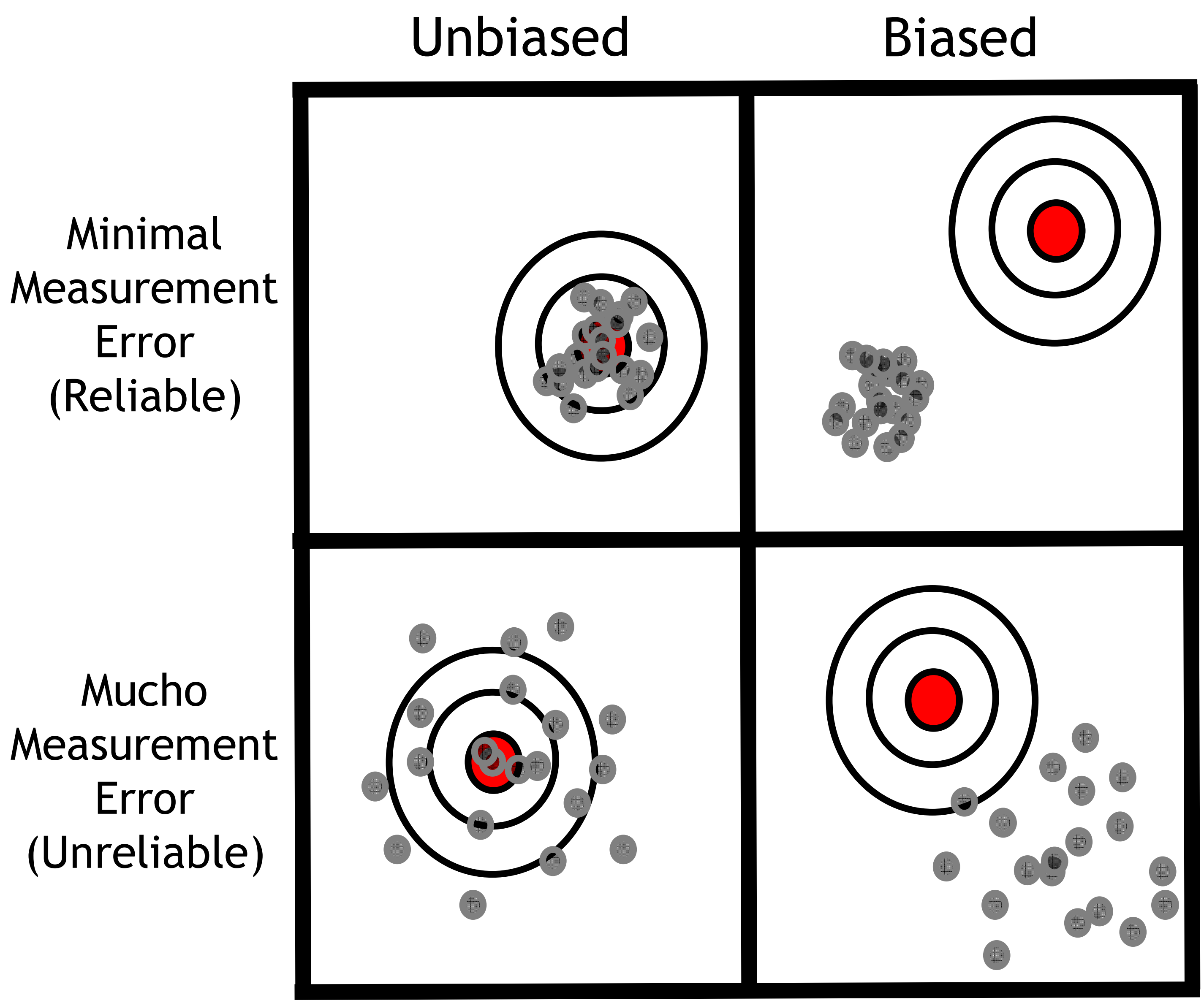
# Jittering From a Univariate Perspective




By adding measurement error, we increase the spread of the distribution. Often in data analysis, we like spread; after all, it gives us something to analyze. There is no analysis of variance (ANOVA) without variance. However, we data analysts don't like just any spread. We like spread that represents real variation in real life, and within that variation we want to detect patterns, or signals. The spread caused by measurement error is random; it is not real variation in real life. Instead, measurement error is only noise that obscures the signals that we are trying to detect. Jittering is helpful in EDA, but we never want to use jittered data in confirmatory data analysis (CDA).

# Measurement Error: The Dart Metaphor




In the next room, on the wall, there are painted four bull's eyes, the sort of things at which one throws darts. Imagine that you are tasked with, one at a time, putting your finger on each bull's eye. The trick is that you are blindfolded. You can feel the wall with your hands, but you can't sense the differences in paint. From your perspective, the wall is perfectly smooth. Your only guide is a few dart holes, created by four different throwers, each aiming for a different bull's eye.

# Defining our Terms (Measurement Error)

**Measurement Error is inconsistency across instances of measurement. The hallmark of measurement error is that it is random, thus it washes out by averaging over repeated measurements. Three sources of measurement error are occasions, raters and tasks. Unreliable tests are tests with too much measurement error.**

**Measurement Error Due to Occasions: When you take the GRE, sometimes it's your lucky day, and sometimes it's not. On any three given days, you could take the exact some test and get three different scores. Sometimes your are stuck with head aches, stomach aches, or heart aches. Sometimes you are struck with keen insights, trustworthy instincts or energizing elations.**

**Measurement Error Due to Raters: When you submit the GRE, sometimes you get lucky with the rater, and sometimes you don't. Somebody (actually two somebodies, or soon to be one somebody and one robot) reads and rates your essay. You could submit the same exact essay three times and get three different grades. If the GRE is reliable, then those three grades will most likely be close. If not, then not.**

**Measurement Error Due to Tasks: When you take the GRE, sometimes you get lucky with the items, and sometimes you don't. You could get a question that is really easy, but you were sick the week they covered it in 7$^{th}$ grade. You could get a question that is really hard, but you did your 11$^{th}$ grade project exactly on the topic. When it comes to tasks (or items), you get what you get, so don't get upset, but it does matter. The longer the test, however, the more likely that all your good luck will balance out your bad luck. I.e., the longer the test, the more reliable (all else being equal).**

# Defining our Terms (Bias)

Measurement Bias is systematic error that consistently leads one to the wrong inference, even after averaging over repeated measurements.

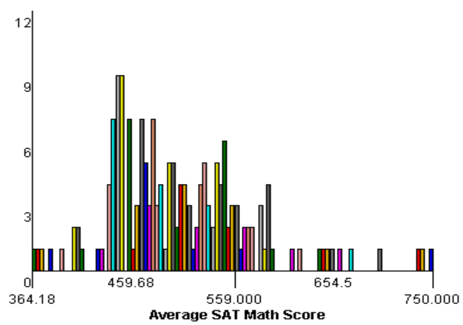
On average, females tend to do worse on multiple choice questions than open response questions and essay questions. Whereas, on average, males tend to do worse on open response questions and essay questions than multiple choice questions. If we take two people of the exact same ability level, one female, Maria, and the other male, Marco, who are typical of their sex when it comes to item format, we can bias the test in favor of one over the other by adjusting the mix of multiple choice, open response and essay questions. Supposing we bias the test against males, by making it all open response, it does not follow, however, that Marco will always do worse than Maria. Marco could always get lucky with respect to the particular rater (e.g., the rater just ate a granola bar before reading Marco's exam), the particular items (e.g., there was a question prompt centered around Marco's favorite subject, skateboarding) or the particular occasion (e.g., Marco was still riding the high from his first 360 flip). Marco's luck could dwarf the bias, but his high score would have been even higher were there no bias against him. Bias is often very difficult to detect, because we need to identify differences in scores of equally able people after accounting for measurement error.

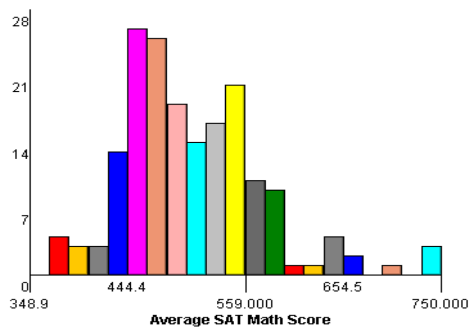
Measurement error is random. Bias is systematic.

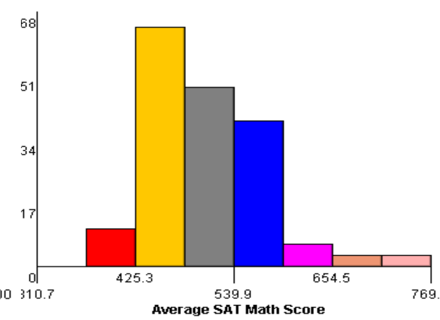
# Artificial Binning

When we take a naturally continuous variable and break it into artificial categories (or bins), we lose information. Not all binning is artificial. *SEX* is naturally binned into "male/female," because SEX, as biologically defined in terms of gametes, is naturally dichotomous. On the other hand, *GENDER* is artificially binned into "masculine/feminine," because, as a social construct, *GENDER* forms a continuum. Note that researchers (among others) often mean "sex" when they use "gender."

For example, *HOMEWORK* has eight bins: 0, 1.25, 2.5, 4.25, 8, 11.75, 15, or 22 hours. A student who does 2 hours of homework per week gets binned into either the 1.25 or 2.5 category. The former category is wrong by 0.75 hours, and the latter is wrong by 0.5 hours. Either way, there is error.

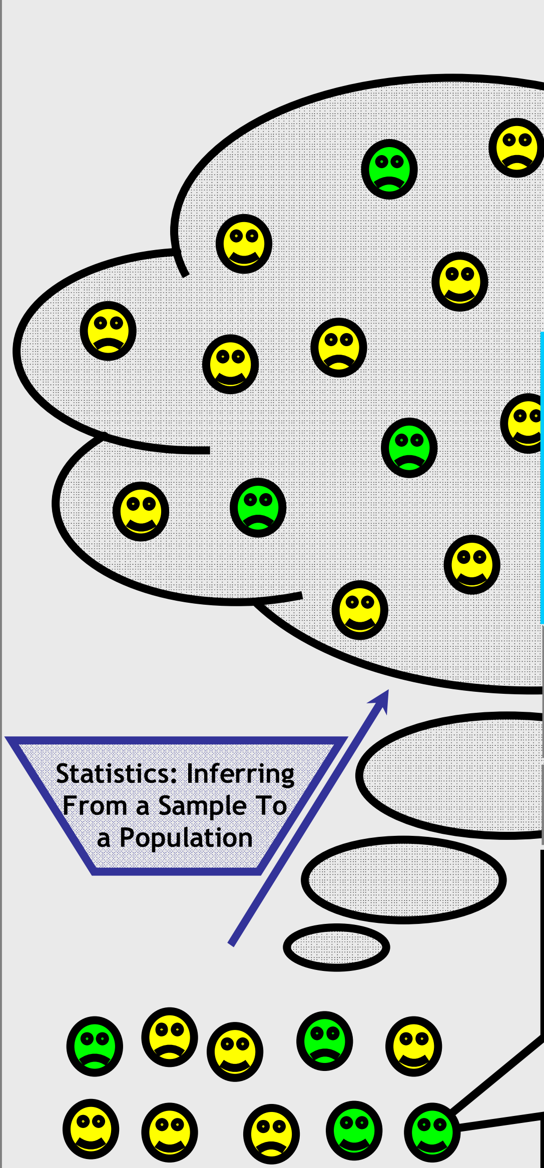



http://www.shodor.org/interactivate/activities/Histogram/

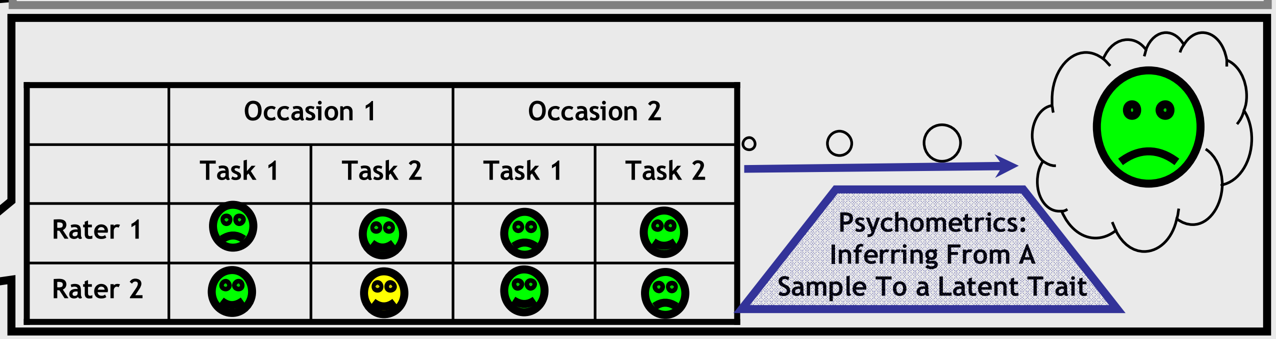
# Statistical and Psychometric Inferences




You might describe statistics as largely concerned with drawing inferences from samples to populations. Likewise, you might describe psychometrics as largely concerned with drawing inferences from samples (of raters, tasks and occasions) to latent traits.

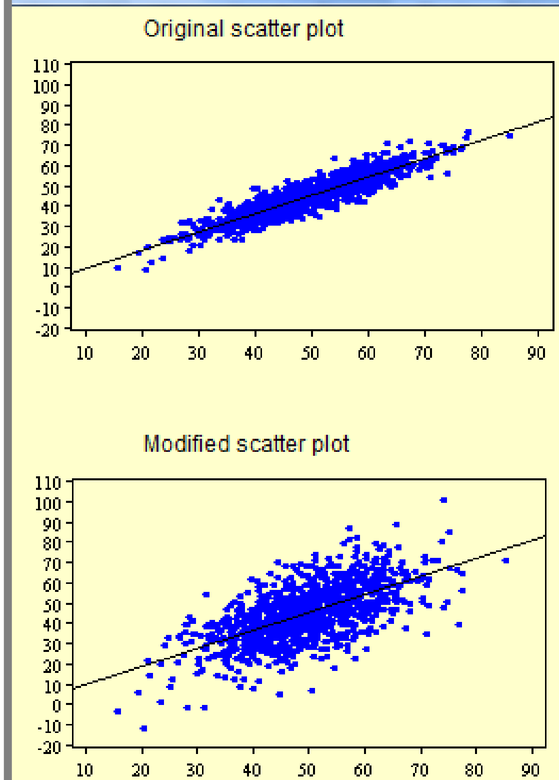
Statistical Conclusion: In the population of Smileys, yellows tend to be happier than greens.

Psychometric Conclusion: George Smiley is green and happy.



# The Upshot I: Measurement Error and Outcomes (ATTENUATION)

**Measurement error in an outcome variable adds noise. It makes effects harder to detect, but statistical methods are designed to handle fallible (i.e., less than perfectly reliable) outcomes. We expect residuals whenever we fit a model, and we know that residuals are due to unaccounted variables, individual variation and <u>measurement error</u>. Measurement error in an outcome leads to attenuated (i.e., weakened) correlations and inflated standard errors, but that's a part of the game; it's a part of the game that we want to minimize, but it's a part of the game that can handle.**




| | Original | Modified |
|---|---|---|
| Pearson's r | 0.905 | 0.628 |
| slope | 0.905 | 0.888 |
| y intercept | 0.235 | 1.080 |
| sd of x | 10.000 | 10.000 |
| sd of y | 10.000 | 14.142 |
| se of the est. | 4.255 | 11.002 |
| Reliability of x | 1.000 | 1.000 |
| Reliability of y | 1.000 | 0.500 |

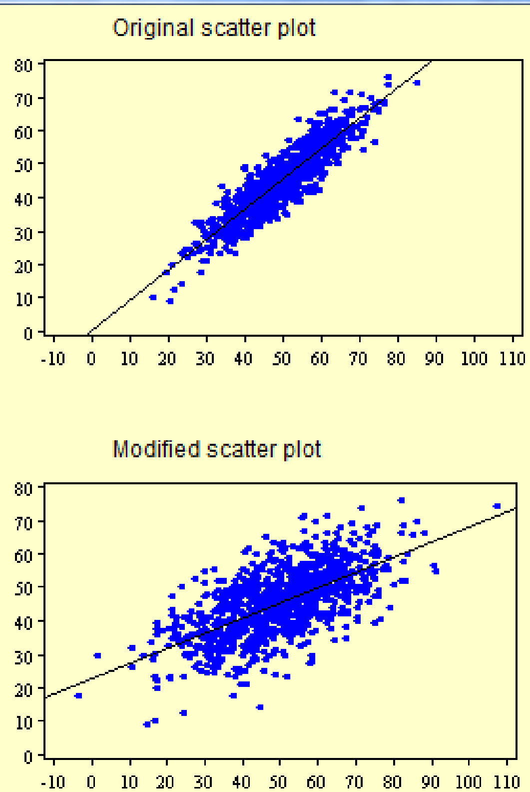
| sd of x error | sd of y error | |
|---|---|---|
| 0 | 10 | OK |

**Notice that the correlation (Pearson's r) changes but the slope estimate does not change when we add measurement error to the outcome (the *y* variable). Recall that Pearson correlation is an indicator of relationship strength, and the slope is an indicator of relationship magnitude. Strength is important for statistical significance. Magnitude is important for policy decisions.**

[http://onlinestatbook.com/stat_sim/reliability_reg/index.html](http://onlinestatbook.com/stat_sim/reliability_reg/index.html)

# The Upshot II: Measurement Error and Predictors (BIAS)

**Measurement error in a predictor variable is problematic. It biases our population estimation (i.e. magnitude) in unpredictable ways. Inferences about latent traits can be biased and inferences about populations can also be biased. For example, a biased sample can lead to a biased inference about the population. Likewise, a biased measurement of an outcome or predictor can lead to a biased inference about a population. Less obvious, but central to this unit: an unreliable measurement of a predictor (but not an outcome!) can lead to a biased inference about a population.**




| | Original | Modified |
|---|---|---|
| Pearson's r | 0.905 | 0.643 |
| slope | 0.905 | 0.455 |
| y intercept | 0.235 | 22.651 |
| sd of x | 10.000 | 14.142 |
| sd of y | 10.000 | 10.000 |
| se of the est. | 4.255 | 7.656 |
| Reliability of x | 1.000 | 0.500 |
| Reliability of y | 1.000 | 1.000 |

| sd of x error | sd of y error | |
|---|---|---|
| 10 | 0 | OK |

**Notice that not only the correlation (Pearson's r) changes but also the slope estimate changes when we add measurement error to the predictor (the *x* variable). As with the previous slide's measurement error in the outcome, the attenuated correlation makes it harder to attain statistical significance. However, unlike earlier, when statistical significance is attained, we end up with the wrong relationship for the purposes of policy decisions.**

# Take-Home Lesson

| | Measurement Error | Measurement Bias |
|---|---|---|
| Regression Outcome | Regression Attenuation<br>•Strength Weakened<br>•Correlation Weakened<br>•Standard Errors Inflated<br>•But, Magnitude Okay | Regression Bias |
| Regression Predictor | Regression Bias<br>•Strength Weakened<br>•Correlation Weakened<br>•Standard Errors Inflated<br>•And, Magnitude Biased | Regression Bias |

Bad: Measurement error in our outcome—ATTENUATION.

Evil: Measurement error in our predictor—BIAS.

Evil: Measurement bias in our outcome—BIAS.

Evil: Measurement bias in our predictor—BIAS.

# GLM and Assumptions About Measurement Error

The general linear model (GLM, encompassing t-tests, ANOVA and regression) is designed to handle measurement error in the outcome, but it assumes no measurement error in the predictor(s).


For every variable in your model, assess its potential for measurement error theoretically. A course in psychometrics will give you the skills to assess measurement error empirically. For now, consider artificial binning, occasions, raters and tasks as potential sources of measurement error.

## ABORT: Artificial Binning, Occasions, Raters, Tasks

- Measurement error in an outcome attenuates the relationship.
- Measurement error in a predictor biases the relationship.

You now have what you need to do the Unit 11 Post Hole: Note the threats to validity posed by measurement error in the outcome and predictor(s) of a model.

# Measurement Error When Regressing *READING* on *HOMEWORK*

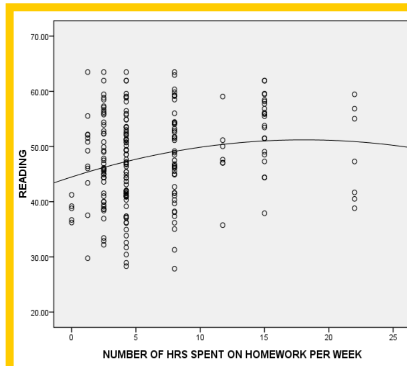
***READING***, our outcome, is a standardized test which has measurement error due to occasion (good test days and bad test days), items (lucky test items and unlucky test items) and, unless the test is completely machine graded, raters (lenient raters and harsh raters). Because this error is in the outcome, we expect an attenuated correlation, which would show up in the r = 0.18.

***HOMEWORK***, our predictor, is probably rife with measurement error (and bias). The unnatural bins will introduce error by forcing responses away from precise times into rough categories. Students may give different answers on different occasions. I also suspect a positive bias due to social expectations. We must therefore question the validity of our finding that each hour of homework per week is associated with 0.33 points on the reading test.

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 45.514 | .154 | | 296.359 | .000 |
| | NUMBER OF HRS SPENT ON HOMEWORK PER WEEK | .332 | .020 | .183 | 16.451 | .000 |

a. Dependent Variable: READING




**Warning! Being too much of a stickler can be hazardous to your career. How much is too much? I recommend that you be slightly less tolerant than your field, which will probably be very tolerant.**

**Note that *READING* also has a ceiling effect, where the best readers can't score higher than the top score. In effect, the top readers get artificially binned at the top score. This artificial binning is not random! It only happens to the best readers, and it always lowers their score. READING is, thus, a biased outcome measure. For future reference, Tobit regression is designed to correct for the bias due to floors and ceilings (i.e. censoring) in our outcomes.**

# Bivariate Exploratory Data Analysis: *READING* vs. *FREELUNCH*

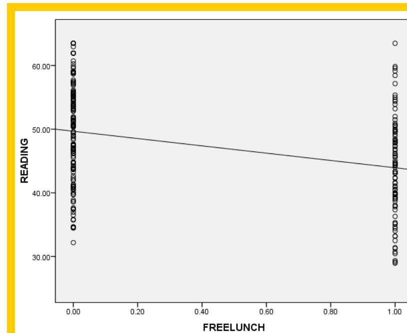
As noted in the previous slide, we expect the correlation to be attenuated because of measurement error in our outcome, *READING*, so our observed Pearson correlation of -0.27 is an underestimate.

Our predictor, *FREELUNCH*, may be biased because it is based on income which people may tend to exaggerate or understate. I do not think there would be much error based on differing income questions or occasions of answering (although fluctuating job markets may contribute to error from occasion as incomes bounce up and down while SES stays fairly stable). As a proxy for socioeconomic status, *FREELUNCH* pulls people into artificially course categories from the refined relative rankings corresponding to a truly continuous latent trait. It is also unclear if income, especially alone, is a valid measure of SES, even if it were treated continuously. In the end, we should seek a better measure of SES. If we use eligibility for free/reduced lunch as a flag only, it is reasonable to conclude that being eligible is associated with 5 fewer points on the test, but we should tread cautiously if we intend to draw inferences about SES.

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 49.118 | .115 | | 428.169 | .000 |
| | FREELUNCH | -4.841 | .198 | -.267 | -24.439 | .000 |

a. Dependent Variable: READING




**Warning! Being too much of a stickler can be hazardous to your career. In this case, however, we have continuous SES data, so it would be foolish to use the dichotomous *FREELUNCH* as a proxy.**

# Answering the Roadmap Question

- ***READING*, *HOMEWORK*, *FREELUNCH*: Previous Slides.**
- ***RACE* (Question Predictor):**
  - Artificial Binning: Is there measurement error in our *RACE* variable that has only four categories: Asian, Black, Latino and White? There appears to be artificial binning. I'm no expert here, but I doubt that race "naturally" falls into only four categories. Is the artificiality of the binning random or systematic? If it is random, then it is measurement error. If it is systematic, then it is bias.
  - Occasion: Do students vary their self-identification from day-to-day? Possibly. Who knows with teenagers and their identity crises? But I wouldn't worry too much here.
  - Rater: These data are based on self-identification, so there is no rater. (Back in the day, researchers would ask school principals, and you can imagine the debacle.)
  - Task: Do students respond differently based on the wording of the question? Possibly, but I'm not going to worry about it. I WILL worry about artificial binning.
  - **Final Conclusion: Some bias for regression, and I would seek more nuanced categories.**
- ***ESL* (Control Predictor):**
  - Artificial Binning: Yes. English language proficiency is a continuum, and we are treating it like a dichotomy.
  - Occasion: This is probably pretty stable from day-to-day.
  - Rater: Some school systems may be more or less strict in their ESL determinations.
  - Task: A student taking the ESL test may get "lucky" or "unlucky" items.
  - **Final Conclusion: Some bias for regression, but I wouldn't let is break me. If I could get the ESL test scores, I would feel better.**

# Unit 11 Appendix: Key Concepts

- Measurement error is random. Bias is systematic.
- Measurement error in an outcome variable adds noise. It makes effects harder to detect, but statistic methods are designed to handle fallible (i.e., less than perfectly reliable) outcomes. We expect residuals whenever we fit a model, and we know that residuals are due to unaccounted variables, individual variation and <u>measurement error</u>. Measurement error in an outcome leads to attenuated (i.e., weakened) correlations, but that's a part of the game; it's a part of the game that we want to minimize, but it's a part of the game that we understand and can handle.
- Measurement error in a predictor variable is problematic. It biases our population estimation in unpredictable ways. Inferences about latent traits can be biased and inferences about populations can also be biased. For example, a biased sample can lead to a biased inference about the population. Likewise, a biased measurement of an outcome or predictor can lead to a biased inference about a population. Less obvious, but central to this unit: an unreliable measurement of a predictor (but not an outcome!) can lead to a biased inference about a population.

# Unit 11 Appendix: Key Interpretations

The distribution of reading scores is approximately normal, except for an obvious ceiling effect at the top score of 63. In our sample, the reading scores range from 24 points to 63 points, thus a range of 39 points, with a midspread of 13 points and a median of 47 points. There does not appear to be any outliers.

The distribution of hours spent on homework is positively skewed. We should note that respondents had only eight choices: about 0, 1.25, 2.5, 4.25, 8, 11.75, 15, or 22 hours. The median and mode is 4.25. Half the students report studying between 2.5 and 8 hours per week.

In our sample of 7,800 8th graders, 2616 (33.5%) were eligible for free/reduced lunch.

The relationship between reading achievement and eligibility for free/reduced lunch is negative in our sample. On average, students who are eligible for free/reduced lunch score 5 points less than their ineligible counterparts. The average eligible student attains a reading score of 44, and the average eligible student attains a reading score of 49. However, due to broad variation among individuals, the relationship is weak, so these group averages should not be the basis for judgments about individuals. There are no apparent outliers.

*READING*, our outcome, is a standardized test which has measurement error due to occasion (good test days and bad test days), items (lucky test items and unlucky test items) and, unless the test is completely machine graded, raters (lenient raters and harsh raters). Because this error is in the outcome, we expect an attenuated correlation, which would show up in the $r = 0.18$.

*HOMEWORK*, our predictor, is probably rife with measurement error (and bias). The unnatural bins will introduce error by forcing responses away from precise times into rough categories. Students may give different answers on different occasions. I also suspect a positive bias due to social expectations. We must therefore question the validity of our finding that each hour of homework per week is associated with 0.33 points on the reading test.

Our predictor, *FREELUNCH*, may be biased because it is based on income which people may tend to exaggerate or understate. I do not think there would be much error based on differing income questions or occasions of answering (although fluctuating job markets may contribute to error from occasion as incomes bounce up and down while SES stays fairly stable). As a proxy for socioeconomic status, *FREELUNCH* pulls people into artificially course categories from the refined relative rankings corresponding to a truly continuous latent trait. It is also unclear if income, especially alone, is a valid measure of SES, even if it were treated continuously. In the end, we should seek a better measure of SES. If we use eligibility for free/reduced lunch as a flag only, it is reasonable to conclude that being eligible is associated with 5 fewer points on the test, but we should tread cautiously if we intend to draw inferences about SES.

# Unit 11 Appendix: Key Terminology

- Measurement Error is inconsistency across instances of measurement. The hallmark of measurement error is that it is random, thus it washes out by averaging over repeated measurements. Three sources of measurement error are occasions, raters and tasks/items. Unreliable tests are tests with too much measurement error.
- When we take a naturally continuous variable and break it into artificial categories (or bins), we lose information. Not all binning is artificial. *SEX* is naturally binned into "male/female," because SEX, as biologically defined in terms of gametes, is naturally dichotomous. On the other hand, *GENDER* is artificially binned into "masculine/feminine," because, as a social construct, *GENDER* forms a continuum. Note that researchers (among others) often mean "sex" when they use "gender."
- Measurement Bias is systematic error that consistently leads one to the wrong inference, even after averaging over repeated measurements.
- You might describe statistics as largely concerned with drawing inferences from samples to populations. Likewise, you might describe psychometrics as largely concerned with drawing inferences from samples (of raters, items/tasks and occasions) to latent traits.

# Unit 11 Appendix: SPSS Syntax

```
*You can use my code by switching out my variables (circled) with your variables.
*You can make a comment by starting with an asterisk and ending with a period.
*SPSS will ignore anything between the asterisk and period.
*SPSS loves/needs to end chunks of command with a period, so if something is acting funky,
   make sure that your periods are in order.
*Also, some command need to be executed by being followed by an "EXECUTE" line
   with period of course.
**************************************************************************.

*Create a table of descriptive statistics.
FREQUENCIES VARIABLES=READING HOMEWORK FREELUNCH
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN
  /ORDER=ANALYSIS.
*Create a histogram.
GRAPH
  /HISTOGRAM(NORMAL)=HOMEWORK.
*Create a bivariate scatterplot.
GRAPH
  /SCATTERPLOT(BIVAR)=HOMEWORK WITH READING
  /MISSING=LISTWISE.
```

# Unit 11 Appendix: SPSS Syntax

```
*The "TEMPORARY" command tells SPSS to do the next thing only for command.
*The SAMPLE command tell SPSS to take a random sample (of 3% in this case) from the full data set.
TEMPORARY.
SAMPLE 0.03.
GRAPH
  /SCATTERPLOT(BIVAR)=HOMEWORK WITH READING
  /MISSING=LISTWISE.
*The "TEMPORARY" command "SAMPLE" is over now.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT READING
  /METHOD=ENTER HOMEWORK.
*Create a jiggered variable by adding a random number drawn from a mean = 0, sd =1 normal distribution.
*You can make the standard deviation any number, but I happened to choose 1 as may not in the parentheses.
COMPUTE JIGGEREDHOMEWORK = HOMEWORK+ NORMAL (1).
EXECUTE.
```

# Preparing for the Memos: Using SPSS

- **As per the programming in Units 1 and 2:**
  - Produce a scatterplot in SPSS or R.
  - Fit a trend line using SPSS chart editor, if you use SPSS.
  - Fit a regression model in SPSS or R to obtain the computer generated y-intercept and slope for the trend lines.
  - Produce descriptive statistics for each variable.
  - Produce a histogram (with a normal curve overlay) for each variable.

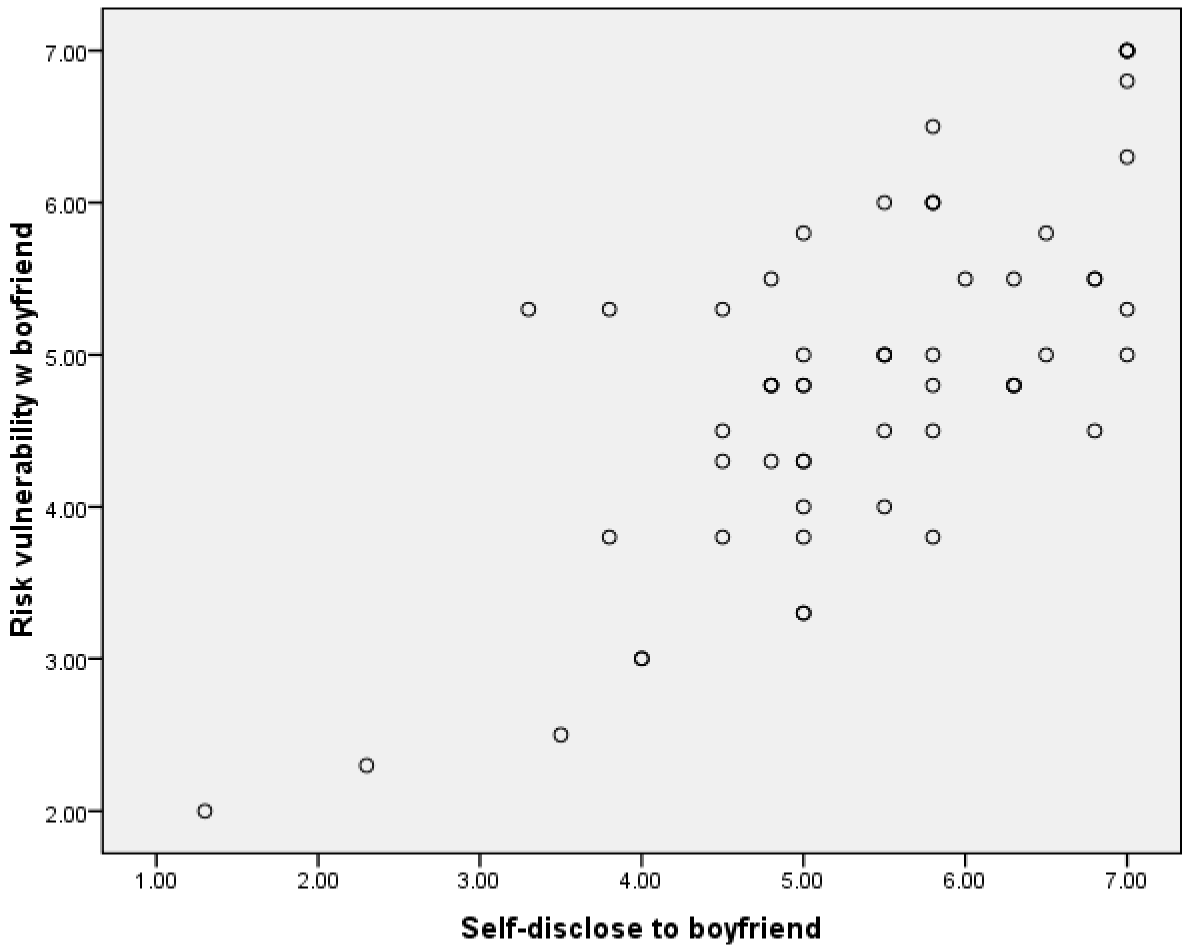
# Perceived Intimacy of Adolescent Girls (Intimacy.sav)

- **Overview: Dataset contains self-ratings of the intimacy that adolescent girls perceive themselves as having with: (a) their mother and (b) their boyfriend.**
- **Source: HGSE thesis by Dr. Linda Kilner entitled *Intimacy in Female Adolescent's Relationships with Parents and Friends* (1991). Kilner collected the ratings using the Adolescent Intimacy Scale.**
- **Sample: 64 adolescent girls in the sophomore, junior and senior classes of a local suburban public school system.**
- **Variables:**

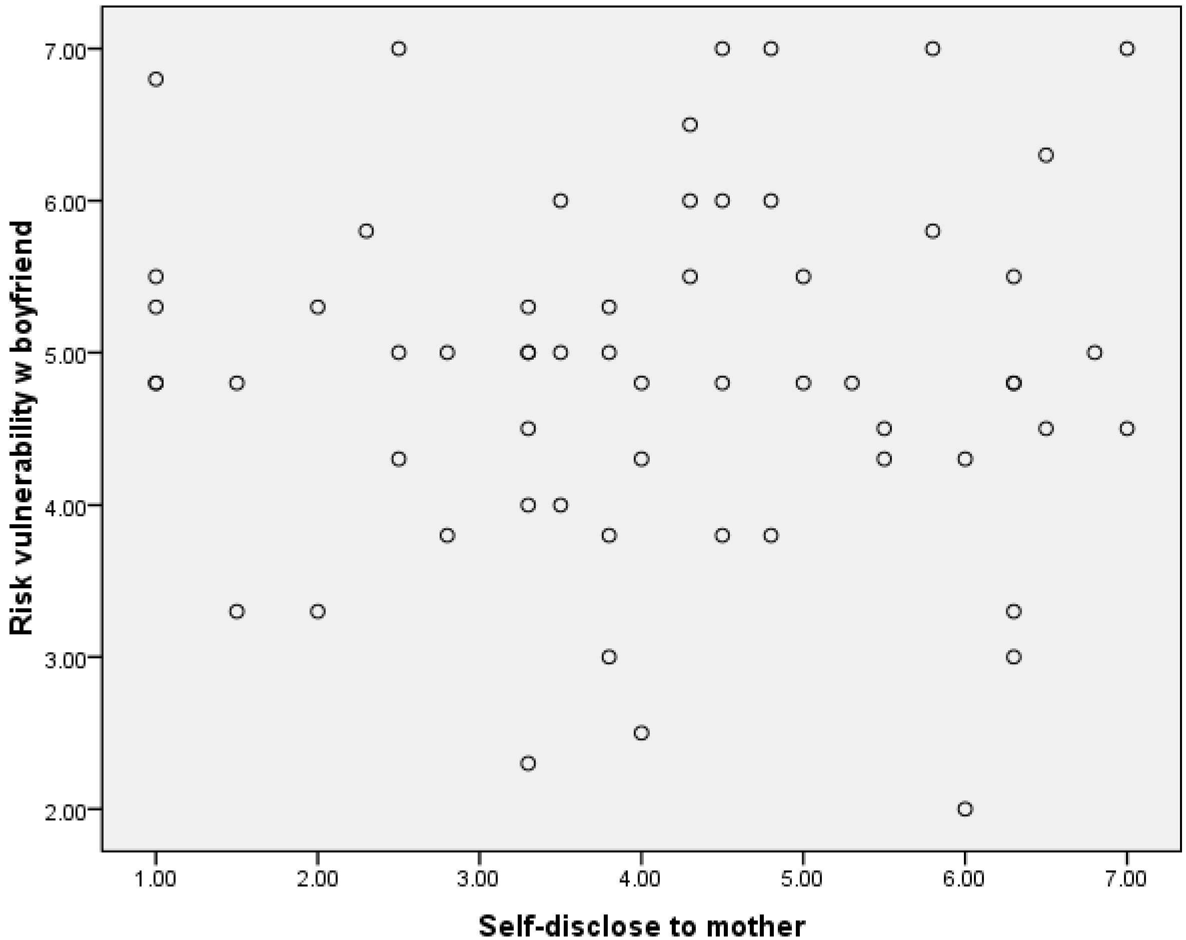
Self Disclosure to Mother (M_Seldis)
Trusts Mother (M_Trust)
Mutual Caring with Mother (M_Care)
Risk Vulnerability with Mother (M_Vuln)
Physical Affection with Mother (M_Phys)
Resolves Conflicts with Mother (M_Cres)

Self Disclosure to Boyfriend (B_Seldis)
Trusts Boyfriend (B_Trust)
Mutual Caring with Boyfriend (B_Care)
Risk Vulnerability with Boyfriend (B_Vuln)
Physical Affection with Boyfriend (B_Phys)
Resolves Conflicts with Boyfriend (B_Cres)

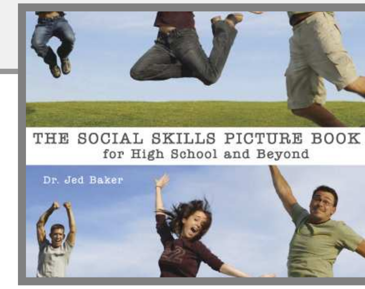
# Perceived Intimacy of Adolescent Girls (Intimacy.sav)



# Perceived Intimacy of Adolescent Girls (Intimacy.sav)



# High School and Beyond (HSB.sav)




- **Overview: High School & Beyond – Subset of data focused on selected student and school characteristics as predictors of academic achievement.**
- **Source: Subset of data graciously provided by Valerie Lee, University of Michigan.**
- **Sample: This subsample has 1044 students in 205 schools. Missing data on the outcome test score and family SES were eliminated. In addition, schools with fewer than 3 students included in this subset of data were excluded.**
- **Variables:**

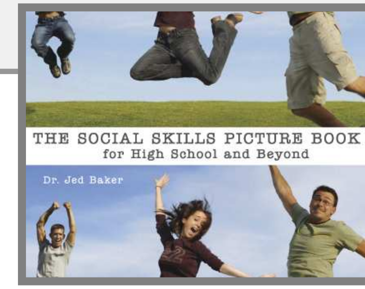
Variables about the student—

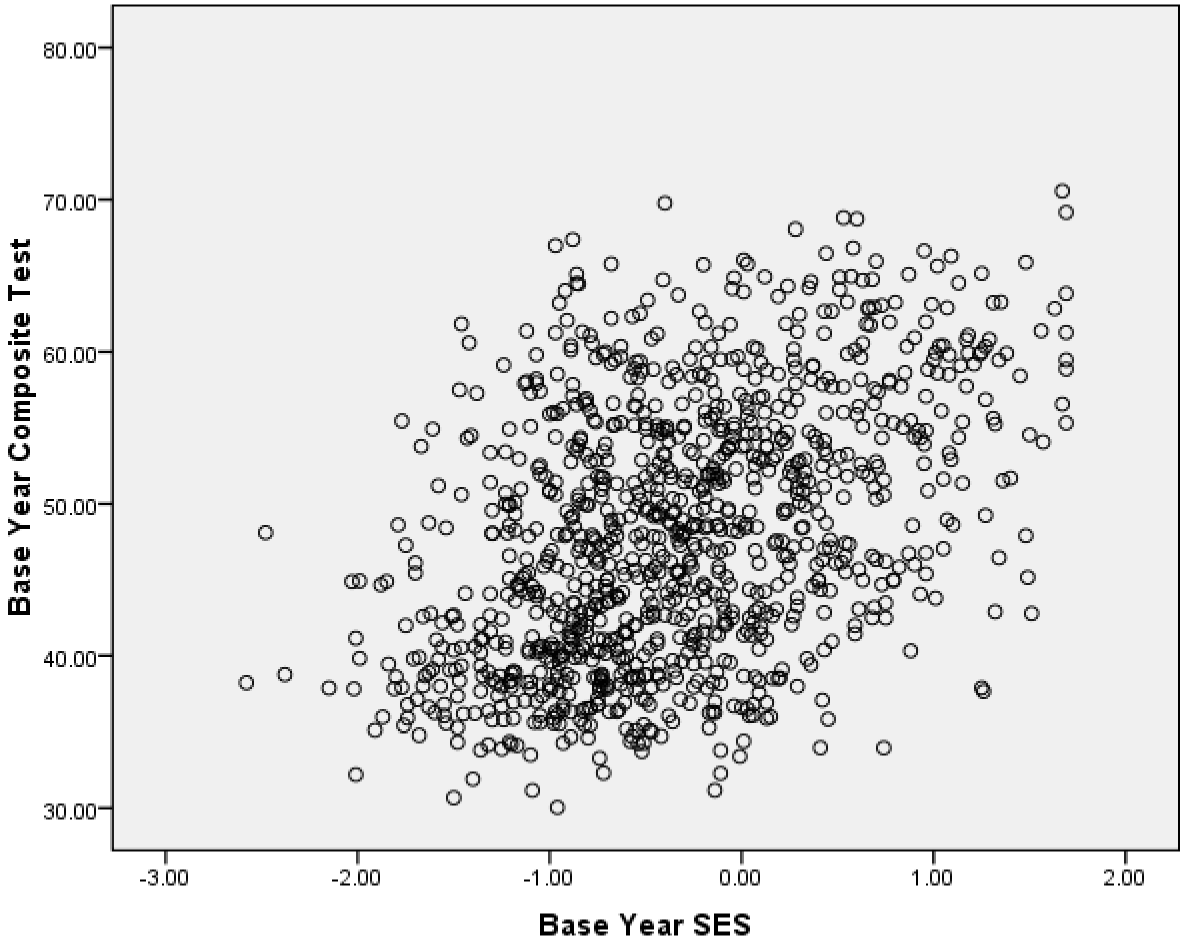
(Black) 1=Black, 0=Other
(Latin) 1=Latino/a, 0=Other
(Sex) 1=Female, 0=Male
(BYSES) Base year SES
(GPA80) HS GPA in 1980
(GPS82) HS GPA in 1982
(BYTest) Base year composite of reading and math tests
(BBConc) Base year self concept
(FEConc) First Follow-up self concept

Variables about the student's school—

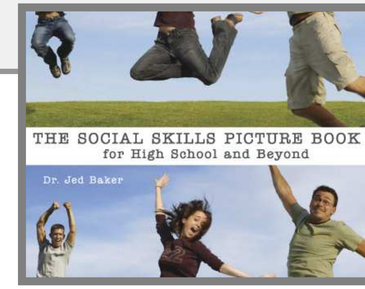
(PctMin) % HS that is minority students Percentage
(HSSize) HS Size
(PctDrop) % dropouts in HS Percentage
(BYSES_S) Average SES in HS sample
(GPA80_S) Average GPA80 in HS sample
(GPA82_S) Average GPA82 in HS sample
(BYTest_S) Average test score in HS sample
(BBConc_S) Average base year self concept in HS sample
(FEConc_S) Average follow-up self concept in HS sample

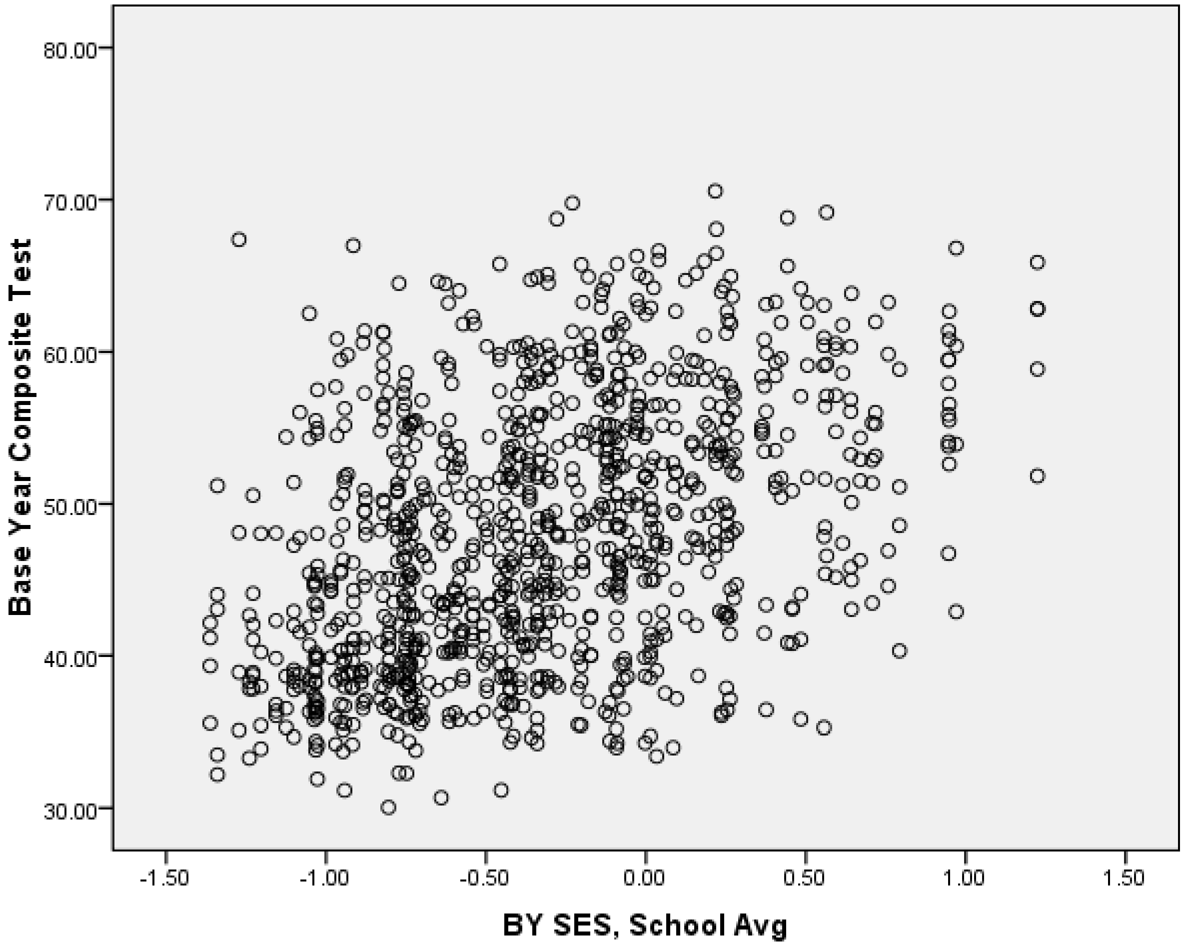
# High School and Beyond (HSB.sav)



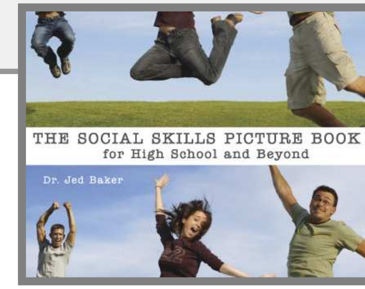



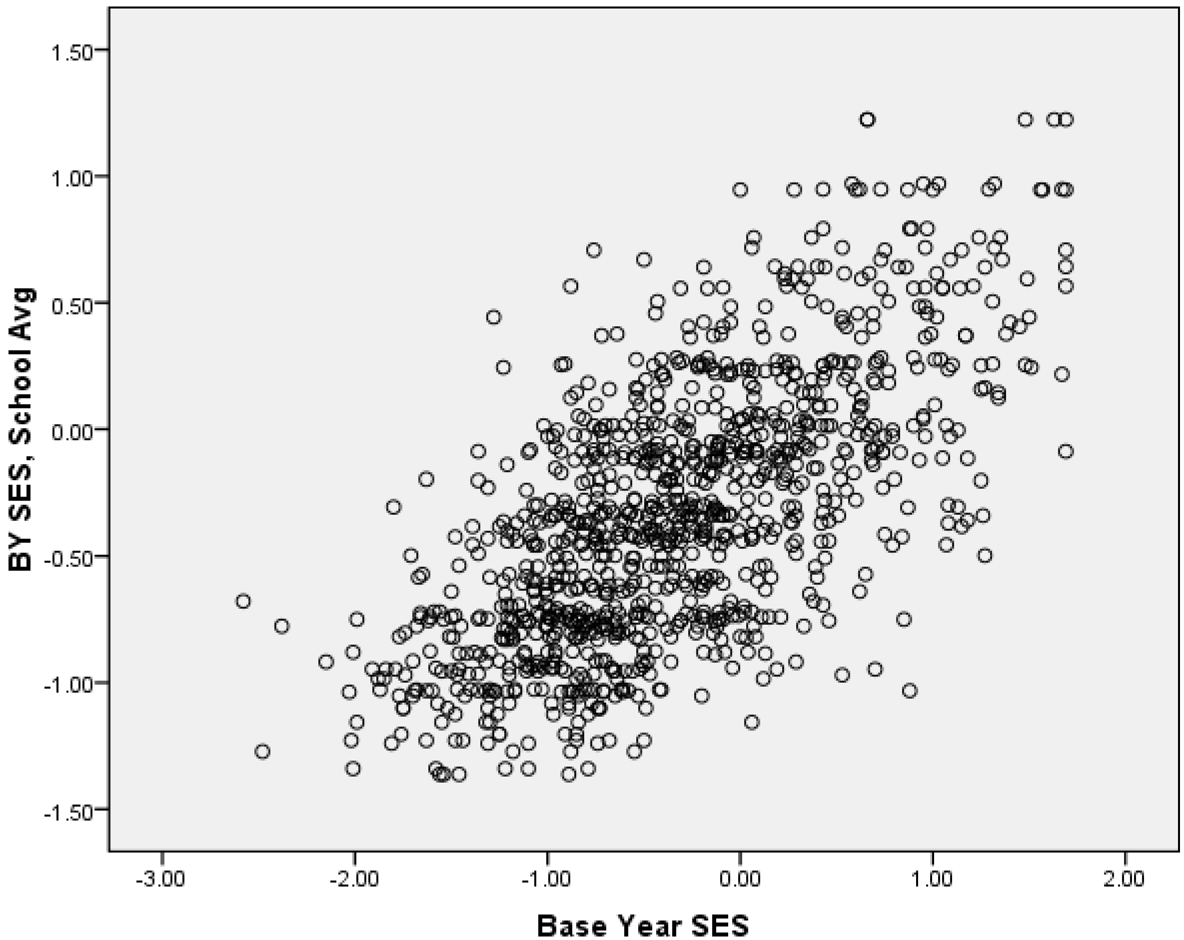
# High School and Beyond (HSB.sav)



# High School and Beyond (HSB.sav)



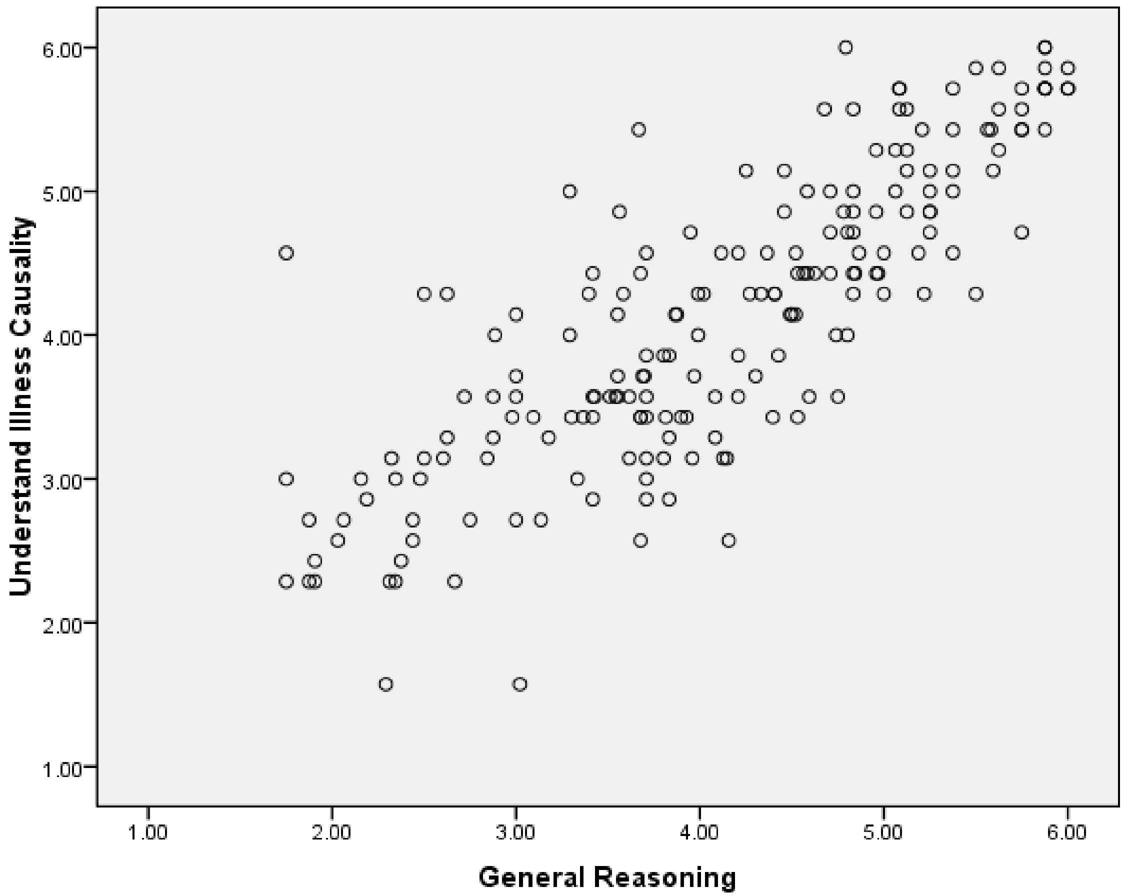
# Understanding Causes of Illness (ILLCAUSE.sav)

- **Overview: Data for investigating differences in children's understanding of the causes of illness, by their health status.**
- **Source: Perrin E.C., Sayer A.G., and Willett J.B. (1991).**

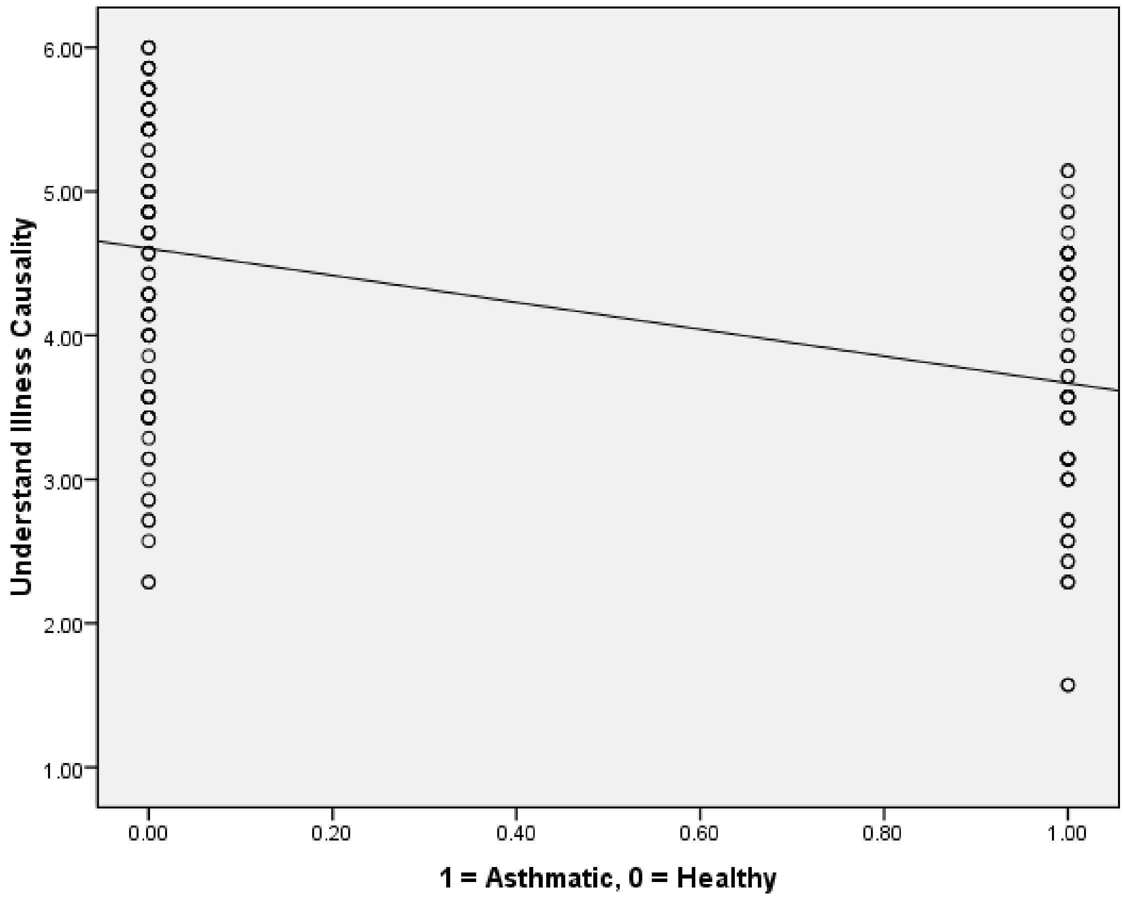
  **Sticks And Stones May Break My Bones: Reasoning About Illness Causality And Body Functioning In Children Who Have A Chronic Illness, *Pediatrics*, 88(3), 608-19.**
- **Sample: 301 children, including a sub-sample of 205 who were described as asthmatic, diabetic,or healthy. After further reductions due to the *list-wise deletion* of cases with missing data on one or more variables, the analytic sub-sample used in class ends up containing: 33 diabetic children, 68 asthmatic children and 93 healthy children.**
- **Variables:**

| | |
|---|---|
| (ILLCAUSE) | Child's Understanding of Illness Causality |
| (SES) | Child's SES (Note that a high score means low SES.) |
| (PPVT) | Child's Score on the Peabody Picture Vocabulary Test |
| (AGE) | Child's Age, In Months |
| (GENREAS) | Child's Score on a General Reasoning Test |
| (ChronicallyIll) | 1 = Asthmatic or Diabetic, 0 = Healthy |
| (Asthmatic) | 1 = Asthmatic, 0 = Healthy |
| (Diabetic) | 1 = Diabetic, 0 = Healthy |

# Understanding Causes of Illness (ILLCAUSE.sav)



# Understanding Causes of Illness (ILLCAUSE.sav)

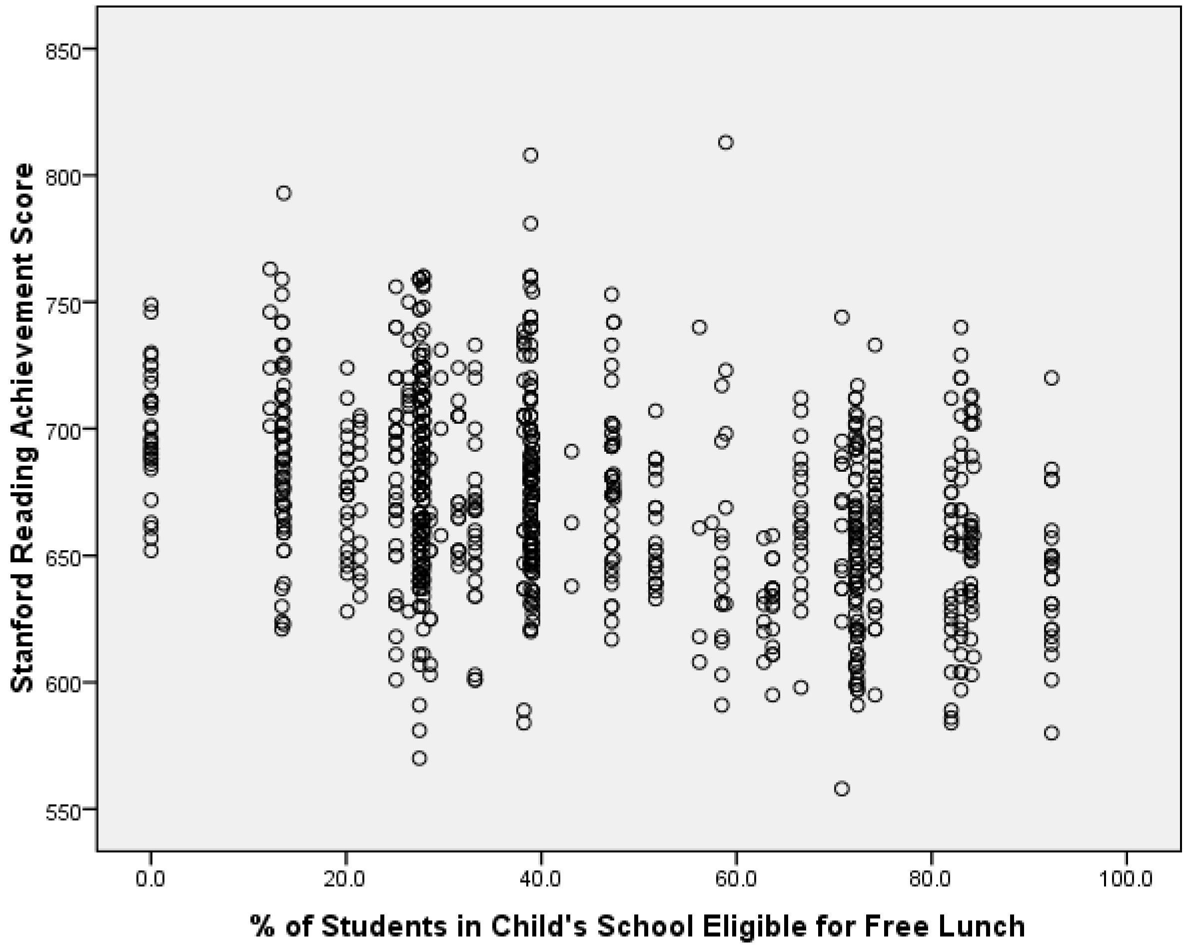


# Children of Immigrants (ChildrenOfImmigrants.sav)

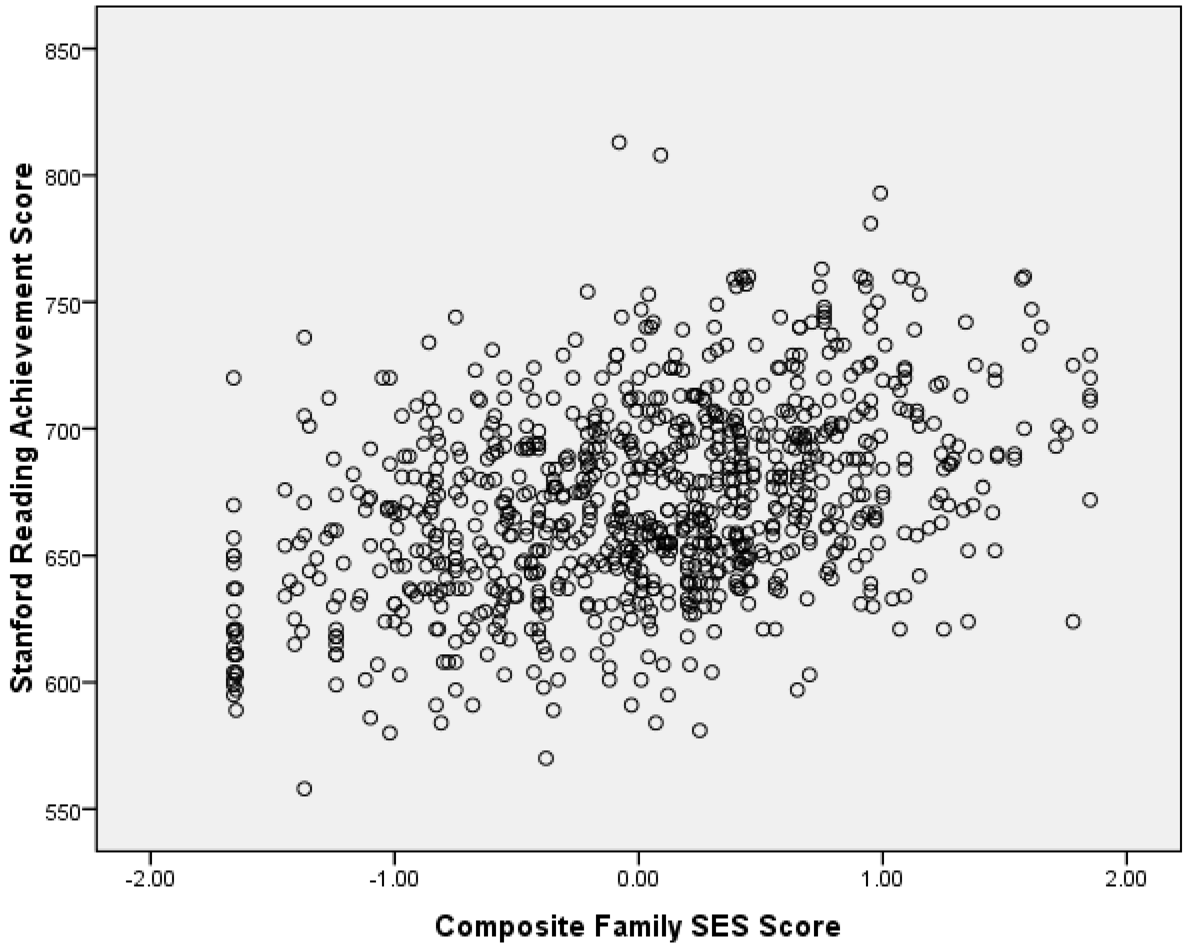
- **Overview: "CILS is a longitudinal study designed to study the adaptation process of the immigrant second generation which is defined broadly as U.S.-born children with at least one foreign-born parent or children born abroad but brought at an early age to the United States. The original survey was conducted with large samples of second-generation children attending the 8th and 9th grades in public and private schools in the metropolitan areas of Miami/Ft. Lauderdale in Florida and San Diego, California" (from the website description of the data set).**
- **Source: Portes, Alejandro, & Ruben G. Rumbaut (2001). *Legacies: The Story of the Immigrant SecondGeneration*. Berkeley CA: University of California Press.**
- **Sample: Random sample of 880 participants obtained through the website.**
- **Variables:**

| | |
|---|---|
| (Reading) | Stanford Reading Achievement Score |
| (Freelunch) | % students in school who are eligible for free lunch program |
| (Male) | 1=Male 0=Female |
| (Depress) | Depression scale (Higher score means more depressed) |
| (SES) | Composite family SES score |

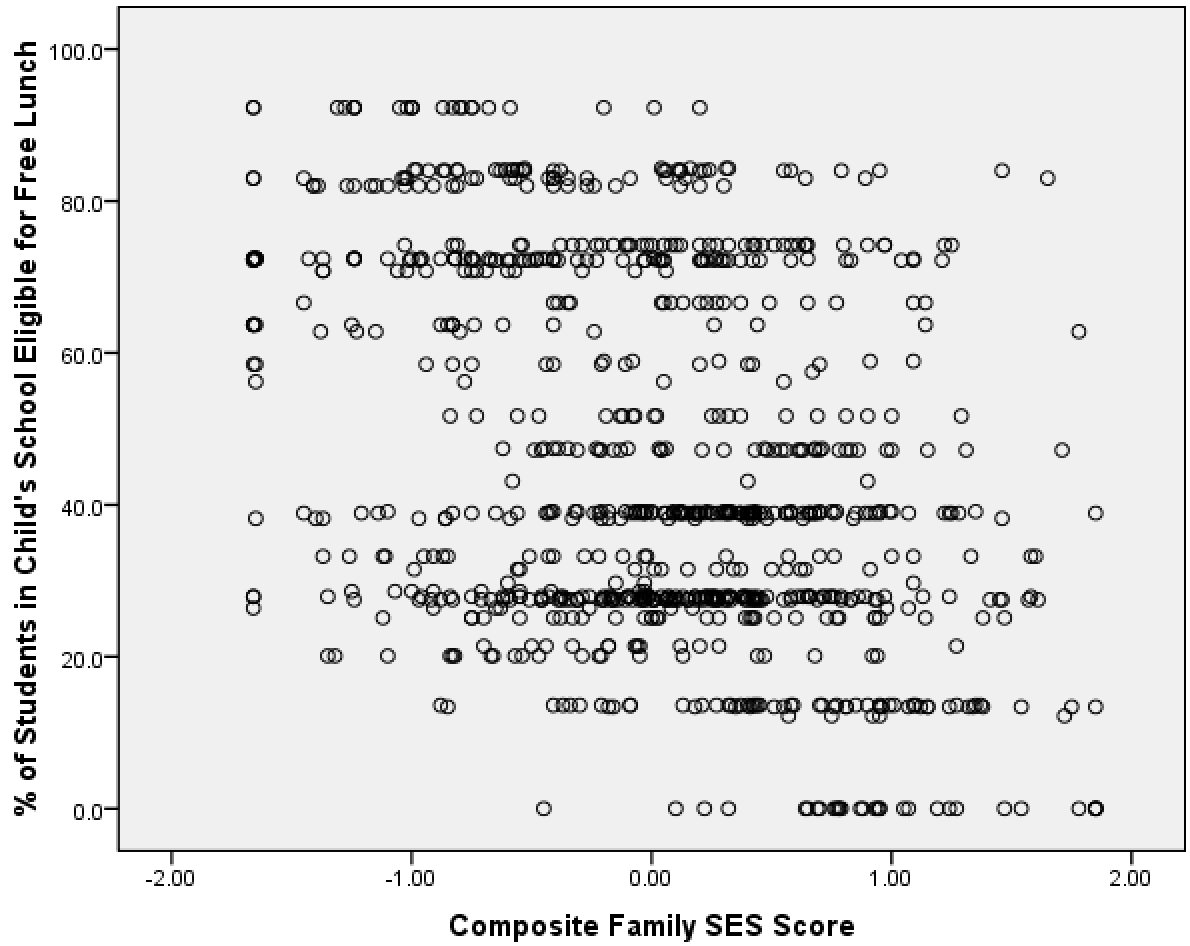
# Children of Immigrants (ChildrenOfImmigrants.sav)



# Children of Immigrants (ChildrenOfImmigrants.sav)



# Children of Immigrants (ChildrenOfImmigrants.sav)

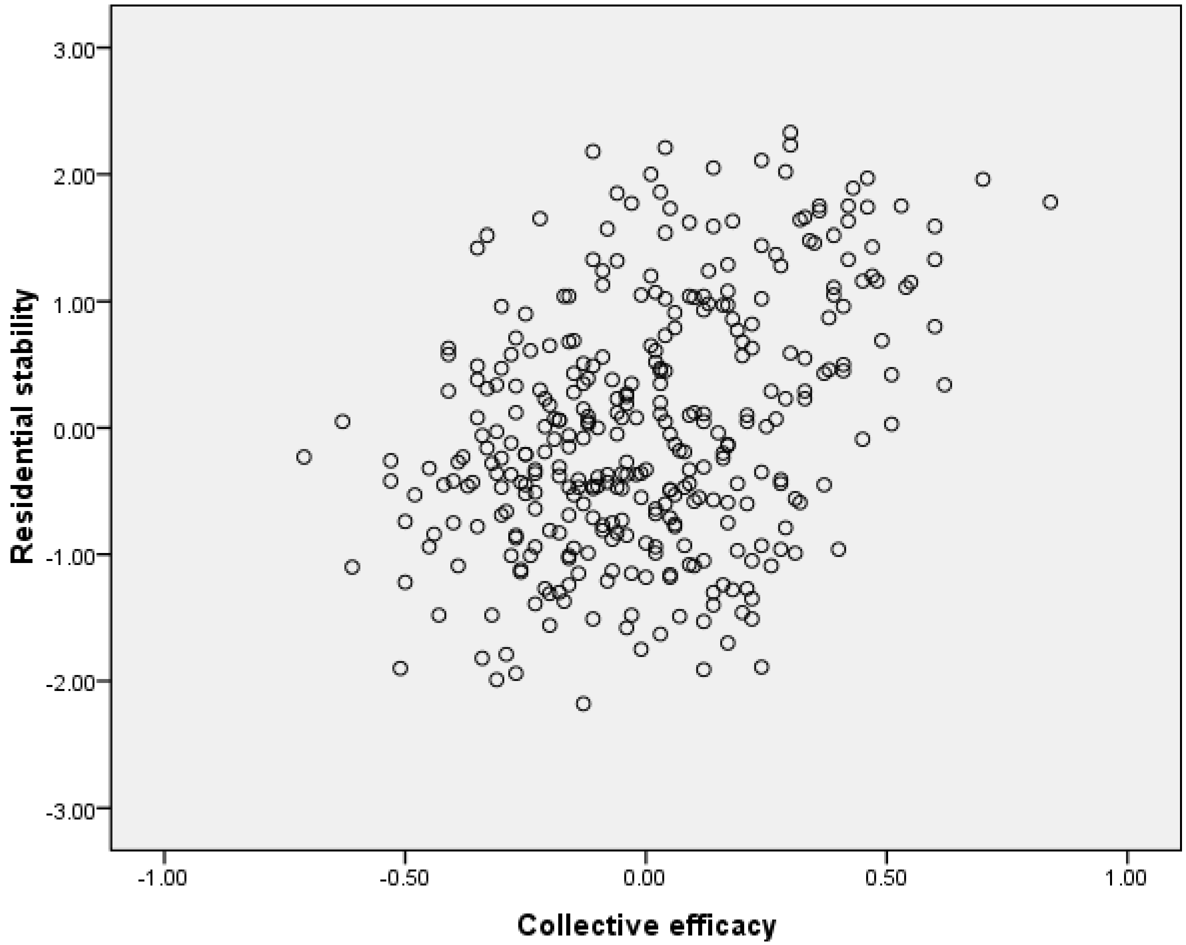


# Human Development in Chicago Neighborhoods (Neighborhoods.sav)

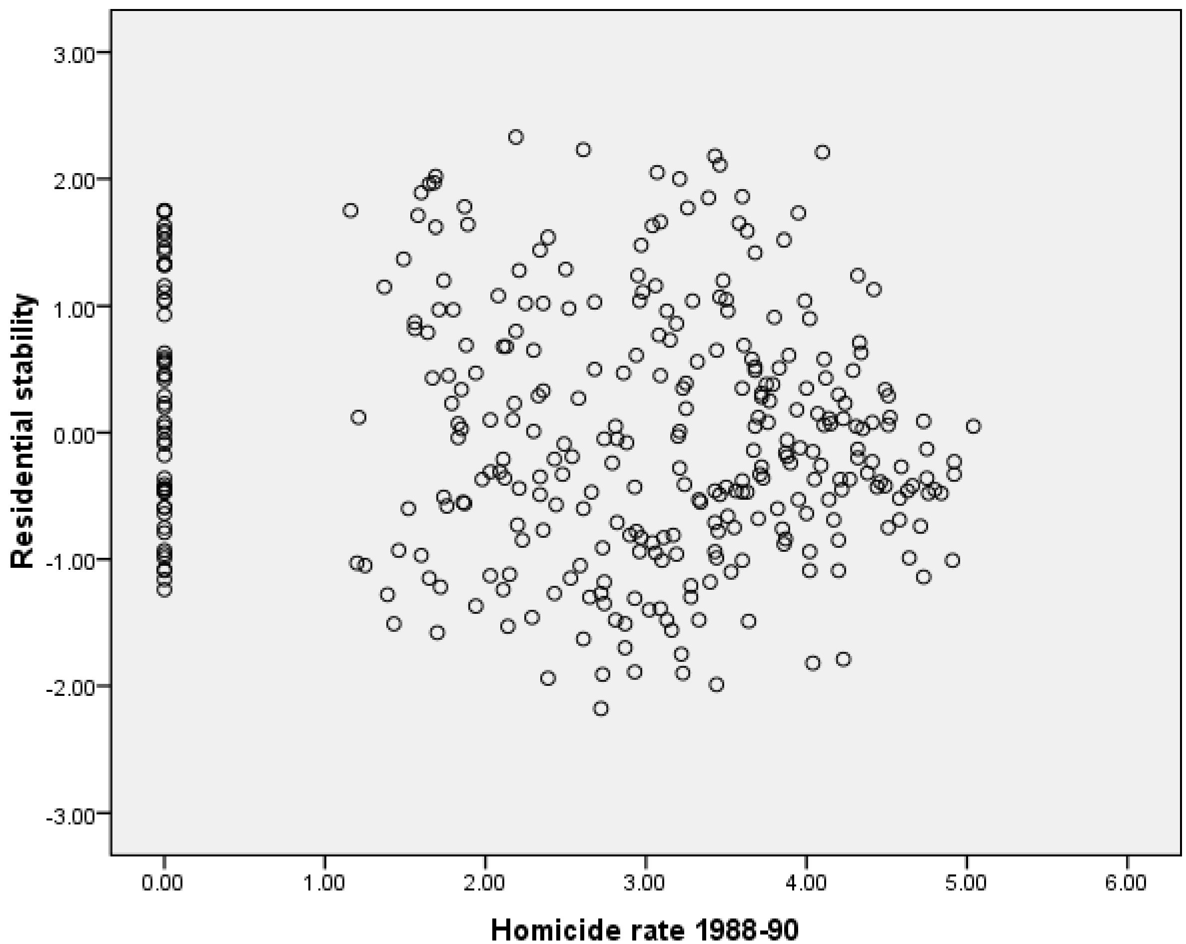
- **These data were collected as part of the Project on Human Development in Chicago Neighborhoods in 1995.**
- **Source: Sampson, R.J., Raudenbush, S.W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277, 918-924.**
- **Sample: The data described here consist of information from 343 Neighborhood Clusters in Chicago Illinois. Some of the variables were obtained by project staff from the 1990 Census and city records. Other variables were obtained through questionnaire interviews with 8782 Chicago residents who were interviewed in their homes.**
- **Variables:**

| | |
|---|---|
| (Homr90) | Homicide Rate c. 1990 |
| (Murder95) | Homicide Rate 1995 |
| (Disadvan) | Concentrated Disadvantage |
| (Imm_Conc) | Immigrant |
| (ResStab) | Residential Stability |
| (Popul) | Population in 1000s |
| (CollEff) | Collective Efficacy |
| (Victim) | % Respondents Who Were Victims of Violence |
| (PercViol) | % Respondents Who Perceived Violence |

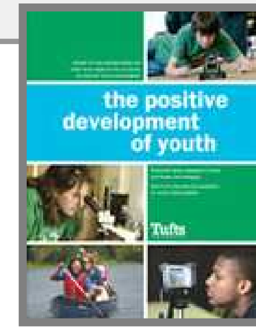
# Human Development in Chicago Neighborhoods (Neighborhoods.sav)



# Human Development in Chicago Neighborhoods (Neighborhoods.sav)



# 4-H Study of Positive Youth Development (4H.sav)

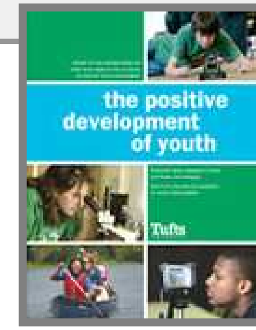


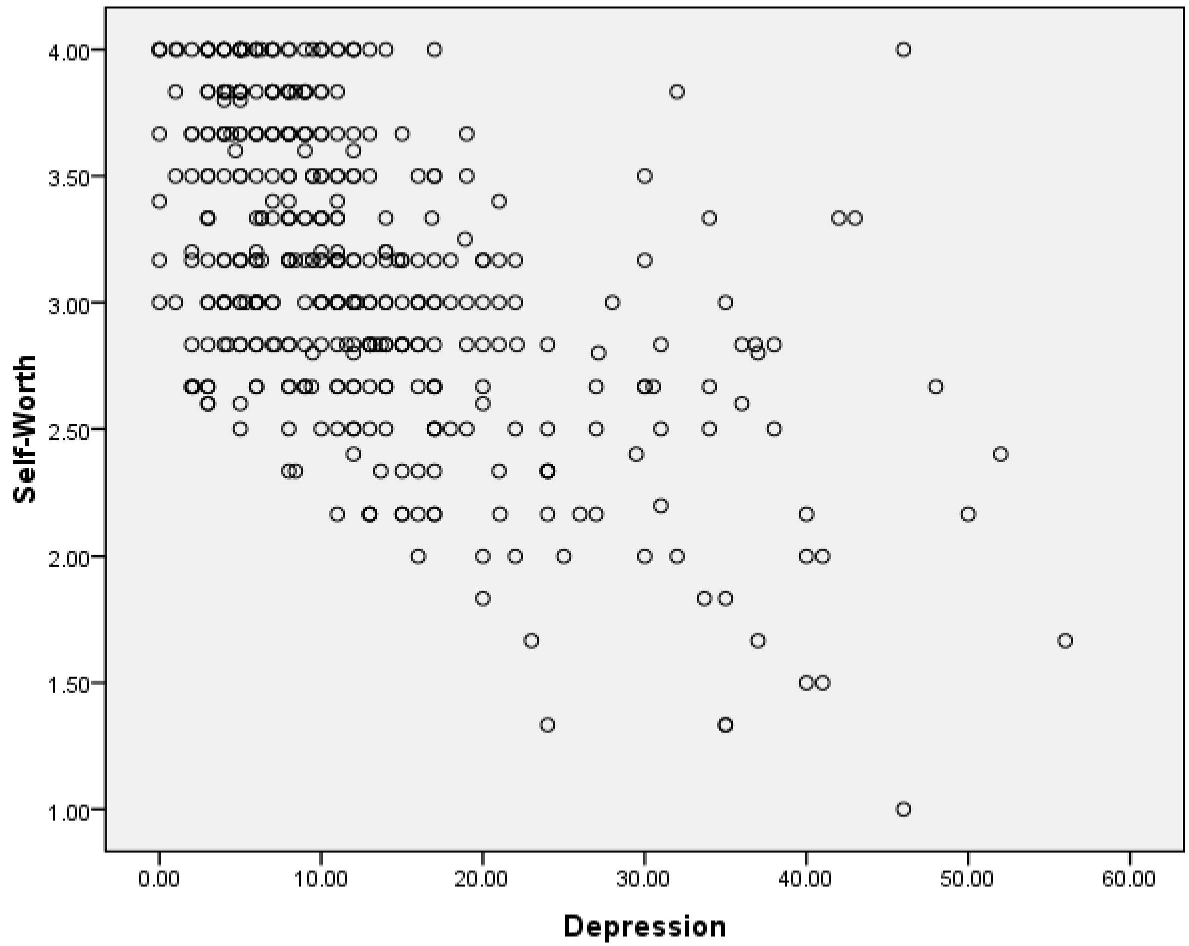

- **4-H Study of Positive Youth Development**
- **Source: Subset of data from IARYD, Tufts University**

- **Sample: These data consist of seventh graders who participated in Wave 3 of the 4-H Study of Positive Youth Development at Tufts University. This subfile is a substantially sampled-down version of the original file, as all the cases with any missing data on these selected variables were eliminated.**
- **Variables:**

| | | | |
|---|---|---|---|
| (SexFem) | 1=Female, 0=Male | (AcadComp) | Self-Perceived Academic Competence |
| (MothEd) | Years of Mother's Education | (SocComp) | Self-Perceived Social Competence |
| (Grades) | Self-Reported Grades | (PhysComp) | Self-Perceived Physical Competence |
| (Depression) | Depression (Continuous) | (PhysApp) | Self-Perceived Physical Appearance |
| (FrInfl) | Friends' Positive Influences | (CondBeh) | Self-Perceived Conduct Behavior |
| (PeerSupp) | Peer Support | (SelfWorth) | Self-Worth |
| (Depressed) | 0 = (1-15 on Depression)<br>1 = Yes (16+ on Depression) | | |

# 4-H Study of Positive Youth Development (4H.sav)



# 4-H Study of Positive Youth Development (4H.sav)